Towards Realistic Few-Shot Relation Extraction

Sam Brody, Sichao Wu, Adrian Benton
Bloomberg
731 Lexington Ave
New York, NY 10022 USA
{sbrody18,swu389,abenton10}@bloomberg.net

Abstract

In recent years, few-shot models have been applied successfully to a variety of NLP tasks. Han et al. (2018) introduced a few-shot learning framework for relation classification, and since then, several models have surpassed human performance on this task, leading to the impression that few-shot relation classification is solved. In this paper we take a deeper look at the efficacy of strong few-shot *classification* models in the more common relation *extraction* setting, and show that typical few-shot evaluation metrics obscure a wide variability in performance across relations. In particular, we find that state of the art few-shot relation classification models overly rely on entity type information, and propose modifications to the training routine to encourage models to better discriminate between relations involving similar entity types.

1 Introduction

Few-shot approaches have been explored in a variety of natural language processing (NLP) tasks, such as machine translation (Gu et al., 2018) and textual entailment (Yin et al., 2020), as well as an assortment of classification and natural language inference tasks (Yan et al., 2018; Brown et al., 2020b). The introduction of large, pretrained transformer language models in NLP increased the promise of building systems that can perform complex NLP tasks from only a small number of training examples (Brown et al., 2020a). Han et al. (2018) introduced a few-shot learning framework for relation classification (FewRel), and recently several systems have achieved near-human performance on this task – in some settings even exceeding it.[1] These stunning results might give the impression that FewRel has been solved, and that these systems can be used to extract any relation of interest from a collection of text (e.g., to populate a knowledge-base from web documents) with only a few example instances.

In this paper, we take a deeper look at the applicability of high performing few-shot relation classification models in a more realistic relation extraction (RE) setting. We find that although transformer-based models achieve high accuracy at the FewRel task, this obscures a wide variability in performance across relation types. In particular, we find that state-of-the-art FewRel models heavily rely on entity type information, and thus are unable to discriminate between many types of relations that are trivial for humans (e.g., spouse-of vs. child-of). However, we find that enriching the training data with relations with similar entity types forces the model to attend less to entity type information, and in a ranking evaluation, improves performance on unseen relations by up to 24% precision at 50, absolute.

2 Few-shot Relation Extraction

2.1 FewRel 1.0

The FewRel challenge (Han et al., 2018) introduced the few-shot paradigm to relation classification. The authors provided a dataset and evaluation framework for measuring performance on the task. They also adapted several few-shot text classification systems to relation classification. The dataset was comprised of 700 example instances from Wikipedia for each of 100 relations. Of these, 64 were designated for training, 16 for validation, and 20 were withheld for testing. The evaluation was structured as follows: for each instance, N relations were chosen from the pool and, for each relation, K support examples were sampled. These examples were the only information provided to the model regarding these relations. In addition, Q query examples were selected from each of the N relations, and the task of the model was to decide the correct label for each query from among N possible answers (N-class classification). This pro-

[1] thunlp.github.io/1/fewrel1.html

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5338–5345
November 7–11, 2021. ©2021 Association for Computational Linguistics

cess was repeated a fixed number of times (1,000 by default) and the results were averaged.

2.2 State of the Art (SOTA)

Although multiple representation and modeling strategies were proposed by the authors and challenge participants, the authors' most successful CNN-based system, as well as the current top performing system (Baldini Soares et al., 2019) employed a simple prototype approach. In this approach, each example was encoded as a vector, and the prototype representation for each relation was derived as the average across all exemplar vectors. The representation of a query example was then compared to the prototypical representations of the candidate relations, and the most similar one (by inner product) was chosen as the label.

2.3 FewRel 2.0

Two main concerns were raised about the FewRel 1.0 setup: the cross-domain applicability of the models, and their performance in the none-of-the-above (NOTA) scenario, where a portion of the query instances may belong to none of the candidate relations. FewRel 2.0 (Gao et al., 2019) was designed to address both of these concerns. Our work addresses orthogonal issues, and we limit ourselves to the FewRel 1.0 dataset.

