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

Elisa Bassignana[🧭][🤖] **Rob van der Goot**[🧭][🤖] **Barbara Plank**[🧭][▲]

[🧭] IT University of Copenhagen, Denmark

[🤖] Pioneer Center for Artificial Intelligence, Denmark

[▲] Center for Information and Language Processing, LMU Munich, Germany

elba@itu.dk robv@itu.dk b.plank@lmu.de

Abstract

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

1 Introduction

A major trend in NLP research aims at designing more sophisticated setups and model architectures in order to improve the state-of-the-art (SOTA) on a target task. The improvements are usually based on intuitions that target limitations of the previous SOTA on the task. The most common procedure follows the steps of *(1)* intuition, *(2)* modeling, *(3)* experiments, *(4)* results, and *(5)* analysis of the results. The latter is occasionally enriched with ablation or case studies with the main aim of proving the validity of the initial intuition and the effectiveness of the proposed methodology. We claim that conducting a preliminary in-depth analysis can help find good intuitions, and therefore guide better modeling and reducing the probability of negative experiments, usually not reported in the paper. Following previous error-guided analysis (Ribeiro et al., 2020; Fu et al., 2020a; Das et al., 2022), we argue in favor of changing the standard chain of actions listed above: *First* perform an exhaustive quantitative analysis of the previous SOTA to identify failure cases and challenging scenarios, and *then* effectively target the baseline improvement in order to tackle those.

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

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 252 - 263
June 20-21, 2024 ©2024 Association for Computational Linguistics

et al., 2023). We find that that the common procedure consists of the five steps earlier mentioned. Specifically, we found that in most cases, the intuition (step 1) that is used as a starting point and as a motivation for the model improvement is based on generic observations of the model architecture, instead of on a quantitative analysis which could lead to more effective targeted improvements.

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

- We provide a tool for comprehensive quantitative analyses of RC model performance.
- We exploit the proposed analysis for investigating the performance of a SOTA RC architecture across 36 in- and cross-domain setups.
- Based on the findings of the analysis, we perform a case study improving the previous SOTA by over 3 points Micro-F1.

2 Related Work

Analysis of NLP Models In this study, we are inspired by the recent trend targeting the evaluation of NLP models. Ribeiro et al. (2020) propose a task-agnostic methodology for testing general linguistic capabilities of NLP models by generating ad-hoc test instances; they test their approach over three tasks: sentiment analysis, Quora question pair, machine comprehension. Liu et al. (2021a) presents a software package for diagnosing the strengths and weaknesses of a single system, allowing for interpretation of relationships between multiple systems, and examining prediction results. They go a bit deeper into the task specificity, therefore their system currently supports the tasks of text classification (sentiment, topic, intention), aspect sentiment classification, Natural Language Inference (NLI), Named Entity Recognition (NER), Part-of-Speech (POS) tagging, chunking, Chinese Word Segmentation (CWS), semantic parsing, summarization, and machine translation. Furthermore, Fu et al. (2020a) and Fu et al. (2020b) introduce the concept of interpretable task-specific evaluation. The first target the comparison of a set of NER systems. The latter, instead, perform a deep evaluation of CWS systems proving that despite the excellent performance achieved on some datasets, there is no perfect system for CWS. This concept has also been applied by Fu et al. (2021) for interpreting the results over a set of sequence tagging setups (NER, CWS, POS, chunking). Within the field of Information Extraction, previous work explored error-driven analysis for the automatic categorization of model prediction errors (Das et al., 2022).

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

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

3 Background

3.1 Cross-domain Relation Classification

Given a sentence and two entity spans within it, the task of RC aims at classifying the semantic relation between them into a type from a pre-defined label

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

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

Table 1: **Relation Classification Attributes.** Description of the 11 RC attributes and categorization in DISCRETE/CONTINUOUS value type, LOCAL/AGGREGATE computation, and ENTITY/RELATION/SENTENCE level. (⋆): We map the original 36 domain-specific entity types defined by Liu et al. (2021b) into five more generic types shared across domains, see Appendix B for details.

