Given a multi-domain training setup, is it possible to improve the performances by adding information about the domain itself?

Germán Buttiero, Hubert Wójcik and Paula Menshikoff

gebu@itu.dk, huwo@itu.dk and pmen@itu.dk

1 Abstract

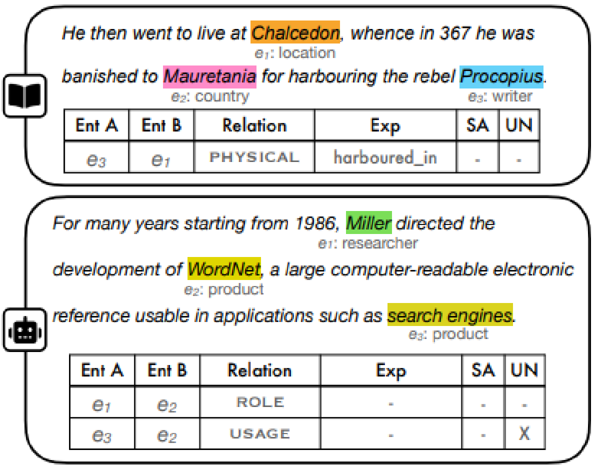
Relation classification is an important task in Natural Language Processing that aims at predicting the relation between two entities in a sentence. The amount of studies on this topic is rapidly growing. However, most of the available research is regarding in-domain setups. Therefore, there is limited knowledge of the performance of models across domains. We use a cross-domain dataset and a state-of-the-art neural network classifier to investigate the performance of the introduction of more context to the dataset. We experiment with the augmentation of the sentences with (special and pre-trained) tokens related to the domain to which they belong to. Results show that augmenting the sentence with a special token increases the performance the most (compared to the rest of the methods). [1]

2 Introduction

Over the last few years, predicting the relation of two entities in a sentence has gained popularity. Many datasets, together with many classification models have been developed and distributed. However, Bassignana and Plank (2022b) found that, when dealing with relation classification, most of the available datasets only include in-domain setups. Therefore, the little development of cross-domain setups for relation classification results in a limited knowledge of the performance of cross-domain relation classification systems (Bassignana and Plank, 2022a).

Identifying the relation of two entities in a Relation Extraction task is fundamental in Natural Language Processing tasks such as question answering, text understanding, etc. (Huang and Wang, 2017). The possible benefits of cross-domain models lie in the fact that they could be generalized instead of them being limited to an in-domain setup. Therefore, investigating whether cross-domain models outperform in-domain ones is crucial.

Because of the little development of cross-domain datasets for relation classification, Bassignana and Plank (2022a) have created a dataset composed of sentences related to six different domains, willing to enlarge the number of studies within this field. The key to this cross relation extraction dataset (CROSS-RE) lies in the annotation system, where the authors have defined the annotation rules and they have labeled the relation of two entities for a total number of 18608 relations (Figure 1).



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: CROSS-RE Samples from Literature and Artificial Intelligence Domains from Bassignana and Plank 2022a

They have defined 17 different labels which are not distributed equally across domains (as can be seen in figure 2). Moreover, most of the entities presented in the dataset have additional meta-data regarding the meaning of the entity.

In their work, Basignana and Plank introduce a state-of-the-art model for relation classification following the model from Baldini Soares et al. 2019. We modify this state-of-the-art model to serve as a cross-domain baseline model and subsequently design different methods to investigate the possibility

[1] https://github.itu.dk/huwo/cross_domain_rc_syp

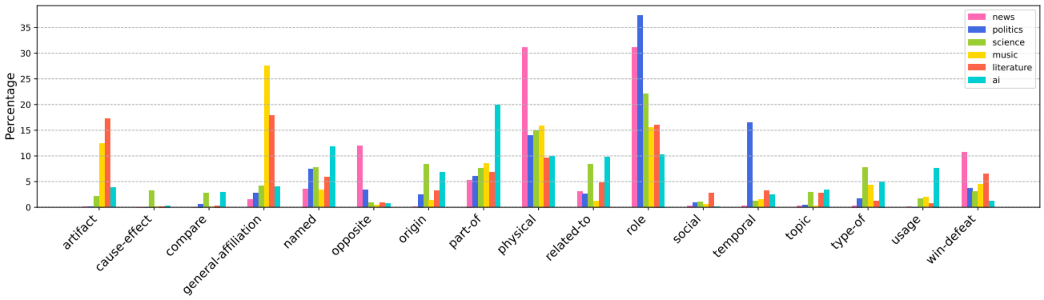


Figure 2: CROSSRE Label Distribution from Bassignana and Plank 2022a

of performance enhancement. Our primary focus centers on introducing more context to the sentences before they are input into the model. Then, the performance of these approaches is evaluated and the results are shown. Through this study, we aim to expand the knowledge of the potential of generalization in relation extraction.

3 Related work

Building upon the foundational work by Bassignana and Plank (2022a), we seek to further investigate the field of cross-domain relation extraction. Despite the increased popularity of Relation Extraction tasks, cross-domain for relation extraction remains relatively unexplored (Zhao and Grishman, 2005). In their research, Bassignana and Plank confronted the limitations of current Relation Extraction evaluations, which are largely confined to in-domain setups. They found that the domains which performed the best were actually the ones with the highest amount of relations annotated.

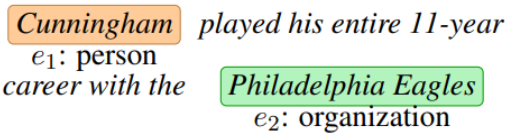
Moreover, the work of (Hausser, 2017) gives an interesting perspective. The authors investigate a concept called "generalized reference" which is a method of adding context to language processing tasks. They propose four ways for adding context: reference by matching, pointing, baptism, and reference by address. Although these methods do not specifically involve the use of a special token at the beginning of a sentence, they do bring some valuable insights into how context can be incorporated into language processing tasks and used to improve models' performance.

In particular, the idea of using a name to refer to something specific, as the baptism method mentioned above, could potentially be adapted to the use of domain-specific tokens. This could serve as a foundation for the strategy of adding a domain-specific token at the start of a sentence to help with understanding the context better.

