

SUMMARIZING TEXT by RANKING TEXT UNITS ACCORDING to SHALLOW LINGUISTIC FEATURES

Pankaj Gupta, Vijay Shankar Pendluru

Wipro Technologies INDIA

pankaj_gupta96@yahoo.com, vijay_shankarece@yahoo.com

Keywords: Text Summarization, Lexical Chains, Vector Space Model, WordNet, lexicographical relationships, topically related.

Abstract

We present an approach of identifying the most prominent text/sentences using various shallow linguistic features, taking degree of connectiveness among the text units into consideration so as to minimize the poorly linked sentences in the resulting summary. As per the limitations of the current summarizing systems, the summary generated by those systems contains poorly linked sentences and are not topically salient. Thus, the paper aims at highlighting the effect of lexical chain scoring after the nouns and compound nouns are chained by searching for lexical cohesive relationships between words in the text using WordNet and using lexicographical relationships such as synonymy and hyponyms. In this paper, our algorithm ranks sentences based on the sum of the scores of the words in each sentence involving approaches like term frequencies, location of sentence in the text, cue words and phrases, word occurrences, and measuring lexical similarity (measuring chain score, word score and finally sentence score) for ranking the text units. We then identified and extracted high scored sentences and then the Vector Space approach is used to measure the relatedness/similarity between the extracted sentence and the topic words involving again the WordNet lexical database relationships to prioritise the topically related sentences. A threshold angle between the two vectors is predefined experimentally to which the ranked/scored sentences to be dropped and which the significant sentences with ranking/scores higher than threshold to be extracted. Note that the value of threshold is predetermined based on the percentage of output summary required to be generated.

Our results show that the agreement among the human subjects and our algorithm is highly significant. We find that the top ranked sentences are most of the time the most important ones that human subjects extract in our experiment. The deviation in percentage for precision and recall as obtained from human generated summary and our system generated summary is 12% and 8 % respectively. Our research work presents the concept of topic word similarity with the high ranking sentences

obtained after applying various heuristics, lexical chaining and the Vector Space approaches. Our approach for summarizing the text determines more significant and topically related sentences in the text and thus outputs a higher summary quality.

I. INTRODUCTION

Today in this Digital World, to avoid drowning in the abundance of information, it must be filtered and extracted. Automatic summarization one such reductive technique allowing the Computer to summarise the longer text to shorter non-redundant form. There are many languages for which large bodies of data aimed at language technology research to a high degree are lacking. Since it is time consuming and expensive, also there may not be resources to develop such bodies of data. . Thus, there is a need for automatic text summarization for the languages in order to subdue this constantly increasing amount of electronically produced text.

In this paper, various shallow linguistic techniques are mentioned that rank the sentences in the text. We focus on lexical cohesion which is the textual property responsible for making the sentences of a text seem to hang together indicated by the use of semantically related vocabulary. Cohesion is thus a surface indicator of the discourse structure of a document. One method of representing this type of discourse structure is through the use of a linguistic technique called lexical chaining. Lexical chains are defined as clusters of semantically related words.

For example: {book, record, sheets, Catching Fire, paper, script} is a chain, where book, script and sheets are synonyms, paper is part of a book and “Catching Fire” is a specialization of book. The lexical chaining algorithm discussed in this paper identify such lexical cohesive relationships between words using the WordNet relationship.

II. IMPLEMENTATION

The following phases of the summarization process can be identified:

A. Document Cleaning

First, we convert documents into internal format (XML). We apply *stemming*, as occurrence of various word forms are not utilized by the summarization algorithm. *Stop words* are then removed from the text. Further processing of document collection is thus simplified.

B. Segmentation into Passages

Documents are segmented into sentences and paragraphs (use of heuristic methods).

C. Indexing

Words are replaced by numeric values to speed up further processing.

D. Sentence Scoring

Each sentence is assigned a numerical score depending on the summarization method being used. Sentences with the highest score are selected.

E. Synthesis (Abstract Generation)

Document abstract is composed of sentences with the highest score. In this phase we can remove unimportant parts of sentences. It is also possible to enhance abstract by using sentences bearing relatively low information value, but improving readability of the final abstract.