3 Performance on Relation Extraction

3.1 Experimental Setup

Our first set of experiments was exploratory. We set out to reproduce the results of several high-scoring models on the FewRel 1.0 dataset, and to examine how they performed in a setting that more closely reflected a few-shot RE use case, where the model is expected to extract instances of a small number of relations from a large corpus, with high precision and recall. We refer to this scenario as "Realistic RE" because it is common in applied RE tasks such as knowledge-base population (KBP) and information extraction. Since prototypical models offered simplicity and best-in-class performance, we focused on those.

In addition to the CNN and BERT (Devlin et al., 2019) models provided by the FewRel framework, we experimented with other extensions of BERT. These included RoBERTa (Liu et al., 2019), SpanBert (Joshi et al., 2020), and LUKE (Peters et al., 2018). We were unable to evaluate the best performing model on the task (Baldini Soares et al., 2019) since the authors have not made this model or training code public.

Encoder	**Acc. ENT**	**Acc. CLS**
CNN	85.24%	83.14%
BERT	93.78%	92.21%
SpanBERT	95.19%	92.73%
RoBERTa-base	95.18%	93.27%
RoBERTa-large	96.23%	93.03%
LUKE-base	95.08%	95.03%

Table 1: Accuracy of trained models on FewRel Wikipedia validation set ($N = 5$, $K = 5$).

For representation of individual relations, we followed two common strategies mentioned in the RE literature and implemented in the FewRel codebase. In both strategies, the subject and object entities in the sentence are enclosed in special tokens. In the first strategy (*CLS*) the encoding of the `[CLS]` token in the subword-tokenized sentence is used as its representation. In the second strategy (*ENT*), the encoding of the two special tokens preceding the entities are concatenated to form the example representation.

We averaged the representations of the examples to create a prototype representation for each relation, and computed similarity between prototype and query by inner product. All models were trained with the default set of hyperparameters provided in the FewRel repository, without any tuning. Since the official test set is not publicly available, we report validation set performance. This set was not used during training or tuning.

For evaluation that more closely mirrors the realistic few-shot RE scenario, we used a total ranking setting. Instead of randomly sampling K examples of N relations for each instance, we compiled a single test set containing 50 instances of each relation. We then evaluated performance on this test set on a relation-by-relation basis. For each relation, we sampled $K = 5$ examples and created a prototypical representation. We then ranked the entire test set by similarity to the prototype representation, and calculated precision at 50 (P@50) for this ranking (Järvelin and Kekäläinen, 2017).[2] We repeated this process 10 times, and reported mean P@50[3]. All models were trained using a single V100 GPU.

3.2 Results

Performance at FewRel and RE In Table 1 we can see the performance of several models in the

[2] We used P@50 in this setting, rather than accuracy, since the latter requires a similarity threshold for classification, and choosing (or tuning) that threshold might affect the results.

[3] Our code is available at `https://github.com/bloomberg/emnlp2021_fewrel`.

Relation	**RoBERTa ENT**	**RoBERTa CLS**	**Span ENT**	**Span CLS**	**BERT ENT**	**BERT CLS**
P59 (celestial sphere)	99.4	100	98.2	93.0	98.0	99.8
P364 (language of work)	98.2	98.0	97.8	97.0	100	98.0
P410 (military rank)	98.0	97.8	99.6	99.0	98.4	97.2
P155 (prior item)	97.0	91.6	92.6	93.0	86.6	88.2
P412 (voice type)	96.8	97.4	95.8	98.0	100	96.6
P2094 (sport classification)	95.6	95.2	98.0	94.4	97.2	96.2
P413 (player position)	92.6	86.2	90.6	94.6	89.8	94.0
P177 (obstacle crossed)	88.4	92.4	94.4	85.0	89.6	85.8
P25 (mother)	84.2	44.2	76.0	68.0	81.6	50.2
P921 (primary topic)	79.8	82.4	77.6	74.6	69.4	74.6
P40 (child)	75.6	59.0	75.4	73.6	66.6	42.2
P641 (sport)	70.2	66.2	64.0	62.8	64.0	75.6
P463 (organization)	69.8	79.4	80.0	65.6	75.8	62.0
P206 (in/near body of water)	51.8	67.4	61.6	53.4	51.0	61.4
P26 (spouse)	50.8	49.0	48.6	39.4	44.6	49.8
P361 (part of)	36.2	31.0	38.0	35.8	25.0	38.4