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

3.2 Experimental Setup

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

We use the baseline model of the original paper.[3] Following the architecture proposed by Baldini Soares et al. (2019), the model by Bassignana and Plank (2022) augments the sentence with four entity markers e_1^{start}, e_1^{end}, e_2^{start}, e_2^{end} surrounding the two entities. The augmented sentence is then passed through a pre-trained encoder, and the classification made by a linear layer over the concatenation of the start markers $[\hat{s}_{e_1^{start}}, \hat{s}_{e_2^{start}}]$. We run our experiments over five random seeds and report the average performance. See Appendix C for reproducibility details.

4 Attribute Guided Analysis

We propose a systematic quantitative analysis of the performance of the CrossRE baseline model's performance across the 36 in- and cross-domain setups derived from training and testing the model on the six domains included in CrossRE. The analysis is performed over the development sets of the dataset. Inspired by the work of Fu et al. (2020a) on Named Entity Recognition, we introduce the first evaluation suite for RC, opening the way to other similar structured prediction tasks. The analysis evaluates the performance of the model over

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

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

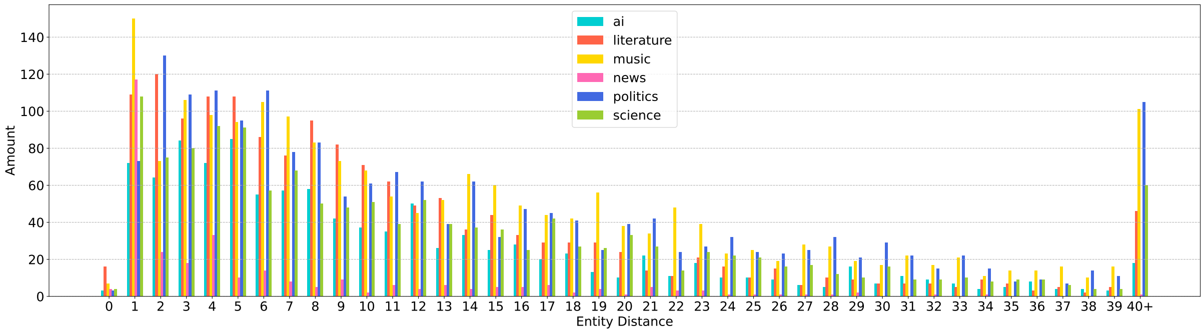


Figure 1: **`entity distance` Distribution.** Distribution of the `entity distance` values across the six development sets of CrossRE (Bassignana and Plank, 2022).

instances grouped by common values of potentially influential attributes (e.g., entity distance, sentence length). In what follows, we will describe the attributes considered and the bucketing strategy employed for splitting the evaluation instances based on the attribute values. Last, we go through the results of our correlation analysis.

4.1 Attributes

In our analysis, we consider 11 different attributes. These are characteristics of the RC instances that could challenge the model and influence its performance. Given an RC instance defined by a triplet *(e1, e2, r)* where *e1* is the head entity, *e2* is the tail entity, and *r* is the relation type connecting *e1* with *e2*; and given a sentence *s* expressing the relation *r* between *e1* and *e2*, we define the attributes listed in Table 1. We categorize each of them in the following three divisions:

- **Value Type:** If the values of the attribute belong to a set of pre-defined values the attribute is DISCRETE (e.g., the entity type), otherwise it is CONTINUOUS (e.g., the entity distance).
- **Computation:** If the attribute is computed by only considering the current instance it is LOCAL, if it is computed over aggregated properties of the train set, it is AGGREGATE; for example, the frequency of entity and relation types refers to the training statistics.
- **Level:** If the attribute value depends on the entities it is at ENTITY LEVEL, if it depends on properties of the entity pair it is at RELATION LEVEL, last if it is related to characteristics of the sentence *s* it is at SENTENCE LEVEL.

As an attribute example, Figure 1 shows the `entity distance` distribution, measured as number of tokens separating *e1* from *e2*. The plot reveals some domain-specific peculiarities, e.g., music and politics have the longest distances. This is mostly due to the long lists present in these domains, where the head entity is mentioned at the beginning and linked to all the elements in the list. For example, a music genre and a list of musical artists representing it; or the artifacts (i.e., songs and albums) of a band. We use the attribute values in order to group the evaluation instances with similar characteristics. We discuss the bucketing strategy in the next section.

