Distributional Semantic Model Based on Convolutional Neural Network for Arabic Textual Similarity

Adnen Mahmoud, Higher Institute of Computer Science and Communication Techniques, Monastir, Tunisia

Mounir Zrigui, Faculty of Science Monastir, Monastir, Tunisia

ABSTRACT

The problem addressed is to develop a model that can reliably identify whether a previously unseen document pair is paraphrased or not. Its detection in Arabic documents is a challenge because of its variability in features and the lack of publicly available corpora. Faced with these problems, the authors propose a semantic approach. At the feature extraction level, the authors use global vectors representation combining global co-occurrence counting and a contextual skip gram model. At the paraphrase identification level, the authors apply a convolutional neural network model to learn more contextual and semantic information between documents. For experiments, the authors use Open Source Arabic Corpora as a source corpus. Then the authors collect different datasets to create a vocabulary model. For the paraphrased corpus construction, the authors replace each word from the source corpus by its most similar one which has the same grammatical class applying the word2vec algorithm and the part-of-speech annotation. Experiments show that the model achieves promising results in terms of precision and recall compared to existing approaches in the literature.

KEYWORDS

Arabic Language, Context Based Approach, Global Vectors Representation, Natural Language Processing, Paraphrase Detection, Semantic Similarity, Word Embedding, Word2vec

1. INTRODUCTION

The rapid development of information and communication technologies has generated a tremendous amount of data which has increased the potential source of plagiarism. This is because of the lack of honesty, irresponsibility and self-confidence due to limited time and competitive pressure to achieve good results. It allows taking the work of others and representing it as one's own work without mentioning the source. Different ways can be applied such as directly copying ideas, adding/deleting of words, or their intelligently substituting them. In this context, we consider the problem of paraphrase detection which requires semantic textual similarity analysis. It has represented an essential problem in many Natural Language Processing (NLP) tasks (e.g. sentiment analysis, question answering, information retrieval, etc.). Often, an important problem to solve is the lack of resources in the publicly available Arabic language. The purpose of this paper is to detect Arabic paraphrase based on global Vector Representation (GloVe) as a feature extraction technique. We apply various supervised machine-learning algorithms e.g. Support Vectors Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR) and Convolutional Neural Network (CNN), and we compare their performances for classification. The remainder of this paper is organized as follows: First, we present

DOI: 10.4018/IJCINI.2020010103

the problem statement in section 2. Next, we make an overview of previous work in section 3. After that, the components of our model are detailed in section 4. The experimental setup and results are discussed in section 5. Finally, we give our conclusions and future work in section 6.

2. PROBLEM STATEMENT

The amount of textual information available and stored electronically has grown at a staggering rate. This has exponentially increased the potential source of paraphrase. More formally, given two sentences S_1 and S_2, such that $S_1 \neq S_2$, when S_1 and S_2 convey the same meaning and are semantically equivalent, they are said to be paraphrased (Agarwal et al. 2017). Many researches on paraphrase detection have focused on the English language, but little effort has been done recently on other languages like Arabic. It is considered as a complex problem because of the challenging features of this language (Mohamed et al 2015). It is Semitic spoken by more than 330 million people and composed of 28 letters written from right to left. In addition, Arabic script has a rich morphologically accentuating by the existence of dots, diacritics and stacked letters (Hkiri et al. 2017, Mansouri et al. 2018, Mahmoud et al. 2018). It is highly inflectional, derivational and non-concatenative compared to other languages (Batita et al. 2018, Mahmoud et al. 2017). To contribute and solve these gaps, recent research has been advancing to propose semantic-similarity-based approaches that have more flexibility and expressiveness compared to syntactic ones. The main objective was to measure the degree of relationship between textual units and cover the maximum of Arabic specificities in terms of word construction and diversity meanings.

3. STATE OF THE ART

Word-embedding models aim a dense representation of words in the form of digital vectors and learned using a variety of language models. In addition, semantic vector representation is able to reveal many hidden relationships between words to enhance the performance of semantic computation and paraphrase detection in different languages, for instance count-based and context-based vector space models.

3.1. Count Based Vector Space Model

Count based vector space models are unsupervised. They rely heavily on the matrix of frequency and the co-occurrence of words. This is done by assuming that words in the same contexts share similar ones or related semantic meanings: Latent Semantic Analysis (LSA) based on the co-occurrence matrix makes it possible to measure the similarity between texts. It represents the meaning not only of individual words, but also of the whole passages, such as sentences, paragraphs and short texts. Based on this idea, Li et al. (2017) used Singular Value Decomposition (SVD) to reduce the dimensionality and suppress the noise of text representation models. They analyzed the optimal number of singular values and calculated the semantic relevance between words combining Term Frequency-Inverse Document Frequency (TF-IDF) weighting and cosine similarity. For experiments, Reuters-21578 data were used with 20 newsgroups and this system achieved about 0.7% of the F-measure. The Latent Dirichlet Allocation (LDA) technique has been one of the most common way of clustering texts. It was a probabilistic model for capturing polysemy (each word has multiple meanings), for example by associating a context with a document. The objective was to reduce the dimensionality of topics as it was used in the work of Dai et al. (2018). They explored semantic topics and author communities for citation recommendation. The experiments were based on the ANN and DBLP datasets and showed that this model could generate qualified author communities and topics. Furthermore, Abdelrahman et al. (2017) detected plagiarism in electronic Arabic resources using heuristic based algorithms, as follows: First, word synonyms were retrieved utilizing the WordNet dictionary. Afterwards,