Table 2: P@50 for individual relations (Rob/Span/Bert: RoBERTa-base/SpanBERT/BERT-base). See Appendix A for full descriptions of these relations.

FewRel 5-way-5-shot setup. We chose these values for K and N since they were the most similar to the few-shot-RE scenario described above. All models were trained on the FewRel training set for 10,000 iterations and with default parameters.

As seen in the table, performance increases with model complexity and size, with a large gap between the CNN model and the transformer-based models, and smaller gaps between the latter group. Entity representations consistently outperform sentence-level, `[CLS]` token representations across model classes, suggesting that entity representations provide a powerful signal for FewRel models.

Table 2 shows P@50 model performance in the RE setting. Performance varies widely between relations and, for many relations, is much lower than one would expect from the numbers in Table 1. We see that the order between models shown in Table 1 is not maintained on individual relations. There is, however, a rough ordering of difficulty among relations, with all models achieving $> 85\%$ precision on the top half of relations.

For the rest of the paper, we report results from a single model, RoBERTa-base with ENT representation, due to the tractable size of the model and its strong overall performance.

Difficult Relations Table 3 displays the two most frequent confounders for the RoBERTa-base ENT representation for the least accurate relations. We can see that on these "hard" relations the model gets confused with one or two primary confounders a significant portion of the time.

Relation	**P@50**	**Top 2 Confounders**	
P25 (mother)	84.2	P26 (12.8)	P40 (3)
P921 (main topic)	79.8	P361 (11.2)	P641 (3)
P40 (child)	75.6	P26 (23.8)	P25 (0.6)
P641 (sport played)	70.2	P413 (14)	P2094 (13)
P463 (member org)	69.8	P361 (17.8)	P59 (5.2)
P206 (nearby water)	51.8	P177 (48.2)	–
P26 (spouse)	50.8	P25 (41.2)	P40 (8)
P361 (part of)	36.2	P59 (29.6)	P463 (25.8)

Table 3: P@50 for individual relations with top-2 confounders (and their percentage) for ENT RoBERTa-base.

For many of these relations, the confusion is justified, and probably does not represent actual false-positives. For example, when extracting relation P206 (the relation between a location and a nearby body of water), the model selects instances of P177 (the relation between a road or bridge and the natural obstacle is crosses) as positives. The most difficult relation, P361 ("part of"), is confused with instances of P463 ("member of organization") and P59 ("star's constellation"), which are both subtypes of that P361. These types of errors are also likely to be made by humans, and may account for the somewhat low human performance on the FewRel task.

However there is one group of relations: *mother*, *child*, *spouse*, which are easily distinguished by humans, but are confused by the models. These relations all share similar entity type signatures – both entities are people. Since several recent papers (see Section 5) demonstrate that supervised RE models rely heavily on entity type information, we hypothesize that few-shot models do the same.

