International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Text Summarization Using Synset Ranking

Samuel V Thomas[1], Jayvardhan[2], Madan Lal Yadav[3]

[1]*MTech CS&E, ASET, Amity University, India*
[2]*MTech CS&E, ASET, Amity University, India*
[3]*Asst. Professor, ASET, Amity University, India*

Abstract— **Text summarization has gained its popularity over the Internet due to its abundance of resources which are not being absorbed efficiently. This is because relevant information retrieval from huge corpuses has become difficult. We propose a novel approach to extractive text summarization using ontology. This ontology based approach helps in finding the most relevant passage.**

Keywords— **text summarization, ontology, text mining, natural language processing, information retrieval, synset ranking**

I. INTRODUCTION

There has been an abundance of resources over the internet, and more than ever there has been a great need for retrieving relevant information to a particular user query because searching of documents in a large corpus was quite overwhelming. Text had to be summarized over the internet for various reasons such as results provided by a search engine, comparison purposes of text and so on. It is nearly impossible to manually summarize large texts over the internet and thus there is a need for automation in this field to reduce overhead in time and effort. Automatic Text Summarization has played a major role in the field of Information Retrieval. Summarized text refers to a shorter version of the original text without losing its overall concept or meaning. There are majorly two approaches to text summarization which is extractive and abstractive summarization [5].

The first approach, i.e. extractive summarization involves the selection of relevant sentences from the text and condensing the text into its corresponding shorter version without losing its overall meaning [5]. The relevance of the sentence could be measured on the basis of statistical features. The second approach, i.e. abstractive summarization involves the concept extraction from the text and re-expressing it in a shorter form [5]. This includes linguistic methods for extracting the concepts and conversion of text into its shorter version.

In this research paper, we have proposed a generic algorithm i.e. not language or domain dependant, based on extractive methodologies which use ontology, in our case WordNet[2] to identify semantic relations in order to strengthen the relevance features extracted in the text. The relevance features are identified using the traditional statistical approach such as term frequency-inverse document frequency.

II. LITERATURE SURVEY

A lot of research has already been done in the case of text summarization. The approaches can be majorly divided into:

A. Statistical based approach

Weights are assigned to the words, sentences or phrases, etc included in the document which is assumed to find more relevance. These weights can be assigned using a number of algorithms which are term frequency-inverse document frequency, mutual information, information gain and residual inverse document frequency. Term frequency-inverse document frequency assigns weights on the basis of the frequency counts of the words. Mutual information helps in identifying the dependency between two words or in other words it helps in the identification of common information. Information gain helps in evaluating an attribute to its relevance in the document and is a metric based technique. The Residual inverse document frequency is quite similar to inverse document frequency where it uses the Poisson distribution.

B. Topic based approach

It follows the top-down methodology as the topic is initially identified around which the concept of the text is built. The topics are identified by topic signatures which are a collection of terms identified to represent the topic, enhanced topic signatures provide important relations between the concepts, thematic signatures help in the ranking of documents as the documents are segmented based on these themes, modelling the content structure of the documents with templates helps in identifying certain specific entities.

C. Graph based approach

This approach is unique as it helps in the visualisation of the terms and their associated semantic links between them. The nodes of the graph could be terms or phrases where their corresponding edges could be represented by the semantic relation between them. One example of such an approach is LexRank.

D. Discourse based approach

Linguistic perspective is required for such an approach and helps in the summarisation of text in more generic fashion. Thus the focus is on identifying lexical chains of the terms. It is assumed that with the help of linguistic knowledge the summarisation algorithm can improve its efficiency.

E. Machine learning based approach

A learning algorithm can help in gaining intelligence from a training set to rank and select the sentences based on relevance. Support Vector Regression (SVR) and Least Angle Regression (LAR) are examples of such an approach. However such an approach would require a huge training corpus to obtain conclusive results. The learning algorithm is probabilistic in nature.

III. PROPOSED METHODOLOGY

Our methodology consists of four major phases.

A. Text Pre-processing

Input text that is received is not normalized in nature. This kind of data is difficult to be worked upon. Thus this phase helps in removing the noise from the input text and to present the most relevant information from the text for processing in the later phases. This phase in itself has four steps:

1. Parsing the text simply involves identifying words and sentences in the text. This can be done by identifying the spaces, punctuations, and other non-alphanumeric characters. Most programming and statistical languages contain character procedures that can be used to parse the text data. The result from this phase is a bag of words collected from the text.
2. Removal of stop words from the bag of words retrieved in the previous stage helps in eliminating noise too. These stop words could refer to articles, prepositions, conjunctions, and other commonly occurring words. These have to be removed as they do not form the concept of the text.
3. Morphological analysis of the words to reduce the word to its stem word. This process is also known as stemming. A stem of a word is its corresponding root word that would provide the same meaning as the word itself. This helps in identifying and grouping the common stems from the bag of words, thus removing redundant words.
4. Filtering of nouns and verbs from the bag of words as they majorly constitute the concept of the text. This is done by using part of speech tagging which helps in identifying to what part of speech a particular word belongs to.

B. Weighted Terms Selection

The previous phase returns a bag of words which are repetitive in nature. The objective of this phase is to select all the unique occurring words with its corresponding frequency by which they occur in the text. Term frequency – Inverse Document Frequency is quite popular in calculating a term's frequency. Based on the frequency identified the terms are rearranged in their descending order of frequency. These terms are also known weighted terms with their frequency being their weight. At this stage, the list of weighted terms that have been identified could be large in number. To reduce overhead in processing these terms, the list is reduced from a fixed depth. This depth is dependent on the size of the text. For example in our case we only select the first 20% of the weighted terms from the list.

