Cross-domain Relation Extraction

Elisa Bassignana

This thesis has been submitted to the Ph.D. School of the
IT University of Copenhagen on 2 May 2024

Committee

Advisor	Prof. Dr. B. Plank	Ludwig-Maximilians-Universität München
		IT Universitetet i København
Co-Advisor	Dr. R.M. van der Goot	IT Universitetet i København
Members	Prof. Dr. I. Augenstein	København Universitetet
	Prof. Dr. R. Klinger	University of Bamberg
	Dr. L. M. Aiello	IT Universitetet i København

Abstract

Language technologies are widely spreading over a diverse range of applications. Therefore, the ability of computational systems to easily adapt to new unseen situations is becoming more and more important.

In this thesis, we explore the task of Relation Extraction (RE) from a cross-domain perspective, in order to push the boundaries of model robustness across domains of application. RE is a key task in the automatic extraction of structured information from unstructured text. The goal of RE is the extraction of semantic triplets where two entities mentioned in the input text are connected by a semantic relation. The main challenge to the robustness of RE across domains is that depending on the downstream application the relevant information to extract differs (i.e., the entities and the types of semantic connections between them).

The work of this thesis covers the whole experimental pipeline for RE: First, given the lack of previous work in cross-domain RE, we outline several challenges characterizing the research area, from the scarcity of available resources for studying cross-domain RE, to the lack of standards in annotation guidelines and experimental settings. Second, to address the aforementioned challenges, we describe the creation of CrossRE, a multi-domain dataset for RE in English, and its subsequent expansion to 26 languages. Third, we propose two methodologies to boost the performance of RE in this multi-domain setup. Last, we present two frameworks for the analysis of the RE pipeline in terms of model performance and presence of socio-demographic biases.

Resumé

Sprogteknologier breder sig over en bred vifte af anvendelsesområder. Derfor bliver computersystemers evne til nemt at tilpasse sig nye, usete situationer vigtigere og vigtigere.

I denne afhandling udforskes Relation Extraction (RE) på tværs af domæner med henblik på at forøge sprogmodellers robusthed over forskellige anvendelsesområder. RE udgør et kerneelement i forhold til automatisk udtrækning af struktureret information fra ustruktureret tekst. Formålet med RE er at udtrække semantiske tripletter, hvor to enheder, der er nævnt i et givent input, er forbundet med en semantisk relation. Den største udfordring for robustheden af RE på tværs af domæner er, at de relevante oplysninger, der skal udtrækkes, varierer afhængigt af downstreamanvendelsen (dvs. enhederne og typerne af den semantiske forbindelse mellem dem).

Denne afhandling dækker alle eksperimentelle anvendelsesområder for RE: For det første, i betragtning af manglen på tidligere studier i RE på tværs af domæner, skitseres flere udfordringer, der karakteriserer forskningsområdet, fra knapheden af tilgængelige ressourcer, til manglen på ensrettede standarder for annotering og eksperimentelle design. For det andet beskriver vi oprettelsen af CrossRE, et multidomæne-datasæt til RE på engelsk, og dets efterfølgende udvidelse til 26 sprog for at imødegå de førnævnte udfordringer. For det tredje foreslår vi to metoder til at øge RE's ydeevne i denne multidomæneopsætning. Til sidst præsenterer vi to metoder til analyse af RE-forsøgsdesign med hensyn til modellernes resultater og tilstedeværelse af sociodemografiske bias.

Acknowledgements

... and that's a wrap!

Time flies and it seems surreal that I'm here writing the acknowledgments of my PhD thesis. It has been an incredible journey, sometimes challenging, but I'm grateful that many people were there supporting me all time along.

First of all, Barbara. You taught me how to think deeply about problems, how to be scientifically critical and how to formulate good research questions while following my own personal interest, which make me the researcher that I am today. Most importantly, you taught me how to always be optimistic, even in front of failures which inevitably happen every now and then. You taught me to patiently wait for the good times to come, while never stop trying. I will treasure all what I learnt from you in these years.

Rob, you joined the journey half way, and since then it could not have been more fun. Our endless discussions taught me how to learn from disagreement and the value in always seeking for a different (contrasting) perspective in order to strengthen my work.

Ancora una volta, ringrazio Viviana per avermi aperta al mondo della ricerca quando ero ancora giusto in triennale e da allora aver sempre tenuto un occhio, vicino o lontano, al mio percorso. Insieme a Viviana, ringrazio Valerio, perché loro due sono stati i primi in assoluto a credere in me come ricercatrice e per questo gliene sarò per sempre grata. Grazie a Valerio per aver revisionato la mia application (vincente) di dottorato, e a Viviana per averci messo una buona parola.

A big journey is made by everyday steps, and I want to thank my (past and present) colleagues from the NLPnorth for sharing this journey with me day by day: Dennis, Daniel, David, Marija, Maria, Joris, Kristian, Peter, Anna, Christian and Stefan.

During my PhD, I visited the TurkuNLP group at the University of Turku, and experienced the chill vibes of Finland. I really enjoyed my time there, and I thank the whole group for hosting me.

During my PhD, I was also lucky to get an internship at Alexa, Amazon.

Table of Contents

Chapter 1

Introduction

According to the Oxford English Dictionary the term "information explosion" was first introduced in the 1940s. In 2011, Hilbert and López (2011) reported that from 1986 until 2007 the world's technological capacity to store information grew from 2.6 to 295 exabytes. The advent of information technology has been a primary driver in the rapid increase of published information and data, usually referred as "information explosion". While the abundance of information means having access to more knowledge, it can be hard to retrieve the right information when needed. For textual data, which substantially contributes to the amount of available (digital) data, Natural Language Processing (NLP) technologies can help mitigate this challenge. Specifically, in response to the "information explosion" mentioned above, Information Extraction (IE) technologies able to automatically extract structured knowledge from unstructured text have been an active field of research in the last decades (Okurowski, 1993). One of the main challenges for these systems is that depending on the context in which they are used, the information which they are required to extract can vary a lot. For example, in the context of news articles, the information to extract may concern a person (e.g., where and when the subject was born). In the context of scientific papers, instead, the information to extract will likely include experimental setups and results. Therefore, it is important that IE systems are able to deal with the variability brought up by their
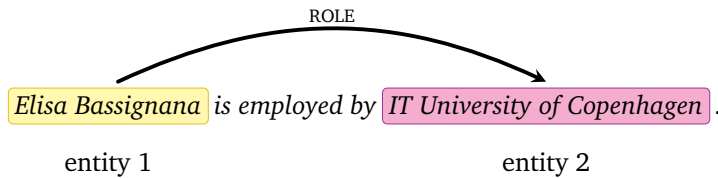
Figure 1.1: **Relation Extraction Example.** The goal of RE is the extraction semantic triplets from unstructured text. From the example above: `<Elisa Bassignana, ROLE, IT University of Copenhagen>`.

wide range of possible applications. Within current NLP methodologies, this variability can be addressed from two perspectives. Either with labeled data in every new domain of application for re-training the systems and adjust them to perform in the current context of interest. Or from the modeling perspective, by enhancing the models with the ability to generalize across the extraction of different types of information from data coming from new, previously unseen, contexts.

In the work performed this thesis, we focus on the task of Relation Extraction (RE), and address the robustness of systems against domain shift. RE is a specific case within IE which consists of extracting structured semantic triplets from unstructured text. As represented in Figure 1.1, the goal is to identify ordered pairs of entities (entity 1: *Elisa Bassignana*; entity 2: *IT University of Copenhagen*) and assign them a relation type from a set of labels (relation type: *ROLE*).

The task of RE is often employed in the pipeline of other Natural Language Understanding (NLU) tasks, in which it is relevant to extract the meaning of text. The most prominent is knowledge base population, where the knowledge base can be directly built from the extracted triples. But RE can also be an intermediate step for question answering or summarization, for example. The fact that RE does not have a final domain of application on its own, but rather it is employed as an intermediate step in other NLP tasks, adds up to the versatility of this task and the consequent relevance of the ability of RE systems to be able to adapt and generalize to many different domains of application.

1.1 Challenges

The task of RE presents many nuances and consequent challenges. Bellow we identify five of them:

1. **There are no standards in naming and experimental setups.** By design, RE includes multiple sub-tasks (i.e., identification of entities, identification of entity pairs semantically connected, and relation labeling). However, there are no standards about which sub-tasks the "Relation Extraction" naming should exactly include (Chapters 2, 5). As a consequence, the experimental setups are often misaligned across different research works, making it hard to fairly compare results (Chapter 5).

2. **There is no unified annotation standard.** Therefore, substantial misalignment arises in the annotation scheme adopted by different RE datasets (Chapters 3, 5).

3. **RE is domain dependent.** Depending on the domain of application, the relation types to extract (i.e., the label space) can vary a lot. Therefore, the ability of RE models to generalize across different domains of application is often a challenge.

4. **The domains covered by current datasets are limited.** As mentioned above, RE has many domain dependent nuances. However, while RE is an active field of research with a fair amount of labeled datasets for exploring the task, the domains covered by them are overall limited (Chapter 5). This restricts the representability of current datasets with respect to real-world applications (Chapter 6).

5. **Long tail relation types.** In addition to the domain specificity of relation types, often many of them are rare, independently of the domain of application. The limitation in the amount of instances makes it challenging for RE models to learn to identify them.

We will expand and address most of the challenges above in the next chapters. Despite being an influential challenge on the performance of RE, the long tail relation type issue is out of scope from the work of this thesis.

1.2 Chapter Guide

Below we describe the structure of the thesis, which is organized in four parts, and the research questions addressed in each of them.

1.2.1 Part I: Background

In the first part, we provide the foundational concepts of the thesis. These include a description of the RE task, how it is formalized, the methodologies for approaching it, and the evaluation metrics (Chapter 2). Then, we introduce the possible dimensions of variation in RE, along with common methodologies for tackling them, and the cross-domain RE setup investigated in this thesis (Chapter 3). The term "domain" is widely adopted in the NLP community, and a central concept in this thesis. However, it is loosely defined in the community (Plank, 2016) and typically used to refer to any non-typological property that influences model performance (e.g., genre, medium, style). In Chapter 3 we describe our use in this thesis of the term "domain" as *topic*, which is the variation dimension along which we investigate the robustness of RE models. Finally, in Chapter 4 we consider a bigger picture including both the interpretations as *topic* and as *genre* and explore human ability to identify domains and their agreement on the task. In this first part we investigate the following research question (RQ):

Part I

RQ1 To what extent can humans identify domains, and how much do humans agree on this task? (Chapter 4)

1.2.2   Part II: Data

In the second part of the thesis, we look at the cross-domain RE problem from the data perspective. In Chapter 5 we present a survey of RE datasets in order to find suitable options for exploring the cross-domain scenario. Within this chapter, we identify and carefully analyze challenges 1, 2 and 4 (see Section 1.1). As a response to challenge 4 ("The domains covered by current datasets are limited."), in Chapter 6 we introduce CrossRE, a novel multi-domain dataset for RE including six diverse text domains. In the development of the annotation guidelines of CrossRE, we aim to address challenge 2 ("There are no unified annotation standards.") by creating a unified annotation scheme able to cover all six domains included in the dataset. Last, in Chapter 7 we extend the previous work to a new dimension of variation, i.e., multi-linguality. We introduce Multi-CrossRE, a multi-lingual version of CrossRE including 26 new languages in addition to the original English. In the second part of this thesis we seek to answer the following research questions:

Part II

RQ2 Which challenges emerge by surveying and analyzing the landscape of existing RE datasets? (Chapter 5)

RQ3 What are important considerations to make when developing a unified annotation scheme for RE that covers multiple domains? (Chapter 6)

RQ4 When considering languages other than English, how does training and evaluating on automatically translated data influence the performance and the evaluation of RE? (Chapter 7)

1.2.3   Part III: Modeling

In the third part of this thesis, we inspect the cross-domain and the multi-domain training setups for RE. While in the former the domain of evaluation is excluded from the training, in the latter the domain of evaluation is included in the training data, along with other text domains (see Section 3.1.1). In the cross-domain setup, the model is trained on data from domains other than the domain of evaluation. Therefore, in Chapter 8 we look for a way to obtain low-cost silver data in a task related to RE to use for pre-training the RE model. In the multi-domain setup of Chapter 9, instead, the model is trained on data coming from different domains, including the one of evaluation. In this context, we compare different ways of encoding domain information in order to allow the model to learn useful commonalities between different domains, while still encoding domain specific knowledge. In the third part of the thesis we explore the following research questions:

Part III

RQ5 Can we exploit the affinity between semantic RE and syntactic parsing in order to obtain large amounts of (low-cost) silver syntactic data for pre-training RE models to improve the performance? (Chapter 8)

RQ6 How can we encode domain information in a multi-domain training setup, and how does it affect performance? (Chapter 9)

1.2.4   Part IV: Model Analysis

In the fourth part of the thesis, we propose two analysis suites for RE. In Chapter 10 we present a tool for quantitative analysis of the performance of Relation Classification models (RC: the task of classifying the relation

type between a pair of entities; see Chapter 2). We exploit the proposed analysis for investigating the performance of a state-of-the-art (SOTA) RC architecture over multiple cross-domain setups. Based on the findings of the analysis, with a simple and targeted improvement of the original architecture we achieve new SOTA performances.

Finally, in Chapter 11, we perform an analysis of biases in RE. As mentioned above, RE is often employed in the pipeline of other NLU tasks like knowledge base population. However, it has been demonstrated that NLP technologies are often affected by the presence of gender and racial biases (Kurita et al., 2019; Tan and Celis, 2019), which may emerge at any stage of the NLP pipeline (Hovy and Prabhumoye, 2021). In Chapter 11 we introduce a procedure for the analysis of socio-demographic biases both at the level of RE datasets and models. We also introduce a taxonomy of relation types for mapping the label sets of different RE datasets into a unified label space, and performing cross-dataset experiments (see issues about inconsistent annotation guidelines across RE datasets in Chapter 3). In the last part of the thesis we investigate the following research questions:

Part IV

RQ7 Is it possible to automatically identify groups of hard-to-handle cases for a SOTA RC model in order to increase the performance of cross-domain RC? (Chapter 10)

RQ8 To what extent is the RE pipeline (data and models) biased with respect to people's gender and origin? (Chapter 11)

1.3 Contributions

In summary, this thesis provides the following contributions:

1. We highlight **current challenges** and issues in the field of RE, in particular the lack of unified standards in annotation guidelines and experimental setups (Chapter 5).

2. The broadest **multi-domain and multi-lingual dataset for RE** to date, with a unified annotation scheme across six domains and parallel data in 27 languages (Chapters 6 and 7).

3. Two **methodologies for enhancing the performance of RE** in the cross-domain and in the multi-domain setups (Chapters 8 and 9).

4. A suite for **quantitative analysis** of the performance of RE models (Chapter 10).

5. A suite for the **analysis of socio-demographic biases** in RE (Chapter 11).

1.4   Publications

In this thesis, we include the following peer-reviewed publications:
(*): equal contribution

1. Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, Mike Zhang, and Rob van der Goot. Can humans identify domains? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2745–2765, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.245
Contribution: Framing the problem, data collection, writing.

2. Elisa Bassignana and Barbara Plank. What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification. In Samuel Louvan, Andrea Madotto, and

Brielen Madureira, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.7. URL https://aclanthology.org/2022.acl-srw.7

Contribution: Framing the problem, dataset survey, conduction of the experiments, writing.

3. Elisa Bassignana and Barbara Plank. CrossRE: A Cross-Domain Dataset for Relation Extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.263. URL https://aclanthology.org/2022.findings-emnlp.263

Contribution: Framing the problem, data collection, conduction of the experiments, writing.

4. Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot, and Barbara Plank. Multi-CrossRE A Multi-Lingual Multi-Domain Dataset for Relation Extraction. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 80–85, Tórshavn, Faroe Islands, May 2023a. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.9

Contribution: Framing the problem, conduction of the experiments, writing.

5. Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot, and Barbara Plank. Silver Syntax Pre-training for Cross-Domain Relation Extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6984–6993, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.

436. URL https://aclanthology.org/2023.findings-acl.436
Contribution: Framing the problem, data preparation, conduction of
the experiments, writing.

6. Elisa Bassignana, Viggo Unmack Gascou, Frida Nøhr Laustsen, Gustav
   Kristensen, Marie Haahr Petersen, Rob van der Goot, and Barbara
   Plank. How to encode domain information in relation classification.
   In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro
   Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the
   2024 Joint International Conference on Computational Linguistics,
   Language Resources and Evaluation (LREC-COLING 2024)*, pages 8301–
   8306, Torino, Italia, May 2024a. ELRA and ICCL. URL https://
   aclanthology.org/2024.lrec-main.728
   Contribution: Framing the problem, data collection, analysis of the
   results, writing.

7. Elisa Bassignana, Rob van der Goot, and Barbara Plank. What's wrong
   with your model? a quantitative analysis of relation classification. In
   Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the
   13th Joint Conference on Lexical and Computational Semantics (*SEM
   2024)*, pages 252–263, Mexico City, Mexico, June 2024b. Association
   for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.20.
   URL https://aclanthology.org/2024.starsem-1.20
   Contribution: Framing the problem, conduction of the experiments,
   writing.

8. Marco Antonio Stranisci*, Pere-Lluís Huguet Cabot*, Elisa Bassig-
   nana*, and Roberto Navigli. Dissecting Biases in Relation Extraction:
   A Cross-Dataset Analysis on People's Gender and Origin. In *Pro-
   ceedings of the 4th Workshop on Gender Bias in Natural Language
   Processing (GeBNLP)*, Bangkok, Thailand, 2024. Association for Com-
   putational Linguistics
   Contribution: Framing the problem, writing.

During the PhD, I was also involved in the following works which are not part of this thesis:

9. Elisa Bassignana*, Max Müller-Eberstein*, Mike Zhang*, and Barbara Plank. Evidence > Intuition: Transferability Estimation for Encoder Selection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4218–4227, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.283. URL https://aclanthology.org/2022.emnlp-main.283
Contribution: Framing the problem, conduction of the experiments, writing.

10. Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental Standards for Deep Learning in Natural Language Processing Research. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.196. URL https://aclanthology.org/2022.findings-emnlp.196
Contribution: Framing the problem, writing.

Part I

Background

Chapter 2

The Task of Relation Extraction

In this chapter we provide the fundamental concepts about the Relation Extraction (RE) task. We define a formalization of the problem, give an overview of methodologies for approaching RE, and conclude by describing the evaluation schema adopted in this research area.

2.1 Formalization

RE is defined as the task of extracting semantic relations connecting entities in a text (Bach and Badaskar, 2007). In traditional RE, the entities can either be named entities (e.g., *IT University of Copenhagen*) or mentions of named entities (e.g., *the university*). The semantic relations are picked from a predefined set of relation labels (e.g., *located-in*, *part-of*).[1] In this thesis we refer to binary RE, where the number of entities involved in the relation are always exactly two. Parallel work (mostly) in the biomedical domain has explored n-ary RE, where more than two entities can be involved in the relation (Peng et al., 2017). In binary RE the goal is to extract semantic triplets including the text span of entity 1, the text span of entity 2, and the relation label expressing the semantic connection between them (see example in Figure 2.1).

---

[1]Note that this is the key difference with respect to Open Relation Extraction, where there is not a predefined set of labels.

Dune tied with Roger Zelazny's This Immortal for the Hugo Award in 1966.
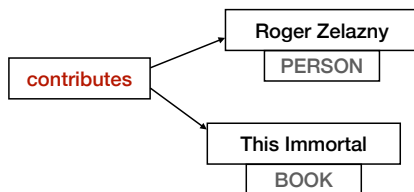
Figure 2.1: **Relation Extraction Example.** The goal of RE is the identification of semantic triplet expressed in a given text.

Formally, given a sequence of tokens $[t_0, t_1, ..., t_n]$ and two entity spans $s_A = [t_i, \ldots, t_j]$ and $s_B = [t_u, \ldots, t_v]$ with $0 \leq i, j, u, v \leq n$, $i \leq j$ and $u \leq v$, RE triples are in the form $\langle s_A, r, s_B \rangle$, where $r \in R$ and R is the predefined set of relation labels. Because of the directionality of the relations, $\langle s_B, r, s_A \rangle$ represents a different triple. We consider sentence-level RE, therefore in our setups the sequence of tokens correspond to one sentence. Next, we are going to present the experimental setups for extracting these triplets.

2.1.1   Relation Extraction Setups

RE is characterized by an intrinsic compositionality including the identification of the entities and the extraction of the semantic relations between them. This leads to the definition of three possible task-setups when approaching RE (see Figure 2.2):

End-to-end Relation Extraction.   Includes the whole pipeline from identifying the entity spans, and eventually assigning them a type (e.g., *person*, *location*), to identifying and classifying the semantic relations into types. End-to-end RE can be approached in two different ways. As a pipeline of tasks: First a Named Entity Recognition (NER) module to identify the entity spans, and then a RE module as described in Section 2.2. Or with an end-to-end architecture able to extract both entities and relations in one step.
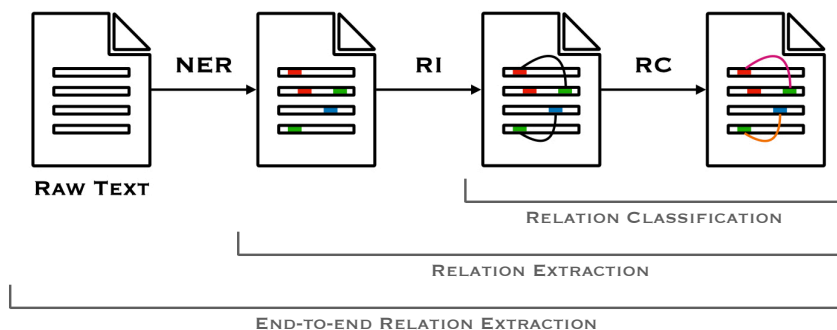
Figure 2.2: **Relation Extraction Pipeline.** The three steps composing the RE pipeline: Named Entity Recognition (NER), Relation Identification (RI) and Relation Classification (RC).

For example, using a table filling methodology (Miwa and Sasaki, 2014; Gupta et al., 2016), or a sequence to sequence approach (Huguet Cabot and Navigli, 2021).

Relation Extraction (RE). In this setup the entity spans are given. The task is to identify which of them are semantically connected and classify the relation between them into the given types.

Relation Classification (RC). Finally, a simplified version of the above is when the pairs of entity spans which are semantically connected are given, and the task (only) consists of the last step of the pipeline, i.e., assigning a relation type to those entity pairs.

Different standards used for naming the aforementioned setups have led to confusion in this research area with respect to the experimental design across different studies, and subsequent difficulties in fair comparison of the results. For example, the use of "Relation Extraction" for referring to the last "Relation Classification" step only (Cui et al., 2021), or at the opposite for referring to the whole end-to-end pipeline (Dixit and Al-Onaizan, 2019). We will discuss the issues related to the lack of standard in RE in more details in Chapter 5. In what follows, we will refer to the naming described above in order to distinguish the three setups. In the

17
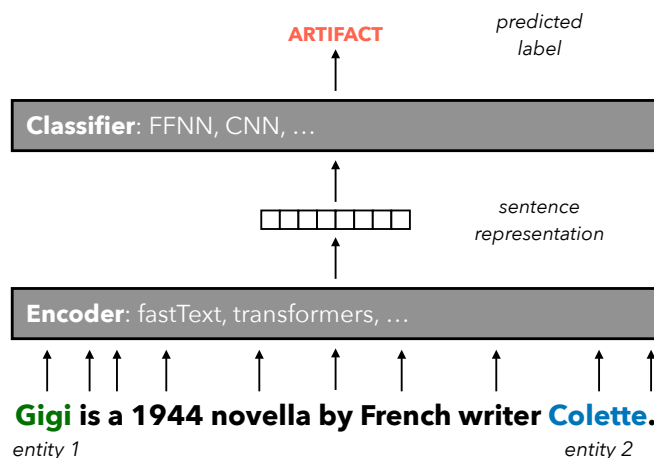
Figure 2.3: **Sentence-level Relation Extraction.** Abstraction of a traditional model architecture for sentence-level RE.

experimental papers included in this thesis we focus on RE (Chapter 8, 11) and RC (Chapters 5, 6, 7, 9 and 10), which will therefore be the focus of the next Sections 2.2 and 2.3 about methodologies and evaluation respectively.

2.2 Methodologies

In RE the goal is to predict a relation label conditioning on the two entity spans and the context around them.[2] The work of this thesis is focused on sentence-level RE, therefore the task is framed as a sentence-level classification task. Traditional architectures for sentence level classification are based on a two-module model architecture. One for encoding the sentence into a numeric representation (embedding), and the second for classifying the output of the first module (see Figure 2.3). The encoding can be performed using static embeddings—e.g., fastText (Bojanowski et al., 2017)—or, more recently, using contextualized word embedding obtained with transformer models —e.g., BERT (Devlin et al., 2019). The classification, instead, is traditionally done with a neural network—

---

[2]From a methodological perspective, RE and RC are approached in the same way, with the inclusion of the 'no-relation' class in the label set when approaching RE.

Figure 2.4: **Convolutional Neural Network for RE.** Visualization of the input of a CNN for RE as proposed by Nguyen and Grishman (2015b).

e.g., a Feed-Forward Neural Network (FFNN), or a Convolutional Neural Network (CNN). In order to specify the position of the entities in the sentence, two main approaches have been used in the field, namely position embeddings (often employed into CNNs) and entity markers. Whenever there are multiple entity pairs within the same sentence, the sentence is repeated multiple times with reference to the different entities. In what follows, we are going into the details of these two widely adopted model architectures which we use in the experimental papers included in this thesis: CNNs (Chapter 5), and the architecture with entity markers proposed by Baldini Soares et al. (2019) (Chapters 6, 7, 8, 9 and 10).

2.2.1 Convolutional Neural Networks

After the seminal work by Kim (2014) CNNs became a popular method for approaching sentence-level classification tasks. They are composed of three layers: A convolutional layer, a pooling layer, and a fully connected layer (FFNN). The first layer is where the core convolutional operation takes place. Here the aim is to optimize the filters (or kernels) used to

automatically extract the relevant features. The output of this layer is called 'feature map'. The second layer consists of a pooling mathematical operation used to reduce the dimensions of the feature map. There are two common types of pooling: *max* (takes the max values in a feature map) and *average* (computes the average across the values in the feature map). The last FFNN layer takes the processed input from the previous layers, and makes the classification.

Nguyen and Grishman (2015b) adapted the use of CNNs to the task of RE by concatenating two *position embeddings* to each word representation in order to provide information to the model about where the entity spans are.[3] Position embeddings are used to specify the relative position of elements in a text, for example the relative distance of a token from an entity. In RE, the standard is to include two position embeddings for indicating the relative distance of each token with respect to entity 1 and entity 2 respectively. They are randomly initialized and learned during training. Figure 2.4 visualizes the representation of the input of a CNN for RE, as described above.

2.2.2 Entity Markers

Baldini Soares et al. (2019) introduced the architecture represented in Figure 2.5, which became very popular for approaching RE (more than 800 citations on Google Scholar at the time of writing). In this approach, four entity markers are placed at the beginning and at the end of the two entity spans. They contain information about the directionality of the entity pair by specifying [E1] and [E2] (e.g., "[E1] *IT University of Copenhagen* [/E1]"). The markers are treated as special tokens by the tokenizer of the transformer encoder, meaning that they are not split into sub-words. They are randomly initialized and the model learns a representation of them during training. Zhong and Chen (2021) further proposed to enrich the markers with entity type information, if available (e.g., [E1:ORGANIZATION], [/E1:ORGANIZATION]). After the sentence is

---

[3]Note that these are not the 'position embeddings' as defined in the transformer models.
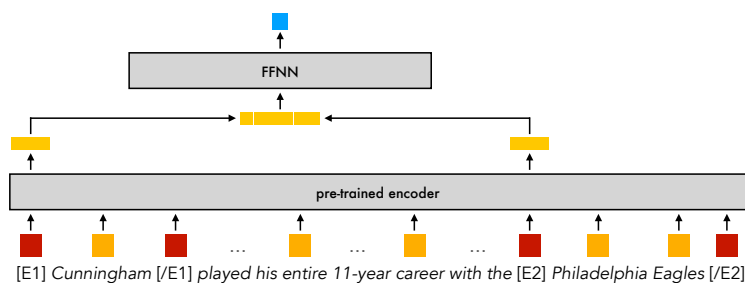
Figure 2.5: **Entity Markers Architecture.** In the architecture introduced by Baldini Soares et al. (2019) the entity spans are surrounded by the entity markers. After encoding the input, the classification is performed over the concatenation of the representations of the two start entity markers.

encoded, the classification into relation types is done with a FFNN over the concatenation of the representations of the two start entity markers.

2.2.3   Generative Information Extraction

While it is not the main focus of this thesis, it is worth mentioning the influence of the latest trend of generative Artificial Intelligence (AI) on the field of RE. In 2021, two works opened the way to sequence to sequence approaches for RE: Huguet Cabot and Navigli (2021), which is based on BART (Lewis et al., 2020) (employed in the experimental paper of Chapter 11), and Paolini et al. (2021), which is based on T5 (Raffel et al., 2020). The idea of these approaches is to input a query sentence and let the model output the list of triplets $\langle s_A, r, s_B \rangle$ following the format used for Instruction Tuning (IT) the base pre-trained language model. For example, Huguet Cabot and Navigli (2021) minimize the number of tokens to be generated in order to make the decoding more efficient:

Given the sentence:

"*This Must Be the Place*" *is a song by* new wave *band* Talking Heads, *released in November 1983 as the second single from its fifth album* "*Speaking in Tongues*"

And given the relation triplets:

(This Must Be the Place, performer, Talking Heads)
(Talking Heads, genre, new wave)
(This Must Be the Place, part of, Speaking in Tongues)
(Speaking in Tongues, performer, Talking Heads)

The triplets are encoded as:

`<triplet>` This Must Be the Place `<subj>` Talking Heads `<obj>`
performer `<subj>` Speaking in Tongues `<obj>` part of `<triplet>`
Talking Heads `<subj>` new wave `<obj>` genre `<triplet>` Speaking in Tongues `<subj>` Talking Heads `<obj>` performer

(Example from Huguet Cabot and Navigli (2021))

Where `<triplet>` marks the start of a new triplet and is followed by a new head entity; `<subj>` marks the end of the head entity and the start of the tail entity; `<obj>` marks the end of the tail entity and the start of the relation between the head and tail entities. The triplets are sorted by their order of appearance in the input text.

Even more recently, since ChatGPT (OpenAI, 2023) came out, a new wave in the field of generative RE has started. Several work has explored the potential of ChatGPT, and that of other generative Large Language Models (LLMs) in the context of RE (e.g., ChatGPT: Han et al. (2023); Li et al. (2023); Wei et al. (2023); Yuan et al. (2023); GPT-3 (Brown et al., 2020) and Flan-T5 (Chung et al., 2022): Wadhwa et al. (2023)). Two main differences characterize this second wave. First, LLMs have been used as a general framework for performing multiple IE tasks at the same time, including for example NER, RE, Event Extraction, and Aspect Level Sentiment Classification. Second, IT has been mostly replaced with In-Context Learning (ICL) methodologies (Pang et al., 2023). This is mainly

because of the increasing size of LLMs in terms of parameters, making it more challenging to fine-tune them, and because of the wide spread use of closed access models. However, at the time of writing there is no experimental evidence of the generative approaches overcoming the discriminative ones in terms of performance, especially when considering comparable resources (i.e., compute and training data). For example, the comparison by Meng et al. (2023) of gpt-3.5-turbo[4] against their proposed approach based on a BERT-base encoder (Devlin et al., 2019) in a few-shot document level RE setup, shows lower performance for the former.

2.3 Evaluation

RE and RC are traditionally evaluated using the Micro and Macro F1 scores. The former can be used to get a general idea of the overall performance of the model, while the latter is used for analyzing the model by taking into account the performance with respect to each individual class. Both of them are based on the *precision* and *recall* metrics:

$$\text{Precision} = \frac{\text{TP}}{TP + FP} \tag{2.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.2}$$

where

TP is the amount of true positive predictions

FP is the amount of false positive predictions

FN is the amount of false negative predictions.

The F1 score is then computed as the harmonic mean between precision and recall:

---

[4]https://platform.openai.com/docs/models/gpt-3-5-turbo

23

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.3}$$

As mentioned above, the Macro-F1 score uniformly averages the performance over every individual class without taking into account the label distribution:

$$\text{Macro-F1} = \sum_{i=1}^{n} \frac{\text{F1-score}_i}{n} \tag{2.4}$$

where

n is the number of relation types in the label set.

The Micro-F1 score, instead, is computed by aggregating the total number of TP, FP, FN across all relation types.

Chapter 3

Cross-domain Relation Extraction

In this chapter, we introduce the main topic addressed in this thesis: Cross-domain RE. We start by discussing two dimensions along which variation between training and evaluation can happen (i.e., input data and output space). Then, within the variation in the input data, we discuss how the concept of 'domain' is typically interpreted in the context of RE. Last, we conclude with an overview of methodologies to tackle the challenges related to the variation between training time and evaluation time.

3.1 Variation in Relation Extraction Setups

Variation between training and evaluation can be defined within two dimensions, as shown in Figure 3.1: Input data (training D, evaluation D') and output space (training Y, evaluation Y'). Variation in the input data means that the domain of the data changes from training time to evaluation time (more details about what is meant by *domain* in Section 3.2). Variation in the output space, instead, means that at least one between the label set and the annotation guidelines change from training to evaluation time (see examples of variation in Figure 3.2). The *situation (a)* in Figure 3.1 (often referred to as 'in-domain') represents the typical situation in which NLP

Input data (D)		
	stable	varies
stable	*situation (a)* D = D' Y = Y'	*situation (b)* D ≠ D' Y = Y'
varies	*situation (c)* D = D' Y ≠ Y'	*situation (d)* D ≠ D' Y ≠ Y'

Figure 3.1: **Variation Dimensions between Training and Evaluation.** Four possible experimental scenarios derived by the combination of variance in the input data (training D, evaluation D') and in the output space (training Y, evaluation Y').

models are trained and evaluated on a dataset assumed to stem from the same underlying data distribution. In the experimental applications of RE, the variation in output space can be caused by two factors. The first is the misalignment of the label sets and/or of the annotation guidelines across datasets collected from the same underlying distribution, which falls into *situation (c)*. For example, as we will discuss in more details in Chapter 5, the case of SemEval-2018 Task 7 (Gábor et al., 2018) and SciERC (Luan et al., 2018) which independently annotated the same data (NLP papers) using two different sets of relation types:

- SemEval-2018 Task 7: `compare`, `usage`, `part_whole`, `model-feature`, `result`, `topic`;

- SciERC: `compare`, `used-for`, `part-of`, `feature-of`, `evaluated-for`, `hyponym-of`, `conjunction`.

The second factor which can cause variation in the output space, instead, is the introduction of domain-specific relation types dependent on the input data, which falls into *situation (d)*. For example, if we train a model on the SemEval-2018 Task 7 dataset mentioned above (with textual data from scientific papers), and evaluate it on the CoNLL04 (Roth and Yih, 2004) dataset composed of news articles, which includes the relation types `kill`,
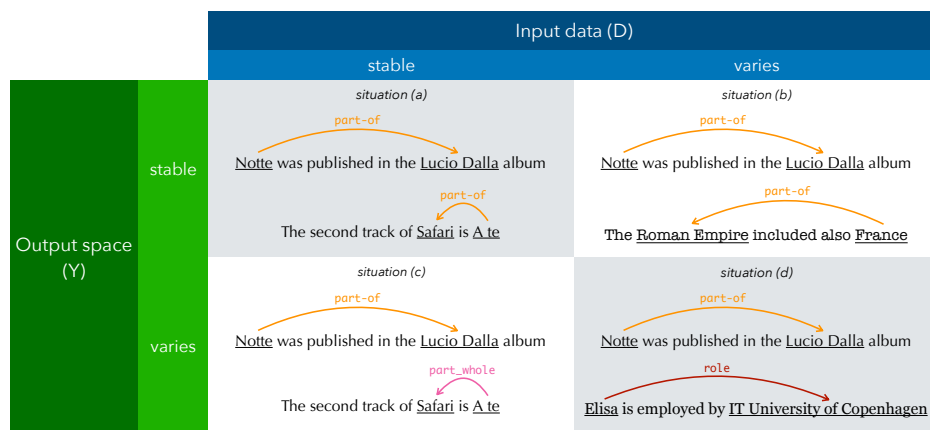
Figure 3.2: **Examples of Variation in Relation Extraction.** Examples of the four possible experimental scenarios derived by the combination of variance in the input data and in the output space.

`work-for`, `organization based-in`, `live-in` and `located-in`.

In the work presented in this thesis, we first point out at the lack of unified annotation standards in RE (see Chapter 5), and then in Chapter 6 we propose a unified annotation scheme that we employ for the annotation of the six domains included in CrossRE (Bassignana and Plank, 2022a). In all the subsequent experimental papers using CrossRE, we mainly focus on the variation within the input data.[1] As we will expand on in Chapter 6, our vision with the annotation of CrossRE is to set the base for the future creation of a "universal taxonomy" of labels, inspired by Universal Dependency (Nivre et al., 2016) for syntactic connections. CrossRE's label set includes relation types which are enough coarse-grained to be universally present in all the domains (e.g., `part-of`, `physical`). This setup falls into *situation (b)* in Figure 3.1, and sets a solid base for potential future expansions of the guidelines to include domain-specific types—i.e., *situation (d)* which is out of scope for the experimental part of this thesis. For example, in the 'music' domain `part-of` could be broken down into `song-part-of-album` and `musician-part-of-band`.

---

[1]The only exception is the paper included in Chapter 11 where we perform cross-dataset experiments by mapping different label sets into a unified taxonomy of relation types.
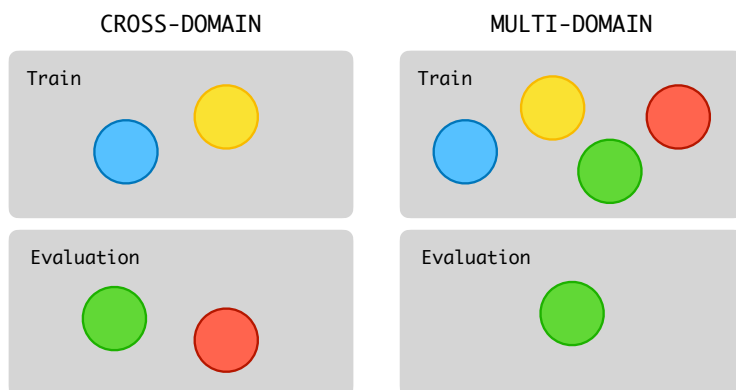
Figure 3.3: **Cross-domain versus Multi-domain setup.** In the *cross-domain* setup the target domain/s is/are not included in the training data; while it/they is/are included in the *multi-domain* training. Different colors represent different domains.

3.1.1 Cross-domain versus Multi-domain

Typically, *situations (b)* and *(d)* in Figure 3.1, where the input data varies from training time to evaluation, are referred to as *cross-domain* (see Figure 3.3 left). This indicates that the domain(s) included in the evaluation is/are not included in the training set. An orthogonal setup that we explore in Chapter 9, is the *multi-domain* training where the evaluation domain(s) is/are included in the training, together with other ones (see Figure 3.3 right). While this setup allows to maximize the training data by including data coming from multiple domains, the challenge lies in retrieving domain-specific information at inference time.

3.2 *Domains* in Relation Extraction

As we mentioned so far, variation in the input data is often referred to as *domains*. In NLP the term *domain* is used to refer to different kinds of variation which can characterize textual data (Biber, 1988). The concept is loosely defined in the field (Plank, 2016), but typically these variations are delineated, for example, with respect to the topic, the data source,

the medium or the style of a text. In Chapter 4 we include an in-depth discussion of the use of the term *domain* in the literature. In the paper included in Chapter 4 we consider the two widely adopted interpretations as 'topic' and 'source type', and present an extensive analysis exploring human misalignment when annotating for these two concepts.

In the context of RE, and broadly speaking in IE, the notion of *domains* typically refers to *topics*. This is because the topic of a text determines the information to extract (i.e., the label space). For example, within the 'music' domain we could have relation types connecting songs and artists, e.g., `<song, authored_by, artist>`, while in the 'news' domain we could likely find triplets as `<person, born_in, location>`. In order to measure the similarity between two domains and get a sense of the distance between them a standard way is computing the vocabulary overlap (Gururangan et al., 2020; Liu et al., 2021b; Bassignana and Plank, 2022b).

3.3 Tackling Variation in Relation Extraction

Following the two types of variations described in Section 3.1 which can occur from training to evaluation time, namely variation in the input data, and variation in the output space (see Figure 3.1), different methodologies have been used to tackle the different challenges.

3.3.1 Variation in the Input Data

Research in the field of Domain Adaptation (DA) which aims at developing methodologies for diminishing the negative effect that variation in the input data may introduce has a long history in NLP (Ben-David et al., 2006). The early work by Blitzer et al. (2006) introduces Structural Correspondence Learning (SCL) to automatically identify correspondences among features from different domains, and investigate its use in part of speech tagging. Later, Blitzer et al. (2007) adapted the SCL algorithm to sentiment classification. Daumé III (2007) proposes to augment the feature space of both the source and target data and use the result as input

to a standard fully supervised learning algorithm. Within the field of RE, previous work include feature-based systems (Nguyen and Grishman, 2014; Nguyen et al., 2014), requiring a few labels in the target domain; other work focuses on unsupervised DA methods (Plank and Moschitti, 2013; Fu et al., 2017). Plank and Moschitti (2013) propose syntactic tree kernels enriched by lexical semantic similarity to learn cross-domain patterns. Fu et al. (2017), instead, introduce a domain adversarial neural network to learn domain-independent representations.

Since the introduction of pre-trained language models, a common practice in DA is to include an additional training step between the traditional pre-training of the language model and the final fine-tuning on the target task. Because the data from the target domain annotated for the target task is usually scarce, the second pre-training step is meant to start adapting the language model towards a related task and/or towards a related domain. For example, Phang et al. (2018) introduce STILT (Supplementary Training on Intermediate Labeled-data Tasks), an intermediate training step on labeled data in a task for which ample data is available. Inspired by Phang et al. (2018), in the paper included in Chapter 8 we experiment with including an additional training step before fine-tuning on data labeled for RE in the target domain. We exploit the affinity between syntactic structure and semantic RE by considering the shortest dependency path between two entities. The additional training step is performed on silver syntax data in the target domain obtained using an out-of-the-box syntactic parser on unlabeled data (more details in Chapter 8)

Because labeled data is often difficult and expensive to obtain, different types of unsupervised domain adaptation approaches have been developed (Ramponi and Plank, 2020). For example, Gururangan et al. (2020) introduce two techniques where, similar to STILT, the base pre-trained language model is additionally trained, this time on a second masked language modeling objective. The first technique, DAPT (Domain Adaptive Pre-Training) uses a large corpus of unlabeled domain-specific text (e.g., biomedical, news, or reviews); while the second one, TAPT (Task Adaptive

Pre-Training) applies masked language modeling on the training set for a given task. These two approaches strikes a different trade-off: TAPT uses a smaller pre-training corpus than DAPT, but one that is much more task-relevant.

3.3.2 Variation in the Output Space

In the field of RE, before the widespread use of generative approaches (Huguet Cabot and Navigli, 2021; Xu et al., 2023), the way for dealing with new unseen relation types with the discriminative methods mainly consists of few-shot solutions. In this setup, a couple of instances with the new relation types are annotated or created ad-hoc. This is not the focus of this thesis, which mainly addresses the variation in the input space. However, it is worth mentioning that since the seminal work by Han et al. (2018) which introduced FewRel, the first RC dataset specifically designed for exploring few-shot RC, more work followed in this direction. Gao et al. (2019) published the FewRel 2.0 dataset, in which—by building upon FewRel— they added a new test set from a different domain (biomedical literature, while the original FewRel includes data from Wikipedia). Finally, Sabo et al. (2021) criticize the unrealistic (synthetic) evaluation setup of FewRel and FewRel 2.0 in terms of distribution of the labels in the datasets, and introduce a few-shot version of TACRED (Zhang et al., 2017b), which follows a real-world distribution of the relation types. More recently, generative AI has revolutionized the way of approaching few-shot and zero-shot RE, and the focus has been largely shifted towards prompt engineering strategies (Wei et al., 2023; Wadhwa et al., 2023), but their effectiveness is part of an on-going debate (see discussion in Section 2.2.3).

Chapter 4

Can humans identify domains?

The work presented in this chapter is based on the paper: Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, Mike Zhang, and Rob van der Goot. Can humans identify domains? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2745–2765, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.245

Abstract

Textual *domain* is a crucial property within the Natural Language Processing (NLP) community due to its effects on downstream model performance. The concept itself is, however, loosely defined and, in practice, refers to any non-typological property, such as genre, topic, medium or style of a document. We investigate the core notion of domains via human proficiency in identifying related intrinsic textual properties, specifically the concepts of genre (communicative purpose) and topic (subject matter). We publish our annotations in **TGeGUM**: A collection of 9.1k sentences from the GUM dataset (Zeldes, 2017) with single sentence and larger context (i.e., prose) annotations for one of 11 genres (source type), and its topic/subtopic as per the Dewey Decimal library classification system (Dewey, 1979), consisting of 10/100 hierarchical topics of increased granularity. Each instance is annotated by three annotators, for a total of 32.7k annotations, allowing us to examine the level of human disagreement and the relative difficulty of each annotation task. With a Fleiss' kappa of at most 0.53 on the sentence level and 0.66 at the prose level, it is evident that despite the ubiquity of domains in NLP, there is little human consensus on how to define them. By training classifiers to perform the same task, we find that this uncertainty also extends to NLP models.

