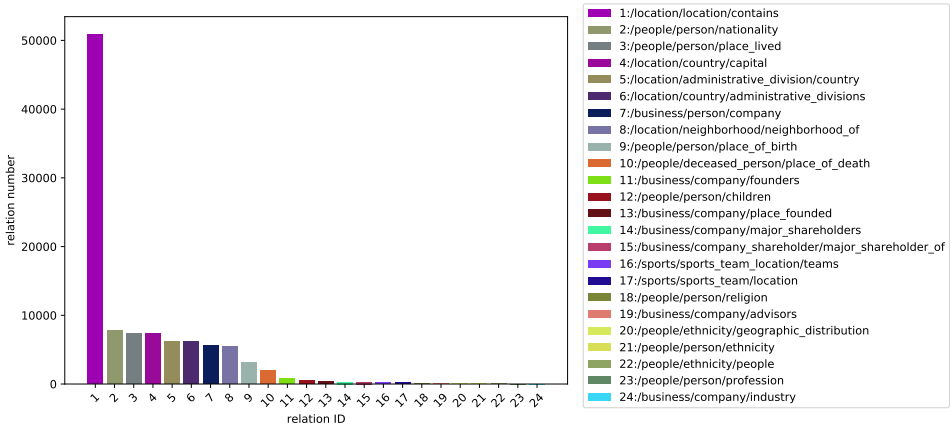
Appendix A. Dataset Overview

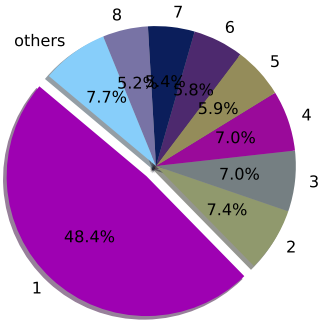
Figure A.1 (a) shows the relation distribution of the NYT dataset. NA instances occupy the main part, but here we focus on analyzing the pre-defined relations. The x- and y-axes are the relation ID and relation number respectively. We first observe there are more than 50,000 "*contains*" relations while others are less than 9,000. Each relation from 11 to 24 occurs less than 1,000 times. Figure A.1 (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 A.1 (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 A.1 (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 A.2 (a) shows the relation distribution of the SemEval-2018 dataset. "*USAGE*" relation occupies the main part (39.3%). "*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 A.2 (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 A.2 (d) shows the token number of samples. We observe the average token number is 25.8.

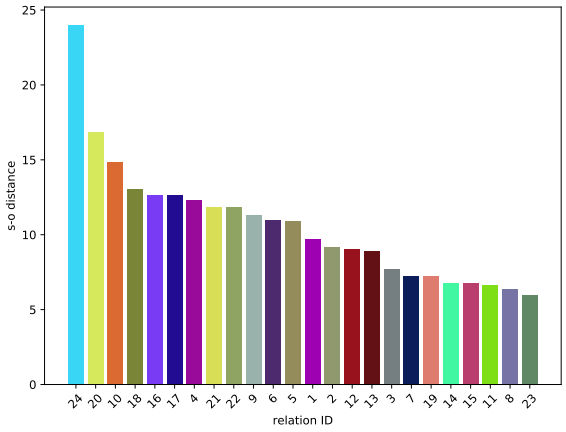
Figure A.3 (a) shows the relation distribution of the TACRED. "*per:title*" relation occupies the main part (17.7%). Figure A.3 (c) shows the "*per:stateorprovince_of_death*"