4.2 Methodology

Once identified the potential influential attributes for the task of RC, the next step is splitting the evaluation sets depending on the attributes values (i.e., bucketing). For the attributes with DISCRETE value types (see Table 1) the bucketing creates one subset for each attribute values—e.g., one subset for each entity type for the `entity type` attribute. For the attributes with CONTINUOUS value types, instead, we set the number of buckets to four in order to maintain a reasonable size for each bucket. We then split the instances by equally distributing them across subsets—except for the two AGGREGATE attributes, which by definition are computed over properties of the train set. Note that the `entity type` and `entity type frequency` have each instance placed into two buckets, one considering the type of *e1* and one considering the type of *e2*.

We measure the performance of the model over the subsets, and compute the Spearman's rank correlation coefficient with respect to the average attribute values of the buckets. Since `entity type` and `relation type` have categorical values, we cannot compute the correlation coefficient and analyze these two attributes separately in Section 4.3.1.

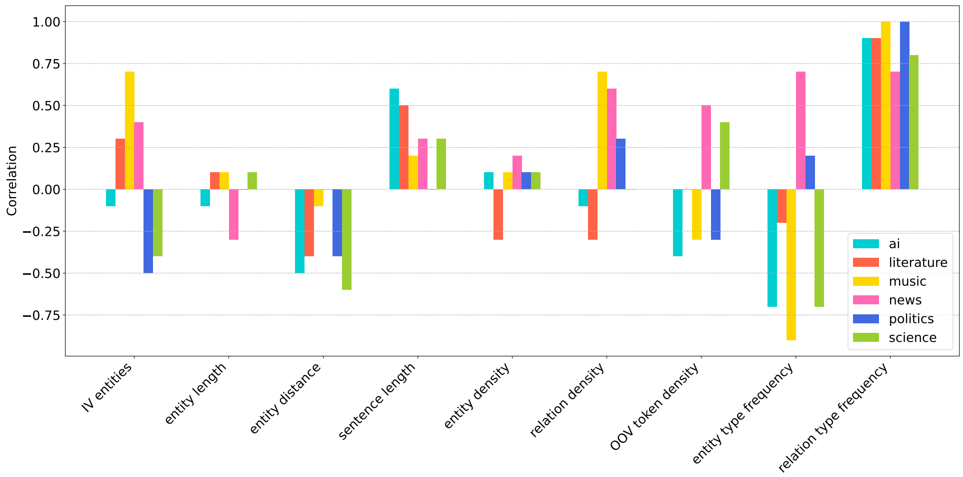


Figure 2: **Per-domain Correlation Analysis.** Spearman's rank correlation coefficient of the the 36 considered setups, averaged over the dev sets.

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

Table 2: **Overall Results.** Average correlation and average standard deviation of the Micro-F1 scores of the buckets (within attribute), averaged over the 36 train-test setups.

4.3 Results

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

Overall Table 2 reports the correlations for the proposed attributes (Section 4.1) averaged across all 36 setups. We also report the average standard deviation across the Micro-F1 scores achieved within attribute and computed separately for each train-test setup. The relation type frequency is by far the most influential attribute: It reports the highest absolute correlation value, and the highest standard deviation between buckets including low- and high-frequent relations types in the train sets. In the current setups with relatively small training sets (see CrossRE statistics in Appendix A) the amount of training instances have an high impact on the final performance of the model. In addition, this is also influenced by the cross-domain setup, with diverse relation label distributions over the six domains (see Figure 3). The second most relevant attribute is entity distance, with the second highest absolute value in correlation and a 6.1 average standard deviation across buckets containing entity pairs at different distances. The entity type frequency presents a weaker correlation, confirming the findings that we will discuss in Section 4.3.1 about the entity type. All the other attributes report an absolute correlation value ranging between 0.2 and 0.0 indicating that within the overall overview of the considered setups they have a lower impact on the model's performance.

