Appendix A. Analysis

Appendix A.1. Vertex Representation Analysis

To analyze the vertex embeddings learned for relation classification, Figure A.1 visualizes vertex embeddings distribution. We pick some samples from the NYT dataset and use PCA and t-SNE for dimensionality reduction. As shown in Figure A.1 (a), dominant vertices for relation classification are represented as colored points placed according to their coordinates. Different colors denote different relation types. PCA tends to reveal the semantic relationship between relation types. We first observe that there are two main clusters, where the cluster of "*contains*" and "*neighbor_of*" relations describes the relationship between locations, while the other cluster describes the relations about person and organization/location. The lower part of the diagram shows the relations of "*place_lived*", "*company*" and "*nationality*" are closely related, but the relations are mixed. t-SNE tends to show a more accurate relationship between relations. As shown in Figure A.1 (b), the "*placed_lived*" is closer to "*nationality*". "*nationality*" is closely related to "*company*", while the "*place_lived*" is not always close to "*company*". This approach is instructive for exploring interesting phenomena in social and economic development. This GRNN, by leveraging vertex embeddings, provides better understanding about relations. This means the vertex embeddings encode semantic relationship between relation types.

Appendix A.2. Dataset Overview

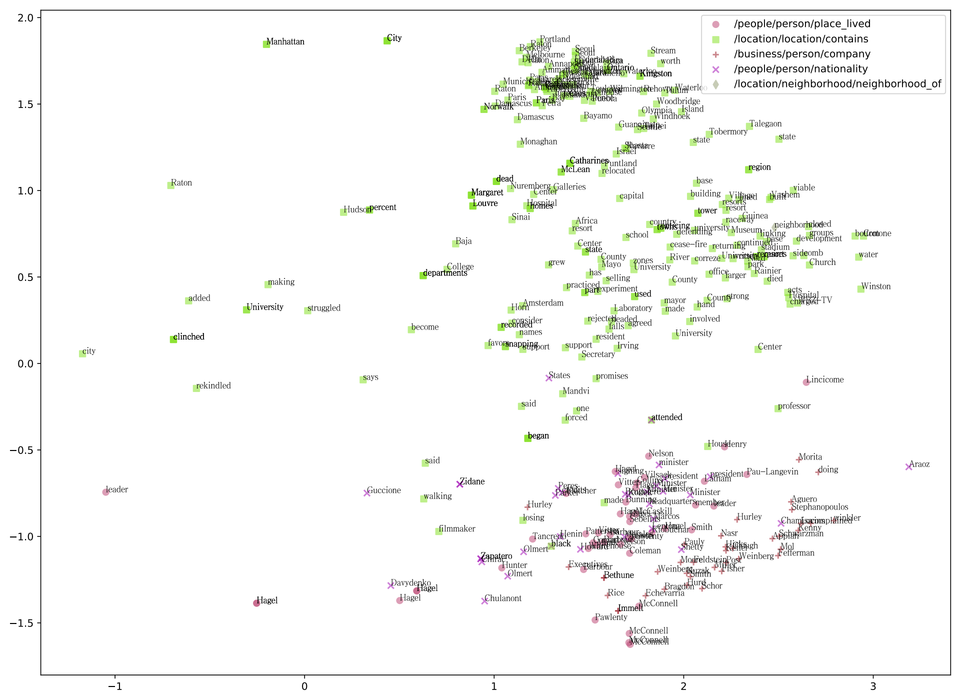
Figure A.2 (a) shows the relation distribution of the NYT dataset. NA instances occupy the main part, but here we focus on analyzing the pre-defined relations. The x- and y- axes are the relation ID and relation number respectively. We first observe there are more than 50,000 "*contains*" relations while others are less than 9,000. Each relation from 11 to 24 occurs less than 1,000 times. Figure A.2 (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. It is important to use distant supervision technique to increase the training data of sparse classes or to introduce entity embeddings of knowledge graph to compensate for data sparsity. Figure A.2 (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 longer context, and "*profession*", "*neighborhood_of*" and "*founders*" are more likely described in a short context. Figure (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.3 (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 fewer types of relations than NYT dataset, the data sparsity problem poses a challenge to the model. Figure A.3 (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.3 (d) shows the token number of samples. We observe the average token number is 25.8.