Keywords: domain, genre, topic, multi-annotation

4.1   Introduction

The concept of "domain" is ubiquitous in Natural Language Processing (NLP), as differences between "sublanguages" have strong effects on model transferability (Kittredge and Grisham, 1986). This issue of domain divergence has prompted comprehensive surveys on how to best adapt language models (LMs) trained on one or more source domains to more specific targets (Ramponi and Plank, 2020; Ramesh Kashyap et al., 2021; Saunders, 2022), and remains an open issue, even with LMs of increasing

size (Ling et al., 2023; Singhal et al., 2023; Wu et al., 2023). Despite its importance, what constitutes a domain remains loosely defined, typically referring to any non-typological property that degrades model transferability. In practice, textual properties with the largest domain effects relate to a document's genre/medium/style (McClosky, 2010; Plank, 2011; Müller-Eberstein et al., 2021b), topic (Lee, 2001; Karouzos et al., 2021), or mixtures thereof (Aharoni and Goldberg, 2020). More broadly, domains can be viewed as a high-dimensional space with variation across the aforementioned properties, plus factors such as author personality, age, or gender (Plank, 2011, 2016).

We attempt to gain a better understanding of the foundational concept of domain, by taking a step back from modeling this phenomenon, and instead investigating whether humans themselves can distinguish between different instantiations of domain-related properties of textual data. In linguistics literature, these properties are separated into register, style and genre (Biber, 1988; Biber and Conrad, 2009, 2019), of which we choose to focus on *genre*, as it distinguishes itself from register and style by remaining consistent across complete texts. In addition, we examine the orthogonal factor of *topic*, i.e., the subject matter of a text, which can be expressed independently of genre (Kessler et al., 1997; Lee and Myaeng, 2002; Stein and Zu Eissen, 2006; Webber, 2009). We operationalize these two factors analogously to van der Wees et al. (2015) as genre stemming from different source types with distinct communicative styles, and topic being the principal subject matter of a given text.

More formally, our main research question is: *To what extent can humans detect genres and topics from text alone, and how does this align with machines?* We investigate the human proficiency in detecting these intrinsic properties by turning our attention to the Georgetown University Multilayer Corpus (GUM; Zeldes, 2017),[1] a large-scale multi-layer corpus consisting of texts from 11 different source types (henceforth *genre*). These act as gold annotations against which we compare the manual genre labels

---

[1] https://gucorpling.org/gum/
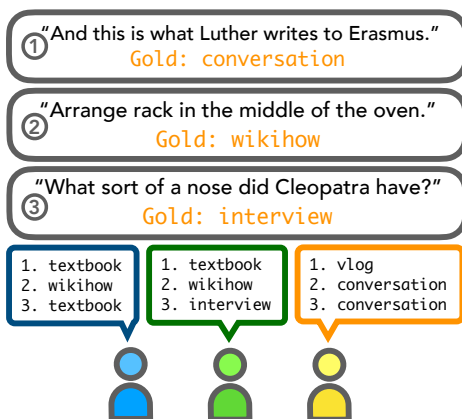
Figure 4.1: Graphical illustration of our triple-annotation setup with gold genre labels.

provided by 12 human annotators for the entirety of the corpus (Figure 4.1). In addition, the annotators supply a new annotation layer regarding the texts' subject matter (henceforth *topic*). As no gold labels are available for topic, they are annotated according to Dewey Decimal Classification (DDC; Dewey, 1979), a library classification system that allows new books to be added to a collection based on the subject matter. The DDC consists of 10 topics, 100 fine-grained topics, and 1,000 even finer-grained topics, of which we investigate the former two in detail and provide a preliminary study on the latter.

To understand the importance of context, we have annotators label genre and topic at both the sentence and prose level (defined as sequences of five sentences), and compare annotator agreement. Due to the subjective uncertainty associated with these types of characteristics, we gather three annotations per instance, measure their agreement, and release them in their unaggregated form as multi-annotations for future research.

Finally, we investigate the ability of machines to identify the same characteristics by training multiple ablations of genre and topic classifiers. Concretely, these experiments examine the difficulty of discerning each property, whether metadata or human notions of genre are more easily

recoverable, as well as which level of context is most appropriate for the different ways in which the genre and topic label distributions can be represented.

Overall, this work is the first to explore the discernability of domain by both humans *and* machines. In Section 4.5, we further discuss the implications of our findings, both with respect to domain-sensitive downstream applications, as well as for the NLP community's more general definition of domain. Our contributions thus include:

- **TGeGUM** (Topic-Genre GUM), a multi-layer extension of GUM, covering 9.1k sentences triple-annotated for a diverse set of 11 genres and 10/100 topics (Section 4.3).[2]

- An in-depth exploratory data analysis of the human annotations concerning annotator disagreement, uncertainty, and overall trends for domain characteristics across different context sizes (Section 4.4).

- A case study on the capability of NLP models to discern the human notions of genre and topic, as well as an analysis of which factors affect classification performance (Section 4.5).

4.2 Related Work

Domains Initially coined as "sublanguages" (Kittredge and Lehrberger, 1982; Kittredge and Grisham, 1986), domains have long been a topic of study in traditional linguistics and NLP (Lee, 2002; Lee and Myaeng, 2002; Stein and Zu Eissen, 2006; Eisenstein et al., 2014; van der Wees et al., 2015; Plank, 2016). Some of the early work mentioning domains as textual categories include Sekine (1997); Ratnaparkhi (1999), which categorize texts into, e.g., "general fiction", "romance & love", and "press:reportage". However, as also mentioned by Lee (2002); Lee and Myaeng (2002); Plank (2011); van der Wees et al. (2015), the concept of domain is under-defined. Plank (2011) considers domains as a multi-dimensional space, spanning

---

[2]Data and code can be found at `bitbucket.org/robvanderg/humans-and-domains`.

all kinds of variability between texts, such as genre, topic, style, medium, etc. In this work, we follow a definition of domains similar to van der Wees et al. (2015), focusing on two of the largest dimensions of variability: i.e., *genres* (the communicative purpose and style) as well as *topics* (the subject matter). The former is closely tied to the source of a text, such as academic papers versus fiction books, while the latter may include subjects such as sports, politics, and philosophy, which can occur in multiple genres.

Automatic Domain Detection   In NLP, automatic domain detection is essential for ensuring robust downstream performance, as it degrades with increasing levels of domain shift (Ramponi and Plank, 2020). Since this issue occurs independently of the application, domain classification has been explored in many contexts. Generally, the problem is either phrased in terms of a binary task, i.e., whether a target text matches the domain of the training data or not (e.g., Tan et al., 2019; Pokharel and Agrawal, 2023), or a multi-label classification task, in which the exact domain is to be determined (e.g., Müller-Eberstein et al., 2021a). Here, we use the latter approach as it requires a more formalized operationalization of domain.

At a broader level, genre is frequently used as a proxy for domain, as it has lower internal variability than many more specific dimensions, including topic (Kessler et al., 1997; Webber, 2009). Its automatic detection has been leveraged for selecting training data for transfer learning across a broad range of applications, such as classification (Ruder and Plank, 2017; van der Goot et al., 2021a; Gururangan et al., 2020) and generative tasks (Aharoni and Goldberg, 2020). Beyond English, genre has further been shown to provide a cross-lingually consistent signal for enabling more robust transfer in syntactic parsing (Müller-Eberstein et al., 2021a).

Topics provide a more granular differentiation between texts, also with close ties to domain. Automatically detecting topics has more immediate practical implications, as knowledge of the subject matter is critical for many downstream information extraction systems (Liu et al., 2021b; Bassignana and Plank, 2022a) and more datasets with topic annotations are available (Sandhaus, 2008; Maas et al., 2011; Wang and Manning,

2012; Zhang et al., 2015); however, these works typically contain source data from only a single corpus.

Going beyond prior work with limited sets of post-hoc topic labels for single-genre corpora, we build on the general-purpose DDC system (Dewey, 1979) for libraries and apply its hierarchical set of 10/100 topics to a corpus containing data from 11 genres. By building on the existing annotations of the GUM dataset (Zeldes, 2017), we further enable research not only ascertaining to domain classification for its own sake, but also with applications to other downstream NLP tasks.

Multi-annotations   Given the subjective nature of domains and their associated properties of genre and topic, each text in our dataset is annotated multiple times and retains individual labels without aggregating them. This approach of *multi-annotations* (Plank, 2022) avoids obscuring human uncertainty in the annotation process and has benefits both for tasks with high variability, such as ours, as well as tasks for which a ground truth is typically assumed.

E.g., Plank et al. (2014) map part-of-speech (POS) tags from Gimpel et al. (2011) to the universal 12-tag set by Petrov et al. (2012), retaining five *crowdsourced* POS labels per token.

For Relation Classification (RC), Dumitrache et al. (2018) obtained annotations for 975 sentences for medical RC, where each sentence is annotated by at least 15 annotators on average.

For Natural Language Inference (NLI), Nie et al. (2020) released ChaosNLI: A dataset with 4,645 examples and 100 annotations per example for some existing data points in the development set of SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018a), and Abductive NLI (Bhagavatula et al., 2020). For a more in-depth overview of multi-annotation datasets, we refer to Uma et al. (2021b).

4.3 The Dataset

4.3.1 Source Data

The source dataset on top of which we build our domain-related anno-
tations is the GUM corpus which in turn incorporates data from a wide
variety of sources. We use the portion of the GUM corpus released as
part of the Universal Dependencies project (UD; Nivre et al., 2017), i.e.,
excluding Reddit. Since a text's source is closely tied to its communicative
purpose, we consider GUM's *data source* metadata field of each instance as
the gold genre label. For the topic, no equivalent gold label is discernible
from the metadata.

The entire dataset is annotated both at sentence and prose level to
investigate the importance of context for genre and topic annotation. For
this purpose, we follow the gold sentence segmentation provided by GUM.
We opted for these blocks instead of paragraphs, as the latter are not natural
dividers for all text types and can have a high variety of conventions and
functions across genres. To avoid the same annotator observing the same
sentence individually as well as in prose, we shuffle the dataset such that
annotations of a sentence with and without context are distributed across
different annotators, while maintaining coverage of the full dataset.

4.3.2 Annotation Procedure

Since there are no official descriptions of the genres in GUM, our an-
notation guidelines refer to the descriptions from the homepages of the
websites of the source or the corresponding abstracts from Wikipedia.
For topic annotation, we follow the Dewey Decimal library classification
system (Dewey, 1979) consisting of 10/100/1,000 hierarchical topics of
increased granularity. We consider the 10 high-level and the 100 mid-level
classes for the coarse- and fine-grained topic annotations. We constrain
our guidelines such that topic-100 should always be a sub-type of topic-10.
For example, if topic-100 is "520 Astronomy", then topic-10 should be
"500 Science". When none of the topic-100 labels fit the fine-grained topic

of the instance, the annotators were allowed to leave the more specific topic blank, i.e., annotating topic-100 with the same label as topic-10. In addition, we include the *no-topic* label for when it is not possible to identify a specific topic from the provided text., such as for very short sentences, like "Ok" or "I agree with that."

We completed an initial annotation round of 20 instances with all annotators and authors of this paper to evaluate the guidelines and annotation setup. None of this data is included in the final dataset. We continued with groups of three annotators annotating different subsets of the data. After an introductory meeting, further unclarities were discussed asynchronously throughout the process. Annotators were asked to pose their questions in general terms and to not use direct examples as to not bias the other annotators on specific instances. We did not conduct inter-annotator studies over the course of annotation and only had minor guideline revisions during the annotation process since we are mostly interested in human intuitions of genre and topic, and there are no gold labels for the topic task.

Annotators could indicate whether they were unsure about the annotation of a specific instance, and were also asked to provide notes/comments, if applicable. The annotation rate started at approximately 80–150 instances per hour. To ensure a similar amount of effort across annotators, we asked them to aim for approximately 150 instances per hour (also considering that annotation speed increases over time).

In total, we hired 12 annotators, who were paid 34,21 EUR per hour (before tax) for a total of 32 hours per person over a period of 4 weeks. The mean age was 27 ($\pm$2), and their highest completed education was equally split between a bachelor's and a master's degree. All rated their English skills as either C2/proficient or native. Seven annotators were reported to be female, three male, and two other/non-binary.

	Instances		Annotations	
	Sentence	Prose	Sentence	Prose
Train	6,911	1,358	20,733	4,074
Dev.	1,117	217	3,351	651
Test	1,096	221	3,288	663
Total	9,124	1,796	27,372	5,388

Table 4.1: Dataset Statistics: Note that each instance has three associated annotations.

4.3.3 Dataset Statistics

Table 4.1 shows the final dataset statistics of **TGeGUM**. The dataset includes around 9.1K sentences, and 1.8K prose, each of them annotated by three individual annotators for genre, coarse-grained topic, and fine-grained topic.

In Figure 4.2, we report the sentence-level distribution of gold labels and human annotations, reporting the average number of annotations per label (total number of annotations divided by three annotators) to align with the singular gold genre metadata. For topic-10 and topic-100 we only report the human annotations as no gold labels exist.

Comparing gold and annotated genre labels, we observe a skew towards *conversation* and *textbook*. We hypothesize that this is due to the small amount of context an annotator receives. For example, the sentence "Is that all that's left?" with the gold genre label *fiction* is annotated by all annotators as *conversation*. Another example is the sentence "Some of the greatest poetry has been born out of failure and the depths of adversity in the human experience." with gold label *interview*. All annotators annotated this example as *textbook*.

For topic, we note that despite skewness, almost all 100 topics are used. The *300 Social sciences* including, e.g., *320 Political science* and *370 Education*, stand out as being the most prevalent topics. The most frequent label, however, is *no-topic*, indicating that it is challenging to identify a
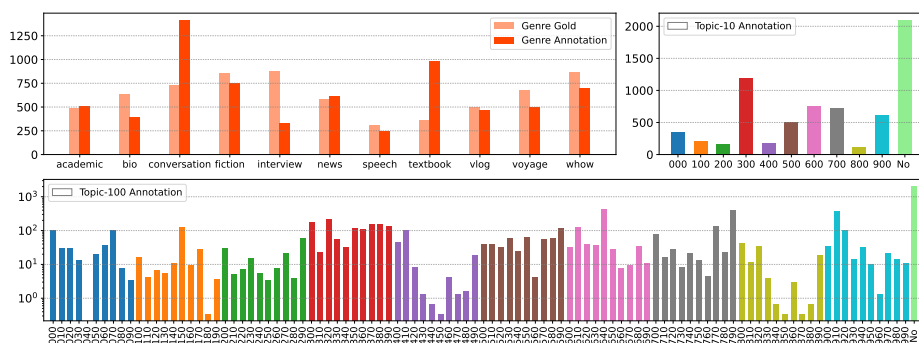
Figure 4.2: Frequency distributions of the labels in gold genre labels, annotations of genres, annotations of topic-10, and annotations of topic-100 (log scale) on sentence level. For the human annotations, the number is divided by three in order to align with the (unique) gold label. The mapping of topic-10 and topic-100 labels can be found in **??**. The tag "No" in the topic annotations refers to *no-topic*.

specific topic given only one sentence and that individual sentences can be associated with different topics, depending on the surrounding context.

The genre distribution at the prose level (Subsection 4.7.4) reveals a more accurate distribution for *conversation*-like utterances; however, the general skew towards *textbook* remains. Concerning topic, the main contrast to the sentence-level distributions is the reduction of the *no-topic* label, confirming that more context is crucial for this task.

4.4 Exploratory Data Analysis

In addition to the previous aggregated overview, we are interested in exploring whether domain characteristics are recoverable by humans in a consistent manner. While we can compare human annotations to the original gold labels for genre, no equivalent is available for topic. Therefore, we place more emphasis on inter-annotator agreement, in the form of Fleiss' Kappa (Fleiss, 1971), to measure intuitive alignment and ease of identification. Table 4.2 and Figure 4.3 shows this agreement across the different genres, topics and levels of available context.

		Kappa		Maj. Acc.
	Genre	Topic-10	Topic-100	Genre
Sentences	0.5260	0.5213	0.4239	67.68
Prose	0.6582	0.5238	0.3838	81.11

Table 4.2: Agreement scores across annotators, and accuracy of majority vote among annotators compared to gold genre labels.

4.4.1 Human Genre Detection

Accuracy and Agreement   Considering that annotation guidelines were phrased to avoid any intentional alignment to an existing ground truth (i.e., annotators were unaware of the existence of gold genre labels), an accuracy of 67.68% at the sentence level shows that genre is recoverable to a far higher degree than by random chance or by a majority baseline. This further increases to 81.11% given more context at the prose level and is also reflected in the increase from moderate inter-annotator agreement (0.53) to substantial agreement (0.66).

The additional context appears to help differentiate genres that have more similarities to each other. This phenomenon is especially pronounced for spoken-language data, such as *conversation*, *interview* and *vlog,* which differ with respect to genre-specific conventions such as who the speech is directed towards (i.e., bi-directional, interviewee, video viewer), or how formal the register is. Both properties are more easily discernible across multiple turns.

Nonetheless, even given more context, high amounts of confusion remain between certain genres such as non-fiction texts of the type *academic*, *biography*, and *textbook*. These are intuitively similar to each other and may require even more context to distinguish. Generally, genres appear to lie on a more continuous spectrum that is difficult to discretize in conceptually similar cases.

Human Uncertainty   In case of uncertainty, annotators were encouraged to select a "best guess" label and to indicate uncertainty by ticking a check-

44
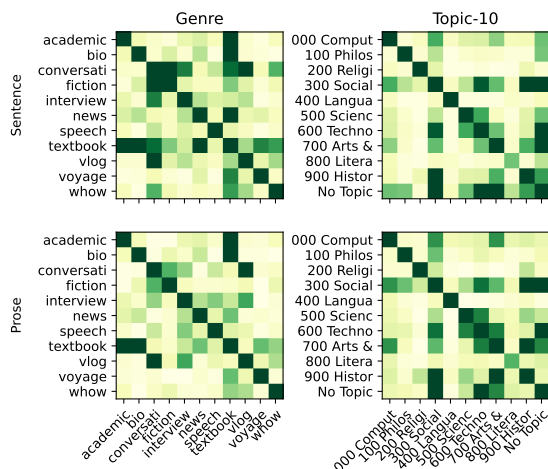
Figure 4.3: Confusion matrix with all annotated pairs of labels for Genre and Topic-10 (across all annotators) in our training data: The darker the color, the higher the number of annotations for that label pair. The diagonal can be seen as agreement, whereas off-diagonal is a proxy for disagreement.

box. In addition to overall uncertainty, we also hypothesize that sentence length affects accuracy due to the amount of information available. To evaluate these two effects for genre detection, we measure the Pearson correlation between human accuracy concerning the gold label, with 1) sentence length, 2) the number of uncertainty flags (Table 4.3). As expected, longer sentences are annotated correctly more often. Figure 4.4 further highlights how spoken-language genres have a strong skew towards shorter sentences, and for which annotators have the lowest agreements. Additionally, sentences marked as "unsure" align with gold labels less often, showing that annotators appear to have well-calibrated judgments of their own uncertainty, even for this relatively difficult task.

4.4.2   Human Topic Detection

Agreement   In the absence of gold labels, inter-annotator agreement allows us to estimate the difficulty of discerning broader vs granular topics.

	Sent	Prose
length vs unsure	-0.1126*	-0.0474
length vs correct	0.1267*	-0.0385
unsure vs correct	-0.2948*	-0.3411

Table 4.3: Correlations across utterance length, correct predictions of human majority vote, and the number of unsure annotations. * indicates statistical significance for $p < 0.05$.

For the 10 broader topics, Table 4.2 shows a moderate agreement of 0.52 for both the sentence and prose levels. As expected with an order of magnitude more labels, Topic-100 sees a drop in agreement to 0.42 and an additional drop to 0.38 at the prose level. While this may seem counter-intuitive due to topic's higher specificity compared to genre, Figure 4.3 sheds some light on this peculiarity: In contrast to genre, topic has a *no-topic* label (Subsection 4.3.2), which, in turn, is used frequently by all annotators at the sentence level, due to the absence of any subject matter in many shorter utterances—especially in speech. Given the additional context, topic becomes more apparent, and agreement spreads toward more topics along the diagonal. As such, sentence-level agreement mainly hinges on *no-topic*, while prose-level annotations agree more with respect to actual topics. This is less apparent for 10-topic kappa, for which this effect cancels out, but is more prevalent with 100 topics, where the shift away from *no-topic* at the prose level comes with a much wider spread of topics, thereby reducing overall agreement, despite having a higher level of true topic annotations.

Overall, topics which were most consistently identified include *social sciences*, *arts & recreation*, *technology*, *science* and *history & geography*. On the other hand, *literature* was least consistently annotated and most frequently confused with the aforementioned topics, potentially due to its broader scope compared to the others.
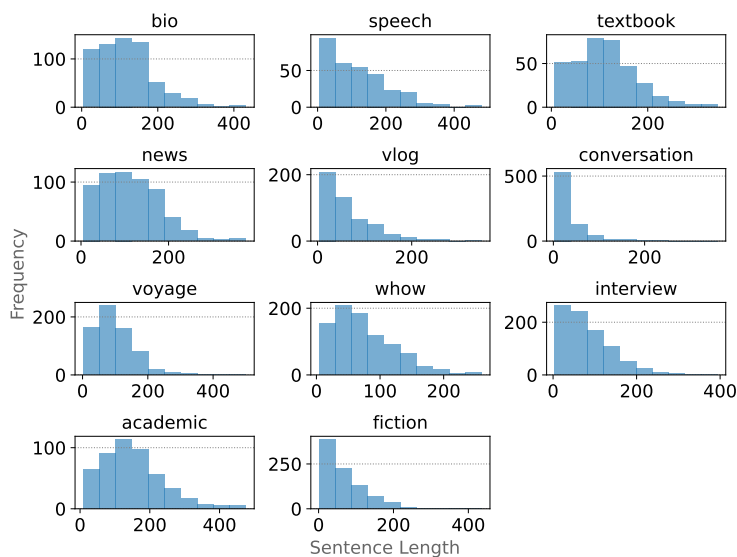
Figure 4.4: Frequency of sentence lengths, measured by the number of characters, per gold genre.

1,000 Topics   After completing the full set of genre and topic-10/100 annotations with three annotators per instance, the remaining time of the annotators was spent on a preliminary study to label the most fine-grained categories of DDC. With 1,000 labels, this task is substantially more difficult. We obtained a total of 904 sentences and 172 prose sequences with three annotations each.[3] Measuring inter-annotator agreement at this level of granularity, we find a Fleiss' Kappa of 0.32 for sentences and 0.26 for prose. Although substantially lower than for coarser topic granularities as well as genre, this score still indicates above-random agreement among annotators. Similarly to the previous topic results, prose-level context allows humans to detect more actual topics than *no-topic*, leading to lower overall agreement but a broader coverage of actual topics.

    In general, despite the importance of topic to downstream applications (i.e., topic classification as a task in itself), there is no clear human con-

---

[3]From 3,918 total annotations, we discarded instances with less than three completed annotations.

sensus regarding discrete topic classification. Similarly to genre, topic appears to be a concept for which human intuition shares some agreement at a broader level, but is also spread along a continuum—especially as granularity increases.

4.5   Modeling Domain

Following our examination of human notions of genre and topic, we investigate automatic methods' ability to model the same properties. Ablating across different setups for representing the multiple annotations per instance (Subsection 4.5.1), we train models to classify genre and topic at different levels of granularity (Subsection 4.5.2) and evaluate their ability to learn the underlying distribution (Subsection 4.5.3). While pre-neural work typically performed document-level classification (Webber, 2009; Petrenz and Webber, 2011), contemporary trends have shifted towards the sentence-level (Aharoni and Goldberg, 2020; Müller-Eberstein et al., 2021b). Leveraging our multi-level annotations, we investigate genre and topic classification at both the sentence and prose-level, mirroring our human annotation setup.

4.5.1   Setup

Most work on modeling multiple annotators is based on tasks consisting of only two or three labels, e.g., hate speech detection, or RTE (Uma et al., 2021b). An exception is Kennedy et al. (2020), who use multiple classification heads to predict a score for a variety of aspects of hate speech, which are then used to predict a final floating point score for hate speech detection. Other related work predicts multiple task labels simultaneously (e.g., Demszky et al., 2020; Kiesel et al., 2023; Piskorski et al., 2023), however these are typically discrete and do not model annotator certainty. We propose a variety of methods to model the distribution of the annotations (overview in Figure 4.5):
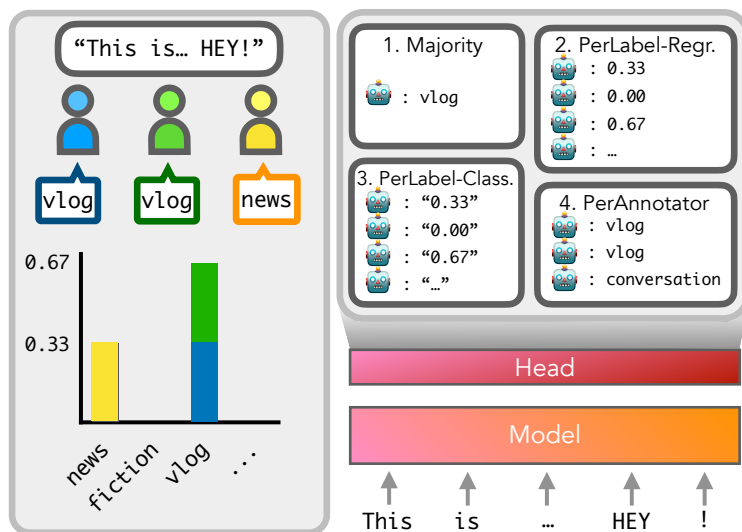
Figure 4.5: The target value each model variant is trained to predict: 1) Majority vote. 2) PerLabelRegr(ession) on label distributions. 3) PerLabel-Class(ification), on score bins per label. 4) PerAnnotator, three different annotations.

Majority   Discretizes the labels using a majority vote, and uses a single classification head to predict it. For the distribution similarity metric (see below), we assign a score of 1.0 to the chosen label.

PerLabel-Regression   Converts the human annotations to scores per label and then predicts these as a regression task. Each label has its own decoder head, trained using an MSE loss, and mapped to the [0;1] range afterwards.

PerLabel-Classification   Converts the human annotations into score bins and predicts them as four possible labels: "0.0", "0.33", "0.66", "1.0".

PerAnnotator   One decoder head modeling each annotator, that predicts their annotation as a discrete label. Afterwards, the three predictions are converted to a distribution.

We evaluate these models using the standard accuracy over each singular predicted label (i.e., highest score or majority). In addition, we conduct a finer-grained evaluation that takes the multi-annotations into account. For this purpose, we propose a similarity metric for comparing the predicted and annotated label distribution per instance. Let n be the number of label types, and X and Y are label distributions that sum to 1, with a score for each label. Then, the distributional similarity per instance can be computed as:

$$distr\_sim = 1 - \frac{\sum_{l=0}^{n} |X_n - Y_n|}{2} \quad .$$

The resulting score between 0 and 1 represents the distributions' similarity. Note that we compare model predictions to the human annotations, which are not a gold standard; here, we aim to determine whether the human ability to discern these concepts is easy to model.

We implement all our model variants in the MaChAmp (van der Goot et al., 2021b) toolkit v0.4 using default parameters. MaChAmp is a toolkit focused on multi-task learning for NLP, and allowed us to implement all varieties of the tasks described earlier. Each way of phrasing the task is implemented on top of a single language model for fair comparison. From an initial evaluation of the bert-large-cased (Devlin et al., 2019), luke-large-lite (Yamada et al., 2020), deberta-v3-large (He et al., 2021), xlm-roberta-large (Conneau et al., 2020) LMs on the gold genre labels, we identify that DeBERTa has the highest accuracy; hence we use it in the following experiments.

4.5.2   Classification Results

We examine which notion of domain is more learnable and distinguishable for a model; genre or topic? Since genre has associated ground truth labels, we additionally examine whether the human annotators' perception of genre or the ground truth genre is easier to learn.

We establish a majority vote based on the human annotations; in case of a tie, the first element in the annotation list is chosen as the label, both

	Accuracy	Macro-F1	$|N|$
Sentence	67.68	59.92	1,117
Prose	81.11	74.75	217

Table 4.4: Performance of annotators' majority vote compared with the gold genre (development set).

for sentences and prose. This happens in ∼10% of cases for genre and topic-10 (sentence and prose), and ∼20% cases for topic-100.

Table 4.4 shows accuracy and macro-F1 scores of the annotators' majority vote evaluated against the gold genre. As noted previously, more context (prose level) helps disambiguate the genre.

To evaluate how well a model can align with the human intuition of genres and topics, we fine-tune an LM on the majority labels of the annotators. We compare the performance on the gold genre labels (the only task for which we have gold labels) and compare the accuracy and macro-F1 scores (Table 4.5). We notice the following:

Sentences 1) Unsurprisingly, DeBERTa fine-tuned on the gold genre labels (gold_genre) is better aligned with the ground truth genre than the human majority vote, i.e., 73.20 (Table 4.5) versus 67.68 (Table 4.4) accuracy at the sentence level (note that other LMs performed worse). 2) In contrast, the fine-tuned DeBERTa model has higher accuracy when trained and tested on the human majority vote (maj_gerne) than when using gold genre labels (gold_genre), i.e., 75.88 versus 73.20, although macro-F1 is lower. This indicates that less common genre labels are easier to learn from gold labels, while more frequent genres are easier to learn based on human intuitions. 3) Despite topic-10 having fewer classes than genre, the notion of topic appears to be more difficult for a model to learn (lower F1). 4) The skew of the fine-grained topics (maj_topic-100) and the difficulty of the long tail become apparent in the large divergence across the accuracy and macro-F1 score.

		Accuracy	Macro-F1
Sent.	gold_genre	73.20± 0.02	70.74± 0.02
	maj_genre	75.88± 0.01	67.04± 0.01
	maj_topic-10	75.56± 0.02	60.54± 0.07
	maj_topic-100	64.55± 0.00	18.43± 0.02
Prose	gold_genre	89.49± 0.02	88.02± 0.03
	maj_genre	80.83± 0.01	74.97± 0.03
	maj_topic-10	67.74± 0.01	50.35± 0.03
	maj_topic-100	52.35± 0.01	16.04± 0.02

Table 4.5: Accuracy and Macro-F1 on test split, for DeBERTa models fine-tuned and evaluated on gold genre, human majority vote for genre, and human majority vote for topic-10/100 (standard deviations across five seeds).

Prose   5) In contrast to the sentence level, our fine-tuned DeBERTa model generalizes better to the gold genre labels (gold_genre) than the human majority vote (maj_genre). At this level of context, the majority vote topic is also harder for a model to learn than the majority vote genre.

4.5.3   Distributional Results

In Figure 4.6, we report the results of the models trained on all instances (sentences and prose) with DeBERTaV3-large.[4] The main trends show that the model performs better on the genre task. Unsurprisingly, for topics, the granularity of the labels impacts performance. By modeling the annotation distributions (i.e., PerLabel-Regression/Classification), we can outperform the majority vote model. However, distributional similarity decreases with increased label granularity (i.e., from topic-10 to topic-100), showing that it is difficult for models to calibrate to diverging human judgments. Interestingly, the per-label models achieve comparable or higher scores on the $distr\_sim$ metric, showing that the examined LMs model label distributions more easily than annotator behavior.

---

[4]Training on sentences and prose separately leads to similar trends (Subsection 4.7.2).
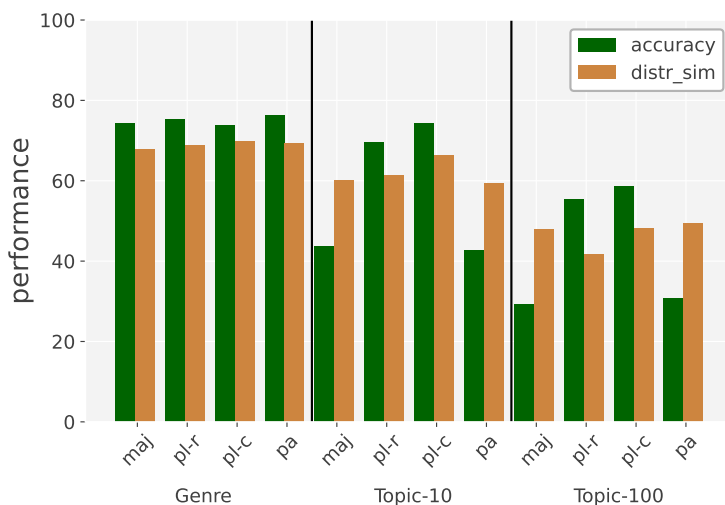
Figure 4.6: Accuracy and distributional similarity on test split, for DeBERTa models trained on target labels based on Majority vote (maj), PerLabel-Regression/Classification (pl-r/c), PerAnnotator labels (pa); standard deviations across five seeds.

4.6   Conclusion

To examine the widely used but scarcely defined notion of *domain*, this work provides the first investigation of human intuitions of this property in the form of **TGeGUM**: a collection of 9.1k sentences annotated with 11 genres and 10/100 topics by three annotators per instance, using an annotation procedure designed to capture human variability instead of forcing alignment (Section 4.3).

Our exploratory analysis (Section 4.4) shows that despite the subjective nature of this task, as reflected in a Fleiss' Kappa of 0.53–0.66, humans can identify certain domain characteristics consistently from one sentence alone. Nonetheless, genres with a high similarity benefit substantially from added context. This is even more crucial for identifying topics, where we observe a shift from annotators not being able to discern any topic at all to being able to reach an above-random agreement, even when presented with 100 or 1,000 topics.

Finally, our experiments of modeling these domain characteristics automatically (Section 4.5) show that genre is easier to model than topic. For both the agreements between human annotators, and the performance from the automatic model, we see that context is crucial for the genre classification task, but not for topic classification, where adding context even leads to decrease in scores if the label space is large.

Overall, this work highlights that despite the importance of "domain", there is little consensus regarding its definition, both in the NLP community as well as in our human annotations. Taking a closer look at what intuition predicted, further reveals that genres and topics are difficult to discretize completely, and that a continuous space of domain variability may be more suited for characterizing these phenomena.

Ethics Statement

Our approach to modeling human label variation is intrinsically linked to the larger issue of human social bias. As highlighted by Plank (2022), significant social implications are tied to the study of label variation. In the context of our research, it is essential to acknowledge that variations in labeling might stem from societal biases and disparities. To address this, we recognize the necessity of addressing bias mitigation techniques as we aim to create more equitable and just models. However, we also contend that our focus on modeling generic subjects, such as genre and topic, may carry less severe implications compared to more subjective tasks like hate speech detection (Akhtar et al., 2021; Davani et al., 2022). The differences in annotations within our work may primarily relate to two categories: "Missing Information" and "Ambiguity" (Sandri et al., 2023).

Another ethical facet we must address is the potential biases present in the classification system we use. In particular, the Dewey Decimal Classification System, which is the de-facto standard for libraries worldwide, has been found to exhibit prejudice (Gooding-Call, 2021). For example, the classification of information related to religion, specifically within class 200, demonstrates a clear skew, with a majority of subjects (six out of

ten) reserved for Christianity-related topics. The remaining four slots are designated for other dominant religions, with an *other* section meant to encompass all other belief systems. This reveals an inherent bias toward Christianity, which can affect the accessibility of non-dominant religions and belief systems. There are alternatives to knowledge organization systems like the Dewey Decimal Classification, as suggested by Franzen (2022), to promote a more inclusive and equitable information landscape.

Acknowledgments

4.7 Appendix

4.7.1 Confusion Matrices Genre

In Figure 4.7-Figure 4.9 we plot the confusion matrices of our DeBERTa model trained on the gold genre labels. The conversation genre shows to be the most difficult label; it is commonly confused with fiction, interview and vlog; which also overlap in length (Section 4.4).
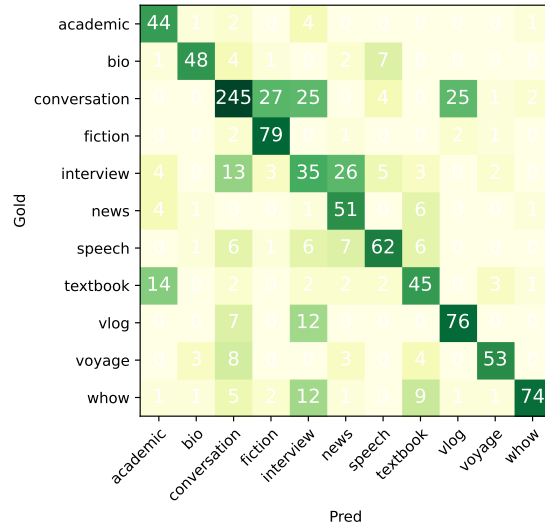
Figure 4.7: Confusion matrix on the sentence level, numbers are summed over all five random seeds.
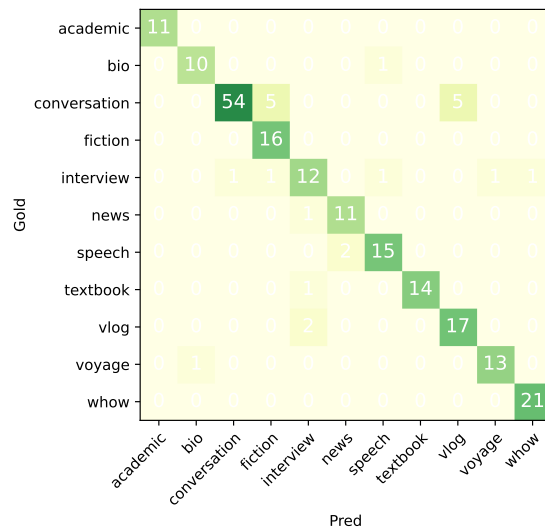
Figure 4.8: Confusion matrix on the prose level, numbers are summed over all five random seeds.
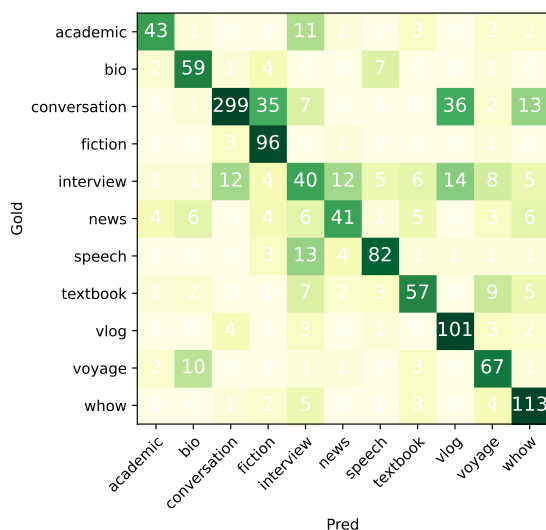
Figure 4.9: Confusion matrix on all data, numbers are summed over all five random seeds.

4.7.2 Sentence and Prose Results

In Figure 4.10 we show the results of our proposed models trained and evaluated only on the sentence level data. Figure 4.11 has the same evaluation on the prose level data.

4.7.3 Visualization of Embeddings

We encode sentences using Sentence-BERT (Reimers and Gurevych, 2019), apply a PCA-downprojection, and color each sentence according to gold genres, our majority-vote genre annotations, as well as majority-vote topic-10 annotations. The results are shown in Figures 4.12–4.14.

4.7.4 Prose-level Statistics

Label statistics on the prose level are shown in Figure 4.15. While general trends, such as the majority genres and topics remain the same as on the sentence level, additional context spreads annotations more evenly,
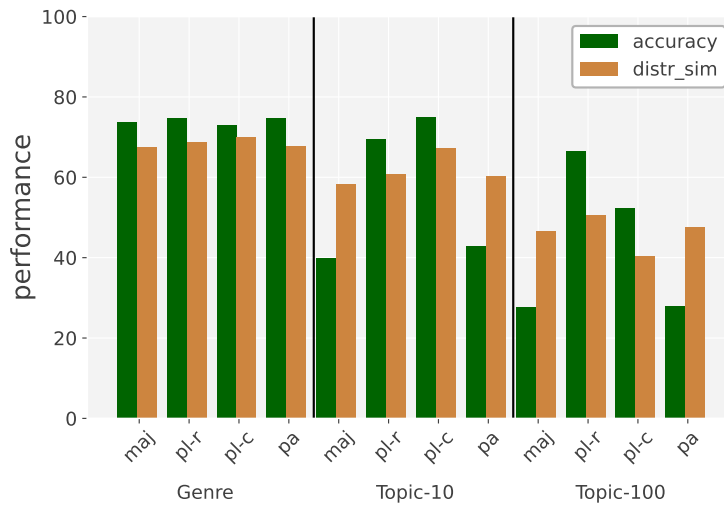
Figure 4.10: Results of our proposed models on the sentence level data.
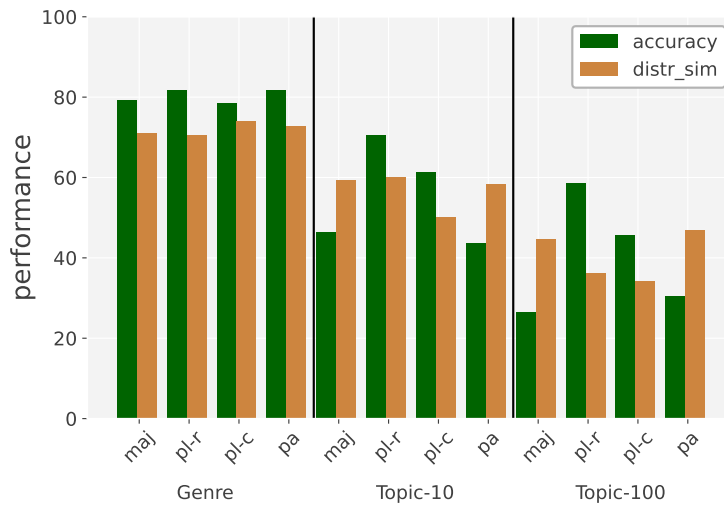
Figure 4.11: Results of our proposed models on the prose level data.
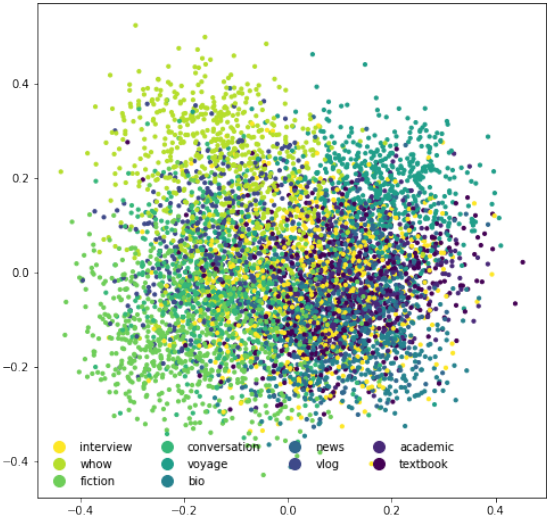
Figure 4.12: PCA plot of sentence embeddings with the gold genres.
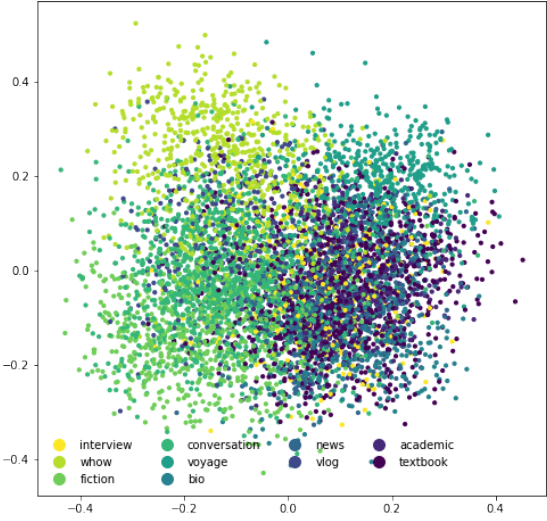
Figure 4.13: PCA plot of sentence embeddings with our annotation for genres, majority vote is used for each instance.
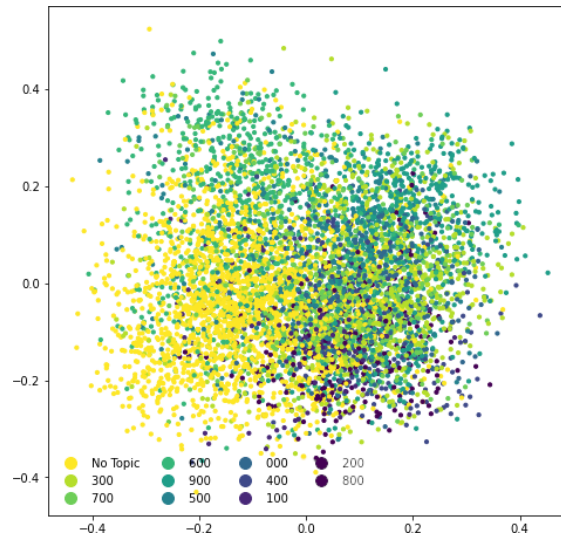
Figure 4.14: PCA plot of sentence embeddings with our annotation for coarse topics, majority vote is used for each instance.

and allows for disambiguations such as for spoken data genres. This is also reflected in the higher alignment between gold and annotated genre labels—both in terms of number, but also in terms of accuracy (Table 4.2). For topic, we further observe almost an order of magnitude fewer no-topic annotations, which are consequently distributed across the spectrum of actual topics.

4.7.5 Annotator Comments

Annotators were provided with a free-form field to provide optional comments regarding each annotation. Of the final dataset, 3.9% of annotations have an annotator comment attached, with a median length of 38 characters. They primarily contain explanations of annotations which were marked with high annotator uncertainty.
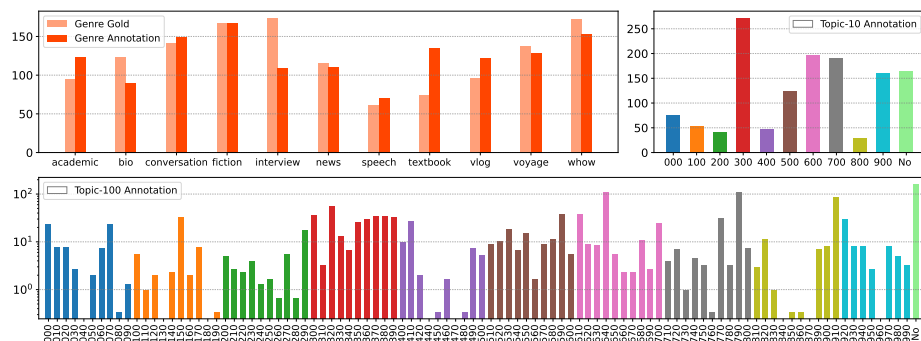
Figure 4.15: **Distribution of Labels (Prose).** Frequency distributions of the labels in gold genre labels, annotations of genres, annotations of topic-10, and annotations of topic-100 (log scale). For the annotations, the number is divided by three to get an average distribution. The mapping of topic-10 and topic-100 labels can be found in **??**. The tag "No" in the topic annotations means "No topic".

