Vol 8. No. 2 Issue 2 – August, 2015
African Journal of Computing & ICT

www.ajocict.net

An Ontology Based Text Document Summarisation Using Statistical Approach

K-K. A. Abdullah
Olabisi Onabanjo University
Ago Iwoye, Ogun State. Nigeria.
+2348060046592,
uwaizabdullah9@gmail.com

ABSTRACT

Due to the development in technology, huge amount of information is available in all over the places. Reading the entire document to understand the content of the document pose a lot of problem, consume space and time considering the amount of information available on the text document. Most existing text summarisation is based on convention term weight but with WordNet text summarisation, the semantic analysis of the text is of major importance. In this paper, hierarchical concepts of the ontology are extracted to interpret the text document for summarisation with concept weighting model.

Keywords: Text Summarisation, WordNet Ontology, Concept, Hierarchical

African Journal of Computing & ICT Reference Format:
K-K. A. Abdullah (2015): An Ontology Based Text Document Summarisation Using Statistical Approach.
Afr J. of Comp & ICTs. Vol 8, No. 2, Issue 2. Pp 1-10.

1. INTRODUCTION

Automatic text summarization is the process of creating a document from one or more document textual sources by reducing the size of text and preserved the information content in the original source documents [1]. Therefore, the information and other characteristics of the text documents depend on the use of summary, however, summary construction is a complex task that involves the use of natural language capabilities [2]. This involves text analysis, text understanding and the use of domain knowledge to tackle the problem. The domain knowledge approach build a semantic representation of the summarisation task such as ontology knowledge [3] and the pattern describing concepts structure [4]. The knowledge from WordNet as well as from UML (a medical ontology) is shown to improve the performance of summarization [5].

The summarization approaches that use statistical method are based on term or concept relevance, combining the approach with thesaurus such as WordNet to extracting semantic relations of words (synonym, antonym) give a better summarisation. The semantic relations are constructed and are used for extract important sentences in a document. In [6], summarisation method that combines several domain specific features with some other known features such as term frequency, title and position are used. Automatic text summarization can be classified into two categories: abstraction and extraction [7]. The extraction method selects a subset of the word, phrase or sentence in the original text document to form summary. The extraction method is easier but the summaries produce much less accuracy compared to human made summaries.

The abstraction method builds an internal semantic representation and the use of natural language capability to create a summary that is closer to what a human might generate. The abstractive summary summary might contain words not explicitly present in the original text document [8]. Text document summarisation can be done on single or multiple documents. Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic.

Extraction of the keyword in the text document should be based on noun which is a subtask of part-of-speech (POS). This enable the carry of most sentence meaning, however lead to a better semantic representation of text. This is based on the ideal that nouns are used as the most important terms (features) that express a documents meaning in Natural Language Processing. The preprocessing is done on the text document to reduce the number of feature. However, different algorithms of text summarisation are based on supervised and unsupervised techniques. The Unsupervised technique use linguistic and statistical information that are obtained from the document itself., which will be our main focus. There are different algorithms which are based on supervised or unsupervised techniques. Supervised techniques use data sets that are labelled by human annotators. Unsupervised approaches do not use annotated data, but use linguistic and statistical information that are obtained from the document itself. However, with ontology-based summarization, computations of the set of features for each sentence are based on the output of the hierarchical structure.

The rest of the paper, we discuss background information about WordNet and related work in section 2. Section 3 outlines the overview of text summarisation different approaches and deduces some useful conclusions about them. Section 4, our proposed method are presented. In section 5, we conclude and discuss our future work.

2. RELATED WORK

Research in automatic text document summarization has received considerable attention in today's fast growing age due to the increase growth in the volume and complexity of information sources on the web. The information and knowledge gained can be used for applications such as business management, production control, and market analysis, to engineering design and science exploration. Reduction in summarisation can significantly improve the conciseness of automatic summaries. Text summarisation is texts that are produced from one or more documents, which contains a fundamental portion of the information from the original document with a reduced size. Human summarisation reflects the understanding of the topic, synthesis of concepts, evaluation, and other processing. Consequently, the result is different from the original document but not explicitly stated, this requires the access to knowledge separate from the input. However, computers do not yet have the language capabilities of human, therefore there is need for alternative methods. By using the automatic summarization, these problems are solved. In automatic text summarization, selection-based approach has so far been obsolete. In some approach, summaries are generated by extracting keyword text segments from the text, based on analysis of features such as term frequency and location of sentence to locate the sentences to be extracted. Keyword extraction from a body of text relies on an evaluation of the importance of each candidate keyword [9].

The term-based mapping of sentences to ontologies was proposed by [10]. It exploits the concept generalization method offered by WordNet relations to find the most informative concepts contained in a text. However, [11] presented the semantic free-text summarization system using ontology knowledge to generate the summarisation. The system retrieves and ranks information according to a user's query with no threshold value for selecting the sentences in the documents. Sentence reduction systems for automatically removal of extraneous phrases from sentences are presented in [12] which are extracted from a document for summarisation purpose. A multiple sources of knowledge were used to decide which phrases in an extracted sentence can be removed such as syntactic knowledge, context information and statistics computed from a Corpus. A query based document summarizer based on similarity of sentences and word frequency is presented in [13].The system find similar sentences to the query and sum focus to find word frequency using Vector Space Model (VSM). Grouping similar sentences and word frequency removes redundancy and collect the required documents and produces summary.