III. SHALLOW LINGUISTIC FEATURES

(HEURISTICS):

There are several approaches to Text Summarization, such as those proposed by (Hovy and Marcu 1998; Mani and Maybury 1999; Tucker 1999; Radev 2000; Maybury and Mani 2001; Mani 2001; Alonso and Castellon 2001). In this paper, we use (Mani and Maybury 1999)'s classification to describe the level at which several summarization systems process texts. Based on this classification, automatic summarizers can be characterized as approaching the problem at the surface, entity, or discourse level. The simplest type of summarizers is the ones classified at surface level. Surface-level approaches represent information in terms of shallow linguistic features that are then selectively combined together to yield a salience function used to extract information. This kind of shallow features were first used by (Luhn 1958) and

(Edmundson 1969). The first one calculated sentence weights using the frequency of terms in the text and the second one taking into account sentence position in text, words occurring in the title and heading and a list of cue words.

Shallow linguistic features used in surface-level approaches can be classified as follows:

A. Term Frequency statistics provides clues that important sentences are the ones that contain words that occur frequently. The salience of a sentence increases for each frequent word it contains, according to the frequencies of words in the text. The assumption is that frequent content words indicate the discussed topic, usually, in this process word stemming is used and stop words are filtered out. As said before, (Luhn 1958) proposes to extract sentences which have the highest salience value, obtaining summaries with a reasonable quality. Various frequency measures are also used by (Edmundson 1969; Kupiec et al. 1995; Teufel and Moens 1999; Hovy and Lin 1999). But the combination of word frequency with other measures does not always produce an improvement, as shown by (Kupiec et al. 1995) and (Teufel and Moens 1999). (Witbrock and Mittal 1999) obtain a statistical model to determine the likelihood that each individual word in the text has to appear in the summary.

B. Location provides clues that important sentences are located at pre-defined positions in the text. Instances of that assumption are the boost method and the head method. The boost method consists of taking the first sentences to create an extract based summary. The head method says that words in titles and headings are more relevant to summarization. Some variants of the location method are used in (Baxendale 1958; Edmundson 1969; Donlan 1980; Kupiec et al. 1995; Teufel and Moens 1997). A generalization of location methods is the Optimum Position Policy (OPP) used by (Lin and Hovy 1997) in their SUMMARIST system, where they exploit Machine Learning (ML) techniques to identify the positions where relevant information is placed within different textual genres and domains.

C. Bias reflects the fact that the importance of meaning units (sentences, clauses, etc.) is determined by the terms from the title or headings, initial part of text, or user's query. On the one hand, (Kupiec et al. 1995; Teufel and Moens 1997; Hovy and Lin 1999) use those words contained in the text's title or headings. On the other hand, (Buckley and Cardie 1997; Strzalkowski et al. 1999; Hovy and Lin 1999;

Conroy and O'Leary 2001; Schlesinger et al. 2002) use the words in user's query to produce query-based summaries.

D. Cue Words and Phrases are typically meta-linguistic markers (e.g., cues: "in summary", "in conclusion", "our investigation", "the paper describes"; or emphasizes: "significantly", "important", "in particular", "hardly", "impossible"), as well as domain-specific bonus phrases and stigma terms. (Kupiec et al. 1995) or (Teufel and Moens 1997) use a manually built list of cue phrases. (Aone et al. 1997) detect such a list by gathering knowledge automatically from a corpus. The method started in (Teufel and Moens 1997) is expanded in (Teufel and Moens 1999), exploiting cue phrases that contribute in the ranking of the sentences bearing these cue phrases.

E. Word Co-occurrence

Words can be related if they occur in common contexts. Some example applications are presented in (Baldwin and Morton 1998; McKeown et al. 1999). (Salton et al. 1997; Mitra et al. 1997) apply IR methods at the document level, treating paragraphs in texts as documents are treated in a collection of documents. Using IR-based method, a word similarity measure is used to determine the set S_i of paragraphs that each paragraph P_i is related to. After determining relatedness scores S_i for each paragraph, the largest S_i paragraphs scores are extracted.

F. Lexical Similarity

Words can be related by WordNet relationships (synonymy, hyponymy, metonymy relations). (Barzilay 1997) details a system where Lexical Chains are used, based on (Morris and Hirst 1991). The assumption is that important sentences are those that are identified by strong chains.

1. Measuring lexical similarity in the text:

The basic chaining algorithm follows the following steps. First, we select a set of candidate words, generally nouns. Then the list of chains is searched and if a word satisfies the relatedness criteria (mentioned latter) with a chain word then the word is added to the chain, otherwise a new chain is created. The algorithms use the WordNet lexical database for determining relatedness of the words (Miller et al. 1990). Word relations/associability in the WordNet database are represented by synonym sets having the words sharing a common sense. For example two senses of "table" are represented as: desk, wooden product, furniture, table (i.e., a physical material) and "array, matrix, content, numeric table". WordNet

contains more than 118,000 different word forms. Words of the same category are linked through semantic relations like synonymy and hyponymy.