Domain Level We visualize the average across the test domains in Figure 2. As previously noted, the relation type frequency trend confirms that the amount of training instances is the most influential attribute within the current setup. The entity distance and sentence length also present a similar trend across all six domains. The negative correlation of the first indicates that, as we could intuitively expect, it is more challenging to identify the semantic relation connecting two entities which are far apart in the sentence, with respect to entity pairs separated by only a couple of tokens. The positive trend within the sentence length attribute, instead, suggests that entity pairs belonging to long sentences (i.e., where more context is given) are easier to classify than the ones from short sentences. The entity density, and relation density attributes present a general positive trend in correlation, but with some outliers (literature and AI). High values in these attributes refer to sentences with many instances, e.g., lists of

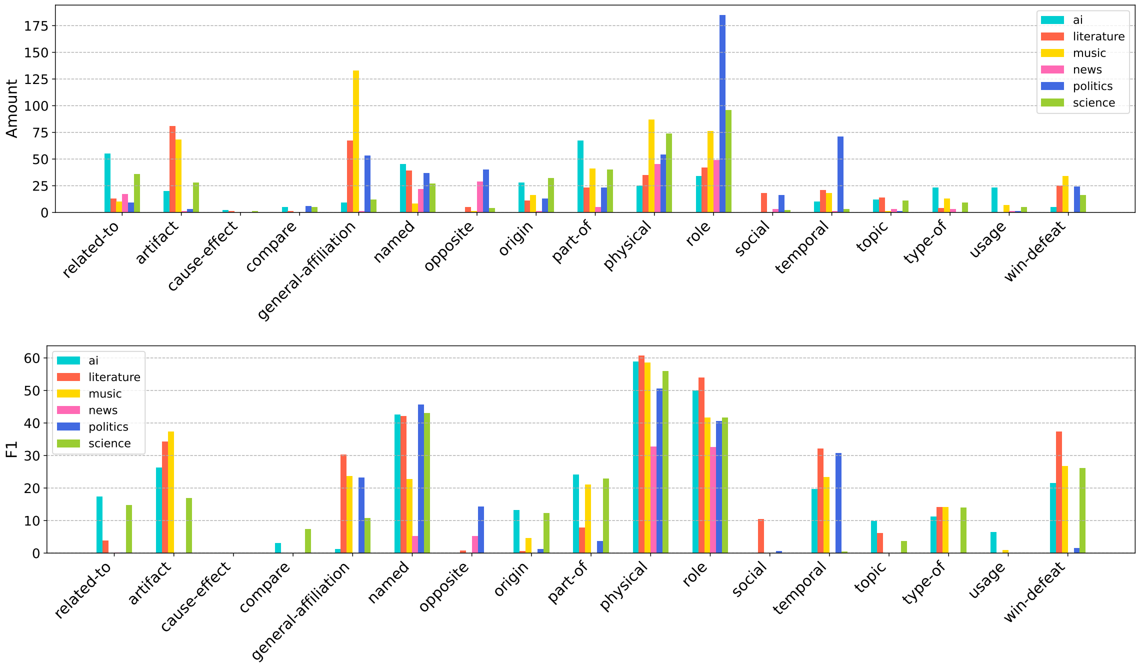


Figure 3: **relation type Analysis.** Distribution of the relation types in the train sets of CrossRE (Bassignana and Plank, 2022) (above), and F1 per label (bottom).

entities which are all linked to an head entity with a similar structure and (most likely) with the same relation type. For example, in the music domain, a list of songs authored by a music artist, or by a band. We speculate these to be easy patterns to identify and learn by a deep learning model.

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

4.3.1 Categorical analysis

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

`relation type` The results in Figure 3 reveal that some of the types are easier to learn across all domains than others (i.e. have higher scores, despite their lower frequency). These can be explained because they occur in very similar linguistic constructions, like "named", which often connects an entity to the consecutive acronym in brackets. Or because they mostly occur with the same entity types, like "temporal" with "event" and "physical" with "location". On the other hand, some relation labels have different performances across domains. For example "win-defeat" which in the domains of AI, literature, music, and science mostly links a person winning an award. In the politics domain, instead, it refers to more complex scenarios where one out of multiple mentioned political parties wins the election. Or, in a completely different semantic context, a country wins a war against another country. Unsurprisingly the most difficult are clearly the infrequent ones, like "cause-effect".

