APPENDIX

A. Experimental Settings

1) Dataset: **NYT** dataset[1] is developed by [1] by aligning Freebase relations with New York Times news articles. We use the data pre-processed by [2][2]. The training data contains 1.18M sentences with 47 entity types and 24 relation types, e.g., */people/person/nationality*, */business/company/place_founded*. We exclude the *None* label (*NA*) relation, as [3] and [2], since the relation positions are uncertain. The test set contains 395 samples manually annotated by [4]. This dataset is created for DS-RE, but the data format satisfies the settings of the text-bound TE. We use this dataset to evaluate the performance of sentence-level TE. This enhances the use of automatically annotated data in supervised relation extraction.

NYT Large dataset is further developed by [5] and [6][3] based on Riedel NYT. This dataset has 53 relation labels including the *NA* labels. The training set contains 522,611 sentences, 279,786 pairs of entity and 18,252 facts which cover all sentences in Reidel NYT. We use this dataset to evaluate the DS-RE task.

SemEval-2018 Task 7 dataset [7] is provided for the task of semantic relation extraction and classification in scientific papers. This dataset defines 6 relation types, e.g., *PART_WHOLE*, *USAGE*. We adopt the dataset for subtask 2. The training/test data is composed of 350/150 abstracts of scientific publications from the ACL Anthology with manually annotated entities and relations. One of the main challenges is the limited size of the training data. To overcome it, we also select a part of the data from the noisy data of subtask 1.2 to extend the training set which finally contains 3,419 sentences.

TACRED dataset is introduced in [8]. It contains 106k sentences with entity mention pairs drawn from the yearly TAC KBP[4] challenge. Sentences are annotated with 41 person- and organization-oriented relation types, e.g., *per:employee_of*, *org:founded*, and no relation for negative examples. Entity mentions are typed, with subjects classified into person and organization, and objects classified into 16 fine-grained types (e.g., date and location).

2) Training and Hyperparameters: We conduct experiments based on the 200-D pre-trained GloVe [9], the 300-D pre-trained FastText [10], the 300-D randomly initialized word vectors and the uncased BERT-base [11] representation respectively. For the SemEval-2018 dataset, we train the domain-specific word embeddings, like (Rotsztejn et al., 2018), using GloVe and Fasttext respectively. Our corpus is composed of all the abstracts since 2001 (5.4 million tokens) collected using the API [5] on arXiv.org and the ACL ARC corpus [6] (90 million tokens). We trained the word embeddings for 500 epochs with 60 threads. We use 50-D randomly initialized character embeddings. For the TACRED and NYT Large datasets, we use the 300-D GloVe embeddings, like [8].

For regularization we apply dropout with $p = 0.5$. The output dimension of character CNN is 100-D. We set the hidden size of the LSTM unit to 300D. We set $\lambda = 1.0$. To select better models, we divide the training data of NYT into 100 parts and the training data of SemEval-2018 into 4 parts. For the NYT and SemEval-2018 dataset, we use the Adam optimization algorithm to update the model parameters with an initial learning rate of 0.001 and a decay rate of 0.9. For the TACRED dataset, we use Stochastic Gradient Descent (SGD) with the initial learning rate of 0.3 and a decay rate of 0.9. We use a cutoff of 5.0 for gradient clipping. For NYT Large, we use SGD with the learning rate 0.5. We conduct experiments on an Intel(R) Xeon(R) CPU E7-4830 v3 @ 2.10GHz (Mem: 976G) and the GPU Tesla K40c and TITAN RTX.

3) Evaluation: For the NYT dataset, we adopt the standard micro F1 score, recall (R) and precision (P) as metrics for NER and RC subtasks. A correct prediction is that the extracted triple matches the ground truth including two entities, relation direction and type. For the SemEval-2018 dataset, we use the official script. The evaluation is carried out in two steps. First, the relation label and directionality are ignored, so it only evaluates the quality of entity pairs by the F1 score. Second, the evaluation of relation classification is the macro-average F1 score. For the TACRED dataset, we report the micro F1 score as [8]. For the DS-RE task, we adopt the held-out evaluation as [6], which is an effective evaluation method for a large dataset without costly human intervention. We compare the precision and recall curve. The curve is drawn by ranking all predicted instances according to their confidence scores and traversing the ranking list from the high score to low score to measure the precision and recall at each position.

