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

Elisa Bassignana♣ and **Barbara Plank**♣◇
♣Department of Computer Science, IT University of Copenhagen, Denmark
◇Center for Information and Language Processing (CIS), LMU Munich, Germany
{elba, bapl}@itu.dk

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 sub-domains 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.

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 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
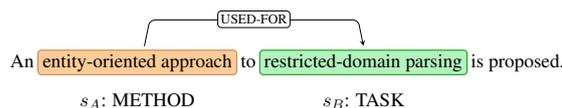
Figure 1: RE annotation sample. The sentence contains two annotated spans denoting two entities, with respective types METHOD and TASK, and a semantic relation between them labeled as USED-FOR.

the task definition.[1] Existing RE surveys mainly 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., 2018; Yu et al., 2019; Zhu et al., 2020a; Liu et al., 2021), 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;

---

[1] We refer the reader to Luo et al. (2016) for a survey on biomedical RE and event extraction.

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

and (*c*) sub-domains impact Relation Classification (RC)—the task of classifying the relation type held between a pair of entities (details in Section 3).

The contributions of this paper are:

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

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 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 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 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 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 pretrained 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

---

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

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

Dataset	Paper	Data Source	# Relation 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◇
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 literature	100 + 25
DocRED	Yao et al. (2019)	Wikipedia and Wikidata	96
SMiLER	Seganti et al. (2021)	Wikipedia	36

Table 1: Overview of the RE datasets for the English language grouped by macro domains. (⋆): Multilingual datasets. (◇): The original paper does not state the number of considered relations and different work describe different dataset setups.

(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 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 TA-CRED to make a FSL RC dataset and compared it to FewRel 1.0 and FewRel 2.0, aiming at a 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.

3 The Relation Extraction Task

Conceptually, RE involves a pipeline of steps (see Figure 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 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). 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
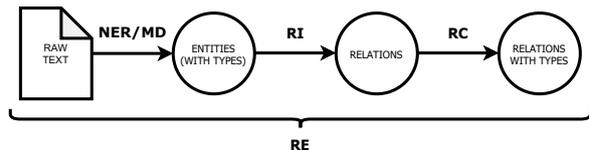
Figure 2: Relation Extraction pipeline. NER: Named Entity Recognition; MD: Mention Detection; RI: Relation Identification; RC: Relation Classification.

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 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.

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., 2018a,b; 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 pro-

---

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

[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 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.

posed by Bekoulis et al. (2018a). 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 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 2 —specially regarding RI—leads to problematic evaluation comparisons; *2)* A lack of cross-domain studies, for all three setups outlined in Section 2.

4 Scientific Domain Data Analysis

In this section, we present the two English corpora involved in the experimental study (Section 4.1), explain the label mapping adopted for the cross-dataset experiments (Section 4.2), discuss the overlap between the datasets and the annotation divergence between them (Section 4.3), and introduce the sub-domains considered (Section 4.4).

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 9 in Appendix A. To the best of our knowledge, we are the first to analyze the corpus at this sub-domain level.

4.2 Cross-dataset Label Mapping

We homogenize the relation label sets via a manual analysis performed after an exploratory data analy-

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

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 4: SemEval-2018 and SCIERC annotation comparison. The common relations are the ones with a direct correspondent in both datasets (see Table 3).

sis, 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 3. The gold label distribution of the relations on the two datasets is shown in Figure 4 in Appendix B. We 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.

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):

- **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 annotated 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 annotations differ.* This involves conflicting annotations, e.g., the bold arrow shows the same entity pair annotated with a different relation label.

Table 4 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 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 (Sample 3).

4.4 Experimental Sub-domains

We use the metadata described in Section 4.1 to divide SCIERC into four sub-domains. Figure 5 in Appendix B 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. The considered sub-domains and their relative amount of data are reported in Table 6.

5 Experiments

5.1 Model Setup

Since the seminal work by Nguyen and Grishman (2015), Convolutional Neural Networks (CNNs) are widely used for IE tasks (Zeng et al., 2014; Nguyen and Grishman, 2015; Fu et al., 2017; Augenstein et al., 2017; Gábor et al., 2018; Yao et al.,

---

[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: Different number of entity (and relation) annotations

SemEval-2018	We evaluate the utility of this <u>constraint</u> in two different algorithms.
SCIERC	We evaluate the utility of this <u>constraint</u> in two different <u>algorithms</u> .