4.7.6   Guidelines

Goal/Task

In this annotation project, we are interested in knowing what the topic and genre is of a sentence and whether we humans can identify these. For Topics, we make use of the Dewey Decimal Classification (DDC) system. For genres, we make use of the genres provided in the Georgetown University Multilayer Corpus (GUM) corpus. The goal is to put the sentence/paragraph at hand into the most probable class (determined by you).

Genre has a *one-layer* annotation scheme, while **Topic** has a *two-layer* annotation scheme, which we will refer to as L1 and L2. We want to annotate for all three. There is an option for "Not Sure" (abbreviated to "NS"). This is when you feel that the label for the sentence is not present in the options. In addition, feel free to add any notes for clarification (e.g., clarify your choice or something else).

Preliminaries

Below we give an introduction to the topics and genre labels of this annotation project. It takes around 15-20 minutes to read. Note that you don't have to remember the label numbers. This introduction is to make you aware of the definition of the classes. All the labels are present in the annotation spreadsheet

Introduction Genres

We make use of the text types (genres) in the GUM corpus. These genres do not have a specific number like the topics above. Therefore we simply enumerate them. The genres are the following:

- Academic
- Bio
- Conversation
- Fiction

- Interview
- News
- Speech
- Textbook

- Vlog
- Voyage
- Whow

Brief explanation of the genre classes

- **Academic** (writing) is nonfiction writing adhering to academic standards and disciplines. It includes research reports, monographs, and undergraduate versions. It uses a formal style, references other academic work, and employs consistent rhetorical techniques to define scope, situate in research, and make new contributions.

- A **bio**graphy is a detailed description of a person's life. It involves more than just basic facts like education, work, relationships, and death; it portrays a person's experience of these life events. Unlike a profile or curriculum vitae (résumé), a biography presents a subject's life story, highlighting various aspects of their life, including intimate

details of experience, and may include an analysis of the subject's personality. Biographical works are usually non-fiction, but fiction can also be used to portray a person's life. One in-depth form of biographical coverage is called legacy writing. Works in diverse media, from literature to film, form the genre known as biography. An authorized biography is written with the permission, cooperation, and at times, participation of a subject or a subject's heirs. An autobiography is written by the person themselves, sometimes with the assistance of a collaborator or ghostwriter.

- **Conversation**: naturally occurring spoken interaction. Represents a wide variety of people of different regional origins, ages, occupations, genders, and ethnic and social backgrounds. The predominant form of language use represented is face-to-face conversation, but lso documents many other ways that that people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story-telling, town hall meetings, tour-guide spiels, and more. Fiction refers to creative works, particularly narrative works, that depict imaginary individuals, events, or places. These portrayals deviate from history, fact, or plausibility. In our data, fiction pertains to written narratives like novels, novellas, and short stories.

- An **interview** is a structured conversation where one person asks questions and another person answers them. It can be a one-on-one conversation between an interviewer and an interviewee. The information shared during the interview can be used or shared with others.

- **News** is information about current events, shared through various media like word of mouth, printing, broadcasting, electronic communication, and witness testimonies. It covers topics such as war, government, politics, education, health, environment, economy, business, fashion, entertainment, sports, and unusual events. Government an-

nouncements and technological advancements have accelerated news dissemination and influenced its content.

- A (political) **speech** is a public address given by a political figure or a candidate for public office, usually with the aim of persuading or mobilizing an audience to support their ideas, policies, or campaigns. Political speeches are an essential tool for politicians to communicate their vision, articulate their positions, and connect with voters or constituents.

- A **textbook** is a book containing a comprehensive compilation of content in a branch of study with the intention of explaining it. Textbooks are produced to meet the needs of educators, usually at educational institutions. Schoolbooks are textbooks and other books used in schools. Today, many textbooks are published in both print and digital formats.

- A **vlog**, also known as a video blog or video log, is a form of blog for which the medium is video. The dataset contains transcripts of the speech occurring in the video.

- A travel/**voyage** guide is a wiki providing information for visitors or tourists about a particular place. It typically includes details about attractions, lodging, dining, transportation, and activities. It may also contain maps, historical facts, and cultural insights. Guide wikis cater to various travel preferences, such as adventure, relaxation, budget, or specific interests like LGBTQ+ travel or dietary needs.

- A Wikihow how-to (**whow**) guide is an instructional document that offers step-by-step guidance on accomplishing a specific task or reaching a particular goal. It aims to assist individuals in learning and comprehending the process involved in successfully completing the task. These guides are typically written in a clear and concise manner, simplifying complex processes into manageable steps. They often include detailed explanations, diagrams, illustrations, or examples to

enhance understanding. How-to guides cover various topics, such as technical tasks, practical skills, creative endeavors, troubleshooting, and more.

Introduction Topics

The DDC system is a widely used library classification system developed by Melvil Dewey in the late 19th century. The DDC is based on the principle of dividing knowledge (in our case sentences) into ten main classes, each identified by a three-digit number; we only focus on the first two:

1. The ten main classes in the Dewey Decimal Classification system are as follows:

   - 000 Computer science, information & general works
   - 100 Philosophy & psychology
   - 200 Religion
   - 300 Social sciences
   - 400 Language
   - 500 Science
   - 600 Technology
   - 700 Arts & recreation
   - 800 Literature
   - 900 History & geography

   These higher level classes belong to L1 in the annotation spreadsheet, and we added the NO-TOPIC label (see description below)

2. Each main class is further divided into subclasses using additional digits (10s). For example, in the 500s (natural sciences and mathematics), you'll find 510 for mathematics, 520 for astronomy, 530 for

physics, and so on. The system allows for more specific classification of books and materials based on their subject matter.

See the following page: https://www.oclc.org/content/dam/oclc/dewey/ddc23-summaries.pdf

This page separates the ten classes above into more finer-grained classes. There is not an explanation for each of them, but usually the name of the label encapsulates the subclass already. Note that the subclasses overwrite the main classes (so you can't pick 400 and 510, then you'd have to change 510 to 500).

These subclasses belong to L2 in the annotation spreadsheet.

Note that for each fine-grained class we deem the main number/code (e.g., 100, 200, 300) in L2 as the No-topic/Other category. The "Other" class can only be chosen in the fine-grained label classes (L2). Choosing this means that you believe that the current sentence belongs to a specific class. But the label is not present.

The Dewey Decimal Classification system is used in many libraries around the world to organize their collections and make it easier for users to locate resources. It provides a systematic way of arranging materials and enables efficient browsing and retrieval of information based on subject areas.

Brief explanation of the topic classes (L1)

- 000 Computer science, information & general works is the most general class and is used for works not limited to any one specific discipline, e.g., encyclopedias, newspapers, general periodicals. This class is also used for certain specialized disciplines that deal with knowledge and information, e.g., computer science, library and information science, journalism. Each of the other main classes (100-900) comprises a major discipline or group of related disciplines. Note that

in our experiments, we do not consider this a miscellaneous category, we have "No-topic" for this.

- 100 Philosophy & psychology covers philosophy, parapsychology and occultism, and psychology.

- 200 Religion is devoted to religion.

- 300 Social sciences covers the social sciences. Class 300 includes sociology, anthropology, statistics, political science, economics, law, public administration, social problems and services, education, commerce, communications, transportation, and customs.

- 400 Language comprises language, linguistics, and specific languages. Literature, which is arranged by language, is found in 800.

- 500 Science is devoted to the natural sciences and mathematics.

- 600 Technology is technology.

- 700 Arts & recreation covers the arts: art in general, fine and decorative arts, music, and the performing arts. Recreation, including sports and games, is also classed in 700.

- 800 Literature covers literature, and includes rhetoric, prose, poetry, drama, etc. Folk literature is classed with customs in 300.

- 900 History & geography is devoted primarily to history and geography. A history of a specific subject is classed with the subject.

- No topic: For cases where the topic can not be determined, or even guessed. For example for utterances that contain no natural language or do not have enough context.

FAQ

- Should the colors of L1 and L2 in the annotation spreadsheet match?

Yes, apart from that the colours should match, the first number of the class to which the sentence belongs should also match.

For example, a sentence that belongs to Arts (700), is restricted to anything in the 700 class, e.g., a painting (750).

- If a sentence has a clear topic in general, but the L2 category does not match, how do we annotate?

  The fine-grained (L2) topics have the priority, and since they have to match you adjust the main topic accordingly.

- Does my choice of Topic depend on the Genre or vice versa?

  No, by default, annotating for genre and topic should be a separate task and should not influence each other.

- How do we distinguish between something that is in the No-topic (or Others) class and NS ("not sure")?

  Use the "others" category when you believe the current instance to belong to a class which is not in the listed ones. Mark your choice with "NS" when you have a guess, but you are not confident about it (e.g., because the instance is very short, or you are not familiar with the genre/topic)

  If you are able to find L1, but none of the labels fit for the sentence in L2, you should choose "Other" (e.g, 000, 100, 200, etc.) in the same colour (class) of L2. The "Other" class can only be chosen in the fine-grained label classes (L2). Choosing this means that you believe that the current sentence belongs to a specific class. But the label is not present. Otherwise, mark your best guess with "NS".

- Is it better to label a sentence as "NO-TOPIC" if there is not a clear label associated with it or are we encouraged to take a guess?

  You are encouraged to take a guess. However, for cases where you have no preference for any of the labels (i.e. a wild guess), label it as NO-TOPIC.

68

- There is already another "Other" class in Religion/Language (e.g., 290 Other religion).

  Good catch, imagine this situation. Let's say the sentence is talking about Buddhism. This falls under 290, because we're talking about another religion. However, if the sentence is "vaguely" talking about religion and doesn't fit within any of the labels, then choose 200 (Other).

- Where do ads/exam questions fit?

  In whichever of the genres you would expect to come across advertisements/exam questions. However, note that the data is scraped from the main information channel of source (i.e., advertisements next to a news text or before a vlog are not included).

- Can we use external resources?

  External resources are allowed, but do not look up the literal sentence.

- How to pick topics (L1/L2) for fiction (genre)?

  Note that the genre and topic tasks should be seen as distinct tasks. So, the genre fiction should not automatically lead to a literature topic label (unless the fiction work is about literature).

- Some utterances seem to be taken from the same text; do we have to give them the same label, or take the contexts into account?

  No, each utterance should be judged independently.

Note for L3:

- For each L2, there is a finer-grained class namely L3. These numbers go in the thousands. Now, try to pick the most likely thousands' topic:

  – You will have to refer to the PDF (L3-1000.pdf) for the right classes.

Figure 4.16: Example of annotation in Google Spreadsheets. NS = Not Sure

- **–** Please write the class number in the spreadsheet cell. There is no dropdown menu.

- The "no-topic" option still exists. Use "NT";

- You should pick the fine-grained L3 topic that best fits the utterance. This time you don't have to match the L1-L2 categories, but we ask you to NOT update your previous L1-L2 annotations, and just annotate L3 independently.

4.7.7   Annotation Tool

We used Google Spreadheets for annotation. The setup is shown in Figure 4.16.

Part II

Data

Chapter 5

What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification

Abstract

Over the last five years, research on Relation Extraction (RE) witnessed extensive progress with many new dataset releases. At the same time, setup clarity has decreased, contributing to increased difficulty of reliable empirical evaluation (Taillé et al., 2020). In this paper, we provide a comprehensive survey of RE datasets, and revisit the task definition and its adoption by the community. We find that cross-dataset and cross-domain setups are particularly lacking. We present an empirical study on scientific Relation Classification across two datasets. Despite large data overlap, our analysis reveals substantial discrepancies in annotation. Annotation discrepancies strongly impact Relation Classification performance, explaining large drops in cross-dataset evaluations. Variation within further subdomains exists but impacts Relation Classification only to limited degrees. Overall, our study calls for more rigour in reporting setups in RE and evaluation across multiple test sets.

5.1 Introduction

Information Extraction (IE) is a key step in Natural Language Processing (NLP) to extract information, which is useful for question answering and knowledge base population, for example. Relation Extraction (RE) is a specific case of IE (Grishman, 2012) with the focus on the identification of semantic relations between entities (see Figure 5.1). The aim of the most typical RE setup is the extraction of informative triples from texts. Given a sequence of tokens $[t_0, t_1..., t_n]$ and two entities (spans), $s_A = [t_i, \ldots, t_j]$ and $s_B = [t_u, \ldots, t_v]$, RE triples are in the form (s_A, s_B, r), where $r \in R$ and R is a pre-defined set of relation labels. Because of the directionality of the relations, (s_B, s_A, r) represents a different triple.

We survey existing RE datasets—outside the biomedical domain—with an additional focus on the task definition.[1] Existing RE surveys mainly

---

[1] We refer the reader to Luo et al. (2016) for a survey on biomedical RE and event extraction.
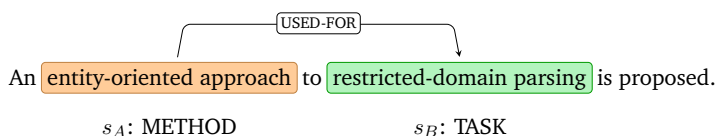
Figure 5.1: RE annotation sample. The sentence contains two annotated spans denoting two entities, with respective types METHOD and TASK, and a semantic relation between them labeled as USED–FOR.

focus on modeling techniques (Bach and Badaskar, 2007; Pawar et al., 2017; Aydar et al., 2021; Liu, 2020). To the best of our knowledge, we are the first to give a comprehensive overview of available RE datasets. We also revisit RE papers from the ACL community over the last five years, to identify what part(s) of the task definition recent work focuses on. As it turns out, this is often not easy to determine, which makes fair evaluation difficult. We aim to shed light on such assumptions.[2]

Moreover, recent work in NLP has shown that single test splits and in-distribution evaluation overestimate generalization performance, arguing for the use of multiple test sets or split evaluation (Gorman and Bedrick, 2019; Søgaard et al., 2021). While this direction has started to be followed by other NLP tasks (Petrov and McDonald, 2012; Pradhan et al., 2013; Williams et al., 2018b; Yu et al., 2019; Zhu et al., 2020a; Liu et al., 2021b), for RE *cross-dataset* and *cross-domain* evaluation have received little attention. We explore this direction in the scientific domain and propose to study the possible presence of distinctive *sub-domains* (Lippincott et al., 2010). Sub-domains are differences between subsets of a domain that may be expected to behave homogeneously. Using two scientific datasets, we study to what degree: (*a*) they contain overlapping data; (*b*) their annotations differ; and (*c*) sub-domains impact Relation Classification (RC)—the task of classifying the relation type held between a pair of entities (details in Section 5.3).

The contributions of this paper are:

- To the best of our knowledge, we are the first to provide a compre-

---

[2]Pyysalo et al. (2008) discuss similar difficulties in the biomedical domain.

hensive survey on currently available RE datasets.

- We define RE considering its modularity. We analyze previous works and find unclarity in setups; we call for more rigour in specifying which RE sub-part(s) are tackled.

- We provide a case study on Relation Classification in the scientific domain, to fill a gap on cross-domain and cross-dataset evaluation.

5.2 Relation Extraction Datasets Survey

RE has been broadly studied in the last decades and many datasets were published. We survey widely used RE datasets in chronological order, and broadly classify them into three domains based on the data source: (1) news and web, (2) scientific publications and (3) Wikipedia. An overview of the datasets is given in Table 5.1. Our empirical target here focuses on the scientific domain as so far it has received no attention in the cross-domain direction; a similar investigation on overlaps in data, annotation, and model transferability between datasets in other domains is interesting future work.

The CoNLL 2004 dataset (Roth and Yih, 2004) is one of the first works. It contains annotations for named entities and relations in news articles. In the same year, the widely studied ACE dataset was published by Doddington et al. (2004). It contains annotated entities, relations and events in broadcast transcripts, newswire and newspaper data in English, Chinese and Arabic. The corpus is divided into six domains.

Another widely used dataset is The New York Times (NYT) Annotated Corpus,[3] first presented by Riedel et al. (2010). It contains over 1.8 million articles by the NYT between 1987 and 2007. NYT has been created with a distant supervision approach (Mintz et al., 2009), using Freebase (Bollacker et al., 2008) as knowledge base. Two further versions of it followed recently: Zhu et al. (2020b) (NYT-H) and Jia et al. (2019) published

---

[3]`http://iesl.cs.umass.edu/riedel/ecml/`

Dataset	Paper	Data Source	# Rel. Types
News and Web			
CoNLL04	Roth and Yih (2004)	News articles	5
ACE*	Doddington et al. (2004)	News and conversations	24
NYT	Riedel et al. (2010)	New York Times articles	24-57$^\diamond$
SemEval-2007	Girju et al. (2007)	Sentences from the web	7
SemEval-2010	Hendrickx et al. (2010)	Sentences from the web	10
TACRED	Zhang et al. (2017b)	Newswire and web text	42
FSL TACRED	Sabo et al. (2021)	TACRED data	42
DWIE	Zaporojets et al. (2021)	Deutsche Welle articles	65
Scientific publications			
ScienceIE	Augenstein et al. (2017)	Scientific articles	2
SemEval-2018	Gábor et al. (2018)	NLP abstracts	6
SciERC	Luan et al. (2018)	Abstracts of AI proceedings	7
Wikipedia			
GoogleRE	-	Wikipedia	5
mLAMA*	Kassner et al. (2021)	GoogleRE data	5
FewRel	Han et al. (2018)	Wikipedia	100
FewRel 2.0	Gao et al. (2019)	FewRel data + Biomedical lit.	100 + 25
DocRED	Yao et al. (2019)	Wikipedia and Wikidata	96
SMiLER	Seganti et al. (2021)	Wikipedia	36

Table 5.1: Overview of the RE datasets for the English language grouped by macro domains. ($\star$): Multilingual datasets. ($\diamond$): The original paper does not state the number of considered relations and different work describe different dataset setups.

manually annotated versions of the test set in order to perform a more accurate evaluation.

RE has also been part of the SemEval shared tasks for four times so far. The two early SemEval shared tasks focused on the identification of semantic relations between nominals (Nastase et al., 2021). For SemEval-2007 Task 4, Girju et al. (2007) released a dataset for RC into seven generic semantic relations between nominals. Three years later, for SemEval-2010 Task 8, Hendrickx et al. (2010) revised the annotation guidelines and published a corpus for RC, by providing a much larger dataset (10k instances, in comparison to 1.5k of the 2007 shared task).

Since 2017, three RE datasets in the scientific domain emerged, two of the three as SemEval shared tasks. In SemEval-2017 Task 10 Augenstein et al. (2017) proposed a dataset for the identification of keyphrases and considered two generic relations (HYPONYM-OF and SYNONYM-OF). The dataset is called ScienceIE and consists of 500 journal articles from the

Computer Science, Material Sciences and Physics fields. The year after, Gábor et al. (2018) proposed a corpus for RC and RE made of abstracts of scientific papers from the ACL Anthology for SemEval-2018 Task 7. The data will be described in further detail in Section 5.4.1. Following the same line, Luan et al. (2018) published SCIERC, which is a scientific RE dataset further annotated for coreference resolution. It contains abstracts from scientific AI-related conferences. From the existing three scientific RE datasets summarized in Table 5.1, in our empirical investigation we focus on two (SemEval-2018 and SCIERC). We leave out ScienceIE as it focuses on keyphrase extraction and it contains two generic relations only.

The Wikipedia domain has been first introduced in 2013. Google released GoogleRE,[4] a RE corpus consisting of snippets from Wikipedia. More recently, Kassner et al. (2021) proposed mLAMA, a multilingual version (53 languages) of GoogleRE with the purpose of investigating knowledge in pre-trained language models. The multi-lingual dimension is gaining more interest for RE. Following this trend, Seganti et al. (2021) presented SMiLER, a multilingual dataset (14 languages) from Wikipedia with relations belonging to nine domains.

Previous datasets were restricted to the same label collection in the training set and in the test set. To address this gap and make RE experimental scenarios more realistic, Han et al. (2018) published Few-Rel, a Wikipedia-based few-shot learning (FSL) RC dataset annotated by crowdworkers. One year later, Gao et al. (2019) published a new version (Few-Rel 2.0), adding a new test set in the biomedical domain and the `None-Of-The-Above` relation (cf. Section 5.3).

Back to the news domain, Zhang et al. (2017b) published a large-scale RE dataset built over newswire and web text, by crowdsourcing relation annotations for sentences with named entity pairs. This resulted in the TACRED dataset with over 100k instances, which is particularly well-suited for neural models. Sabo et al. (2021) used TACRED to make a FSL RC dataset and compared it to FewRel 1.0 and FewRel 2.0, aiming at a

---

[4]`https://code.google.com/archive/p/relation-extraction-corpus/downloads`

more realistic scenario (i.e., non-uniform label distribution, inclusion of pronouns and common nouns).

All datasets so far present a sentence level annotation. To address this, Yao et al. (2019) published DocRED, a document-level RE dataset from Wikipedia and Wikidata. The difference with a traditional sentence-level corpus is that both the intra- and inter-sentence relations are annotated, increasing the challenge level. In addition to RE, DocRED annotates coreference chains. DWIE by Zaporojets et al. (2021) is another document-level dataset, specifically designed for multi-task IE (Named Entity Recognition, Coreference Resolution, Relation Extraction, and Entity Linking).

Lastly, there are works focusing on creating datasets for specific RE aspects. Cheng et al. (2021), for example, proposed a Chinese document-level RE dataset for *hard cases* in order to move towards even more challenging evaluation setups.

Domains in RE   Given our analysis, we observe a shift in target domains: from news text in seminal works, over web texts, to emerging corpora in the scientific domain and the most recent focus on Wikipedia. Similarly, we observe the emerging trend for FSL.

Different datasets lend themselves to study different aspects of the task. Concerning cross-domain RE, we propose to distinguish three setups:

1. Data from different domains, but same relation types, which are general enough to be present in each domain (limited and often confined to the ACE dataset) (e.g., Plank and Moschitti, 2013).

2. Stable data domain, but different relation sets (e.g., FewRel by Han et al., 2018). Note that when labels change, approaches such as FSL must be adopted.

3. A combination of both: The data changes and so do the relation types (e.g., FewRel 2.0 by Gao et al., 2019).

In the case study of this paper, given the scientific datasets available, we focus on the first setup.
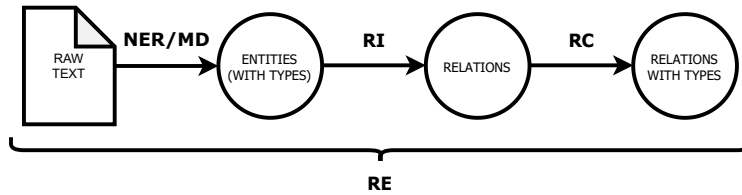
Figure 5.2: Relation Extraction pipeline. NER: Named Entity Recognition; MD: Mention Detection; RI: Relation Identification; RC: Relation Classification.

5.3   The Relation Extraction Task

Conceptually, RE involves a pipeline of steps (see Figure 5.2). Starting from the raw text, the first step consists in identifying the entities and eventually assigning them a type. Entities involve either nominals or named entities, and hence it is either Named Entity Recognition (NER) or, more broadly, Mention Detection (MD).[5] After entities are identified, approaches start to be more blurry as studies have approached RE via different angles.

One way is to take two steps, Relation Identification (RI) and subsequent Relation Classification (RC) (Ye et al., 2019), as illustrated in Figure 5.2. This means to first identify from all the possible entity pairs the ones which are in some kind of relation via a binary classification task (RI). As the proportion of positive samples over the negative is usually extremely unbalanced towards the latter (Gormley et al., 2015), a priori heuristics are generally applied to reduce the possible combinations (e.g., entity pairs involving distant entities, or entity type pairs not licensed by the relations are not even considered). The last step (RC) is usually a multi-class classification to assign a relation type r to the positive samples from the previous step. Some studies merge RI and RC (Seganti et al., 2021) into one step, by adding a `no-relation` (`no-rel`) label. Other studies instead reduce the task to RC, and assume there exists a relation between two entities and the task is to determine the type (without a `no-rel` label).

---

[5]Some studies divide the entity extraction into two sub-steps: identification (often called MD), and subsequent classification into entity types.

Regardless, RI is influenced by the RC setup: Relations which are not in the RC label set are considered as negative samples in the RI phase. Some studies address this approximation by distinguishing between the `no-rel` and the `None-Of-The-Above` (`NOTA`) relation (Gao et al., 2019). Note that, in our definition, the `NOTA` label differs from `no-rel` in the sense that a relation holds between the two entities, but its type is not in the considered RC label set.[6]

What Do You Mean by Relation Extraction?   RE studies rarely address the whole pipeline. We analyze all the ACL papers published in the last five years which contain the *Relation Extraction* keyword in the title and determine which sub-task is performed (NER/MD, RI, RC). Table 5.2 shows such investigation. We leave out from this analysis (*a*) papers which make use of distant supervision or which somehow involve knowledge bases, (*b*) shared task papers, (*c*) the bioNLP field, (*d*) temporal RE, and (*e*) Open RE.

The result shows that gold entities are usually assumed for RE, presumably given the complexity of the NER/MD task on its own. Most importantly, for end-to-end models, recent work has shown that ablations for steps like NER are lacking (Taillé et al., 2020). Our analysis further shows that it is difficult to determine the RI setup. While RC is always performed, the situation is different for RI (or `no-rel`). Sometimes RI is clearly not done (i.e., the paper assumes a scenario in which every instance contains at least one relation), but most of the times it is either not clear from the paper, or done in a simplified scenario (e.g., datasets which already clear out most of the `no-rel` entity pair instances). As this blurriness hampers fair evaluation, we propose that *studies clearly state which step they include*, i.e., whether the work focus is on RC, RI+RC or the full RE pipeline and how special cases (`no-rel` and `NOTA`) are handled. These details are utterly important as they impact both model estimation and evaluation.

---

[6]Some studies name such relation `Other` (Hendrickx et al., 2010).

Relation Extraction Paper	Task Performed		
	NER/MD	RI	RC
2021			
Wang et al. (2021)	✓	✓	✓
Cui et al. (2021)			✓
Tang et al. (2021)		(?)	✓
Xie et al. (2021)	✓	(?)	✓
Tian et al. (2021)			✓
Ma et al. (2021)		✓	✓
Mathur et al. (2021)			✓
Yang et al. (2021)			✓
Huang et al. (2021b)		(?)	✓
Huang et al. (2021a)		(?)	✓
2020			
Kruiper et al. (2020)	✓		✓
Nan et al. (2020)			✓
Alt et al. (2020)		✓	✓
Yu et al. (2020)		✓	✓
Shahbazi et al. (2020)		(?)	✓
Pouran Ben Veyseh et al. (2020)			✓
2019			
Trisedya et al. (2019)	✓	(?)	✓
Guo et al. (2019)		✓	✓
Yao et al. (2019)			✓
Zhu et al. (2019)		✓	✓
Li et al. (2019)	✓	(?)	✓
Ye et al. (2019)		✓	✓
Fu et al. (2019)	✓	✓	✓
Dixit and Al-Onaizan (2019)	✓	✓	✓
Obamuyide and Vlachos (2019)		(?)	✓
2018			
Christopoulou et al. (2018)		✓	✓
Phi et al. (2018)			✓
2017			
Lin et al. (2017)		(?)	✓

Table 5.2: ACL paper analysis: over the last 5 years, which RE sub-task is performed. *(?)* indicates that either the paper does not state if the step is considered, either it is performed, but in a simplified scenario.

Pipeline or Joint Model?   The traditional RE pipeline is, by definition of pipeline, prone to error propagation by sub-tasks. Joint entity and relation extraction approaches have been proposed in order to alleviate this problem (Miwa and Bansal, 2016; Zhang et al., 2017a; Bekoulis et al., 2018b,a; Wang and Lu, 2020; Wang et al., 2021). However, Taillé et al. (2020) recently discussed the challenge of properly evaluating such complex models. They surveyed the evaluation metrics of recently published works on end-to-end RE referring to the *Strict*, *Boundaries*, *Relaxed* evaluation setting proposed by Bekoulis et al. (2018b). They observe unfair comparisons and overestimations of end-to-end models, and claim the need for more rigorous reports of evaluation settings, including detailed datasets statistics.

While some recent work shifts to joint models, it is still an open question which approach (joint or pipeline) is the most robust. Zhong and Chen (2021) found that when incorporating modern pre-trained language models (e.g., BERT) using separate encoders can surpass existing joint models. Since the output label space is different, separate encoders could better capture distinct contextual information. At the moment it is not clear if one approach is more suitable than the other for RE. For this reason and because of our final goal, which is a closer look to sub-domains in the scientific field, we follow the pipeline approach and, following most work from Table 5.2, we here restrict the setup by focusing on the RC task.

Open Issues   To summarize, open issues are: *1)* The unclarity of RE setups, as illustrated in Table 5.2 —specially regarding RI—leads to problematic evaluation comparisons; *2)* A lack of cross-domain studies, for all three setups outlined in Section 5.2.

5.4   Scientific Domain Data Analysis

In this section, we present the two English corpora involved in the experimental study (Section 5.4.1), explain the label mapping adopted for the cross-dataset experiments (Section 5.4.2), discuss the overlap between the

datasets and the annotation divergence between them (Section 5.4.3), and introduce the sub-domains considered (Section 5.4.4).

5.4.1 Datasets

SemEval-2018 Task 7 (Gábor et al., 2018)   The corpus contains 500 abstracts of published research papers in computational linguistics from the ACL Anthology. Relations are classified into six classes. The task was split into three sub-tasks: (1.1) RC on clean data (manually annotated), (1.2) RC on noisy data (automatically annotated entities) and (2) RI+RC (identifying instances + assigning class labels). For each sub-task, the training data contains 350 abstracts and the test data 150. The train set for sub-task (1.1) and (2) is identical.

SCIERC (Luan et al., 2018)   The dataset consists of 500 abstracts from scientific publications annotated for entities, their relations and coreference clusters. The authors define six scientific entity types and seven relation types. The original paper presents a unified multi-task model for entity extraction, RI+RC and coreference resolution. SCIERC is assembled from different conference proceedings. As the data is released with original abstract IDs, this allows us to identify four major sub-domains: AI and ML, Computer Vision (CV), Speech Processing, and NLP, sampled over a time frame from 1980 to 2016. Details of the sub-domains are provided in Table 5.9 in Appendix 5.7.1. To the best of our knowledge, we are the first to analyze the corpus at this sub-domain level.

5.4.2 Cross-dataset Label Mapping

We homogenize the relation label sets via a manual analysis performed after an exploratory data analysis, as we find that most of the labels in SemEval-2018 and SCIERC have a direct correspondent, and hence we mapped them as shown in Table 5.3. The gold label distribution of the relations on the two datasets is shown in Figure 5.4 in Appendix 5.7.2. We

	SemEval-2018	**SCIERC**
Considered in this study	COMPARE USAGE PART_WHOLE MODEL-FEATURE RESULT	COMPARE USED-FOR PART-OF FEATURE-OF EVALUATE-FOR[*]
Not-considered	TOPIC - -	- HYPONYM-OF CONJUNCTION

Table 5.3: Label mapping. (*): Same semantic relation, but inverse direction. We homogenized the two versions by flipping the head with the tail.

Table 5.4: Annotated sentence pairs from SemEval-2018 and SCIERC. The underlined spans are the entities.

decided to leave out the two generic labels from SCIERC and one relation from SemEval-2018 which does not have any correspondent and is rare.

5.4.3 Overlap of the Datasets and Annotation Divergences

Our analysis further reveals a high overlap in articles between SemEval-2018 and SCIERC corresponding to 307 ACL abstracts.[7] Interestingly, the overlap contains a huge annotation divergence. In more detail, we identify three main annotation disagreement scenarios between the two datasets (represented by the 3 samples in Table 5.4):

---

[7]Note that in our study, regarding SemEval-2018, for fair comparison with SCIERC, which is manually annotated, we consider the dataset related to sub-task (1.1).

- **Sample 1**: *The annotated entities differ and so the annotated relations do as well.* SemEval-2018 annotates just one entity and thus there can not even exist a relation; as the corresponding sentence in SCIERC is annotated with two entities, it contains a relation.

- **Sample 2**: *The amount of annotated entities and the amount of anno- tated relations are the same, but the annotations do not match.* The relations involve non-mutual entities and so do not correspond.

- **Sample 3**: *The annotated entities are the same, but the relation an- notations differ.* This involves conflicting annotations, e.g., the bold arrow shows the same entity pair annotated with a different relation label.

Table 5.5 shows the annotation statistics from the two corpora and their overlap. Overall both datasets contain the same amount of abstracts, but the amount of annotated relations differs substantially. The overlap between the two corpora reports a similar trend. Even the fairer count of the common labels (see Table 5.3) reveals that the annotation gap still holds (ratio of 1:1.8). In more detail, the entity pairs annotated in both dataset by using a strict criterion (i.e., entity spans with the same boundaries) are only 394 (considering relations from the whole relation sets). Out of them, only 327 are labeled with the same relation type, meaning that there are 67 conflicting instances as the bold arrow in Table 5.4 (Sample 3).

5.4.4 Experimental Sub-domains

We use the metadata described in Section 5.4.1 to divide SCIERC into four sub-domains. Figure 5.5 in Appendix 5.7.2 shows the label distribution over the new SCIERC split. As we are particularly interested in the annotation divergence impact, we leave out of this study 193 abstracts from SemEval-2018 which are not in overlap with SCIERC.

We assume a setup which takes the NLP domain as source training domain in all experiments, as it is the largest sub-domain in both datasets.

Whole corpus		
	SemEval-2018	SCIERC
# abstracts	500	500
# relations	1,583	4,648
Datasets Overlap (307 abstracts)		
# relations	1,087	2,476
# common relations	1,071	1,922
Same entity pair		394
Same entity pair + same relation type		327

Table 5.5: SemEval-2018 and SCIERC annotation comparison. The common relations are the ones with a direct correspondent in both datasets (see Table 5.3).

Dataset	Sub-domain	train	dev	test
SemEval-2018	NLP	257	50	50
SCIERC	NLP	257	50	50
	AI-ML	-	-	52
	CV	-	-	105
	SPEECH	-	-	35

Table 5.6: Sub-domains and relative amount of abstracts.

The considered sub-domains and their relative amount of data are reported in Table 5.6.

5.5 Experiments

5.5.1 Model Setup

Since the seminal work by Nguyen and Grishman (2015b), Convolutional Neural Networks (CNNs) are widely used for IE tasks (Zeng et al., 2014; Nguyen and Grishman, 2015b; Fu et al., 2017; Augenstein et al., 2017; Gábor et al., 2018; Yao et al., 2019). Similarly, since the advent of contextualized representations (Peters et al., 2018; Devlin et al., 2019),

BERT-like representations are commonly used (Seganti et al., 2021), but non-contextualized embeddings (i.e., GloVe, fastText) are still widely adopted (Yao et al., 2019; Huang et al., 2021b). We compare the best CNN setup to fine-tuning a full transformer model. For the latter we use the MaChAmp toolkit (van der Goot et al., 2021b)

Our CNN follows Nguyen and Grishman (2015b). We tests both non-contextualized word embeddings—fastText (Bojanowski et al., 2017)—and contextualized ones—BERT (Devlin et al., 2019) and the domain-specific SciBERT (Beltagy et al., 2019). Further details about the model implementation and hyperparameter settings can be found in Appendix 5.7.3. We use macro F1-score as evaluation metric. All experiments were run over three different seeds and the results reported are the mean. [8]

5.5.2   Cross-dataset Evaluation

We test the following training configurations:[9] (1) *cross-dataset*: Training on SemEval-2018 and testing on SCIERC, and vice versa; (2) *cross-annotation*: Training on a mix of SemEval-2018 and SCIERC overlap: (2.1) *exclusive*: Considering either abstracts from the two corpora, (2.2) *repeated labeling*: Including every abstract twice, once from each dataset; this approach repeats instances with different annotations and is a simple method to handle divergences in annotation (Sheng et al., 2008; Uma et al., 2021a), (2.3) *filter*: Double annotation of the abstracts as in (2.2), but filtering out conflicting annotations.

Results   Table 5.7 reports the results of the experiments. The *cross-dataset* experiments (1) confirm the expected drop across datasets, in both directions (Sem: 40.28 → 34.81 and SCI: 34.29 → 31.37). Considering the *cross-annotation* setups, results are mixed in the *exclusive* version (2.1). The overall amount of training data is the same as the cross-dataset experiments, but there is less dataset-specific data, which hurts SemEval-2018.

---

[8] https://github.com/elisabassignana/scientific-re
[9] The development set follows the train set distributions.

Model						CNN		Transformer [tuned]	
Word embedding			FastText			BERT	SciBERT	SciBERT	SciBERT
↓Test | Train (NLP) →	Sem	Sci	[½ + ½]	2A	2A w/o CR	2A w/o CR	2A w/o CR	2A	2A w/o CR
SemEval NLP	40.28	34.81	39.91	50.17	48.95	42.54	49.27	79.16	77.79
SciERC NLP	31.37	34.29	36.29	39.36	41.48	38.63	51.99	67.36	69.90
SciERC AI-ML	37.00	50.44	46.78	49.52	49.66	40.81	51.14	72.48	76.80
SciERC CV	33.32	41.30	37.24	44.59	45.60	38.51	48.18	73.55	76.11
SciERC SPEECH	29.60	35.00	33.71	35.39	35.11	31.62	42.72	64.17	65.21
avg.	34.31	39.17	38.79	43.81	**44.16**	38.34	**48.66**	71.34	**73.56**

Table 5.7: Macro F1-scores of the cross-dataset and cross-domain experiments. (2.1) [½ + ½] refers to the case in which the train is made half by SemEval-2018 and half by SciERC; (2.2) 2A means double annotation from the two datasets; (2.3) CR are the conflicting relations (bold sample in Table 5.4).

In contrast, regarding (2.2) and (2.3), in both setups improvements are evident on both test sets. Compared to (2.1), the training data amount is effectively doubled and the model benefits from it. Removing the conflicting instances results in a slightly smaller train set, but an overall higher average performance ($43.81 \rightarrow 44.16$). The improvement of (2.3) over (2.2) is significant, which we test by the *almost stochastic dominance* test (Dror et al., 2019). Details about significance are in Appendix 5.7.4.

5.5.3 Contextualized Word Embeddings

We pick the best performing training scenario (*cross-annotation filter*, 2.3) and compare fastText with contextualized embeddings: BERT and the domain-specific SciBERT. The central columns of Table 5.7 report the results. While BERT does not bring relevant improvements over the best fastText setup, SciBERT confirms the strength of domain-specific trained language models (improvement of 4.5 F1 points and *almost stochastic dominance*). Compared to the CNN, full transformer fine-tuning results in the best model (rightmost columns). We tested different setups to feed the input to the transformer (see appendix 5.7.5), finding two entity spans and the full sentence as best setup. The full fine-tuned transformer model confirms the *dominance* of training setup (2.3) over (2.2).
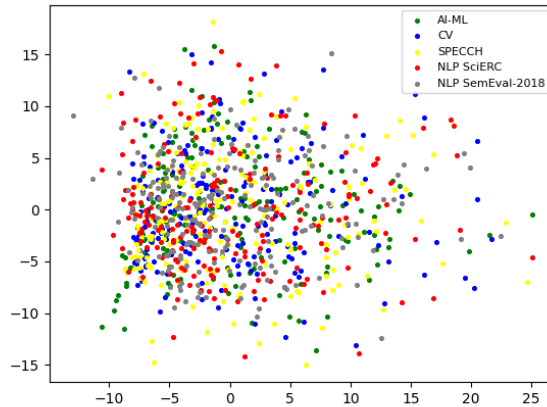
Figure 5.3: PCA representation of the CNN hidden state (just before the linear layer) using SciBERT.

5.5.4   Cross-domain Evaluation

Next, we look at *cross-domain* variation: Training on NLP, and testing on all sub-domains. The lower rows in Table 5.7 show the results. If we focus on the SciBERT models, we observe that there is some drop in performance from NLP, but mostly to CV and SPEECH. Interestingly, in some cases, AI-ML even outperforms the in-domain performance. Over all models, the SPEECH domain shows the clearest drop in transfer from NLP.[10] From an analysis of the predictions of the RC trained on SciBERT, we notice that the classifier struggles with identifying the most frequent USAGE relation (see Appendix 5.7.2) across sub-domains (confusion from lowest to highest: AI-ML, CV and SPEECH), and it is most confused with MODEL–FEATURE. Figure 5.6 in Appendix 5.7.6 contains the detailed confusion matrices. The overall evaluation suggests that in this setup sub-domain variation impacts RC performance to a limiting degree only.

In order to confirm this qualitatively, we (1) inspect whether model-internal representations are able to capture sub-domain variation, and we

---

[10]We note that the data amount for speech is the smallest in respect to the other sub-corpora, which might have an impact.

Domain	# word types	# overlap	% overlap
NLP	5,646	-	-
AI-ML	1,895	917	48.39%
CV	3,387	1,205	35.58%
SPEECH	1,398	715	51.14%

Table 5.8: Vocabulary overlap between NLP and the other sub-domains. *# word types*, *# overlap* in word types, and *% overlap* as relative percentages. Note that the amount of abstracts varies, cf. Table 5.6.

(2) test whether sub-domain variation is identifiable. To answer (1), we visualise the PCA representation of the CNN trained on setup (2.3) with SciBERT. The result is shown in Figure 5.3. The plot confirms that the representations do not contain visible clusters: The relation instances from each sub-domain are equally spread over it, and thus the performance of the relation classifier is similar for each of them. Our intuition is that the unified label set contains relations general enough to be equally covered by every sub-domain.

We explore the sub-domains more deeply apart from the RC task. To answer (2), we built a domain classifier to investigate how hard it is to tear apart the sub-domains. We hypothesize that, if sub-domains are distinguishable, a classifier should be able to easily distinguish them by looking at the coarsest level (the abstract). The classifier consists of a linear layer on top of the SciBERT encoder and achieves a F1-score of 62.01, over a random baseline of 25.58. This shows that the sub-domains are identifiable at the abstract level but with modest performance. As we would expect, SPEECH and NLP are highly confused (Figure 5.7 in Appendix 5.7.7 reports the confusion matrix) and the large vocabulary overlap shown in Table 5.8 between these sub-domains confirms this observation. Overall, sub-domains are identifiable but have limited impact on the RC task in the setup considered.

5.6 Conclusions

We present a survey on datasets for RE, revisit the task definition, and provide an empirical study on scientific RC. We observe a domain shift in RE datasets, and a trend towards multilingual and FSL for RE. Our analysis shows that our surveyed ACL RE papers focus mostly on RC and assume gold entities. Other steps are more blurry, concluding with a call for reporting RE setups more clearly.

As testing on only one dataset or domain bears risks of overestimation, we carry out a cross-dataset evaluation. Despite large data overlaps, we find annotations to substantially differ, which impacts classification results. Sub-domains extracted from meta-data instead only slightly impact performance. This finding on sub-domain variation is specific to the explored RC task on the scientific setup considered. Our study contributes to the first of three cross-domain RE setups we propose (Section 5.2) to aid further work on generalization for RE.

Limitations and Ethical Considerations

This work focuses on a limited view of the whole RE research field. Our dataset survey excludes specific angles of RE such as temporal RE or bioNLP, as they are large sub-fields which warrant a dedicated analysis in itself. From a methodological point of view, in our analysis we did not further cover weakly-supervised (e.g., distant supervision) and unsupervised approaches. Finally, given that our study points out gaps in RE, specifically cross-dataset, our experiments are still limited to RC only and next steps are to extend to the whole pipeline and to additional datasets and domains.

The data analyzed in this work is based on existing publicly-available datasets (based on published research papers).

Acknowledgements

5.7 Appendix

5.7.1 SCIERC Conference Division

The metadata relative to the IDs of the SCIERC abstracts contains information about the proceedings in which the papers have been published. We use this information to divide SCIERC into four sub-domains as shown in Table 5.9.

5.7.2 Data Analysis

Figure 5.4 reports the gold label distribution over SemEval-2018 and SCIERC respectively.

Figure 5.5, instead, contains the gold label distributions of SCIERC sub-domains over the five matching labels between the two datasets (see Table 5.3).

5.7.3 Model Details

Our RC model is a CNN with four layers (Nguyen and Grishman, 2015b). The layers consist of lookup embedding layers for word embeddings and entity position information (detailed below), convolutional layers with n-gram kernel sizes (2, 3 and 4), a max-pooling layer and a linear softmax relation classification layer with dropout of 0.5. Each input to the network

Conference	# abs
Artificial Intelligence - Machine Learning (AI-ML)	52
NeurIPS Neural Information Processing Systems	20
IJCAI International Joint Conference on Artificial Intelligence	14
ICML International Conference on Machine Learning	10
AAAI Association for the Advancement of Artificial Intelligence	8
Computer Vision (CV)	105
CVPR Conference on Computer Vision and Pattern Recognition	66
ICCV International Conference on Computer Vision	23
ECCV European Conference on Computer Vision	16
Speech	35
INTERSPEECH Annual Conference of the International Speech Communication Association	25
ICASSP International Conference on Acoustics, Speech, and Signal Processing	10
Natural Language Processing (NLP)	308
ACL Association for Computational Linguistics	307
IJCNLP International Joint Conference on Natural Language Processing	1

Table 5.9: SCIERC division into conferences and relative amount of abstracts for each of them.

