International Journal of Cognitive Informatics and Natural Intelligence
Volume 17 • Issue 1

Efficacy of Deep Neural Embeddings-Based Semantic Similarity in Automatic Essay Evaluation

Manik Hendre, Ramanbyte Pvt. Ltd., India*
Prasenjit Mukherjee, Ramanbyte Pvt. Ltd., India
Raman Preet, Ramanbyte Pvt. Ltd., India
Manish Godse, Pune Institute of Business Management, India

ABSTRACT

Semantic similarity is used extensively for understanding the context and meaning of the text data. In this paper, use of the semantic similarity in an automatic essay evaluation system is proposed. Different text embedding methods are used to compute the semantic similarity. Recent neural embedding methods including Google sentence encoder (GSE), embeddings for language models (ELMo), and global vectors (GloVe) are employed for computing the semantic similarity. Traditional methods of textual data representation such as TF-IDF and Jaccard index are also used in finding the semantic similarity. Experimental analysis of an intra-class and inter-class semantic similarity score distributions shows that the GSE outperforms other methods by accurately distinguishing essays from the same or different set/topic. Semantic similarity calculated using the GSE method is further used for finding the correlation with human rated essay scores, which shows high correlation with the human-rated scores on various essay traits.

KEYWORDS

ELMo, Embedding, Essay Grading, Global Vectors, Semantic Similarity, Sentence Encoder

INTRODUCTION

Automatic Essay Evaluation is one of the oldest research area in the field of Natural Language Processing (NLP). Unlike multiple choice questions and short question answers, an essay is an open ended question. There is no fixed format and one can have multiple ways of writing an essay. Manually grading the essays is a very resource intensive task from the perspective of time and labour. Teachers have to spend their valuable time on grading the essays written by the students. If we have an automatic essay grading system then teachers can devote more time on the teaching part. An essay is used to assess one's understanding of the particular language. Because of which, TOEFL (2019) and GRE (2019) like exams has essay writing as one of the main component. Since last 5 decades researchers are developing solutions for automatic essay grading systems (Page, 1968; Christie, 1999;

DOI: 10.4018/IJCINI.323190 *Corresponding Author

Rudner et al., 2006). In Natural Language Processing field there has been many advancements in last couple of years. We have more powerful language models which can perform various tasks as par with humans (Young et al., 2018). In tasks like sentiment Analysis, Chatbot, Question Answering, Automatic Essay Evaluation, Dialogue Systems, Parsing, Word-sense disambiguation, Named-Entity Recognition, POS Tagging and many more, we are observing good results (Young et al., 2018; Khurana et al., 2017; Cambria & White, 2014). The computing resources are more available and affordable now, as compared with couple of years back. Due to this, the research in NLP using Deep Learning Techniques is taking new leap in every field (Otter et al., 2020; Young et al., 2018; Deng & Liu, 2018).

In this paper, different neural embeddings are used to check their efficacy in automatically evaluating the essays by considering the semantic similarity. Survey of different word embedding methods have been performed by Wang (2019). Specifically, six word embedding models have been evaluated on different natural language processing tasks. In this, authors points out that currently there are no metrics available for evaluation of word embedding models. The semantic and syntactic relations captured by word embedding models are different from how human beings, understand languages (Wang et al., 2019). Most of the Word embedding models are task specific because they are trained for specific natural language processing tasks. In many NLP tasks, we need to compare different set of texts. For natural language understanding, only keyword matching while similarity checking is not sufficient. The semantics have very important role to play in natural language generation and understanding. There are many ways in which same text can be written having the same meaning. The semantics tries to capture this meaning from different text data. In this paper, we are going to calculate semantic similarity using different neural embedding techniques on an essay data. Automatic Essay Evaluation using Word-Mover Distance is proposed by Tashu & Horvath (2018). In This Semantic similarity of text is given more weightage than the syntax and vocabulary. For calculating essay score, the word-mover distance between Normalized Continuous Bag-of-word features is calculated (Wang et al., 2019). Semantic similarity based on knowledge graphs is proposed in (Zhu & Iglesias, 2016). Most of the semantic similarity techniques uses only surrounding words while computing semantic similarity. Knowledge graphs represent concepts and complex relationships can be extracted from them (Zhu & Iglesias, 2016). Semantic similarity in academic articles where length of the document is more based on word embeddings, is proposed in (Liu et al., 2017). To improve the accuracy, authors have proposed to create semantic profile for each article which then will be used along with word embeddings to calculate similarity. Semantic similarity between two words, sentences and paragraphs is presented by Pawar & Mago (2019). In this, sentence similarity is computed in two phases, first phase the similarity is maximized using word, sentence and word-order similarity. In second phase, the skewness is removed which was introduced because of deviation from actual similarity. Automatic evaluation of text using word and sentence embeddings is proposed by Clark et al. (2019). Authors have introduced a new metric sentence mover's similarity which is the extension of word mover distance for multiple sentences. Sentence mover's similarity metric has improved correlation with the human judgment scores on automatic text evaluation task (Clark et al., 2019). In semantic similarity context of a word is important. Context Representation method using bi-direction LSTM is proposed in (Melamud et al., 2016). Few of the recent and notable contribution in the field of an automatic essay evaluation are reviewed in Table 1.

The main contribution of this paper is to calculate neural embeddings based semantic similarity score to be used in an automatic essay evaluation. The rest of the paper is organized as follows. In Section 2, we list all the studied neural embedding techniques. Datasets used are explained in the Section 3. Proposed methodology and the performance evaluation techniques are explained in Section 4. Experimental results are presented in Section 5. Finally, the conclusions are drawn in Section 6.

NEURAL EMBEDDING TECHNIQUES

For every NLP task the numerical vector representation of text data is very important. Most of the machine learning and deep learning techniques require numeric vectors as an input to the system.