4 Experiments

4.1 Baseline model

Each of the sentences from the CROSS-RE dataset is augmented with entity marks marking the start and end position of each entity (Figure 3). Additional tokens "[CLS]" and "[SEP]" are also appended to the sentences at the beginning and end of the sentences, respectively. Moreover, the metadata, if existent, is also included within the entity marks.



s is augmented as:

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

Figure 3: Sample of sentence augmented from Bassignana and Plank 2022a

The augmented sentences from the different domains but from the same sets (train, dev) are all merged to make up the new datasets used for the baseline. As in Bassignana and Plank (2022a) work, the model then passes these sentences into a BERT model in order to obtain the embeddings of the two starting markers (Figure 4).

The embeddings of the two start markers serve as the input of a one-layer Feed Forward Neural Network that once trained, predicts the relation of the two entities.

Metric	News	Politics	N. Science	Music	Literature	Artificial Science
Baseline	0.57	0.61	0.47	0.79	0.69	0.48
Method 1	0.59	0.62	0.49	0.79	0.72	0.51

Table 1: Comparison of the Micro-F1 metric between models

Model	News	Politics	N. Science	Music	Literature	Artificial Science
Baseline	0.15	0.28	0.37	0.44	0.41	0.33
Method 1	0.21	0.27	0.40	0.47	0.46	0.40

Table 2: Comparison of the Macro-F1 metric between models

Model	News	Politics	N. Science	Music	Literature	Artificial Science
Baseline.	0.49	0.56	0.43	0.77	0.65	0.44
Method 1.	0.54	0.59	0.40	0.77	0.69	0.49

Table 3: Comparison of the Weighted-F1 metric between models

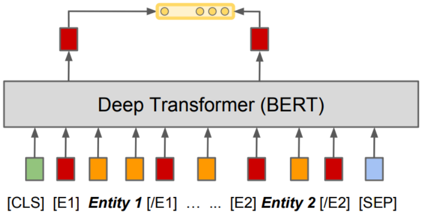


Figure 4: Architecture for extracting relation representation from Baldini Soares et al. 2019

Table 1, 2 and table 3 show the performance of the state-of-the-art model per domain compared to the performance of method 1.

4.2 Methods

We have identified 4 different approaches to introduce more context to the sentences by augmenting them with different tokens. We aim at guiding the model to recognize to which domain the sentences belong so it can follow the label distribution of the in-domain set instead of the general distribution. All the models have been trained on the complete train set (where all the domains have been merged) and tested individually on each domain test set.

For a better understanding, we will use as an example the following sentence: *"The Belgium group Technotronic scored an international hit with the song Pump Up the Jam."*.

The baseline model gets the sentence augmented as:

"*[CLS] The* <E1:country> *Belgium* </E1:country> *group* <E2:band> *Technotronic* </E2:band> *scored an international hit with the song Pump Up the Jam. [SEP]*".

4.2.1 Method 1

The first approach consists of augmenting the sentence with a special token that matches the sentence with the domain they belong to. In this method, the model learns that these special tokens only appear in sentences from the same domain. In our example, the sentence is augmented to:

"*[CLS] [music] The* <E1:country> *Belgium* </E1:country> *group* <E2:band> *Technotronic* </E2:band> *scored an international hit with the song Pump Up the Jam. [SEP]*".

4.2.2 Method 2

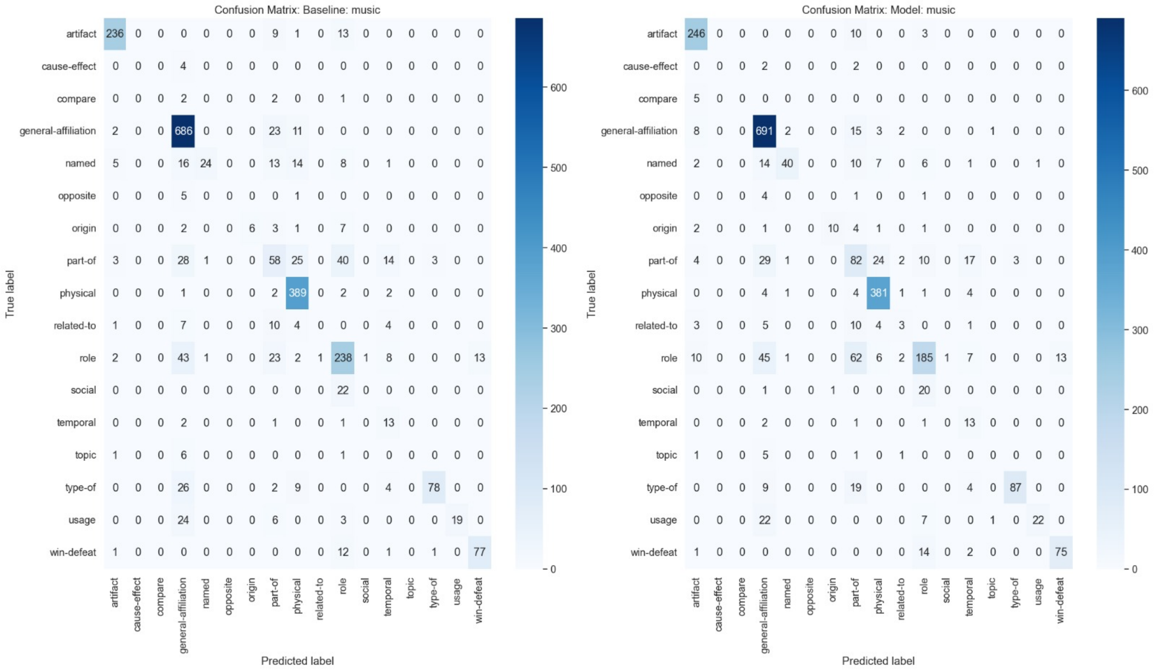
A variation of the first method, where the appended tokens are already part of the existing vocabulary. Therefore, the model learns that the "domain" token occurs mostly within the domain dataset, but its occurrence is not domain specific. This method augments the sentence as:

"*[CLS] music The* <E1:country> *Belgium* </E1:country> *group* <E2:band> *Technotronic* </E2:band> *scored an international hit with the song Pump Up the Jam. [SEP]*".

