A Combined Method of Text Summarization via Sentence Extraction

CHENGHUA DANG
Hebei University of Engineering
Handan, Hebei, 056038
PEOPLE'S REPUBLIC OF CHINA

XINJUN LUO
Hebei University of Engineering
Handan, Hebei, 056038
PEOPLE'S REPUBLIC OF CHINA

Abstract: - In this paper, we propose a practical approach for extracting the most relevant sentences from the original document to form a summary. We present this summarization procedure based upon statistical selection and WordNet. Experimental results show that our approach compares favourably to a commercial text summarizer.

Key-Words: - Text Summarization, Key Sentence, Keyword, WordNet, Synset

1 Introduction

Text summarization is the process of condensing a source text while preserving its information content and maintaining readability. The main (and large) difference between automatic and human-based text summarization is that humans can capture and convey subtle themes that permeate documents, whereas automatic approaches have a large difficulty to do the same. Nonetheless, as the amount of information available in electronic format continues to grow, research into automatic text summarization has taken on renewed interest.

A summary can be employed in an indicative way – as a pointer to some parts of the original document, or in an informative way – to cover all relevant information of the text [1]. In both cases the most important advantage of using a summary is its reduced reading time. Summary generation by an automatic procedure has also other advantages: (i) the size of the summary can be controlled; (ii) its content is determinist; and (iii) the link between a text element in the summary and its position in the original text can be easily established.

Technology of automatic summarization of text is maturing and may provide a solution to this problem [2, 3]. Automatic text summarization produces a concise summary by abstraction or extraction of important text using statistical approaches [4], linguistic approaches [5] or combination of the two [3, 6, 7].

In this paper, we propose a practical approach for extracting the most relevant sentences from the original document to form a summary. The idea of our approach is to exploit sentences from both the Keywords extraction based on statistics and Synsets extraction using WordNet. These two properties can be combined and tuned for ranking and extracting sentences. We provide experimental evidence that our approach achieves reasonable performance compared with a commercial text summarizer (Microsoft Word summarizer).

This paper is structured as follows. In Section 2 we review previous research related to the problem of text summarization and summary evaluation. Section 3 presents our combined method of key-sentence extraction. Section 4 provides

experiments comparing our method to ten other summarization approaches. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Summarization Techniques

Text summarization by extraction can employ various levels of granularity, e.g., keyword, sentence, or paragraph.

MEAD [8], a state of the art sentence-extractor and a top performer at DUC, aims to extracts sentences central to the overall topic of a document. The system employs (1) a centroid score representing the centrality of a sentence to the overall document, (2) a position score which is inversely proportional to the position of a sentence in the document, and (3) an overlap-with-first score which is the inner product of the tf * idf with the first sentence of the document. MEAD attempts to reduce summary redundancy by eliminating sentences above a similarity threshold parameter.

Other approaches for sentence extraction include NLP methods [9, 10] and machine-learning techniques [11, 12]. These approaches tend to be computationally expensive and genre-dependent even though they are typically based on the more general tf * idf framework. Work on generative algorithms includes sentence compression [13], sentence fusion [14], and sentence modification [15].

2.2 Keywords Extraction Techniques

Traditionally, keywords are extracted from the documents in order to generate a summary. In this work, single keywords are extracted via statistical measures. Based on such keywords, the most significant sentences, which best describe the document, are retrieved.

Keyword extraction from a body of text relies on an evaluation of the importance of each candidate keyword [16]. A candidate keyword is considered a true keyword if and only if it occurs frequently in the document, i.e., the total frequency of occurrence is high. Of course, stop words like “the”, “a” etc are excluded.

2.3 WordNet in Text Classification

WordNet [17] is an online lexical reference system in which English nouns, verbs, adjectives and adverbs are grouped organized into synonym sets or synsets, each representing one underlying lexical concept. A synset is a set of synonyms (word forms that relate to the same word meaning) and two words are said to be synonyms if their mutual substitution does not alter the truth value of a given sentence in which they occur, in a given context. Noun synsets are related to each other through hypernymy (generalization), hyponymy (speciali-zation), holonymy (whole of) and meronymy (part of) relations. Of these, (hypernymy, hyponymy) and (meronymy, holonymy) are complementary pairs.

The verb and adjective synsets are very sparsely connected with each other. No relation is available between noun and verb synsets. However, 4500 adjective synsets are related to noun synsets with pertainyms (pertaining to) and attra (attributed with) relations.

Scott and Matwin [18] propose to deal with text classification within a mixed model where WordNet and machine learning are the main ingredients. This proposal explores the hypothesis that the incorporation of structured linguistic knowledge can aid (and guide) statistical inference in order to classify corpora. Other proposals have the same hybrid spirit in related areas: Rodriguez, Buenaga, Gómez-Hidalgo, Agudo [19] and Vorhees [20] use the WordNet ontology for Information Retrieval; Resnik [21] proposes another methodology that index corpora to WordNet with the goal of increasing the reliability of Information Retrieval results.

Scott and Matwin [18], however, use a machine learning algorithm elaborated for WordNet (more specifically, over the relations of synonymy and hyperonymy). This aims to alter the text representation from a non-ordered set of words (bag-of-words) to a hyperonymy density structure.

3 Our Algorithms

3.1 Preprocessing of the text

1) Break the text into sentences.
2) Stop-word elimination – common words with no semantics and which

do not aggregate relevant information to the task (e.g., "the", "a") are eliminated;

3) Case folding: consists of converting all the characters to the same kind of letter case - either upper case or lower case;
4) Stemming: syntactically similar words, such as plurals, verbal variations, etc. are considered similar; the purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics.