(a) SemEval-2018
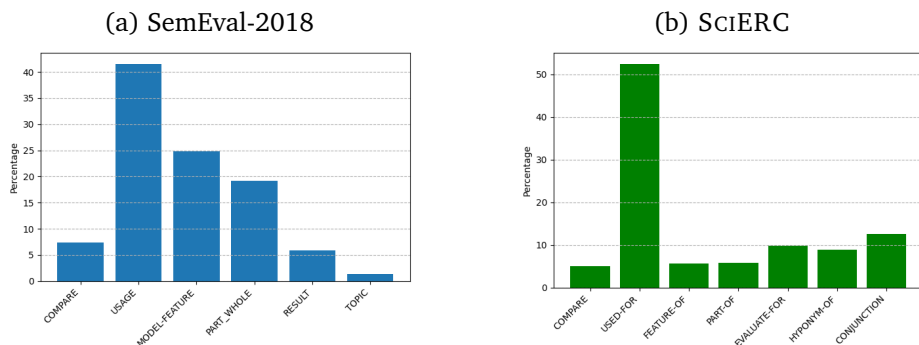
(b) SCIERC

Figure 5.4: Gold label distribution in the SemEval-2018 sub-task (1.1) and SCIERC datasets.
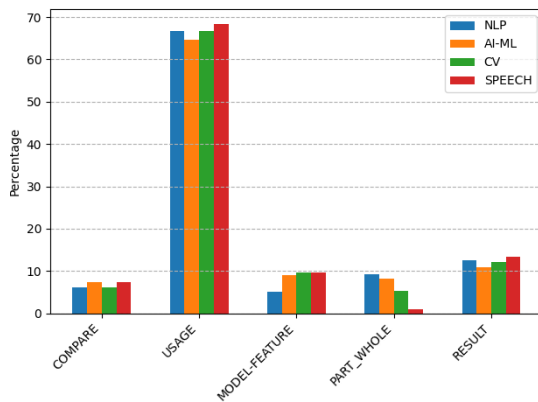
Figure 5.5: Gold label distribution of the five considered relations over SCIERC sub-domains.

is a sentence containing a pair of entities—which positions in the sentence are given—and a label within R, the set of five considered relations.

We experiment with three types of pre-trained *word embeddings*: one non-contextualized, fastText (Bojanowski et al., 2017), and two contextualized representations, BERT (Devlin et al., 2019) and the domain-specific SciBERT (Beltagy et al., 2019). For word split into subword-tokens, we adopt the strategy of keeping only the first embedding for each token. For every token we also consider two *position embeddings* following Nguyen and Grishman (2015b). Each of them encodes the relative distance of the token from each of the two entities involved in the relation.

Hyperparameters were determined by tuning the model on a held-out development set.

All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU.

5.7.4   Significance Testing

We compare our setups using Almost Stochastic Order (ASO; Dror et al., 2019).[11] Given the results over multiple seeds, the ASO test determines whether there is a stochastic order. The method computes a score (ϵ_{min}) which represents how far the first is from being significantly better in respect to the second. The possible scenarios are therefore (*a*) $\epsilon_{min} = 0.0$ (*truly stochastic dominance*) and (*b*) $\epsilon_{min} < 0.5$ (*almost stochastic dominance*). Table 5.10 reports the ASO scores with a confidence level of $\alpha = 0.05$ adjusted by using the Bonferroni correction (Bonferroni, 1936). See Section 5.5 for the setup details.

5.7.5   Transformer setups

The MaChAmp toolkit (van der Goot et al., 2021b) allows for a flexible amount of textual inputs (separated by the [SEP] token) to train the transformer and test the fine-tuned model on. We used SciBERT (Beltagy et al., 2019) and tested the following input configurations:

---

[11]Implementation by Ulmer (2021).

	2A [fastText]*	2A w/o CR [fastText]*	2A w/o CR [BERT]*	2A w/o CR [SciBERT]*	2A [SciBERT]†	2A w/o CR [SciBERT]†
2A [fastText]*	-	1.0	**0.0**	1.0	1.0	1.0
2A w/o CR [fastText]*	**0.0**	-	**0.0**	1.0	1.0	1.0
2A w/o CR [BERT]*	1.0	1.0	-	1.0	1.0	1.0
2A w/o CR [SciBERT]*	**0.0**	**0.0**	**0.0**	-	1.0	1.0
2A [SciBERT]†	**0.0**	**0.0**	**0.0**	**0.0**	-	1.0
2A w/o CR [SciBERT]†	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	-

Table 5.10: ASO scores of the main experimental setups described in Section 5.5. (*) CNN model. (†) full fine-tuned transformer model. Read as row → column.

1. The two entities:
   [*ent-1* [SEP] *ent-2*]

2. The sentence containing the two entities:
   [*sentence*]

3. The two entities and the sentence containing them:
   [*ent-1* [SEP] *ent-2* [SEP] *sentence*]

4. For the third setup, we introduce a marker between the two entities, resulting in a 2-inputs configuration:
   [*ent-1* [MARK] *ent-2* [SEP] *sentence*]

5. Finally—following Baldini Soares et al. (2019)—we augment the input sentence with four word pieces to mark the beginning and the end of each entity mention ([E1-START], [E1-END], [E2-START], [E2-END]):
   [*sentence-with-entity-markers*]

↓Test | Input Setup →	①	②	③	④	⑤
SEMEVAL NLP	58.15	42.08	77.79	74.85	75.12
SCIERC NLP	51.42	42.16	69.90	69.09	71.32
SCIERC AI-ML	54.63	40.35	76.80	75.08	74.93
SCIERC CV	53.16	41.09	76.11	74.73	74.21
SCIERC SPEECH	49.59	40.42	67.21	66.78	67.56
avg.	53.39	41.22	**73.56**	72.11	72.63

Table 5.11: Macro F1-scores of the RC using SciBERT (Beltagy et al., 2019) within the MaChAmp toolkit (van der Goot et al., 2021b). Setups 1-5 described in Appendix 5.7.5.

Table 5.11 reports the results of the experiments using MaChAmp on the setups described above.

5.7.6   Scientific Sub-domain Analysis

Figure 5.6 contains the confusion matrices of the CNN trained with SciBERT for the AI-ML, CV and SPEECH sub-domains. For fair comparison between the different data amounts the numbers reported are percentages.

5.7.7   Conference Classifier

Figure 5.7 represents the confusion matrix relative to the conference classifier described in Section 5.5.4.

(a) AI-ML (52 abstracts)

(b) CV (105 abstracts)

(c) SPEECH (35 abstracts)
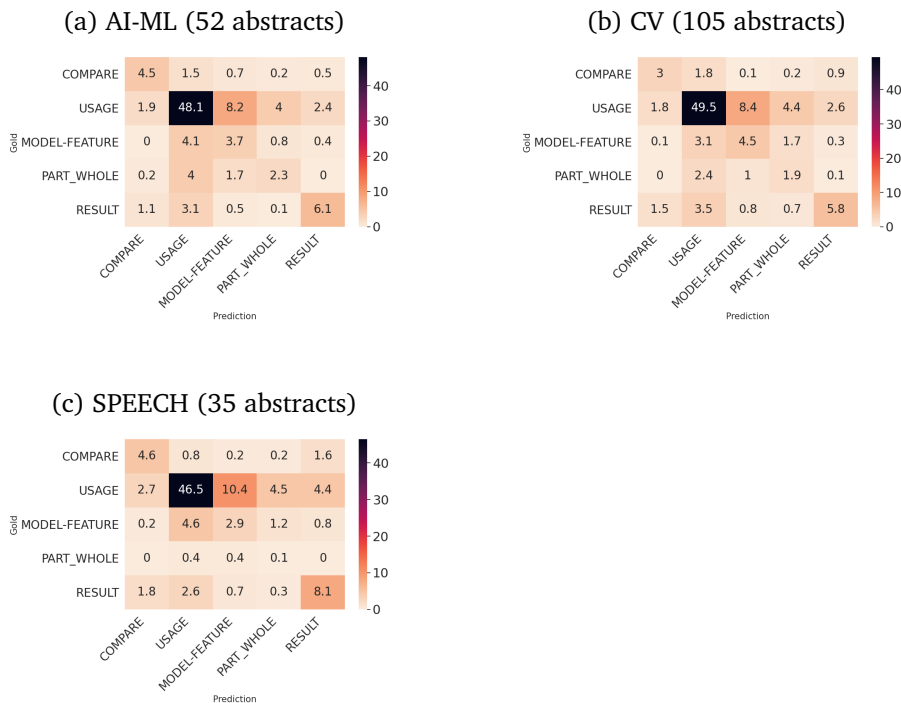
Figure 5.6: Percentage confusion matrices of the CNN on SCIERC sub-domains.
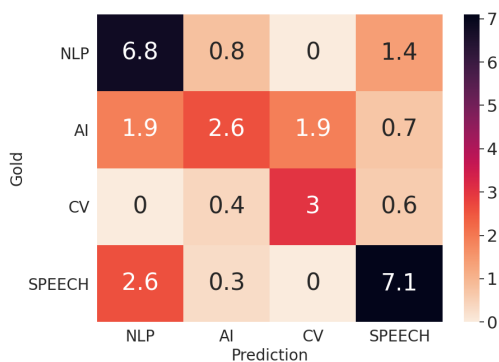
Figure 5.7: Confusion matrix of the conference classification experiment. The numbers reported are the average over three runs on different seeds.

Chapter 6

CrossRE: A Cross-Domain Dataset for Relation Extraction

Abstract

Relation Extraction (RE) has attracted increasing attention, but current RE evaluation is limited to in-domain evaluation setups. Little is known on how well a RE system fares in challenging, but realistic out-of-distribution evaluation setups. To address this gap, we propose CROSSRE, a new, freely-available cross-domain benchmark for RE, which comprises six distinct text domains and includes multi-label annotations. An additional innovation is that we release *meta-data* collected during annotation, to include explanations and flags of difficult instances. We provide an empirical evaluation with a state-of-the-art model for relation classification. As the meta-data enables us to shed new light on the state-of-the-art model, we provide a comprehensive analysis on the impact of difficult cases and find correlations between model and human annotations. Overall, our empirical investigation highlights the difficulty of cross-domain RE. We release our dataset, to spur more research in this direction.[1]

6.1   Introduction

Relation Extraction (RE) is the task of extracting structured knowledge from unstructured text. Although the fact that the task has attracted increasing attention in recent years, there is still a large gap in comprehensive evaluation of such systems which include out-of-domain setups (Bassignana and Plank, 2022b). Despite the drought of research on cross-domain evaluation of RE, its practical importance remains. Given the wide range of applications for RE to downstream tasks which can vary from question answering, to knowledge-base population, to summarization, and to all kind of other tasks which require extracting structured information from unstructured text, out-of-domain generalization capabilities are extremely beneficial. It is essential to build RE models that transfer well to new unseen domains, which can be learned from limited data, and work well

---

[1] https://github.com/mainlp/CrossRE

even on data for which new relations or entity types have to be recognized.

One direction which is gaining attention is to study RE systems under the assumption that new relation types have to be learned from few examples (*few-shot learning*; Han et al., 2018; Gao et al., 2019). One other direction is to study how sensitive a RE system is under the assumption that the input text features change (*domain shift*; Plank and Moschitti, 2013). There exists a limited amount of studies that focus on the latter aspect, and—to the best of our knowledge—there exists only one paper that proposes to study both, few-shot relation classification under domain shift (Gao et al., 2019). However, this last work considers only two domains—Wikipedia text for training and biomedical literature for testing—and has been criticized for its unrealistic setup (Sabo et al., 2021). In this paper, we propose CROSSRE, a new challenging cross-domain evaluation benchmark for RE for English (samples in Figure 6.1). CROSSRE is manually curated with hand-annotated relations covering up to 17 types, and includes multi-label annotations. It contains six diverse text domains, namely: news, literature, natural sciences, music, politics and artificial intelligence. One of the challenges of CROSSRE is that both entities and relation type distributions vary considerably across domains. CROSSRE is heavily inspired by CrossNER (Liu et al., 2021b), a recently proposed challenging benchmark for Named Entity Recognition (NER). We extend CrossNER to RE and collect additional meta-data including explanations and flags of difficult instances. To the best of our knowledge, CROSSRE is the most diverse RE datasets available to date, enabling research on domain adaptation and few-shot learning. In this paper we contribute:

- A new, comprehensive, manually-curated and freely-available RE dataset covering six diverse text domains and over 5k sentences.

- We release meta-data collected during annotation, and the annotation guidelines.

- An empirical evaluation of a state-of-the-art relation classification model and an experimental analysis of the meta-data provided.
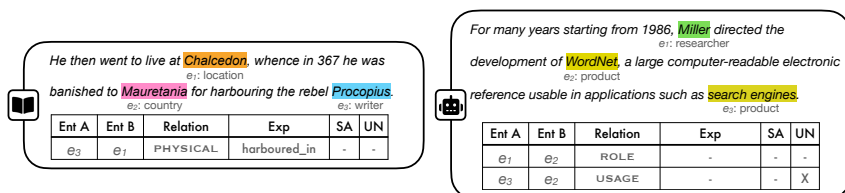
Figure 6.1: **CROSSRE Samples from Literature and Artificial Intelligence Domains.** At the top, the relation is enriched with the EXPLANATION (Exp) "harboured_in". At the bottom, instead, the second relation is marked with UNCERTAINTY (UN) by the annotator.

6.2 Related Work

Despite the popularity of the RE task (e.g. Nguyen and Grishman, 2015b; Miwa and Bansal, 2016; Baldini Soares et al., 2019; Wang and Lu, 2020; Zhong and Chen, 2021), the cross-domain direction has not been widely explored. There are only two datasets which can be considered an initial step towards cross-domain RE. The ACE dataset (Doddington et al., 2004) has been analyzed considering its five domains: news (broadcast news, newswire), weblogs, telephone conversations, usenet and broadcast conversations (Plank and Moschitti, 2013; Nguyen and Grishman, 2014, 2015a). In contrast to ACE, the domains in CROSSRE are more distinctive, with specific and more diverse entity types in each of them.

More recently, the FewRel 2.0 dataset (Gao et al., 2019), has been published. It builds upon the original FewRel dataset (Han et al., 2018)— collected from Wikipedia—and adds a new test set in the biomedical domain, collected from PubMed.

6.3 CrossRE

6.3.1 Motivation

RE aims to extract semantically informative triples from unstructured text. The triples comprehend an ordered pair of text spans which represent
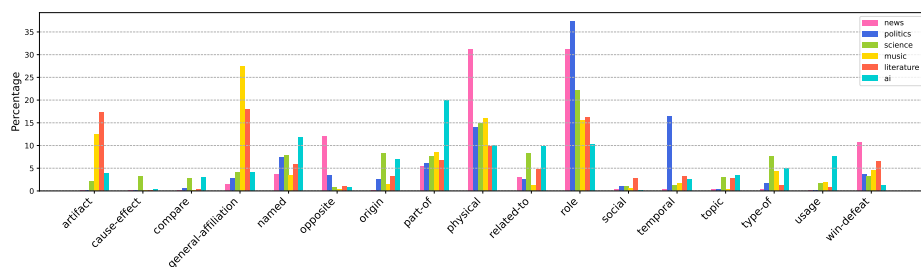
Figure 6.2: **CROSSRE Label Distribution.** Percentage label distribution over the 17 relation types divided by CROSSRE's six domains. Detailed counts and percentages in Appendix 6.7.4.

named entities or mentions, and the semantic relation which holds between them. The latter is usually taken from a pre-defined set of relation types, which typically changes across datasets, even within the same domain. The absence of standards in RE leads to models which are designed to extract specific relations from specific datasets. As a consequence, the ability to generalize over out-of-domain distributions and unseen data is usually lacking. While such specialized models could be useful in applications where particular knowledge is required (e.g. the bioNLP field), in most of the cases a more generic level is enough to supply the information required for the downstream task. In conclusion, RE models that are able to generalize over domain-specific data would be beneficial in terms of both costs of developing and training RE systems designed to work in pre-defined scenarios. To fill this gap, and in order to encourage the community to explore more the cross-domain RE angle, we publish CROSSRE, a new dataset for RE which includes six different domains, with a unified label set of 17 relation types.[2]

6.3.2 Dataset Overview

CROSSRE includes the following domains: news (▦), politics (🏛), natural science (🍃), music (🎵), literature (📖) and artificial intelligence (🤖;

---

[2]Our data statement (Bender and Friedman, 2018) can be found in Appendix 6.7.1.

AI). Our semantic relations are annotated on top of CrossNER (Liu et al., 2021b), a cross-domain dataset for NER which contains domain-specific entity types.[3] The news domain (collected from Reuters News) corresponds to the data released for the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003), while the other five domains have been collected from Wikipedia. The six domains have been proposed and defined by previous work, and shown to contain diverse vocabularies. We refer to Liu et al. (2021b) for details on e.g. vocabulary overlap across domains.

During our relation annotation process, we additionally correct some mistakes in named entities previously annotated in CrossNER (entity type, entity boundaries), but only revise existing entity mentions involved in a semantic relation, as well as add new entities involved in semantic relations (see samples in Appendix 6.7.3).

The final dataset statistics are reported in Table 6.1. We keep the train/dev/test data split by Liu et al. (2021b) and because of resource constraints, we fix as lower bound the sentence amount of the smallest domain (AI). We pursue their design choice of making training sets relatively small as cross-domain models are expected to do fast adaptation with a small-scale of target domain data samples. Our annotations are at the sentence-level, and the number of relations indicates the amount of directed entity pairs which are annotated with at least one of the 17 relation labels.

The final dataset contains 17 relation labels for the six domains: PART-OF, PHYSICAL, USAGE, ROLE, SOCIAL, GENERAL-AFFILIATION, COMPARE, TEM-PORAL, ARTIFACT, ORIGIN, TOPIC, OPPOSITE, CAUSE-EFFECT, WIN-DEFEAT, TYPE-OF, NAMED, and RELATED-TO. The latter, very generic, encapsulates all the semantic relations occurring with an extremely low frequency. With this label we make a step forward in respect to Sabo et al. (2021) which merge the 'other' and 'no-relation' cases into the 'None-of-the-above' (NOTA) label. We provide the description of each relation type in Appendix 6.7.2, and the full annotation guidelines in our repository. The resulting label distribution

---

[3]https://github.com/zliucr/CrossNER/tree/main/ner_data

	SENTENCES				RELATIONS			
	train	dev	test	**tot.**	train	dev	test	**tot.**
📧	164	350	400	914	175	300	396	871
🏛	101	350	400	851	502	1,616	1,831	3,949
🍃	103	351	400	854	355	1,340	1,393	3,088
🎵	100	350	399	849	496	1,861	2,333	4,690
📖	100	400	416	916	397	1,539	1,591	3,527
🤖	100	350	431	881	350	1,006	1,127	2,483
tot.	668	2,151	2,446	**5,265**	2,275	7,662	8,671	**18,608**

Table 6.1: **CROSSRE Statistics.** Number of sentences and number of relations annotated for each domain.

is illustrated in Figure 6.2, showing that relations vary substantially across domains. We will return to this point in the experimental section and provide further details in the next Section. After that, we describe the process that resulted in the final annotation guidelines and relation types. This includes the details on annotation agreement.

As mentioned, our guidelines allow for *multi-label annotations* (Jiang et al., 2016). This means that each entity pair can be assigned to multiple relation types—except for the RELATED-TO label which is exclusive and has to be used when none of the other 16 labels fit the data (see example in Appendix 6.7.5). The combination of labels enables more precise annotations which better represent the meaning expressed in the text (e.g. domain-specific scenarios), by keeping the relation label set relatively small and generic, as motivated in Section 6.3.1. Overall, 6% of the relations in CROSSRE are annotated with multiple labels, specifically: 📧 2%, 🏛 15%, 🍃 5%, 🎵 4%, 📖 2%, and 🤖 4%. Note that because of the directionality of the relations, entity pairs containing the same entities, but reverse order, do not count as multi-labeled.

6.3.3 Label Distributions

CROSSRE includes the same label set over its six domains. This implementation choice is motivated by the aim of studying cross-domain RE models which are able to generalize over domain-specific data, and abstract to

non-domain-specific relations. The result is a dataset with divergent label distributions across the different domains. Figure 6.2 shows the label distribution over CROSSRE.

From the individual distributions emerges the distinctiveness of each domain. News includes mainly OPPOSITE and WIN-DEFEAT relations referring to wars, countries being against each other, or sport news about matches between different teams; PHYSICAL, as many instances include the actual location of the news, and ROLE given that most instances in news are about describing business relationships between organizations or countries.

The politics domain contains OPPOSITE and WIN-DEFEAT, typically political parties and politician being against each other and winning, or losing the elections; the elections, mentioned quite often, usually supply information about the time and so are linked to other entities with the TEMPORAL relation. Last, the politics domain presents a high amount of ROLE relations as most of the sentences describe business relations between politicians and political parties or organizations.

Natural science presents a more homogeneous distribution. Distinctively, but similar to AI, which also contain technical text, a higher percentage of relations in respect to the other domains are annotated as RELATED-TO, as they would require specialized labels. Furthermore, similar to AI, the ORIGIN label stands out by linking ideas, algorithms, and inventions described in such domains to scientists and researchers. In AI the NAMED relation is also distinctively used, given the wide use in this field of acronyms preceded by their extension.

Last, music and literature have a particular high number of ARTIFACT labels describing songs, albums and books made and written by bands, musicians and writers, and GENERAL-AFFILIATION relations linking songs, albums, musicians, books and writers to specific music and literary genres.

6.3.4 Annotation Guidelines Definition Process

We bootstrap the dataset starting with a traditional top-down process, using an initial set of existing labels (Doddington et al., 2004; Hendrickx et al., 2010; Gábor et al., 2018; Luan et al., 2018), but continue by following a bottom-up approach (*data-driven annotation*), with the goal to annotate all the semantic relations present in the data, while balancing a trade-off between specificity (to domain-specific labels) and generalizability (Pustejovsky and Stubbs, 2012). The whole process (annotation guideline definition and data annotation) lasted around seven months, and is depicted next.

The guidelines have been defined via an iterative process including a total of seven annotation rounds (two preliminary and five official rounds). The two preliminary rounds have been completed by in-house NLP experts, with one round in the entire lab. The latter has been particularly crucial for collecting different points of view about the relations present in the dataset. After those, a hired expert with a linguists degree (who is the official annotator of the dataset) entered the process and the five official rounds began. These last rounds have been performed by the linguist together with one NLP expert, in consultation with a third NLP expert during the plenary discussion rounds.

The annotators in the official rounds were allowed to use the labels from the defined set, and were asked to explain their choice with a more fine-grained type (written in free text, typically as a predicate like 'won_award'). In addition, they were initially allowed to define new relation labels if a case was not fitting in any of the proposed ones. Each annotation round was carried out individually by each annotator and was followed by a plenary discussion. During the latter the given guidelines were reviewed and modified for the next annotation round. The process continued until the current high annotation agreement was achieved (see Section 6.3.5), after which the professional annotator continued to annotate the rest. This took close to 5 months of near full-time (0.8 fte) employment.

EXPLANATION (EXP)					

On 12 April 2019 a new Eurosceptic party, the **Brexit Party** *was officially launched by former* **UK Independence Party** *Leader* **Nigel Farage** .

	e_1: political party		e_2: political party	e_3: politician

$(e_1, e_3,$ ORIGIN, EXP: founded_by) $(e_3, e_1,$ ROLE, EXP: founder_of) $(e_3, e_2,$ ROLE, EXP: former_leader_of)

SYNTAX AMBIGUITY (SA)				

Variants of the **back-propagation algorithm** *as well as* **unsupervised methods** *by* **Geoff Hinton** *and colleagues at the* **University of Toronto** *can be used [...]*

e_1: algorithm	e_2: misc	e_3: researcher	e_4: university

$(e_1, e_3,$ ORIGIN, SA: True) $(e_2, e_3,$ ORIGIN) $(e_3, e_4,$ ROLE) $(e_3, e_4,$ PHYSICAL)

UNCERTAINTY of the annotator (UN)			

DNA methyltransferase *is recruited to the site and adds* **methyl groups** *to the* **cytosine** *of the* **CpG dinucleotides** .

e_1: enzyme	e_2: misc	e_3: chemical compound	e_4: misc

$(e_1, e_2,$ RELATED-TO, UN: True) $(e_2, e_3,$ PART-OF, UN: True) $(e_3, e_4,$ PART-OF, UN: True)

Table 6.2: **Samples of Meta-data Annotations.** Annotation samples from CROSSRE which have been enriched with meta-data: EXPLANATION of the relation type assigned, SYNTAX AMBIGUITY which poses a challenge for the annotator, and UNCERTAINTY of the annotator.

6.3.5 Annotation Agreement

With the aim of a more fine-grained analysis of the annotation agreement, we split RE into its two task components: Relation Identification (RI) and Relation Classification (RC). The first is the identification task which given a sentence and two marked entities determines if there exist one of the 17 semantic relation between them. The second, more fine-grained, takes the positive sample from RI and, given the label set, classifies the instances into the specific relation types. Such division supported the guideline definition process in order to understand whether the label descriptions were not specific enough, or whether there was unclarity in detecting the presence of a relation at all.

As described in Section 6.3.4, the guideline definition has been an iterative process with five annotation rounds and Figures 6.3 and 6.4 report the annotation agreement between the linguist and the NLP expert. As the entity order is part of the annotation guidelines, we furthermore tease apart the directionality component for a deeper analysis of the annotation agreement.

In Figure 6.3 we see that when considering the direction—$(e_1, e_2) \neq (e_2, e_1)$—the RI agreement is lower as we are considering one additional constraint in respect to the looser setup where $(e_1, e_2) = (e_2, e_1)$. In

Figure 6.3: **RI Annotation Agreement.** F1 score of the identified relations during the official annotation rounds.

Figure 6.4: **RC Annotation Agreement.** Macro-F1 score of the assigned labels over the entity pairs identified by both annotators during the official annotation rounds.

Figure 6.4 RC presents, instead, an inverse trend which is motivated by the fact that if the annotators agree on the direction, they will more likely assign the same relation label.

Several interesting observation emerge during the process. First, the drop in round 2 for RC indicates that it was at first easier to identify a relation between two entities (as RI agreement increases) than determining the exact label (RC agreement decreases). Therefore, between round 2 and 3 the discussion was centered around specifying the relation type descriptions and their respective directionality in more detail. The effect of this is visible in the next rounds, which resulted at first in an annotation agreement drop for RI (and consequently slight drop in RC agreement), but starting from round 3 onwards we observe a steady increase: This is also the point that marked the final version of the annotation guidelines,

which remained stable and the annotators were trained to use them over rounds 3, 4, 5. The converging agreements (w/ and w/o direction) of round 5 for both RI and RC indicate that the annotators achieved high data quality, annotating relations correctly.

The last annotation round (Round 5) included 72 sentences (12 from each of the six domains) for a total of 2,284 tokens resulting in high agreement. In particular, RI agreement considering the direction of the entity pairs is 94.16 F1 and without considering it 96.44 F1. The RC agreement considering the direction is 86.65 Macro-F1, and without considering it 86.39 Macro-F1. Furthermore, as we check and correct the entity spans from the previous NER datasets (see Section 6.3.2), we additionally compute the entity annotation agreement. Regarding entities, the Span-F1 with respect to the original data source is 90.79 and 91.81 respectively for the official annotator and the NLP expert, while the Span-F1 between them increases to 94.43, indicating that there is high consistency in correcting the entities. In light of the increasing interest to question the strong assumption of one unique gold label (Plank et al., 2014; Basile et al., 2021), we also release the doubly-annotated data from the last round in our repository to spurge research on learning with human label variation.

6.3.6 Meta-data Annotation

By embracing the subjectivity of manually-curated datasets, we collect *metadata* (see data samples in Table 6.2). We hope this facilitates future analyses of the dataset, including new annotation iterations, and interpretability of the predictions.

We include an EXPLANATION field for adding notes or specifications regarding the label assigned. In the first example in Table 6.2, the first relation (ORIGIN) is motivated by e_1 having been founded by e_3. Similarly the second relation, which includes the same entities, but with inverse order given the directionality of the ROLE label—note that this is not counted as multi-labeled as also the order has to match. In the last triple ROLE assumes a different meaning and it is specified in the EXP field by

	📧	🏛	🍃	🎵	📖	🤖	tot.
Exp	138	479	421	777	1,036	448	3,299
SA	0	32	20	169	31	25	277
Un	6	17	126	23	37	238	447

Table 6.3: **Meta-data Statistics.** Amount of annotations which have been marked with the following metadata: Explanation (Exp), Syntax Ambiguity (SA), and Uncertainty of the annotator (Un). The counts refer to the sum over train, dev, and test.

'former_leader_of'. Furthermore, we include two check-boxes. One is for identifying the presence of Syntax Ambiguity, which poses a challenge for the annotator. In the second example in Table 6.2, while we can confidently state that e_2 has been originated by e_3, the scenario for e_1 is ambiguous, and therefore the first triple is marked with 'SA: True'. The other check-box, named Uncertainty, allows the indication of low confidence by the annotator on the relation identified or on the label assigned. For instance, the third example in Table 6.2 (from the science domain) contains technical text which may require deeper knowledge of an expert in the field, and so our annotator (a linguist) flagged the relations in it as Uncertainty. The meta-data described have been extremely useful for the guideline definition process.

Table 6.3 reports the statistics of the meta-data annotations. The domains where our annotator is less confident are natural science and AI, and these are also the ones which contain more technical text specific to the two respective fields.

6.4 Baseline Experiments

We provide the evaluation of a state-of-the-art model on the proposed dataset. To establish baselines, we train models over each of the proposed domains. Two major challenges affecting the dataset are the multi-label annotation setup and the highly sparse label distribution distinctive of each domain.

6.4.1 Experimental Setup

Within this first empirical evaluation of CROSSRE, and given the challenges highlighted above, we follow previous work (Han et al., 2018; Baldini Soares et al., 2019; Gao et al., 2019) and focus on Relation Classification (RC) only, leaving the complete RE task for future work. The goal of RC is to assign the correct relation types to the ordered entity pairs which have been identified as being semantically connected.

6.4.2 Model

Our RC model follows the current state-of-the-art by Baldini Soares et al. (2019). Given a sentence s and an ordered pair of entity mentions (e_1, e_2), we augment s with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} which delimit the start and end of the entity spans. Following Zhong and Chen (2021) we enrich the entity markers with information about the entity types. For example, given the following sentence s and entity mention pair (e_1, e_2):

Cunningham *played his entire 11-year career with the* *Philadelphia Eagles*
e_1: person e_2: organization

 s is augmented as:

<E1:person> *Cunningham* </E1:person> *played his entire 11-year career with the* <E2:organization> *Philadelphia Eagles* </E2:organization>

The above version of s is then fed into a pre-trained encoder (BERT; Devlin et al., 2019) and we denote the output representation by $\hat{s}$. The output representations of the two start markers are concatenated in $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$ and used for the relation type classification via a feed-forward neural network. Given a set of n relation labels, the latter consists of a linear layer with output size n, followed by a softmax activation function. Considering the amount of multi-labeled instances being only around the 6% over the whole dataset, ignoring them by using a single-head model which can be trained more easily resulted in the best choice.[4] We run our experiments over five random seeds. See Appendix 6.7.6 for hyperparameters settings.

---

[4]We previously tested a multi-head model for enabling multi-label predictions, but the per-label data is not enough to effectively train each of the head classifier.

	📧	🏛	🍃	🎵	📖	🏭	avg.
MICRO-F1	46.36	58.26	40.10	75.96	67.70	45.40	55.63
MACRO-F1	16.52	20.33	25.29	39.19	37.74	30.66	28.29
WEIGH.-F1	37.59	53.53	35.84	73.16	63.08	41.52	50.79

Table 6.4: **CROSSRE Baselines.** Results achieved by our baseline model on the RC task. Reported are the averages over five random seeds (see Table 6.11).

6.4.3   Results

Evaluation   For a better evaluation of the baseline, given the highly imbalanced label distributions of the six domains, we follow Harbecke et al., 2022 and compute the micro-averaged F1, as well as the macro-averaged F1 and the weighted F1. The macro-average does not consider the classes with a support set of 0 in the test set.[5] The per-class data scarcity of most of the labels over the different domains (see Table 6.9) means the Macro-F1 is lower with respect to the other two metrics. However, it provides a more realistic scenario of the per-class performance of the model, and of the difficulty that the sparsity of the relation types adds in an already challenging classification task with 17 labels.

General Scores   Table 6.4 reports the scores achieved by our RC model. The news domain is the only one based on CoNLL-2003 as opposed to the other five domains (CrossNER). The instances are mostly news headlines or very short news reports and so, even if the amount of annotated sentences is comparable with the other domains, the semantic relations present in these data are considerably fewer (see Table 6.1). This, in addition to the most imbalanced label distribution—predominantly ROLE, PHYSICAL, OPPOSITE, WIN-DEFEAT and PART-OF (see Figure 6.2)—leads news to be one the most challenging domain in term of Macro-F1. In contrast, the music domain, with the highest amount of annotated relations, achieves the highest scores in respect to the other domains.

---

[5]Evaluation code in our repository.

	📰	🏛	🌿	🎵	📖	🤖
ARTIFACT	-	0.0	17.93	85.74	86.13	52.55
CAUSE-EFFECT	-	-	0.0	0.0	0.0	0.0
COMPARE	-	0.0	45.39	0.0	0.0	0.0
GEN.-AFF.	0.0	24.49	29.19	87.04	84.46	3.07
NAMED	34.67	54.56	53.53	10.3	48.66	65.11
OPPOSITE	9.38	4.41	0.0	0.0	2.67	0.0
ORIGIN	-	0.0	26.7	32.79	0.0	42.51
PART-OF	0.0	2.2	19.71	38.33	11.06	49.11
PHYSICAL	45.8	71.12	73.06	91.23	76.43	76.79
RELATED-TO	0.0	0.0	41.51	11.81	8.98	27.44
ROLE	58.84	59.67	40.15	65.57	63.38	61.56
SOCIAL	-	0.0	0.0	0.0	50.34	0.0
TEMPORAL	0.0	85.72	0.0	32.68	63.45	51.66
TOPIC	-	0.0	1.14	0.0	9.48	30.73
TYPE-OF	-	0.0	6.57	79.29	59.73	18.68
USAGE	-	-	0.0	56.15	0.0	12.2
WIN-DEFEAT	0.0	2.77	75.06	75.31	76.75	29.78

Table 6.5: **Per-class Results.** Detailed F1 scores for each relation type. Reported are the averages over five random seeds (see Table 6.11). '-' indicates the class is not present in the test set.

Per-label Performance   In Table 6.5 we report the per-label F1 scores for a more detailed analysis. Several labels have just few samples in the training sets and so are very difficult to learn, leading to an F1 of 0.0. These cases push down the Macro-F1 scores in Table 6.4. Overall, the amount of instances per-label—see Figure 6.2 for percentages and Table 6.9 for counts—are good indicators for the individual scores in Table 6.5. For example GENERAL-AFFILIATION achieves high F1 both in the music and in the literature domains (87.04 and 84.46 respectively). This is similar in TEMPORAL in the politics domain (85.72). However, we notice that some labels are more challenging than others: While the ROLE label contains more instances than the TEMPORAL one in the politics domain, it only achieves a score of 59.67. Given the imbalanced train/dev/test split over the six domains, and in order to give a more realistic idea of the distributions, we report as an example the label distribution over the train/dev/test split of the politics domain in Appendix 6.7.7. We additionally notice that the same label can have different levels of challenge depending on the

	📧	🏛	🌿	🎵	📖	🤖
SA	0	15	8	150	6	20
Un	1	9	62	19	8	68
SA or Un	1	23	69	167	12	88

Table 6.6: **Test Set Statistics.** Amount of annotations which have been marked with SYNTAX AMBIGUITY (SA) and with UNCERTAINTY (UN) in the test sets.

domain. For example, NAMED corresponds to similar percentages in the domains of news and music (3.62% and 3.34% respectively), but given the disparate total amount of in-domain relations these correspond to very different amounts: 32 in news and 164 in music. However, the NAMED label achieves an F1 score of 34.67 in the news domain, and only 10.3 in the music domain.

6.5   Meta-data Analysis

In this section, we use the meta-data collected during the annotation of the dataset for further analysis. We consider SYNTAX AMBIGUITY (SA) and UNCERTAINTY of the annotator (UN) and examine the performance of our baseline model on such instances. Table 6.6 reports the meta-data statistics on the six test sets. Given the almost absence of samples in the news domain, we leave it out from this analysis. Table 6.7 shows the results of our model when evaluated on samples only with SA and UN, both, or none, compared to ALL. For this ablation study we do not report the Macro-F1 because changing the evaluation set would mislead the analysis (as mentioned, the Macro-F1 only considers classes present in the evaluation set).

With the low amount of instances in politics an literature, results are less pronounced and differences with the overall scores are absent in most cases. Therefore, we focus here on the remaining three domains—natural science 🌿, music 🎵, AI 🤖. We observe slightly but consistently higher

scores when taking out the cases marked with UN, showing that they are challenging not only for the human but also the system. Those are the cases identified as most challenging, specially considering the annotator's background (i.e. natural science and AI, mostly on CAUSE-EFFECT, PART-OF, USAGE). The results in respect to the SA annotations are mixed: There is not a unified trend over domains or metrics. We attribute this to the fact that our model does not explicitly build upon syntactic features (e.g. syntactic trees; Plank and Moschitti, 2013). Finally, the scores from the data which consider the combination of SA and UN increase over the baseline in the science domain, where taking out both SA and UN individually increase over the ALL setup. In the music domain, where SA are frequent, excluding them result in a little drop of Micro-F1 (75.96→74.67). In fact, the model is good on SA in the music domain: The majority of cases are on the GENERAL-AFFILIATION label, which achieves high per-label F1 (87.04). We attribute it to the fact that in this domain there are many lists of entities and relative attributes, which structurally can be ambiguous, but often involve a similar relation structure. AI presents a similar trend as music, but the scores from the combination of SA and UN increase a bit in both metrics.

In conclusion, we do gain informative insights from the collected meta-data—especially when the annotator is unsure about the annotated relation, and also to understand whether syntactic ambiguity detected by the annotator impacts system accuracy.

6.6 Conclusion

We present CROSSRE, a new challenging manually-curated corpus for RE. It is the first dataset for RE covering six diverse text domains (news, politics, natural science, music, literature, AI) with annotations spanning 17 relation types. Some annotations are enriched with meta-data information (explanation for the choice of the assigned label, identification of syntax ambiguity, and uncertainty of the annotator). Throughout the annotation process and in the empirical validation, this meta-data proves to be use-

			🏛	🌿	🎵	📖	🤖
Micro F1	ALL		58.26	40.10	75.96	67.70	45.40
	SA	w/	54.67	12.50	95.25	76.67	87.00
		w/o	58.27	40.26	74.67	67.66	44.68
	Un	w/	66.67	20.97	27.00	67.50	27.34
		w/o	58.21	40.97	76.38	67.70	46.58
	SA or Un	w/	57.39	20.29	87.04	70.00	40.67
		w/o	58.26	41.11	75.12	67.68	45.82
Weighted F1	ALL		53.53	35.84	73.16	63.08	41.52
	SA	w/	57.16	13.33	94.91	74.57	87.00
		w/o	53.62	36.00	71.66	63.06	40.97
	Un	w/	61.78	12.91	30.46	64.46	19.13
		w/o	53.50	36.68	73.59	63.11	42.95
	SA or Un	w/	55.62	13.18	86.79	64.24	32.43
		w/o	53.62	36.83	72.13	63.10	42.38

Table 6.7: **Meta-data Analysis.** F1 scores on the instances which have been marked with SYNTAX AMBIGUITY (SA), UNCERTAINTY (UN), or at least one of the two. We report also the baselines of Table 6.4 (ALL).

ful and insightful. As it aids the analysis of the provided baseline, we invite the research community to both collect and release such additional information.

We perform an empirical evaluation of CROSSRE by applying state-of-the-art RC methods (Baldini Soares et al., 2019; Zhong and Chen, 2021), and show the challenges of its highly imbalanced label distributions over the domains.

The cross-domain dimension is currently under-explored in the RE field. With this dataset we invite future work on cross-domain RE evaluation, the exploration of domain-adaptive techniques (e.g. DAPT; Gururangan et al., 2020) and other adaptation methods to improve the baseline set out in this work for the different data domains.

Limitations

Because of resource constraints (time and costs, see next Section), the proposed dataset is limited to one annotator. However, as our annotation process details show, we expect the quality to be high, nevertheless, preferably if resources were available, gaining larger subsets with multiple annotations would be a promising next step. Crucially, we involved the annotator in the guideline definition process, which was very fruitful and inspired us to collect syntax ambiguity information as well.

We identify as a second limitation the fact that five out of six of our domains belong to the same data source (Wikipedia). However, the advantage is that Wikipedia data can be redistributed freely. We acknowledge the already challenging setup of our dataset, but invite future work on the inclusion of different data sources whenever possible.

Ethics Statement

The data included in our newly proposed dataset correspond to a sub-set of the data collected and freely published by Liu et al. (2021b) within the CrossNER project.

Our dataset is annotated by a hired expert with a linguists degree employed on 0.8fte for this project following national salary rates. The total costs for data annotation amount to roughly 19,000 USD, amounting to $\approx$ 1\$ per annotated relation.

Acknowledgements

First, we would like to thank our annotator for the great job and substantial help given to this project. We thank Filip Ginter and Sampo Pyysalo for insightful discussion at the last stage of this work. Furthermore, we thank the NLPnorth group for feedback on an earlier version of this paper— in particular Rob van der Goot, Mike Zhang, and Max Müller-Eberstein. Last, we also thank the ITU's High-performance Computing cluster for

6.7 Appendix

6.7.1 Data Statement CROSSRE

Following (Bender and Friedman, 2018) we outline below the data statement fo CROSSRE.

A. CURATION RATIONALE: Collection of Reuters News and Wikipedia pages annotated with the aim of studying Relation Extraction.

B. LANGUAGE VARIETY: The language is English. For additional details we refer to Tjong Kim Sang and De Meulder, 2003 and to Liu et al., 2021b who did the data collection.

C. SPEAKER DEMOGRAPHIC: Unknown.

D. ANNOTATOR DEMOGRAPHIC: One sprofessional annotator with a background in Linguistics and one NLP expert with a background in Computer Science. Age range: 25–30; Gender: both female; Race/ethnicity: white European; Native language: Danish, Italian; Socioeconomic status: higher-educated.

E. SPEECH SITUATION: We refer to Tjong Kim Sang and De Meulder, 2003 and to Liu et al., 2021b.

F. TEXT CHARACTERISTICS: The texts are news from Reuters News, and Wikipedia pages about politics, natural science, literature, artificial intelligence.

G. RECORDING QUALITY: N/A

H. OTHER: N/A

I. PROVENANCE APPENDIX: The data statements of the previous datasets (Tjong Kim Sang and De Meulder, 2003; Liu et al., 2021b) are not available.

6.7.2 Relation Label Description

Below we report the description of each relation type we use to annotate CROSSRE. We refer to our repository for the complete annotation guidelines, including directionality of the relations, samples. and instruction on what to annotate.

- PART-OF Something that is part of something else (e.g. song_part_of_album, task_ part_of_field).

- PHYSICAL Answer the question *Where?* (e.g. location, near, destination, located_in, based_in, residence, released_in, come_from).

- USAGE Something which make use of something else in order to accomplish its scope, includes an agent using an instrument.

- ROLE Two entities which are linked by a *business related* role (e.g. management, founder, affiliate_partner, member_of, citizen_of, participant, nominee_of).

- SOCIAL Two entities linked by a *non-business related* role (e.g. parent, sibling, spouse, friend, acquaintance).

- GENERAL-AFFILIATION Religion, ethnicity, genre (e.g. book_genre, music_genre).

- COMPARE Something that is compared with something else.

- TEMPORAL Something that happens or exist during an event.

- ARTIFACT Something *concrete* which is the result of the work of someone (e.g. written_by, made_by).

- ORIGIN Something *abstract* which is originated by something else (e.g. invented, idea, title_obtained_by).

- TOPIC The topic or focus of something.

- OPPOSITE Something that is physically or idealistically opposite, contrary, against or inverse of something else.

- CAUSE-EFFECT An event or object which leads to an effect.

- WIN-DEFEAT Someone or something who has won or lost a competition, an award or a war (default is victory, in case of defeat it is specified in the 'Explanation' field).

- TYPE-OF The type, property, feature or characteristic of something.

- NAMED Two spans which refer to the same entity (e.g. nickname, acronym, second name or abbreviation of something or someone).

- RELATED-TO Two semantically connected entities which do not fall in any of the previous cases.

6.7.3   Entity Alteration Samples

In Table 6.8 we report one sample for each entity alteration type that we perform in respect to the original entity annotations from CrossNER (Liu et al., 2021b) and CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003). In the first sample, we correct the entity type of e_1 from 'conference' to 'organisation'. In the second sample, we extend e_2—which originally only contains an adjective—in order to include also the following noun. We do this because, following previous work on RE, our relation labels do not hold between adjectives only. Last, in the third sample we add the annotation for marking 'Squealer' as an entity.

6.7.4   Detailed Label Statistics

Table 6.9 contains the detailed label statistics (counts and percentages) for each domain.

ENTITY TYPE

Finally, every other year, [ELRA] organizes a major conference [LREC] , the [International Language Resources and Evaluation Conference] .

e_1: conference e_2: conference e_3: conference

Finally, every other year, [ELRA] organizes a major conference [LREC] , the [International Language Resources and Evaluation Conference] .

e_1: organisation e_2: conference e_3: conference

ENTITY BOUNDARIES

[China] controlled most of the match and saw several chances missed until the 78th minute when [Uzbek] striker [Igor Shkvyrin] took advantage [...]

e_1: country e_2: misc e_3: person

[China] controlled most of the match and saw several chances missed until the 78th minute when [Uzbek striker] [Igor Shkvyrin] took advantage [...]

e_1: country e_2: misc e_3: person

NEW ENTITIES

[Tamsin Greig] narrated, and the cast included [Nicky Henson] as [Napoleon] , [Toby Jones] as the propagandist Squealer, and [Ralph Ineson] as [Boxer] .

e_1: person e_2: person e_3: person e_4: person e_5: person e_6: person

[Tamsin Greig] narrated, and the cast included [Nicky Henson] as [Napoleon] , [Toby Jones] as the propagandist [Squealer] , and [Ralph Ineson] as [Boxer] .

e_1: person e_2: person e_3: person e_4: person e_4: person e_5: person e_6: person

Table 6.8: **Samples of Modified Entity Annotations.** Instances with the original annotations from CrossNER (Liu et al., 2021b) and corresponding sentences from CROSSRE with the corrected entities.

