Abstract Retrieval over Wikipedia Articles Using Neural Network

Falah Hassan Ali Al-akashi, University of Kufa, Najaf, Iraq

ABSTRACT

In this article, we propose a neural network model to create a Wikipedia article summarization for each query to allow users to find summary of the topic without going through the whole content in the article. Often, Wikipedia returns the articles related to a search query that makes obvious finding the relevant topic for the user. Text summarization is generated by extracting all those important sentences that are most significant in its topics and have a strong match in its content. Experimentally, each sentence in the article content is encoded as a set of features and presented as an input to the network. The proposed neural network is trained using a set of randomly selected typical articles from Wikipedia. The network output is then used to predict the sentences as a summary of content from the searched query. The results showed that the proposed approach is robust and efficient at finding relevant summaries for most searched queries. Evaluation of the proposal yields accuracy scores of 0.10317 in ROUGE-N and 0.13998 in ROUGE–L.

KEYWORDS

Ensemble Neural Network, Kohonen Map, Parts of Speech, Text Mining, Wikipedia

1. INTRODUCTION

Text summarization (TS), sometimes called text simplification, deals with transforming the original text into simplified variants to increase its readability, usability, and understandability (Surya & Mishra, 2018). It is an information retrieval method to computationally generate a summary of text. The technique has been studied for decades but today with the exponential increasing of data in Internet and World Wide Web it has become much more important than ever. Most text summarizers are extraction-based algorithms which means generating summary by extracting sentences from the text. Some summarization algorithms are capable of producing summaries that contain not only sentences that are presented in the article but also new automatically constructed phrases that are added to the summary to make it more readable. Sometimes, this functionality makes the summarization algorithm more powerful due to improving the comprehensibility of the output summary. In practice, the automatic construction of phrases is fairly difficult task and there is no guarantee that the new phrases will be really meaningful for the user. Moreover, it mostly seems that it is impossible for a web user to find and utilize information relevant to the user interest since the dramatic increasing of data available on the Internet.

Query dependent multi-document summarization that used graph-based approach has been showed the efficiency of document summarization (Bhaskar & Bandyopadhyay, 2010) (Chali & Hasan, 2012). Also, cosine similarity in multi-document summarization has been improved its capability of producing relevant abstract summary (Ranjan et al., 2016); whilst, Wikipedia has been...
showed its ability to provide strong topics for different documents summarization (Gong et al., 2014). Wikipedia is a largest repository and a rich source of information and knowledge made user possible for finding information and knowledge (Al-akashi, 2014). Most researchers preferred Wikipedia as a good reference for document summarization and simplification. Ramanathan et al. (2009) used an algorithm for mapping document sentences to semantic concepts in Wikipedia and select sentences for summary using the frequency of the mapped-to-concepts.

Nevertheless, user needs to find information from the plethora of strong resources; e.g. Wikipedia. Generally, search engines response to the query search consisting of list of matching knowledge retrieved from resources; but still somewhat requires on the part of the user to go through the content to check whether information or knowledge is relevant to his/her perceptions. However, text summarization is not limited to the syntax similarity but also to the semantic similarity or sometimes to the role of labelling. Semantic levels deploy ‘WordNet’ and sentence compression for enabling sentence generation, or semantic role labelling to get semantic representation of text, and then using segmentation to form clusters of related pieces of text (Bhartiya & Singh, 2014).

Neural Network is another solution for text summarization, simplification, and abstracting to tackle the obstacles of data bulk (Kaikakh, 2004) or news articles (Kaikakh, 2015). Ensemble neural networks are neural networks consist of cluster of networks each having the same form where the goal is to extract all those sentences in the article that related to the user expectations. Words are tagged in the sentence and in the query with the parts-of-speech and the process identifier to compare them to extract related sentences. This helps a user to find all those abstracts that are related to the user’s query. User perception can predict whether particular information and knowledge is relevant or solely needs to be summarized to circumvent its reading difficulties. It is commonplace to define a part-of-speech as a linguistic category of lexical items (generally words) that share common syntactic or morphological characteristics. Generally, grammarians often divide the English language into eight parts-of-speech: the noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. The deployed neural network is explained in depth in the following Sections. Words “text” and “document” are used interchangeably throughout this manuscript and they both refer to the Wikipedia article.

2. ENSEMBLE NEURAL NETWORKS

Ensemble Neural Network combines the outputs of several neural networks. The output of an ensemble is average weight of the output of each network. The ensemble weights determined as a function of the relative error of each training network. Individual networks in the ensemble are trained and employed the same data but their synaptic weights are initialized with different values (Perrone, 1993). In several papers and previous work, it has been shown that the ensemble network has a generalization error and generally smaller than that obtained with a single network and the variance of ensemble is less than the variance of a single network. Thus, the ensemble neural network provides better generalization in classification problems than a single network. However, the output of basic ensemble method consists of the output of all the networks weighted equally, as stated in the following function.

\[
F(n) = \frac{2}{M} + \sum_{i=1}^{M} F_i(n)
\]

where \( M \) is the number of the individual neural networks in the ensemble, \( F_i(n) \) is the output of network \( i \) on the \( n^{th} \) training pattern, and \( F(n) \) is the output of the ensemble on the \( n^{th} \) training pattern.
Related Content

Application of Artificial Neural Computation in Topex Waveform Data: A Case Study on Water Ratio Regression
Bo Zhang, Franklin W. Schwartz and Daoqin Tong (2009). International Journal of Software Science and Computational Intelligence (pp. 81-91).
www.igi-global.com/article/application-artificial-neural-computation-topex/34090?camid=4v1a

Classifier Ensemble Based Analysis of a Genome-Wide SNP Dataset Concerning Late-Onset Alzheimer Disease
Lúcio Coelho, Ben Goertzel, Cassio Pennachin and Chris Heward (2010). International Journal of Software Science and Computational Intelligence (pp. 60-71).
www.igi-global.com/article/classifier-ensemble-based-analysis-genome/49132?camid=4v1a
Machine Learning
[www.igi-global.com/chapter/machine-learning/56127?camid=4v1a](www.igi-global.com/chapter/machine-learning/56127?camid=4v1a)

Aircraft Aerodynamic Parameter Estimation Using Intelligent Estimation Algorithms
[www.igi-global.com/chapter/aircraft-aerodynamic-parameter-estimation-using-intelligent-estimation-algorithms/213039?camid=4v1a](www.igi-global.com/chapter/aircraft-aerodynamic-parameter-estimation-using-intelligent-estimation-algorithms/213039?camid=4v1a)