CrossRE: A Cross-Domain Dataset for Relation Extraction

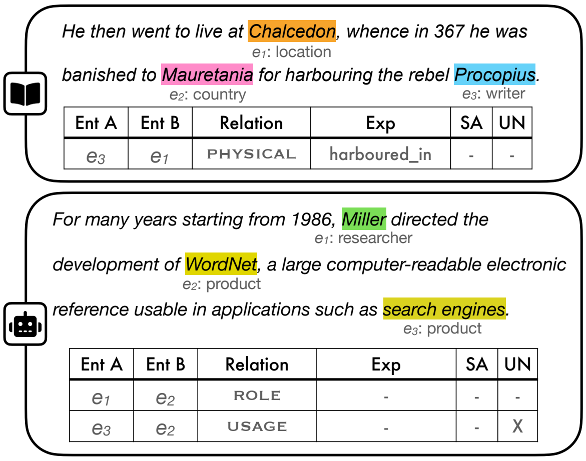
Elisa Bassignana and **Barbara Plank**
Department of Computer Science, IT University of Copenhagen, Denmark
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
elba@itu.dk bplank@cis.lmu.de

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]

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, 2022). 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 even on data for which new relations or entity types have to be recognized.



He then went to live at Chalcedon, whence in 367 he was banished to Mauretania for harbouring the rebel Procopius.

Ent A	Ent B	Relation	Exp	SA	UN
e_3	e_1	PHYSICAL	harboured_in	-	-

For many years starting from 1986, Miller directed the development of WordNet, a large computer-readable electronic reference usable in applications such as search engines.

Ent A	Ent B	Relation	Exp	SA	UN
e_1	e_2	ROLE	-	-	-
e_3	e_2	USAGE	-	-	X

Figure 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.

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 1). CROSSRE

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

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3592–3604
December 7-11, 2022 ©2022 Association for Computational Linguistics

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., 2021), 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.

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.

3 CrossRE

3.1 Motivation

RE aims to extract semantically informative triples from unstructured text. The triples comprehend an ordered pair of text spans which represent 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]

3.2 Dataset Overview

CrossRE includes the following domains: news (📰), politics (🏛), natural science (🍃), music (♫), literature (📖) and artificial intelligence (🤖; AI). Our semantic relations are annotated on top of CrossNER (Liu et al., 2021), 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. (2021) for details on e.g. vocabulary overlap across domains.

During our relation annotation process, we additionally correct some mistakes in named enti-

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

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

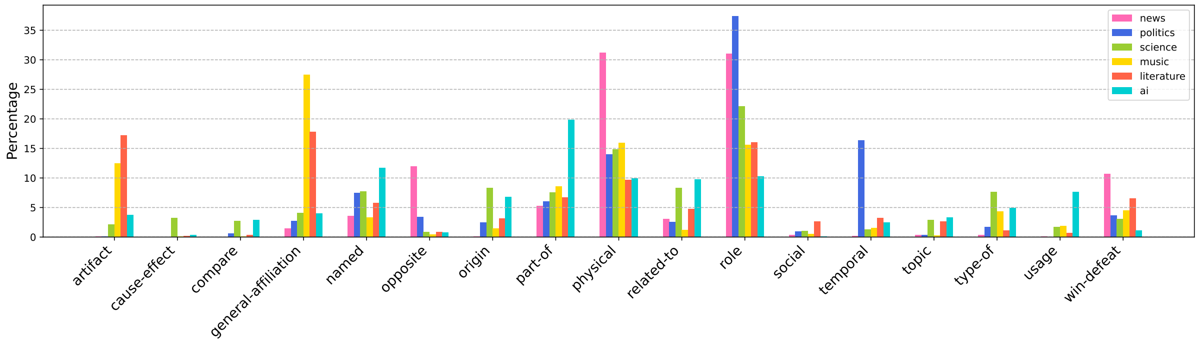


Figure 2: **CROSSRE Label Distribution.** Percentage label distribution over the 17 relation types divided by CROSSRE's six domains. Detailed counts and percentages in Appendix D.

ties 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 C).

The final dataset statistics are reported in Table 1. We keep the train/dev/test data split by Liu et al. (2021) 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, TEMPORAL, 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 B, and the full annotation guidelines in our repository. The resulting label distribution is illustrated in Figure 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.

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

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 E). 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 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.

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 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.

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 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.

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 3.4, the guideline definition has been an iterative process with five annotation rounds and Figures 3 and 4 report the annotation agreement between the linguist and the NLP expert. As the entity order is part of the an-

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 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.

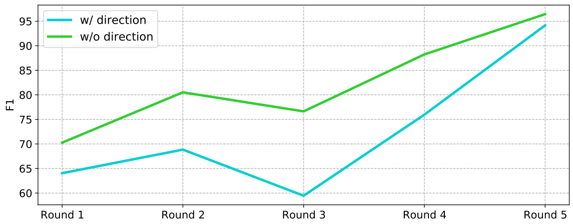


Figure 3: **RI Annotation Agreement.** F1 score of the identified relations during the official annotation rounds.