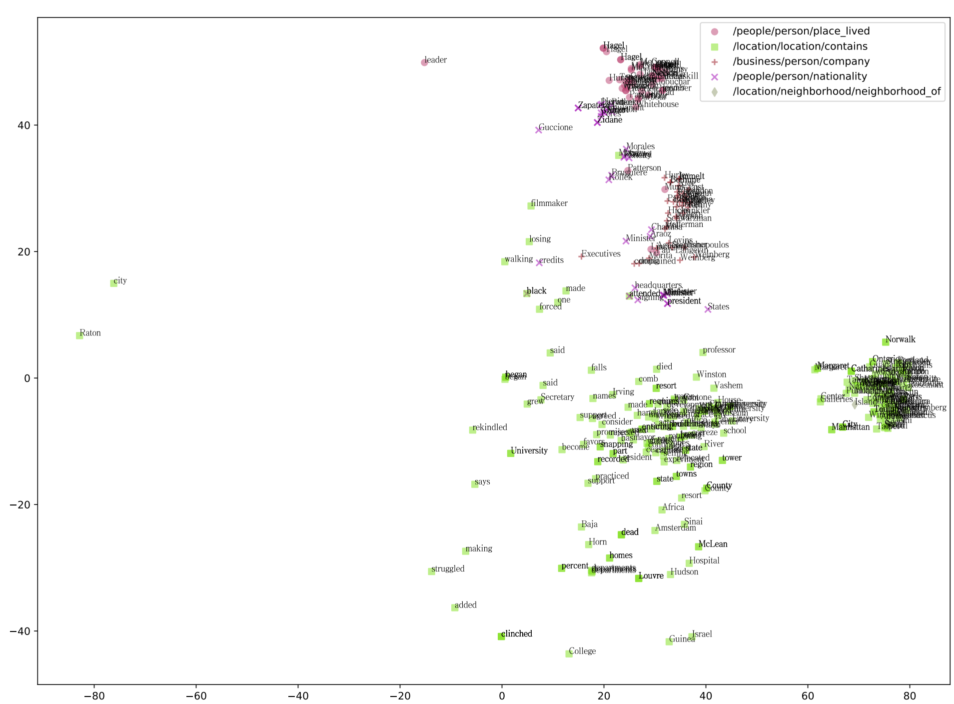
Appendix A.3. Edge Types of Dependency Parsing

