Politicians Are Also People: Mapping is All You Need

Clustering Entity Types in Cross-Domain Relation Classification Setups

David Peter Süle and **Mie Jonasson** and **Nicklas Koch Rasmussen**
BSc. in Data Science, IT University of Copenhagen
{dasy,miejo,nicra}@itu.dk

Abstract

Relation Extraction is an evolving field within natural language processing. As its last step, Relation Classification (RC) aims to identify the relation type to which two semantically related named entities belong. Cross-domain setups are especially challenging, even more so when domain-specific entity types are used. Research is scarce in the area and mostly focuses on using generic entity types or simply fine-tune the model on a single target domain. This might still offer challenges when annotated data is not accessible for fine-tuning.

In this paper we explore ways of clustering domain-specific named entity types to reduce cross-domain complexity and improve performance on previously unseen domains. We propose five different methods of grouping entity types and evaluate them in multi-domain and out-of-domain scenarios using our two new benchmarks. In conclusion, we find that all our entity mapping methods outperform the baseline in the out-of-domain setting, with the best performing model improving on the baseline by 8.6 percentage points in weighted F1.[1]

1 Introduction

Relation extraction (RE) is an evolving field within natural language processing that aims to discover and classify relations between named entities in a given sentence. Despite out-of-domain generalization being an extremely desirable quality of RE models, research in this direction is lacking (Bassignana and Plank, 2022b). One complicating factor is that named entity types vary across domains and the most common solution is to fine-tune the model on target domain data. In many cases only limited amount of training data is available which poses another challenge. Research combining these two scenarios – *few-shot learning* and *domain shift* – is limited (Gao et al., 2019).

[1]Our code can be found at: https://github.itu.dk/dasy/2yp_project

We suggest and evaluate methods for cross-domain clustering of named entity (NE) types to help relation classification (RC) models generalize better while preserving the necessary level of granularity. We work on the research question: *What is the performance impact of clustering domain-specific named entity types in cross-domain relation-classification setups and what benchmark can be established for future research?*

Our contributions are twofold:

- We propose two benchmarking methods for cross-domain RC models based on the publicly available CrossRE dataset (Bassignana and Plank, 2022a), matching two real life use cases, namely, multi-domain and out-of-domain evaluation.
- We evaluate five different NE type clustering methods with our proposed benchmarks.

2 Related Work

The CrossNER project (Liu et al., 2021) produced a human-annotated NER dataset of six diverse domains with domain-specific entity types. They highlight the lack of domain-specific entities in cross-domain benchmark datasets, leading to less generalizable evaluation results.

The CrossRE project (Bassignana and Plank, 2022a) builds upon the CrossNER data to bring a new cross-domain dataset for relation classification with an established baseline to be improved upon. The project notes the difficulty of the different label distributions across domains, and further highlights the need for research in the underexplored cross-domain dimension of RE.

3 Data

To test our proposed named entity grouping methods and the resulting relation classification model, we use the publicly available cross-domain

CrossRE dataset (Bassignana and Plank, 2022a) which itself is based upon the CrossNER project data (Liu et al., 2021). CrossNER consists of a curated set of sentences in six domains, namely AI, literature, music, natural science, news, and politics, each with domain-specific named entity labels and boundary annotations.

The two sources are Reuters News from the CoNLL-2003 shared task by Tjong Kim Sang and De Meulder (2003) (news domain), and Wikipedia. For details on the data collection, annotation processes, the vocabulary and data statistics we refer to Liu et al. (2021) and the published data.[2]

CrossRE uses the named entities identified in CrossNER (with minor changes) to hand-annotate semantic relations between them, using 17 domain-agnostic relation labels. The dataset models a low-resource scenario for domain-specific fine-tuning, with ~100 / ~350 / ~400 examples in the train/dev/test splits, respectively. This matches the original, resulting in over five thousand sentences.

Both entity and relation type distributions vary across domains; for statistics and annotation details we refer to Bassignana and Plank (2022a) and the published data.[3] For the named entity and relation labels see Appendix A.[4]

4 Entity Type Clustering Methods

Examples and technical details for all clustering methods can be found in Appendices B through G.

4.1 Manual Grouping

For the first method, we started with a list of all entity types and attempted to create meaningful categories for subsets of the entities. We based our initial clusters on a grouping obtained from Elisa Bassignana (Appendix B.1). Throughout the process, we made individual categorizations, used the definitions of the entity types by Liu et al. (2021) to settle doubts, and discussed ambiguous cases.

4.2 Embedding Space Based Grouping

In our second method, we utilized word embedding vectors for algorithmic clustering using the google-news-300 (Google, 2013) word2vec model (Mikolov et al., 2013). This was a suitable choice as Liu et al. (2021) used the news as the source domain. To obtain the embeddings, we simplified named entity labels consisting of two words (e.g. `musical artist` → `musician`) or replaced them with a specific example (e.g. `programming language` → `javascript`); see Appendix G for the full list of substitutions.

We applied a two-phase dimensionality reduction step consisting of PCA and UMAP (Uniform Manifold Approximation and Projection, McInnes et al., 2020). Using UMAP was motivated by our aim to more accurately capture fine detail and favour local information over preserving the global structure. To avoid the risk of picking up patterns from random noise we applied PCA before UMAP (Lee and Verleysen, 2007 and van der Maaten and Hinton, 2008). Finally, we used affinity propagation (Frey and Dueck, 2007) to cluster the results. This has the advantage of automatically determining the number of clusters, and being relatively stable across random seeds.

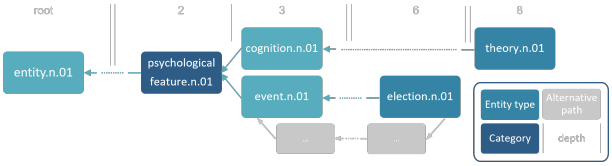


Figure 1: DAG structure of WordNet nouns. Both *theory* and *election* have *psychological feature* as their depth 2 hypernym, which is also their LCS.

4.3 Topological Grouping

We use WordNet (Fellbaum, 1998 and Princeton, 2010), a large lexical database of English, to extract the topological categorization. WordNet groups words into synsets, sets of cognitive synonyms, and models semantic relationships between them, one type of which is the `is-a` relation (for nouns), that forms a directed acyclic graph (DAG) with one root entity. For this reason, it is possible to traverse the hypernyms of the selected synsets to the root, and represent each word with the entity at a given depth in this path. If two words' least common subsumer (LCS) is of equal or higher depth (further from the root) than the chosen depth level, they will be represented by the same label. By varying the chosen distance from the root, we can tune the granularity of the categorization, e.g. the number of labels; we chose a depth of 2, resulting in 9 labels. See Figure 1 for an illustration.