B. Analysis

In this section, we performed ablation studies on the text-bound and distant supervision relation extraction respectively. Then, we analyze the influence of relation candidate selection on model training. Finally, we use case study to visualize the model's attention distribution.

1) Ablation Study: **Text-bound Relation Extraction** Table III shows an ablation study of multi-task training and pipeline training on the NYT dataset. Two systems denote a fine-tuned NER system and a SwitchNet system. Here we use a BERT model as the NER system. We first compare the multi-task training in different settings. A first observation is that relying on the fine-tuned NER system slightly reduces the final results by 1.72% F1. Although the NER system is enhanced, some entities still hurt the final precision. Because extracting more entities does not always help the relation extraction subtask and some entities may increase the risk of predicting false positives. The above training process is a multi-stage pipeline and relation candidates are determined by the NER system. We first fine-tune the NER system and then predict the label sequence $\hat{Y}_{ner}$ which are also written to disk. Then we train the joint entity and relation extraction model that can use the $\hat{Y}_{ner}$ information. When we replace $\hat{Y}_{ner}$ with the ground truth NER labels, the result is significantly improved (5.03% F1). This means that a high-quality label sequence can help improve the final results.

Removing the multi-task training degrades the performance by 0.94% F1. This means that multi-task training benefits the RE subtask from the NER subtask. When we remove the NER system, the performance improves by 1.72% F1. Because the global optimization process creates an organism that does not rely on the NER system. Then, we remove the NER system and multi-task training. We first train the NER subtask, then freeze the parameters of the shared layer, and then train the RE subtask. The F1 score drops by 2.51% because the RE subtask cannot be encoded at the lower layer to interact with the NER subtask. This means that multi-task joint training is critical for subtask interaction. When we do not freeze the lower layer, the result drops significantly. Because when we train the RE subtask, the memory of the NER subtask is weakening. This phenomenon is also known as catastrophic forgetting [12], [13], [14], [15].

We also apply our training pipeline to BERT. We fine tune a simple BERT model for TE. Table II shows the results. When

[1] http://iesl.cs.umass.edu/riedel/ecml/
[2] https://github.com/shanzhenren/CoType
[3] https://github.com/thunlp/NRE
[4] https://tac.nist.gov/2017/KBP/index.html
[5] https://arxiv.org/help/api/index
[6] http://acl-arc.comp.nus.edu.sg/

TABLE I: SwitchNet setting ablation

Model	P	R	F1
Two systems + multi-task	56.78	51.89	54.23
+ NER label	61.18	57.46	59.26
–multi-task	55.64	51.13	53.29
–NER system	60.28	52.27	55.95
–multi-task, NER system	50.33	56.96	53.44
–multi-task, NER system, Frozen NER	55.88	43.29	48.78

TABLE II: BERT model setting ablation

Model	P	R	F1
Two systems + multi-task	51.30	54.93	53.05
+ NER label	54.89	61.01	57.79
–multi-task	50.46	54.43	52.37
–NER system	50.68	56.20	53.30
–multi-task, NER system	46.46	54.93	50.34
–multi-task, NER system, Frozen NER	–	–	–

only using multi-task training, BERT achieves slightly lower results than our SwitchNet. When we remove the multi-task, NER system and Frozen NER, this model almost forgets all the NER memory, so the RE subtask fails. Fine-tuning models can achieve different functions, which also means that BERT is sensitive to parameter changes.

Figure 1 shows the class-aware prediction result. Figure 1 (a) and 1 (b) denote the model with or without the ground truth NER labels when extracting triples. We observe "*place_lived*", "*nationality*", "*place_of_death*", "*contains*" and "*company*" relations are easier to extract. Other relations in the test set sometimes are not extracted. This is because some sparse relations are not fully learned and there is the class imbalance problem in the training data. We visualize and analyze this dataset in Appendix A. We observe that providing NER labels enhances the extraction of well-trained relations, while other relations are not obviously improved.

Distant Supervision Relation Extraction We performed ablation studies on DS-RE. Table III shows the results of different settings. A first observation is that adding representations of subject and object to the multi-head attention reduces the results. This is because all positive and noisy samples in the bag contain the same subject and object, so using this information may hurt model training. When we remove the multi-head attention and keep only one head, the AUC decreases, but the P@N of the top-ranked relations improves. When we also remove the sentence-level attention, the result drops. This means that sentence-level attention is important in this framework. When we remove only the word-level attention, the result also drops. This means the word- and sentence- level attention mechanisms complement each other to form the hierarchical attention mechanism. When we remove the two-level attention, the results will drop further.