`entity type` The results in Figure 4 show that there is not a strong link between the amount of training instances and the performance achieved, confirming the findings from Figure 2. This is because in the CrossRE guidelines there are no constraints linking the relation types to specific entity types. The higher scoring types are mostly the ones that are implicitly associated with specific relation types, e.g., "location" with the "physical" relation type, and "event" with "temporal". On the other hand, the most varied category "misc" is the most challenging (see entity mapping in Table 5).

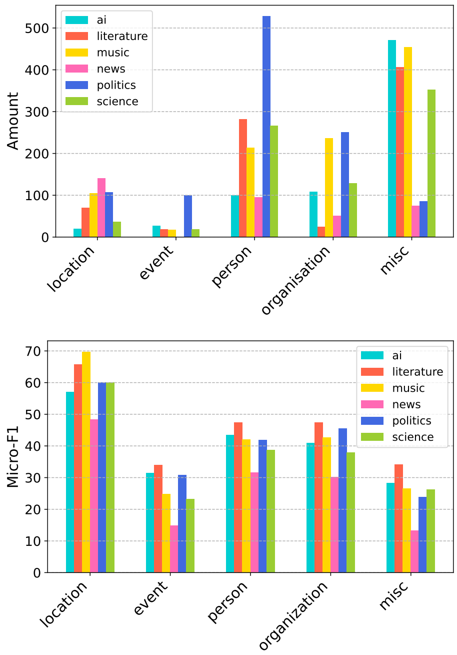


Figure 4: `entity type` **Analysis.** Distribution of the entity types in the train sets of CrossRE (Bassignana and Plank, 2022) (above), and Micro-F1 achieved on each bucket (bottom).

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

Table 3: **Performance Comparison Across Setups.** Micro-F1 scores achieved with the baseline architecture, and with the three proposed variants. (**bold**): Scores beating the baseline; (underline): Highest scores within the four setups.

5 Application: Model Improvement

As mentioned in the introduction, our final aim is to guide better modeling by targeting quantitatively measured weaknesses of the model. Here we present a case study which exploits the findings of our proposed analysis. From the overall results in Table 2 we can derive that the most influential attribute is the `relation type frequency`, with a correlation of 0.9 and the highest standard deviation of 24.9. Targeting this factor would mean obtaining additional training data by manual annotation or via some data augmentation techniques. Within this case study, we aim to focus on improving the model architecture. Therefore, here we target the `entity distance` attribute, which holds the second highest absolute correlation (0.4), for improving the model performance.