[2] https://github.com/zliucr/CrossNER

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

[4] As we didn't modify the data set, we did not prepare a data statement (Bender and Friedman, 2018) and refer the reader to Appendix A of Bassignana and Plank (2022a) instead.

4.4 Thesaurus-Affinity Based Grouping

We can extend this notion of similarity to calculate pairwise similarities between synsets using the Wu-Palmer Similarity measure (Wu and Palmer, 1994) and use the resulting similarity matrix as an input to a clustering algorithm. For consistency, we applied the affinity propagation algorithm (Section 4.2).

4.5 Out-of-Domain Clustering

The methods discussed so far build on the assumption that we have general information about target domains, that are otherwise not used to train the model. Namely, we build the categories either knowing the named entity labels occurring in such domains or create the categories such that it is likely that entity labels will have a sensible category to be grouped under.

This is a realistic scenario in cases where target domains are known in advance but no training data is available for them. However, to remove this assumption we assign unique entities for each domain (Appendix F.1) based on the occurrence count of the entity labels in different domains, except for the news domain that has only generic entity types.

For each domain (*evaluation domain*) we remove unique entity labels from the pool of all labels, cluster the remaining ones using the embedding space method described in Section 4.2, then predict which cluster the unique entity labels belong to. This results in one grouping per *evaluation domain* that will be applied only in the out-of-domain evaluation setting. (Section 5.2)

5 Experimental Setup

5.1 Model Implementation

To isolate the effect of our interventions, we base our implementation on the CrossRE project which follows the state-of-the art by Baldini Soares et al. (2019).

In the model, a sentence s containing an ordered pair of entities (e_1, e_2) is augmented by entity span markers e_1^{start}, e_1^{end}, e_2^{start}, and e_2^{end}. We enrich the delimiters (Zhong and Chen, 2021) with the cluster label of the entity. This is different from CrossRE, where the entity label itself is used. For example, following the replacements in the manual grouping `magazine` → `artifact` and `writer` → `person`:

> <E1:artifact> *The New Yorker* </E1:artifact> *wrote that* <E2:person> *Kubrick* </E2:person> *has taken* [. . .]

As described in CrossRE, s is then fed into a pre-trained BERT encoder (Devlin et al., 2019) and the concatenated output representation of the entity start markers become the input to a single-layer feed-forward neural network with softmax activation function for the relation classification.

Since the models were trained on several domains at once, we implemented shuffling at training time to reduce bias and aid with generalization. The training parameters are specified in Appendix H.

5.2 Benchmark

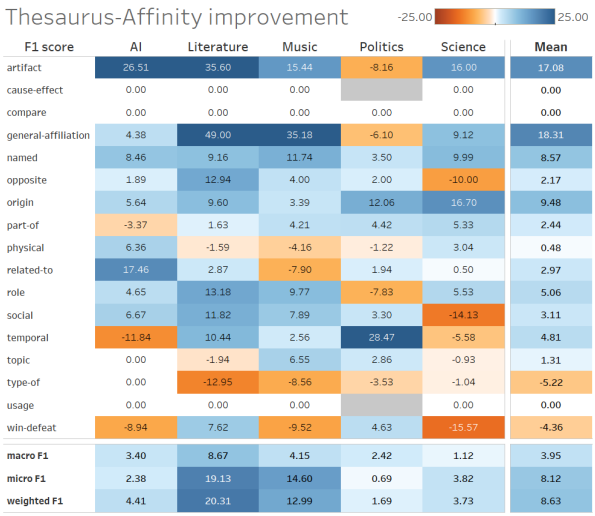
We establish two baseline models that we propose to be used as benchmarks for further research. The first one combines cross-domain fine-tuning with *multi-domain evaluation*. In this case we train the model on the combined training dataset from all six domains and also test it on the combined test corpus. Therefore, it can be considered in-domain evaluation in a cross-domain setting. This setting models the few-shot learning scenario where all target domains are known in advance.

The second benchmark combines cross-domain training on five domains and *out-of-domain evaluation* on the sixth (test / target), done six times, resulting in one baseline for each domain. This mimics the use case where the model might be used on a previously unknown domain with unique entity types, or the target domain lacks training data.

5.3 Evaluation

As baseline models we use our benchmarks as described in Section 5.2 with a *no mapping* scenario, injecting the entity types themselves, not the clusters names. We compare these baselines to all combinations of mapping types and evaluation methods, except for the OOD Clustering, which is not relevant to the multi-domain evaluation setting.

When examining entity labels unique to a domain for OOD Clustering, we noted that some were present in the dataset of other domains. To realize the scenario for which we set up this mapping type (Section 4.5), we removed the sentences containing these entities from the corpus in this particular use case. For comparability, we preserved relevant training and test set sizes by randomly reallocating sentences from the dev set. This affected 0.44% of the total corpus, see Appendix F.3 for the detailed numbers.



F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	26.51	35.60	15.44	-8.16	16.00	17.08
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	4.38	49.00	35.18	-6.10	9.12	18.31
named	8.46	9.16	11.74	3.50	9.99	8.57
opposite	1.89	12.94	4.00	2.00	-10.00	2.17
origin	5.64	9.60	3.39	12.06	16.70	9.48
part-of	-3.37	1.63	4.21	4.42	5.33	2.44
physical	6.36	-1.59	-4.16	-1.22	3.04	0.48
related-to	17.46	2.87	-7.90	1.94	0.50	2.97
role	4.65	13.18	9.77	-7.83	5.53	5.06
social	6.67	11.82	7.89	3.30	-14.13	3.11
temporal	-11.84	10.44	2.56	28.47	-5.58	4.81
topic	0.00	-1.94	6.55	2.86	-0.93	1.31
type-of	0.00	-12.95	-8.56	-3.53	-1.04	-5.22
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	-8.94	7.62	-9.52	4.63	-15.57	-4.36
macro F1	3.40	8.67	4.15	2.42	1.12	3.95
micro F1	2.38	19.13	14.60	0.69	3.82	8.12
weighted F1	4.41	20.31	12.99	1.69	3.73	8.63

Figure 2: Difference in F1 scores for Thesaurus Affinity compared to baseline for each relation type and domain.

6 Results

6.1 OOD Evaluation Results

We chose to exclude the news domain from the OOD evaluation, due to the challenges described by Bassignana and Plank, 2022a, otherwise the highly imbalanced label distribution and many relation types having a support of zero would skew the results. Based on Harbecke et al., 2022 we used weighted F1 as our main metric, however the trends are equal for micro and macro F1. Detailed results can be found in Appendices I through J.