EVALUATE-FOR

Sample 2: Different entity annotations

SemEval-2018	We propose a <u>detection method</u> for orthographic variants caused by <u>transliteration</u> in a large <u>corpus</u> .
SCIERC	We propose a <u>detection method</u> for <u>orthographic variants</u> caused by <u>transliteration</u> in a large corpus.

PART_WHOLE USED-FOR

Sample 3: Different relation annotations

SemEval-2018	The <u>speech-search algorithm</u> is implemented on a <u>board</u> with a single <u>Intel i860 chip</u> , which provides a factor of 5 speed-up over a <u>SUN 4</u> for straight <u>C code</u> .
SCIERC	The <u>speech-search algorithm</u> is implemented on a <u>board</u> with a single <u>Intel i860 chip</u> , which provides a factor of 5 speed-up over a <u>SUN 4</u> for <u>straight C code</u> .

MODEL-FEATURE USED-FOR PART-OF COMPARE USED-FOR

Table 5: Annotated sentence pairs from SemEval-2018 and SCIERC. The underlined spans are the entities.

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 6: Sub-domains and relative amount of abstracts.

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., 2021)

Our CNN follows Nguyen and Grishman (2015). 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 C. We use macro F1-score as evaluation metric. All experiments were run over three different seeds and the results reported are the mean.[8]

5.2 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., 2021), (2.3) *filter*: Double annotation of the abstracts as in (2.2), but filtering out conflicting annotations.

Results Table 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. 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 → 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 D.

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 7 report the results. While BERT does not bring relevant improvements over the best fastText setup, SciBERT confirms the strength of domain-

---

[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 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).

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 E), 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).

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 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 B) across sub-domains (confusion from lowest to highest: AI-ML, CV and SPEECH), and it is most confused with MODEL-FEATURE. Figure 7 in Appendix F 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 (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 3. The plot confirms

Figure 3: PCA representation of the CNN hidden state (just before the linear layer) using SciBERT.

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 6 in Appendix G reports the confusion matrix) and the large vocabulary overlap shown in Table 8 between these sub-

---

[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 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 6.

domains confirms this observation. Overall, subdomains are identifiable but have limited impact on the RC task in the setup considered.

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 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 un-supervised approaches. Finally, given that our study points out gaps in RE, specifically crossdataset, 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).

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. 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. Association for Computational Linguistics.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. 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. Association for Computational Linguistics.

Mehmet Aydar, Özge Bozal, and Furkan Özbay. 2021. Neural relation extraction: a review. *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(2):1029–1043.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. 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. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. 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. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

75

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciB-ERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. 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.

C.E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. 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. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. 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. Association for Computational Linguistics.

Li Cui, Deqing Yang, Jiaxin Yu, Chengwei Hu, Jiayang Cheng, Jingjie Yi, and Yanghua Xiao. 2021. 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Kalpit Dixit and Yaser Al-Onaizan. 2019. Span-level model for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5308–5314, Florence, Italy. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. 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. European Language Resources Association (ELRA).

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. 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. Association for Computational Linguistics.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. 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. Asian Federation of Natural Language Processing.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. 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. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. 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. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. 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. Association for Computational Linguistics.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. 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. Association for Computational Linguistics.

Kyle Gorman and Steven Bedrick. 2019. 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. Association for Computational Linguistics.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. 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. Association for Computational Linguistics.

Ralph Grishman. 2012. Information extraction: Capabilities and challenges.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. 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. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. 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. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. 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. Association for Computational Linguistics.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021a. 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. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021b. 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. Association for Computational Linguistics.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. 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. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. 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. Association for Computational Linguistics.

Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas. 2020. 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. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. 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. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. 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. Association for Computational Linguistics.

Tom Lippincott, Diarmuid Ó Séaghdha, Lin Sun, and Anna Korhonen. 2010. Exploring variation across biomedical subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 689–697, Beijing, China. Coling 2010 Organizing Committee.

Kang Liu. 2020. A survey on neural relation extraction. *Science China Technological Sciences*, 63(10):1971–1989.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yuan Luo, Özlem Uzuner, and Peter Szolovits. 2016. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting

biomedical relations. *Briefings in Bioinformatics*, 18(1):160–178.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. 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. Association for Computational Linguistics.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. 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. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. 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. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. 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. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. 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. Association for Computational Linguistics.

Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séagdha. 2021. Semantic relations between nominals. *Synthesis Lectures on Human Language Technologies*, 14(1):1–234.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2019. Meta-learning improves lifelong relation extraction. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 224–229, Florence, Italy. Association for Computational Linguistics.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), NAACL-HLT*.

Van-Thuy Phi, Joan Santoso, Masashi Shimbo, and Yuji Matsumoto. 2018. 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. Association for Computational Linguistics.

Barbara Plank and Alessandro Moschitti. 2013. 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. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. 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. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

S. Pyysalo, R. SÃ¦tre, J. Tsujii, and T. Salakoski. 2008. 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.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dan Roth and Wen-tau Yih. 2004. 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. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. 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. Association for Computational Linguistics.

Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli. 2020. Relation extraction with explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6488–6494, Online. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. 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. Association for Computing Machinery.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. 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. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. 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. Association for Computational Linguistics.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. 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. Association for Computational Linguistics.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. 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. Association for Computational Linguistics.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. 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. Association for Computational Linguistics.

Dennis Ulmer. 2021. deep-significance: Easy and Better Significance Testing for Deep Neural Networks. Https://github.com/Kaleidophon/deep-significance.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *The Journal of Artificial Intelligence Research*, Forthcoming.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. 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. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. 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. Association for Computational Linguistics.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. 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. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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. Association for Computational Linguistics.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. 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. Association for Computational Linguistics.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. 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. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. 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. Association for Computational Linguistics.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. 2019. 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. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

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. 2019. 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. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing Management*, 58(4):102563.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. 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. Dublin City University and Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017a. 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. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017b. 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. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. 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. Association for Computational Linguistics.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. 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. Association for Computational Linguistics.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020a. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

Wanrong Zhu, Xin Wang, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2020b. 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. Association for Computational Linguistics.

Appendix

A 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 9.

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

B Data Analysis

Figure 4 reports the gold label distribution over SemEval-2018 and SCIERC respectively.

Figure 5, instead, contains the gold label distributions of SCIERC sub-domains over the five matching labels between the two datasets (see Table 3).

C Model Details

Our RC model is a CNN with four layers (Nguyen and Grishman, 2015). 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

(a) SemEval-2018
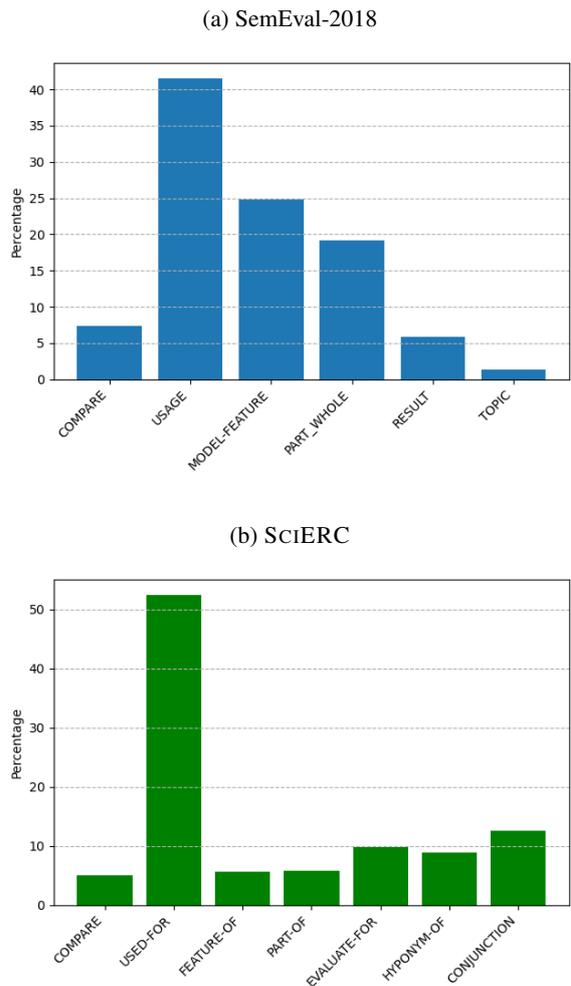
(b) SCIERC

Figure 4: Gold label distribution in the SemEval-2018 sub-task (1.1) and SCIERC datasets.
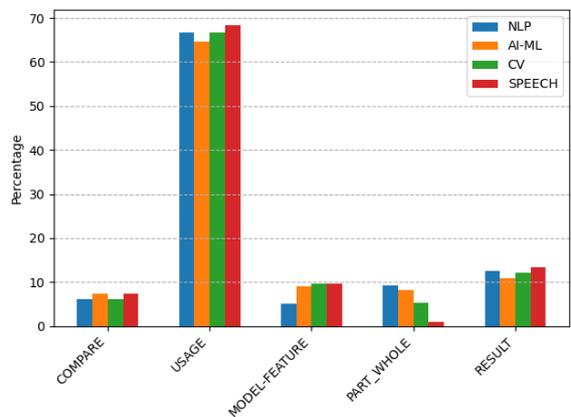
Figure 5: Gold label distribution of the five considered relations over SCIERC sub-domains.