To test this hypothesis, we evaluated the models on TACRED data. In this evaluation, family relations are confused, as are other groups of relations which share a similar type signature. Table 4

Rel. types	Relation	P@50	Top 2 Confounders	
per:family	children	72.0	other_family (14.0)	spouse (8.4)
	siblings	69.6	other_family. (16.0)	children (11.0)
	parents	62.2	spouse (20.0)	other_family. (8.6)
	spouse	43.2	children (31.0)	parents (9.8)
	other_family	23.6	spouse (21.2)	siblings (15.0)
per: date	birth_date	77.2	death_date (22.8)	–
	death_date	62.4	birth_date (35.6)	cause_of_death (1.4)
per:location	birth_city	52.8	birth_state (13.2)	death_city (11.4)
	birth_state	41.4	birth_city (20.2)	death_state (15.4)
	death_state	35.8	birth_city (18.0)	death_city (14.4)
	death_city	35.6	birth_city (25.0)	death_state(16.8)
org:date	founded_date	72.6	dissolved_date (27.2)	alt_name (0.2)
	dissolved_date	50.6	founded_date (48.2)	alt_name (0.3)
org:location	hq_state	68.2	hq_city (30.8)	member_of (0.8)
	hq_city	59.6	hq_state (40.0)	member_of (0.4)

Table 4: TACRED: P@50 for a subset of relations along with top-2 confounders.

shows the results for a subset of these groups. See Appendix C for the full confusion matrix for all TACRED person and organization relations. This confirms that even high-scoring few-shot RE models rely primarily on entity type information, and find it hard to distinguish between relations with similar type signatures.

4 Overcoming Entity Type Bias

4.1 Alternate Representations

Under the hypothesis that the choice of the *ENT* representation strategy was responsible for the entity-type bias, we experimented with other representations, including *CLS* (as described above), concatenation of *ENT* and *CLS*, and entity masking. Our results showed no significant improvement in distinguishing confusable relations. This indicates that the models are still focusing on the entity-type information, even if they are getting it indirectly (e.g., from the [CLS] tag attending to the entity parts of the sentence).

4.2 Data Augmentation

We note that the FewRel *training* dataset does not contain confusable relations. Even if a few such relations existed in the dataset, the training procedure, which randomly samples a small subset of relation from a large pool, would rarely result in a difficult example where two or more sampled relations share the same type.

In the supervised setting, Rosenman et al. (2020) show that model performance can be improved by introducing (manually-created) challenging examples into the training data. We attempted a similar remedy in the FewRel setting, by adding examples from TACRED, which contains a large number of confusable relations. In order to avoid overlap between training and test data, we split the TACRED relations in two: all person relations were added to the FewRel training set, and all organization relations were used for testing.

4.3 Results

Relation	FewRel	FR+per
founded_date	72.6	**74.8**
dissolved_date	50.6	**57.3**
hq_city	59.6	**84.4**
hq_state	68.2	**92.4**

Table 5: P@50 for selected TACRED organization relations when training on FewRel alone, and after augmenting with TACRED person relations.

Table 5 shows the results of training on the FewRel dataset, augmented with TACRED person relations and testing on TACRED organization. We can see improved results and less confusion among the relations sharing similar entity type signatures. This means that with the addition of more challenging examples, the model was forced to look beyond the entity-type signature, and incorporate other information suitable for distinguishing confusable relations. Note that organization TACRED relations were *never* observed during training. In addition, augmenting the training data with TACRED person relations improves the overall accuracy from 95.18% to 96.16% on the original

FewRel validation set and from 82.54% to 85.48% on the TACRED organization relations.

5 Related Work

Several recent papers have analyzed the weaknesses of SOTA *supervised* relation extraction systems, primarily on the TACRED dataset (Zhang et al., 2017). Rosenman et al. (2020) list several "lazy" strategies employed by supervised SOTA models in the TACRED challenge, including the "entity-type heuristic" which relies solely on entity types, ignoring context. Alt et al. (2020) perform in-depth error analysis and show that many errors stem from confusing relations with identical (coarse) type signatures, and ignoring context. Tran et al. (2020) present a system that uses only entity information (in the form of unsupervised cluster IDs) to match SOTA results on TACRED. To the best of our knowledge, there has not been a similar error analysis for few-shot classification models.