	News		Politics		Nat. Science		Music		Literature		AI	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
ARTIFACT	1	0.11	6	0.13	70	2.16	612	12.48	620	17.26	99	3.8
CAUSE-EFFECT	0	0.0	1	0.02	106	3.28	4	0.08	7	0.19	9	0.35
COMPARE	0	0.0	29	0.64	90	2.78	7	0.14	13	0.36	76	2.92
GENERAL-AFFILIATION	13	1.47	123	2.71	133	4.11	1,349	27.5	642	17.87	104	3.99
NAMED	32	3.62	338	7.45	251	7.76	164	3.34	209	5.82	306	11.74
OPPOSITE	106	11.98	154	3.4	28	0.87	21	0.43	32	0.89	21	0.81
ORIGIN	1	0.11	114	2.51	270	8.35	71	1.45	114	3.17	178	6.83
PART-OF	47	5.31	273	6.02	246	7.61	421	8.58	243	6.76	517	19.83
PHYSICAL	276	31.19	634	12.98	481	14.87	782	15.93	348	9.69	259	9.93
RELATED-TO	27	3.05	116	2.56	270	8.35	62	1.26	173	4.81	254	9.75
ROLE	275	31.07	1,695	37.38	716	22.14	767	15.64	578	16.09	269	10.32
SOCIAL	3	0.34	42	0.93	33	1.02	27	0.55	97	2.7	2	0.08
TEMPORAL	2	0.23	744	16.41	41	1.27	78	1.59	117	3.26	65	2.49
TOPIC	3	0.34	17	0.37	95	2.94	13	0.27	97	2.7	88	3.38
TYPE-OF	3	0.34	80	1.76	249	7.7	214	4.36	42	1.17	130	4.99
USAGE	1	0.11	1	0.02	55	1.7	95	1.94	25	0.7	199	7.63
WIN-DEFEAT	95	10.73	167	3.68	100	3.09	222	4.53	236	6.57	30	1.15
total	885		4,534		3,234		4,909		3,593		2,606	

Table 6.9: **Relation Label Statistics.** Absolute count and relative percentage of each relation label. Note that, because of the multi-label setup, these numbers are higher in respect to the relation counts in Table 6.1.

MULTI-LABEL ANNOTATION

He was the last former Prime Minister to lose his seat until [Tony Abbott] *lost his seat of* [Warringah] *in* [2019 Australian federal election] *, [...]*

e_1: politician e_2: location e_3: election

$(e_1, e_3,$ TEMPORAL$)$ $(e_1, e_3,$ WIN-DEFEAT$)$

Table 6.10: **Example of Multi-label Annotation.** Example from CROSSRE of an ordered entity pair which has been annotated with multiple relation labels.

Parameter	Value
Encoder	`bert-base-cased`
Classifier	1-layer FFNN
Loss	Cross Entropy
Optimizer	Adam optimizer
Learning rate	$2e^{-5}$
Batch size	32
Seeds	4012, 5096, 8878, 8857, 9908

Table 6.11: **Hyperparameters Setting.** Model details for reproducibility of the baseline.

6.7.5   Multi-label annotation

In Table 6.10 we report an example of multi-label annotation in which e_1, a politician entity, is related to e_3, an election. The entity pair is annotated both as TEMPORAL because it provides temporal information about *Tony Abbott*'s existence, and also as WIN-DEFEAT, to capture the fact that he lost the election mentioned in e_3.

6.7.6   Reproducibility

We report in Table 6.11 the hyperparameter setting of our RC model (see Section 6.4.2). All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU.

6.7.7   Label Distribution Per-Domain

Given the imbalance of the label distribution (see Figure 6.2) and of the train/dev/test splits (see Table 6.1), we report in Figure 6.5 as a sample the specific label distribution of the politics domain.
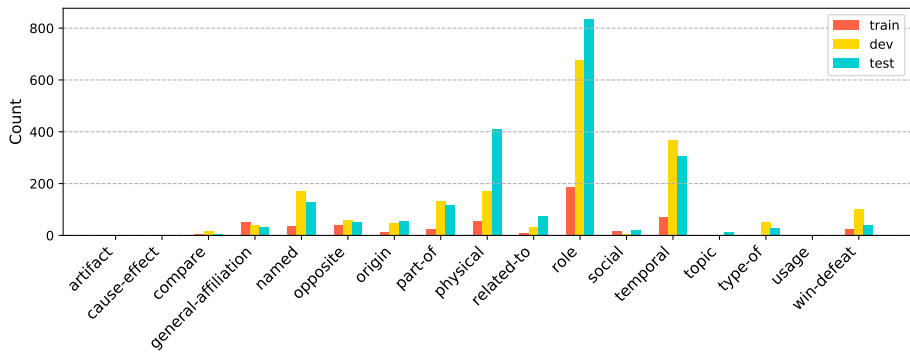
Figure 6.5: **Label Distribution of the Politics Domain.** Distribution of the 17 relation types over the train/dev/test split.

Chapter 7

Multi-CrossRE A Multi-Lingual Multi-Domain Dataset for Relation Extraction

Abstract

Most research in Relation Extraction (RE) involves the English language, mainly due to the lack of multi-lingual resources. We propose MULTI-CROSSRE, the broadest multi-lingual dataset for RE, including 26 languages in addition to English, and covering six text domains. MULTI-CROSSRE is a machine translated version of CrossRE (Bassignana and Plank, 2022a), with a sub-portion including more than 200 sentences in seven diverse languages checked by native speakers. We run a baseline model over the 26 new datasets and—as sanity check—over the 26 back-translations to English. Results on the back-translated data are consistent with the ones on the original English CrossRE, indicating high quality of the translation and the resulting dataset.

7.1 Introduction

Binary Relation Extraction (RE) is a sub-field of Information Extraction specifically aiming at the extraction of triplets from text describing the semantic connection between two entities. The task gained a lot of attention in recent years, and different directions started to be explored. For example, learning new relation types from just a few instances (few-shot RE; Han et al., 2018; Gao et al., 2019; Sabo et al., 2021; Popovic and Färber, 2022), or evaluating the models over multiple source domains (cross-domain RE; Bassignana and Plank, 2022b,a). However, a major issue of RE is that most research so far involves the English language only.

After the very first multi-lingual work from the previous decade—the ACE dataset (Doddington et al., 2004) including English, Arabic and Chinese—recent work has started again exploring multi-lingual RE. Seganti et al., 2021 published a multi-lingual dataset, built from entity translations and Wikipedia alignments from the original English version. The latter was collected from automatic alignment between DBpedia and Wikipedia. The result includes 14 languages, but with very diverse relation type distributions: Only English contains instances of all the 36 types, while

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with learning algorithms that analyze data used for classification and regression analysis.

Beim maschinellen Lernen sind Support-Vektor-Maschinen (SVMs, auch Support-Vektor-Netzwerke) überwachte Lernmodelle mit Lernalgorithmen, die Daten für Klassifizierungs- und Regressionsanalysen analysieren.

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with learning algorithms that analyse data for classification and regression analysis.

Figure 7.1: **Example sentence with color-coded entity markup.** From top to bottom: The original English text, its translation to German, and translation back to English. In the first translation step the entity *classification* is not transferred to German. In the second translation step the entity *machine learning* is (wrongly) expanded by a comma—later corrected in our post-processing.

the most low-resource Ukrainian contains only 7 of them (including the 'no_relation'). This setup makes it hard to directly compare the performance on different languages. Kassner et al., 2021 translated TREx (Elsahar et al., 2018) and GoogleRE,[1] both consisting of triplets in the form (object, relation, subject) with the aim of investigating the knowledge present in pre-trained language models by querying them via fixed templates. In the field of distantly supervised RE, Köksal and Özgür, 2020 and Bhartiya et al., 2022 introduce new datasets including respectively four and three languages in addition to English.

In this paper, we propose MULTI-CROSSRE, to the best of our knowledge the most diverse RE dataset to date, including 27 languages and six diverse text domains for each of them. We automatically translated CrossRE (Bassignana and Plank, 2022a), a fully manually-annotated multi-domain RE corpus, annotated at sentence level. We release the baseline results on the proposed dataset and, as quality check, on the 26 back-translations to English. Additionally, we report an analysis where native speakers in seven diverse languages manually check more than 200 translated sentences and the respective entities, on which the semantic relations are based. MULTI-CROSSRE allows for the investigation of sentence-level RE in the 27 languages included in it, and for direct performance comparison between them. Our contributions are: ① We propose a practical

---

[1] https://code.google.com/archive/p/relation-extraction-corpus

	SENTENCES				RELATIONS			
	train	dev	test	**tot.**	train	dev	test	**tot.**
🖥	164	350	400	914	175	300	396	871
🏛	101	350	400	851	502	1,616	1,831	3,949
🍃	103	351	400	854	355	1,340	1,393	3,088
🎵	100	350	399	849	496	1,861	2,333	4,690
📖	100	400	416	916	397	1,539	1,591	3,527
🤖	100	350	431	881	350	1,006	1,127	2,483
tot.	668	2,151	2,446	**5,265**	2,275	7,662	8,671	**18,608**

Table 7.1: **CrossRE Statistics.** Number of sentences and number of relations for each domain.

approach to machine-translate datasets with span-based annotations and apply it to produce MULTI-CROSSRE, the first multi-lingual and multi-domain dataset for RE including 27 languages and six text domains.[2] ② Multi-lingual and multi-domain baselines over the proposed dataset. ③ Comprehensive experiments over the back-translations to English. ④ A manual analysis by native speakers over more than 200 sentences in seven diverse languages.

7.2 MULTI-CROSSRE

CrossRE   As English base, we use CrossRE (Bassignana and Plank, 2022a),[3] a recently published multi-domain dataset. CrossRE is entirely manually-annotated, and includes 17 relation types spanning over six diverse text domains: artificial intelligence (🤖), literature (📖), music (🎵), news (🖥), politics (🏛), natural science (🍃). The dataset was annotated on top of CrossNER (Liu et al., 2021b), a Named Entity Recognition (NER) dataset. Table 7.1 reports the statistics of CrossRE.

Translation Process   With the recent progress in the quality of machine translation (MT), utilizing machine-translated datasets in training and evaluation of NLP methods has become a standard practice (Conneau et al.,

---

[2]https://github.com/mainlp/CrossRE
[3]Released with a GNU General Public License v3.0.

2018; Kassner et al., 2021). As long as the annotation is not span-bound, producing a machine-translated dataset is rather straightforward. The task however becomes more involved for datasets with annotated spans, such as the named entities in our case of the CrossRE dataset, or e.g. the answer spans in a typical question answering (QA) dataset. Numerous methods have been developed for transferring span information between the source and target texts (Chen et al., 2022). These methods are often tedious and in many cases rely on language-specific resources to obtain the necessary mapping. Some methods also require access to the inner state of the MT system, e.g. its attention activations, which is generally not available when commercial MT systems are used.

In this work, we demonstrate a practical and simple approach to the task of machine translating a span-based dataset. We capitalize on the fact that DeepL,[4] a commercial machine translation service very popular among users thanks to its excellent translation output quality, is capable of translating document markup. This feature is crucial for professional translators—the intended users of the service—who need to translate not only the text of the source documents, but also preserve their formatting. In practice, this means that the input of DeepL can be a textual document with formatting (a Word document) and the service produces its translated version with the formatting preserved.

For the CrossRE dataset, we only need to transfer the named entities, which can be trivially encoded as colored text spans in the input documents, where the color differentiates the individual entities. This is further facilitated by the fact that the entities do not overlap in the dataset, allowing for a simple one-to-one id-color mapping. Observing that oftentimes the entities are over-extended by a punctuation symbol during translation, the only post-processing we apply is to strip from each translated entity any trailing punctuation not encountered in the suffix of the original named entity. The process is illustrated in Figure 7.1, with details about two typical issues with this approach (later analysed in Section 7.4).[5]

---

[4] https://www.deepl.com/translator
[5] The overall translation process cost is ≈ 60€.

	🏛️	📖	🎵	📑	🏛	🍃	avg.
English	20.8	36.4	30.7	10.1	20.0	21.6	23.3

Table 7.2: **CrossRE Baseline Results.** Macro-F1 scores of the RC baseline over the original CrossRE English dataset.

7.3 Experiments

Model Setup   In order to be able to directly compare our results with the original CrossRE baselines on English, we follow the model and task setup used by Bassignana and Plank, 2022a. We perform Relation Classification (Han et al., 2018; Baldini Soares et al., 2019; Gao et al., 2019), which consists of assigning the correct relation types to the ordered entity pairs which are given as semantically connected. The model follows the current state-of-the-art architecture by Baldini Soares et al., 2019 which augments the sentence with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} surrounding the two entities. Following Zhong and Chen (2021) the entity markers are enriched with information about the entity types. The augmented sentence is then passed through a pre-trained encoder (XLM-R large; Conneau et al., 2020), and the classification made by a linear layer over the concatenation of the start markers $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$. We run all our experiments over five random seeds. See Appendix 7.6.1 for reproducibility and hyperparameters settings.

Results   The original CrossRE study reports the baseline experiments by using the mono-lingual BERT (Devlin et al., 2019) language encoder. In order to be able to compare the original baseline with the results on our MULTI-CROSSRE dataset, we re-run the English experiments by using the multi-lingual XLM-R large (Conneau et al., 2020) language encoder, and report the results in Table 7.2.

Table 7.3: **MULTI-CROSSRE Baseline Results.** Macro-F1 scores of the baseline model ordered by increasing lang2vec distance from English. Δ_{BT}: delta between back-translated and original evaluation when model trained on back-translated data. Δ_{OR}: delta between model trained on back-translated data and on original CrossRE data when evaluated on original CrossRE English.

Language	lang2vec	TRANSLATION (EN → X)							BACK-TRANSLATION (X → EN) — EVAL ON BACK-TRANSLATED DATA							EVAL ON ORIGINAL CROSSRE DATA							Δ_{BT}	Δ_{OR}
		🌐	▬	⌐	▦	⌂	♪	avg.	🌐	▬	⌐	▦	⌂	♪	avg.	🌐	▬	⌐	▦	⌂	♪	avg.		
German	0.18	24.6	27.6	29.6	9.7	19.7	21.1	22.0	24.9	31.5	27.9	10.5	19.3	21.2	22.5	25.1	30.7	27.7	10.4	19.6	21.5	22.5	0.0	0.8
Danish	0.18	25.5	30.8	33.0	11.9	19.8	21.4	23.7	25.6	31.4	34.6	8.4	20.0	21.4	23.6	25.6	30.6	33.8	8.6	20.1	20.6	23.2	0.4	0.1
Portuguese_BR	0.18	26.2	30.7	29.2	10.7	20.0	21.2	23.0	24.9	34.7	32.1	10.1	18.2	21.5	23.6	25.3	32.5	32.5	10.1	17.9	21.4	23.3	0.3	0.0
Portuguese_PT	0.18	28.2	32.9	31.7	10.5	20.1	22.9	24.4	24.4	34.7	28.0	10.1	19.9	21.9	23.2	25.1	34.5	28.9	10.0	19.7	22.3	23.4	0.2	0.1
Dutch	0.19	25.8	30.9	29.3	9.7	18.5	20.7	22.5	25.0	32.1	30.3	10.5	19.9	21.6	23.2	25.7	32.2	30.3	10.7	20.4	21.8	23.5	0.3	0.2
Ukrainian	0.21	26.7	29.1	27.5	9.0	19.4	20.4	22.0	24.8	31.4	29.9	10.4	16.1	21.6	22.5	24.6	30.9	30.5	10.8	16.2	23.3	22.7	0.2	0.6
Swedish	0.21	25.8	33.4	31.1	10.6	18.6	21.6	23.5	25.7	32.1	33.4	8.0	17.4	20.5	22.9	25.2	31.3	32.4	8.3	17.8	20.2	22.5	0.4	0.8
Slovenian	0.22	27.0	32.3	28.1	7.9	15.0	20.1	21.7	25.3	32.4	28.4	10.5	19.8	21.1	22.9	25.1	31.3	30.2	10.1	20.0	20.2	22.8	0.1	0.5
Italian	0.22	27.1	32.5	31.3	12.8	19.1	22.3	24.2	26.3	34.6	32.0	11.3	19.9	19.7	24.0	26.7	34.3	31.5	11.3	20.2	20.0	24.0	0.0	0.7
Romanian	0.23	26.5	33.0	30.2	10.3	16.6	21.3	23.0	24.0	33.7	29.8	10.8	20.7	19.4	23.1	24.3	30.5	30.4	10.8	20.0	19.2	22.5	0.6	0.8
Bulgarian	0.23	28.1	34.4	27.2	9.0	20.1	20.9	23.3	24.3	31.5	29.2	10.8	19.1	21.4	22.7	24.3	31.1	30.9	10.9	19.0	21.5	22.9	0.2	0.4
French	0.23	29.6	33.5	32.3	11.3	19.3	23.5	24.9	25.5	33.5	31.4	11.2	19.8	21.8	23.9	25.5	32.1	31.2	10.9	20.1	21.7	23.6	0.3	0.3
Slovak	0.23	23.1	32.7	28.2	9.2	18.6	18.2	21.7	24.4	32.6	31.6	10.2	19.2	19.8	23.0	24.1	33.6	31.7	10.3	17.8	20.1	22.9	0.1	0.4
Indonesian	0.24	26.0	34.6	33.2	9.6	19.7	20.7	24.0	25.2	32.9	32.6	9.7	16.9	20.9	23.0	26.1	32.9	32.4	9.8	16.5	20.7	23.1	0.1	0.2
Latvian	0.25	24.8	32.3	29.0	11.0	15.9	19.1	21.4	25.2	32.6	27.6	8.7	18.8	20.5	22.1	24.4	30.9	28.7	9.8	19.1	20.5	22.0	0.1	1.3
Spanish	0.27	27.6	32.2	29.9	9.7	19.2	22.5	23.5	24.5	32.4	29.1	9.2	19.5	23.9	23.1	24.6	31.9	28.6	9.5	20.2	23.3	23.0	0.1	0.3
Hungarian	0.27	22.4	28.9	26.0	8.5	19.2	18.4	20.6	21.2	31.0	28.5	8.6	18.5	21.2	21.5	22.2	30.2	29.1	8.5	19.3	21.3	21.8	0.3	1.5
Greek	0.27	28.3	33.3	31.8	9.1	20.3	22.7	24.2	24.1	30.7	32.9	11.2	18.6	19.8	22.9	24.7	31.9	33.6	10.9	19.2	20.8	23.5	0.6	0.2
Estonian	0.27	23.4	29.3	27.4	8.3	17.1	19.0	20.8	22.7	31.8	29.2	8.5	15.8	19.4	21.2	23.8	30.6	30.4	8.5	16.4	18.4	21.3	0.0	2.0
Lithuanian	0.27	26.2	31.5	26.3	9.9	18.9	16.2	20.8	24.5	31.3	26.4	10.8	18.8	21.4	22.2	25.3	30.0	27.6	10.3	18.6	21.2	22.2	0.0	1.1
Polish	0.27	24.6	34.3	28.7	10.4	19.5	19.9	22.9	24.4	31.6	27.9	9.7	16.6	20.4	21.8	24.5	30.9	28.6	9.6	16.6	20.8	21.8	0.0	1.5
Finnish	0.28	22.9	30.2	24.7	8.8	17.0	18.1	20.3	21.4	29.5	27.1	8.8	17.4	20.5	20.8	24.9	34.7	32.1	10.1	18.2	21.5	23.6	2.8	0.3
Czech	0.29	25.0	30.1	28.4	10.1	19.4	18.1	21.8	23.8	30.8	29.0	9.8	20.2	19.6	22.2	24.4	31.9	29.5	9.7	19.6	20.0	22.5	0.3	0.8
Chinese	0.30	22.2	33.4	25.0	9.0	20.1	18.7	21.4	23.1	28.4	27.1	9.5	18.9	22.0	21.5	23.8	28.7	27.4	9.9	18.7	21.3	21.6	0.1	1.7
Turkish	0.38	23.8	29.4	26.7	10.6	20.4	18.2	21.5	23.4	23.2	28.4	9.3	17.6	19.1	20.2	24.5	23.2	29.8	9.2	17.9	20.3	20.8	0.6	2.5
Japanese	0.41	22.6	29.2	20.1	8.9	19.5	12.9	18.9	21.1	27.4	21.7	8.0	16.1	15.2	18.3	20.5	27.9	23.4	8.1	16.1	16.2	18.7	0.4	4.6

In Table 7.3 we report the results of our experiments over MULTI-CROSSRE. The left-most columns are the results of the models trained and evaluated over the translated data (from English to language X). As a sanity check, we back-translated the data from each of the 26 new languages to English (from language X to English). We train and evaluate new models on this data in the middle columns. Finally, on the right-most columns we evaluate the same models—trained on back-translated data—over the original CrossRE test sets. We sort the languages by increasing distance to English, computed as the cosine distance between the syntax, phonology and inventory vectors of lang2vec (Littell et al., 2017).

For our analysis we consider the average of the six domains.[6] Our scores on the translated data reveal a relatively small drop in respect to the English baseline in Table 7.2. The difference range goes from an improvement of $+1.6$ Macro-F1 points on French, to a maximum drop of -4.4 on Japanese—which has the largest lang2vec distance with respect to English (0.41). The results of the models trained on the back-translated data present essentially the same trend between evaluating on the back-translations and on the original CrossRE English data—with a Pearson's correlation coefficient of 0.88—confirming the high quality of the proposed translation. The only exception if Finnish, with a difference of 2.8 points between the two evaluations. All the other languages report a smaller difference in a range between 0.0 and 0.6. The lang2vec distance is not informative of the quality of the individual translations (Pearson's correlation -0.59). However, other factors should be taken into account, e.g. the language model performances on each individual language.

7.4 Manual Analysis

We performed a manual analysis for further inspecting the quality and usability of MULTI-CROSSRE for studying multi-lingual RE. We manually

---

[6]Bassignana and Plank, 2022a discuss the lower scores of news (▤) attributing them to the data coming from a different data source and the fewer amount of relation instances with respect to the other domains.

checked 210 sentences from a diverse set of seven languages, including one North Germanic (Danish), one Uralic (Finnish), one West Slavic (Czech), two Germanic (German and Dutch), one Latin (Italian), and one Japonic (Japanese). For each of them, native speakers annotated the following: ① In how many sentences is the overall meaning preserved? ② How many entities are transferred to language X? ③ How many entities are correctly translated? ④ How many entities are marked with the correct entity boundaries?

We annotated 30 sentences for each language. Table 7.4 reports the statistics of our analysis. Overall, we find a surprisingly high quality of entity translations (96% are judged as correct by our human annotators). Out of the seven languages, Japanese is the one suffering the most by the translation process and, as we discussed above, this is reflected in the lowest scores in Table 7.3. Some entities are not transferred. These are mostly due to compounds typical for some languages. For example, the English snippet "the *Nobel* laureate" (where only *Nobel* is marked as entity), is translated to Danish as "nobelpristageren", and to Dutch as "Nobelprijswinnaar". In Italian, which in this regard behaves more similarly to English, all the entities are correctly transferred. In Appendix 7.6.2 we report the total per-language percentages of transferred entities and relations. Regarding the entity translations and the entity boundaries, the latter is a bigger challenge for the translation tool, often including surrounding function words—e.g. the writer *Pat Barker* in Danish is extended to the entity *Pat Barker er*. These could easily be post-processed, but since the Relation Classification model relies on the injected entity markers, it is not much influenced by this type of error (see baseline discussion in Section 7.3).

7.5   Conclusion

We introduce MULTI-CROSSRE, the most diverse RE dataset to date, including 26 languages in addition to the original English, and six text domains. The proposed span-based MT approach could be easily applied to similar cases. We report baseline results on the proposed resource and, as quality

Language	Sent. Transl.	# entities	Ent. Transl.	Ent. Bound.
English	30	160	-	-
Czech	28	158	152	143
Danish	27	158	143	136
Dutch	28	158	156	141
Finnish	30	150	141	137
German	27	151	148	139
Italian	29	160	157	152
Japanese	19	150	145	82

Table 7.4: **Statistics of the Manual Analysis.** At the top, total amount of original English sentences and annotated entities within them. Below, for each sample set, amount of correct instances in the four categories of sentence translation, number of entities, entity translations, and entity boundaries.

check, we back-translate MULTI-CROSSRE to English and run the baseline model again over it. Our manual analysis reveals that the higher challenge during the translation is transferring the correct entity boundaries. However, given the model architecture, this does not influence the scores.

Acknowledgments

Parameter	Value
Encoder	xlm-roberta-large
Classifier	1-layer FFNN
Loss	Cross Entropy
Optimizer	Adam optimizer
Learning rate	$2e^{-5}$
Batch size	32
Seeds	4012, 5096, 8878, 8857, 9908

Table 7.5: **Hyperparameters Setting.** Model details for reproducibility of the baseline.

7.6 Appendix

7.6.1 Reproducibility

We report in Table 7.5 the hyperparameter setting of our RC model (see Section 7.3). All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU.

7.6.2 Per-language Analysis

In table 7.6 we report the percentages of entities which are transferred during the translation process from the original English to language X, and the percentage of relations which do not involve missing entities (i.e. are transferred during the translation process).

Language	% Entities	% Relations
German	96.7	91.4
Danish	97.5	93.9
Portuguese_BR	99.8	99.5
Portuguese_PT	99.8	99.6
Dutch	98.5	95.8
Ukrainian	99.1	97.7
Swedish	97.6	94.1
Slovenian	99.1	98.0
Italian	99.8	99.5
Romanian	98.8	96.7
Bulgarian	99.5	98.9
French	99.6	99.4
Slovak	99.2	98.1
Indonesian	99.8	99.5
Latvian	99.4	98.6
Spanish	99.3	98.3
Hungarian	98.2	95.8
Greek	98.8	98.0
Estonian	97.9	94.6
Lithuanian	99.4	98.8
Polish	99.4	98.6
Finnish	96.0	90.7
Czech	99.0	98.0
Chinese	99.3	98.4
Turkish	99.4	98.5
Japanese	94.9	88.9

Table 7.6: **Transferred Entities and Relations.** Percentages of entities and of relations transferred during the translation process for each language.

Part III

Modeling

Chapter 8

Silver Syntax Pre-training for Cross-Domain Relation Extraction

Abstract

Relation Extraction (RE) remains a challenging task, especially when considering realistic out-of-domain evaluations. One of the main reasons for this is the limited training size of current RE datasets: obtaining high-quality (manually annotated) data is extremely expensive and cannot realistically be repeated for each new domain. An intermediate training step on data from related tasks has shown to be beneficial across many NLP tasks. However, this setup still requires supplementary annotated data, which is often not available. In this paper, we investigate intermediate pre-training specifically for RE. We exploit the affinity between syntactic structure and semantic RE, and identify the syntactic relations which are closely related to RE by being on the shortest dependency path between two entities. We then take advantage of the high accuracy of current syntactic parsers in order to automatically obtain large amounts of low-cost pre-training data. By pre-training our RE model on the relevant syntactic relations, we are able to outperform the baseline in five out of six cross-domain setups, *without* any additional annotated data.

8.1   Introduction

Relation Extraction (RE) is the task of extracting structured knowledge, often in the form of triplets, from unstructured text. Despite the increasing attention this task received in recent years, the performance obtained so far are very low (Popovic and Färber, 2022). This happens in particular when considering realistic scenarios which include out-of-domain setups, and deal with the whole task—in contrast to the simplified Relation Classification which assumes that the correct entity pairs are given (Han et al., 2018; Baldini Soares et al., 2019; Gao et al., 2019). One main challenge of RE and other related Information Extraction tasks is the "domain-specificity": Depending on the text domain, the type of information to extract changes. For example, while in the news domain we can find entities like *person* and *city,* and relations like *city of birth* (Zhang et al., 2017b), in scien-

142
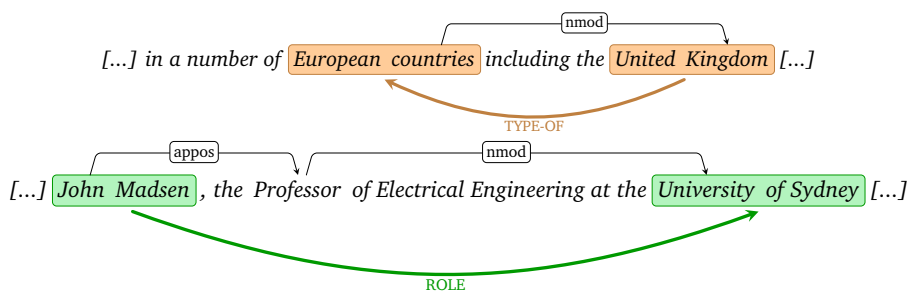
Figure 8.1: **Syntactic and Semantic Structures Affinity.** Shortest dependency path (above), and semantic relation (below) between two semantic entities.

tific texts, we can find information about *metrics*, *tasks* and *comparisons* between computational models (Luan et al., 2018). While high-quality, domain-specific data for fine-tuning the RE models would be ideal, as for many other NLP tasks, annotating data is expensive and time-consuming.[1] A recent approach that leads to improved performance on a variety of NLP tasks is intermediate task training. It consists of a step of training on one or more NLP tasks between the general language model pre-training and the specific end task fine-tuning (STILT, Supplementary Training on Intermediate Labeled-data Tasks; Phang et al., 2018). However, STILT assumes the availability of additional high quality training data, annotated for a related task.

In this paper, we explore intermediate pre-training specifically for cross-domain RE and look for alternatives which avoid the need of external manually annotated datasets to pre-train the model on. In particular, we analyze the affinity between syntactic structure and semantic relations, by considering the shortest dependency path between two entities (Bunescu and Mooney, 2005; Fundel et al., 2006; Björne et al., 2009; Liu et al., 2015). We replace the traditional intermediate pre-training step on additional annotated data, with a *syntax pre-training* step on silver data. We exploit

---

[1]For example, Bassignana and Plank, 2022a report a cost of 19K USD (≈ 1$ per annotated relation) and seven months of annotation work for an RE dataset including 5.3K sentences.
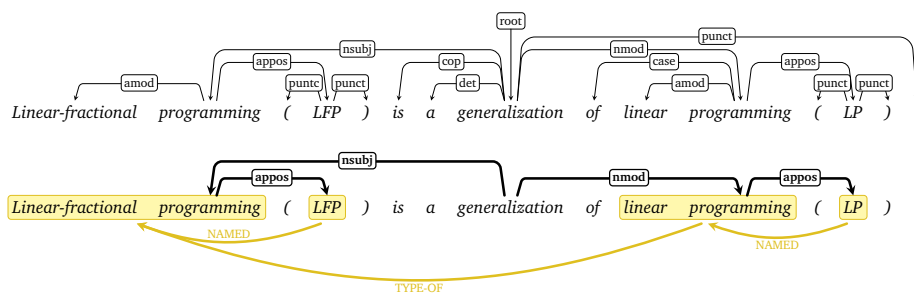
Figure 8.2: **Pre-training Example**. Given the dependency tree (above), we filter in for pre-training only the UD labels which are on the shortest dependency path between two semantic entities (below).

the high accuracy of current syntax parsers, for obtaining large amount of low-cost pre-training data. The use of syntax has a long tradition in RE (Zhang et al., 2006; Qian et al., 2008; Nguyen et al., 2009; Peng et al., 2015). Recently, work has started to infuse syntax during language model pre-training (Sachan et al., 2021) showing benefits for RE as well. We instead investigate dependency information as silver data in intermediate training, which is more efficient. To the best of our knowledge, the use of syntax in intermediate pre-training for RE is novel. We aim to answer the following research questions: ① Does syntax help RE via intermediate pre-training (fast and cheap approach)? and ② How does it compare with pre-training on additional labeled RE data (expensive)? We release our model and experiments.[2]

8.2 Syntax Pre-training for RE

Syntactic parsing is a structured prediction task aiming to extract the syntactic structure of text, most commonly in the form of a tree. RE is also a structured prediction task, but with the aim of extracting the semantics expressed in a text in the form of triplets—entity A, entity B,

---

[2]https://github.com/mainlp/syntax-pre-training-for-RE

and the semantic relation between them.[3] We exploit the affinity of these two structures by considering the shortest dependency path between two (semantic) entities (see Figure 8.1).

The idea we follow in this work is to pre-train an RE baseline model over the syntactic relations—Universal Dependency (UD) labels—which most frequently appear on the shortest dependency paths between two entities (black bold arrows in Figure 8.2). We assume these labels to be the most relevant with respect to the final target task of RE. In order to feed the individual UD relations into the RE baseline (model details in Section 8.3.1) we treat them similarly as the semantic connections. In respect to Figure 8.2, we can formalize the semantic relations as the following triplets:

- `NAMED(LFP,Linear-fractional programming)`
- `TYPE-OF(linear programming,Linear-fractional programming)`
- `NAMED(LP,linear programming)`.

Accordingly, we define the syntax pre-training instances as:

- `appos(programming,LFP)`
- `nsubj(generalization,programming)`
- `nmod(generalization,programming)`
- `appos(programming,LP)`.

In the next section we describe the detailed training process.

8.3 Experiments

8.3.1 Setup

Data   In order to evaluate the robustness of our method over out-of-domain distributions, we experiment with CrossRE (Bassignana and Plank,

---

[3]In this project, we follow previous work, and assume gold entities, leaving end-to-end RE for future work.
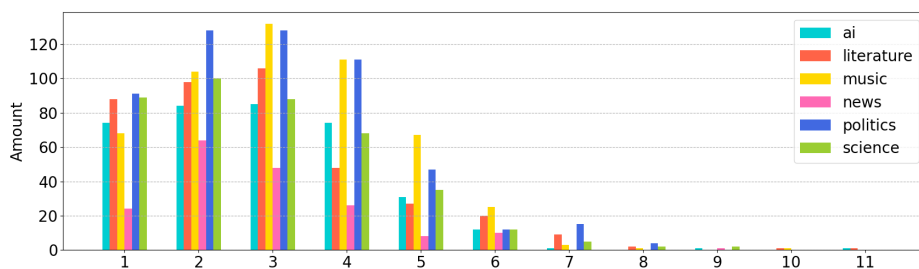
Figure 8.3: **Shortest Dependency Path Length.** Statistics of the shortest dependency path length between two entities over the train sets of CrossRE (Bassignana and Plank, 2022a).

2022a),[4] a recently published multi-domain dataset. CrossRE includes 17 relation types spanning over six diverse text domains: news, politics, natural science, music, literature and artificial intelligence (AI). The dataset was annotated on top of a Named Entity Recognition dataset—CrossNER (Liu et al., 2021b)—which comes with an unlabeled domain-related corpora.[5] We used the latter for the *syntax pre-training* phase.

UD Label Selection   In order to select the UD labels which most frequently appear on the shortest dependency path between two semantic entities, we parsed the training portions of CrossRE. Our analysis combines RE annotations and syntactically parsed data. We observe that the syntactic distance between two entities is often higher than one (see Figure 8.3), meaning that the shortest dependency path between two entities includes multiple dependencies—in the examples in Figure 8.1, the one above has distance one, the one below has distance two. However, the shortest dependency paths contain an high frequency of just a few UD labels (see Figure 8.4) which we use for *syntax pre-training*: nsubj, obj, obl, nmod, appos. See Appendix 8.6.1 for additional data analysis.

---

[4]Released with a GNU General Public License v3.0.
[5]Released with an MIT License.
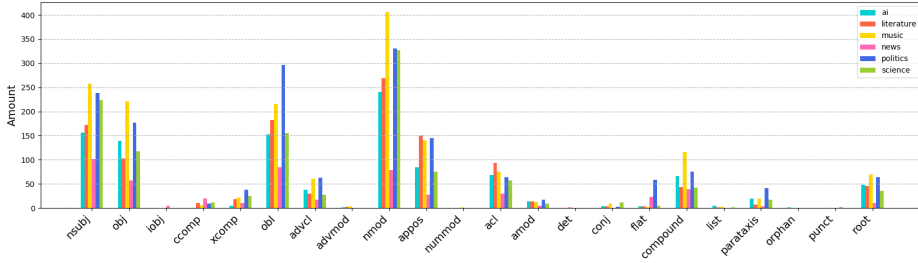
Figure 8.4: **UD Label Distribution Over the Shortest Dependency Paths.** Statistics of the UD labels which are on the shortest dependency path between two entities over the six train sets of CrossRE (Bassignana and Plank, 2022a).

Model Our RE model follows the current state-of-the-art architecture by Baldini Soares et al., 2019 which augments the sentence with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} before feeding it into a pre-trained encoder (BERT; Devlin et al., 2019). The classification is then made by a 1-layer feed-forward neural network over the concatenation of the start markers $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$. We run our experiments over five random seeds and report the average performance. See Appendix 8.6.2 for reproducibility and hyperparameters settings of our model.

Training The training of our RE model is divided into two phases. In the first one—which we are going to call *syntax pre-training*—we use the unlabeled corpora from CrossNER for pre-training the baseline model over the *RE-relevant* UD labels. To do so, ① we sample an equal amount of sentences from each domain[6] (details in Section 8.4), and ② use the MaChAmp toolkit (van der Goot et al., 2021b) for inferring the syntactic tree of each of them. We apply an additional sub-step for disentangling the conj dependency, as illustrated in Appendix 8.6.3. Then, ③ we filer in only the nsubj, obj, obl, nmod, and appos UD labels and ④ feed those connections to the RE model (as explained in the previous section). Within

---

[6]Regarding the news domain, which does not have a corresponding unlabeled corpus available, we sampled from the train set of CrossNER which is not included in CrossRE.

	TRAIN	news	politics	science	music	literature	AI	avg.
					TEST			
BASELINE	news	10.98	1.32	1.24	1.01	1.49	1.42	**2.91**
	politics	16.07	11.30	6.74	7.24	7.29	5.54	9.03
	science	6.54	5.95	8.57	7.13	6.65	7.29	7.02
	music	3.99	9.91	9.22	19.01	10.43	8.53	10.18
	literature	11.30	9.60	9.79	12.49	17.17	9.79	11.69
	AI	6.58	7.42	11.03	7.11	6.15	15.57	8.98
SYNTAX	news	6.67	1.15	0.72	0.61	1.13	0.75	1.84
	politics	13.72	12.09	7.47	7.15	7.78	6.24	**9.08**
	science	8.46	7.08	8.69	8.19	7.52	8.91	**8.14**
	music	3.35	10.65	9.35	18.63	11.62	10.34	**10.66**
	literature	11.85	9.84	10.35	13.58	18.64	9.94	**12.37**
	AI	8.87	8.59	11.87	8.29	7.68	15.93	**10.21**
SCIERC	news	11.88	2.30	2.09	1.13	1.82	2.16	3.56
	politics	14.25	13.55	6.52	7.12	7.42	7.07	9.32
	science	8.27	10.31	13.59	9.09	7.78	11.11	10.03
	music	5.41	11.84	10.85	21.39	12.26	11.22	12.16
	literature	12.36	8.05	8.87	13.13	16.44	9.40	11.37
	AI	11.00	10.12	14.03	8.93	8.50	18.89	11.91

Table 8.1: **Performance Scores.** Macro-F1 scores of the baseline model, compared with the proposed *syntax pre-training* approach, and—as comparison—with the traditional pre-training over the manually annotated SciERC dataset (Luan et al., 2018).

the RE model architecture described above, each triplet corresponds to one instance. In this phase, in order to assure more variety, we randomly select a maximum of five triplets from each pre-train sentence.

In the second training phase—the *fine-tuning* one—we replace the classification head (i.e. the feed-forward layer) with a new one, and individually train six copies of the model over the six train sets of CrossRE. Note that the encoder is fine-tuned in both training phases. Finally, we test each model on in- and out-of-domain setups.

8.3.2 Results

Table 8.1 reports the results of our cross-domain experiments in terms of Macro-F1. We compare our proposed approach which adopts *syntax*

pre-training with the zero-shot baseline model.[7] Five out of six models outperform the average of the baseline evaluation, including in- and out-of-domain assessments. The average improvement—obtained without any additional annotated RE data—is 0.71, which considering the low score range given by the challenging dataset (with limited train sets, see dataset size in Appendix 8.6.4), and the cross-domain setup, is considerable. The model fine-tuned on the news domain is the only one not outperforming the baseline. However, the performance scores on this domain are already extremely low for the baseline, because news comes from a different data source with respect to the other domains, has a considerable smaller train set, and present a sparse relation types distribution, making it a bad candidate for transferring to other domains (Bassignana and Plank, 2022a).

As comparison, we report the scores obtained with the traditional intermediate pre-training which includes additional annotated data. We pre-train the language encoder on SciERC (Luan et al., 2018), a manually annotated dataset for RE. SciERC contains seven relation types, of which three overlap with the CrossRE relation set. In this setup, the improvement over the baseline includes the news, but not the literature domain. Nevertheless, while the gain is on average slightly higher with respect to the proposed *syntax pre-training* approach, it comes at a much higher annotation cost.

8.4   Pre-training Data Quantity Analysis

We inspect the optimal quantity of syntactic data to pre-train our RE model on by fine-tuning this hyperparameter over the dev sets of CrossRE. The plot in Figure 8.5 reports the average performance of the six models when pre-trained on increasing amounts of syntactic dependencies.[8] Starting from 8.4K instances onward, the performance stabilizes above the baseline.

---

[7]While utilizing the model implementation by Bassignana and Plank, 2022a, our score range is lower because we include the *no-relation* case, while they assume gold entity pairs.
[8]Pre-training performance in Appendix 8.6.5.
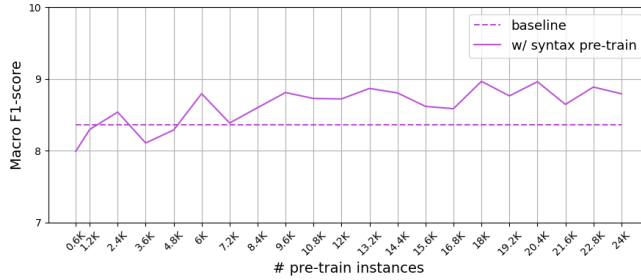
Figure 8.5: **Pre-train Data Quantity Analysis.** Average (dev) performance of the six models when pre-trained over increasing amounts of syntactic instances.
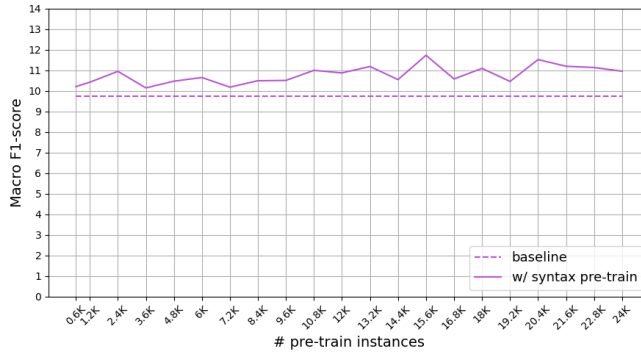
Figure 8.6: **Per-Domain Pre-train Data Quantity Analysis.** Individual (dev) performance of the model fine-tuned on AI when pre-trained over increasing amounts of syntactic instances.

We select the peak (20.4K, albeit results are similar between 18-20.4K) for reporting our test set results in Table 8.1. While we are interested in the robustness of our method across multiple domains, and therefore consider the average (Figure 8.5), domain-optima could be achieved by examining individual domain performance. As example, we report in Figure 8.6 the plot relative to the model fine-tuned on AI, which is the one obtaining the highest gain. The model fine-tuned on AI generally gains a lot from the *syntax pre-training* step, with its peak on 15.6K pre-training instances.

8.5   Conclusion

We introduce *syntax pre-training* for RE as an alternative to the traditional intermediate training which uses additional manually annotated data. We pre-train our RE model over silver UD labels which most frequently connect the semantic entities via the shortest dependency path. We test the proposed method over CrossRE and outperform the baseline in five out of six cross-domain setups. Pre-training over a manually annotated dataset, in comparison, only slightly increases our scores in five out of six evaluations, but at a much higher cost.

Limitations

While we already manage to outperform the baseline, the pre-training data quantity is relatively small ($\sim$20K instances). Given the computational cost of training 30 models—six train sets, over five random seeds each—and testing them within in- and cross- domain setups, we break the inspection of the optimal pre-training data amount at 24K instances. However we do not exclude that more pre-training instances would be even more beneficial for improving even more over the baseline.

Related to computation cost constrains, we test our *syntax pre-training* approach over one set of UD labels only (`nsubj`, `obj`, `obl`, `nmod`, `appos`). Different sets could be investigated, e.g. including `acl` and `compound`, which present a lower, but still considerable amount of instances (see Figure 8.4).

Finally, while approaching RE by assuming that the gold entities are given is a common area of research, we leave for future work the inspection of the proposed method over end-to-end RE.

Acknowledgments

	rel-to	artifact	cause-eff	compare	gen-aff	named	opposite	origin	part-of	physical	role	social	temporal	topic	type-of	usage	win-def
nsubj	89	106	2	12	120	54	61	53	75	115	248	33	54	10	18	30	68
obj	78	51	1	6	76	36	48	41	83	55	129	9	48	17	14	37	86
iobj	0	0	0	0	1	0	0	0	0	1	3	0	0	0	0	0	0
ccomp	5	7	0	4	7	10	8	2	2	2	9	0	0	0	0	0	0
xcomp	6	9	0	3	15	5	5	9	5	11	17	1	16	3	1	2	10
obl	88	62	5	14	92	53	25	44	77	202	224	19	121	17	26	6	11
advcl	10	9	4	8	47	21	19	10	18	14	41	3	15	2	6	7	2
advmod	0	3	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
nmod	100	140	2	12	181	57	47	58	148	276	386	29	72	43	35	19	48
appos	26	89	0	2	85	108	11	23	41	72	112	9	12	6	6	1	20
nummod	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
acl	40	24	0	0	39	30	10	25	48	33	74	0	11	24	2	13	15
amod	5	1	0	2	31	5	3	3	5	2	3	0	3	2	0	3	4
det	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
conj	1	4	0	0	3	1	0	1	2	3	11	0	1	1	0	0	0
flat	2	3	0	0	1	12	8	0	2	11	37	8	7	1	0	0	3
compound	29	24	0	5	70	27	5	7	54	53	57	2	9	2	5	10	22
list	0	1	0	0	2	2	0	0	0	0	2	0	2	0	0	0	0
parataxis	5	7	0	0	30	14	0	0	14	5	17	1	8	1	5	0	1
orphan	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
punct	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 8.2: **UD Label Distribution Over the Shortest Dependency Paths per Relation Type.** Statistics of the UD labels which are on the shortest dependency path between two entities divided by the 17 relation types of CrossRE (Bassignana and Plank, 2022a).