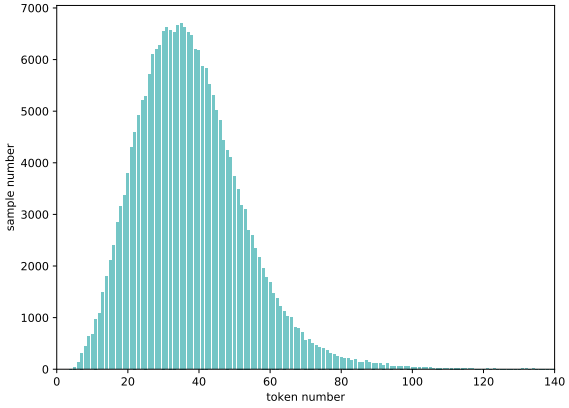
(a) Relation number



(b) Relation portion

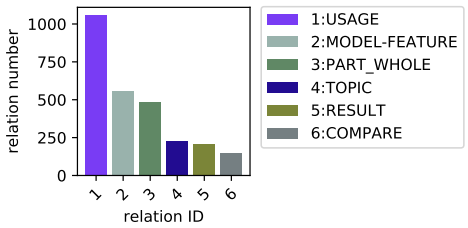


(c) s-o distance

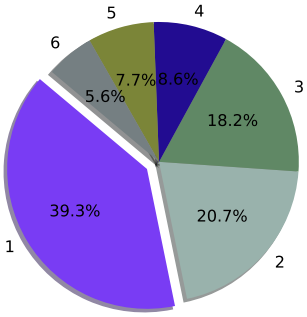


(d) Token number

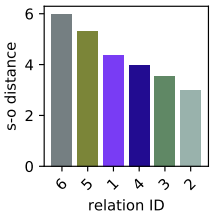
Figure A.1: NYT Dataset visualization



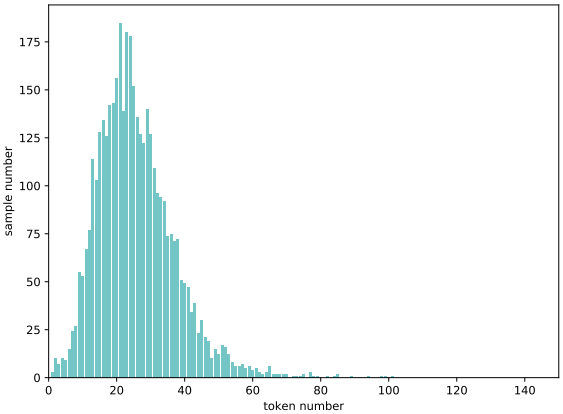
(a) Relation number



(b) Relation portion



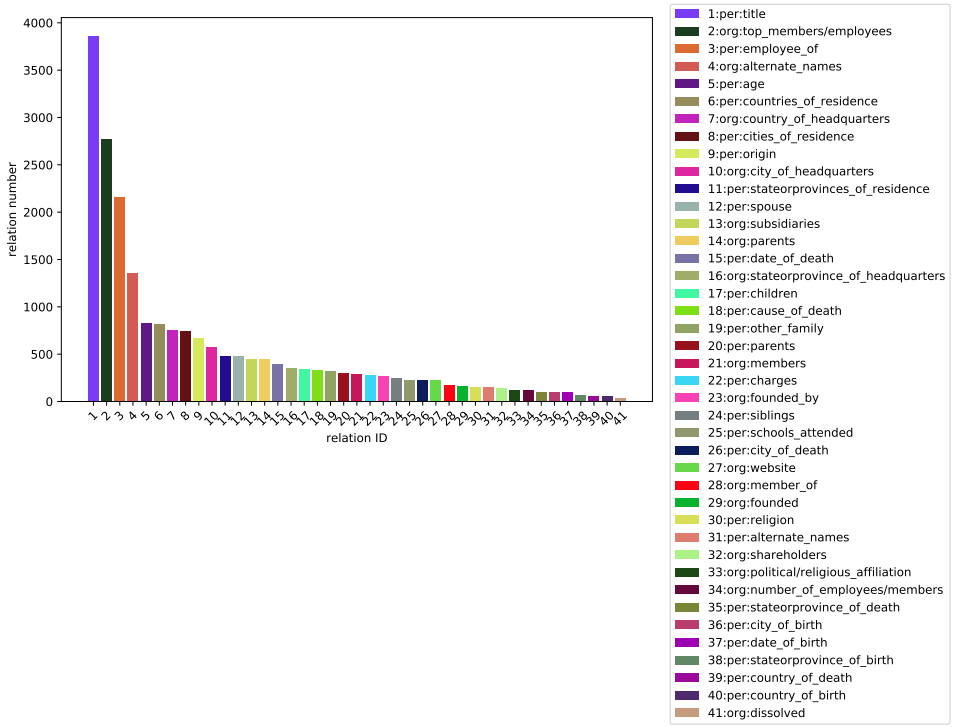
(c) s-o distance



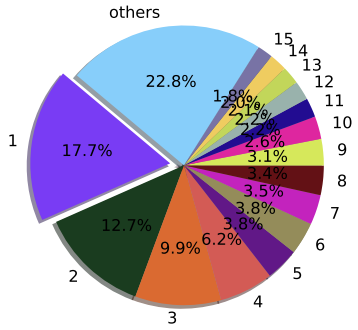
3

(d) Token number

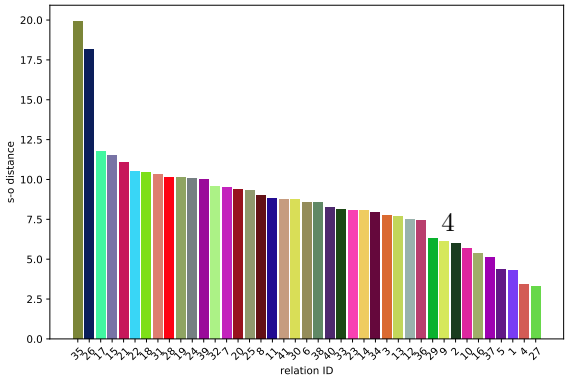
Figure A.2: SemEval-2018 dataset visualization



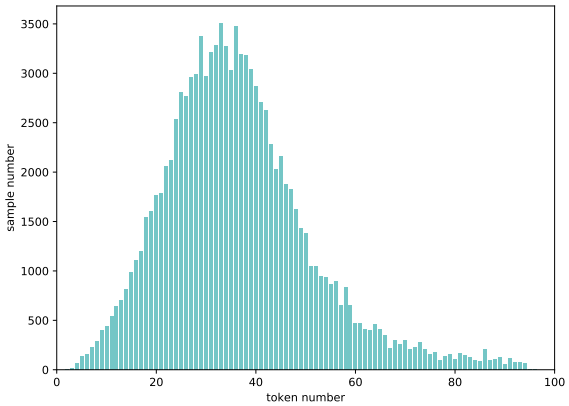
(a) Relation number



(b) Relation portion



(c) s-o distance



(d) Token number

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 A.3 (d) shows the token number of samples. We observe the average token number is 36.4.

References