4.2.3 Method 3

A variation of the second model where we also include the token [SEP] to separate the domain token from the rest of the sentence. For example:

"*[CLS] music [SEP] The* <E1:country> *Belgium* </E1:country> *group* <E2:band> *Technotronic* </E2:band> *scored an international hit with the song Pump Up the Jam. [SEP]*".



Confusion Matrix: Baseline: music

True label \ Predicted label	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	236	0	0	0	0	0	0	9	1	0	13	0	0	0	0	0	0
cause-effect	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
compare	0	0	0	2	0	0	0	2	0	0	1	0	0	0	0	0	0
general-affiliation	2	0	0	686	0	0	0	23	11	0	0	0	0	0	0	0	0
named	5	0	0	16	24	0	0	13	14	0	8	0	1	0	0	0	0
opposite	0	0	0	5	0	0	0	0	1	0	0	0	0	0	0	0	0
origin	0	0	0	2	0	0	6	3	1	0	7	0	0	0	0	0	0
part-of	3	0	0	28	1	0	0	58	25	0	40	0	14	0	3	0	0
physical	0	0	0	1	0	0	0	2	389	0	2	0	2	0	0	0	0
related-to	1	0	0	7	0	0	0	10	4	0	0	0	4	0	0	0	0
role	2	0	0	43	1	0	0	23	2	1	238	1	8	0	0	0	13
social	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0
temporal	0	0	0	2	0	0	0	1	0	0	1	0	13	0	0	0	0
topic	1	0	0	6	0	0	0	0	0	0	1	0	0	0	0	0	0
type-of	0	0	0	26	0	0	0	2	9	0	0	0	4	0	78	0	0
usage	0	0	0	24	0	0	0	6	0	0	3	0	0	0	0	19	0
win-defeat	1	0	0	0	0	0	0	0	0	0	12	0	1	0	1	0	77

Confusion Matrix: Model: music

True label \ Predicted label	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	246	0	0	0	0	0	0	10	0	0	3	0	0	0	0	0	0
cause-effect	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0
compare	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
general-affiliation	8	0	0	691	2	0	0	15	3	2	0	0	0	1	0	0	0
named	2	0	0	14	40	0	0	10	7	0	6	0	1	0	0	1	0
opposite	0	0	0	4	0	0	0	1	0	0	1	0	0	0	0	0	0
origin	2	0	0	1	0	0	10	4	1	0	1	0	0	0	0	0	0
part-of	4	0	0	29	1	0	0	82	24	2	10	0	17	0	3	0	0
physical	0	0	0	4	1	0	0	4	381	1	1	0	4	0	0	0	0
related-to	3	0	0	5	0	0	0	10	4	3	0	0	1	0	0	0	0
role	10	0	0	45	1	0	0	62	6	2	185	1	7	0	0	0	13
social	0	0	0	1	0	0	1	0	0	0	20	0	0	0	0	0	0
temporal	0	0	0	2	0	0	0	1	0	0	1	0	13	0	0	0	0
topic	1	0	0	5	0	0	0	1	0	1	0	0	0	0	0	0	0
type-of	0	0	0	9	0	0	0	19	0	0	0	0	4	0	87	0	0
usage	0	0	0	22	0	0	0	0	0	0	7	0	0	1	0	22	0
win-defeat	1	0	0	0	0	0	0	0	0	0	14	0	2	0	0	0	75

Figure 5: Confusion Matrix for "Music" test set for baseline and Method 1

4.2.4 Method 4

This is a variation of method 1, whereas, in method 3, we include the [SEP] token after the special domain token, in order to split the sentence in two. In this case, we augment the sentence like:

"*[CLS] [music] [SEP] The* <E1:country> *Belgium* </E1:country> *group* <E2:band> *Technotronic* </E2:band> *scored an international hit with the song Pump Up the Jam. [SEP]*".

5 Results

Table 1, 2 and 3 show the results of method 1 per domain. The other methods' results can be found in tables 4, 5, and 6. We have calculated the Micro-F1, Macro-F1, and weighted-F1 to get a better understanding of the performance of the different methods.

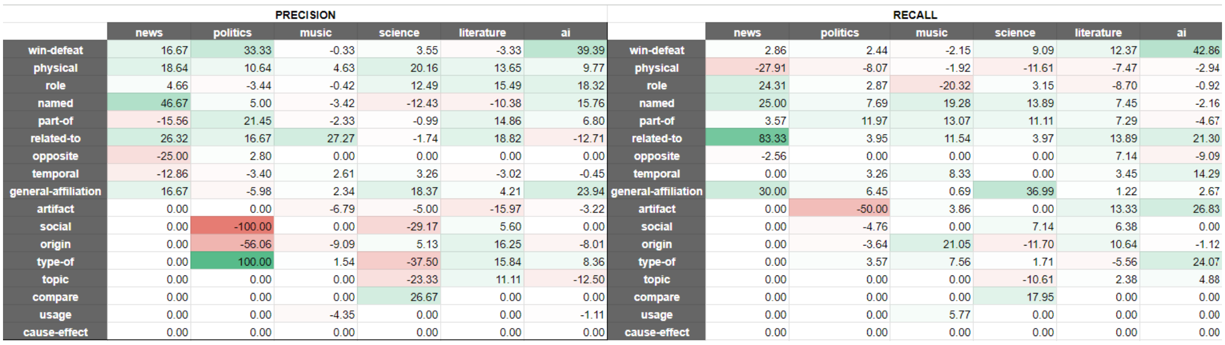
Due to the high-class imbalance of the dataset (there are only 871 "news" relations in total while there are 4690 total relations belonging to the domain "music"), we mostly focus on the macro-f1 scores that only consider the classes that appear in the test set.

In general, method 1, improves the macro-F1 score for most of the six domains. However, the value is still quite low for domains like "news" where there was not sufficient data compared to the rest of the domains. But in domains like "music" and "literature" which presented the highest amount of relations, our approach improves performance by 3 percentage points and 5 percentage points, respectively. Even though the performance improvement is slight, it serves as a starting point to keep investigating the benefits of providing more context to sentences.