8.6 Appendix

8.6.1 UD Analysis for RE

We inspect the same statistics as Figure 8.4 and Figure 8.3—UD labels on the shortest dependency paths, and shortest dependency path lengths respectively—but instead of at domain level, at semantic relation type level. Table 8.2 and Table 8.3 report this analysis, revealing similar trends over the 17 types.

	related-to	artifact	cause-eff	compare	gen-aff	named	opposite	origin	part-of	physical	role	social	temporal	topic	type-of	usage	win-def
1	3	43	0	0	73	92	1	19	26	72	64	0	5	7	17	0	12
2	29	43	0	3	59	23	25	28	51	85	127	10	41	6	11	6	31
3	36	71	2	1	33	14	21	29	57	75	136	17	30	14	3	14	34
4	42	24	2	4	52	20	13	12	37	41	104	7	28	6	17	12	17
5	17	12	0	5	36	12	14	8	18	28	33	5	10	8	2	4	3
6	9	7	0	4	18	6	4	5	8	12	10	0	2	1	2	1	2
7	4	0	0	0	4	3	0	0	2	6	4	0	5	0	0	0	5
8	0	0	0	0	0	3	1	0	0	1	1	0	3	0	0	0	0
9	0	1	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0
10	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0

(Rows labelled "Shortest Dependency Path Length")

Table 8.3: **Shortest Dependency Path Length per Relation Type.** Statistics of the shortest dependency path length between two semantic entities divided by the 17 relation types of CrossRE (Bassignana and Plank, 2022a).

Parameter	Value
Encoder	`bert-base-cased`
Classifier	1-layer FFNN
Loss	Cross Entropy
Optimizer	Adam optimizer
Batch size	12, 24
Learning rate	$1e^{-5}$ (pre-train)
Learning rate	$2e^{-5}$ (fine-tuning)
Seeds	4012, 5096, 8878, 8857, 9908

Table 8.4: **Hyperparameters Setting.** Model details for reproducibility of the baseline.

8.6.2   Reproducibility

We report in Table 8.4 the hyperparameter setting of our RE model (see Section 8.3.1). All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU. Within this computation infrastructure the baseline converges in ∼ 7 minutes. The the *syntax pre-training* step takes ∼ 10 minutes, to which we have to add ∼ 7 minutes in order to obtain the complete training time.

We train MaChAmp v0.4 on the English Web Treebank v2.10 with XLM-R large (Conneau et al., 2020) as language model with all default hyperparameters of MaChAmp.
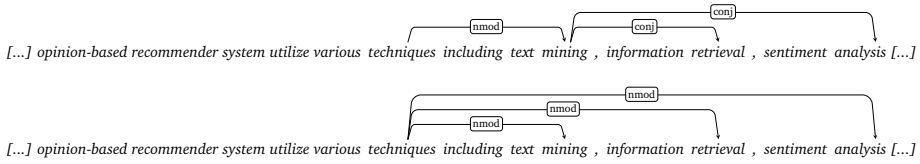
Figure 8.7: **Example of** `conj` **Alteration**. Original UD dependencies (above), and disentangled conjunction dependencies reflecting the semantic relation annotations (below).

	SENTENCES				RELATIONS			
	train	dev	test	**tot.**	train	dev	test	**tot.**
news	164	350	400	914	175	300	396	871
politics	101	350	400	851	502	1,616	1,831	3,949
science	103	351	400	854	355	1,340	1,393	3,088
music	100	350	399	849	496	1,861	2,333	4,690
literature	100	400	416	916	397	1,539	1,591	3,527
AI	100	350	431	881	350	1,006	1,127	2,483
tot.	668	2,151	2,446	**5,265**	2,275	7,662	8,671	**18,608**

Table 8.5: **CrossRE Statistics.** Number of sentences and number of relations for each domain of CrossRE (Bassignana and Plank, 2022a).

8.6.3 Handling of `Conj`

In UD, the first element in a conjuncted list governs all other elements of the list via a `conj` dependency and represents the list syntactically w.r.t. the remainder of the sentence. CrossRE (Bassignana and Plank, 2022a) relations, on the other hand, directly link the two entities involved in the semantic structure. To account for this difference, we propagate the conjunction dependencies in order to reflect the semantic relations, as shown in Figure 8.7.

8.6.4 CrossRE Size

We report in Table 8.5 the dataset statistics of CrossRE (Bassignana and Plank, 2022a) including the number of sentences and of relations.

Figure 8.8: **Pre-train Performance.** Pre-train performance of the RE model over increasing amounts of dependency instances

8.6.5 Syntax Pre-training Performance

Figure 8.8 reports the performance of the RE model during the *syntax pre-training* phase, over increasing amounts of pre-training dependency instances. The scores are computed on a set including 600 sentences (100 per domain) not overlapping with the train set used in the syntax pre-training phase.

Chapter 9

How to Encode Domain Information in Relation Classification

The work presented in this chapter is based on the paper: Elisa Bassignana, Viggo Unmack Gascou, Frida Nøhr Laustsen, Gustav Kristensen, Marie Haahr Petersen, Rob van der Goot, and Barbara Plank. How to encode domain information in relation classification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8301–8306, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.728

Abstract

Current language models require a lot of training data to obtain high performance. For Relation Classification (RC), many datasets are domain-specific, so combining datasets to obtain better performance is non-trivial. We explore a multi-domain training setup for RC, and attempt to improve performance by encoding domain information. Our proposed models improve > 2 Macro-F1 against the baseline setup, and our analysis reveals that not all the labels benefit the same: The classes which occupy a similar space across domains (i.e., their interpretation is close across them, for example *physical*) benefit the least, while domain-dependent relations (e.g., *part-of*) improve the most when encoding domain information.

Keywords. Relation Classification, Domain, Multi-domain training, Robustness

9.1 Introduction

Relation Classification (RC) is the task of identifying the semantic relation between two given entities. The task is beneficial for many different downstream tasks which involve Natural Language Understanding. For example, question answering, knowledge base population, or summarization. In addition to the wide variety of downstream applications, as most information extraction tasks, RC is topic-specific, meaning that depending on the topic the information to extract can vary a lot. For example, in the music domain we may want to extract that a song is included in a musical album, while in the politics domain we may have a politician winning a political election. While current deep learning models require a lot of training data, collecting and annotating text from every domain is time-consuming and expensive.

In this project, we explore the critical setup of multi-domain training with the aim of identifying the best setup for maximizing the training data (by including data coming from different domains), without losing domain-

specific information. To do so, we compare multiple ways of enriching the input instances with domain information (see Section 9.2).

Encoding information about where a certain utterance originates from has been previously explored in other Natural Language Processing fields. In the multi-lingual space, Conneau and Lample (2019) exploited language embeddings for multi-lingual model training. Ammar et al. (2016) first proposed to use language embeddings for training a multi-lingual syntactic parser for seven European languages, and showed improved performance. Later work also successfully trained parsers with the so called treebank embeddings for datasets within the same language (Stymne et al., 2018) or language family (Smith et al., 2018). Other work have used special language ids to mark the language of each instance in the context of machine translation (Liu et al., 2020). To the best of our knowledge, these approaches have been exploited mostly in multi-lingual setups and syntactic tasks. In this work, we explore a gap and test their effectiveness for encoding domain information in a semantic setup: Relation Classification. We compare "dataset embeddings" and "domain markers" from previous work with a new approach exploiting domain-specific entity types. Our contributions are:

- CrossRE 2.0, an extension of the CrossRE dataset (Bassignana and Plank, 2022a) with 3.3k new annotations in the news domain in order to balance data across domains;

- We propose the first multi-domain training baseline on CrossRE;

- We test previous work for encoding dataset information in RC, and compare it with a new RC-specific technique; We present an in-depth analysis of the results obtained.

159

9.2   Domain Encoding for Relation Classification

9.2.1   Dataset Embeddings

The dataset embedding model tries to encode information about the domain with ad-hoc embeddings on the encoder side. Dataset embeddings are vector representations learned at training time that aim at capturing distinctive properties of multiple data sources into a continuous vector, without losing their heterogeneous characteristics. Originally they were often concatenated to the word embedding and then used in e.g., a Bi-LSTM (Stymne, 2020; Wagner et al., 2020; van der Goot et al., 2021a). However, since large language models have become the standard, this has become trickier, as they have a pre-determined input size. To enable usage of dataset embeddings, van der Goot and de Lhoneux (2021) propose to sum them to the input representation. In our setup, we treat each domain as a separate data source.

9.2.2   Special Domain Markers

An intuitive and simple alternative way of encoding the domain is by using special tokens appended to the input text itself. This has been previously done in machine translation in order to mark the different languages (Liu et al., 2020). We concatenate a special token at the beginning of each instance containing the corresponding domain (e.g., [MUSIC] or [NEWS]). These domain markers are treated by the tokenizer as special tokens, i.e., they are not tokenized into subwords, so the model learns a representation for each of them during training.

9.2.3   Entity Type Information

The domain-specific entity types carry out information which can be relevant for identifying the correct relation label. Following Zhong and Chen (2021) we add entity type information in the representation of the input (see model description in Section 9.3.2). We test two different approaches

to do this. First, we use the 39 fine-grained types proposed by Liu et al. (2021b) including e.g., *musician* or *political party*, which are domain-specific. In the second setup we map these fine-grained types into five coarse-grained types. For example, *musician* and *political party* are mapped to *person* and *organization* respectively. While this last approach shades domain information, it guarantees a more condensed entity type distribution, and it can be combined with the other two setups.

9.3 Experimental Setup

9.3.1 Data

CrossRE (Bassignana and Plank, 2022a),[1] is a manually-annotated dataset for multi-domain RC including 17 relation types spanning over six diverse text domains: news (▤), politics (🏛), natural science (🍃), music (🎵), literature (📖), and artificial intelligence (🤖). The dataset was annotated on top of CrossNER (Liu et al., 2021b), a Named Entity Recognition (NER) dataset. Table 9.1 reports the statistics of CrossRE. While the train, dev, and test splits include similar amounts of sentences across the six domains, the number of annotated relations varies over a wider range. The reason for this includes different average sentence lengths, and different relation densities across the domains. In the original dataset, the news domain is particularly small. This domain comes from a different source with respect to the other five—the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) and Wikipedia (Liu et al., 2021b) respectively.

9.3.1.1 CrossRE 2.0

With the aim of mitigating the effect of dataset size on the model performance, which influences the comparison of results across domains, we expand the news domain of CrossRE. We follow the annotation guidelines

---

[1]Released with a GNU General Public License v3.0.

	SENTENCES				RELATIONS			
	train	dev	test	**tot.**	train	dev	test	**tot.**
🖥	217	1,320	3,053	4,590	156	1,043	2,115	3,314
🖥	164	350	400	914	175	300	396	871
🏛	101	350	400	851	502	1,616	1,831	3,949
🌿	103	351	400	854	355	1,340	1,393	3,088
🎵	100	350	399	849	496	1,861	2,333	4,690
📖	100	400	416	916	397	1,539	1,591	3,527
🤖	100	350	431	881	350	1,006	1,127	2,483
tot.	885	3,471	5,499	**9,855**	2,431	8,705	10,786	**21,922**

Table 9.1: **CrossRE 2.0 Statistics.** Number of sentences and number of relations of the news extension (first row), and statistics of the original domains of CrossRE (below).

by Bassignana and Plank (2022a)[2] and manually annotate more than 4.5k sentences from the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003)—the original data source of this domain. The data is annotated by a hired linguist compensated fairly according to national salary scales, who got extensive training for the task. We refer to Bassignana and Plank (2022a) for the discussion on the annotation agreement because for consistency we employed the same annotator who annotated the original version of CrossRE. Table 9.1 reports the statistics of our extension, with over 3k annotated relations and an overall total in news (including the original dataset) of 4.1k, which is in line with the other domains. The dataset extension is public available in the CrossRE repository.[3] We train the model in a multi-domain setup, i.e., mixing the six training sets of CrossRE.

9.3.2 Model Architecture

We use the baseline model of the original CrossRE paper.[4] Following the model architecture first proposed by Baldini Soares et al. (2019), the implementation by Bassignana and Plank (2022a) augments the sentence

---

[2]https://github.com/mainlp/CrossRE/blob/main/crossre_annotation/CrossRE-annotation-guidelines.pdf

[3]https://github.com/mainlp/CrossRE/

[4]https://github.com/mainlp/CrossRE

		📰	🏛	🍃	🎵	📖	🎰	avg.
	BASELINE	25.45	31.35	39.46	39.69	38.84	38.09	35.48
dev	DATASET EMB.	15.38	22.22	24.77	32.64	30.95	29.80	25.96
	DOMAIN MARK.	26.36	32.77	40.31	42.65	40.59	38.71	**36.90**
	FINE-GRAIN.	23.67	32.67	35.35	38.76	38.23	35.94	34.10
	COARSE-GRAIN.	24.46	31.56	38.59	39.33	38.09	37.90	34.99
	DOM. + COARSE	24.52	32.02	39.63	42.19	40.01	37.17	35.92
test	BASELINE	24.73	34.12	39.67	39.96	44.64	35.71	36.47
	DOMAIN MARK.	26.72	37.62	43.57	41.48	44.88	37.69	**38.66**

Table 9.2: **Performance Scores.** Macro-F1 scores of the explored setups. DOM. + COARSE indicates the combination of special domain markers with the coarse-grained entity types.

with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} surrounding the two entities. When exploiting the entity types, the information is injected in the entity markers (e.g., [E1:person]) The augmented sentence is then passed through a pre-trained encoder, and the classification is made by a linear layer over the concatenation of the start markers $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$. We run our experiments over five random seeds and report the average. All hypermarameters follow Bassignana and Plank (2022a). Our code is available on GitHub.[5]

9.4 Results

Table 9.2 reports the Macro-F1 results of our experiments. The dataset embeddings setup fails to beat the baseline. The main reason for this is the limited amount of training data in our setup, which challenges the model in learning them. [6] The dataset embeddings are summed to the word, segment, and position embeddings, which are then updated all at once in the forward pass. Additionally, in settings where they are successful, these embeddings are usually used to disambiguate datasets coming from difference data sources or languages. Here instead we are at a more fine-

---

[5]Project repository https://github.com/viggo-gascou/multi-domain-rc
[6]We manually inspected the dataset embeddings before and after training.

grained level, trying to model different topics, with data extracted from the same source (except for news).

Concatenating a special domain marker at the beginning of the sentence results in the best performance (36.90 Macro-F1), with the highest improvement in the music domain (+2.96) and sometimes small yet consistent improvements across all domains. The fine-grained entity types lead to decreased performance, because their distribution is very sparse across the six domains. For example, out of the 39, the news domain from CoNLL 2003 only includes *person*, *location*, *organization* and *miscellaneous*, resulting in a performance decrease of -1,78. Using the coarse-grained entity types—shared across all the domains—results in a slightly better average Macro-F1 (34.99) than employing the fine-grained ones (34.10), but it does not improve over the baseline either. As this setup lacks domain information, we try combining the coarse-grained entity representation with the special domain markers. Within this setup results are mixed across the domains: While most of them (except AI) improve over the coarse-grained entity type (without domain information), only politics, science, music and literature overcome the baseline. The overall average across the domains results in a minor improvement of +0.44.

We evaluate the best setup (the special domain markers) on the test set in order to confirm our findings. Following the trend on the development set, the improvement over the baseline is +2.19 Macro-F1. The lower performance range of news over the other domains (both in dev and test) indicates that the different data source has a high impact even with the more uniform data distribution across domains proposed with our dataset extension.

9.5   Analysis

Domain Representation   To inspect how much domain information the out-of-the-box embeddings contain, in Figure 9.1 we plot the PCA representations of the untrained embeddings (with `bert-base-cased`, the encoder used by the RC model) of the instances in the development set.
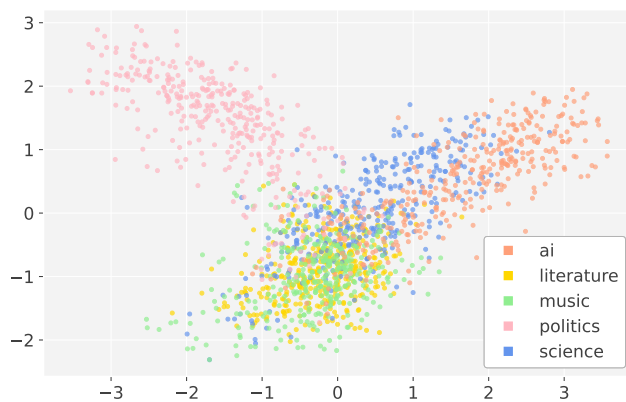
Figure 9.1: **Domain Representation.** PCA plot of the untrained embeddings of the instances in the development set, colored by domain.
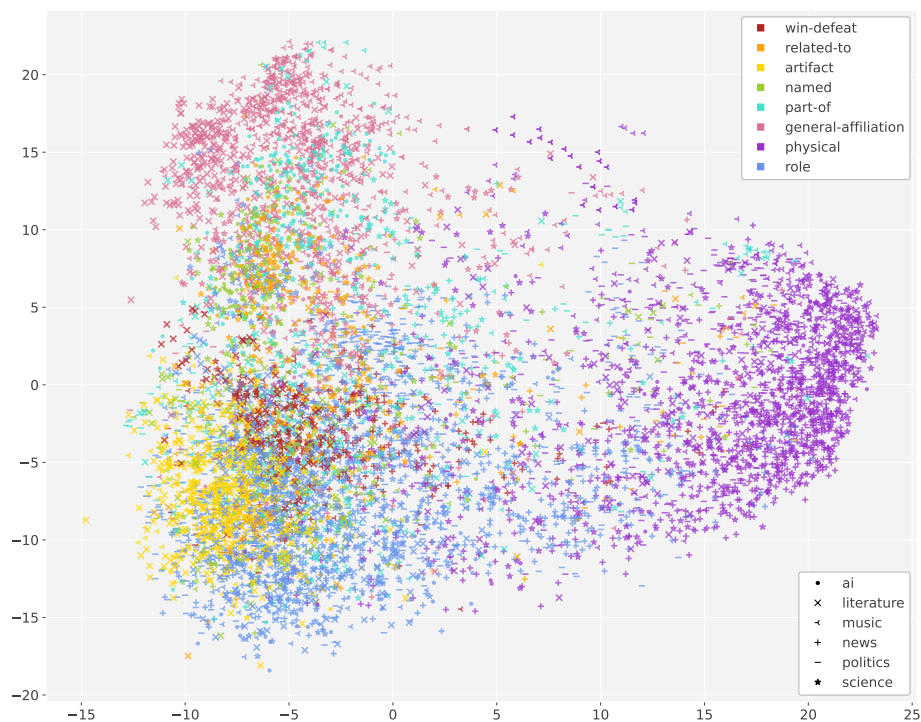
Figure 9.2: **Relation Representation.** PCA plot of the trained embeddings of the most frequent relation labels in the development set, colored by relation labels and shaped by domain.

We do not include the news domain in this plot because given its different data source (consisting of shorter sentences, typically news headlines), the news instances are very distant from the other domains, resulting in an high overlap of the latter. The current setup allows us to shed light on the remaining five domains, besides news which we already know is very distinctive. The domains are already relatively distinguishable with the untrained encoder. The two technical domains, science and AI, marginally overlap; politics is completely detached; and only music and literature overlap significantly. Our intuition is that encoding additional domain information (see Section 9.2) may not be particularly relevant.

Relation Representation   We dive deeper into analyzing the relations and explore whether in the baseline setup (i.e., without additional domain information) the representation of instances coming from different domains, but belonging to the same class, are close to each other. In Figure 9.2 we plot the PCA representation of the trained baseline model of the instances with the most frequent relation labels in the development set, separated by class and domain. The main finding from this plot is that most of the classes are quite clustered, independently from the domain they belong to. For example, the *physical* relation on the right side has instances from all the domains. Similarly, the *artifact* and the *role* labels towards the bottom-left corner of the plot. Interestingly, the *general-affiliation* relation presents clustered representations of the instances in the literature and music domains, but it still dominates the upper left side of the plot. Less surprisingly, the *related-to* label, listed as None-Of-The-Above (NOTA) in the guidelines, has a more sparse distribution across the plot.

The labels which present a less defined cluster (i.e., the ones whose meaning shifts the most across domains) are the ones which benefit the most from the special domain markers. For example, *related-to* improves from a baseline value of 20.99 F1 up to 24.21 in the special domain markers setup; *named* goes from 68.25 to 71.30 F1; and *part-of* improves from 32.79 to 35.54 F1. In contrary, the relation labels which present a better defined cluster already within the baseline (see Figure 9.2) do not benefit

much from the additional encoding of domain information. For example, the per-label F1 scores of the *physical*, *general-affiliation*, and *role* relations in the baseline and special domain markers setups are respectively 77.16 and 77.51, 54.09 and 54.46, 65.60 and 65.11.

9.6   Conclusion

We explore how to encode domain information in a multi-domain training setup for the domain-specific task of RC. We propose CrossRE 2.0, a dataset extension of CrossRE (Bassignana and Plank, 2022a) for balancing the amount of data across the six domains included in it. We manage to improve the multi-domain training baseline by > 2 Macro-F1 with a simple, but effective technique which encodes domain information in special domain markers concatenated at the beginning of each input. Our analysis reveals that not all of the relation labels benefit the same from the domain encoding: The most generic, with a diverse interpretation across domains (e.g., *part-of*) are the ones which gain the most in terms of per-label F1.

9.7   Ethics Statement

We do not foresee any potential risk related to this work. The data we use is published freely by Liu et al. (2021b) and Bassignana and Plank (2022a).

For the dataset extension, we hired an expert with a linguistics degree employed following national salary rates. The cost of the annotation process amounts to $\approx 1\$$ per annotated relation.

Acknowledgments

Part IV

Model Analysis

Chapter 10

What's wrong with your model? A Quantitative Analysis of Relation Classification

Abstract

With the aim of improving the state-of-the-art (SOTA) on a target task, a standard strategy in Natural Language Processing (NLP) research is to design a new model, or modify the existing SOTA, and then benchmark its performance on the target task. We argue in favor of enriching this chain of actions by a preliminary error-guided analysis: *First*, explore weaknesses by analyzing the hard cases where the existing model fails, and *then* target the improvement based on those. Interpretable evaluation has received little attention for structured prediction tasks. Therefore we propose the first in-depth analysis suite for Relation Classification (RC), and show its effectiveness through a case study. We propose a set of potentially influential attributes to focus on (e.g., entity distance, sentence length). Then, we bucket our datasets based on these attributes, and weight the importance of them through correlations. This allows us to identify highly challenging scenarios for the RC model. By exploiting the findings of our analysis, with a carefully targeted adjustment to our architecture, we effectively improve the performance over the baseline by >3 Micro-F1.

10.1 Introduction

A major trend in NLP research aims at designing more sophisticated setups and model architectures in order to improve the state-of-the-art (SOTA) on a target task. The improvements are usually based on intuitions that target limitations of the previous SOTA on the task. The most common procedure follows the steps of *(1)* intuition, *(2)* modeling, *(3)* experiments, *(4)* results, and *(5)* analysis of the results. The latter is occasionally enriched with ablation or case studies with the main aim of proving the validity of the initial intuition and the effectiveness of the proposed methodology. We claim that conducting a preliminary in-depth analysis can help find good intuitions, and therefore guide better modeling and reducing the probability of negative experiments, usually not reported in the paper. Following previous error-guided analysis (Ribeiro et al., 2020; Fu et al.,

2020a; Das et al., 2022), we argue in favor of changing the standard chain of actions listed above: *First* perform an exhaustive quantitative analysis of the previous SOTA to identify failure cases and challenging scenarios, and *then* effectively target the baseline improvement in order to tackle those.

We introduce an in-depth performance analysis suite in the context of Relation Classification (RC). Within the field of Information Extraction (IE), which broadly aims at extracting structured knowledge from unstructured text, the goal of RC aims at classifying the semantic relation between two named entities. We pick this task because, despite its popularity, the task is far from being solved or reaching high performance, especially when considering realistic challenging setups—e.g. cross-domain (Bassignana and Plank, 2022a), or document-level (Popovic and Färber, 2022). We inspect the research approach of some of the most cited papers in the field from recent years, on top of which current SOTA are based: Baldini Soares et al. (2019) introducing the widely adopted entity markers, Zhong and Chen (2021) introducing the typed entity markers and proposing a pipeline approach for end-to-end Relation Extraction (RE), and Ye et al. (2022) at the time of writing holding the SOTA on three of the most established datasets in the field. We also inspect the research approach of papers published in the last year at major NLP conferences (ACL, NAACL, EMNLP, AACL, EACL) that propose new SOTA models for RC, or for the related tasks of end-to-end RE and few-shot RC (Tan et al., 2022; Liu et al., 2022; Zhou and Chen, 2022; Wang et al., 2022b; Zhenzhen et al., 2022; Guo et al., 2022; Wang et al., 2022c; Zhang et al., 2022c; Zhang and Lu, 2022; Tang et al., 2022; Zhang et al., 2022a; Wang et al., 2022a; Duan et al., 2022; Guo et al., 2023; Wan et al., 2023). We find that that the common procedure consists of the five steps earlier mentioned. Specifically, we found that in most cases, the intuition (step 1) that is used as a starting point and as a motivation for the model improvement is based on generic observations of the model architecture, instead of on a quantitative analysis which could lead to more effective targeted improvements.

In this work, we propose a systematic quantitative analysis of a SOTA

RC model to detect sets of challenging instances sharing common characteristics (e.g., entity distance). The goal is to identify hard-to-handle setups for the SOTA architecture. Importantly, our approach is easily reproducible in future setups with different models, and/or on different datasets. The relevance of performing an in-depth analysis is supported by a demonstration of how the acquired information can help to effectively address the weaknesses of the baseline and design a new SOTA. Our contributions are:[1]

- We provide a tool for comprehensive quantitative analyses of RC model performance.

- We exploit the proposed analysis for investigating the performance of a SOTA RC architecture across 36 in- and cross-domain setups.

- Based on the findings of the analysis, we perform a case study improving the previous SOTA by over 3 points Micro-F1.

10.2   Related Work

Analysis of NLP Models   In this study, we are inspired by the recent trend targeting the evaluation of NLP models. Ribeiro et al. (2020) propose a task-agnostic methodology for testing general linguistic capabilities of NLP models by generating ad-hoc test instances; they test their approach over three tasks: sentiment analysis, Quora question pair, machine comprehension. Liu et al. (2021a) presents a software package for diagnosing the strengths and weaknesses of a single system, allowing for interpretation of relationships between multiple systems, and examining prediction results. They go a bit deeper into the task specificity, therefore their system currently supports the tasks of text classification (sentiment, topic, intention), aspect sentiment classification, Natural Language Inference (NLI), Named Entity Recognition (NER), Part-of-Speech (POS) tagging, chunking, Chinese Word Segmentation (CWS), semantic parsing, summarization, and machine translation. Furthermore, Fu et al. (2020a) and Fu et al. (2020b)

---

[1]Project repository: `https://github.com/mainlp/RC-analysis`

introduce the concept of interpretable task-specific evaluation. The first target the comparison of a set of NER systems. The latter, instead, perform a deep evaluation of CWS systems proving that despite the excellent performance achieved on some datasets, there is no perfect system for CWS. This concept has also been applied by Fu et al. (2021) for interpreting the results over a set of sequence tagging setups (NER, CWS, POS, chunking). Within the field of Information Extraction, previous work explored error-driven analysis for the automatic categorization of model prediction errors (Das et al., 2022).

Analysis of RC Models   Error analysis and in-depth evaluations of NLP systems are tied to specific tasks because of the peculiarities of each of them in terms of linguistic challenges, input type, and expected output. Relation Classification and related tasks (e.g., end-to-end RE) have received little attention in the context of systematic quantitative evaluation. Pre-Large Language Models, Katiyar and Cardie (2016) performed a manual evaluation of bi-directional LSTMs for the extraction of opinion entities and relations ("is-from", "is-about") by discussing the model output of a couple of instances. The same authors (Katiyar and Cardie, 2017) performed an error analysis, also based on a manual evaluation, comparing their model with Miwa and Bansal (2016). More recently, instead, some work has inspected the quality of RC corpora. Alt et al. (2020) analyze the impact of potentially noisy crowd-based annotations in the widely adopted TACRED (Zhang et al., 2017b). Lee et al. (2022) target the specific problem of overlapping instances between train and test sets in two popular RC benchmarks, namely NYT (Riedel et al., 2010) and WebNLG (Gardent et al., 2017).

Driven by the popularity of the task, and the contrasting lack of in-depth quantitative evaluation of RC systems, we fill this gap with an evaluation analysis suite for RC, and a case study including 36 in- and cross-domain setups.

10.3   Background

10.3.1   Cross-domain Relation Classification

Given a sentence and two entity spans within it, the task of RC aims at classifying the semantic relation between them into a type from a pre-defined label set. The task is currently far from being solved, in particular when considering realistic challenging setups, for example document-level RC (Yao et al., 2019), or few-shot RC (Han et al., 2018; Gao et al., 2019). In this study, we consider the cross-domain setup, where the challenge lies in different text types and label distributions from train to evaluation set. The cross-domain setup is important for testing the robustness of models aginst data shift. Despite the research in this direction from previous years, mainly evaluated on ACE (Doddington et al., 2004) where the domains are not particularly distinctive (Fu et al., 2017; Pouran Ben Veyseh et al., 2019), recent work on more challenging scenarios show very low performance due to data sparsity across domains. For example, cross-dataset (Popovic and Färber, 2022), or evaluated on the recently published CrossRE dataset (Bassignana and Plank, 2022a) which consists of data from six diverse text domains. In this study, we aim at improving the CrossRE baseline by systematically identifying challenging scenarios for the model.

10.3.2   Experimental Setup

CrossRE (Bassignana and Plank, 2022a),[2] is a manually-annotated dataset for cross-domain RC including 17 relation types spanning over six diverse text domains: artificial intelligence (🤖), literature (📖), music (🎵), news (📰), politics (🏛), natural science (🍃). The dataset was annotated on top of CrossNER (Liu et al., 2021b), a Named Entity Recognition (NER) dataset. Appendix 10.7.1 reports the statistics of CrossRE.

We use the baseline model of the original paper.[3] Following the archi-

---

[2]Released with a GNU General Public License v3.0.
[3]https://github.com/mainlp/CrossRE

tecture proposed by Baldini Soares et al. (2019), the model by Bassignana and Plank (2022a) augments the sentence with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} surrounding the two entities. The augmented sentence is then passed through a pre-trained encoder, and the classification made by a linear layer over the concatenation of the start markers $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$. We run our experiments over five random seeds and report the average performance. See Appendix 10.7.3 for reproducibility details.

10.4 Attribute Guided Analysis

We propose a systematic quantitative analysis of the performance of the CrossRE baseline model's performance across the 36 in- and cross-domain setups derived from training and testing the model on the six domains included in CrossRE. The analysis is performed over the development sets of the dataset. Inspired by the work of Fu et al. (2020a) on Named Entity Recognition, we introduce the first evaluation suite for RC, opening the way to other similar structured prediction tasks. The analysis evaluates the performance of the model over instances grouped by common values of potentially influential attributes (e.g., entity distance, sentence length). In what follows, we will describe the attributes considered and the bucketing strategy employed for splitting the evaluation instances based on the attribute values. Last, we go through the results of our correlation analysis.

10.4.1 Attributes

In our analysis, we consider 11 different attributes. These are characteristics of the RC instances that could challenge the model and influence its performance. Given an RC instance defined by a triplet *(e1, e2, r)* where *e1* is the head entity, *e2* is the tail entity, and *r* is the relation type connecting *e1* with *e2*; and given a sentence *s* expressing the relation *r* between *e1* and *e2*, we define the attributes listed in Table 10.1. We categorize each of them in the following three divisions:

Attribute	Description	Value Type		Computation		Level		
		DISCR.	CONT.	LOCAL	AGGR.	ENT.	REL.	SENT.
entity type*	the types of *e1* and *e2*	✓		✓		✓		
relation type	the type of *r*	✓		✓			✓	
IV entities	in-vocabulary entities: the amount of entities which appear in the train set (values 0, 1, or 2)	✓		✓		✓		
entity length	the sum of the number of tokens in *e1* and *e2*		✓	✓		✓		
entity distance	the number of tokens separating *e1* from *e2*		✓	✓			✓	
sentence length	the number of tokens in *s*		✓	✓				✓
entity density	the total number of entities in *s* over the sentence length (in percentage)		✓	✓				✓
relation density	the total number of semantic relations in *s* over the sentence length (in percentage)		✓	✓				✓
OOV token density	the amount of out-of-vocabulary tokens in *s* with respect to the train set over the sentence length (in percentage)		✓	✓				✓
entity type frequency*	the frequencies of the types of *e1* and *e2* in the train set		✓		✓	✓		
relation type frequency	the frequency of the type of *r* in the train set		✓		✓		✓	

Table 10.1: **Relation Classification Attributes.** Description of the 11 RC attributes and categorization in DISCRETE/CONTINUOUS value type, LOCAL/AGGREGATE computation, and ENTITY/RELATION/SENTENCE level. (⋆): We map the original 36 domain-specific entity types defined by Liu et al. (2021b) into five more generic types shared across domains, see Appendix 10.7.2 for details.

- **Value Type:** If the values of the attribute belong to a set of pre-defined values the attribute is DISCRETE (e.g., the entity type), otherwise it is CONTINUOUS (e.g., the entity distance).

- **Computation:** If the attribute is computed by only considering the current instance it is LOCAL, if it is computed over aggregated properties of the train set, it is AGGREGATE; for example, the frequency of entity and relation types refers to the training statistics.

- **Level:** If the attribute value depends on the entities it is at ENTITY LEVEL, if it depends on properties of the entity pair it is at RELATION LEVEL, last if it is related to characteristics of the sentence *s* it is at SENTENCE LEVEL.

As an attribute example, Figure 10.1 shows the `entity distance` distribution, measured as number of tokens separating *e1* from *e2*. The plot
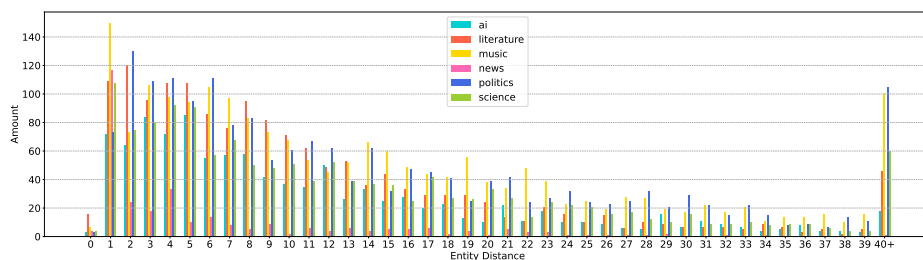
Figure 10.1: `entity distance` **Distribution.** Distribution of the `entity distance` values across the six development sets of CrossRE (Bassignana and Plank, 2022a).

reveals some domain-specific peculiarities, e.g., music and politics have the longest distances. This is mostly due to the long lists present in these domains, where the head entity is mentioned at the beginning and linked to all the elements in the list. For example, a music genre and a list of musical artists representing it; or the artifacts (i.e., songs and albums) of a band. We use the attribute values in order to group the evaluation instances with similar characteristics. We discuss the bucketing strategy in the next section.

10.4.2 Methodology

Once identified the potential influential attributes for the task of RC, the next step is splitting the evaluation sets depending on the attributes values (i.e., bucketing). For the attributes with DISCRETE value types (see Table 10.1) the bucketing creates one subset for each attribute values—e.g., one subset for each entity type for the `entity type` attribute. For the attributes with CONTINUOUS value types, instead, we set the number of buckets to four in order to maintain a reasonable size for each bucket. We then split the instances by equally distributing them across subsets—except for the two AGGREGATE attributes, which by definition are computed over properties of the train set. Note that the `entity type` and `entity type frequency` have each instance placed into two buckets, one considering the type of *e1* and one considering the type of *e2*.

	IV entities	entity length	entity distance	sentence length	entity density	relation density	OOV token density	entity type frequency	relation type frequency
avg. correl	0.1	0.0	-0.4	0.3	0.1	0.2	0.0	-0.3	0.9
avg. stdev	22.2	7.1	6.1	6.4	7.0	5.9	9.9	14.6	24.9

Table 10.2: **Overall Results.** Average correlation and average standard deviation of the Micro-F1 scores of the buckets (within attribute), averaged over the 36 train-test setups.

We measure the performance of the model over the subsets, and compute the Spearman's rank correlation coefficient with respect to the average attribute values of the buckets. Since `entity type` and `relation type` have categorical values, we cannot compute the correlation coefficient and analyze these two attributes separately in Section 10.4.3.1.

10.4.3 Results

In this section we are going to present the results of our analysis, first looking at the overall correlation study, and then at the per-domain results.

Overall Table 10.2 reports the correlations for the proposed attributes (Section 10.4.1) averaged across all 36 setups. We also report the average standard deviation across the Micro-F1 scores achieved within attribute and computed separately for each train-test setup. The `relation type frequency` is by far the most influential attribute: It reports the highest absolute correlation value, and the highest standard deviation between buckets including low- and high-frequent relations types in the train sets. In the current setups with relatively small training sets (see CrossRE statistics in Appendix 10.7.1) the amount of training instances have an high impact on the final performance of the model. In addition, this is also influenced
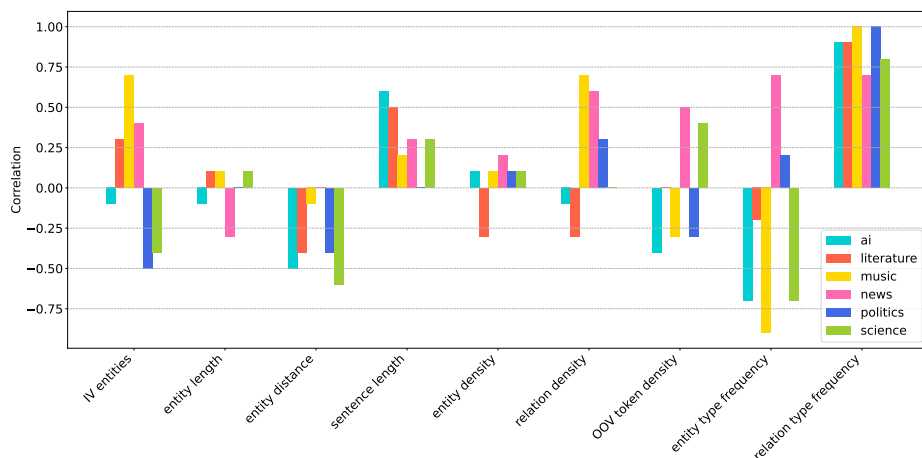
Figure 10.2: **Per-domain Correlation Analysis.** Spearman's rank correlation coefficient of the the 36 considered setups, averaged over the dev sets.

by the cross-domain setup, with diverse relation label distributions over the six domains (see Figure 10.3). The second most relevant attribute is `entity distance`, with the second highest absolute value in correlation and a 6.1 average standard deviation across buckets containing entity pairs at different distances. The `entity type frequency` presents a weaker correlation, confirming the findings that we will discuss in Section 10.4.3.1 about the `entity type`. All the other attributes report an absolute correlation value ranging between 0.2 and 0.0 indicating that within the overall overview of the considered setups they have a lower impact on the model's performance.

Domain Level   We visualize the average across the test domains in Figure 10.2. As previously noted, the `relation type frequency` trend confirms that the amount of training instances is the most influential attribute within the current setup. The `entity distance` and `sentence length` also present a similar trend across all six domains. The negative correlation of the first indicates that, as we could intuitively expect, it is more challenging to identify the semantic relation connecting two entities which

are far apart in the sentence, with respect to entity pairs separated by only a couple of tokens. The positive trend within the `sentence length` attribute, instead, suggests that entity pairs belonging to long sentences (i.e., where more context is given) are easier to classify than the ones from short sentences. The `entity density`, and `relation density` attributes present a general positive trend in correlation, but with some outliers (literature and AI). High values in these attributes refer to sentences with many instances, e.g., lists of entities which are all linked to an head entity with a similar structure and (most likely) with the same relation type. For example, in the music domain, a list of songs authored by a music artist, or by a band. We speculate these to be easy patterns to identify and learn by a deep learning model.

News is often an outlier with respect to the other domains. When training on this domain the performance drops with higher values of `entity length` (instead of improving as for most of the other domains), and for `entity type frequency` is exactly the reverse. The latter is probably due to the entity type hierarchy adopted, which maps the domain-specific entity types defined by Liu et al. (2021b) for the other five domains into the types included in the news domain. However, it should be noted that news comes from a different data source and has ∼4 times fewer relations compared to the other domains, which makes the results more unstable (Bassignana and Plank, 2022a).

10.4.3.1 Categorical analysis

For the two categorical attributes it is not possible to compute the correlation coefficients.

`relation type`   The results in Figure 10.3 reveal that some of the types are easier to learn across all domains than others (i.e. have higher scores, despite their lower frequency). These can be explained because they occur in very similar linguistic constructions, like "named", which often connects an entity to the consecutive acronym in brackets. Or because
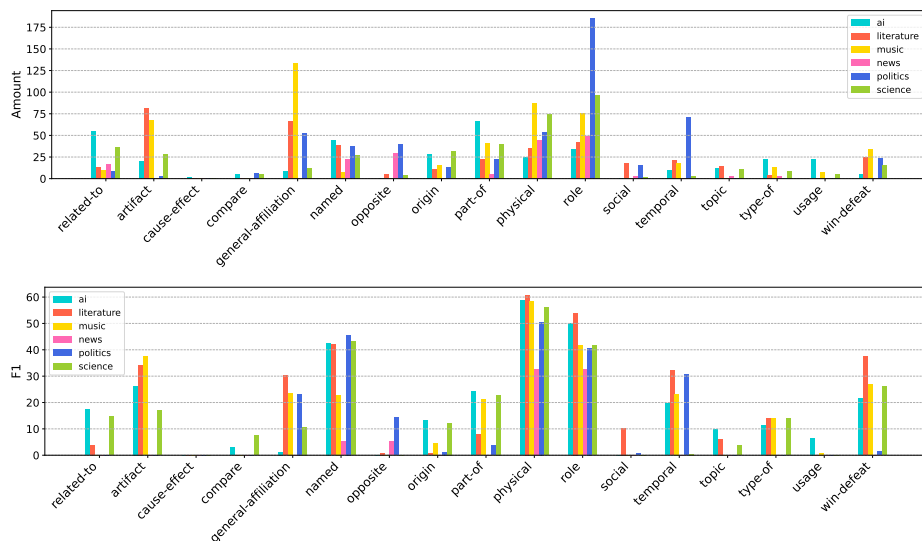
Figure 10.3: `relation type` **Analysis.** Distribution of the relation types in the train sets of CrossRE (Bassignana and Plank, 2022a) (above), and F1 per label (bottom).

they mostly occur with the same entity types, like "temporal" with "event" and "physical" with "location". On the other hand, some relation labels have different performances across domains. For example "win-defeat" which in the domains of AI, literature, music, and science mostly links a person winning an award. In the politics domain, instead, it refers to more complex scenarios where one out of multiple mentioned political parties wins the election. Or, in a completely different semantic context, a country wins a war against another country. Unsurprisingly the most difficult are clearly the infrequent ones, like "cause-effect".

`entity type` The results in Figure 10.4 show that there is not a strong link between the amount of training instances and the performance achieved, confirming the findings from Figure 10.2. This is because in the CrossRE guidelines there are no constraints linking the relation types to specific entity types. The higher scoring types are mostly the ones that are implicitly associated with specific relation types, e.g., "location" with the "physical"
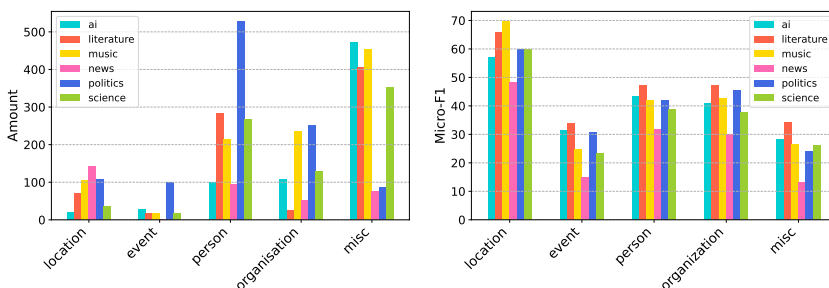
Figure 10.4: `entity type` **Analysis.** Distribution of the entity types in the train sets of CrossRE ([Bassignana and Plank, 2022a](#)) (above), and Micro-F1 achieved on each bucket (bottom).

relation type, and "event" with "temporal". On the other hand, the most varied category "misc" is the most challenging (see entity mapping in Table 10.5).

10.5 Application: Model Improvement

As mentioned in the introduction, our final aim is to guide better modeling by targeting quantitatively measured weaknesses of the model. Here we present a case study which exploits the findings of our proposed analysis. From the overall results in Table 10.2 we can derive that the most influential attribute is the `relation type frequency`, with a correlation of 0.9 and the highest standard deviation of 24.9. Targeting this factor would mean obtaining additional training data by manual annotation or via some data augmentation techniques. Within this case study, we aim to focus on improving the model architecture. Therefore, here we target the `entity distance` attribute, which holds the second highest absolute correlation (0.4), for improving the model performance.