Figure 2 shows the precision and recall curves of different settings. We observe that when we do not use the **w**ord-level **att**ention (wATT), the curves generally move down. As the recall increases, the precision decreases more rapidly. This means that word-level attention can help achieve higher performance. When we remove the sentence-level attention (ATT), the curve fluctuations increase. This means that the sentence-level selection enhances the stability of the model.

2) Analysis of Relation Candidates Selection: Candidates' feedback can be positive and negative and both types of feedback have great potentials to boost recall and precision. However, the number of possible relations is $\mathcal{O}(n^2)$, where n is the number of entities, which potentially increases computational complexity and may lead to class imbalance problem. Existing approaches for this question typically perform random sampling, which might include some inefficient relation candidates. Inspired by the idea of SVM [16], a few support vectors are effective to decide the classification hyperplane. We assume that difficult candidates that are closer to the classification hyperplane can improve the classifier more effectively. This can reduce the computational complexity. During model training, we rank the negative candidates according to their prediction probability of being a None relation and only keep top-k negative ones with the least probabilities. These relations are more likely not to be predicted as NA. k is hyper-parameter, and we experimented with $k \in \{1, 2, \infty\}$. ∞ denotes all candidates. This ranking mechanism filters a vast number of negative candidates leaving the classifier with a small set.

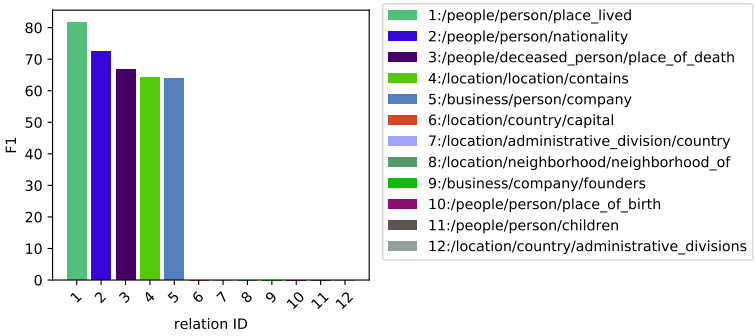
We compared several selection strategies, i.e. random selection (RANDOM), top-k selection (TOP-K) and all candidates (ALL), as shown in Figure 3. We find that using top-k selection can achieve comparable F1 scores to using all candidates, but the precision and recall might be not balanced enough. When more negative candidates are retained, the recall will decrease and the precision will increase. This method potentially enhances the use of ranking mechanism to reduce the computational complexity of RE models.

C. Case Study

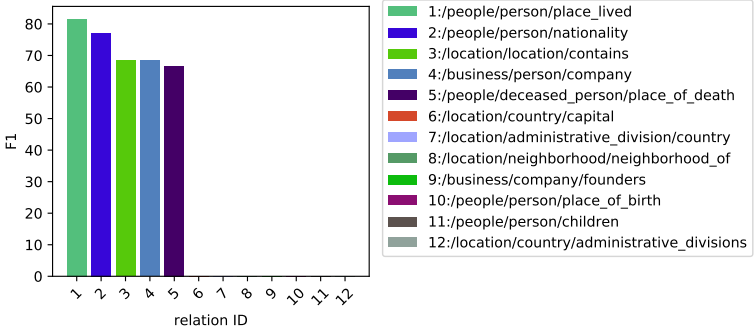
It is instructive to analyze which words the model attends to when classifying relations. We hand-picked some examples in the SemEval-2018 dataset and visualize the attention patterns of these samples.

Figure 4 shows how our model extracts informative words for relation representation. The first column is the sample id. The second column contains the extracted triples, and the third column shows the textual input. The underlined phrase represents an entity in the extracted triple. The red degree denotes the word weight for the relation representation. The first sentence shows that this model is capable of focusing on informative words to identify the "*PART_WHOLE*" relation type for "*English-Chinese Bitexts*" and "*Web*". The second sentence shows this model resolves the comparative relation by attending to "*narrower than*". The third sentence shows that "*are described in*" means the "*TOPIC*" relation.