We identify that our model tends to consider the domain class distribution whereas the baseline model is guided by a general class distribution. Due to the high-class imbalance, the baseline classifier tends to assign relations to the classes that appear more often in the dataset like "role", "physical", etc. On the other hand, our model, as it gets more context regarding the domain the sentences belong to, tends to better classify relations considering the domain class distribution. Besides, it is better at identifying classes that do not appear often.

For example, for the domain "music" (Figure 5), it can be seen that the baseline most likely classifies the relations as either "general/affiliation", "physical" or "role". These classes ("role" and "physical") are the most common classes when considering all the domains together. On the other hand, our model mostly classifies the entity relations as either "general/affiliation", "physical" or "artifact". Moreover, it improves the amount of correct classified sentences for those classes that do not appear often.

In general, we notice that our model gets a higher recall on the most common domain classes compared to the baseline (as it can be seen in figure 6).



	PRECISION							RECALL					
	news	politics	music	science	literature	ai		news	politics	music	science	literature	ai
win-defeat	16.67	33.33	-0.33	3.55	-3.33	39.39	win-defeat	2.86	2.44	-2.15	9.09	12.37	42.86
physical	18.64	10.64	4.63	20.16	13.65	9.77	physical	-27.91	-8.07	-1.92	-11.61	-7.47	-2.94
role	4.66	-3.44	-0.42	12.49	15.49	18.32	role	24.31	2.87	-20.32	3.15	-8.70	-0.92
named	46.67	5.00	-3.42	-12.43	-10.38	15.76	named	25.00	7.69	19.28	13.89	7.45	-2.16
part-of	-15.56	21.45	-2.33	-0.99	14.86	6.80	part-of	3.57	11.97	13.07	11.11	7.29	-4.67
related-to	26.32	16.67	27.27	-1.74	18.82	-12.71	related-to	83.33	3.95	11.54	3.97	13.89	21.30
opposite	-25.00	2.80	0.00	0.00	0.00	0.00	opposite	-2.56	0.00	0.00	0.00	7.14	-9.09
temporal	-12.86	-3.40	2.61	3.26	-3.02	-0.45	temporal	0.00	3.26	8.33	0.00	3.45	14.29
general-affiliation	16.67	-5.98	2.34	18.37	4.21	23.94	general-affiliation	30.00	6.45	0.69	36.99	1.22	2.67
artifact	0.00	0.00	-6.79	-5.00	-15.97	-3.22	artifact	0.00	-50.00	3.86	0.00	13.33	26.83
social	0.00	-100.00	0.00	-29.17	5.60	0.00	social	0.00	-4.76	0.00	7.14	6.38	0.00
origin	0.00	-56.06	-9.09	5.13	16.25	-8.01	origin	0.00	-3.64	21.05	-11.70	10.64	-1.12
type-of	0.00	100.00	1.54	-37.50	15.84	8.36	type-of	0.00	3.57	7.56	1.71	-5.56	24.07
topic	0.00	0.00	0.00	-23.33	11.11	-12.50	topic	0.00	0.00	0.00	-10.61	2.38	4.88
compare	0.00	0.00	0.00	26.67	0.00	0.00	compare	0.00	0.00	0.00	17.95	0.00	0.00
usage	0.00	0.00	-4.35	0.00	0.00	-1.11	usage	0.00	0.00	5.77	0.00	0.00	0.00
cause-effect	0.00	0.00	0.00	0.00	0.00	0.00	cause-effect	0.00	0.00	0.00	0.00	0.00	0.00

Figure 6: Comparison of Confusion Matrix for "precision" and "recall" on baseline and Method 1 approach. The values are percentages, a positive value means that the model is performing better than the baseline

For example, our model has improved the recall of the class artifact by 13.33 percentage points for the "literature" domain. Moreover, for the "politics" domain, the class "temporal" has improved by 3.26 percentage points, meaning that it is better at identifying these classes. Intuitively, this improvement comes sometimes with a decrease in the domain's most common classes' precision.

On the other hand, when we consider the general most common classes like "physical" we see an improvement in their precision. Our model does not classify as many sentences as these labels compared to the baseline. However, this translates into a decrease in their recall.

Lastly, for the labels that do not appear often either in the general or within the domain, our model often improves their recall (as for "name" and "part-of") more than their precision. However, these results might change if we were to have more sentences in our dataset.

6 Conclusion

In conclusion, different methods are presented to augment the input sentences with more context regarding the domain they belong to and how it performs when compared with a slightly modified state-of-the-art relation classification model by Bassignana and Plank 2022a.

We notice that the introduction of a special token that is only common between in-domain sentences has a positive impact on almost every domain compared to the state-of-the-art model. However, we also show how complicated it is to work with a highly imbalanced label distribution over domains.

Although the improvement, as previously stated, might not be sufficient, it serves as evidence that a cross-domain classifier could be possible in the future.

7 Limitations

Given imperfect labeling from annotators and the inherent ambiguity of language, a single label is not sufficient to learn the spectrum of language interpretation. (Zhang et al., 2021) In addition, the uneven distribution of labels across different domains is also a major limitation as it could make our model skewed and affect its ability to generalize.

Moreover, certain domains, such as news, are underrepresented in the dataset, with a relatively small number of samples. This was detrimental for the model, and we decided to focus more on the domains that have a higher number of samples as they present a broader scope for analysis and offer a more representative view.

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.

Elisa Bassignana and Barbara Plank. 2022a. 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.

Elisa Bassignana and Barbara Plank. 2022b. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification.

Ronald Hausser. 2017. A computational treatment of generalized reference 5, 2. In *Complex Adaptive Systems Modeling*.

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. *CoRR*, abs/2109.04408.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 419–426, Ann Arbor, Michigan. Association for Computational Linguistics.

8 Appendix

8.1 Participation

All the members of the team participated equally in the programming process, analysis, and writing of the report. Due to issues with pushing to GitHub, we worked on one computer.

8.2 Chatbots Usage

As we struggled to find related work using ordinary search engines, we used ChatGPT's scholar.ai plugin to efficiently query through the database of academic papers and find those that could be relevant to our study.