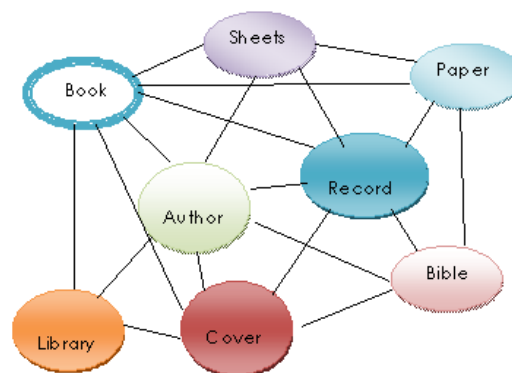


Figure 1. Visual Representation of Lexical Chains

The relatedness criteria for the words are the relationships outlined by St. Onge. St. Onge used WordNet as the knowledge source for lexical chaining. He devised three different relationships between candidate words: extra-strong, strong and medium-strong. Extra-strong relations are represented by lexical repetitions of a word and strong relations are by synonyms or near-synonyms. Medium strength relations follow sets of rules laid out by St. Onge. These rules direct the paths that are allowable in the WordNet structure while navigating for the lexical relations.

Most lexical chain based summarizers follow the same approach by generating lexical chains and then the strongest of these chains are used to weight and extract key sentences in the text for summary generation, Barzilay and Elhadad. For Chain Scores, they calculated the product of two chain characteristics: the length of the chain, which is the total number of words in the chain plus repetitions and, the homogeneity of the chain, which is equal to 1 minus the number of distinct words divided by the length of the chain. Chain score that exceeds an average chain score plus twice the standard deviation are considered 'strong' chains and the preferred candidates for the text extraction. Barzilay et al. then selects the first sentence that contains a 'representative' word from a 'strong' chain determined, where a 'representative' word has a frequency/repetition more than or equal to the average frequency of words in that chain determined.

Our chaining algorithm uses a modified lexical chaining topic-driven approach. The first step in our

chain formation is to perform parts-of-speech(POS) tagging to an incoming document. We have used the General Architecture for Text Engineering (GATE) tool for POS tagging and searching the patterns of tags corresponding to these types of phrases e.g. presidential/JJ campaign/NN, or U.S/NN President/NN Bush/NP where /NN is a noun tag and /NP is a proper noun tag.

The nouns and compound nouns are chained by looking for lexical cohesive relationships between words in the text using WordNet using lexicographical relationships such as synonymy (*book, stationery*), specialization/generalization (*calculator, machine*), part-whole/whole-part (*politicians, government*). Unlike Barzilay et al.'s approach, the algorithm by William Doran, Nicola Stokes, Joe Carthy, and John Dunnion calculates chain scores based on the word frequencies and the type of WordNet relations between chain members. More specifically, as shown in equation below, the chain score is the sum of each score assigned to each word pair in the chain. The score is calculated as the sum total of the frequencies of the two words, multiplied by the relationship score between them as,

$$chain_score(chain) = \sum (reps_i + reps_j) * rel(i, j)$$

where $reps_i$ is the frequency of word i th in the text, and $rel(i, j)$ is a relationship score assigned based on the strength of the relationship between word i th and j th, where a synonym relationship gets assigned a value of 0.9, specialization / generalization and part-whole/whole-part 0.7.

The next step is to rank sentences based on the sum of the scores of the words in each sentence in the text, where a word's score is a scaled version of its chain's score. The scaling factor is calculated as the minimum distance between a word and its successor or its predecessor in the chain.

$$word_score(word_i) = \alpha \times chain_score(chain(word_i))$$

$$\alpha = 1 - (\min[dist(w_{i-1}, w_i), dist(w_i, w_{i+1})] / dist(w_1, w_n))$$

where $dist(w_i, w_j)$ is the distance i.e. the number of words that separate two words in the text while the $chain(word_i)$ is the chain $word_i$ belongs to. The sentence score is the sum of these word scores normalized with respect to the length of the sentence and the number of chain words it contains.

VI. USING TOPIC DRIVEN APPROACHES for RANKING LEXICALLY SALIENT TEXT UNITS

Topic-focused summarization tasks are related with a specific information need expressed by the user. This need is usually expressed by a natural language query or, as in World Wide Web search engines, by a list of keywords. This kind of tasks are also known as user-focused or query-driven summarization, in contrast to generic or unbiased summarization, which is text driven.