2.1 Approaches of Text Document Summarisation

2.1.1 Statistical Approaches

The text summarization approaches that use statistics can be based on two different methods: The first method is based on concept relevance and Bayesian classifier. This approach uses word frequency, uppercase words, sentence length, keywords and phrase structure. the second one used algebraic statistical methods such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), and Semi-discrete Matrix Decomposition (SDD) for document summarization. The LSA, algorithms are mostly used which is based on singular value decomposition (SVD). In this algorithm similarity among sentences and similarity among words are extracted.

2.1.2 Text Connectivity Based Approaches

The approach referenced the already mentioned parts of a document. The methods are used by lexical chains and Rhetorical Structure Theory (RST). Lexical chain extracts semantic relations of words (synonym, antonym) that is the concepts using dictionaries and WordNet. Using semantic relations lexical chains are constructed and used for extracting important sentences in a document while RST organizes text units into a tree like structure. Then this structure is used for summarization purposes.

2.1.3 Graph Based Approaches

The approach is similar to text connectivity based approach but the concepts of the semantic relation are in form of node and edges. The nodes in graph based summarization approaches represent the sentences, and the edges represent the similarity among the sentences. The similarity values are calculated using the overlapping words or phrases. The sentences with highest similarity to the other sentences are chosen as a part of the resulting summary. Example such as TextRank and Cluster LexRank are methods that use graph based approach for document summarization.

2.1.4 Non-Extractive Summarization Methods

Abstractive summarization methods try to fully understand the given documents, even non-explicitly mentioned topics and generate new sentences for the summary. This approach is very similar to the way of human summarization. There are approaches that create summaries in a non-extractive manner, using information extraction, ontological information, information fusion and compression.

2.2 WordNet Ontology

WordNet [14] is a machine-readable lexical database for English widely used in computational linguistics community developed at Princeton University. It aims to represent some aspects of the semantics of the lexicon, and the relationships of different lexicalized concepts. The database consists of linked words, primarily nouns, verbs, adjectives and adverbs. These words are organized into synonym sets called synsets, and connected by three lexicon-semantic relations. hypernym, and meronym.

Vol 8. No. 2 Issue 2 – August, 2015
African Journal of Computing & ICT

www.ajocict.net

The nodes of each ontology hierarchy control the level of generalisation and give a chain of synsets connected by hypernym. This defines the limit in the chain where the synsets are used. The need to ensure that whenever related words with different domains share the same hypernym, this common hypernym is not too high in the chain. Otherwise, this common synset could be too abstract to yield any usefulness. However, the hypernym is too high in the hierarchy to have interesting consequences Using WordNet ontology model, the hierarchical representation is generated for the concept terms. Therefore, collection of keyphrase from the text document are compared with Synsets

3. THE PROPOSED METHODOLOGY

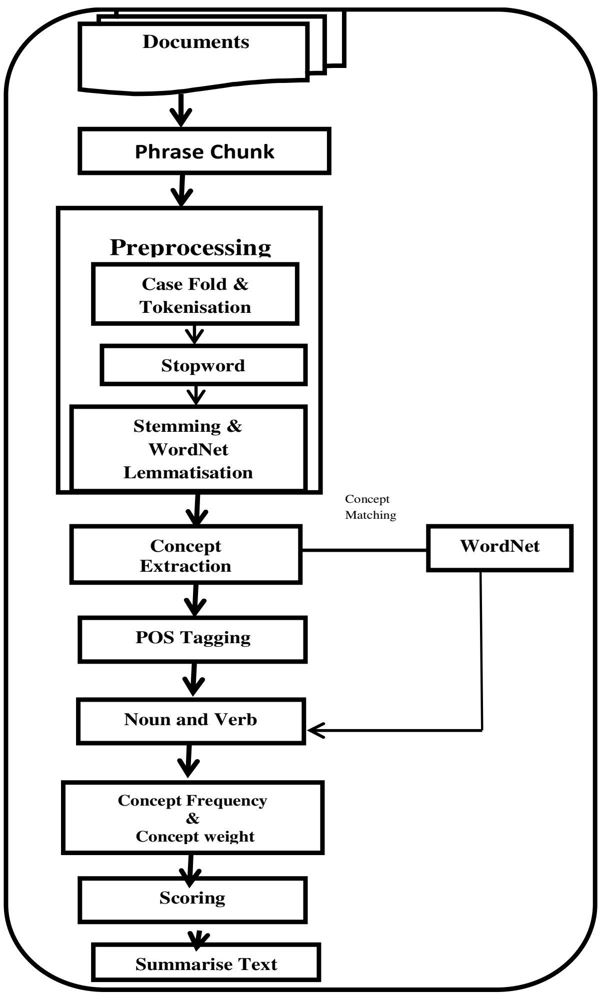

Figure 1: The Proposed Text Document Summarisation Architecture

The Figure 1 represents the proposed system which is based on extraction summarisation method to perform the automatic summarization.