8.3 Results for the rest of methods

Metrics	News	Politics	N. Science	Music	Literature	Artificial Science
Micro-F1	0.59	0.61	0.47	0.78	0.71	0.50
Macro-F1	0.19	0.27	0.39	0.43	0.42	0.37
Weigh. F1	0.53	0.56	0.43	0.75	0.66	0.46

Table 4: Baseline model: Results achieved by Method 2

Metrics	News	Politics	N. Science	Music	Literature	Artificial Science
Micro-F1	0.47	0.59	0.43	0.75	0.67	0.46
Macro-F1	0.13	0.25	0.31	0.38	0.36	0.28
Weigh. F1	0.39	0.54	0.40	0.72	0.62	0.40

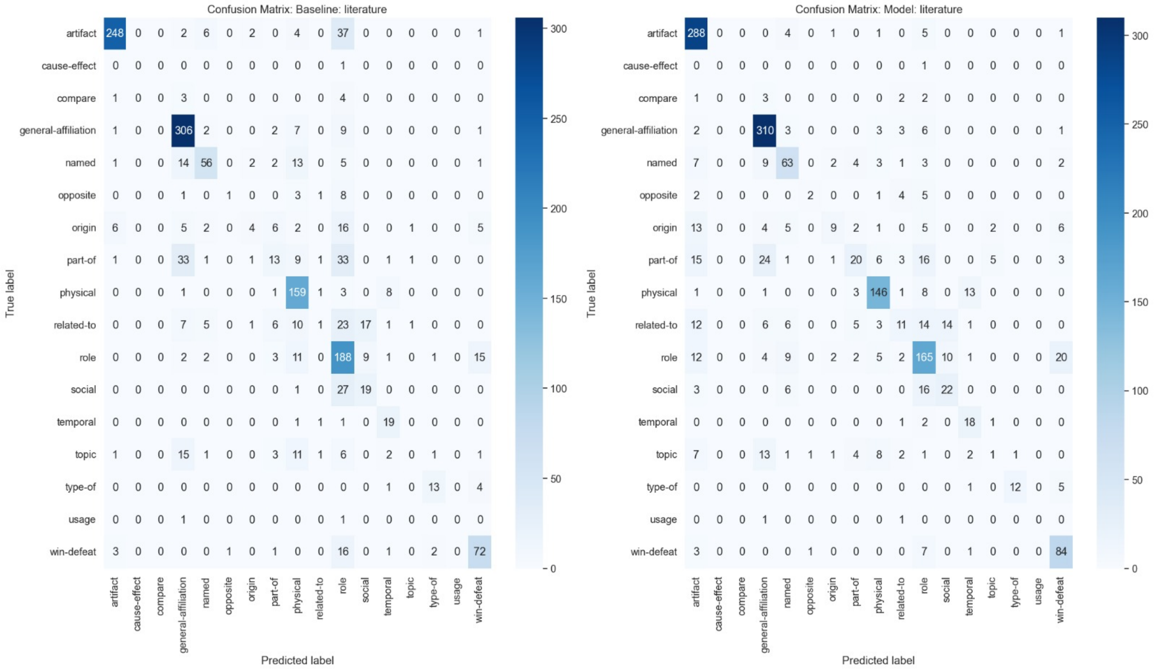
Table 5: Baseline model: Results achieved by Method 3

Metrics	News	Politics	N. Science	Music	Literature	Artificial Science
Micro-F1	0.55	0.61	0.46	0.78	0.70	0.50
Macro-F1	0.17	0.24	0.36	0.43	0.40	0.35
Weigh. F1	0.49	0.57	0.42	0.76	0.65	0.47

Table 6: Baseline model: Results achieved by Method 4

8.4 Confusion Matrix

Here we provide the confusion matrix for the rest of the domains for the baseline and Method 1.