The fourth sentence shows that "*produced by means*" for "*Translations*" and "*beam search decoder*" denotes the "*MODEL-FEATURE*" relation, while the fifth sentence shows this model extracts the relation type "*RESULT*" by attending to "*produce best*". This suggests that this model considers the entity semantics and sentence context. The sixth sentence shows that this model can extract informative tokens "*resource*



(a) Multi-task training



(b) Multi-task training + NER label

Fig. 1: Class-aware result

TABLE III: SwitchNet setting ablation

Model	100	200	300	Mean	AUC(%)
M-SwitchNet+ATT	81.2	74.1	72.1	75.8	37.9
+ subject, object	79.2	72.1	69.8	73.7	35.0
–Multi-head	82.2	77.1	71.4	76.9	36.1
–Multi-head, ATT	81.1	75.1	68.4	74.9	35.0
–Multi-head, Single-head	73.3	73.6	69.8	72.2	35.1
–Multi-head, Single-head, ATT	75.2	71.6	70.8	72.6	34.7

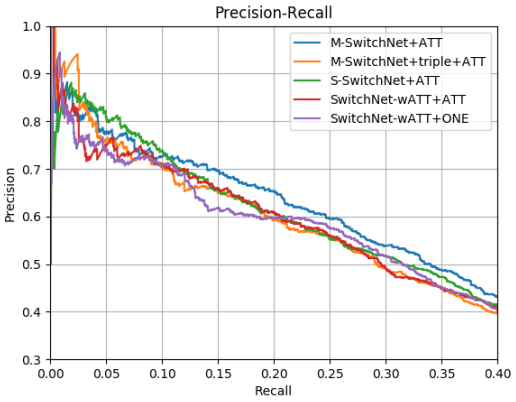


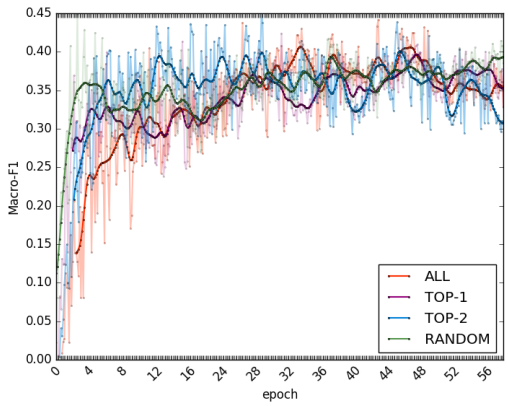
Fig. 2: Precision and recall curves of different settings

for" instead of the important verb in a long context. These results indicate that the model considers the entity semantics and the relation context while attending to the informative words for relation representation.

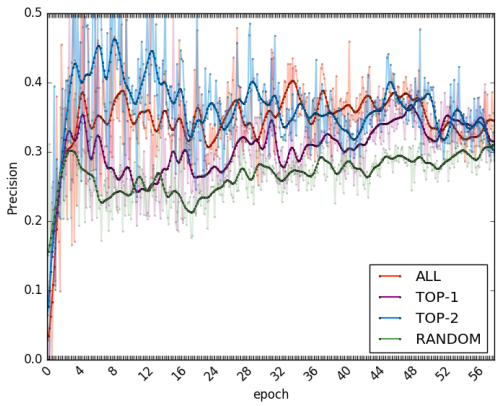
D. Dataset Overview

Figure 5 (a) shows the relation distribution of the NYT dataset. NA instances occupy the main part, but here we focus on analyzing the pre-defined relations. The x- and y-axes are the relation ID and relation number respectively. We first observe there are more than 50,000 "*contains*" relations while others are less than 9,000. Each relation from 11 to 24 occurs less than 1,000 times. Figure 5 (b) shows that "*contains*" relation occupies the main part (48.4%), while the relations from 9 to 24 account for 7.7%. This dataset has class imbalance problem, which poses a challenge to model performance. This means we can use distant supervision to generate more training data for sparse classes or we can introduce structured knowledge using KG embeddings. Figure 5 (c) shows the average distance of each relation type between subject and object. The x- and y-axes are relation ID and token number respectively. We observe that "*industry*", "*geographic_distribution*", "*place_of_death*" relations are often described in a long context, and "*profession*", "*neighborhood_of*" and "*founders*" are more likely described in a short context. Figure 5 (d) shows the token number of samples. The x- and y- axes denote the token number and sample number respectively. We observe the average token number is 37.8.