Here we approach to identify the similarity between the topic word and each salient sentence that we have extracted using the modified lexical chaining, by using common procedures such as tree similarity (Schilder and McInnes 2006), QA systems (Fuentes et al. 2005; S. Blair-Goldensohn 2006; Lacatusu et al. 2006; Molla and S. 2006) or the vector space model, syntactic and semantic relations (Lacatusu et al. 2006) or LSA and other kinds of semantic resources (Miller 2003; Jagadeesh et al. 2005; Hachey et al. 2005).

Implementing the basic formula for similarity between Topic words and the salient extracted text is given by:

The Vector Space Model

The Vector Space model is a document similarity model used in Information Retrieval (Salton 1971). In this model the document signatures are represented as feature vectors consisting of the words that occur within the documents, with weights attached to each word denoting its importance for the document. For example, for each term/word we can record the frequency in each document.

To measure the similarity between two documents, we can represent these with the following two vectors:

$$\bar{D}_i = (w_{i,1}, w_{i,2}, w_{i,3}, \dots, w_{i,n})$$

$$\bar{D}_j = (w_{j,1}, w_{j,2}, w_{j,3}, \dots, w_{j,m})$$

Here ; $\bar{D}_i$ (topic words) and $\bar{D}_j$ (salient text) denote the two documents, with n and m being the total number of index terms occurring in these documents. We can now compute the similarity by calculating the cosine angle between the two:

$$\text{Cos}(\bar{D}_i, \bar{D}_j) = \frac{\sum_{k=1}^n (w_{k,i} \times w_{k,j})}{\sqrt{\sum_{k=1}^n (w_{k,i})^2} \times \sqrt{\sum_{k=1}^m (w_{k,j})^2}}$$

Here n is the total number of terms recognized by the matching system, while $w_{k,j}$ and $w_{k,j}$ represent the importance of the index

term k to $\bar{D}_i$ and $\bar{D}_j$, respectively.

V. HEURISTICS for EXTRACTING

THE SIGNIFICANT SENTENCES:

Most of the times, the same topic is discussed in a number of places in the text document, so its chain is distributed across the whole text. Still, in some text unit, this global topic is the central topic (focus) of the segment. We try to identify extract high scored sentences and then the above mentioned Vector Space approach is used to measure the relatedness/similarity between the extracted sentence and the topic words involving again the WordNet lexical database relationships. A threshold angle between the two vectors is predefined experimentally below which the ranked/scored sentences are dropped and the significant sentences with ranking/scores higher than threshold are extracted. Note that the value of threshold is predetermined based on the percentage of output summary required to be generated.

VI. EVALUATION & RESULTS

For evaluation of our summarization algorithm, we make use of intrinsic method (Edmunson 1969; Paice 1990; Kupiec, Pedersen, & Chen 1995; Marcu 1997; Salton et al. 1997; Ono, Sumita, & Miike 1994). Our approach is to generate an 'ideal' summary, written by multiple human subjects. The output of our summarizing system is then compared with the 'ideal' summary. Precision (the fraction of the text portions retrieved that are relevant to user's information need) and recall (the fraction of the text portions that are relevant to the topic that are successfully retrieved) are the two performance measure used for determining the quality of the summary. We also compared the summaries with those obtained by the Microsoft Summarizer available in Word97 as the same used by Brazilay & Elhada97 for measuring the quality of summaries generated.

For studying agreement among the human subjects, 20 documents were selected; for each document; 40 summaries were constructed by 20 human subjects. Each subject constructed 2 summaries of the document: one for 20% and another for 30%. The percent of length is calculated in terms of number of sentences.

In addition to these 40 human constructed summaries, we built 20 % and 30% summaries using our system and Microsoft summarizer. Documents selected are news articles on computer, terrorism, pollution, health, Technology, Business and Arts. The average length of articles is 50 sentences. The human subjects chosen for manual summarization are Graduate students in Department of Information Technology at Amity University, UP INDIA.

	Microsoft	
	Prec	Recall
20%	32	34
30%	37	39

	Our Algorithm	
	Prec	Recall
20%	81	86
30%	83	89

	Human Subjects	
	Prec	Recall
20%	91	93
30%	94	96

TABLE 1. EVALUATION RESULTS

Our results show that the agreement among the human subjects and our algorithm is highly significant. We find that the top ranked sentences are most of the time the most important ones that human subjects extract. Furthermore, extraction of sentences that contain word sense ambiguity is still problematic in our system.

