Yiran Mao 260850827
Silan He 260738985

Final Project Proposal

GloVe is an unsupervised word vector learning algorithm.

The model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix instead of the entire sparse matrix. In other words, GloVe generates a word-word co-occurrence weight table that it uses to generate the word vectors. The weighting should not overweigh rare co-occurrences or frequent co-occurrences due to the weight function being a piecewise logarithmic and constant function. This model performs very well and produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a word analogy task described by Mikolov (2013).

We propose to generate a Chinese word vector in order to compare this model’s performance to native Chinese word vector models on a standardized word analogy task.

Chinese is very different from English in many ways. For example, a Chinese sentence has no spaces so that we need to additionally preprocess the corpus with some word segmentation tools. Currently we plan to use the “jieba” segmentation toolkit. It is very popular and useful in many Chinese NLP tasks.

The Chinese corpus we intend to use to generate our word vectors include documents from many areas such like Weibo, Wikipedia_zh, Baidu Encyclopedia as well as some of the Chinese Literature. This rich corpus will definitely cover most Chinese words, as well as implied semantic and syntactic properties.

We want to evaluate the word vectors using the Chinese analogy reasoning data set CA8 and its evaluation toolkit developed by Shen Li (2018). Using this standardized word analogy task, we can compare the GloVe word vector performance with existing native Chinese word vector performance.

Additionally, we can discuss the final results by exploring the advantages and disadvantages of the usage of GloVe with Chinese with respect to the morphology and syntactic differences Chinese has with English. We also seek to discuss the final results with respect to the GloVe’s differences with native Chinese word vector learning algorithms.

References:

Tomas Mikolov, Kai Chen, Greg Corrado, and Jef- frey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR Work- shop Papers*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations