COMP 550 Project Proposal

Group 47

Group Member Tianyi Wu 260714699
Yue Ma 260759490
Jingyuan Wang 260860682

Project Summary SWEM[1] is a text classification model which used simple pooling functions over word-embeddings before training the parameters with multilayer perceptron. This techniques yields performance equal to or better than sophisticated neural network models, such as CNN and LSTM, while having considerably fewer parameters and thus is more efficient. In our project, we are going to implement the SWEM model and further test its feasibility on Chinese and Japanese text datasets.

Resources and Dataset According to the paper[1], SWEM produces best performance on document classification tasks. Hence, we will experiment on similar corpus in other languages: for Chinese we will use "THUCTC(THU Chinese Text Classification)" dataset while for Japanese "Reputation Analysis dataset from Japanese Twitter"[2]. All texts will be preprocessed and converted to romanization . Furthermore, since the original paper used pre-trained word-embeddings as input, we are also going to use generated word vectors[3][4].

Proposed Experiment Firstly we will implement the SWEM model. With pre-generated word-embeddings as initialization, our model will apply pooling function over the whole sequence and train MLP afterwards. The experiment will be performed from the following perspectives:
1. Comparison of SWEM and other baseline on English, Chinese and Japanese datasets: A main difference between English and the other two languages in terms of nlp tasks is that Chinese and Japanese texts need manual segmentation. In addition, Chinese and Japanese has many homonyms which indicates two words with the same romanization may have completely different meanings and the languages contains certain combination ambiguities. Consequently it is significant to evaluate if SWEM is feasible on Chinese and Japanese corpus.
2. Functionality of different pooling methods: SWEM proposed four pooling functions which are average pooling, max pooling, concatenated pooling and hierarchical pooling. The authors found that each method outperforms the others in certain tasks. We are going to investigate which function has the best performance with Chinese and Japanese dataset since each language has special word order.
3. Model performance on document categorization vs. other nlp tasks: The model can also be utilised in short text classification, sentiment analysis, etc. If possible, we will acquire varied corpus and explore how good it works on those tasks for Chinese and Japanese language.

References

[1] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Kyoko Ikuo, Suzuki Yu, Yoshino Koichiro, New Big Glam, Ohara Hitoshi, Mukai Riro, and Nakamura Satoshi. Extracting japanese reputation information from twitter using word semantic vector dictionary.

[3] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018.

[4] Hiroyuki Shinnou, Masayuki Asahara, Kanako Komiya, and Minoru Sasaki. nwjc2vec: Word embedding data constructed from ninjal web japanese corpus. *Journal of Natural Language Processing*, 24(5):705–720, 2017.