The extraction method filter uninformative words of low semantic content that is most grammatical words ("the"", "of " ,"as", "in"). The preprocessing can be perform by using Natural Language ToolKit (NLTK) which is a primary step. The text document are load the into the proposed system and decompose the given text into its constituent sentences. The process such as case folding transforms the text into its lower case to improve the accuracy of the system. The stopwords are are remove from text to reduce the size of the candidate features. Consequently, each word are reduce to its respective word form, however, WordNet lemmatisation are perform on the words that are not in the stemming form such as "teeth and tooth" . Word tags assign part-of-speech (POS) such as (noun, verb, and pronoun, etc.) to each word in a sentence to give word class. The input to a tagging algorithm is a set of words in a natural language and specified tag to each in the sentence. In the tagging process look for the token in a lookup WordNet dictionary.

Concepts are extracted using concept extraction algorithm. By using the ontology model the hierarchical representation is generated for the concept terms. The system uses the concept frequency to determine how important a features is and distinguish relevant word or phrase.D

$$Weight(w) = cf * idf$$

Eqn. 1

$$idf = \log\left(\frac{|D|}{df + 1}\right)$$

Eqn. 2

$$cf\,/\,idf = cf \; * \log\left(\frac{|D|}{df + 1}\right)$$

Eqn 3.

Where

$cf = concept\ frequency$

$idf = inverse\ document\ frequency$

$D = Document\ collection$

Concept weighting approach in equation 1 is used for picking the valid sentences in the text document. In this approach, a set of frequencies and concept weights based on the number of occurrences of the words is calculated. Summarization methods based on semantic analysis also use concept weights for final sentence selection. The model is then used to extract important sentences from the documents. The system computes the weight for each concept keyphrase using all the features. The weight represents the strength of the keyphrase, the more weight value the more likely to be a good keyword (keyphrase).

African Journal of Computing & ICT

www.ajocict.net

The results of the extracted keyphrases are the input to the text summarization. The range of scores depends on the input text. The system selects keywords with the highest values. The ranked sentences have been selected and the summary is generated based on weight of the sentences for a given sample document. The system works dynamically and was implemented in Python.

5. CONCLUSION AND FUTURE WORK

This paper described the technique to generate document summarisation with details of each step. It presented ontology-based summarization system using statistical method. Ontology knowledge is proven to be an effective concept based methods. It shows how sentences can be mapped to nodes of a ontology. The mapping provides a semantic representation of the information content of sentences that improves summarization quality. Another interesting research is the transfer of the approach to the more general case of a non-hierarchical ontology.

REFERENCES

[1] Sparck-Jones, K. 1999. Automatic summarizing: factors and directions. In Mani, I.; Maybury, M. *Advances in Automatic Text Summarization*. The MIT Press 1-12.

[2] Mitra, M., Singhal, A. and Buckley, C. 1997. Automatic text summarization by paragraph extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*. Madrid

[3] Fiszman, M., Rindflesch T., Kilicoglu, H. 2004. Abstraction Summarization for Managing the Biomedical Reearch Literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004.

[4] Barzilay, R. Elhadad, N. and McKeown. K. 2002, Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17:3-8

[5] Verma, R. Chen, P. and Lu. W. 2007. A semantic free-text summarization system using ontology knowledge. In Proc. of the 2007 Document Understanding Conf. (DUC 07).

[6] Kamal Sarkar, 2009. Using Domain Knowledge for Text Summarization in Medical Domain", International Journal of Recent Trends in Engineering, 1.1:200-205.

[7] Gupta, V., Singh Lehal, G. 2010. A Survey of Text Summarization Extractive Techniques, Journal of emerging technologies in web intelligence, Aug. 2 3,.

[8] Rafeeq Al-Hashemi 2010. Text Summarization Extraction System (TSES) Using Extracted Keywords. International Arab Journal of e-Technology, June 1.4:164-168.

[9] Buyukkokten, O. , Garcia-Molina, H. and Paepcke, A. 2001. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In Proceedings of Tenth International World Wide Web Conference, 652–662.

[10] Hovy E. and. Lin, . C.-Y 1999 Automated text summarization in summarist. In I. Mani and M. Maybury, editors, Advances in Automated Text Summarization, MIT Press. 18–94.,

[11] Verma, R., Chen, P. and Lu, W. 2009. A Semantic Free text Summarization System Using Ontology Knowledge, IEEE Transactions on Information Technology in Biomedicine. 5.4: 261- 270.

[12] Hongyan Jing. 2000. Sentence Reduction for Automatic Text Summarization"; Proceedings of the sixth conference on Applied natural language processing, Seattle, Washington, 310 – 315.

[13] Siva kumar A. P.,. Premchand , P. and. Govardhan, A. 2011 Query-Based Summarizer Based on Similarity of Sentences and Word Frequency, International Journal of Data Mining & Knowledge Management Process, May 1..3.

[14] Miller George A. (1990) *WordNet: An On-Line Lexical Database*. In Special Issue of International Journal of Lexicography, 3. 4.