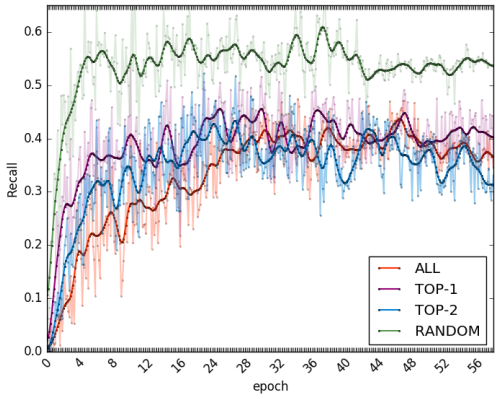
Figure 6 (a) shows the relation distribution of the SemEval-2018 dataset. "*USAGE*" relation occupies the main part (39.3%).



(a) Macro-F1



(b) Precision



(c) Recall

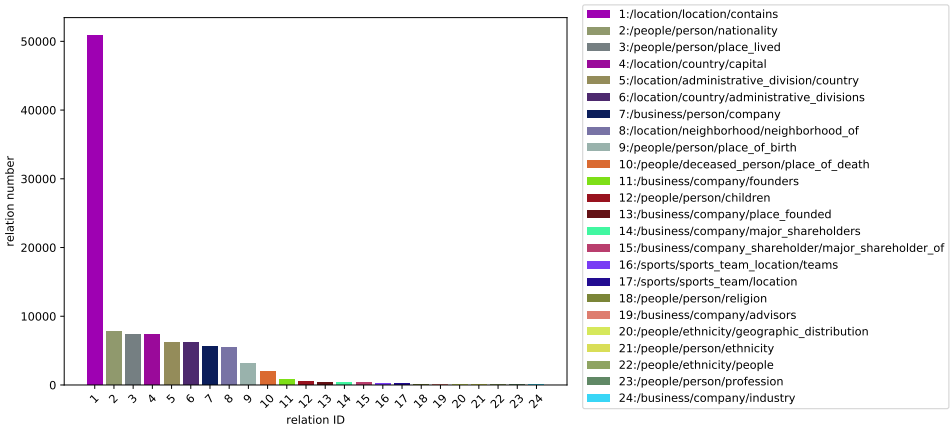
Fig. 3: Prediction results on SemEval-2018 based on different selection strategies for relation candidates

Id	Triple	Sentences
1	(English-Chinese bitexts, PART_WHOLE, Web)	This piece of work has also laid a foundation for exploring and harvesting English-Chinese bitexts in a larger volume from the Web .
		This piece of work has also laid a foundation for exploring and harvesting English-Chinese bitexts in a larger volume from the Web .
2	(domains, COMPARE, MUC-4 terrorism domain)	These previous domains were much narrower than the MUC-4 terrorism domain .
		These previous domains were much narrower than the MUC-4 terrorism domain .
3	(paper, TOPIC, overview)	An overview of HowNet and information structure are described in this paper .
		An overview of HowNet and information structure are described in this paper .
4	(beam-search decoder, MODEL-FEATURE, Translations)	Translations are produced by means of a beam-search decoder .
		Translations are produced by means of a beam-search decoder .
5	(Bayesian classifiers, RESULT, recall performance)	In our evaluation , Bayesian classifiers produce the best recall performance of 80 % but the precision is low (60%) .
		In our evaluation , Bayesian classifiers produce the best recall performance of 80 % but the precision is low (60%) .
6	(WordNet, USAGE, Word Sense Disambiguation (WSD) task)	WordNet has been used extensively as a resource for the Word Sense Disambiguation (WSD) task , both as a sense inventory and a repository of semantic relationships .
		WordNet has been used extensively as a resource for the Word Sense Disambiguation (WSD) task , both as a sense inventory and a repository of semantic relationships .

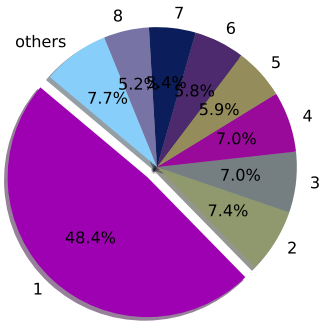
Fig. 4: Visualization of some cases

"*TOPIC*" relation is sparse so it has more impact for the final results since the evaluation process uses a macro-averaged F1 score. Although this dataset has less types of relations than NYT dataset, the data sparsity problem poses a challenge to the model. Figure 6 (c) shows the "*COMPARE*" relationship is more likely described with more words, while the "*MODEL-FEATURE*" and "*PART_WHOLE*" relations are more likely expressed with less words. We observe the average distance between subject and object entities is shorter than NYT dataset. Figure 6 (d) shows the token number of samples. We observe the average token number is 25.8.