C. Synset Ranking

Each word constitutes a list of synsets or a synonym ring. Each synset of a word corresponds to the different kind of contexts the word may appear in. Synsets are identified from the list of weighted words. The objective of this phase is to select all the unique occurring synsets with its corresponding frequency by which they occur in the text. Term frequency – Inverse Document Frequency is quite popular in calculating a synset's frequency. Based on the frequency identified the synsets are rearranged in their descending order of frequency. These synsets are also known weighted terms with their frequency being their weight. At this stage, the list of weighted synsets that have been identified could be large in number. To reduce overhead in processing these terms, the list is reduced from a fixed depth. This depth is dependent on the size of the text. For example in our case like the previous phase we only select the first 20% of the weighted terms from the list.

D. Sentence Ranking and Selection

Sentences are ranked on the basis of the terms relevance to the selected synset list and their associated ranks. Based on these ranks the algorithm selects the most relevant sentences. Now for each sentence, $S[i]$ is the rank of the i^{th} sentence calculated and N is the total number of unique terms occurring in the sentence. The synset_frequency is the frequency of the synset calculated in the previous phase and synset_count is the number of occurrences of the same synset.

$$S[i] = \sum_{n=1}^{N} (synset\_frequency \times synset\_count) \div N$$

E. Final Filtering

This phase focuses on the linguistic perspective. Initially sentences with the same semantics are identified as redundant, so they are removed as they do not add up to the overall meaning of the content. These redundant sentences are removed using a technique called sentence disambiguation. The simplest approach which we have proposed for sentence disambiguation is to identify the triplets in each sentence. These triplets identify mainly comprises of verbs and nouns. The nouns signify the subject and the object whereas the verbs represent the action. In short the sentences with similar action, subject and object should be treated as redundant.

The next step includes the removal of unlinked references in a sentence such as a sentence starting a pronoun. These pronouns have to be replaced by their respective proper nouns. This can be done by maintaining a list of lexical chains while parsing.

Finally filtered sentences are reduced by traversing the parse tree of the sentence and by removing the child elements that are grammatically not relevant to the meaning of the sentence [9].

IV. IMPLEMENTATION

We implemented our approach with the help of several open source libraries available. We chose python as our programming platform as it had a great support for natural language processing. One of its most popular natural language processing libraries is NLTK[3]. This library provides a huge variety of text processing utilities such as parsing, word tokenization, part-of-speech tagging, chunking, named entity recognition and also has a support for popular corpora such as WordNet[2].

Synset ranking requires the usage of ontology for example WordNet to club the similar terms together. MultiWordnet[6] provides with an API in java to find the synset, also it provides with the SQL dump files of the English and the Spanish WordNet corpus. For building a WordNet API from scratch, the WordNet SQL builder helps in providing the support for WordNet corpus and its associated semantic utilities. NLTK already provides a plethora of corpuses including WordNet and its associated semantic associations and functions.

V. CONCLUSION

We have successfully proposed a novel approach for extractive text summarization that is based on extracting and highlighting the most relevant information in the text. This relevance is improved by an ontological approach in combination to the traditional statistical one.

V. FUTURE WORK

This research paper proposes a novel approach whose other interesting possibilities could be worked upon in the future. These possibilities could be in the category of refining the ranking criteria. For example, the synset ranking algorithm can be refined using a probabilistic approach over a skewed dataset. This would also improve the overhead of the system and increase its speed of execution. This probabilistic approach would help in reducing the complexity of traversing the whole WordNet ontology.

The other perspective that can be viewed on an exploratory analysis is the linguistic behavior. Theoretically it is assumed that the overall performance of summarization can be improved by acquiring the knowledge of linguistics. Such knowledge can be used in the final filtering by the help of lexical chains of the terms [5].

Since our approach is generic in fashion, i.e. we have not uses anything that is language dependant, thus we can also use an ontological resource such as WordNet for other languages and extend this approach of text summarization to them.

REFERENCES

[1] R. V. V Murali Krishna, and Ch. Satyananda Reddy, "A Sentence Scoring Method for Extractive Text Summarization based on Natural language queries," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May2012.

[2] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princetonedu

[3] Steven Bird, Ewan Klein, and Edward Loper, "Natural Language Processing with Python". O'Reilly Media Inc. http://nltk.org.book, 2009.

[4] Christos Bouras, and Vassilis Tsogkas, "Improving Text Summarization Using Noun Retrieval Techniques", Springer, I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 593-600, 2008

[5] Andreas Hotho, and Andreas Nurnberger, "A Brief Survey of Text Mining", May 2005.

[6] Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal, Vaibhav Mehta, Ganesh Ramakrishnan, and Pushpak Bhattacharyya, "Generic Text Summarization using WordNet", 2004.

[7] Emanuele Pianta, Luisa Bentivogli and Christian Girardi, "MultiWordNet: Developing and Aligned Multilingual Database". In Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21-25, 2002, pp. 293-302.

[8] Bernardo Magini and Gabriela Cavaglià, "Integrating Subject Field Codes into WordNet" in Gavrilidou M., Crayannis G., Markantonatu S. and Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May – 2 June, 2000, pp. 1413-1418.

[9] Hongyan Jing, "Sentence Reduction for Automatic Text Summarization", Published in ANLC '00 Proceedings of the sixth conference on Applied natural language processing Pages 310-315.

[10] George A. Miller (1995). Wordnet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41

[11] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press