3.2 Synsets Ranking

The basic motivation of this step is to rank the synsets based on their relevance to the text. So, if lots of words in the text correspond to the same synset, that synset or 'meaning' is more relevant to the text, and thus, it must get a higher rank. This idea has been borrowed from [22], which details the use of WordNet Synsets as a mode of text representation.

3.3 Refinement of Keywords

The collection of Keywords are refined as compared with Synsets obtained above. The comparison is conducted by calculating the similarity between Keywords and Synsets. According to the vectorial model, this feature is obtained by using the Synsets of the document as a "query" against all the Keywords of the document; then the similarity of the document's Synsets and each Keyword is computed by the cosine similarity measure [23]. Then we retain those Keywords which have the closest similarity to the Synsets.

3.4 Key-Sentence Selection

Once the keywords are identified, the most significant sentences for summary generation can be retrieved from all narrative paragraphs based on the presence of keywords [24]. The significance of a sentence is measured by calculating a weight value, which is the maximum of the weights for word clusters within the sentence. A word cluster is defined as a list of words which starts and ends with a keyword and less than 2 non-keywords must separate any two neighboring keywords [16]. The weight of a word cluster is computed by adding the weights of all keywords within the word cluster, and dividing this sum by the total number of keywords within the word cluster.

The weights of all sentences in all narrative text paragraphs are computed and the top five sentences (ranked according to sentence weight) are the key sentences to be included in the summary.

The overall summary is formed by the top 25 keywords and the top 5 key sentences. These numbers are determined based on the fact that key sentences are more informative than keywords, and the whole summary should fit in a single page.

4 Experiments

Summaries can be evaluated using intrinsic or extrinsic measures [25]. While intrinsic methods attempt to measure summary quality using human evaluation thereof, extrinsic methods measure the same through a task-based performance measure such the information retrieval-oriented task.

Intrinsic approach was utilized in our experiments. However, it is a time-consuming process to identify important units in documents by humans, therefore, we chose the Microsoft Word summarizer of MS Office 2000 to output summary baselines.

The comparison between our algorithm and the summarization algorithm for MS Word 2000 demonstrates that our experimental results give the best summarization at around 35% summary of a document.

5 Conclusion

We have presented a combined technique for the extraction of key-sentences from a document, and use such sentences as a summary of the same document. Refining Keywords against WordNet Synsets comprehensively improve the correctness of automatic summary.

References:

[1] Mani, I. Automatic Summarization. J.Benjamins Publ. Co. Amsterdam Philadelphia (2001).

[2] I. Mani and M. Maybury. Advances in Automatic Text Summarization. MIT Press, ISBN 0-262-13359-8, 1999.

[3] I. Mani. Recent developments in text summarization. In ACM Conference on Information and Knowledge Management, CIKM'01, pages 529–531, 2001.

[4] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In Proceedings of 10th International World-Wide Web Conference, 2001.

[5] C. Aone, M.E. Okurowski, J. Gorlinsky, and B. Larsen. A scalable summarization system using robust NLP. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 66–73, 1997.

[6] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain, 1997.

[7] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In Proceedings of SIGIR, pages 121–128, 1999.

[8] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In ANLP/NAACL Workshop on Automatic Summarization, pages 21–29, 2000.

[9] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. A Scalable Summarization System Using Robust NLP. In Proceedings of the Intelligent Scalable Text Summarization Workshop, pages 66–73, 1997.

[10] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop*, pages 10–17, 1997.

[11] C. Nobata and S. Sekine. Results of CRL/NYU System at DUC-2003 and an Experiment on Division of Document Sets. In 2003 Document Understanding Conference Draft Papers, pages 79–85, 2003.

[12] S. Teufel and M. Moens. Sentence extraction as a classification task. In ACL/EACL Workshop on Intelligent and Scalable Text Summarization, 1997.

[13] C.-Y. Lin. Improving Summarization Performance by Sentence Compression - A Pilot Study. In Proceedings of the International Workshop on Information Retrieval with Asian Language, pages 1–8, 2003.

[14] K. Han, Y. Song, and H. Rim. KU Text Summarization System for DUC 2003. In Document Understanding Conference Draft Papers, pages 118–121, 2003.

[15] A. Nenkova, B. Schiffman, A. Schlaiker, S. Blair-Goldensohn, R. Barzilay, S. Sigelman, V. Hatzivassiloglou, and K. McKeown. Columbia at the Document Understanding Conference 2003. In 2003 Document Understanding Conference Draft Papers, pages 71–78, 2003.

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

[17] Fellbaum, Christiane (Ed.), Wordnet : An Electronic Lexical Database (Language, Speech and Communication). MIT Press, 1998.

[18] S. Scott and S. Matwin, Text classification using WordNet hypernyms. In "Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems", Montreal, 1998.

[19] Buenaga M. Rodríguez., J. M. Gómez-Hidalgo, B. Díaz Agudo, Using WordNet to complement training information in text categorization. In "Proceedings of the International Conference on Recent Advances in Natural Language Processing", Tzigov Chark, 1997.

[20] M. Vorhees, Ellen, Using WordNet for text retrieval. In Fellbaum C. (ed.) "WordNet: An Electronic Lexical Database", MIT Press, 1998.

[21] Resnik Philip, Using information content to evaluate semantic similarity in a taxonomy. In "Proceedings of the 14th International Joint Conference on Artificial Intelligence", Montreal, 1995.

[22] Ramakrishnan and Bhattacharya, Text representation with wordnet synsets. Eight International Conference on Applications of Natural Language to Information Systems (NLDB2003), 2003.

[23] Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513-523. 1988.

[24] Chuang, W., and Yang, J. Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 152–159, 2000.

[25] D. R. Radev, E. Hovy, and K. McKeown. Introduction to the special issue on summarization. Computational Linguistics, 28(4):399–408, 2002.