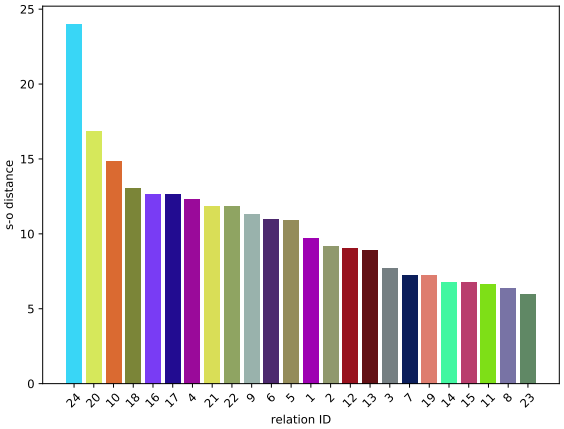
Figure 7 (a) shows the relation distribution of the TACRED. "*per:title*" relation occupies the main part (17.7%). Figure 7 (c) shows the "*per:stateorprovince_of_death*" relation is more likely described with more words, while the "*org:website*" and "*org:alternate_names*" relations are more likely expressed with less words. We observe similar properties between NYT and TACRED datasets. For example, they have some similar relations, i.e., "*per:country_of_birth*" $\sim$ "*/people/person/place_of_birth*", "*per:employee_of*" $\sim$ "*/business/person/company*". "*per:city_of_death*" and "*/people/deceased_person/place_of_death*" both have longer s-o distance. It seems promising to combine the two datasets to address the data sparsity problem. Figure 7 (d) shows the token number of samples. We observe the average token number is 36.4.



(a) Relation number



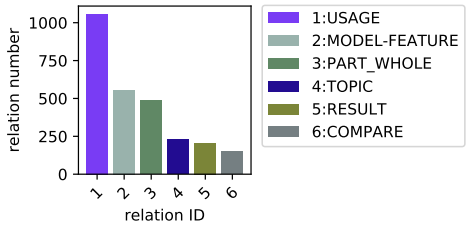
(b) Relation portion



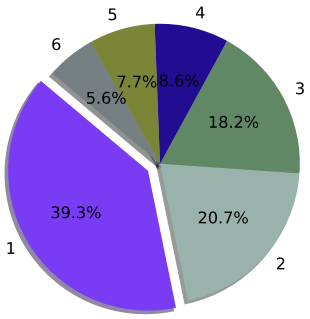
(c) s-o distance

(d) Token number

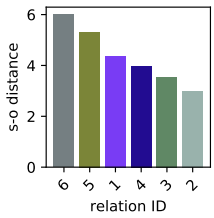
Fig. 5: NYT Dataset visualization



(a) Relation number



(b) Relation portion



(c) s-o distance

(d) Token number

Fig. 6: SemEval-2018 dataset visualization

REFERENCES

[1] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010.

[2] X. Ren, Z. Wu, W. He, M. Qu, C. R. Voss, H. Ji, T. F. Abdelzaher, and J. Han, "CoType: Joint extraction of typed entities and relations with knowledge bases," in *Proceedings of WWW 2017*, 2017.

[3] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction of entities and relations based on a novel tagging scheme," *arXiv preprint arXiv:1706.05075*, 2017.

[4] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proceedings of the ACL 2011*, 2011.

[5] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of EMNLP 2015*, 2015.

[6] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proceedings of ACL 2016*, 2016.

[7] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, "Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proceedings of SemEval 2018*, 2018.

[8] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proceedings of EMNLP 2017*, 2017.