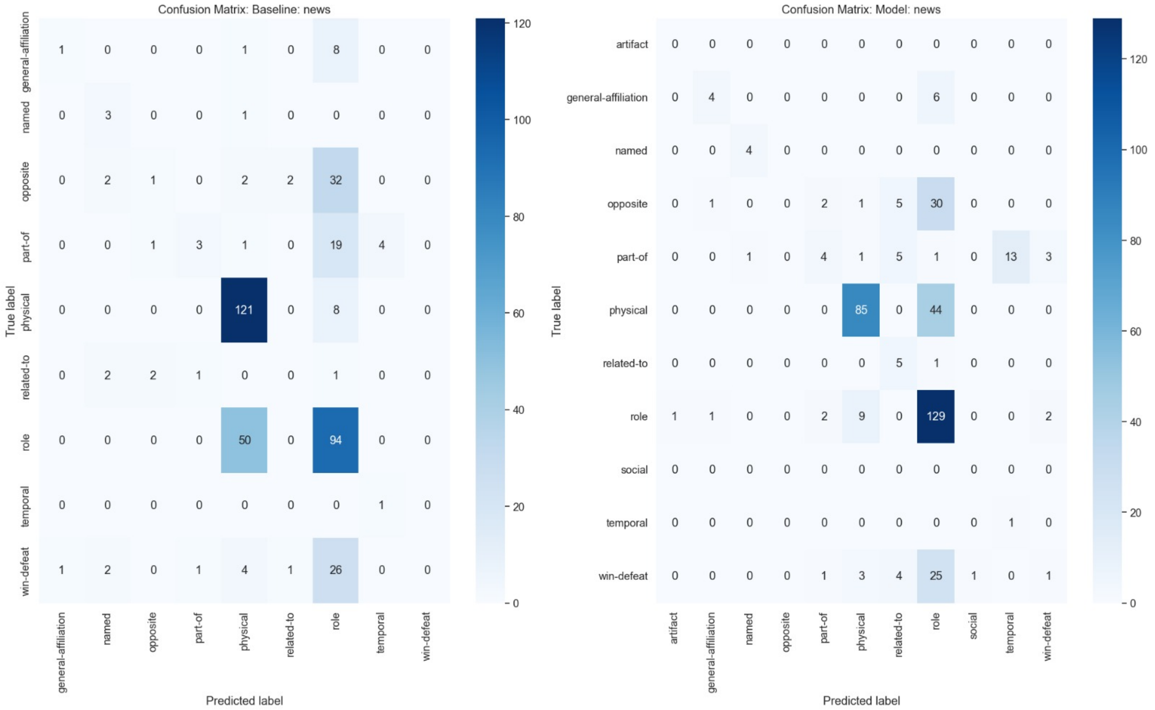
Confusion Matrix: Baseline: literature

	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	248	0	0	2	6	0	2	0	4	0	37	0	0	0	0	0	1
cause-effect	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
compare	1	0	0	3	0	0	0	0	0	0	4	0	0	0	0	0	0
general-affiliation	1	0	0	306	2	0	0	2	7	0	9	0	0	0	0	0	1
named	1	0	0	14	56	0	2	2	13	0	5	0	0	0	0	0	1
opposite	0	0	0	1	0	1	0	0	3	1	8	0	0	0	0	0	0
origin	6	0	0	5	2	0	4	6	2	0	16	0	0	1	0	0	5
part-of	1	0	0	33	1	0	1	13	9	1	33	0	1	1	0	0	0
physical	0	0	0	1	0	0	0	1	159	1	3	0	8	0	0	0	0
related-to	0	0	0	7	5	0	1	6	10	1	23	17	1	1	0	0	0
role	0	0	0	2	2	0	0	3	11	0	188	9	1	0	1	0	15
social	0	0	0	0	0	0	0	0	1	0	27	19	0	0	0	0	0
temporal	0	0	0	0	0	0	0	0	1	1	1	0	19	0	0	0	0
topic	1	0	0	15	1	0	0	3	11	1	6	0	2	0	1	0	1
type-of	0	0	0	0	0	0	0	0	0	0	0	0	1	0	13	0	4
usage	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
win-defeat	3	0	0	0	0	1	0	1	0	0	16	0	1	0	2	0	72

Confusion Matrix: Model: literature

	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	288	0	0	0	4	0	1	0	1	0	5	0	0	0	0	0	1
cause-effect	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
compare	1	0	0	3	0	0	0	0	0	2	2	0	0	0	0	0	0
general-affiliation	2	0	0	310	3	0	0	0	3	3	6	0	0	0	0	0	1
named	7	0	0	9	63	0	2	4	3	1	3	0	0	0	0	0	2
opposite	2	0	0	0	0	2	0	0	1	4	5	0	0	0	0	0	0
origin	13	0	0	4	5	0	9	2	1	0	5	0	0	2	0	0	6
part-of	15	0	0	24	1	0	1	20	6	3	16	0	0	5	0	0	3
physical	1	0	0	1	0	0	0	3	146	1	8	0	13	0	0	0	0
related-to	12	0	0	6	6	0	0	5	3	11	14	14	1	0	0	0	0
role	12	0	0	4	9	0	2	2	5	2	165	10	1	0	0	0	20
social	3	0	0	0	6	0	0	0	0	0	16	22	0	0	0	0	0
temporal	0	0	0	0	0	0	0	0	0	1	2	0	18	1	0	0	0
topic	7	0	0	13	1	1	1	4	8	2	1	0	2	1	1	0	0
type-of	0	0	0	0	0	0	0	0	0	0	0	0	1	0	12	0	5
usage	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
win-defeat	3	0	0	0	0	1	0	0	0	0	7	0	1	0	0	0	84

Figure 7: Confusion Matrix for "Literature" test set for baseline and Method 1



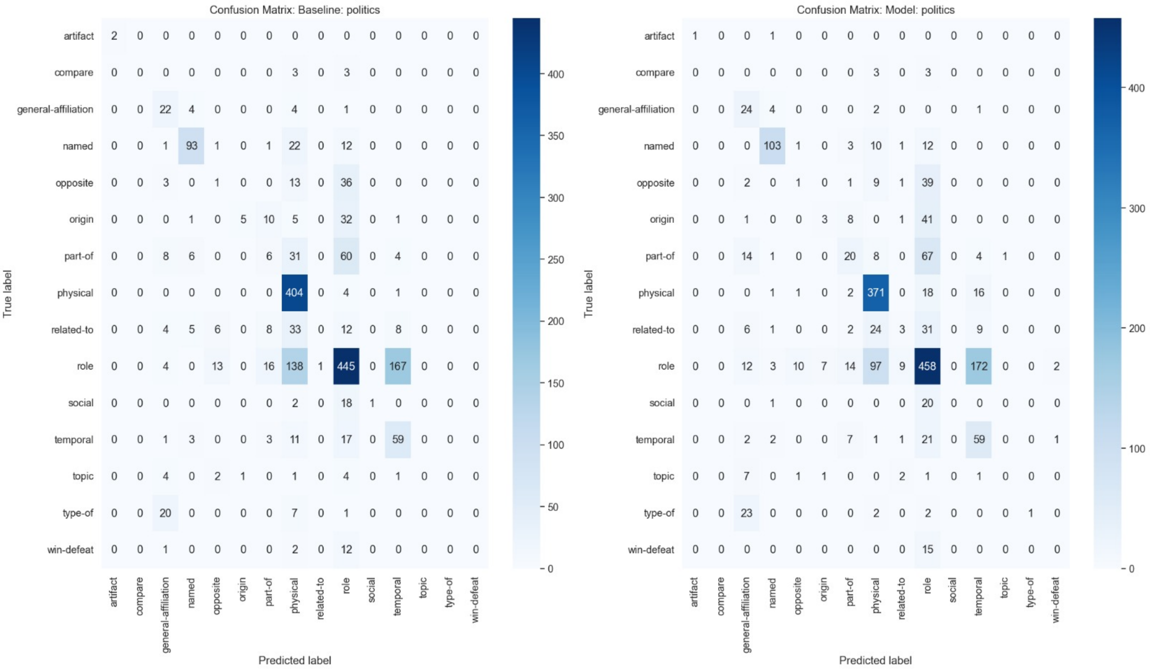
Confusion Matrix: Baseline: news

	general-affiliation	named	opposite	part-of	physical	related-to	role	temporal	win-defeat
general-affiliation	1	0	0	0	1	0	8	0	0
named	0	3	0	0	1	0	0	0	0
opposite	0	2	1	0	2	2	32	0	0
part-of	0	0	1	3	1	0	19	4	0
physical	0	0	0	0	121	0	8	0	0
related-to	0	2	2	1	0	0	1	0	0
role	0	0	0	0	50	0	94	0	0
temporal	0	0	0	0	0	0	0	1	0
win-defeat	1	2	0	1	4	1	26	0	0

