Assignment 2 Report

Manas Joshi, 2015ME10108

May 8, 2019

1 Preprocessing

- Apostrophe expansion and lowercasing of tokens
- Removal of all digit or all punctuation tokens
- Replacement of tokens with POS tags ”NNP” or ”NNPS” with ”-pro-”

2 Model

- Gensim Word2Vec model initialized with pretrained embeddings given as input.
- Trained on corpus + eval_data given as input with <<target>> replaced by actual target value.
- Inference using ”score” function of gensim Word2Vec model which returns the log likelihood of a sentence.
- For every test target word *t*, a sentence is created with <<target>> replaced by *t*. The target words are ranked based on the log likelihood of the this sentence.

3 Postprocessing

- Scores of target words not in vocabulary are reduced and are ranked last.
- If POS tag of <<target>> comes out to be ”NNP” or ”NNPS”, target word ”-pro-” if present, is ranked first.