[9] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the EMNLP 2014*.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, vol. 5, pp. 135–146, 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

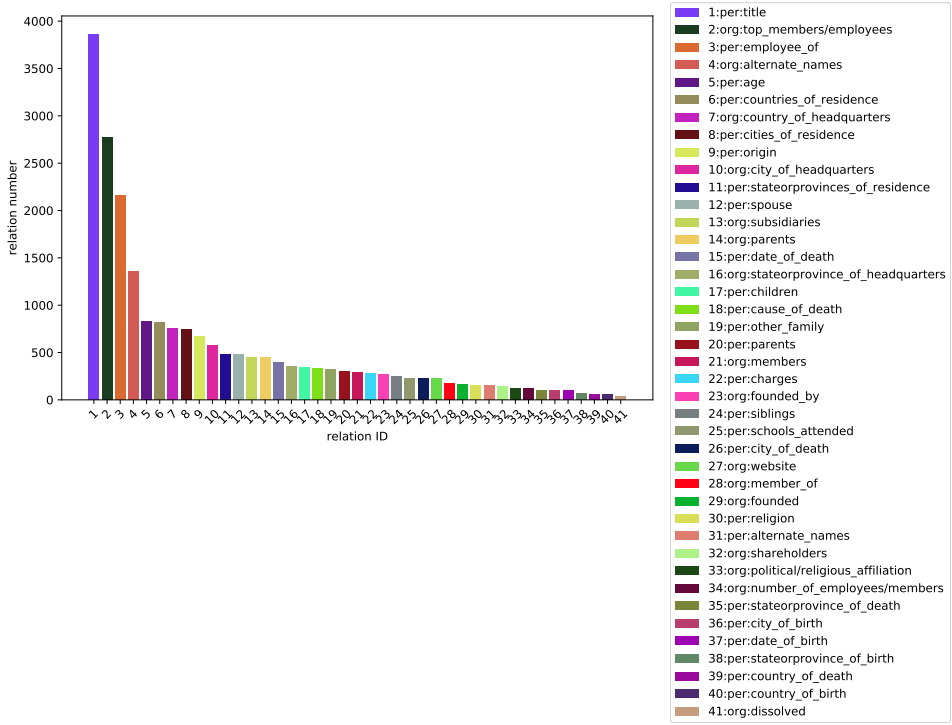
[12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[13] Z. Chen and B. Liu, "Lifelong machine learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 10, no. 3, pp. 1–145, 2016.

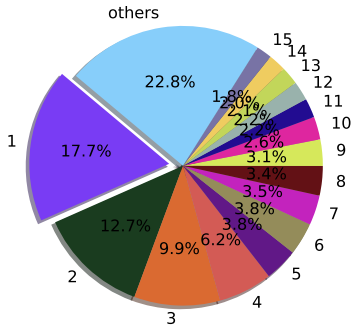
[14] W. Hu, Z. Lin, B. Liu, C. Tao, Z. Tao, J. Ma, D. Zhao, and R. Yan, "Overcoming catastrophic forgetting for continual learning via model adaptation," in *Proceedings of the ICLR 2018*, 2018.

[15] G. Zeng, Y. Chen, B. Cui, and S. Yu, "Continual learning of context-dependent processing in neural networks," *Nature Machine Intelligence*, vol. 1, no. 8, pp. 364–372, 2019.

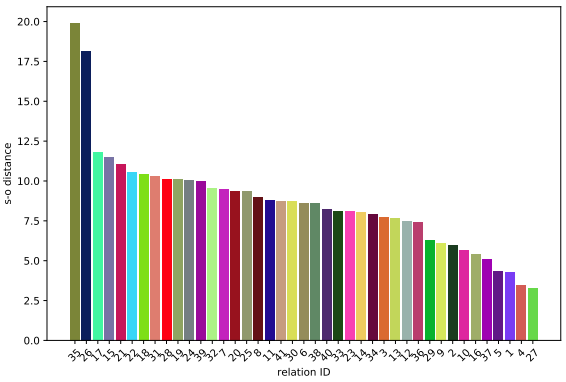
[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.



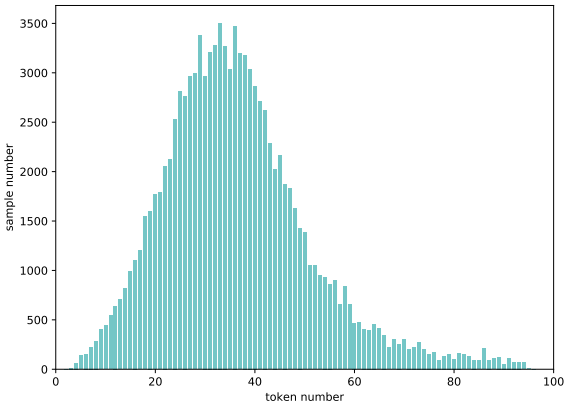
(a) Relation number



(b) Relation portion



(c) s-o distance



(d) Token number

Fig. 7: TACRED visualization