Project Abstract
Foundations of Machine Learning 2016
Professor. Mohri
Due Date: 12/20/2016

Yves Greatti
Utku Evci

Problem Statement

A new feature space for vector embedding models and its application in Natural Language Processing Tasks

Investigation

Word2vec [1] is a popular software from Google available in the Tensorflow framework, which has given existence to many research papers. Word2vec is a shallow neural network with one hidden layer allowing one-hot encoding of the words. It uses a vector representation of the words, and each word id (0 to the number of words in the vocabulary) is mapped to a low-dimensional vector-space from their distributional properties observed in the text corpus given as input to word2vec. The output feature vectors are the probabilities that each word is embedded near each other (semantic similarity). The weights or feature vectors are learned using two methods: CBOW or Skip-Gram [2] . Although different in nature, they both have the same end result: vectors of words judged similar by their context are nudged closer together. For a faster training, these models instead of computing similarities of every word against every context of the text corpus, selects noise contexts randomly, a method known as negative sampling. However, words have more than one meaning, which is captured in a lexical database like WordNet [1]. And the goal of this project is to change the output vectors of word2vec by using WordNet or other semantic networks such as BabelNet [3].

At the core of wordnet is a synset, an unordered set of synonyms which could be nouns, verbs or adjectives. A synset contains a definition (a gloss), and can have different types of relations with other synsets like ISA relation. The main idea of the project is to use synsets or glosses as inputs of word2vec. The "tokens" of word2vec are the synset or gloss ids. Therefore, given such input, word2vec generates latent vectors in the synset or gloss feature space. With this new set of features vectors, we evaluate the quality of our word representation on a natural language understanding evaluation dataset such as Google datasets SEM_REL and SYN_REL (Mikolov), depending availability, or SemEval. To generate the training data which consists of the sequence of synset or gloss ids corresponding to the initial document, we use a tagged database like BabelNet, or a boosting algorithm with base classifier Lesk. We run an experimentation with the new set of vectors generated by word2vec in a word disambiguation task using the energy based algorithm MaxEnt [4]. And we may also generate gloss vectors by using the SVD based model, GloVe [2]. Finally we may conduct, if time permits, more investigations using syn2vec or gloss2vec with deep learning networks.

References

[1] T Mikolov - 2013 *Efficient Estimation of Word Representations in Vector Space*.

[2] Xin Rong - 2013 *word2vec Parameter Learning Explained*

[1] `https://wordnet.princeton.edu/`
[2] `http://nlp.stanford.edu/projects/glove/`

[3] Roberto Navigli and Simone Paolo Ponzetto - 2012 *Multilingual WSD with Just a Few Lines of Code: the BabelNet API*

[4] Adam Berger, Stephen Della Pietra, and Vincent Della Pietra - 1996 *A maximum entropy approach to natural language processing*