Text coherence new method using word2vec sentence vectors and most likely n-grams

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

Morteza Zahedi
Kharazmi International Campus Shahrood University
Shahrood, Iran
zahedi@ganjineh.co.ir

***Abstract*— Discourse coherence modeling evaluation remains a challenge task in all Natural Language Processing subfields. Most proposed approaches focus on feature engineering, which accepts the sophisticated features to capture the logic, syntactic or semantic relationships between all sentences within a text. This paper investigates the automatic evaluation of text coherence. We introduce a fully-automatic rich statistical model of local and global coherence that uses word2vec approach to assess the coherence a document. Our modeling approach relies on numerical vectors derived from word2vec algorithm applied on a very large collection of texts. We successfully combined the word2vec vectors and most likely n-grams with cohesive LD-n-grams perplexity to assess the coherence and topic integrity of document. We present experimental results that assess the predictive power that it does not depend on the language and its semantic concepts. So it has the ability to apply on any language. Our model achieves state-of-the-art performance in coherence evaluation and order discrimination task on two datasets widely used in the previous methods.**

Keywords—: global text coherence; local text coherence; language models; word embeddings