6 Conclusions

In this work we explored the applicability of few-shot relation classification models in a relation extraction setting. We showed that high classification accuracy does not translate to high extraction performance, due to the reliance of few-shot models on entity type information. As a result, the models tend to perform poorly on relations involving broad entity types, such as people, locations, or dates. By explicitly adding confusable relations at training time, we force the model to rely less heavily on entity types, and consequently discriminate between relations with similar argument types. Further modifications to the training sampler that encourage the model to downweight entity type information are the subject of ongoing work.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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.

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 6251–6256, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

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.

Kalervo Järvelin and Jaana Kekäläinen. 2017. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3702–3710, Online. Association for Computational Linguistics.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online. Association for Computational Linguistics.

Leiming Yan, Yuhui Zheng, and Jie Cao. 2018. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239.

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 (EMNLP 2017)*, pages 35–45.

A FewRel Relations - Validation Set

P59 The area of the celestial sphere of which the subject is a part (from a scientific standpoint, not an astrological one).

P364 Language in which a film or a performance work was originally created.

P410 Military rank achieved by a person, or military rank associated with a position.

P155 Immediately prior item in a series of which the subject is a part.

P412 Person's voice type. Expected values: soprano, mezzo-soprano, contralto, countertenor, tenor, baritone, bass (and derivatives).

P2094 Official classification by a regulating body under which the subject (events, teams, participants, or equipment) qualifies for inclusion.

P413 Position or specialism of a player on a team, e.g., Small Forward.

P177 Obstacle (body of water, road, ...) which this bridge crosses over or this tunnel goes under.

P25 Female parent of the subject.

P921 Primary topic of a work.

P40 Subject has object as biological, foster, and/or adoptive child.

P641 Sport in which the subject participates or belongs to.

P463 Organization or club to which the subject belongs.

P206 Sea, lake or river.

P26 The subject has the object as their spouse (husband, wife, partner, etc.).

P361 Object of which the subject is a part.

B TACRED Relations

Person relations	
per:employee_of	per:cities_of_residence
per:children	per:title
per:siblings	per:religion
per:age	per:stateorprovinces_of_residence
per:countries_of_residence	per:spouse
per:origin	per:other_family
per:stateorprovince_of_birth	per:date_of_death
per:alternate_names	per:parents
per:schools_attended	per:cause_of_death
per:city_of_death	per:stateorprovince_of_death
per:country_of_birth	per:date_of_birth
per:city_of_birth	per:charges
per:country_of_death	
Organization relations	
org:founded_by	org:alternate_names
org:website	org:member_of
org:top_members/employees	org:city_of_headquarters
org:members	org:country_of_headquarters
org:stateorprovince_of_headquarters	org:number_of_employees/members
org:parents	org:subsidiaries
org:political/religious_affiliation	org:dissolved
org:shareholders	org:founded

C Few-shot Confusion Matrices

Figure 1 and Figure 2 are the confusion matrices for TACRED person and organization relations from a RoBERTa-base model trained on FewRel alone (values in percent).