The best performing mapping methods are Manual and Thesaurus Affinity, outperforming the baseline by 7.6 and 8.6 percentage points, respectively. These methods perform significantly better than the baseline in every domain, as can be seen in Table 1, and all mapping methods perform better than the baseline overall. Figure 2 shows the F1 score difference for all relations in each domain for the Thesaurus Affinity compared to the baseline models in the OOD evaluation setting. Absolute results as heatmaps for all domains can be found in Appendix I.1, difference maps in Appendix I.2 and per domain performance charts in Appendix I.3.

From Figure 3 it can be seen that almost all mappings perform better than the baseline in each domain. The only exceptions are OOD Clustering in the AI domain as well as OOD Clustering & Embedding in politics. Specifics of the performance will be discussed further in section 7.1.

6.2 Multi-Domain Evaluation Results

The average weighted F1 scores for the Multi-Domain evaluation can be found in Table 1. It is clear that the performance of the baseline is in the middle of the field, with the leaders Manual and Thesaurus Affinity not far ahead. A breakdown of F1 scores on relation label level can be found in Appendix J, and a summary table for both evaluation methods in Appendix K.

7 Discussion

7.1 OOD Evaluation

In general, all methods suffer from the label distribution not being uniform, as discussed by Bassignana and Plank, 2022a. This leads to relations with F1 scores of zero and makes it difficult to improve performance on these relation types. Future work might benefit from training with higher penalty on less frequent relation labels.

The **OOD Clustering** is the method with the least overlap of information between training and target domain, and thus also the hardest test case to beat. It is notable that it still performs significantly better overall than the baseline in most domains, thus it is the clearest indicator that entity type clustering generalizes better for out-of-domain use cases.

As expected, it performs worse than other grouping methods. This is partially due it being a more challenging subset of the **Embedding** clustering method, which itself performed only slightly better, leaving a significant performance gap compared to other mapping methods. Surprisingly, OOD Clustering performed better than Embedding in the science domain, which has the most unique entities, making it especially hard for the former. We leave detailed analysis of the connection between the number of unique entity types and the performance of OOD Clustering to future work. We also note affinity propagation's tendency to group words by topic, contributing to its middling performance.

Manual mapping, Topological and Thesaurus Affinity methods are based on linguistics and how humans interpret language, which led to the best performance across domains. In **Topological**, WordNet's hierarchical modeling of word senses is aligned with our aim to group entities based on semantic relations. However, commonality expressed through Wu-Palmer similarity appears to be an even better basis for clustering. Since **Thesaurus Affinity** uses the same clustering algorithm as Embedding, the performance difference shows that WordNet is able to capture semantic relations better than the word2vec model and the limiting

Eval. Type	Baseline	Manual	Embedding	OOD Clustering	Topological	Thesaurus Affinity
OOD eval.	43.54	51.15	47.94	46.22	51.03	52.16
Multi-Domain	61.86	62.19	61.15	-	61.38	61.98

Table 1: Weighted F1 scores for each mapping method and evaluation type.

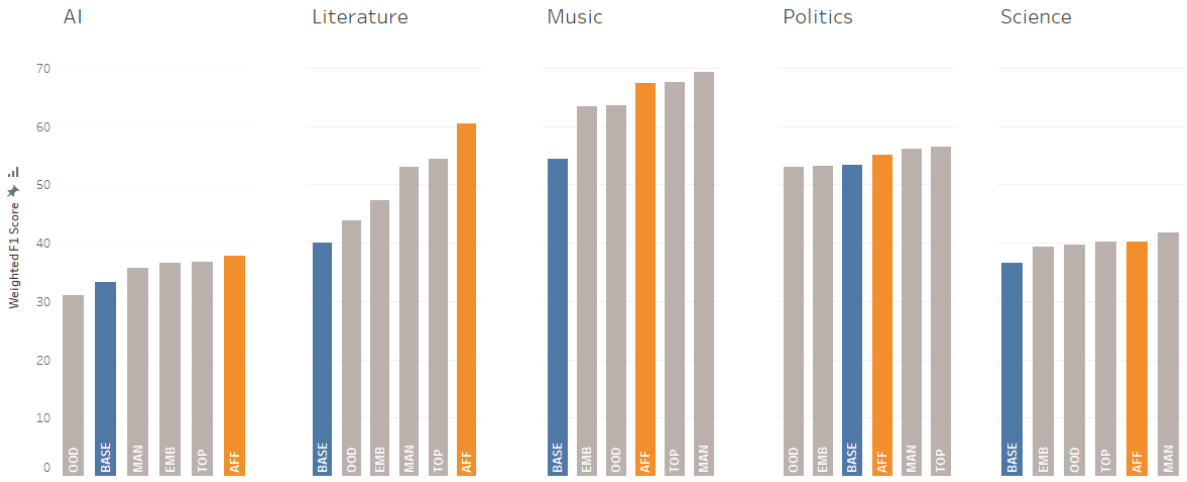


Figure 3: Weighted F1 scores for each mapping method over the five domains in OOD evaluation.

factor is not the algorithm.

In the **Manual** method we grouped entity types on the basis of our understanding of the labels and descriptions, leading to a performance nearly on par with the best result. We theorize however that this method is the most generalizable and with expert knowledge a mapping could be made which would suit numerous domains. Additionally, this does not suffer from the updatability issues of a thesaurus.

7.2 Multi-Domain Evaluation

As expected, all clustering methods perform better in this setting than their OOD counterparts. In Table 1 we can see that the performance is virtually the same across the board.

We find that the results, as displayed in Table 1, makes sense, as domain-specific entity types interact in a way that might not be caught by the less granular categories. Grouping entity types together in categories involves removing information and might therefore not capture the way entity types interact properly.

8 Conclusion

We researched how clustering named entity types in cross-domain relationship classification setups can increase performance and lead to more generalizable models.

We find that significant performance increases can be achieved by using entity mapping in such scenarios, where training data for the target domain is unavailable or the model is used for inference on a previously unseen domain. It is particularly relevant for use cases with unique entity types in the target domain.

On the other hand, it is not relevant in the case of target domain training data being available, as the granularity of the original entity types offer valuable domain-specific information.

We find that the combination of human centered word-understanding using WordNet and clustering (Thesaurus Affinity) was the best performing mapping method in the OOD evaluation setting, but that manually curated groups would create the most generalizable and sustainable mapping.

Limitations

The result are limited to the English language and the source data is mostly generated in the Global North by people with internet access.

The Manual grouping and simplifications are created by us, limiting us to our interpretation of the entities and descriptions.