We list the main edge types of dependency parsing in Stanford CoreNLP, i.e., "acomp, advcl, acl, advmod, agent, amod, appos, aux, auxpass, ccomp, cc, compound, name, mwe, conj:and/or, conj:and, conj:both, conj:but, conj:nor, conj:or, conj:plus, conj, conj_x, cop, csubj, csubjpass, dep, det, discourse, dobj, expl, goeswith, infmod, iobj, mark, mwe, neg, nn, npadvmod, nsubj, nsubjpass, number, num, op, parataxis, partmod, pcomp, pobj, possessive, poss, nmod:poss, preconj, predet, case, nmod:aboard, nmod:about, nmod:above, nmod:according_to, nmod:across_from, nmod:across, nmod:after, nmod:against, nmod:ahead_of, nmod:along, nmod:alongside_of, nmod:alongside, nmod:along_with, nmod:amid, nmod:among, nmod:anti, nmod:apart_from, nmod:around, nmod:as_for, nmod:as_from, nmod:aside_from, nmod:as_of, nmod:as_per, nmod:as, nmod:as_to, nmod:at, nmod:away_from, nmod:based_on, nmod:because_of, nmod:before, nmod:behind, nmod:below, nmod:beneath, nmod:beside, nmod:besides, nmod:between, nmod:beyond, nmod:but, nmod:by_means_of, nmod:by, nmod:depending_on, nmod:dep, nmod:despite, nmod:down, nmod:due_to, nmod:during, nmod:en, nmod:except_for, nmod:excepting, nmod:except, nmod:excluding, nmod:exclusive_of, nmod:followed_by, nmod:following, nmod:for, nmod:from, nmod:if, nmod:in_accordance_with, nmod:in_addition_to, nmod:in_case_of, nmod:including, nmod:in_front_of, nmod:in_lieu_of, nmod:in_place_of, nmod:in, nmod:inside_of, nmod:inside, nmod:in_spite_of, nmod:instead_of, nmod:into, nmod:irrespective_of, nmod:like, nmod:minus, nmod:near, nmod:near_to, nmod:next_to, nmod:off_of, nmod:off, nmod:of, nmod:on_account_of, nmod:on_behalf_of, nmod:on, nmod:on_top_of, nmod:onto, nmod:opposite, nmod:out_of, nmod:out, nmod:outside_of, nmod:outside, nmod:over, nmod:owing_to, nmod:past, nmod:per, nmod:plus, nmod:preliminary_to, nmod:preparatory_to, nmod:previous_to, nmod:prior_to, nmod:pursuant_to, nmod:regarding, nmod:regardless_of, nmod, nmod:round, nmod:save, nmod:since, nmod:subsequent_to, nmod:such_as, nmod:thanks_to, nmod:than, nmod:throughout, nmod:through, nmod:together_with, nmod:to, nmod:toward, nmod:towards, nmod:underneath, nmod:under, nmod:unlike, nmod:until, nmod:upon, nmod:up, nmod:versus, nmod:via, nmod:vs., nmod:whether, nmod:within, nmod:without, nmod:with_regard_to, nmod:with_respect_to, nmod:with, prt, punct, purpcl, quantmod, rcmod, root, tmod, vmod, xcomp".

The frequent edge types contain "acl, acl:relcl, advcl, advmod, amod, appos, aux, auxpass, case, cc, cc:preconj, ccomp, compound, compound:prt, conj, cop, csubj, csubjpass, dep, det, det:predet, discourse, dobj, expl, iobj, mark, mwe, neg, nmod, nmod:npmod, nmod:poss, nmod:tmod, nsubj, nsubjpass, nummod, parataxis, punct, root, xcomp".