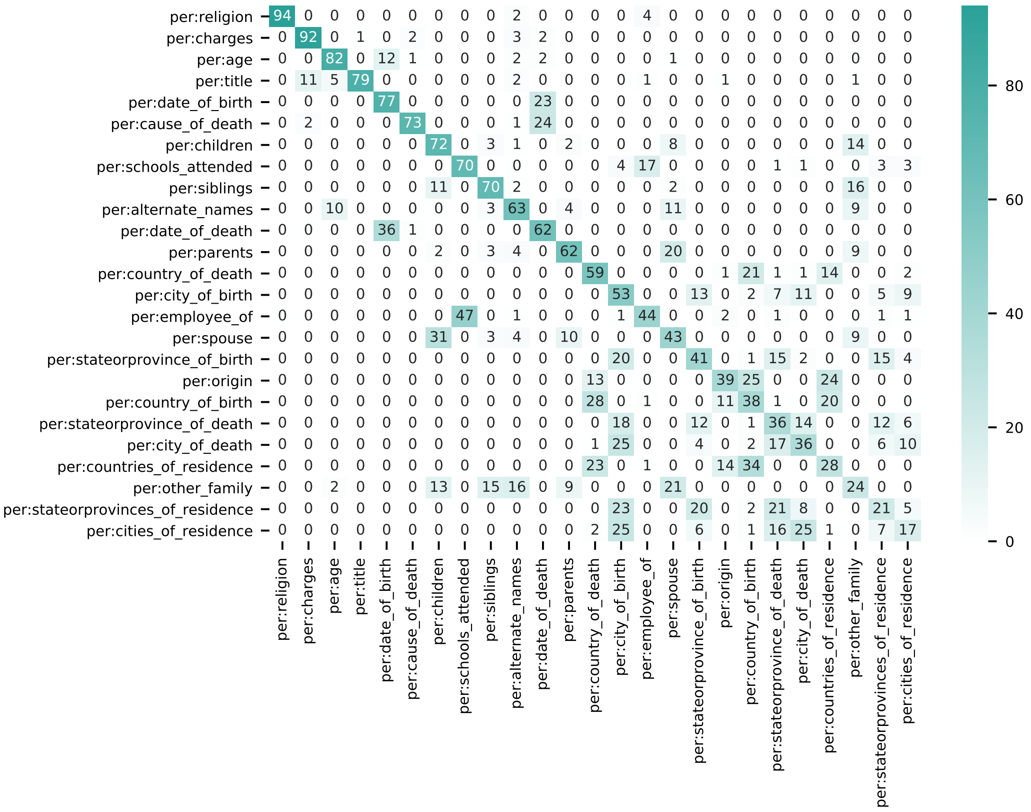

Figure 1: Confusion matrix over TACRED person relations.

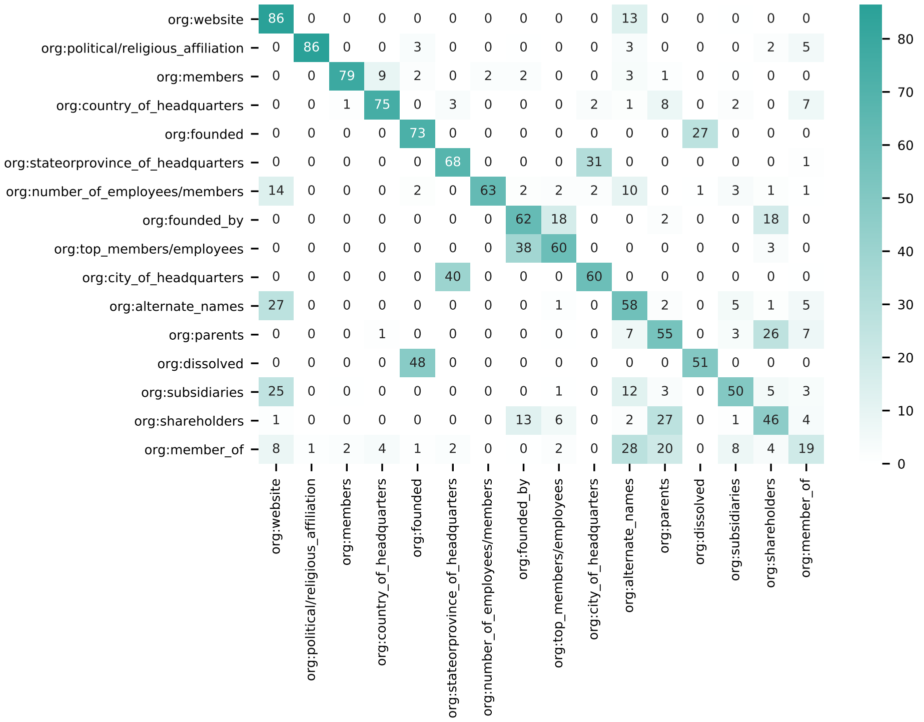

Figure 2: Confusion matrix over TACRED organization relations.