Automatic text Summarization for Abstract Generation

Yogeswari Magar

ABSTRACT-

Now- a- days the volume of data is increasing from variety of sources. So this volume of text data needs to be summarized effectively to be useful. This paper is a comprehensive literature review of Automatic text summarization, an algorithm that extracts sentences from a text document, determines which are most important, and returns them in a readable and structured way.

Automatic text summarization design process includes word frequencies count for the entire text document, sorting of the words, each sentence score calculation, relationship between the sentences and reframing the summary.

Index terms-

Text Summarization, lexical similarity, semantic similarity, summary extraction, sentence ordering and selection.

1. INTRODUCTION-

Automatic text summarization is part of the field of natural Language processing. The main idea of summarization is to find a subset of data which contains the "information" of the entire set. Document summarization tries to create a representative summary or abstract of the entire documents, by finding the most informative sentences. Which is how computers can analyses, understand and derive meaning from human language.

It is also useful for students and authors. Imagine being able to tactically generate an abstract based for your research paper or chapter in a book in a clear and concise way that is faithful to the original source material.

2. PRAPOSED FRAME WORK-

2.1WORD FREQUENCY COUNT FOR THE ENTIRE TEXT-

Automatic summarization of text works by first calculating the word frequencies for the entire text document. The occurrence of each word is calculated, which is present in the document. Then the 100 most common words are stored and sorted. Each sentence is then scored based on how many high frequency words it contains, with higher frequency word being worth more. Here those sentence score more which contain more high frequency words.

Then the sentences are sorted according to the score. Finally the top X sentences are then taken and sorted based on their position in the original text. These are the sentences which are going to be used in the Abstract construction.

2.2 SIMILARITY TEST BETWEEN THE SENTENCES-

Sentences are extracted from the text and then a graph is built linking sentences that are similar. The similarity between the sentences depends on many factors.

1. Lexical similarity-If the words present in sentences are similar, then lexical similarity exists between the sentences.

Here the similarity can be calculated by using the **Jaccard Index**

When taken as a measure of string similarity, the coefficient may be calculated for two strings X and Y using bigrams as follows.

$$S=2n_t / n_x+n_y$$

Where n_t is the n umber of character bigrams found in both strings, n_x is the number of bigrams in string x and n_y is the number of bigrams in string y. For example, to calculate the similarity between:

night
nacht

We would find the set of bigrams in each word:

{ni,ig, gh,ht}
{na,ac,ch,ht}

Each set has four elements, and the intersection of these two sets has only one element: ht.
Inserting these numbers into the formula, we calculate, $s = (2 \cdot 1) / (4 + 4) = 0.25$.

2. Semantic Similarity-If sentences contain words which are similar in meaning, then there exist semantic similarity between the sentences.

The semantic similarity can be find out in the following ways

i) **Structure-**Here Synonyms-words that denote the same concept and are interchangeable in many contexts are grouped into unordered set (Synset). So if the sentences contain the same word then it can be concluded that the two sentences are similar.

ii) **Encoded relation- ER** among synsets is the super subordinate relation (also called hyperonymy, hyponymy or ISA relation).Hyponymy relation is transitive relation which juge the relationship between the sentences.

Ex- if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture.

3. Structural Similarity- If sentences contain similar words as well as structure is also similar, and then there exist structural similarity between sentences. You need to consider relationship exist between words of sentences, word order for structural similarity.

2.3 EXATRACTING THE MOST SIMILAR SENTENCES-

Here the similarity between the sentence is find out.
Here we choose the different similarity measures between the sentences .The probability of similarity between the sentences is calculated. The top most sentences are selected for the summary.

We can here also use the sentence extraction to get new meaningful sentences. Some approaches for extracting meaningful summary are available. One of the Techniques are Natural Language Processing and Text Mining. In this way we can generate a good abstract from a paragraph.

2.4. CONCLUSION-

As the demand for compressive, meaningful abstract is increasing day by day, there should be a appropriate method for Text summarization .Text Summarization can be helpful for student, researcher, business analyst and it can be used in many more fields. Here we have represented a method for this. Till now the methods developed are not sufficient for summary generation and there is still a lot of scope for exploring such method for more meaningful summarization.

References

[1] Smart Computing and Informatics: Proceedings of the First ..., Volume 1,edited by Suresh Chandra Satapathy, Vikrant Bhateja, Swagatam Das.
[2] Abstractive Text summarization by somye Singhal and Arnab Bhattacharya
[3] Abstractive Text summarization using Attentive Sequence –to –sequence RNNs by Elliott Jobson & Abiel Gutiérrez Rajaraman Kanagasabai, Zhuo Zhang
[4] https://en.wikipedia.org/wiki/Integer programming
[5] https://github.com/aneesha
[6] Mari-SannaPaukkeri and TimoHonkela,'Likey: Unsupervised Language-independent KeyphraseExtraction,Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 162–165,Uppsala, Sweden, 15-16 July 2010.
[7] Letian Wang, Fang Li, SJTULTLAB: Chunk Based Method for Keyphrase Extraction, Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010,pp 158– 161,Uppsala, Sweden, 15-16 July 2010.
[8] International Journal on Soft Computing (IJSC) Vol.2, No.4, November 2011.
[9] Key phrase Extraction Based on Core Word Identification and Word Expansion', Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 142–145, Uppsala, Sweden, 15-16 July 2010. Su Nam Kim, ÄOlenaMedelyan,~ Min-Yen Kan} and Timothy BaldwinÄ,'SemEval-2010
[10] Automatic Keyphrase Extraction from Scientific Articles', Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pp 21–26,Uppsala, Sweden, 15-16 July 2010.

[11] FumiyoFukumotoAkina Sakai Yoshimi Suzuki,'Eliminating Redundancy by Spectral Relaxation for Multi-Document Summarization',Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pp 98–102,Uppsala, Sweden, 16 July 2010.
[12]Michael,ChengXiangZhai,RoxanaGirju,'Summarizing Contrastive Viewpoints in Opinionated Text',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 66–76,MIT, Massachusetts, USA, 9-11 October 2010.
[13] You Ouyang , Wenjie Li Qin Lu Renxian Zhang, 'A Study on Position Information in Document Summarization', C Xiaojun Wan, Huiying Li and JianguoXiao,'Cross-Language
[14] Document Summarization Based on Machine Quality Prediction',Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 917–926,Uppsala, Sweden, 11-16 July 2010.
[15] Ahmet Aker Trevor Cohn,'Multi-document summarization using A* search and discriminative training',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 482–491,MIT, Massachusetts, USA, 9-11 October 2010.
[16] Hal Daum´eIII*,DanielMarcu*, 'Induction of Word and Phrase Alignments for Automatic Document Summarization', Computational Linguistics, 31 (4), pp. 505-530, December, 2006 .