4), a max-pooling layer and a linear softmax relation classification layer with dropout of 0.5. Each input to the network 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 (2015). 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.

D 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 10 reports the ASO scores with a confidence level of $\alpha = 0.05$ adjusted by using the Bonferroni correction (Bonferroni, 1936). See Section 5 for the setup details.

E Transformer setups

The MaChAmp toolkit (van der Goot et al., 2021) 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:

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

---

[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 10: ASO scores of the main experimental setups described in Section 5. (*) CNN model. (†) full fine-tuned transformer model. Read as row → column.

↓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 11: Macro F1-scores of the RC using SciBERT (Beltagy et al., 2019) within the MaChAmp toolkit (van der Goot et al., 2021). Setups 1-5 described in Appendix E.

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*]

Table 11 reports the results of the experiments using MaChAmp on the setups described above.

F Scientific Sub-domain Analysis

Figure 7 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.

G   Conference Classifier

Figure 6 represents the confusion matrix relative to the conference classifier described in Section 5.4.
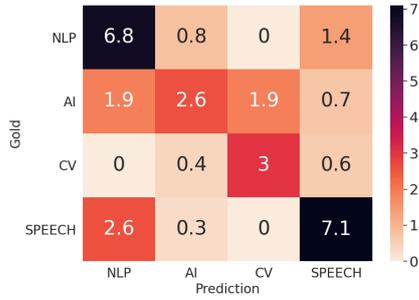
Figure 6: Confusion matrix of the conference classification experiment. The numbers reported are the average over three runs on different seeds.

(a) AI-ML (52 abstracts)
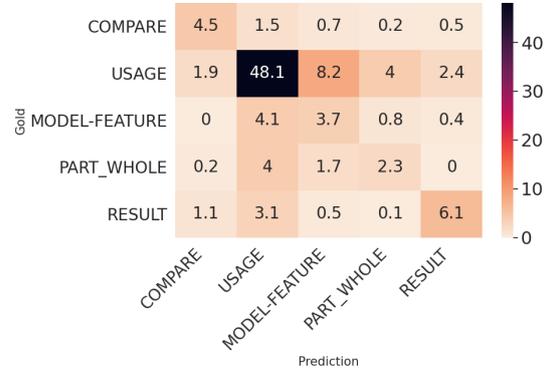
(b) CV (105 abstracts)
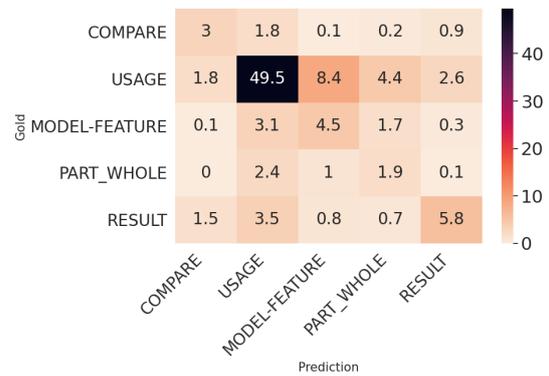
(c) SPEECH (35 abstracts)
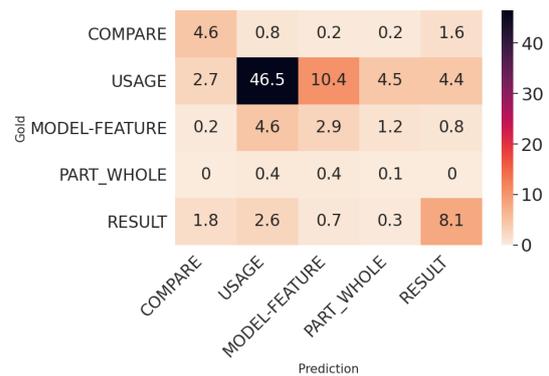
Figure 7: Percentage confusion matrices of the CNN on SCIERC sub-domains.