Detailed analysis of metadata – syntax ambiguity and uncertainty – collected during the annotation process is outside of the scope of this project, leaving it for future work.

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. 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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

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.

Google. 2013. word2vec - Google Code Archive. Last accessed 11 May 2023.

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.

J.A. Lee and M. Verleysen. 2007. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer New York. Section 7.2.

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.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

University Princeton. 2010. About wordnet. `https://wordnet.princeton.edu/`. Accessed: 2023-04-23.

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.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605. Section 4.2.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. 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.

Appendix

A Named Entity and Relation Types

A.1 Named Entity Labels

See Table 2 for the original labels as published in Liu et al. (2021) and labels added by Bassignana and Plank (2022a).

A.2 Semantic Relation Labels

As published in Bassignana and Plank (2022a).

`Part-Of`, `Physical`, `Usage`, `Role`, `Social`, `General-Affiliation`, `Compare`, `Temporal`, `Artifact`, `Origin`, `Topic`, `Opposite`,

Domain	CrossNER	CrossRE (added)
News (Reuters)	person, organization, location, miscellaneous	country
Politics	politician, person, organization, political party, event, election, country, location, miscellaneous	album, book, musical artist, musical instrument, song, university
Natural Science	scientist, person, university, organization, country, location, discipline, enzyme, protein, chemical compound, chemical element, event, astronomical object, academic journal, award, theory, miscellaneous	book
Music	music genre, song, band, album, musical artist, musical instrument, award, event, country, location, organization, person, miscellaneous	magazine
Literature	book, writer, award, poem, event, magazine, person, location, organization, country, miscellaneous	album, band, literary genre, music genre, programming language, scientist, song, university
Artificial Intelligence	field, task, product, algorithm, researcher, metrics, university, country, person, organization, location, miscellaneous	academic journal, conference, event, programming language

Table 2: Named Entity labels used in CrossNER (Liu et al., 2021) and added in CrossRE (Bassignana and Plank, 2022a) per domain.

`Cause-Effect`, `Win-Defeat`, `Type-Of`, `Named`, `Related-To`.

B Manual Groupings

B.1 Groups Obtained from Elisa Bassignana

person: person, researcher, musical artist, politician, scientist, writer

location: location, country

organisation: organisation, university, political party, band

event: event, election, conference

miscellaneous: miscellaneous, program language, metrics, product, field, poem, musical instrument, song, theory, academic journal, chemical element, protein, discipline, enzyme, chemical compound, astronomical object, award, music genre, album, book, magazine, task, algorithm, literary genre

B.2 Proposed Manual Grouping

person: person, researcher, musical artist, politician, scientist, writer

location: location, country

organisation: organisation, university, political party, band

event: event, election, conference, award

artifact: album, song, academic journal, poem, magazine, book

scientific: metrics, enzyme, protein, chemical compound, chemical element, astronomical object

concept: theory, music genre, field, discipline, algorithm, literary genre

brand: product, programming language

miscellaneous: miscellaneous, musical instrument, task

C Embedding Space Based Grouping

Category names were obtained automatically by finding the word closest to the mean of the entity labels in the category in the google-new-300 word2vec embedding space, excluding words containing punctuation and special characters. This categorization is an example, as the result depends on the random state.

essay: academic journal, book, magazine, poem, theory

songs: album, band, music genre, literary genre, musical instrument, song

proteins: algorithm, astronomical object, chemical compound, chemical element, enzyme, metrics, programming language, protein

convention: award, conference, election, event, location, organisation, person, political party

TELW: country, discipline, field, miscellaneous, product, task, university

journalist: musical artist, politician, researcher, scientist, writer

D Topological Grouping

Since one word can belong to several synsets, carrying different meanings, it is possible to disambiguate the meanings (e.g. `java` the programming language, not the island). We theorize that the effect of these substitutions is negligible because of the directed acyclic nature of the `is-a` relationships (Figure 1).

The category names are the synset identifier names on the selected level, with the number removed from the end (except if a name clash were to occur).

object: album, book, musical instrument, academic journal, location, magazine, product

psychological feature: algorithm, discipline, election, event, field, music genre, literary

communication: award, programming language, poem, song

group: band, conference, country, astronomical object, organisation, political party, university

matter: chemical compound, chemical element, enzyme

measure: metrics

assorted: miscellaneous

casual agent: musical artist, person, politician, researcher, scientist, writer

thing: protein

E Thesaurus-Affinity Based Grouping

Category names were obtained the same way as described in Appendix C. This categorization is an example, as the result depends on the random state.

journals: album, book, musical instrument, academic journal, magazine, product

pleiotropy: algorithm, discipline, election, event, field, music genre, literary genre, metrics, miscellaneous, task, theory

poems: award, programming language, poem, song

group: band, conference, country, astronomical object, organisation, political party, university

enzymes: chemical compound, chemical element, enzyme, protein

journalist: location, musical artist, person, politician, researcher, scientist, writer

F Out-of-Domain Clustering

F.1 Unique Entities

Artificial Intelligence: algorithm, conference, field, metrics, product, researcher, task, programming language

Literature: poem, writer, literary genre, book, magazine

Music: band, song, musical artist, music genre, album, musical instrument

Politics: political party, election, politician

Science: chemical compound, chemical element, astronomical object, discipline, enzyme, protein, theory, scientist, academic journal

F.2 OOD Clustering Example

An example for OOD clustering with Artificial Intelligence as evaluation domain. The predicted entities (unique to AI) are italicized. Category names were obtained the same way as described in Appendix C. This categorization is an example, as the result depends on the random state.

songs: album, band, music genre, literary genre, poem, song, musical instrument

group: award, political party, event, location, organisation, person, *conference*, *task*

novelist: book, election, academic journal, magazine, musical artist, politician, scientist, writer, *researcher*

proteins: chemical compound, chemical element, enzyme, astronomical object, miscellaneous, protein, *algorithm*, *programming language*

disciplines: country, discipline, theory, university, *field*, *metrics*, *product*

F.3 Sentences removed in OOD Clustering training

Domain	Removed	%
AI	3	0.34
Literature	10	1.09
Music	1	0.12
News	0	0
Politics	2	0.24
Science	7	0.82
Total	23	0.44

Table 3: Number and percentage of sentences removed in OOD Clustering training.

G Substitutions for clustering

Named entity label substitutions. Substitutions specific to Embedding Space Based Grouping and Out-of-Domain Clustering are marked with `word2vec`, while those specific to Topological Grouping and Thesaurus-Affinity Based Grouping with `WordNet`.