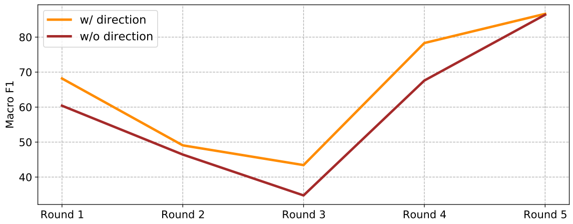


Figure 4: **RC Annotation Agreement.** Macro-F1 score of the assigned labels over the entity pairs identified by both annotators during the official annotation rounds.

notation guidelines, we furthermore tease apart the directionality component for a deeper analysis of the annotation agreement.

In Figure 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 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 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.

3.6 Meta-data Annotation

By embracing the subjectivity of manually-curated datasets, we collect *meta-data* (see data samples in Table 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 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 ‘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 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 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 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.

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.

	📰	🏛	🍃	♫	📖	🤖	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 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.

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.

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*
e_1: person
career with the *Philadelphia Eagles*
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

	📰	🏛	🍃	🎵	📖	🤖	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 4: **CROSSRE Baselines.** Results achieved by our baseline model on the RC task. Reported are the averages over five random seeds (see Table 8).

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 F for hyperparameters settings.

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 10) 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 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 1). This, in addition to the most imbalanced label distribution—predominantly ROLE, PHYSICAL, OPPOSITE, WIN-DEFEAT and PART-OF (see Figure 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.

	📰	🏛	🍃	🎵	📖	🤖
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 5: **Per-class Results.** Detailed F1 scores for each relation type. Reported are the averages over five random seeds (see Table 8). '-' indicates the class is not present in the test set.

Per-label Performance In Table 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 4. Overall, the amount of instances per-label—see Figure 2 for percentages and Table 10 for counts—are good indicators for the individual scores in Table 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 G. We additionally notice that the same label can have different levels of challenge depending on the 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.

[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.

[5] Evaluation code in our repository.

	📰	🏛	🍃	♫	📖	🤖
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: **Test Set Statistics.** Amount of annotations which have been marked with Syntax Ambiguity (SA) and with Uncertainty (Un) in the test sets.

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 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 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.

			🏛	🍃	♫	📖	🤖
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 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 4 (All).

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 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 useful 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 underexplored 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. (2021) 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 ≈ 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 computing resources. Elisa Bassignana and Barbara Plank are supported by the Independent Research Fund Denmark (Danmarks Frie Forskningsfond; DFF) Sapere Aude grant 9063-00077B. Barbara Plank is supported by the ERC Consolidator Grant DIALECT 101043235.

References

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.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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.

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).

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. 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.

David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. 2022. 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. 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.

Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. 2016. 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. The COLING 2016 Organizing Committee.

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.

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.

Thien Huu Nguyen and Ralph Grishman. 2014. 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. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015a. 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. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015b. 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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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. 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.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

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.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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.

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.

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.

A 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., 2021 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., 2021.

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., 2021) are not available.

B 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.

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

- 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.

C Entity Alteration Samples

In Table 9 we report one sample for each entity alteration type that we perform in respect to the original entity annotations from CrossNER (Liu et al., 2021) 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.

D Detailed Label Statistics

Table 10 contains the detailed label statistics (counts and percentages) for each domain.

E Multi-label annotation

In Table 11 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.

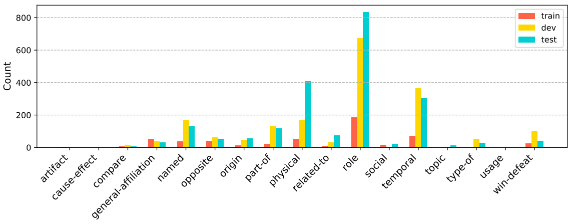


Figure 5: **Label Distribution of the Politics Domain.** Distribution of the 17 relation types over the train/dev/test split.

F Reproducibility

We report in Table 8 the hyperparameter setting of our RC model (see Section 4.2). All experiments were ran on an NVIDIA® A100 SXM4 40 GB GPU and an AMD EPYC™ 7662 64-Core CPU.

G Label Distribution Per-Domain

Given the imbalance of the label distribution (see Figure 2) and of the train/dev/test splits (see Table 1), we report in Figure 5 as a sample the specific label distribution of the politics 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_A: person e_5: person e_6: person

Table 9: **Samples of Modified Entity Annotations.** Instances with the original annotations from CrossNER (Liu et al., 2021) 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 10: **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 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 11: **Example of Multi-label Annotation.** Example from CrossRE of an ordered entity pair which has been annotated with multiple relation labels.