Confusion Matrix: Model: news

	artifact	general-affiliation	named	opposite	part-of	physical	related-to	role	social	temporal	win-defeat
artifact	0	0	0	0	0	0	0	0	0	0	0
general-affiliation	0	4	0	0	0	0	0	6	0	0	0
named	0	0	4	0	0	0	0	0	0	0	0
opposite	0	1	0	0	2	1	5	30	0	0	0
part-of	0	0	1	0	4	1	5	1	0	13	3
physical	0	0	0	0	0	85	0	44	0	0	0
related-to	0	0	0	0	0	0	5	1	0	0	0
role	1	1	0	0	2	9	0	129	0	0	2
social	0	0	0	0	0	0	0	0	0	0	0
temporal	0	0	0	0	0	0	0	0	0	1	0
win-defeat	0	0	0	0	1	3	4	25	1	0	1

Figure 8: Confusion Matrix for "News" test set for baseline and Method 1



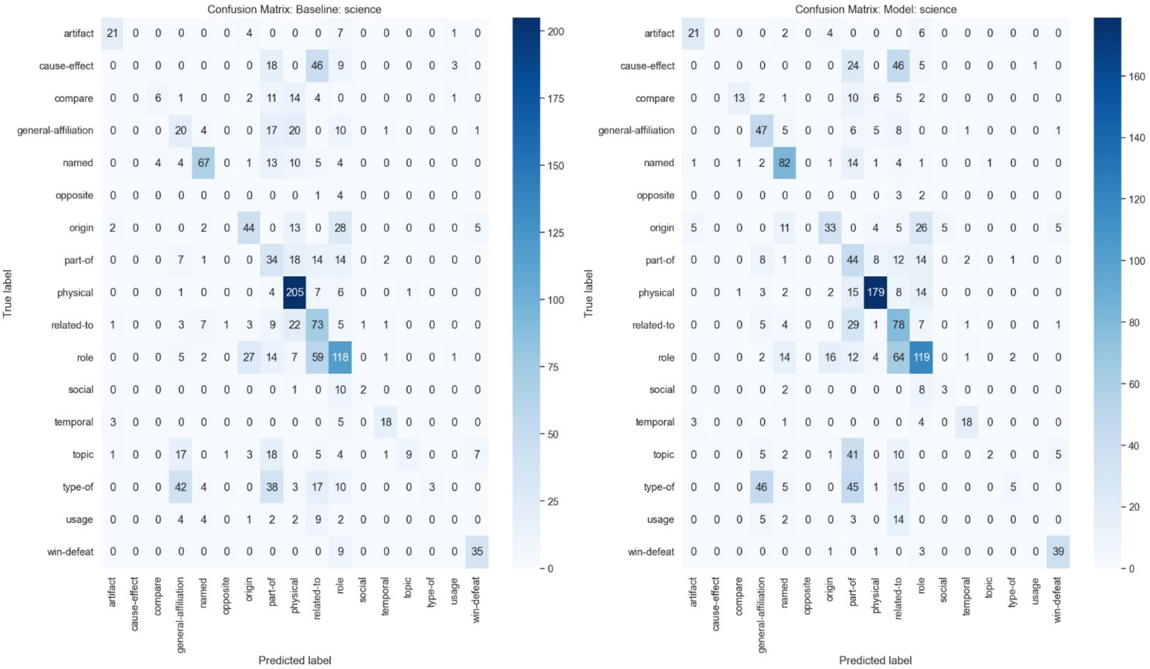
Confusion Matrix: Baseline: politics

	artifact	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	win-defeat
artifact	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
compare	0	0	0	0	0	0	0	3	0	3	0	0	0	0	0
general-affiliation	0	0	22	4	0	0	0	4	0	1	0	0	0	0	0
named	0	0	1	93	1	0	1	22	0	12	0	0	0	0	0
opposite	0	0	3	0	1	0	0	13	0	36	0	0	0	0	0
origin	0	0	0	1	0	5	10	5	0	32	0	1	0	0	0
part-of	0	0	8	6	0	0	6	31	0	60	0	4	0	0	0
physical	0	0	0	0	0	0	0	404	0	4	0	1	0	0	0
related-to	0	0	4	5	6	0	8	33	0	12	0	8	0	0	0
role	0	0	4	0	13	0	16	138	1	445	0	167	0	0	0
social	0	0	0	0	0	0	0	2	0	18	1	0	0	0	0
temporal	0	0	1	3	0	0	3	11	0	17	0	59	0	0	0
topic	0	0	4	0	2	1	0	1	0	4	0	1	0	0	0
type-of	0	0	20	0	0	0	0	7	0	1	0	0	0	0	0
win-defeat	0	0	1	0	0	0	0	2	0	12	0	0	0	0	0

Confusion Matrix: Model: politics

	artifact	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	win-defeat
artifact	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
compare	0	0	0	0	0	0	0	3	0	3	0	0	0	0	0
general-affiliation	0	0	24	4	0	0	0	2	0	0	0	1	0	0	0
named	0	0	0	103	1	0	3	10	1	12	0	0	0	0	0
opposite	0	0	2	0	1	0	1	9	1	39	0	0	0	0	0
origin	0	0	1	0	0	3	8	0	1	41	0	0	0	0	0
part-of	0	0	14	1	0	0	20	8	0	67	0	4	1	0	0
physical	0	0	0	1	1	0	2	371	0	18	0	16	0	0	0
related-to	0	0	6	1	0	0	2	24	3	31	0	9	0	0	0
role	0	0	12	3	10	7	14	97	9	458	0	172	0	0	2
social	0	0	0	1	0	0	0	0	0	20	0	0	0	0	0
temporal	0	0	2	2	0	0	7	1	1	21	0	59	0	0	1
topic	0	0	7	0	1	1	0	0	2	1	0	1	0	0	0
type-of	0	0	23	0	0	0	0	2	0	2	0	0	0	1	0
win-defeat	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0

