Sentence matrix normalization using most likely n-grams vector

Mohamad Abdolahi
Kharazmi International Campus Shahrood University Shahrood, Iran
mabdolahi512@yahoo.com
09153140551

Moreza Zahedh
Kharazmi International Campus Shahrood University Shahrood, Iran
zahedi@ganjineh.co.ir
09124738644

***Abstract*— Word embeddings is one of the interesting natural language processing filed and has been shown to be a great asset for a large variety of NLP tasks. N-gram language models are also an important text processing methods and are based on statistics of how likely words are to follow each other. A big problem in all text processing approaches based on word vectors are different size of sentences matrices. In most of suggested approaches, the average value of columns is calculated and a one dimensional vector including N elements is created. But the approach has some disadvantages such as its resistance against word ordering, ignoring sentence length feature and their vector sentences are very close to each other. We introduce an efficient and very simple rich statistical model of Word2Vec approach and n-grams language model to assess unique size sentence matrices. The unique size resulting matrix does not depend on the language and its semantic concepts. Our results demonstrate that certain models capture complementary aspects of coherence evaluation, text summarization, automatic essay scoring, detecting fake and copied texts, text topic comparison and thus can be combined to improve performance.**

Keywords- n-grams normalization; natural language processing (NLP); sentence matrix; text preprocessing; word2vec algorithm word vector.