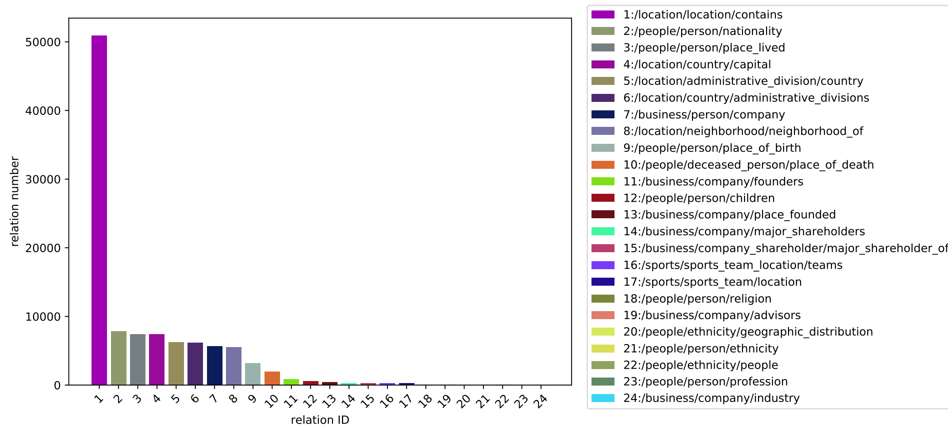


(a) Dimensionality reduction with PCA

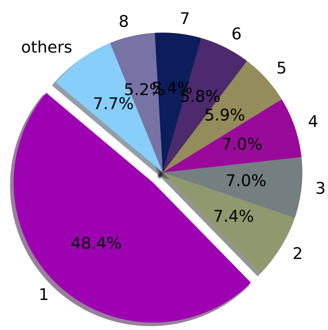


(b) Dimensionality reduction with t-SNE

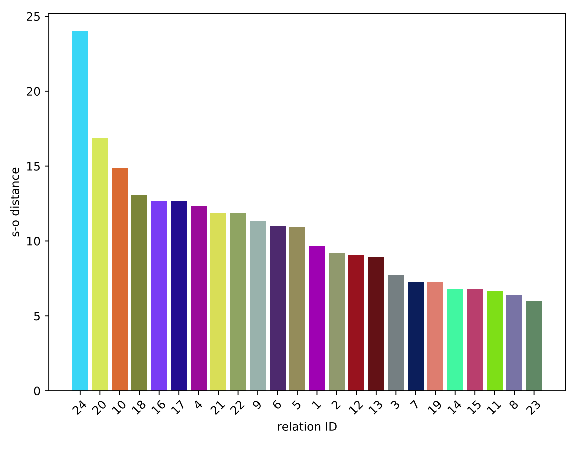
Figure A.1: Vertex embeddings visualization



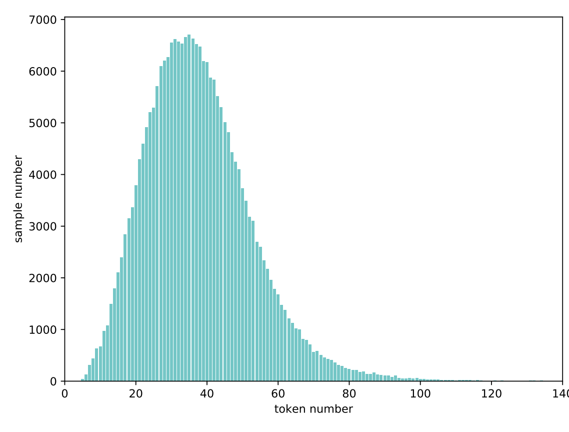
(a) Relation number



(b) Relation portion

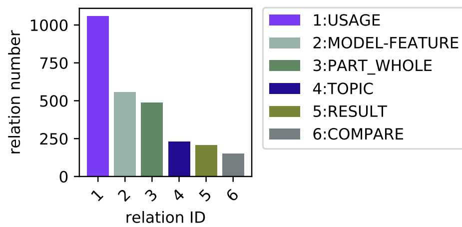


(c) s-o distance

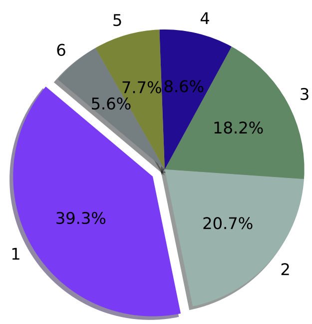


(d) Token number

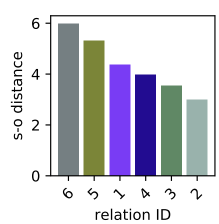
Figure A.2: NYT Dataset visualization



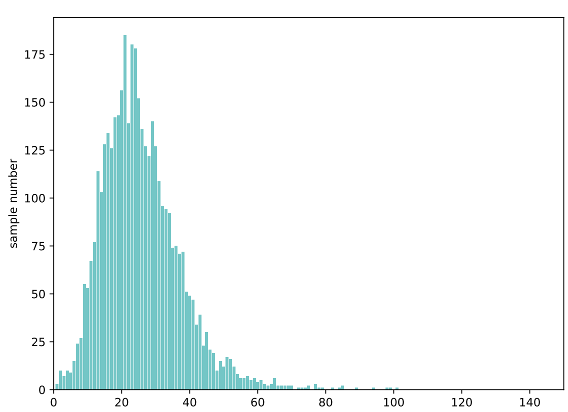
(a) Relation number



(b) Relation portion



(c) s-o distance



(d) Token number

Figure A.3: SemEval-2018 dataset visualization

References