Figure 9: Confusion Matrix for "Politics" test set for baseline and Method 1



Confusion Matrix: Baseline: science

	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	21	0	0	0	0	0	4	0	0	0	7	0	0	0	0	1	0
cause-effect	0	0	0	0	0	0	0	18	0	46	9	0	0	0	0	3	0
compare	0	0	6	1	0	0	2	11	14	4	0	0	0	0	0	1	0
general-affiliation	0	0	0	20	4	0	0	17	20	0	10	0	1	0	0	0	1
named	0	0	4	4	67	0	1	13	10	5	4	0	0	0	0	0	0
opposite	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0
origin	2	0	0	0	2	0	44	0	13	0	28	0	0	0	0	0	5
part-of	0	0	0	7	1	0	0	34	18	14	14	0	2	0	0	0	0
physical	0	0	0	1	0	0	0	4	205	7	6	0	0	1	0	0	0
related-to	1	0	0	3	7	1	3	9	22	73	5	1	1	0	0	0	0
role	0	0	0	5	2	0	27	14	7	59	118	0	1	0	0	1	0
social	0	0	0	0	0	0	0	0	1	0	10	2	0	0	0	0	0
temporal	3	0	0	0	0	0	0	0	0	0	5	0	18	0	0	0	0
topic	1	0	0	17	0	1	3	18	0	5	4	0	1	9	0	0	7
type-of	0	0	0	42	4	0	0	38	3	17	10	0	0	0	3	0	0
usage	0	0	0	4	4	0	1	2	2	9	2	0	0	0	0	0	0
win-defeat	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	35

Confusion Matrix: Model: science

	artifact	cause-effect	compare	general-affiliation	named	opposite	origin	part-of	physical	related-to	role	social	temporal	topic	type-of	usage	win-defeat
artifact	21	0	0	0	2	0	4	0	0	0	6	0	0	0	0	0	0
cause-effect	0	0	0	0	0	0	0	24	0	46	5	0	0	0	0	1	0
compare	0	0	13	2	1	0	0	10	6	5	2	0	0	0	0	0	0
general-affiliation	0	0	0	47	5	0	0	6	5	8	0	0	1	0	0	0	1
named	1	0	1	2	82	0	1	14	1	4	1	0	0	1	0	0	0
opposite	0	0	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0
origin	5	0	0	0	11	0	33	0	4	5	26	5	0	0	0	0	5
part-of	0	0	0	8	1	0	0	44	8	12	14	0	2	0	1	0	0
physical	0	0	1	3	2	0	2	15	179	8	14	0	0	0	0	0	0
related-to	0	0	0	5	4	0	0	29	1	78	7	0	1	0	0	0	1
role	0	0	0	2	14	0	16	12	4	64	119	0	1	0	2	0	0
social	0	0	0	0	2	0	0	0	0	0	8	3	0	0	0	0	0
temporal	3	0	0	0	1	0	0	0	0	0	4	0	18	0	0	0	0
topic	0	0	0	5	2	0	1	41	0	10	0	0	0	2	0	0	5
type-of	0	0	0	46	5	0	0	45	1	15	0	0	0	0	5	0	0
usage	0	0	0	5	2	0	0	3	0	14	0	0	0	0	0	0	0
win-defeat	0	0	0	0	0	0	1	0	1	0	3	0	0	0	0	0	39

Figure 10: Confusion Matrix for "Science" test set for baseline and Method 1

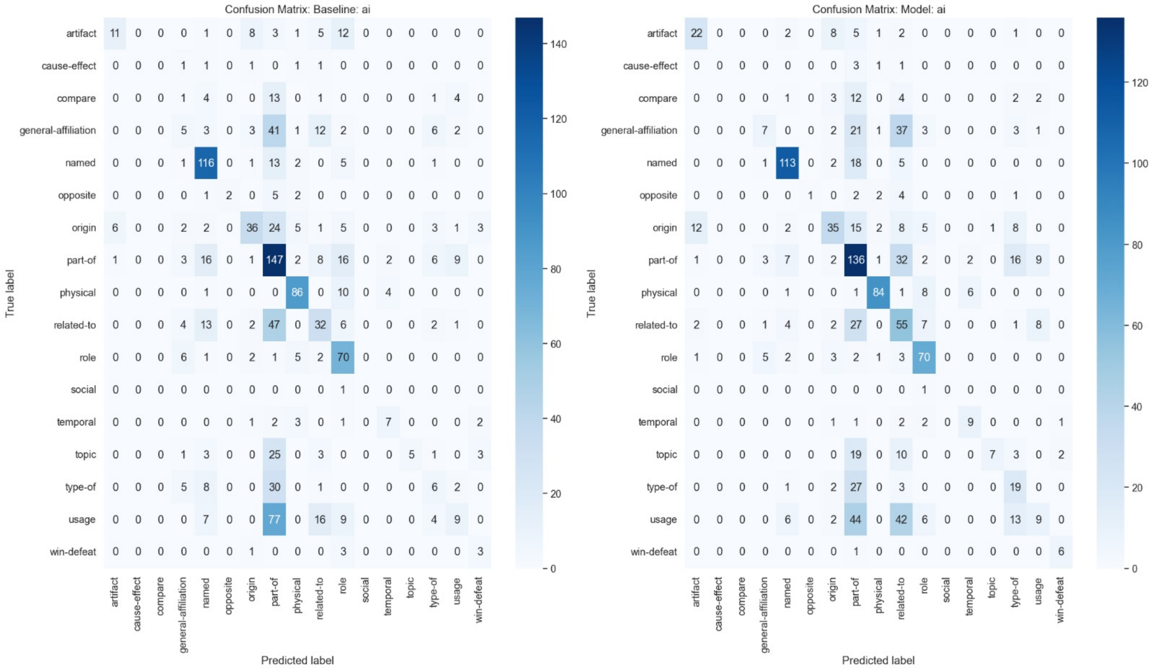


Figure 11: Confusion Matrix for "ai" test set for baseline and Method 1