CS 224U Experimental Protocol

Thomas Boyden
Stanford University
thomasb3@stanford.edu

Robert DiDonato
Stanford University
rdido@stanford.edu

Miles Zoltak
Stanford University
mzoltak@stanford.edu

1 Hypotheses

In our final project we seek to explore the potential benefits of combined usage of knowledge embeddings and word embeddings trained jointly such that their representations occupy the same vector space.

This project is inspired by "Text Classification Using Embeddings: A Survey" (Costa et al., 2023) in which they explore the benefits of using embeddings to assist in text classification. In the article word embeddings are the primary subject (as is the case in most NLP literature relating to embeddings), but knowledge embeddings are also mentioned with enthusiasm.

The authors note that "text embeddings, even context-dependent ones, do not precisely represent semantic relations between concepts and entities as KGs do." By leveraging KGs in the form of knowledge embeddings, the authors argue that text classification may be done more effectively, especially on when texts have "little contextual information."

With respect to the goal of employing jointly trained word and knowledge embeddings, our goal is to create a text classification system and compare performance between 4 different types of embeddings: (1) plain pretrained word embeddings, (2) plain pretrained knowledge embeddings, (3) disjoint word and knowledge embeddings, (4) jointly trained word and knowledge embeddings.

Our hypothesis is that jointly trained word and knowledge embeddings will be the highest performer, disjoint will come shortly behind (perhaps demonstrating that gains of joint vs disjoint training are marginal), word embeddings, and in last will be knowledge embeddings.

It is possible that jointly trained knowledge embeddings will outperform all the others by a significant amount depending on the way in which those embeddings are trained as compared to the pretrained embeddings. To minimize this potentially misleading disparity, we will try to keep the training datasets/methods as similar as possible across the different types of embeddings. Potentially, this could involve not using pretrained embeddings at all and actually training types 1-3 if that is feasible. It is not clear to us yet whether that is indeed feasible in the time we have between now and the time this project must be finished, so that will impact whether we explore that option.

2 Data

For evaluation, we will be using the News Category Dataset from Kaggle (Wang et al., 2022). This dataset contains 42 different categorires of news stories, and each entry represents a news article as a category, headline, and short description, along with a few other fields we do not plan to use in our model.

Additionally, the pretrained embedding models we will be using are a TransE embedding of the Wikidata5m dataset (Wang et al., 2021) and BERT (Devlin et al., 2018). The joint model we will propose in our final project will be trained on the datasets used to pretrain those models.

3 Metrics

Because we are attempting a straightforward text categorization task, we have 42 different categories to choose from. To evaluate our models we will be calculating precision and accuracy over all predictions.

4 Models

For our baseline model we will simply be using a text classification model from PyTorch that will allow us to experiment with different embedding layers. We will not have any baselines that already exist. However, we will be training 4 different models and evaluating each. As stated before, the first three are simple word embeddings, knowledge embeddings, and then a disjointly-trained combination of word and knowledge embeddings. These are a sort of baseline for us, because we will be comparing them to our custom jointly-trained word and knowledge embeddings in terms of performance in this simple model.

This model is appropriate for our use case because it is simple. We are trying to demonstrate the possible power of these jointly-trained embeddings, not the model into which they'll be fed.

As for the actual model for training the joint embeddings, we will be evaluating the approach set forth in KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation (Wang et al., 2019). This is the only paper we have found so far that seems to deal with the problem we are approaching. We may attempt to use these embeddings trained as they have done and then use that as a baseline if we can obtain their model.

5 General Reasoning

Our expectation is that our approach should be effective. We expect it to at the very least outperform both the knowledge-only and word-only embeddings. It would be nice to see it outperform the disjointly trained versions of those embeddings, but depending on our approach this could absolutely go south. Even if our system does not outperform disjoint embeddings, we think that this idea makes a lot of sense and it should not discourage futher work in this field. Our best case scenario is outperforming KEPLER, or at least having comparable performance.

6 Summary of Progress

The extent to which we have begun our experimention is limited to logistical organizing efforts. We have been primarily occupied with getting access to datasets and pretrained models to this point. In the near future we will begin training our systems and devising a more specific approach to joint-training of knowledge and word embeddings.

References

Liliane Costa, Italo Oliveira, and Renato Fileto. 2023. Text classification using embeddings: a survey. *Knowledge and Information Systems*, 65:1–43.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *CoRR*, abs/1911.06136.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Wikidata5m.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.