Original	Substitute
musical artist	musician
organisation	organization
political party	coalition (`word2vec`), party (`WordNet`)
academic journal	journal
chemical compound	chemical
chemical element	chemical
astronomical object	galaxy
music genre	genre
literary genre	genre
programming language	javascript (`word2vec`), java (`WordNet`)
musical instrument	violin (`word2vec`), instrument (`WordNet`)
misc	miscellaneous

Table 4: Named entity label substitutions.

H Reproducibility

The hyperparameters are reported in Table 5. They are used across all mapping types & the baseline, and across both evaluation types. The early stopping parameter stops the training if there are no improvements on the development data for three consecutive epochs. The training were divided between two machines with the following specifications:

- Nvidia® GeForce™ RTX 3090 Ti 24 GB GPU; Intel® Core i9-9900K CPU
- Nvidia® GeForce™ GTX 1660 Super 6 GB GPU; Intel® Core i5-11400F CPU

Parameter	Value
General	
Random Seeds	4012, 5096, 8857, 8878, 9908
Clustering	
PCA `n_components`	35[5]
UMAP `n_components`	3
distance metric	`cosine`
UMAP `n_neighbor`	4
UMAP `min_dist`	0.3
Affinity Prop. damping factor	0.5
Model	
Encode	`bert-base-cased`
Classifier	1-layer FFNN
Loss	Cross Entropy
Optimizer	Adam
Epochs	50
Batch Size	32
Learning Rate	$2e^{-5}$
Early Stop	True
Shuffle Data	True

Table 5: Hyperparameter settings.

I OOD Evaluation Result Figures

I.1 Heatmap over relation and domain

Gray colored fields indicate a case where the support is zero for a relation in the test set.

[5]When the number of entity types in the OOD Clustering case is below 35, then that number is used.

No mapping - Baseline

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	31.65	41.25	64.57	54.10	59.23	50.16
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	14.17	28.87	47.10	44.24	28.08	32.49
named	61.03	49.29	49.48	65.81	49.42	55.01
opposite	11.87	8.79	0.00	3.21	10.00	6.77
origin	29.09	13.98	10.39	9.56	18.26	16.25
part-of	37.74	21.00	15.40	23.67	11.53	21.87
physical	77.28	80.72	86.92	77.20	75.40	79.50
related-to	20.15	7.93	8.53	3.60	33.11	14.66
role	57.21	46.61	52.71	64.50	45.66	53.34
social	0.00	1.60	5.20	10.04	19.06	7.18
temporal	21.21	56.97	48.77	43.01	63.19	46.63
topic	0.00	5.14	10.67	0.00	3.74	3.91
type-of	0.00	76.00	71.63	28.43	5.71	36.35
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	49.76	61.80	77.81	0.00	84.20	54.71
macro F1	24.19	29.41	32.30	28.49	29.80	28.84
micro F1	39.63	43.00	54.56	57.34	38.91	46.69
weighted F1	33.30	40.07	54.40	53.40	36.54	43.54

Figure 4: Baseline Heatmap - Average F1 score

Manual mapping

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	39.11	71.04	82.71	59.49	74.80	65.43
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	16.28	54.08	82.17	48.15	35.50	47.24
named	62.87	53.29	62.12	68.92	48.55	59.15
opposite	18.06	9.99	0.00	7.15	2.67	7.57
origin	33.45	26.35	14.68	20.10	36.85	26.28
part-of	31.22	15.36	15.60	23.48	18.04	20.74
physical	82.66	82.44	89.28	77.02	74.76	81.23
related-to	33.59	5.60	2.20	5.20	36.45	16.61
role	62.07	57.72	66.95	61.12	58.84	61.34
social	0.00	2.47	8.21	16.97	35.93	12.72
temporal	41.53	67.49	47.59	65.80	68.74	58.23
topic	0.00	1.48	6.30	2.35	8.70	3.77
type-of	0.00	63.10	63.85	27.80	3.58	31.67
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	46.23	64.97	70.87	4.40	79.83	53.26
macro F1	27.48	33.85	36.03	32.53	34.31	32.84
micro F1	40.10	56.49	71.48	59.63	44.46	54.43
weighted F1	35.63	52.92	69.24	56.16	41.78	51.15

Figure 5: Manual Heatmap - Average F1 score

Embedding mapping

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	40.52	63.67	67.79	30.30	65.21	53.50
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	11.74	38.12	75.93	37.42	31.66	38.98
named	68.07	57.30	59.91	62.62	56.56	60.89
opposite	16.43	29.25	0.00	9.17	10.00	12.97
origin	29.47	17.98	7.60	14.71	28.52	19.66
part-of	35.82	15.71	17.50	16.50	16.32	20.37
physical	82.85	76.39	82.34	74.69	76.92	78.64
related-to	36.43	9.62	5.24	2.15	36.68	18.02
role	59.62	53.49	59.63	59.11	49.62	56.30
social	0.00	2.96	19.73	3.64	6.67	6.60
temporal	40.70	64.39	38.59	62.88	68.97	55.11
topic	0.00	3.58	0.00	2.67	3.69	1.99
type-of	0.00	63.82	64.42	23.93	6.51	31.74
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	45.99	62.86	68.93	5.38	67.97	50.23
macro F1	27.51	32.89	33.39	27.01	30.90	30.34
micro F1	41.14	49.90	65.28	56.49	42.68	51.10
weighted F1	36.57	47.27	63.42	53.18	39.24	47.94

Figure 6: Embedding Heatmap - Average F1 score

OOD Clustering

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	35.17	51.01	71.61	17.80	73.43	49.80
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	12.07	39.30	77.92	41.70	31.88	40.58
named	61.61	52.38	56.25	64.71	53.39	57.67
opposite	8.38	10.58	0.00	8.21	0.00	5.44
origin	24.59	13.56	11.24	13.81	37.00	20.04
part-of	27.04	18.08	5.63	22.07	17.50	18.06
physical	76.84	74.45	77.73	73.01	75.50	75.51
related-to	29.33	6.58	2.09	2.82	36.90	15.54
role	48.50	52.38	65.23	55.17	51.77	54.61
social	0.00	0.83	11.99	3.56	20.39	7.35
temporal	28.79	55.68	42.06	71.56	64.97	52.61
topic	0.00	0.00	14.91	0.00	3.57	3.69
type-of	0.00	67.66	66.13	23.63	6.23	32.73
usage	0.00	0.00	0.00		1.18	0.29
win-defeat	34.30	60.47	67.86	1.90	60.84	45.07
macro F1	22.74	29.59	33.57	26.66	31.44	28.80
micro F1	36.98	44.91	65.69	57.01	42.72	49.46
weighted F1	31.09	43.75	63.56	52.94	39.73	46.22