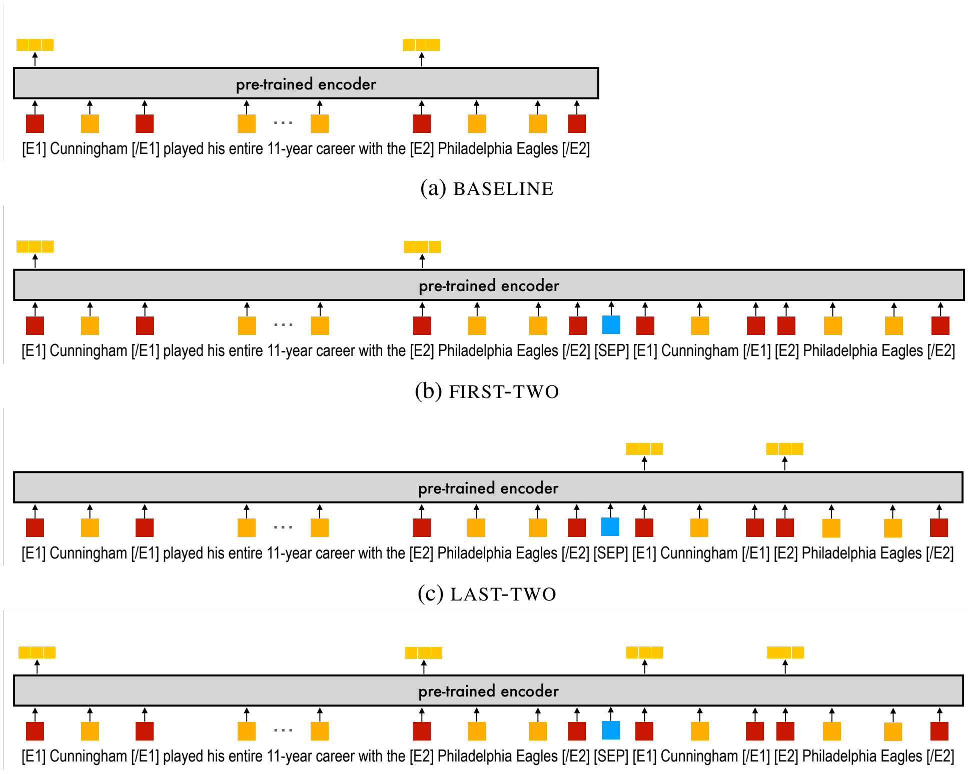
5.1 Improved Experimental Setting

The fact that the `entity distance` (i.e., the number of tokens separating *e1* from *e2*) has a high influence on the RC model performance, means that the tokens between *e1* and *e2* can somehow mislead the prediction. In order to target this issue, we aim at moving the two involved entities closer to each other. We repeat the entities at the end of the original sentence representation augmented with the entity markers. Then, similar to the original CrossRE baseline (Section 3.2), we pass the input through a pre-trained encoder and extract a representation on which we do the classification of the relation with a linear layer. We test out three different representations as illustrated in Figure 5:

- FIRST-TWO concatenation of the representation of the first two entity markers start, as in the original baseline setup;
- LAST-TWO concatenation of the representation of the last two entity markers start, the ones introduced after the [SEP] token;
- ALL-FOUR concatenation of the representation of all four entity markers start, the original ones and the newly introduced.

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

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


pre-trained encoder
[E1] Cunningham [/E1] played his entire 11-year career with the [E2] Philadelphia Eagles [/E2]
pre-trained encoder
[E1] Cunningham [/E1] played his entire 11-year career with the [E2] Philadelphia Eagles [/E2] [SEP] [E1] Cunningham [/E1] [E2] Philadelphia Eagles [/E2]
pre-trained encoder
[E1] Cunningham [/E1] played his entire 11-year career with the [E2] Philadelphia Eagles [/E2] [SEP] [E1] Cunningham [/E1] [E2] Philadelphia Eagles [/E2]
pre-trained encoder
[E1] Cunningham [/E1] played his entire 11-year career with the [E2] Philadelphia Eagles [/E2] [SEP] [E1] Cunningham [/E1] [E2] Philadelphia Eagles [/E2]

(a) BASELINE

(b) FIRST-TWO

(c) LAST-TWO

(d) ALL-FOUR

Figure 5: **Proposed Setups.** Representation of the baseline architecture (a) and of the three proposed setups (b, c, d) which include the repetition of *e1* and *e2* at the end of the sentence.

5.2 New SOTA Results

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

6 Conclusion

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

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

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

Ethics Statement

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

Limitations

In this work we report a case study of our proposed evaluation suite over CrossRE which includes six datasets covering six text domains. We focused mainly on the current SOTA model, future work could consider more models and datasets. The set of attributes is mostly bound to the RC task, but other relation-based tasks could employ similar attributes. More aspects could be included in the analysis in order to inspect specific strengths and weaknesses of the model, or in order to adapt it to other related structured prediction tasks. Last, with respect to the model improvement in Section 5, we focus on the architecture of the RC model, but given the high impact of the `relation type frequency` attribute, data augmentation techniques could be explored in order to further improve the performance of the model.

Acknowledgements

We thank the NLPnorth group at ITU and the MaiNLP group at LMU for feedback on an earlier version of this paper. EB and BP are supported by the Independent Research Fund Denmark (Danmarks Frie Forskningsfond; DFF) Sapere Aude grant 9063-00077B. BP is in parts supported by the European Research Council (ERC) grant agreement No. 101043235.

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.

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.

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Aliva Das, Xinya Du, Barry Wang, Kejian Shi, Jiayuan Gu, Thomas Porter, and Claire Cardie. 2022. Automatic error analysis for document-level information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3960–3975, Dublin, Ireland. Association for Computational Linguistics.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

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.

