Automatic texts similarity metric for large datasets

Aliya Jangabylova*

Institute of Information and Computational Technologies, Kazakhstan

**Corresponding author's e-mail: ajangabylova@gmail.com*

Abstract

The similarity of texts is an essential part that is applied in many natural language processing tasks such as text retrieval, automatic summaries, plagiarism detection and many other where is necessary to catch any semantic relation of texts. There are several approaches that is close to evaluate effectively the true relation of texts. However, some of them are computationally costly and can be used only for small datasets, whereas others are might be fast but not as accurate as it desired. This thesis is aimed to consider other metrics that could be accurate and yet computationally effective to be applied for large datasets.

Keywords: semantic similarity, texts similarity

1 Introduction

The detection of similarity of words, texts or documents is a key element for further broadly used tasks such as topic modelling, texts classification or even recommender systems. Different approaches should be used depending on the task formulation. For simple query tasks might be enough to use feature-based metrics like Jaccard similarity [1] that demonstrates how much common features (words) contribute relatively to the distinct features.

However, Jaccard similarity does not differentiate homonym words, which have the same spelling but different meanings, or synonym words that have identically same meaning but different spellings. That is why a more popular and efficient approach to use is vector-based words. The main idea is that each word can be represented as a vector and then to utilize distance metrics such as Euclidean or Cosine to evaluate how close the words are, where the smaller distance denotes a stronger relation of words. The state-of-the-art vector representations are Bag of Words [2], Term Frequency - Inverse Document Frequency [3] or embedding methods such as Word2Vec [4] or Bert [5].

In [4] was presented a new metric called Word Mover Distance which based on Word2Vec embedding and minimizes a total distance between two group of words placed in two sentences/documents. The main disadvantage of this method is that its complexity time is $O(p^3 \log(p))$. Thus [4] presented an alternative WMD with relaxed boundaries with some reduction in accuracy.

The 2019 was called the year of Bert [5] which is a new stare-of-the-art model and it is based on transfer learning that uses pre-trained models and allows to fine-tune them under a specific task. The main highlight of Bert that order of words does matter. However, in practice it does not always gives promising results as it is not very clear what is the optimal way of extracting embedding of words.

2 Decision

So, this paper suggests to develop a different method that will look to all possible pair of words from two texts/documents and based on some decision function decide whether they are similar enough to contribute to the total similarity of two documents. Moreover, considering the weights of each word in regards to the document it located in or in regard to the whole corpus would play an essential role. This method is very flexible since as input it takes any desired word vectors or embeddings and it could be tuned further by choosing weights depending on the task. In addition, there can be chosen many versions of decision function to evaluate what gives the best result. By this, we will take an advantage of state-of-the-art embeddings and be able to apply for large datasets as it requires less computational time compared to WMD.

3 Conclusion

In this thesis we made a review of the most popular methods in texts similarity, discussed pros and cons of each of them and proposed the view of another method that could outperformed the popular methods and be applied for Big Data.

References

[1] Hamers L, et al. 1989 *Similarity Seasures in Scientometric Research*
[2] Wallach H 2006 *Topic Modelling: Beyond Bag-of-Words*
[3] Aizawa A 2003 *An information theoretic prospectice of tf-idf measures*
[4] Kusner M, et al. 2015 *From word embeddings to distances*
[5] Devlin J, et al. 2018 *Pre-training of deep bidirectional transformers for language understanding*