VII. CURRENT RESEARCH and

FUTURE DIRECTIONS:

One of the goals of this research is to eventually create a system which generates natural language summary. Currently, our approach uses sentence selection as its methods of generation. It seems obvious that sentence selection will not create fluent, coherent text. Here, the limitation to our work is the completeness. Because information extraction is only at the sentence boundary, information which may be very important may be left out if a highly compressed summary is required.

Our current research is examining methods of using all of the important sentences determined by our algorithm. Our goal is to combine and condense all the significant information pertaining to a given concept which can then be used in generation of coherent text. The use of compression techniques will increase the condensation of the summary and improve its quality (Barzilay, McKweon, and Elhadad, 1999; Mani, Gates, and Bloedorn, 1999;

Jing and McKeown, 2000; Knight and Marcu, 2000). We intend to include additional criteria and plan to exploit text coherence patterns for summarization (cf. (Hahn 1990; Hahn & Strube 1997) for texttheoretical models and related proposals to build on text coherence structure by (Alterman 1986) and (Marcu 1997). Our future work eyes at a method for producing query-based summary using these algorithms. The future extension will concentrate on web text retrieval according to a given interest profile. The text will be retrieved for the common search engine and then the text will be analyzed to find out if they are really relevant according to user's interested profile in order to provide concise summary of the retrieved subset of the relevant documents using the multidocument summarization facilities.

CONCLUSIONS

Document summarization methods have been in notice for a long time and new approaches are coming up very sporadically. Current research is therefore focused namely on modifications of the existing approaches, or their combination. The new trends in Summarization approaches include the effort to find an optimum setup of various parameters, but also the use of various thesauruses, or other dictionary-based methods (e.g. dictionaries of geographical locations, well-known personalities, etc.). When summarizing longer documents, these can be split into several sections (clusters), summarizing each section independently. The summarizer can be also improved by utilizing linguistic information contained in text documents. In this paper, we have presented an efficient implementation of the lexical cohesion approach as the driving engine of the summarization system. The ranking procedure is used to select the most salient and best connected sentences in a text corresponding to the summary ratio requested by the user.

Score (sentence_i) = Term_frequency_score + Cue_words_phrases_score + Location_score + Bias_score + word_cooccurrence_score + Lexical_chaining_with_topic term similarity

Readability and brevity of summaries are critical for their usefulness to the user, along with in formativeness and the ability to capture the right aspects of the content. If the main purpose of having summaries is to save the user some time, then their comprehensiveness plays a secondary role - the user can always refer to the original text for details.

REFERENCES

- Adamson, G. and Boreham, J. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Inf. Stor. Retrieval* 10, 253-260.
- Baccchin, M., Ferro, N., and Melucci, M. 2005. A probabilistic model for stemmer generation. *Inf. Process. Manage.* 41, 1, 121-137.
- Barzilay, Regina, Kathleen McKweon, and Micheal Elhadad. (1999). Information fusion in the context of multidocument summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 550-557, College Park, Maryland.
- Barzilay, Regina, Noémie Elhadad, and Kathy McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.
- Brill, E. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the Third Conference on Applied Computational Linguistics*. Trento, Italy: Association for Computational Linguistics.
- Brunn M., Chali Y., and Pinchak C. 2001, Text summarisation using lexical chains, In *Workshop on Text Summarisation in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana, 2001.
- Godsmith, J. A., Higgings, D., and Soglasnova, S. 2000. Automatic language-specific stemming in information retrieval. In *Proceedings of the Workshop on Cross-Language Evaluation Forum (CLEF)*, 273-284.
- Hirst, G., and St-Onge, D. 1998 [to appear]. Lexical chains as representation of context for the detection and correction of malapropisms. In Fellbaum, C., ed.
- McKeown, K., and Radev, D. 1995. Generating summaries of multiple news articles. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74-82. Seattle, Washington: Special Interest Group on Information Retrieval.
- Miller G. A., Beckwith R., Fellbaum C., Gross, D., and Miller, K. 1990, *Five papers on WordNet*. Technical Report, Cognitive Science Laboratory, 1990.
- Morris, Jane and Graeme Hirst. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-48.
- WordNet: An Electronic Lexical Database and Some of its Applications. Cambridge, MA: The MIT Press. Hoey, M. 1991.