Machine Learning Project Proposal

Group Members:

Zhuoru Lin (zl1755)

Zhaopeng Liu (zl1732)

**Business Problem**

The business problem of this project is to alleviate a common problem with modern-day intercultural communication: Word Sense Disambiguation (WSD). Many words have more than one senses, which makes it difficult especially for a language learner to locate the most appropriate sense of a word in a given context. For example, crane can be either a bird or a machine. In the sentence “Crane eat a fish”, one have to manually distinguish the correct sense of the word “Crane”. The objective of this project is to create a language model that is able to process the whole sentence and attempt to yield the correct meaning of each word in a sentence, and therefore improve the efficiency of intercultural communication. This natural language processing task is also named “All-word WSD” in literature.

**Description of the Data**

**Word senses inventory**

We will use WordNet (Fellbaum, 1998) as our word senses inventory. As one of the most widely used and comprehended lexical database, WordNet contains senses knowledge of 155,287 words. Beside senses information, WordNet also contains 117,659 synnets that show the syntactic relations between words. It also contains lexical categories like nouns, verbs, adjectives and adverbs. Words from the same lexical category with similar meaning are grouped into synsets, For example, good, right, ripe are adjective synsets.

**Annotated corpus**

In order to train a supervised language model, an annotated corpus is needed. SemCor (Miller et al., 1993), with 234,000 WordNet annotated words, annotated corpus is currently the most widely used annotated corpus for All-word WSD.

**Word embeddings as features**

Using word embedings instead of simple bag-of-words vector generally allow models to pick up deeper syntactic information between words. The neural network model, widely applied in natural language processing tasks, almost unanimously leverage the advantages of word embeddings. In this project we would use two most popular pretrained word embeddings: Word2Vec (Mikolov, 2013) and GloVe (Pennington, 2014).

**Evaluation**

Different literature have been using different metrics for evaluation. To evaluate our model and compare different models we build. We will use accuracy and F1-Score as our evaluation metrics.

Accuracy is defined by:

accuracy = Number of correct senses assigned / Number of words in total.

Note that the total number of senses assigned could be smaller than the number of words in total. The accuracy defined above is sometimes refered as Recal. The fraction of correct senses assignment and total senses assigned is defined to be the precision of the model:

precission = Number of correct senses assigned /Number of senses assigned. The model’s F1-Score is defined as:

F1 = 2\*Prcession / (Precession + Recall)

Beside using a quantitative measurement of our model accuracy, we also would like to observe our model’s performances in our chosen test context. Since WSD might have a direct application in language education, we would like to know how well our model disambiguates some challenging long sentences in SAT reading corpus.

**Models**

**Baseline models**

• Most frequent senses model(MFS): This model always assigns the most frequent sense to every word in test corpus. Note that senses frequencies information is encoded in most sense inventories since modern dictionaries mostly enumerate senses in a sense-frequency order.

• ExtLesk: This model predicts by comparing the Word2Vec gloss with word context window in a bag-of-word fashion.

**Models to be implemented**

We will implemented two models in this project: a SVM-based model (Iacobacci et al., 2016). and a Long Short Term Memory(LSTM) neural network model proposed by (Yuan et al., 2016). Both models firstly build a language model to predict the hold out word in a sentence and classify the correct sense in a nearest-neighbor fashion by comparing the predicted context embeddings with sense embeddings. Due to the length constraint, please refer to the cited paper for detailed model structure.

**Timeline**

Week 1 (Mar 24- Mar 31): We will investigate the structure of WordNet database and SemCor annotated corpus. All data should be cleaned and restructure so that it is ready to be used in the model.

By the 2nd meeting with advisor (April 19): Implement most-frequent-sense(MFS) baseline model and at least one of the proposed model using training data (SemCor). Start the LSTM Models. List the problems that occur during the implementation process.

By the 3rd meeting with advisor (May 9th): Two proposed models should have already been successfully implemented and are ready for fine tunning of parameters.

Before submission (May 12th): Trying techniques (semi-supervised learning if possible) to aim for better results.

**References**

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Iacobacci, I., Pilehvar, M.T. and Navigli, R., 2016. *Embeddings for word sense disambiguation: An evaluation study.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 897-907).

Miller, G., Leacock, C., Tengi, R., and Bunker, R. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*. 303–308.

Mikolov, T., Le, Q., and Sutskever, L. 2013. *Exploiting similarities among languages for machine translation.* arXiv preprint arXiv:1309.4168.

Pennington, J., Socher, R., and Manning, C. 2014. *GloVe: Global Vectors forWord Representation*. In Proceedings of the 2014 EMNLP , pages 1532–1543, Doha, Qatar.

Yuan, D., Richardson, J., Doherty, R., Evans, C. and Altendorf, E., 2016. *Semi-supervised Word Sense Disambiguation* with Neural Models. arXiv preprint arXiv:1603.07012.