Assignment - 01

Group Number: 08

160007J Abinayan Balarajah

160256U Jeyakeethan Jeyaganeshan

160442L Nirojan Thiyakarasa

Background: Word representations learned from neural language models have been shown to improve many NLP tasks. These representations are learned as parameters in a language model and trained to maximize the likelihood of a large corpus of raw text. They are included as features or used to initialize the parameters of neural networks targeting tasks for which substantially less training data is available. One of the most widely used models used in Word2Vec tool, in particular the “skip-gram” and the “continuous bag-of-words” (CBOW) models. These two models discard word order information in how they account for context. This leads to have many major improvements in many tasks. Though, Word2Vec remains a popular choice due to their efficiency and simplicity.

Problem and Motivation: The main issue with the original models is the fact that they are insensitive to word order which is useful for inducing semantic representations, this leads to suboptimal results when they are used to solve syntax-based problems.

When word order is discarded, the many syntactic relations between words cannot be captured properly. This is supported by empirical evidence that suggests that order-insensitivity does indeed lead to substandard syntactic representations. The systems using pre-trained with Word2Vec models yield slight improvements but more expensive than the model which uses word order information embeddings of Collobert et al. (2011) and yielded much better results.

Contribution: Two simple modifications to Word2Vec, one for the skip-gram model and one for the CBOW model, that improve the quality of the embeddings for syntax-based tasks 1. Our goal is to improve the final embeddings while maintaining the simplicity and efficiency of the original models.

Scope: Natural language to machine readable vector in Natural Language Processing.