10.5.1 Improved Experimental Setting

The fact that the `entity distance` (i.e., the number of tokens separating *e1* from *e2*) has a high influence on the RC model performance, means

(a) BASELINE

(b) FIRST-TWO

(c) LAST-TWO

(d) ALL-FOUR
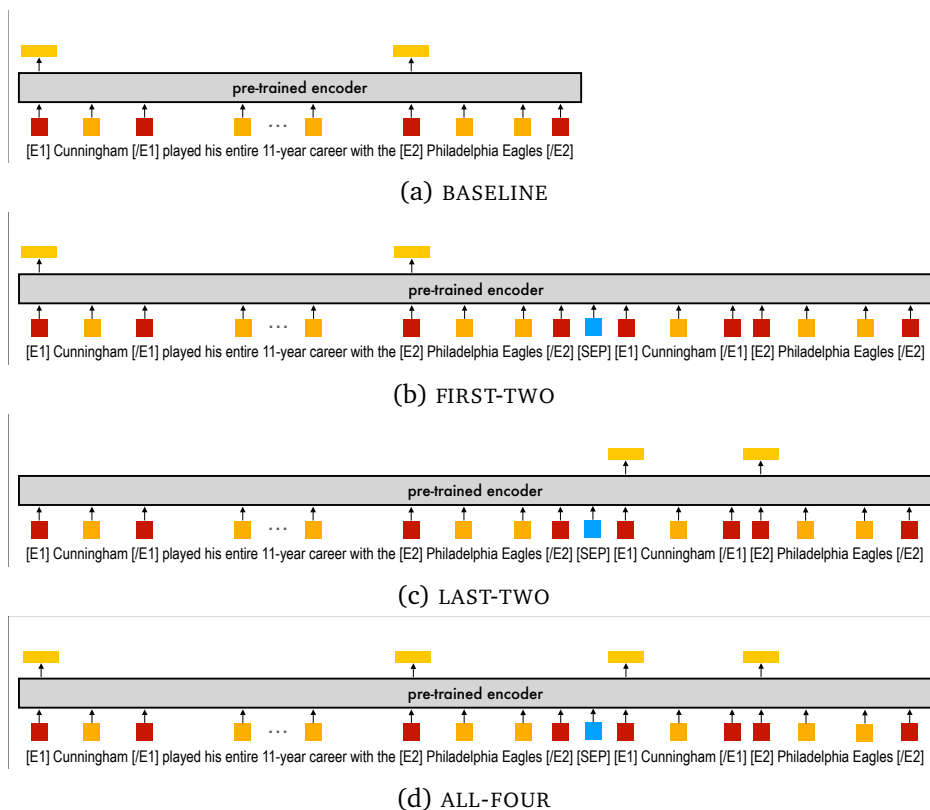
Figure 10.5: **Proposed Setups.** Representation of the baseline architecture (a) and of the three proposed setups (b, c, d) which include the repetition of *e1* and *e2* at the end of the sentence.

that the tokens between *e1* and *e2* can somehow mislead the prediction. In order to target this issue, we aim at moving the two involved entities closer to each other. We repeat the entities at the end of the original sentence representation augmented with the entity markers. Then, similar to the original CrossRE baseline (Section 10.3.2), we pass the input through a pre-trained encoder and extract a representation on which we do the classification of the relation with a linear layer. We test out three different representations as illustrated in Figure 10.5:

- FIRST-TWO concatenation of the representation of the first two entity markers start, as in the original baseline setup;

- LAST-TWO concatenation of the representation of the last two entity markers start, the ones introduced after the [SEP] token;

- ALL-FOUR concatenation of the representation of all four entity markers start, the original ones and the newly introduced.

In what follows, we show the effectiveness of moving the entities closer to each other, and compare the three classification strategies described above. The new model architectures are also included in our project repository.[4]

10.5.2   New SOTA Results

Table 10.3 compares the performance of the original baseline architecture with our proposed settings. In general, performances are higher with the repeated entities, except for the news domain, which achieves the least stable results across all our analyses. As pointed out by the authors of the dataset, this is the most challenging domain because it comes from a different data source and contains the least amount of instances, making the scores more unstable with respect to the other domains (Bassignana and Plank, 2022a). Furthermore, ALL-FOUR consistently outperforms FIRST-TWO and LAST-TWO. The gain of the overall average is even larger compared to the sum of both individual gains, suggesting that they provide highly complementary insights. The obtained improvements are substantial (> 3 points on average), and come at negligible costs—e.g., without drastically increasing the training time with pre-training steps. We perform significance testing in Appendix 10.7.4.

---

[4]https://anonymous.4open.science/r/RC-analysis-sSEM-3B2A

	TRAIN	🤖	📖	🎵	📰	🏛	🌿	avg.
BASELINE	🤖	46.4	32.9	27.5	44.6	36.4	35.3	37.2
	📖	28.0	63.1	55.5	34.7	49.0	35.4	44.3
	🎵	25.3	44.2	70.8	38.8	37.2	29.9	41.0
	📰	12.6	15.8	16.4	52.6	33.5	21.6	<u>25.4</u>
	🏛	20.1	34.0	40.6	40.5	55.8	31.2	37.0
	🌿	35.9	29.0	30.0	41.4	37.8	38.0	35.3
	avg.							36.7
FIRST-TWO	🤖	45.2	33.2	28.4	40.7	35.8	33.7	36.2
	📖	25.7	66.4	64.2	37.8	53.6	35.8	**47.3**
	🎵	27.5	48.4	71.6	36.9	42.2	30.6	**42.8**
	📰	14.1	17.0	18.9	43.6	35.5	23.2	25.3
	🏛	18.4	33.4	41.3	43.2	56.6	31.1	**37.3**
	🌿	36.8	28.6	30.2	40.7	36.3	38.6	35.2
	avg.							**37.4**
LAST-TWO	🤖	45.0	35.1	31.7	41.4	39.7	34.6	**37.9**
	📖	25.1	68.9	68.7	38.6	51.5	34.8	**47.9**
	🎵	28.6	57.6	73.2	38.2	39.1	32.4	**44.8**
	📰	9.9	14.4	17.7	33.3	29.8	19.4	20.8
	🏛	15.7	28.7	38.6	42.2	55.6	29.9	35.1
	🌿	33.2	31.0	35.8	42.0	41.6	40.9	**37.4**
	avg.							**37.3**
ALL-FOUR	🤖	46.5	36.2	32.2	48.1	42.0	37.5	**<u>40.4</u>**
	📖	25.8	69.4	68.2	40.1	53.9	35.8	**<u>48.9</u>**
	🎵	29.6	59.1	74.6	37.7	46.0	33.6	**<u>46.8</u>**
	📰	12.8	16.3	20.5	41.4	32.9	21.4	24.2
	🏛	19.4	32.9	41.9	43.7	58.3	33.1	**<u>38.2</u>**
	🌿	38.0	31.8	34.2	45.8	44.9	41.3	**<u>39.3</u>**
	avg.							**<u>39.6</u>**

Table 10.3: **Performance Comparison Across Setups.** Micro-F1 scores achieved with the baseline architecture, and with the three proposed variants. (**bold**): Scores beating the baseline; (<u>underline</u>): Highest scores within the four setups.

10.6   Conclusion

We present a tool for systematic quantitative analysis of the performance of RC models, and conduct the first in-depth analysis of an RC system, across 36 in- and cross-domain setups. We identify potentially influential attributes, and correlate their value with model performance. Our findings highlight the influence of data scarcity of relation types over the model performance. The second most correlated attribute is the distance between the two entities: The further away, the more challenging it is to classify the semantic relation between them.

Last, we provide a case study exploiting the findings of the analysis for improving the baseline architecture with a simple yet effective method. We target the entity distance weakness, and by repeating the entities closer to each other at the end of the sentence we achieve a new SOTA on CrossRE, with an average improvement > 3 points Micro-F1. We provide code for reproducing the proposed analysis on other RC setups (or related tasks, e.g., end-to-end RE). And we also release the code of the new SOTA architecture.

Our aim is to encourage preliminary quantitative analysis of models prior to designing new architectures. Future work includes expanding the set of attributes proposed in this work for RC in order to comprise other tasks, with different challenges.

Ethics Statement

We do not foresee any potential risk related to this work. The data we use is published freely by Liu et al. (2021b) and Bassignana and Plank (2022a).

Limitations

In this work we report a case study of our proposed evaluation suite over CrossRE which includes six datasets covering six text domains. We focused mainly on the current SOTA model, future work could consider

	SENTENCES				RELATIONS			
	train	dev	test	**tot.**	train	dev	test	**tot.**
AI	100	350	431	881	350	1,006	1,127	2,483
literature	100	400	416	916	397	1,539	1,591	3,527
music	100	350	399	849	496	1,861	2,333	4,690
news	164	350	400	914	175	300	396	871
politics	101	350	400	851	502	1,616	1,831	3,949
science	103	351	400	854	355	1,340	1,393	3,088
tot.	668	2,151	2,446	**5,265**	2,275	7,662	8,671	**18,608**

Table 10.4: **CrossRE Statistics.** Number of sentences and number of relations for each domain of CrossRE (Bassignana and Plank, 2022a).

more models and datasets. The set of attributes is mostly bound to the RC task, but other relation-based tasks could employ similar attributes. More aspects could be included in the analysis in order to inspect specific strengths and weaknesses of the model, or in order to adapt it to other related structured prediction tasks. Last, with respect to the model improvement in Section 10.5, we focus on the architecture of the RC model, but given the high impact of the relation type frequency attribute, data augmentation techniques could be explored in order to further improve the performance of the model.

Acknowledgements

10.7   Appendix

10.7.1   CrossRE Statistics

We report in Table 10.4 the dataset statistics of CrossRE (Bassignana and Plank, 2022a).

person	location	miscellaneous	
researcher	country	field	program language
writer		task	product
musical artist		algorithm	metrics
politician		book	literary genre
scientist		award	poem
organization	**event**	magazine	music genre
university	election	song	album
band	conference	musical instrument	discipline
political party		enzyme	chemical element
		chemical compound	protein
		astronomical object	theory
		academic journal	

Table 10.5: **Entity Hierarchy.** Mapping of the original 39 domain-specific entity types by Liu et al. (2021b) into five domain-agnostic meta types.

Parameter	Value
Encoder	`bert-base-cased`
Classifier	1-layer FFNN
Loss	Cross Entropy
Optimizer	Adam optimizer
Learning rate	$2e^{-5}$
Batch size	32
Seeds	4012, 5096, 8257, 8824, 9908

Table 10.6: **Hyperparameters Setting.** Model details for reproducibility of the experiments.

10.7.2 Entity Type Mapping

The CrossRE dataset adopts the 39 domain-specific entity types initially proposed by Liu et al. (2021b) in CrossNER. When dealing with the `entity type` and `entity type frequency` attributes, in order to perform our cross-domain analysis, we map the original 39 entity types into five domain-agnostic meta entity types as illustrated in Table 10.5.

10.7.3 Reproducibility

We report in Table 10.6 the hyperparameter setting of our RC model (see Section 10.3.2). All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU.

	BASELINE	FIRST-TWO	LAST-TWO	ALL-FOUR
BASELINE	1.0	0.8	0.8	0.9
FIRST-TWO	**0.0**	1.0	**0.1**	1.0
LAST-TWO	**0.0**	**0.3**	1.0	1.0
ALL-FOUR	**0.0**	**0.0**	**0.0**	1.0

Table 10.7: **Significance Testing.** ASO scores comparing the experimental setups described in Section 10.5. Read as row → column.

10.7.4   Significance Testing

We compare our setups using the Almost Stochastic Order test (ASO; Del Barrio et al. (2018); Dror et al. (2019)) implementation by Ulmer et al. (2022b). The method computes a score (ϵ_{min}) which represents how far the first is from being significantly better in respect to the second. The possible scenarios are therefore $\epsilon_{min} = 0.0$ (*truly stochastic dominance*), and $\epsilon_{min} < 0.5$ (*almost stochastic dominance*). Table 10.7 reports the ASO scores with a confidence level of $\alpha = 0.05$ adjusted by using the Bonferroni correction (Bonferroni, 1936). See Section 10.5 for the setup details.

Chapter 11

Dissecting Biases in Relation Extraction: A Cross-Dataset Analysis on People's Gender and Origin

Abstract

Relation Extraction (RE) is at the core of many Natural Language Understanding tasks, including knowledge-base population and Question Answering. However, any Natural Language Processing system is exposed to biases, and the analysis of these has not received much attention in RE. We propose a new method for inspecting bias in the RE pipeline, which is completely transparent in terms of interpretability. Specifically, in this work we analyze biases related to gender and place of birth. Our methodology includes (*i*) obtaining semantic triplets (subject, object, semantic relation) involving 'person' entities from RE resources, (*ii*) collecting meta-information ('gender' and 'place of birth') using Entity Linking technologies, and then (*iii*) analyze the distribution of triplets across different groups (e.g., men versus women). We investigate bias at two levels: In the training data of three commonly used RE datasets (SRED$^{\text{FM}}$, CrossRE, NYT), and in the predictions of a state-of-the-art RE approach (REBEL). To enable cross-dataset analysis, we introduce a taxonomy of relation types mapping the label sets of different RE datasets to a unified label space. Our findings reveal that bias is a compounded issue affecting underrepresented groups within data and predictions for RE.

11.1   Introduction

Language technologies are widely spreading throughout our everyday life. However, it has been demonstrated that these technologies are often affected by the presence of gender and racial biases (Kurita et al., 2019; Tan and Celis, 2019). "Bias" is a cover term for a number of issues, which according to Hovy and Prabhumoye (2021) may emerge at any stage of the Natural Language Processing (NLP) pipeline. They could come from the data curation process (Sap et al., 2019), be intrinsic into the trained model (Zhao et al., 2017), or they could derive from the cultural background of NLP practitioners (Santy et al., 2023). An orthogonal taxonomy of biases distinguishes between *allocative* and *representational*

ones (Suresh and Guttag, 2021). *Allocative* biases regard the unequal distribution of opportunities across different groups, such as disparity in granting loans (Hardt et al., 2016) or the systematic exclusion of certain minorities from public archives (Weathington and Brubaker, 2023). *Representational* biases focus on stereotypical associations between groups and certain features (Caliskan et al., 2017) (e.g., women and lexicon about marriage and parenthood). Blodgett et al. (2020) show that existing works in NLP mainly focus on *representational* biases while the *allocative* ones are often overlooked.

In this context, Relation Extraction (RE) techniques represent a powerful tool to jointly explore the two types of bias described above. RE methods extract fine-grained triples from texts (subject, object, and the semantic relation connecting them), allowing for the discovery of gaps in digital archives. Previous work performed event extraction on Wikipedia biographies to study the presence of systematic gender biases in this archive (Sun and Peng, 2021; Stranisci et al., 2023). Gaut et al. (2020) collected a distantly supervised dataset from Wikipedia for exploring gender bias in RE, but they only include four relation types ('spouse', 'hypernym', 'birthDate', 'birthPlace'). Despite this preliminary work, standards for the adoption and evaluation of RE techniques for bias detection are still missing and are limited to the analysis of gender. Furthermore, before using RE for bias detection there is the pressing need to explore whether these systems portray any themselves.

In this paper, we explore the presence of biases in RE, both at the level of data (by analyzing the training data) and model (by analyzing the model predictions). We illustrate our procedure in Figure 11.1. As a first step, in order to enable cross-dataset analysis, we introduce a taxonomy of relation types mapping the label sets from different RE datasets into a unified label space. Then, as a second and third steps we collect information about people mentioned in a text. This includes semantic relations involving people (from RE), and meta-information related to them (i.e., 'gender' and 'place of birth'; using Entity Linking). As a last step, we explore the
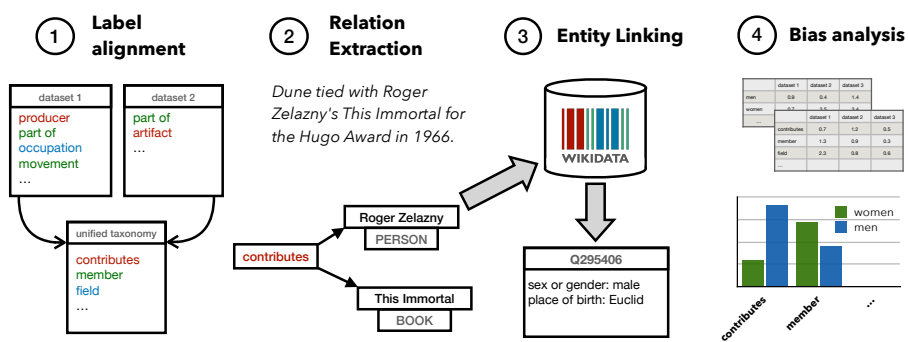
Figure 11.1: **Overview of our Proposed Methodology.** The first step aligns the label sets of different RE datasets into a unified taxonomy of relation types. In the second step, we extract semantic triplets including 'person' entities. Within the third step, we collect socio-demographic information from Wikidata of the people extracted in the second step. Finally, in the last step we analyze potential *allocative* and *representational* imbalances in the distribution of the extracted information (entities and relations) across different social groups (e.g., men versus women).

allocative and *representational* biases by inspecting potential imbalances into the distribution of the extracted triples across different groups (e.g., men versus women). Concretely, we investigate if any relation type (e.g., *member*, *contributes*) is more likely associated with one social group (more details in Section 11.5). We repeat our procedure both on the training sets on three widely adopted RE datasets: SRED^FM (Huguet Cabot et al., 2023), CrossRE (Bassignana and Plank, 2022a), NYT (Riedel et al., 2010); and on the predictions of a state-of-the-art RE approach, REBEL (Huguet Cabot and Navigli, 2021).

Not only do our findings corroborate existing research regarding the prevalence of gender biases in RE but they also broaden the discourse by uncovering biases along additional dimensions, such as origin. To our knowledge, this is the first investigation that examines bias through the lens of transfer learning and reveals the nuanced effects of simplistic interventions like data balancing. While such strategies may reduce biases for certain target groups, they can inadvertently introduce new biases,

underscoring the necessity for a more sophisticated, multi-axial approach for bias mitigation.

The contributions of this paper are:

- We introduce a meticulous bias analysis procedure for RE designed to be applicable across various dimensions, addressing both dataset and model-level biases.

- An in-depth analysis of biases related to 'gender' and 'place of birth' in the train sets of three widely adopted RE datasets and on the predictions of a SotA RE model on those.

- A taxonomy of relation types mapping the label sets of different RE datasets into a unified label space. The taxonomy makes our approach robust and versatile, and opens to cross-dataset analysis.

11.2   Related Work

Sun et al. (2019) and Blodgett et al. (2020) emphasize current issues in the research about bias detection and mitigation. The first presents a survey aimed at identifying research directions for gender bias detection, while the second criticizes how research in bias detection and mitigation is usually conducted. In order to make explicit potential biases in NLP, Bender and Friedman (2018) and Mitchell et al. (2019) propose to better document datasets and Language Models (LMs) respectively.

Some works released ad-hoc datasets to explore bias detection. Zhao et al. (2018) presented WinoBias, a dataset for coreference resolution aimed at testing stereotypical associations between women and certain types of profession. Nadeem et al. (2021) introduced StereoSet, for testing the presence of stereotypical knowledge in LMs while Gehman et al. (2020) released RealToxicityPrompt, a list of annotated prompts that is intended to measure the toxicity of text generated by LMs. Kiritchenko and Mohammad (2018) presented the Equity Evaluation Corpus, designed to measure gender and racial biases in models trained for sentiment analysis.

Several work on bias analysis focuses on inspecting the internal representation of NLP models. Caliskan et al. (2017) proposed two metrics for bias detection from word embeddings; May et al. (2019) from sentence encoders; and Kurita et al. (2019) from contextualized word embeddings. More recent approaches in this direction use probing strategies (Lauscher et al., 2022; Köksal et al., 2023). However, the outcome of these methods is often hard to interpret because of the black box nature of neural models. In order to prioritize interpretability of the results and obtain a more transparent bias analysis, we propose a new procedure for bias detection in RE technologies, which is applicable both at the level of data and model.

11.3 Methodology

We introduce a four-step procedure for detecting biases related to 'gender' and 'place of birth' in the Relation Extraction pipeline (see Figure 11.1). The method can be easily extended to explore other socio-demographic biases.

①First, we align the label spaces of different RE datasets using a unique taxonomy of relations with the aim of performing comparable analysis across corpora (details in Section 11.3.1).

②As a second step, we employ Relation Extraction in order to gather triplets (subject, object, relation) about people mentioned in a text. This can be done by filtering the triplets in which at least one of the two entities has type 'person'. We leverage the triplets in labeled training sets as well as in the predictions of systems trained using them.

③We collect socio-demographic data about people that are included in the biographical triplets extracted in step ②. We use Entity Linking (EL) to disambiguate the entity spans with type 'person' and link them to Wikidata (Vrandečić and Krötzsch, 2014) entries. We collect two types of meta-information from Wikidata: 'gender' and 'place of birth'.

④Last, given the triplets extracted in the second step and the socio-demographic information collected in the third step, we conduct bias analysis by investigating any imbalance in the distribution of relations

across different social groups (e.g., men versus women). Since it has been demonstrated that biases may occur at any stage of the NLP pipeline (Hovy and Prabhumoye, 2021), we applied our procedure for assessing the presence of biases both on the corpora used for training RE models and on the entities and relations predicted by them. Specifically, we investigate *allocative* bias in the training data (Section 11.5.1) and in the predictions made by these models (Section 11.5.3). Similarly, we examine *representational* bias, adapting metrics from earlier studies to evaluate both the training datasets (Section 11.5.2) and the predictions (Section 11.5.4).

11.3.1   Relation Type Taxonomy

RE datasets often include a label set with relation types which are too fine-grained with respect to our objective of exploring social biases related to 'gender' and 'place of birth' (e.g., *field-of-work* and *occupation* from SRED[FM]). Aggregating certain types to broader categories enables a higher-level analysis with enough samples per type that would be otherwise unfeasible with infrequent or narrow ones. In addition, we face the issue of lack of standards in dataset annotation for RE (Bassignana and Plank, 2022b), which prevents the comparison of results across corpora (e.g., the relation type */people/person/profession* in NYT versus *occupation* in SRED[FM]). To overcome these issues we introduce a taxonomy of relation types mapping the original types from the different datasets into a unified label space (e.g., *field-of-work*, *occupation* and */people/person/profession* to *field*). The taxonomy enables cross-dataset comparison and makes our approach versatile. Table 11.1 reports the ten newly introduced labels, with the co-occurring entity types (one entity type is always a person), and a corresponding example. The taxonomy is organized around the entity types that are part of the triplet. For instance, *contributes* is used to identify all triples with a person and a work, while *relationship* represents triplets where both subject and object are persons.

Relation type	Co-occurring entity	Example
contributes	work	In 2018, *Zhao* directed her third feature film, *Nomadland*, starring Frances McDormand
date	date	*Rosa Luxemburg* born Rozalia Luksenburg, *5 March 1871*
field	occupation, discipline	*Stephen William Hawking* was an English *theoretical physicist, cosmologist*
geographical relation	place	Born in *Ogidi*, Colonial Nigeria, *Achebe*'s childhood was influenced by both Igbo traditional culture and postcolonial Christianity
language	language	*Seedorf* speaks six languages fluently: *Dutch*, *English*, *Italian*, *Portuguese*, *Spanish* and *Sranan Tongo*
member	organization	Ahead of the 2009–10 season, *Ronaldo* joined *Real Madrid* for a world record transfer fee at the time of £80 million (€94 million)
participated	event	*Tim Burton* appeared at the *2009 Comic-Con* in San Diego, California, to promote both 9 and Alice in Wonderland
position held	organization	*Meredith Whittaker* is the president of the *Signal Foundation* and serves on their board of directors
relationship	person	*Billy Porter* married *Adam Smith* on January 14, 2017, after meeting him in 2009
topic	work	*Napoleon* appears briefly in the first section of Victor Hugo's *Les Misérables*, and is extensively referenced in later sections

Table 11.1: **Relation Type Taxonomy.** A list of biographical situations designed for RE. Labels are distinguished on the basis of the co-occurring entities in a triple. All examples are derived from the English Wikipedia.

	Train		Validation		Test	
	sent.	rel.	sent.	rel.	sent.	rel.
SRED^{FM}	1,199,046	2,480,098	6,333	13,322	3,015	6,474
CrossRE	297	1,220	835	3,483	891	3,604
NYT	19,709	26,267	1,765	2,318	1,773	2,327

Table 11.2: **Dataset Statistics.** Number of sentences and number of triplets (relations) for each dataset.

11.4 Experimental Setup

We follow the four-step procedure described in Section 11.3 to investigate biases in three commonly adopted RE datasets, and the predictions of a popular RE model. Below, we describe our experimental setup in terms of datasets (Section 11.4.1) and modeling (Section 11.4.2). Details about their licenses can be found in Appendix 11.8.2.

11.4.1 Datasets

SRED^{FM} (Huguet Cabot et al., 2023). The SRED^{FM} dataset is a distantly annotated dataset build on top of Wikipedia pages and Wikidata relations, employing a novel triplet critic filtering. The dataset covers 17 languages, but for the scope of this paper we employ only the English portion. Since this is the larger corpus in our study, we use it as a pre-training stage for the experiments on the other two datasets.

CrossRE (Bassignana and Plank, 2022a). CrossRE is a multi-domain dataset for RE containing data from the news, politics, natural science, music, literature and artificial intelligence domains. This dataset is the only entirely manually-annotated in our study. Given the small size of the six sub-sets, in our experiments we join the data across the different domains.

NYT (Riedel et al., 2010). NYT is a RE dataset consisting of news sentences from the New York Times corpus. It contains distantly annotated

relations using FreeBase. We use the processed version of Zeng et al. (2018) called NYT-multi.

For each of these datasets, we filter the triplets which include at least one entity 'person'. In Table 11.2 we report the statistics of the corpora after the filtering phase. In addition, following step ① in Section 11.3, we map the original relation types of the three datasets, into a unified label space defined by our taxonomy of relation types (Section 11.3.1). We report our mapping in Table 11.8 in Appendix 11.8.1.

11.4.2 Models

In steps ② and ③ of our proposed procedure (described in Section 11.3) we employ a Relation Extraction (RE) and an Entity Linking (EL) model respectively. Below we describe them both.

REBEL (Huguet Cabot and Navigli, 2021).   For RE, we employ the same setup as REBEL, a generative model based on BART (Lewis et al., 2020). We use the same default parameters as the original paper and train on top of BART-large.

EntQA (Zhang et al., 2022b).   To disambiguate the extracted entities 'person' and link them to Wikidata (Vrandečić and Krötzsch, 2014) we use EntQA, a recent state-of-the-art EL system based on the Retriever-Reader paradigm. We employ it to perform entity disambiguation on the entity spans extracted by REBEL. We only default to these predictions when the original dataset does not have a link to Wikidata, either because a span prediction was not labeled as an entity in the dataset, or because the original dataset did not include disambiguated entities. We use EntQA out-of-the-box (i.e., we do not fine-tune it on our datasets).

11.4.3 Relation Extraction Experiments

As mentioned in Section 11.4.1, we use SRED$^{\text{FM}}$ for pre-training REBEL before employing it on the two smaller datasets (CrossRE, NYT). We perform

| | Test | taxonomy | + SRED^{FM} pre-train | | |
			original	taxonomy	balanced
	SRED^{FM}		69.13	71.07	64.84
zero-shot	CrossRE		17.35	20.27	20.07
	NYT		28.58	32.89	33.66
fine-tuned	CrossRE	44.72	51.74	52.04	52.12
	NYT	89.26	88.47	88.52	89.83

Table 11.3: **Experiments Performance.** Micro-F1 scores of REBEL trained and evaluated on SRED^{FM}, zero-shot and fine-tuning evaluation on CrossRE and NYT. 'original' refers to a model trained on the original label set; 'taxonomy' indicates that the model was trained on the taxonomy mapping (see Table 11.8); 'balanced' stands for a gender-balanced version of it (see Section 11.6). First row indicates performance after pre-training on SRED^{FM} test set.

two categories of experiments: 'Zero-shot', where REBEL is pre-trained on SRED^{FM} and directly evaluated on CrossRE and NYT; and 'fine-tuning', where REBEL is both pre-trained on SRED^{FM} and fine-tuned on the target dataset.

Zero-shot Experiments.   In Table 11.3 we report the scores of REBEL trained on SRED^{FM} and evaluated on CrossRE and NYT in a zero-shot fashion. Evaluation is always done in the coarse-grained space of the taxonomy, either on the predictions of a model trained on SRED^{FM} mapped to the taxonomy (column 'taxonomy'), or by mapping the predictions of a model trained on the original labels to the taxonomy (column 'original'). Training on the taxonomy relation types improves the performance for both datasets. These results validate our proposed mapping as a way to unify label sets from different datasets.

Fine-tuning Experiments.   Similarly to the previous experiment, in Table 11.3 we report the scores of REBEL trained on SRED^{FM} and then fine-tuned on CrossRE or NYT, as well as regular fine-tuning without pre-training (left column). These experiments allow us to explore the use

	SRED^{FM}	CrossRE	NYT
Women	20.0%	11.8%	17.3%
Global South	18.9%	10.0%	12.2%

Table 11.4: **Allocative Bias in Training Data.** The percentage of women and Global South people in SRED^{FM}, CrossRE, and NYT corpora.

of our shared label space as a means of transfer learning across datasets and later on study how transfer learning affects the bias distribution (see Sections 11.5.3 and 11.5.4). Differences in performance are smaller than in the zero-shot counterpart, especially when enough data is available in the target dataset (NYT). Still, this experiment showcases that pre-training on the taxonomy improves performance on low data regimes while it has a small difference on larger ones.

11.5   Social Bias Analysis

In this section we report our bias analysis conducted on the training sets of the datasets described in Section 11.4.1 and on the predictions obtained with our trained models. In line with previous work on 'gender' bias analysis, we consider *men* versus *women* (Zhang and Terveen, 2021). For biases related to the 'place of birth', instead, we follow previous work and consider *Global North* versus *Global South* (Dirlik, 2007). We discuss more in details these division in the Limitation Section. We maintain the distinction between *allocative* and *representational* biases and explore both bias types at the level of training sets (Sections 11.5.1 and 11.5.2) and in the predictions (Sections 11.5.3 and 11.5.4).

11.5.1   Allocative Bias in Training Data

To assess the *allocative* bias in training data we compare the distributions across two axes between entities that are included in SRED^{FM}, CrossRE, and NYT: The distribution of women against men, and of people born in a Global South countries against ones born in the Global North. As ex-

plained in Section 11.3 we gather this meta-information about people from Wikidata, a collaborative knowledge graph that is part of the Wikimedia ecosystem. Since the analysis relies on metadata extracted from Wikidata, we are only able to compare people whose information about their 'gender' (Wikidata ID P21) and 'place of birth' (Wikipedia ID P19) are available. This did not have an impact on the analysis of 'gender', while the Wikidata gap with respect to 'place of birth' is 31% of people from SRED$^{\text{FM}}$, 8% from CrossRE and 11% from NYT. Once we obtained this information, in Table 11.4 we observe the distribution of women and Southern people in order to understand to which extent they are underrepresented in RE corpora. CrossRE is the corpus where both categories are less represented while in SRED$^{\text{FM}}$ they benefit from a higher representation. Overall, the analysis shows a significant underrepresentation of women and people born in the Global South across all corpora, always falling in a range between 10% and 20% of the total. This is even more daring when considering that the Global South accounts for around 80% of the world population. We also want to stress that these allocative biases are compounded from several sources. All our datasets are in English, and from sources that target an English speaking audience. Wikidata and Wikipedia showcase a skewed gender distribution where only 25% and 20% respectively of people's pages are women (Zhang and Terveen, 2021), furthermore Wikipedia collaborators are 83% male.[1] The annotation process for each of the datasets we analyze may also introduce further biases. Our goal here is not to pinpoint where these biases originated but rather how they are reflected in RE resources.

11.5.2   Representational Bias in Training Data

The analysis of *representational* biases relies on a Monte Carlo experiment that simulates a balanced distribution of people along the axes of 'gender' (men *vs* women) and 'place of birth' (Global North *vs* Global South). For each training set we perform an experiment structured in three parts: (*i*) We randomly pick 100 individuals for each group and average the number

---

[1] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians

	SRED^{FM}		CrossRE		NYT		SRED^{FM}		CrossRE		NYT	
	M	W	M	W	M	W	N	S	N	S	N	S
contributes	0.28	0.475	**0.407**	0.291	–	–	**0.758**	0.162	**0.447**	0.333	–	–
date	**1.038**	0.926	–	–	–	–	**1.07**	0.993	–	–	–	–
field	**0.388**	0.291	–	–	–	–	0.394	0.451	–	–	–	–
geographical	**0.469**	0.368	0.218	0.218	3.251	2.164	0.501	**0.64**	0.198	**0.644**	0.965	**1.019**
language	0.013	0.006	–	–	–	–	0.025	0.024	–	–	–	–
member	0.21	0.164	0.229	0.218	**0.739**	0.283	0.252	0.201	0.300	0.222	**0.169**	0.121
participated	0.088	0.049	**0.278**	0.145	–	–	0.052	0.08	**0.218**	0.133	–	–
position held	0.091	0.038	0.745	0.727	**0.085**	0.012	0.144	**0.196**	0.742	**1.200**	0.036	0.009
relationship	0.124	**0.215**	**0.098**	0.036	0.078	**0.211**	0.132	0.119	0.093	0.111	**0.077**	0.025
topic	0.001	0.001	0.018	0.018	–	–	0.002	0.002	**0.013**	0.0	–	–

Table 11.5: **Representational Bias in Training Data.** Results of the experiment aimed at identifying statistically-significant differences between social groups for each relation and across corpora. Values represent the proportion of each relation type per person. First six columns report the comparison between men (M) and women (W); last six between Global North (N) and South (S) people. For each relation, we report the group that is significantly more associated with it in bold, if neither is it means that there is not a statistically significant difference ($p \geq 0.05$).

of relation in which they are subject or object. (*ii*) We repeat the sampling 10 times for each distribution. (*iii*) For each relation type we calculate the t-test statistics between the 10 mean scores of a majority and a minority group. Results are reported in Table 11.5. For each relation we report the average per social group and whether there is a significant difference between the two groups. The comparison between genders shows that *member* and *position held* are significantly related to men in the NYT corpus, perhaps due to its nature as a news corpus, along with *geographical* (also in SRED^{FM}). *Relationship* is instead skewed towards women in SRED^{FM} and NYT, and towards men in CrossRE. From the comparison between Global North and South it emerges that the latter are always more associated to *geographical*. The *position held* property behaves differently across corpora: It is mostly related to South in SRED^{FM} and CrossRE, and to North people in NYT, which is also skewed towards this group for the *member* relation. *Relationship* is significantly associated to Global South people only in NYT.

In general, some trends emerge when comparing across datasets. The only gender bias that favors women concerns *relationship*, while all the other types (when significant) skew towards men, independently of the

	SRED^{FM}	CrossRE	NYT
Women	–	- 2.2%	+ 5.6%
+ SRED^{FM}	- 3.5%	- 5.8%	+ 0.6%
+ gen. balanced	- 2.9%	- 4.4%	0.0%
Global South	–	- 8.3%	- 2.1%
+ SRED^{FM}	- 1.7%	- 6.7%	- 1.6%
+ gen. balanced	- 0.3%	- 9.9%	- 5.9%

Table 11.6: **Allocative Bias in Prediction.** Percentage difference of women and Global South people in false positive and true positive predictions of the model when trained on each dataset (first row), fine-tuned on top of SRED^{FM} pre-training (second row) or fine-tuned on top of a gender-balanced SRED^{FM} pre-training (third row).

dataset. On the other hand, with respect to the North/South analysis, biases are more widespead and of different nature. Of the three datasets, SRED^{FM} shows less biases on this dimension, and coincidentally it is the one having a higher percentage of people from the Global South (see Table 11.4). It is worth noticing how the only bias favoring North shared across datasets (with a very high degree in SRED^{FM}) is *contributes*, which may be reflective of an overall cultural bias within the English Wikipedia, from which both SRED^{FM} and CrossRE are collected.

Summarizing, the analysis shows the presence of recurring *representational* biases against underrepresented groups, specifically for certain relation types: *relationship* for women, *geographical* for Global South. NYT includes the highest number of biases, where men and Northern people mostly appear in relations that emphasize their profession (*member*, *position held*).

11.5.3   Allocative Bias in Prediction

Our analysis on bias in predictions follows that of Gaut et al. (2020). For *allocative* bias we rely on the False Positive Balance score (FP_{Bal}) inspired by Hardt et al. (2016). This metric is a comparison between the percentage of entities belonging to an underrepresented group in the model's wrong

predictions and their distribution in the test and evaluation sets. A positive delta between these two percentages is interpreted as the model tendency to recognize entities from an underrepresented group. The analysis is performed on predictions obtained with and without SRED$^{\text{FM}}$ pre-training, while always fine-tuning on the target dataset (Table 11.3). This allows to assess the impact of SRED$^{\text{FM}}$ pre-training on the distribution of bias. Table 11.6 shows that women and Global South people are affected by *allocative* harms in different proportions and that these vary across corpora. The FP_{Bal} score is negative for women in CrossRE, while in NYT it is positive. Using the pre-trained model before fine-tuning amplifies this bias in CrossRE (from -2.2 to -5.8), while it lowers it in the NYT (from +5.6 to +0.06). The opposite happens if Global South people are considered. Given the fact that a negative FP_{Bal} emerges in all distributions, the pre-training step reduces this bias from -8.3 to -6.7 in CrossRE and from -2.1 to -1.6 in NYT.

In summary, while adopting SRED$^{\text{FM}}$ for transfer learning to CrossRE and NYT has a positive effect on the performance (CrossRE goes from 44.72 to 52.04, see Table 11.3), it has a mixed effect with respect to the biases. On one side, it amplifies the *allocative* biases for women in predictions, on the other it introduces a mitigation in favor of people from Global South. This could be explained by SRED$^{\text{FM}}$ showing a lower starting bias of -1.7 compared to the other datasets, and therefore acting as a mitigator when used as a pre-trained model. The opposite is observed for women, where SRED$^{\text{FM}}$ has a higher starting bias (-3.5).

11.5.4 Representational Bias in Prediction

We perform the *representational* bias analysis on the predictions by adopting the *Minority Recall Gap* metric (Rec_{Gap}). Inspired by the 'true positive rate gender gap' from De-Arteaga et al. (2019), our metric measures the differences in recall for predictions of two groups. Since the data used for evaluation is unbalanced and some relation types are rare, we only compute the Rec_{Gap} for types appearing at least 10 times in each corpus.

	gender			place of birth		
	SRED$^{\text{FM}}$	CrossRE	NYT	SRED$^{\text{FM}}$	CrossRE	NYT
contributes	+ 0.03	- 0.01	–	+ 0.04	- 0.30	–
date	+ 0.03	–	–	- 0.05	–	–
field	+ 0.05	–	–	- 0.03	–	–
geographical	- 0.09	+ 0.16	+ 0.04	+ 0.23	+ 0.15	+ 0.05
language	–	–	–	+ 0.29	–	–
member	- 0.12	- 0.10	–	–	- 0.10	–
participated	- 0.07	- 0.01	–	+ 0.10	- 0.06	–
position held	- 0.17	0.00	- 0.01	+ 0.10	- 0.04	- 0.02
relationship	+ 0.07	+ 0.14	- 0.17	+ 0.15	+ 0.03	+ 0.16
topic	–	–	–	–	–	–

Table 11.7: **Representational Bias in Prediction.** The Rec_{Gap} on the evaluation triples with respect to the underrepresented groups (i.e., positive values for women and people from the Global South). '–' means that the relation type appears less than 10 times.

Table 11.7 shows the Rec_{Gap} for each relation throughout all datasets. A positive value means that the model is more likely to retrieve a relation if it is associated to an underrepresented group (i.e., women and people from the South); on the opposite, a negative value means that the model is more likely to retrieve the relation type if it includes men or people from the Global North respectively. The analysis shows patters that already emerged in the training sets (Section 11.5.2). *Relationship* and *geographical* triples are more often retrieved when a woman or a Global South person represents its subject or object in five out of six cases. The only exceptions are SRED$^{\text{FM}}$, which achieves a Rec_{Gap} score of -0.09 in favor of men for *geographical*, and NYT, with a score of -0.17 in favor of men for *relationship*. The opposite happens for *position held*, which is mostly retrieved for Global South ($+0.10$) only in SRED$^{\text{FM}}$, while in all the other cases it always leans towards Global North. *Contributes* achieves a positive Rec_{Gap} in SRED$^{\text{FM}}$ and a negative one in CrossRE for both bias analysis, while *member* is always mostly associated with men or people from the North. The same happens for *participated*, except for 'place of birth' in SRED$^{\text{FM}}$. Finally, *field* and *date* are more associated with women and Global North.

These results mostly follow the trends in the training datasets (Section 11.5.2). Representational biases in predictions regard similar associations between certain categories of people and relation types: Women with *relationship*, Southern people with *geographical*, men and Northern people with *member*. However, the model seems to have its own impact on the propagation of biases. For instance, *field* does not present statistically significant differences between Global North and Global South in the training sets (see Table 11.5), but it is mostly associated to Northern people in the predictions. This behavior underlines the need of designing approaches for bias detection that encompass all the stages of the RE task.

11.6 Bias Mitigation

In this section we look at a common approach to tackle skewed distributions of data by balancing the pre-training data (SRED$^{\text{FM}}$) in order to obtain fairer representations of underrepresented groups. This mitigating strategy was the only one shown to be effective in Gaut et al. (2020). Since in Table 11.6 the 'gender' bias of SRED$^{\text{FM}}$ is more pronounced with respect to the bias related to the 'place of birth' (-3.5% versus -1.7%), we consider the 'gender' axis and re-train REBEL on a dataset with a balanced distribution across genders. In order to do so, we gather from SRED$^{\text{FM}}$ all triplets involving at least one woman, and then we add triplets involving men until we reach an equal amount. As a results, we obtained a dataset of $836,638$ instances, of which 50.7% involves at least one woman.

As it can be observed in the bottom line of Table 11.6, the adoption of a gender balanced pre-training dataset has a mitigation effect on the *allocative* biases against both underrepresented groups in SRED$^{\text{FM}}$. The FP_{Bal} decreases from -3.5% to -2.9% against women and from -1.7% to -0.3% against Southern people. The effect on the gender bias of the other datasets is also positive. The balanced distribution improves the FP_{Bal} score from -5.8% to -4.4% in CrossRE, and from $+0.06$ to 0 in the NYT corpus. However, balancing the gender axis has a negative impact on the *allocative* bias against people from the Global South both in CrossRE and

NYT. In CrossRE, it amplifies them from -8.3% to -9.9%, while in the NYT corpus from -2.1% to -5.9%. This could be explained by the drop of presence of Southern people in SRED^{FM} from 18.9% (see Table 11.4) to 16.9% in the balanced version. An intersectional approach (Crenshaw, 2017) that jointly considers these sources of underrepresentation could be explored to better understand how to mitigate biases from multiple angles.

11.7   Conclusion

In this paper we address the critical matter of biases within RE data and systems, and propose a four-step procedure to analyze them. Our approach showcases the widespread nature of biases in the life-cycle of RE systems, encompassing datasets, transfer learning and model predictions. Our findings reveal a concerning underrepresentation of women and individuals from the Global South as well as undesired biases for specific relation types. We demonstrate that tackling bias is a complex and compounded issue which requires careful thought. Simple techniques, such as balancing the data for an underrepresented group, may introduce other unwanted biases. We also provide a carefully designed taxonomy of relation types that enables comparison and effective transfer across RE datasets.

In conclusion our work serves a dual purpose: On one side, it sheds light on the pervasive biases related to gender and origin within RE datasets and systems, on the other it offers a critical perspective on the use of Information Extraction (IE) techniques for bias exploration. This study emphasizes the need for nuanced, multi-faceted approaches to detect and mitigate biases, urging the community to proceed with caution and depth in developing and applying RE technologies.

Limitations

The first limitation of this work regards the taxonomy adopted for distinguishing people on the basis of their 'place of birth' in the context of a globalized world. We adopt the distinction between Global North and

Global South as it has been recently re-proposed as a framework by the United Nations. However, such a conceptualization has been proposed in a Western context and thus might have an impact on the cultural representation of this underrepresented group. Therefore, we design an operational definition of country belonging to the Global South as being a former colony and having a Human Development Index lower than 0.8.

The second limitation regards the usage of Wikidata for the collection of socio-demographic information about people. The underrepresentation of women and people from the Global South in this knowledge base is a known issue that may have an impact in the analysis. People from the Global South correspond to 85% of the world population, while in Wikidata they represent only the 17.2%. Women are 24.1% in Wikidata, against 49.7% in the real world. Unfortunately, at the time of writing there are no alternative open resources with the same coverage of Wikidata.

A final limitation of our work regards gender. Since we rely on Wikidata to augment corpora with socio-demographic information, we must adopt their P21 property that squeezes biological sex, gender identity, and sexual orientation in a single label. Additionally, the representation of people who do not identify as men or women is statistically irrelevant in our RE corpora. Therefore, we were not able to adopt a non-binary perspective on this aspect.

11.8 Appendix

11.8.1 Relation Type Mapping

In Table 11.8 we report the mapping that we apply from the original labels of SRED^{FM}, CrossRE, NYT to our proposed unified taxonomy of relation types.

11.8.2 Resources

The datasets and models utilized in this paper are governed by the following licenses:

	SRED^FM			CrossRE	NYT
contributes	cast member author producer creator librettist architect	notable work screenwriter composer lyrics by designed by film editor	director performer discoverer or inventor after a work by executive producer voice actor	artifact origin	
date	date of birth work period (end)	date of death time period	work period (start)		
field	occupation field of work	sport instrument	field of this occupation sports discipline competed in		/people/person/profession
geographical relation	place of death country league allegiance	place of birth work location educated at place of burial	country of citizenship country for sport residence indigenous to	physical	/people/person/nationality /people/deceased_person/place_of_death /people/person/place_of_birth /people/ethnicity/geographic_distribution /people/person/place_lived
language	native language	writing language	languages spoken, written or signed		
member	part of member of movement record label	genre crew member(s) ethnic group religious order	member of sports team religion or worldview military branch	part-of general-affiliation	/people/person/religion /people/person/ethnicity /people/ethnicity/people /sports/sports_team_location/teams
participated	participant winner significant event	award received candidate conflict	successful candidate nominated for		
position held	position held chairperson head of state owned by employer	founded by military rank director / manager commanded by	position played on team / speciality office held by head of the organization member of political party head of government	role	/business/company_shareholder/major_shareholder_of /business/person/company /business/company/advisors /business/company/major_shareholders /business/company/founders
relationship	spouse parent relative unmarried partner	sibling family influenced by	child partner in business or sport student	social	/people/person/children
topic	characters	depicts	main subject	topic	

Table 11.8: **Taxonomy Mapping.** Mapping of the original relation types from SRED^FM, CrossRE, NYT into the taxonomy of relation types of Table 11.1.