Zhichao Duan, Xiuxing Li, Zhenyu Li, Zhuo Wang, and Jianyong Wang. 2022. Not just plain text! fuel document-level relation extraction with explicit syntax refinement and subsentence modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1941–1951, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinlan Fu, Liangjing Feng, Qi Zhang, Xuanjing Huang, and Pengfei Liu. 2021. Larger-context tagging: When and why does it work? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1463–1475, Online. Association for Computational Linguistics.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020a. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online. Association for Computational Linguistics.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020b. RethinkCWS: Is Chinese word segmentation a solved task? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5676–5686, Online. 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.

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.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2606–2617, Dubrovnik, Croatia. Association for Computational Linguistics.

Qiushi Guo, Xin Wang, and Dehong Gao. 2022. Dependency position encoding for relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1601–1606, Seattle, United States. 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.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Arzoo Katiyar and Claire Cardie. 2017. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 917–928, Vancouver, Canada. Association for Computational Linguistics.

Juhyuk Lee, Min-Joong Lee, June Yong Yang, and Eunho Yang. 2022. Does it really generalize well on unseen data? systematic evaluation of relational triple extraction methods. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3858, Seattle, United States. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021a. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

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

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.

Nicholas Popovic and Michael Färber. 2022. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5733–5746, Seattle, United States. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Thien Nguyen, and Dejing Dou. 2019. Improving cross-domain performance for relation extraction via dependency prediction and information flow control. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5153–5159. International Joint Conferences on Artificial Intelligence Organization.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

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.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. UniRel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Zhen Wan, Fei Cheng, Qianying Liu, Zhuoyuan Mao, Haiyue Song, and Sadao Kurohashi. 2023. Relation extraction with weighted contrastive pre-training on distant supervision. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2580–2585, Dubrovnik, Croatia. Association for Computational Linguistics.

Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. 2022a. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. 2022b. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States. Association for Computational Linguistics.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, and Bryan Hooi. 2022c. GraphCache: Message passing as caching for sentence-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1698–1708, Seattle, United States. 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.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Liang Zhang, Jinsong Su, Yidong Chen, Zhongjian Miao, Min Zijun, Qingguo Hu, and Xiaodong Shi. 2022a. Towards better document-level relation extraction via iterative inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8306–8317, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. 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.

Yunqi Zhang, Yubo Chen, and Yongfeng Huang. 2022b. RelU-net: Syntax-aware graph U-net for relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4217, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie, and Dongsheng Li. 2022. Improving few-shot relation classification by prototypical representation learning with definition text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 454–464, Seattle, United States. 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.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

A CrossRE Statistics

We report in Table 4 the dataset statistics of CrossRE (Bassignana and Plank, 2022).

B Entity Type Mapping

The CrossRE dataset adopts the 39 domain-specific entity types initially proposed by Liu et al. (2021b) in CrossNER. When dealing with the `entity type` and `entity type frequency` attributes, in order to perform our cross-domain analysis, we map the original 39 entity types into five domain-agnostic meta entity types as illustrated in Table 5.

C Reproducibility

We report in Table 6 the hyperparameter setting of our RC model (see Section 3.2). 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 the Almost Stochastic Order test (ASO; Del Barrio et al. (2018); Dror et al. (2019)) implementation by Ulmer et al. (2022). The method computes a score (ϵ_{min}) which represents how far the first is from being significantly better in respect to the second. The possible scenarios are therefore $\epsilon_{min} = 0.0$ (*truly stochastic dominance*), and $\epsilon_{min} < 0.5$ (*almost stochastic dominance*). Table 7 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.

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

Table 4: **CrossRE Statistics.** Number of sentences and number of relations for each domain of CrossRE (Bassignana and Plank, 2022).

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

Table 5: **Entity Hierarchy.** Mapping of the original 39 domain-specific entity types by Liu et al. (2021b) into five domain-agnostic meta types.

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

Table 6: **Hyperparameters Setting.** Model details for reproducibility of the experiments.

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

Table 7: **Significance Testing.** ASO scores comparing the experimental setups described in Section 5. Read as row → column.