Natural Language Processing
A study of Language Models using word2vec

Isabel B. Amaro[1], **Adriano Veloso**[2]

[1]Department of Computer Science – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil

{isabel.amaro,adrianov}@dcc.ufmg.br

Abstract.

Introduction

With the appearing of social networks, a massive amount of data is constantly being generated weekly in the last few years (CAMBRIA, 2014). Natural Language Processing is a Computer Science field that studies the best ways to retrieve, process and generate from representations of human language (CAMBRIA, 2014). To work with human language, Language Models are used in NLP. A well known Language Model is word2vec, which represents each word from a corpus as a vector. Each word (vector) in the space is related to a context, which are the other words that are around it.

This work starts explaining two implementations of word2vec: Continuous Bag-Of-Words and Skip-Gram. Then, it analyses these two implementations' analogy after training with different sizes of corpus and context.

word2vec

Word2vec represents words in a vector space. To do this representation, word2vec can use Continuous Bag-Of-Words Model or Skip-Gram Model. The Continuous Bag-Of-Words uses continuous distributes representation of the context to predict the current word. Thus, as standard bag-of-words model, the other of words in context doesn't change the prediction (MOKOLOV, 2013). On the other hand, in Skip-Gram, the context is the one to be predicted given the current word. The Skip-Gram model is capable get better results increasing the range of the context, but it is related to more computational cost (MOKOLOV, 2014).

Trained models

There were used three Wikipedia corpus with different sizes to train both CBOW and Skip-Gram language models, with 3 different window sizes (context). The word2vec code was also provided by Google.

Result and analysis

Conclusion