- SRED^{FM} Dataset: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license.

- CrossRE Dataset: GNU General Public License v3.0.

- NYT Dataset: Linguistic Data Consortium (LDC) Data Use Agreement.

- REBEL Model: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license.

- EntQA Model: MIT License.

11.8.3   Hardware

We train every model on a single NVIDIA® RTX 3090 graphic card with 24GB of VRAM. We use the default hyperparameters used in the original paper for REBEL with Adam (Kingma and Ba, 2015) as optimizer.

213

Part V

Conclusion

Chapter 12

Discussion and Conclusion

Automatic extraction of semantic triplets from text remains a very challenging task, especially when considering data coming from different domains. In this thesis, we advance the field of cross-domain RE by highlighting current issues and challenges (Chapters 1 and 5) and by contributing into three experimental areas: First, in **Part II** we address data scarcity by introducing CrossRE, a multi-domain dataset for RE including six text domains (Chapter 6). Furthermore, we extend CrossRE to 26 new languages other then English (Chapter 7). Then, in **Part III** we propose two methodologies to boost the performance in cross-domain RE and multi-domain RC respectively (Chapters 8 and 9). Last, in **Part IV**, we present two frameworks for analyzing the RE pipeline in terms of model performance and socio-demographic biases (Chapters 10 and 11). In what follows we will answer the research questions defined in Section 1.2.

> **RQ1 To what extent can humans identify domains, and how much do humans agree on this task?**

In Chapter 4 we consider two connotations of the term "domain" as *genre* and *topic*. We report an accuracy of human annotation of 67.68% in detecting genre from a pool of 11 options at the sentence level, and 81.11% when the given context includes five consecutive sentences (i.e., prose). This indicates that genre is identifiable by humans to a fairly high

degree. The inter-annotator agreement varies from a Fleiss' Kappa of 0.53 to 0.66 when annotating genre at sentence and prose level respectively. While it reaches a Fleiss' Kappa of 0.52 when annotating the topic (both at sentence and prose level), which indicates a moderate agreement. Note that within this project, because the gold labels for topic did not exist, we could not compute the accuracy of human annotation.

RQ2 Which challenges emerge by surveying and analyzing the landscape of existing RE datasets?

In Chapter 5 we survey the existing RE datasets and identify two main challenges. First, there are no unified annotation standards, resulting in substantial misalignment across RE datasets. Second, despite the presence of multiple datasets, the domain coverage is limited, causing an issue of representativity because of the many domain-specific nuances of RE.

RQ3 What are important considerations to make when developing a unified annotation scheme for RE that covers multiple domains?

In Chapter 6 we present CrossRE, a novel dataset for RE including six diverse text domains (AI, literature, music, science, politics, news) annotated with a unified annotation schema. It is necessary to make a compromise between the diversity of the domains and the domain-specificity of the relation types. In CrossRE, we introduce a set of labels which is of fairly high level (e.g., *part-of*, *role*), but potentially able to cover any domain. The advantage of starting from a high-level perspective is that following a top-down approach it is possible to expand the annotation guidelines at a later time in order to include domain-specific relations specifying the high-level corresponding type, if required by the current application (e.g., *song-in-album*, *manager*). However, this means that the coarse-grained labels often cover slightly different meanings depending on the domain, which can be a challenge for RE models.

RQ4 When considering languages other than English, how does training and evaluating on automatically translated data influence the performance and the evaluation of RE?

In Chapter 7 we introduce Multi-CrossRE, an automatically translated version of CrossRE to 26 languages beyond English. In order to asses the quality of the translation and its influence on the performance of RE, we back-translate all the 26 new versions of CrossRE to English and compare the performance with the original English data. Our experiments show that for RE training and evaluating on automatically translated data does not influence much the performance. Specifically, the delta of the performance between a model trained on the back-translated data and a model trained on the original data, both evaluated on the original English data, is < 1.5 Macro-F1 in 85% of the cases (22 languages out of 26); and is ≤ 4.6 Macro-F1 for the remaining 15% (Hungarian, Polish, Chinese, Japanese). The delta between the performance of a model trained on back-translated data and evaluated on back-translated data versus evaluating the same model on the original data is ≤ 0.6 Macro-F1 for 25 languages out of 26 (and 2.8 for the one outlier, Japanese).

RQ5 Can we exploit the affinity between semantic RE and syntactic parsing in order to obtain large amounts of (low-cost) silver syntactic data for pre-training RE models to improve the performance?

In Chapter 8 we exploit the syntactic connections frequently linking two entities in a RE triplet via the shortest dependency path for intermediate training of our RE model. We obtain large amounts of (low-cost) silver syntactic connections by running an out-of-the-box syntactic parser on raw data. Our experiments show this strategy to be effective, but only with a moderate improvement of 0.71 Macro-F1 on average, in five out of six domains of evaluation.

RQ6 How can we encode domain information in a multi-domain training setup, and how does it affect performance?

In Chapter 9 we test different methodologies for encoding domain information in a multi-domain training setup. We obtain the best performance by concatenating a special domain marker at the beginning of each instance, with an improvement of 2.19 Macro-F1 over the baseline. From an analysis of the internal model representation of the relations we find that the relation types which benefit the most are the ones whose meaning shifts the most across domains (e.g., *part-of*, *related-to*).

> **RQ7 Is it possible to automatically identify groups of hard-to-handle cases for a SOTA RC model in order to increase the performance of cross-domain RC?**

In Chapter 10 we provide a tool for comprehensive quantitative analysis of the performance of RC models. We analyze the effect of 11 attributes characterizing RC instances (e.g., *entity distance, sentence length*) on the performance of the model. The findings of the analysis reveal the *entity distance* (in terms of number of words separating two entities) to be one of the most influential property for the model in identifying the correct relation type between them—the further away, the more challenging. We target this factor and design a new SOTA architecture where the two entities are repeated at the end of the sentence, close to each other. This simple, but targeted approach improves over the baseline by > 3 Micro-F1 on average across six domains of evaluation.

> **RQ8 To what extent is the RE pipeline (data and models) biased with respect to people's gender and origin?**

In Chapter 11 we propose a framework for the analysis socio-demographic biases in the RE pipeline. We apply our methodology for the analysis of allocative and representational biases related to people's gender and origin on three RE datasets (Huguet Cabot et al., 2023; Bassignana and Plank, 2022a; Riedel et al., 2010) and on the output of a RE model (Huguet Cabot and Navigli, 2021). From the analysis of allocative biases in the training data, we find that women and people from the Global South (which accounts for around 80% of the world population) are underrepresented,

always falling in a range between 10% and 20% of the total. From the analysis of allocative biases in predictions, we find that the model overpredicts men (against women) by 3.5% and 2.2% (in two out of three datasets), and people from the Global North (against Global South) between 8.3% and 1.7% (over the three datasets). Last, from the analysis of representational biases, we identify that in the training sets some relation types are mostly associated with people from one of the two extremes of the considered dimensions (gender and origin). For example the relation type *geographical* which connects a person with a place, is mostly associated with people from the Global South. The representational biases in prediction mostly follow the trends in the training sets, but the model has its own influence in the creation of new biases (likely coming from the data used for pre-training the base language model). For example, the relation type *field* which links a person to an occupation or to a discipline, does not present statistically significant differences in the training sets, but it is mostly associated with people from the Global North in the predictions.

12.1   Future Directions

We conclude with outlining some directions of future work for cross-domain RE. First of all, we acknowledge the role large language models (LLMs) and generative AI are going to have in this field. As discussed earlier in Section 2.2.3, a big change is that the latest trends in generative AI are shifting towards comprehensive frameworks able to address multiple IE tasks (NER, RE, Event Extraction, etc.). However, these tasks are very complex and still far from being solved even with these new technologies. We underline the importance of data quality, at least for evaluation, also within this new era of the NLP field. To that we add the importance of data diversity both for training and evaluation in order to ensure model robustness. Within the widespread of NLP technologies in many different domains of application, and their integration in decision-making processes, robustness is a fundamental property of NLP systems. Last, in connection to the challenges identified in Chapter 1 and further discussed in Chapters 2, 3

and 5 regarding the lack of standards in naming, experimental setups and unified annotation scheme in RE, we stress the importance of having unified standards across the field, and their potential impact in contributing to more solid and faster progress.

Bibliography

Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.692. URL https://aclanthology.org/2020.acl-main.692. pages 35, 38, 48

Sohail Akhtar, Valerio Basile, and Viviana Patti. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *ArXiv preprint*, abs/2106.15896, 2021. URL https://arxiv.org/abs/2106.15896. pages 54

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.142. URL https://aclanthology.org/2020.acl-main.142. pages 82, 175

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444, 2016. doi: 10.1162/tacl_a_00109. URL https://aclanthology.org/Q16-1031. pages 159

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting

keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2091. URL https://aclanthology.org/S17-2091. pages 77, 87

Mehmet Aydar, Özge Bozal, and Furkan Özbay. Neural relation extraction: a review. *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(2):1029–1043, 2021. pages 75

Nguyen Bach and Sameer Badaskar. A review of relation extraction. 2007. URL https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf. pages 15, 75

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL https://aclanthology.org/P19-1279. pages 19, 20, 21, 97, 104, 114, 119, 132, 142, 147, 162, 173, 177

Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, Mike Zhang, and Rob van der Goot. Can humans identify domains? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2745–2765, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.245. pages

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop*

on Benchmarking: Past, Present and Future, pages 15–21, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bppf-1.3. URL https://aclanthology.org/2021.bppf-1.3. pages 112

Elisa Bassignana and Barbara Plank. CrossRE: A Cross-Domain Dataset for Relation Extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.263. URL https://aclanthology.org/2022.findings-emnlp.263. pages 27, 38, 128, 129, 130, 132, 134, 143, 145, 146, 147, 149, 152, 153, 154, 159, 161, 162, 163, 167, 173, 176, 177, 179, 182, 183, 184, 186, 188, 189, 196, 201, 220

Elisa Bassignana and Barbara Plank. What Do You Mean by Relation Extraction? A Survey on Datasets and Study on Scientific Relation Classification. In Samuel Louvan, Andrea Madotto, and Brielen Madureira, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.7. URL https://aclanthology.org/2022.acl-srw.7. pages 29, 102, 128, 199

Elisa Bassignana*, Max Müller-Eberstein*, Mike Zhang*, and Barbara Plank. Evidence > Intuition: Transferability Estimation for Encoder Selection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4218–4227, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.283. URL https://aclanthology.org/2022.emnlp-main.283. pages

Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot, and Barbara Plank. Multi-CrossRE A Multi-Lingual Multi-Domain Dataset for Re-

lation Extraction. In Tanel Alumäe and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 80–85, Tórshavn, Faroe Islands, May 2023a. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.9. pages

Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot, and Barbara Plank. Silver Syntax Pre-training for Cross-Domain Relation Extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6984–6993, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.436. URL https://aclanthology.org/2023.findings-acl.436. pages

Elisa Bassignana, Viggo Unmack Gascou, Frida Nøhr Laustsen, Gustav Kristensen, Marie Haahr Petersen, Rob van der Goot, and Barbara Plank. How to encode domain information in relation classification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8301–8306, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.728. pages

Elisa Bassignana, Rob van der Goot, and Barbara Plank. What's wrong with your model? a quantitative analysis of relation classification. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 252–263, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.20. URL https://aclanthology.org/2024.starsem-1.20. pages

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45, 2018a. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.07.

032. URL https://www.sciencedirect.com/science/article/pii/
S095741741830455X. pages 83

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder.
Adversarial training for multi-context joint entity and relation extraction.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural
Language Processing*, pages 2830–2836, Brussels, Belgium, October-
November 2018b. Association for Computational Linguistics. doi:
10.18653/v1/D18-1307. URL https://aclanthology.org/D18-1307.
pages 83

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language
model for scientific text. In *Proceedings of the 2019 Conference on
Empirical Methods in Natural Language Processing and the 9th Interna-
tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,
pages 3615–3620, Hong Kong, China, November 2019. Association
for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL
https://aclanthology.org/D19-1371. pages 88, 96, 98

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando
Pereira. Analysis of representations for domain adaptation. In
B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in
Neural Information Processing Systems*, volume 19. MIT Press,
2006. URL https://proceedings.neurips.cc/paper_files/paper/
2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf. pages
29

Emily M. Bender and Batya Friedman. Data statements for natural language
processing: Toward mitigating system bias and enabling better science.
Transactions of the Association for Computational Linguistics, 6:587–604,
2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/
Q18-1041. pages 105, 121, 197

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sak-
aguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih,

and Yejin Choi. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB. pages 39

Abhyuday Bhartiya, Kartikeya Badola, and Mausam . DiS-ReX: A multilingual dataset for distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 849–863, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022. acl-short.95. URL https://aclanthology.org/2022.acl-short.95. pages 129

Douglas Biber. *Variation across speech and writing*. Cambridge University Press, Cambridge, 1988. ISBN 0521320712. pages 28, 35

Douglas Biber and Susan Conrad. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, 1 edition, 2009. ISBN 9780521677899. pages 35

Douglas Biber and Susan Conrad. *Register, genre, and style*. Cambridge University Press, 2019. pages 35

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 10–18, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1402. pages 143

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-1615. pages 29

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1056. pages 29

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485. pages 195, 197

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://aclanthology.org/Q17-1010. pages 18, 88, 96

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008. pages 76

C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL https://books.google.de/books?id=3CY-HQAACAAJ. pages 96, 191

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, 2015.

Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075. pages 39

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. pages 22

Razvan Bunescu and Raymond Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1091. pages 143

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075. doi: 10.1126/science. aal4230. pages 195, 198

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. Frustratingly easy label projection for cross-lingual transfer, 2022. URL https://arxiv.org/abs/2211.15613. pages 131

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online, August 2021. Association

for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.249.
URL https://aclanthology.org/2021.findings-acl.249. pages 79

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. A walk-based
model on entity graphs for relation extraction. In *Proceedings of the
56th Annual Meeting of the Association for Computational Linguistics
(Volume 2: Short Papers)*, pages 81–88, Melbourne, Australia, July 2018.
Association for Computational Linguistics. doi: 10.18653/v1/P18-2014.
URL https://aclanthology.org/P18-2014. pages 82

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William
Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma,
Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun
Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robin-
son, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent
Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi,
Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and
Jason Wei. Scaling instruction-finetuned language models, 2022. pages
22

Alexis Conneau and Guillaume Lample. Cross-lingual language model
pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-
Buc, E. Fox, and R. Garnett, editors, *Advances in Neural In-
formation Processing Systems*, volume 32. Curran Associates, Inc.,
2019. URL https://proceedings.neurips.cc/paper_files/paper/
2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf. pages
159

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel
Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating
cross-lingual sentence representations. In *Proceedings of the 2018 Con-
ference on Empirical Methods in Natural Language Processing*, pages
2475–2485, Brussels, Belgium, October-November 2018. Association
for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL
https://aclanthology.org/D18-1269. pages 130

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`. pages 50, 132, 153

Kimberlé W Crenshaw. *On intersectionality: Essential writings*. The New Press, 2017. URL `https://scholarship.law.columbia.edu/books/255/`. pages 211

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. Refining sample embeddings with relation prototypes to enhance continual relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 232–243, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.20. URL `https://aclanthology.org/2021.acl-long.20`. pages 17, 82

Aliva Das, Xinya Du, Barry Wang, Kejian Shi, Jiayuan Gu, Thomas Porter, and Claire Cardie. Automatic error analysis for document-level information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3960–3975, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.274. URL `https://aclanthology.org/2022.acl-long.274`. pages 173, 175

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://aclanthology.org/P07-1033`. pages 29

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, pages 92–110, 2022. URL https://transacl.org/index.php/tacl/article/view/3173. pages 54

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287572. URL https://doi.org/10.1145/3287560.3287572. pages 208

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer, 2018. pages 191

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372. pages 48

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. pages 18, 23, 50, 87, 88, 96, 114, 132, 147

Melvil Dewey. Dewey decimal classification and relative index. 1979. pages 34, 36, 39, 40

Arif Dirlik. Global south: Predicament and promise. *The Global South*, 1: 12–23, 01 2007. doi: 10.1353/gbs.2007.0009. pages 204

Kalpit Dixit and Yaser Al-Onaizan. Span-level model for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5308–5314, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1525. URL https://aclanthology.org/P19-1525. pages 17, 82

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf. pages 76, 77, 104, 109, 128, 176

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266. pages 89, 96, 191

Zhichao Duan, Xiuxing Li, Zhenyu Li, Zhuo Wang, and Jianyong Wang. Not just plain text! fuel document-level relation extraction with explicit syntax refinement and subsentence modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1941–1951, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.140. pages 173

Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing semantic label propagation in relation classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5503. URL https://aclanthology.org/W18-5503. pages 39

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014. pages 37

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1544. pages 129

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 1971. pages 43

Mara Franzen. Alternatives to the dewey decimal system, 2022. URL https://bookriot.com/alternatives-to-the-dewey-decimal-system/. pages 55

Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.489. URL https://aclanthology.org/2020.emnlp-main.489. pages 172, 174, 177

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. RethinkCWS: Is Chinese word segmentation a solved task? In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5676–5686, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.457. URL https://aclanthology.org/2020.emnlp-main.457. pages 174

Jinlan Fu, Liangjing Feng, Qi Zhang, Xuanjing Huang, and Pengfei Liu. Larger-context tagging: When and why does it work? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1463–1475, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.115. URL https://aclanthology.org/2021.naacl-main.115. pages 175

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–429, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-2072. pages 30, 87, 176

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1136. URL https://aclanthology.org/P19-1136. pages 82

Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 12 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl616. URL https://doi.org/10.1093/bioinformatics/btl616. pages 143

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang Qasem-iZadeh, Haïfa Zargayouna, and Thierry Charnois. SemEval-2018 task

7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1111. URL https://aclanthology.org/S18-1111. pages 26, 77, 78, 84, 87, 109

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1649. URL https://aclanthology.org/D19-1649. pages 31, 77, 78, 79, 81, 103, 104, 114, 128, 132, 142, 176

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL https://aclanthology.org/P17-1017. pages 175

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.265. URL https://aclanthology.org/2020.acl-main.265. pages 195, 207, 210

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in lan-

guage models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301. pages 197

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-2008. pages 39

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1003. pages 77

Anna Gooding-Call. Racism in the dewey decimal system, 2021. URL https://bookriot.com/racism-in-the-dewey-decimal-system/. pages 54

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1267. URL https://aclanthology.org/P19-1267. pages 75

Matthew R. Gormley, Mo Yu, and Mark Dredze. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784, Lisbon, Portugal, September 2015. Associa-

tion for Computational Linguistics. doi: 10.18653/v1/D15-1205. URL
https://aclanthology.org/D15-1205. pages 80

Ralph Grishman. Information extraction: Capabilities and challenges, 2012.
pages 74

Jia Guo, Stanley Kok, and Lidong Bing. Towards integration of discrim-
inability and robustness for document-level relation extraction. In
*Proceedings of the 17th Conference of the European Chapter of the As-
sociation for Computational Linguistics*, pages 2606–2617, Dubrovnik,
Croatia, May 2023. Association for Computational Linguistics. URL
https://aclanthology.org/2023.eacl-main.191. pages 173

Qiushi Guo, Xin Wang, and Dehong Gao. Dependency position encod-
ing for relation extraction. In *Findings of the Association for Compu-
tational Linguistics: NAACL 2022*, pages 1601–1606, Seattle, United
States, July 2022. Association for Computational Linguistics. doi:
10.18653/v1/2022.findings-naacl.120. URL https://aclanthology.
org/2022.findings-naacl.120. pages 173

Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional
networks for relation extraction. In *Proceedings of the 57th Annual
Meeting of the Association for Computational Linguistics*, pages 241–251,
Florence, Italy, July 2019. Association for Computational Linguistics. doi:
10.18653/v1/P19-1024. URL https://aclanthology.org/P19-1024.
pages 82

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. Table filling multi-
task recurrent neural network for joint entity and relation extraction.
In *Proceedings of COLING 2016, the 26th International Conference on
Computational Linguistics: Technical Papers*, pages 2537–2547, Osaka,
Japan, December 2016. The COLING 2016 Organizing Committee. URL
https://aclanthology.org/C16-1239. pages 17

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo,
Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretrain-

ing: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740. pages 29, 30, 38, 119

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *CoRR*, abs/2305.14450, 2023. doi: 10.48550/ARXIV.2305.14450. URL https://doi.org/10.48550/arXiv.2305.14450. pages 22

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1514. URL https://aclanthology.org/D18-1514. pages 31, 77, 78, 79, 103, 104, 114, 128, 132, 142, 176

David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. Why only micro-f1? class weighting of measures for relation classification. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 32–41, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlppower-1.4. URL https://aclanthology.org/2022.nlppower-1.4. pages 115

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/

`2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf`. pages 195, 207

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. pages 50

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://aclanthology.org/S10-1006`. pages 77, 81, 109

Martin Hilbert and Priscila López. The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011. doi: 10.1126/science.1200970. URL `https://www.science.org/doi/abs/10.1126/science.1200970`. pages 1

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. doi: https://doi.org/10.1111/lnc3.12432. URL `https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432`. pages 7, 194, 199

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.repl4nlp-1.30. URL `https://aclanthology.org/2021.repl4nlp-1.30`. pages 82

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. Three sentences are all you need: Local path enhanced

document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.126. URL https://aclanthology.org/2021.acl-short.126. pages 82, 88

Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204. URL https://aclanthology.org/2021.findings-emnlp.204. pages 17, 21, 22, 31, 196, 202, 220

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. RED^{FM}: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.237. pages 196, 201, 220

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1135. URL https://aclanthology.org/P19-1135. pages 76

Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480, Osaka,

Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1139. pages 107

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potami-anos. UDALM: Unsupervised domain adaptation through language mod-eling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-guage Technologies*, pages 2579–2590, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.203. URL https://aclanthology.org/2021.naacl-main.203. pages 35

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.284. URL https://aclanthology.org/2021.eacl-main.284. pages 77, 78, 129, 131

Arzoo Katiyar and Claire Cardie. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1087. URL https://aclanthology.org/P16-1087. pages 175

Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 917–928, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1085. URL https://aclanthology.org/P17-1085. pages 175

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Va-cano. Constructing interval variables via faceted rasch measurement and

multitask deep learning: a hate speech application. *ArXiv preprint*, abs/2009.10277, 2020. URL https://arxiv.org/abs/2009.10277. pages 48

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain, July 1997. Association for Computational Linguistics. doi: 10.3115/976909.979622. URL https://aclanthology.org/P97-1005. pages 35, 38

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.semeval-1.313. pages 48

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181. pages 19

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*, 2015. pages 213

Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL https://aclanthology.org/S18-2005. pages 197

Richard Kittredge and Ralph Grisham. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, 1986. pages 34, 37

Richard Kittredge and John Lehrberger. *Sublanguage: Studies of language in restricted semantic domains*. Walter de Gruyter, 1982. pages 37

Abdullatif Köksal and Arzucan Özgür. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.32. URL https://aclanthology.org/2020.findings-emnlp.32. pages 129

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. Language-agnostic bias detection in language models with bias probing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.848. URL https://aclanthology.org/2023.findings-emnlp.848. pages 198

Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas. In layman's terms: Semi-open relation extraction from scientific texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1500, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.137. URL https://aclanthology.org/2020.acl-main.137. pages 82

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational

Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823. pages 7, 194, 198

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.539. pages 198

David Lee. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. In *Teaching and Learning by Doing Corpus Analysis*, pages 245–292. Brill, 2002. pages 37

David YW Lee. Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72, 2001. pages 35

Juhyuk Lee, Min-Joong Lee, June Yong Yang, and Eunho Yang. Does it really generalize well on unseen data? systematic evaluation of relational triple extraction methods. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3858, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.282. URL https://aclanthology.org/2022.naacl-main.282. pages 175

Yong-Bae Lee and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150, 2002. pages 35, 37

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.

BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703. pages 21, 202

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633, 2023. URL https://api.semanticscholar.org/CorpusID:258297899. pages 22

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1129. URL https://aclanthology.org/P19-1129. pages 82

Yankai Lin, Zhiyuan Liu, and Maosong Sun. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1004. URL https://aclanthology.org/P17-1004. pages 82

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ArXiv*, 2023. pages 35

Tom Lippincott, Diarmuid Ó Séaghdha, Lin Sun, and Anna Korhonen. Exploring variation across biomedical subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 689–697, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1078. pages 75

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2002. pages 134

Kang Liu. A survey on neural relation extraction. *Science China Technological Sciences*, 63(10):1971–1989, 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1673-6. URL https://doi.org/10.1007/s11431-020-1673-6. pages 75

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.34. URL https://aclanthology.org/2021.acl-demo.34. pages 174

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. A dependency-based neural network for relation classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 285–290, Beijing, China, July 2015. Association for Computational Linguistics. doi:

10.3115/v1/P15-2047. URL https://aclanthology.org/P15-2047.
pages 143

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. A simple yet
effective relation information guided approach for few-shot relation
extraction. In *Findings of the Association for Computational Linguistics:
ACL 2022*, pages 757–763, Dublin, Ireland, May 2022. Association for
Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.62. URL
https://aclanthology.org/2022.findings-acl.62. pages 173

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan
Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denois-
ing pre-training for neural machine translation. *Transactions of the Associ-
ation for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_
a_00343. URL https://aclanthology.org/2020.tacl-1.47. pages
159, 160

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyaw-
ijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-
domain named entity recognition. *Proceedings of the AAAI Confer-
ence on Artificial Intelligence*, 35(15):13452–13460, May 2021b. URL
https://ojs.aaai.org/index.php/AAAI/article/view/17587. pages
29, 38, 75, 103, 106, 120, 121, 122, 123, 124, 130, 146, 161, 167, 176,
178, 182, 188, 190

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-
task identification of entities, relations, and coreference for scien-
tific knowledge graph construction. In *Proceedings of the 2018 Con-
ference on Empirical Methods in Natural Language Processing*, pages
3219–3232, Brussels, Belgium, October-November 2018. Association
for Computational Linguistics. doi: 10.18653/v1/D18-1360. URL
https://aclanthology.org/D18-1360. pages 26, 77, 78, 84, 109, 143,
148, 149

Yuan Luo, Özlem Uzuner, and Peter Szolovits. Bridging semantics and

syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, 18(1):160–178, 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw001. URL https://doi.org/10.1093/bib/bbw001. pages 74

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. SENT: Sentence-level distant relation extraction via negative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6201–6213, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.484. URL https://aclanthology.org/2021.acl-long.484. pages 82

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015. pages 38

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.67. URL https://aclanthology.org/2021.acl-short.67. pages 82

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis,

Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063. pages 198

David McClosky. *Any domain parsing: automatic domain adaptation for natural language parsing*. PhD thesis, Brown University, 2010. pages 35

Shiao Meng, Xuming Hu, Aiwei Liu, Shuang Li, Fukun Ma, Yawen Yang, and Lijie Wen. RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5208–5226, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.316. URL https://aclanthology.org/2023.emnlp-main.316. pages 23

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113. pages 76

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596. pages 197

Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1105. URL https://aclanthology.org/P16-1105. pages 83, 104, 175

Makoto Miwa and Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1200. URL https://aclanthology.org/D14-1200. pages 17

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.393. URL https://aclanthology.org/2021.emnlp-main.393. pages 38

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. How universal is genre in Universal Dependencies? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 69–85, Sofia, Bulgaria, December 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.tlt-1.7. pages 35, 48

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416. pages 197

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. Reasoning with

latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.141. URL https://aclanthology.org/2020.acl-main.141. pages 82

Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séagdha. Semantic relations between nominals. *Synthesis Lectures on Human Language Technologies*, 14(1):1–234, 2021. pages 77

Minh Luan Nguyen, Ivor W. Tsang, Kian Ming A. Chai, and Hai Leong Chieu. Robust domain adaptation for relation extraction via clustering consistency. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–817, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1076. URL https://aclanthology.org/P14-1076. pages 30

Thien Huu Nguyen and Ralph Grishman. Employing word representations and regularization for domain adaptation of relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 68–74, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2012. URL https://aclanthology.org/P14-2012. pages 30, 104

Thien Huu Nguyen and Ralph Grishman. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China, July 2015a. Association for Computational Linguistics. doi: 10.3115/v1/P15-2060. URL https://aclanthology.org/P15-2060. pages 104

Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspec-

tive from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado, June 2015b. Association for Computational Linguistics. doi: 10.3115/v1/W15-1506. URL https://aclanthology.org/W15-1506. pages 19, 20, 87, 88, 93, 96, 104

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1378–1387, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1143. pages 144

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734. pages 39

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1262. pages 27

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe

G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Davide Rovati, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena

255

Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data, 2017. URL http://hdl.handle.net/11234/1-2184. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. pages 40

Abiola Obamuyide and Andreas Vlachos. Meta-learning improves lifelong relation extraction. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 224–229, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4326. URL https://aclanthology.org/W19-4326. pages 82

Mary Ellen Okurowski. Information extraction overview. In *TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993*, pages 117–121, Fredericksburg, Virginia, USA, September 1993. Association for Computational Linguistics. doi: 10.3115/1119149.1119164. URL https://aclanthology.org/X93-1012. pages 1

OpenAI. Introducing chatgpt. *OpenAI Blog*, 2023. URL https://openai.com/blog/chatgpt. pages 22

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. Guideline learning for in-context information extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389, Singapore, December 2023. Association for Computational Linguistics. doi: 10.

18653/v1/2023.emnlp-main.950. URL https://aclanthology.org/
2023.emnlp-main.950. pages 22

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro
Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and
Stefano Soatto. Structured prediction as translation between augmented
natural languages. In *International Conference on Learning Representa-*
tions, 2021. URL https://openreview.net/forum?id=US-TP-xnXI.
pages 21

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation
extraction: A survey. *arXiv preprint arXiv:1712.05191*, 2017. URL
https://arxiv.org/abs/1712.05191. pages 75

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-
tau Yih. Cross-sentence n-ary relation extraction with graph LSTMs.
Transactions of the Association for Computational Linguistics, 5:101–115,
2017. doi: 10.1162/tacl_a_00049. URL https://aclanthology.org/
Q17-1008. pages 15

Yifan Peng, Samir Gupta, Cathy Wu, and Vijay Shanker. An extended
dependency graph for relation extraction in biomedical texts. In *Proceed-*
ings of BioNLP 15, pages 21–30, Beijing, China, July 2015. Association
for Computational Linguistics. doi: 10.18653/v1/W15-3803. URL
https://aclanthology.org/W15-3803. pages 144

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher
Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word rep-
resentations. In *Proceedings of the 2018 Conference of the North American*
Chapter of the Association for Computational Linguistics: Human Language
Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans,
Louisiana, June 2018. Association for Computational Linguistics. doi:
10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.
pages 87

Philipp Petrenz and Bonnie Webber. Squibs: Stable classification of text genres. *Computational Linguistics*, 37(2):385–393, June 2011. doi: 10.1162/COLI_a_00052. URL https://aclanthology.org/J11-2004. pages 48

Slav Petrov and Ryan McDonald. Overview of the 2012 shared task on parsing the web. In *First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), NAACL-HLT*, 2012. pages 75

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf. pages 39

Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018. pages 30, 143

Van-Thuy Phi, Joan Santoso, Masashi Shimbo, and Yuji Matsumoto. Ranking-based automatic seed selection and noise reduction for weakly supervised relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 89–95, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2015. URL https://aclanthology.org/P18-2015. pages 82

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.semeval-1.317. pages 48

Barbara Plank. *Domain adaptation for parsing*. PhD thesis, University of Groningen, 2011. URL https://research.rug.nl/en/publications/domain-adaptation-for-parsing. pages 35, 37

Barbara Plank. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS 2016, Ruhr-University Bochum*. 2016. pages 4, 28, 35, 37

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.731. pages 39, 54

Barbara Plank and Alessandro Moschitti. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1498–1507, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1147. pages 30, 79, 103, 104, 118

Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2083. URL https://aclanthology.org/P14-2083. pages 39, 112

Rhitabrat Pokharel and Ameeta Agrawal. Estimating semantic similarity between in-domain and out-of-domain samples. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 409–416, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.35. URL https://aclanthology.org/2023.starsem-1.35. pages 38

Nicholas Popovic and Michael Färber. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5733–5746, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main. 421. URL https://aclanthology.org/2022.naacl-main.421. pages 128, 142, 173, 176

Amir Pouran Ben Veyseh, Thien Nguyen, and Dejing Dou. Improving cross-domain performance for relation extraction via dependency prediction and information flow control. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5153–5159. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/716. URL https://doi.org/10.24963/ijcai.2019/716. pages 176

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. Exploiting the syntax-model consistency for neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8021–8032, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020. acl-main.715. URL https://aclanthology.org/2020.acl-main.715. pages 82

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3516. pages 75

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.", 2012. pages 109

S. Pyysalo, R. SÃ¦tre, J. Tsujii, and T. Salakoski. Why biomedical relation extraction results are incomparable and what to do about it. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 149–152, 2008. URL http://mars.cs.utu.fi/smbm2008/files/smbm2008proceedings/smbmpaper_33.pdf. pages 75

Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 697–704, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1088. pages 144

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html. pages 21

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. Domain divergences: A survey and empirical analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1830–1849, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.147. URL https://aclanthology.org/2021.naacl-main.147. pages 34

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.603. URL https://aclanthology.org/2020.coling-main.603. pages 30, 34, 38

Adwait Ratnaparkhi. Learning to parse natural language with maximum entropy models. *Machine learning*, 34:151–175, 1999. pages 37

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410. pages 57

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442. pages 172, 174

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15939-8. pages 76, 77, 175, 196, 201, 220

Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2401. pages 26, 76, 77

Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

372–382, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1038. URL https://aclanthology.org/D17-1038. pages 38

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706, 2021. doi: 10.1162/tacl_a_00392. URL https://aclanthology.org/2021.tacl-1.42. pages 31, 77, 78, 103, 106, 128

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2647–2661, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.228. URL https://aclanthology.org/2021.eacl-main.228. pages 144

Evan Sandhaus. The new york times annotated corpus, 2008. pages 38

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.178. pages 54

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.505. pages 194

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163. pages 194

Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424, 2022. pages 34

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.166. URL https://aclanthology.org/2021.eacl-main.166. pages 77, 78, 80, 88, 128

Satoshi Sekine. The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington, DC, USA, 1997. Association for Computational Linguistics. doi: 10.3115/974557.974572. URL https://aclanthology.org/A97-1015. pages 37

Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli. Relation extraction with explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6488–6494, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.579. URL https://aclanthology.org/2020.acl-main.579. pages 82

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA, 2008. Association for Computing Ma-

chinery. ISBN 9781605581934. doi: 10.1145/1401890.1401965. URL
https://doi.org/10.1145/1401890.1401965. pages 88

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei,
Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis,
Stephen Pfohl, et al. Large language models encode clinical knowledge.
Nature, pages 1–9, 2023. pages 35

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan
Shao, and Sara Stymne. 82 treebanks, 34 models: Universal Depen-
dency parsing with multi-treebank models. In *Proceedings of the CoNLL
2018 Shared Task: Multilingual Parsing from Raw Text to Universal De-
pendencies*, pages 113–123, Brussels, Belgium, October 2018. Associa-
tion for Computational Linguistics. doi: 10.18653/v1/K18-2011. URL
https://aclanthology.org/K18-2011. pages 159

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We
need to talk about random splits. In *Proceedings of the 16th Conference
of the European Chapter of the Association for Computational Linguistics:
Main Volume*, pages 1823–1832, Online, April 2021. Association for
Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.156. URL
https://aclanthology.org/2021.eacl-main.156. pages 75

Benno Stein and Sven Meyer Zu Eissen. Distinguishing topic from genre. In
*Proceedings of the 6th International Conference on Knowledge Management
(I-KNOW 06)*, pages 449–456. Citeseer, 2006. pages 35, 37

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti,
Daniele Radicioni, and Tommaso Caselli. WikiBio: a semantic re-
source for the intersectional analysis of biographical events. In *Pro-
ceedings of the 61st Annual Meeting of the Association for Computa-
tional Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto,
Canada, July 2023. Association for Computational Linguistics. URL
https://aclanthology.org/2023.acl-long.691. pages 195

Marco Antonio Stranisci*, Pere-Lluís Huguet Cabot*, Elisa Bassignana*, and Roberto Navigli. Dissecting Biases in Relation Extraction: A Cross-Dataset Analysis on People's Gender and Origin. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. pages

Sara Stymne. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.tlt-1.6. URL https://aclanthology.org/2020.tlt-1.6. pages 160

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2098. URL https://aclanthology.org/P18-2098. pages 159

Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.45. URL https://aclanthology.org/2021.acl-short.45. pages 195

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159. pages 197

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416. 3483305. URL https://doi.org/10.1145/3465416.3483305. pages 195

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.301. URL https://aclanthology.org/2020.emnlp-main.301. pages 74, 81, 83

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1364. URL https://aclanthology.org/D19-1364. pages 38

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.132. URL https://aclanthology.org/2022.findings-acl.132. pages 173

Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances*

in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf. pages 7, 194

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.60. URL https://aclanthology.org/2021.acl-long.60. pages 82

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. UniRel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.477. pages 173

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. Dependency-driven relation extraction with attentive graph convolutional networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4458–4471, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.344. URL https://aclanthology.org/2021.acl-long.344. pages 82

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at*

HLT-NAACL 2003, pages 142–147, 2003. URL https://aclanthology.org/W03-0419. pages 106, 121, 122, 123, 161, 162

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1023. URL https://aclanthology.org/P19-1023. pages 82

Dennis Ulmer. deep-significance: Easy and Better Significance Testing for Deep Neural Networks, 2021. URL https://doi.org/10.5281/zenodo.4638709. https://github.com/Kaleidophon/deep-significance. pages 96

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental Standards for Deep Learning in Natural Language Processing Research. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.196. URL https://aclanthology.org/2022.findings-emnlp.196. pages

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*, 2022b. pages 191

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *The Journal of Artificial Intelligence Research*, Forthcoming, 2021a. ISSN 1076-9757. pages 88

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey.

Journal of Artificial Intelligence Research, 72:1385–1470, 2021b. pages 39, 48

Rob van der Goot and Miryam de Lhoneux. Parsing with pretrained language models, multiple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria, December 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.tlt-1.9. pages 160

Rob van der Goot, Ahmet Üstün, and Barbara Plank. On the effectiveness of dataset embeddings in mono-lingual,multi-lingual and zero-shot conditions. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 183–194, Kyiv, Ukraine, April 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.adaptnlp-1.19. pages 38, 160

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online, April 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.22. URL https://aclanthology.org/2021.eacl-demos.22. pages 50, 88, 96, 98, 147

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. What's in a domain? analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 560–566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2092. URL https://aclanthology.org/P15-2092. pages 35, 37, 38

Denny Vrandečić and Markus Krötzsch.   Wikidata: A free col-
laborative knowledge base.   *Communications of the ACM*, 57:
78–85, 2014.   URL http://cacm.acm.org/magazines/2014/10/
178785-wikidata/fulltext. pages 198, 202

Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation
extraction in the era of large language models. In *Proceedings of
the 61st Annual Meeting of the Association for Computational Linguis-
tics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada,
July 2023. Association for Computational Linguistics.   URL https:
//aclanthology.org/2023.acl-long.868. pages 22, 31

Joachim Wagner, James Barry, and Jennifer Foster. Treebank embedding
vectors for out-of-domain dependency parsing. In *Proceedings of the 58th
Annual Meeting of the Association for Computational Linguistics*, pages
8812–8818, Online, July 2020. Association for Computational Linguistics.
doi: 10.18653/v1/2020.acl-main.778.   URL https://aclanthology.
org/2020.acl-main.778. pages 160

Zhen Wan, Fei Cheng, Qianying Liu, Zhuoyuan Mao, Haiyue Song, and
Sadao Kurohashi. Relation extraction with weighted contrastive pre-
training on distant supervision.   In *Findings of the Association for
Computational Linguistics: EACL 2023*, pages 2580–2585, Dubrovnik,
Croatia, May 2023. Association for Computational Linguistics.   URL
https://aclanthology.org/2023.findings-eacl.195. pages 173

Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su,
Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document
relation extraction. In *Proceedings of the 2022 Conference on Empirical
Methods in Natural Language Processing*, pages 9871–9881, Abu Dhabi,
United Arab Emirates, December 2022a. Association for Computational
Linguistics.   URL https://aclanthology.org/2022.emnlp-main.671.
pages 173

Jue Wang and Wei Lu. Two are better than one: Joint entity and relation

extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.133. URL https://aclanthology.org/2020.emnlp-main.133. pages 83, 104

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.188. URL https://aclanthology.org/2022.findings-naacl.188. pages 173

Sida Wang and Christopher Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-2018. pages 38

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. UniRE: A unified label space for entity relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long. 19. URL https://aclanthology.org/2021.acl-long.19. pages 82, 83

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, and Bryan Hooi. GraphCache: Message passing as caching for sentence-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1698–1708, Seattle, United States, July 2022c. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.128. URL https://aclanthology.org/2022.findings-naacl.128. pages 173

Katy Weathington and Jed R Brubaker. Queer identities, normative databases: Challenges to capturing queerness on wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–26, 2023. URL https://cmci.colorado.edu/idlab/assets/bibliography/pdf/Weathington2023-Wikidata.pdf. pages 195

Bonnie Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1076. pages 35, 38, 48

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. Zero-shot information extraction via chatting with chatgpt, 2023. pages 22, 31

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101. pages 39

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101. pages 75

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebas-

tian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *ArXiv*, 2023. pages 35

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. Revisiting the negative data of distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3572–3581, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.277. URL https://aclanthology.org/2021.acl-long.277. pages 82

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. Large language models for generative information extraction: A survey. *ArXiv*, abs/2312.17617, 2023. URL https://api.semanticscholar.org/CorpusID:266690657. pages 31

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.523. URL https://aclanthology.org/2020.emnlp-main.523. pages 50

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.124. URL https://aclanthology.org/2021.acl-short.124. pages 82

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL https://aclanthology.org/P19-1074. pages 77, 79, 82, 87, 88, 176

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.337. URL https://aclanthology.org/2022.acl-long.337. pages 173

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1351–1360, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1130. URL https://aclanthology.org/P19-1130. pages 80, 82

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020. acl-main.444. URL https://aclanthology.org/2020.acl-main.444. pages 82

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir Radev. SParC: Cross-domain semantic parsing in context. In *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, pages 4511–4523, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1443. URL https://aclanthology.org/P19-1443. pages 75

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.bionlp-1.7. pages 22

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021. 102563. URL https://www.sciencedirect.com/science/article/pii/S0306457321000662. pages 77, 79

Amir Zeldes. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017. pages 34, 35, 39

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1220. pages 87

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:

10.18653/v1/P18-1047. URL https://aclanthology.org/P18-1047.
pages 202

Charles Chuankai Zhang and Loren Terveen. Quantifying the gap: A
case study of wikidata gender disparities. In *Proceedings of the 17th
International Symposium on Open Collaboration*, OpenSym '21, New
York, NY, USA, 2021. Association for Computing Machinery. ISBN
9781450385008. doi: 10.1145/3479986.3479992. URL https://doi.
org/10.1145/3479986.3479992. pages 204, 205

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Min Zijun, Qing-
guo Hu, and Xiaodong Shi. Towards better document-level relation
extraction via iterative inference. In *Proceedings of the 2022 Confer-
ence on Empirical Methods in Natural Language Processing*, pages 8306–
8317, Abu Dhabi, United Arab Emirates, December 2022a. Association
for Computational Linguistics. URL https://aclanthology.org/2022.
emnlp-main.568. pages 173

Meishan Zhang, Yue Zhang, and Guohong Fu. End-to-end neural relation
extraction with global optimization. In *Proceedings of the 2017 Conference
on Empirical Methods in Natural Language Processing*, pages 1730–1740,
Copenhagen, Denmark, September 2017a. Association for Computational
Linguistics. doi: 10.18653/v1/D17-1182. URL https://aclanthology.
org/D17-1182. pages 83

Min Zhang, Jie Zhang, and Jian Su. Exploring syntactic features for relation
extraction using a convolution tree kernel. In *Proceedings of the Human
Language Technology Conference of the NAACL, Main Conference*, pages
288–295, New York City, USA, June 2006. Association for Computational
Linguistics. URL https://aclanthology.org/N06-1037. pages 144

Peiyuan Zhang and Wei Lu. Better few-shot relation extraction with label
prompt dropout. In *Proceedings of the 2022 Conference on Empirical
Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi,
United Arab Emirates, December 2022. Association for Computational

Linguistics. URL https://aclanthology.org/2022.emnlp-main.471.
pages 173

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. EntQA: Entity linking as question answering. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=US2rTP5nm_.
pages 202

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html. pages 39

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004.
pages 31, 77, 78, 142, 175

Yunqi Zhang, Yubo Chen, and Yongfeng Huang. RelU-net: Syntax-aware graph U-net for relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4217, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.282. pages 173

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323. pages 194

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003. pages 197

Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie, and Dongsheng Li. Improving few-shot relation classification by prototypical representation learning with definition text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 454–464, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.34. URL https://aclanthology.org/2022.findings-naacl.34. pages 173

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5. pages 20, 83, 104, 114, 119, 132, 160, 173

Wenxuan Zhou and Muhao Chen. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only, November 2022. Association

for Computational Linguistics. URL https://aclanthology.org/2022.aacl-short.21. pages 173

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1128. URL https://aclanthology.org/P19-1128. pages 82

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295, 2020a. doi: 10.1162/tacl_a_00314. URL https://aclanthology.org/2020.tacl-1.19. pages 75

Wanrong Zhu, Xin Wang, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Towards understanding sample variance in visually grounded language generation: Evaluations and observations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8806–8811, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.708. URL https://aclanthology.org/2020.emnlp-main.708. pages 76