Figure 7: OOD Clustering Heatmap - Average F1 score

Topological mapping

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	57.49	73.21	82.58	48.82	67.89	66.00
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	17.50	60.00	80.50	40.47	30.34	45.76
named	69.03	54.07	63.87	71.85	52.41	62.25
opposite	15.64	12.15	0.00	5.53	0.00	6.66
origin	30.86	27.60	17.47	22.68	37.66	27.25
part-of	39.09	16.70	20.02	24.29	11.68	22.36
physical	82.35	76.32	83.59	77.05	76.09	79.08
related-to	24.58	10.79	2.99	3.91	35.89	15.63
role	59.87	56.45	63.59	58.80	52.08	58.16
social	0.00	3.17	4.56	17.62	35.41	12.15
temporal	14.81	69.94	51.13	73.36	61.40	54.13
topic	0.00	0.00	5.71	5.21	6.49	3.48
type-of	0.00	72.16	63.31	32.17	4.29	34.39
usage	0.00	0.00	0.00		4.29	1.07
win-defeat	44.75	67.04	67.39	2.79	83.81	53.16
macro F1	26.82	35.27	35.69	32.30	32.92	32.60
micro F1	41.61	56.97	68.79	59.64	42.72	53.95
weighted F1	36.76	54.32	67.47	56.48	40.12	51.03

Figure 8: Topological Heatmap - Average F1 score

Thesaurus-Affinity mapping

0.00 100.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	58.16	76.85	80.00	45.94	75.23	67.24
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	18.56	77.86	82.28	38.13	37.19	50.81
named	69.49	58.45	61.22	69.31	59.41	63.58
opposite	13.75	21.73	4.00	5.21	0.00	8.94
origin	34.72	23.57	13.78	21.63	34.95	25.73
part-of	34.36	22.63	19.61	28.09	16.85	24.31
physical	83.64	79.13	82.75	75.98	78.44	79.99
related-to	37.60	10.80	0.63	5.54	33.61	17.64
role	61.86	59.78	62.49	56.67	51.19	58.40
social	6.67	13.42	13.09	13.34	4.93	10.29
temporal	9.37	67.41	51.33	71.48	57.61	51.44
topic	0.00	3.21	17.22	2.86	2.82	5.22
type-of	0.00	63.04	63.07	24.90	4.67	31.14
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	40.82	69.41	68.29	4.63	68.63	50.36
macro F1	27.59	38.08	36.46	30.91	30.91	32.79
micro F1	42.01	62.13	69.15	58.02	42.74	54.81
weighted F1	37.71	60.38	67.38	55.09	40.26	52.16

Figure 9: Thesaurus Affinity Heatmap - Average F1 score

I.2 Heatmap over relation and domain - difference from baseline

Gray colored fields indicate a case where the support is zero for a relation in the test set.

Manual improvement

-25.00 25.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	7.46	29.79	18.14	5.39	15.57	15.27
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	2.11	25.21	35.07	3.91	7.42	14.75
named	1.85	4.00	12.63	3.11	-0.88	4.14
opposite	6.19	1.21	0.00	3.94	-7.33	0.80
origin	4.36	12.37	4.29	10.53	18.59	10.03
part-of	-6.52	-5.64	0.20	-0.19	6.51	-1.13
physical	5.38	1.72	2.36	-0.18	-0.64	1.73
related-to	13.44	-2.33	-6.33	1.61	3.34	1.95
role	4.86	11.11	14.24	-3.38	13.18	8.00
social	0.00	0.87	3.01	6.93	16.88	5.54
temporal	20.32	10.52	-1.18	22.79	5.55	11.60
topic	0.00	-3.66	-4.36	2.35	4.96	-0.14
type-of	0.00	-12.89	-7.77	-0.63	-2.13	-4.69
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	-3.53	3.17	-6.94	4.40	-4.37	-1.46
macro F1	3.29	4.44	3.73	4.04	4.51	4.00
micro F1	0.47	13.49	16.92	2.30	5.54	7.74
weighted F1	2.33	12.85	14.84	2.76	5.24	7.61

Figure 10: Manual Heatmap - difference from baseline

Embedding improvement

-25.00 25.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	8.87	22.42	3.22	-23.79	5.98	3.34
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	-2.43	9.25	28.83	-6.82	3.59	6.48
named	7.04	8.01	10.42	-3.19	7.14	5.88
opposite	4.56	20.46	0.00	5.96	0.00	6.20
origin	0.38	4.00	-2.79	5.15	10.27	3.40
part-of	-1.92	-5.30	2.10	-7.17	4.79	-1.50
physical	5.57	-4.34	-4.57	-2.51	1.52	-0.87
related-to	16.29	1.69	-3.29	-1.45	3.57	3.36
role	2.42	6.89	6.91	-5.39	3.96	2.96
social	0.00	1.36	14.53	-6.40	-12.39	-0.58
temporal	19.49	7.42	-10.18	19.87	5.78	8.48
topic	0.00	-1.56	-10.67	2.67	-0.05	-1.92
type-of	0.00	-12.18	-7.20	-4.50	0.80	-4.62
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	-3.77	1.06	-8.88	5.38	-16.23	-4.49
macro F1	3.32	3.48	1.08	-1.48	1.10	1.50
micro F1	1.50	6.90	10.73	-0.85	3.76	4.41
weighted F1	3.27	7.21	9.02	-0.21	2.70	4.40

Figure 11: Embedding Heatmap - difference from baseline

OOD Clustering improvement

-25.00 25.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	3.52	9.76	7.04	-36.30	14.20	-0.36
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	-2.10	10.43	30.82	-2.53	3.80	8.08
named	0.58	3.09	6.76	-1.10	3.97	2.66
opposite	-3.49	1.80	0.00	5.00	-10.00	-1.34
origin	-4.50	-0.42	0.85	4.25	18.74	3.78
part-of	-10.70	-2.93	-9.77	-1.60	5.97	-3.81
physical	-0.44	-6.27	-9.19	-4.19	0.10	-4.00
related-to	9.18	-1.36	-6.43	-0.77	3.80	0.88
role	-8.71	5.78	12.52	-9.33	6.11	1.28
social	0.00	-0.77	6.80	-6.48	1.33	0.18
temporal	7.59	-1.30	-6.71	28.55	1.77	5.98
topic	0.00	-5.14	4.24	0.00	-0.18	-0.22
type-of	0.00	-8.33	-5.50	-4.80	0.52	-3.62
usage	0.00	0.00	0.00		1.18	0.29
win-defeat	-15.46	-1.33	-9.95	1.90	-23.36	-9.64
macro F1	-1.44	0.18	1.26	-1.83	1.64	-0.04
micro F1	-2.66	1.91	11.13	-0.32	3.80	2.77
weighted F1	-2.21	3.68	9.17	-0.46	3.20	2.68

Figure 12: OOD Clustering Heatmap - difference from baseline

Topological improvement

-25.00 25.00

F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	25.84	31.96	18.02	-5.28	8.65	15.84
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	3.32	31.13	33.40	-3.77	2.26	13.27
named	8.00	4.78	14.39	6.05	2.99	7.24
opposite	3.77	3.36	0.00	2.32	-10.00	-0.11
origin	1.77	13.63	7.07	13.12	19.41	11.00
part-of	1.36	-4.31	4.62	0.62	0.16	0.49
physical	5.07	-4.40	-3.32	-0.15	0.69	-0.42
related-to	4.43	2.86	-5.54	0.32	2.78	0.97
role	2.66	9.84	10.87	-5.70	6.42	4.82
social	0.00	1.57	-0.63	7.59	16.35	4.97
temporal	-6.40	12.97	2.36	30.35	-1.79	7.50
topic	0.00	-5.14	-4.95	5.21	2.75	-0.43
type-of	0.00	-3.83	-8.31	3.74	-1.42	-1.97
usage	0.00	0.00	0.00		4.29	1.07
win-defeat	-5.01	5.24	-10.42	2.79	-0.39	-1.56
macro F1	2.64	5.86	3.39	3.81	3.13	3.76
micro F1	1.97	13.97	14.24	2.31	3.81	7.26
weighted F1	3.46	14.26	13.07	3.08	3.59	7.49

Figure 13: Topological Heatmap - difference from baseline

Thesaurus-Affinity improvement

-25.00 25.00

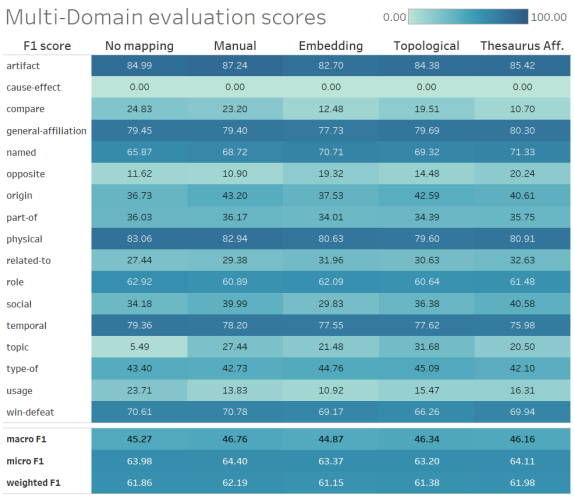
F1 score	AI	Literature	Music	Politics	Science	Mean
artifact	26.51	35.60	15.44	-8.16	16.00	17.08
cause-effect	0.00	0.00	0.00		0.00	0.00
compare	0.00	0.00	0.00	0.00	0.00	0.00
general-affiliation	4.38	49.00	35.18	-6.10	9.12	18.31
named	8.46	9.16	11.74	3.50	9.99	8.57
opposite	1.89	12.94	4.00	2.00	-10.00	2.17
origin	5.64	9.60	3.39	12.06	16.70	9.48
part-of	-3.37	1.63	4.21	4.42	5.33	2.44
physical	6.36	-1.59	-4.16	-1.22	3.04	0.48
related-to	17.46	2.87	-7.90	1.94	0.50	2.97
role	4.65	13.18	9.77	-7.83	5.53	5.06
social	6.67	11.82	7.89	3.30	-14.13	3.11
temporal	-11.84	10.44	2.56	28.47	-5.58	4.81
topic	0.00	-1.94	6.55	2.86	-0.93	1.31
type-of	0.00	-12.95	-8.56	-3.53	-1.04	-5.22
usage	0.00	0.00	0.00		0.00	0.00
win-defeat	-8.94	7.62	-9.52	4.63	-15.57	-4.36
macro F1	3.40	8.67	4.15	2.42	1.12	3.95
micro F1	2.38	19.13	14.60	0.69	3.82	8.12
weighted F1	4.41	20.31	12.99	1.69	3.73	8.63

Figure 14: Thesaurus Affinity Heatmap - difference from baseline

I.3 Micro & Macro F1 Results Across Domains

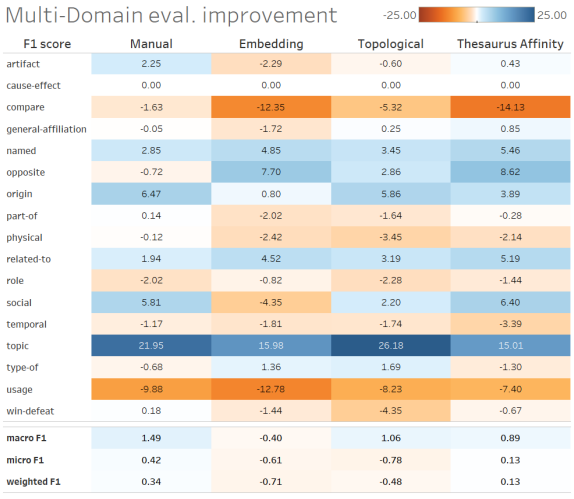
See Figure 15 for macro and Figure 16 for micro F1 scores for each mapping method over the five domains in OOD evaluation. The Baseline (no mapping) is colored blue, and the overall best model, Thesaurus Affinity, is colored orange.

J Multi-Domain Result Figures



F1 score	No mapping	Manual	Embedding	Topological	Thesaurus Aff.
artifact	84.99	87.24	82.70	84.38	85.42
cause-effect	0.00	0.00	0.00	0.00	0.00
compare	24.83	23.20	12.48	19.51	10.70
general-affiliation	79.45	79.40	77.73	79.69	80.30
named	65.87	68.72	70.71	69.32	71.33
opposite	11.62	10.90	19.32	14.48	20.24
origin	36.73	43.20	37.53	42.59	40.61
part-of	36.03	36.17	34.01	34.39	35.75
physical	83.06	82.94	80.63	79.60	80.91
related-to	27.44	29.38	31.96	30.63	32.63
role	62.92	60.89	62.09	60.64	61.48
social	34.18	39.99	29.83	36.38	40.58
temporal	79.36	78.20	77.55	77.62	75.98
topic	5.49	27.44	21.48	31.68	20.50
type-of	43.40	42.73	44.76	45.09	42.10
usage	23.71	13.83	10.92	15.47	16.31
win-defeat	70.61	70.78	69.17	66.26	69.94
macro F1	45.27	46.76	44.87	46.34	46.16
micro F1	63.98	64.40	63.37	63.20	64.11
weighted F1	61.86	62.19	61.15	61.38	61.98

Figure 17: Heatmap over mapping-type and relations - F1 score



F1 score	Manual	Embedding	Topological	Thesaurus Affinity
artifact	2.25	-2.29	-0.60	0.43
cause-effect	0.00	0.00	0.00	0.00
compare	-1.63	-12.35	-5.32	-14.13
general-affiliation	-0.05	-1.72	0.25	0.85
named	2.85	4.85	3.45	5.46
opposite	-0.72	7.70	2.86	8.62
origin	6.47	0.80	5.86	3.89
part-of	0.14	-2.02	-1.64	-0.28
physical	-0.12	-2.42	-3.45	-2.14
related-to	1.94	4.52	3.19	5.19
role	-2.02	-0.82	-2.28	-1.44
social	5.81	-4.35	2.20	6.40
temporal	-1.17	-1.81	-1.74	-3.39
topic	21.95	15.98	26.18	15.01
type-of	-0.68	1.36	1.69	-1.30
usage	-9.88	-12.78	-8.23	-7.40
win-defeat	0.18	-1.44	-4.35	-0.67
macro F1	1.49	-0.40	1.06	0.89
micro F1	0.42	-0.61	-0.78	0.13
weighted F1	0.34	-0.71	-0.48	0.13

Figure 18: Heatmap over mapping-type and relations - F1 score gain over Baseline

K Results Summary

A summary of micro, macro, and weighted F1 scores for each mapping method and evaluation type can be found in Table 6.

L Group Contributions

The overall workload was equally distributed across group members. Throughout the project we have worked both in conjunction and individually, thus all parts of the project has been done and/or reviewed by all group members.

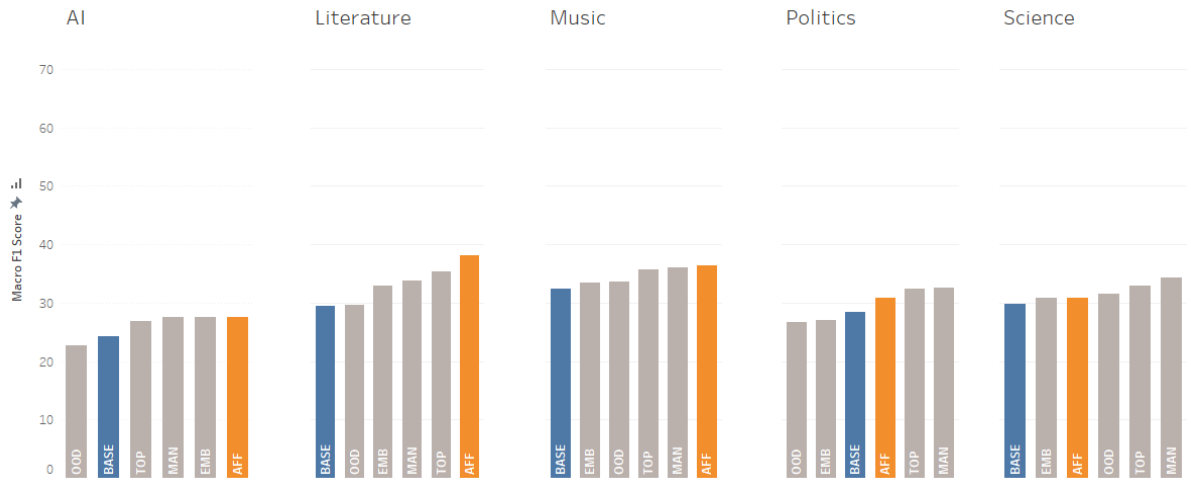


Figure 15: **Macro F1** scores for each mapping method over the five domains in OOD evaluation.

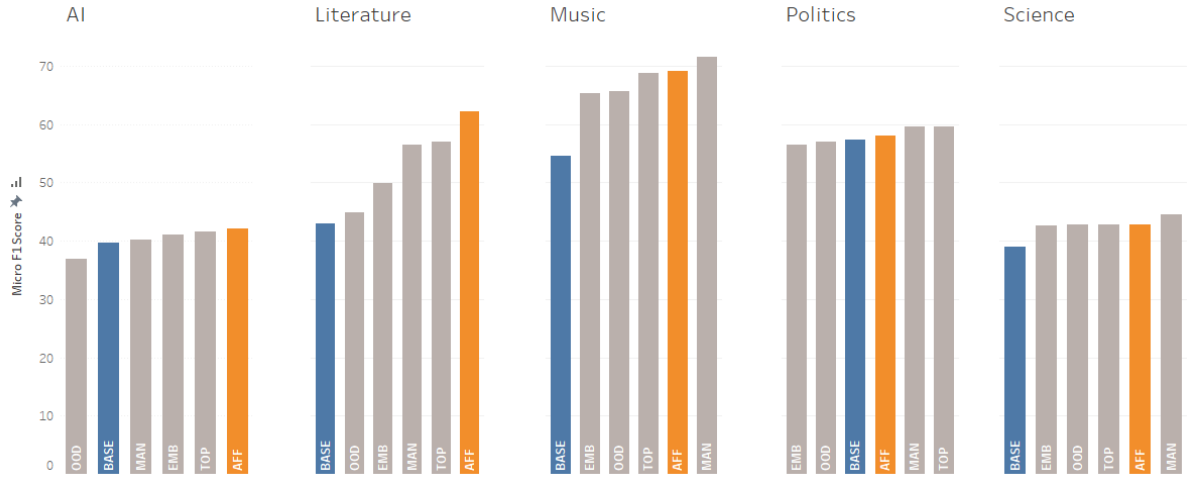


Figure 16: **Micro F1** scores for each mapping method over the five domains in OOD evaluation.

Eval. Type	Metric	Baseline	Manual	Embedding	OOD Clustering	Topological	Thesaurus Affinity
OOD	Weighted F1	43.54	51.15	47.94	46.22	51.03	52.16
	Macro F1	28.84	32.84	30.34	28.80	32.60	32.79
	Micro F1	46.69	54.43	51.10	49.46	53.95	54.81
Multi-Domain	Weighted F1	61.86	62.19	61.15	-	61.38	61.98
	Macro F1	45.27	46.76	44.87	-	46.34	46.16
	Micro F1	63.98	64.40	63.37	-	63.20	64.11

Table 6: F1 scores for each mapping method and evaluation type.