Proceedings of the IASTED International Conference
ARTIFICIAL INTELLIGENCE AND APPLICATIONS
February 16-18, 2004, Innsbruck, Austria

A TURKISH AUTOMATIC TEXT SUMMARIZATION SYSTEM

Zeynep Altan
İstanbul University, Faculty of Engineering, Department of Computer Engineering
34320 - Avcılar, İstanbul, Turkey. *zaltan@istanbul.edu.tr*

ABSTRACT

The system developed in this study uses a Turkish text as input, and after the implementation of a sequence of procedures the summary results accomplishing the target sentence length. The study has been specialized to obtain more significant results for the articles on economic matters. We converted the context of all papers into HTML documents to provide the formal structures. The training system comprehends a corpus including 50 different articles. Moreover, it is possible to add the summarized document to the corpus in case of demand. The choice of the summarization percentage has been made by the user. We have utilized the Internet both for the development environment of the system and its running platform. The program interface runs on the Web through browsers.

This study principally depends on the statistical analysis of the paragraphs, sentences and words in the document according to predefined specific weighting factors. Although these weight points determine the skeleton of summary, the sentences for the summarization are primarily chosen to emphasize the semantic integrity.

KEY WORDS

Natural Language Processing, Text Summarization, Corpus, Frequency

1. Introduction

The number of online articles has excessively increased in the past decade, and the reorganization of information was begun to consider important as a result of this overflow. However, growing demand in number could'nt increase the summarized articles even in English because of the difficulty to get them with human effort. Therefore, the constitution of an automatic summarization system has become one of the major topics in applied natural languages processing, and some of these technologies have rapidly been implemented into the real word problems.

As a result, we can outline the text summarization that produces maximum information in minimum capacity. Since online services on WWW environment has caused the information explosion, the importance of well-arranged information increased.

1.1 The Classification of Techniques

In general there are many categories of automatic summarization systems. We can divide these studies into three fundamental classes as classical approaches, corpus-based approaches and knowledge rich approaches. Classical approaches describe surface-level researches constituting the fundamental studies on summarization. For example; Luhn's article developed in 1958 was basically based on term frequency, Edmundson in 1969 compared the use of term frequency with some other features, and *Chemical Abstract Service* developed by Pollock and his colleague Zamara in 1973 depended on the use of chemical cue-phrases [1].

The second approximation to summarization systems describes different corpus-based methods. The study describing the use of Bayesian classifier to extract sentences is generally called KPC approach [2]; it constitutes a privileged point of view to the corpus based statistics. Cue words, sentence location, title words have an important influence on corpus statistics. Morphological, semantic properties (synonym words), proper name variants, and other linguistic properties strengthen the summarization performance. Thesaurus-based algorithms can also assist the concept in the text.

While the first two classifications focus on surface-level approaches, the third approach is aimed to different applications. Most researches in that category convert information from multiple facts into a single sentence, depending on the various linguistical constrains. *Streak* as the general summarization of sports events, and *Plandoc* as the automated documentation of telephone planning activities are the two examples of those approaches, [3]. *Streak* uses a set of scores for any basketball game as input, and declares a short summary. *Plandoc* using discourse planning produces a report which includes the summaries of telephone network design activity.

Information extraction as a different approximation to summarization systems is such a process that collects information about the specified types of various entities, relationships or events. This discipline has arisen from the influence of *MUC Conferences* (Message Understanding Conferences) [4]. The goal of those conferences was comparatively the evaluation of participants' systems on different extraction tasks. Summarist as a text summarizing system [5] has been divided into three different classes as topic identification, interpretation and generation phases, and was consisted of several independent modules training on large corpora. The researchers utilized information retrieval techniques augmenting them with symbolic/semantic and statistical methods. Another specific project related to that categorization has referred to medical articles [6]. The system input has been preprocessed for the patient records, a set of medical journal articles and for the user query. This system has integrated different techniques studied previously. Journal article results have matched with the patient's record. By merging and ordering the extracted information, summary content was processed.

Multiple documents covering similar information examined according to their similarities and differences as another approach to summarization [7]. Phrase extraction as summary descriptors and synonym/hyponym extraction using *WordNet* constituted the framework of the system. *Mead* [8] *as* an alternative multi-document summarizer generates summaries using cluster centroids produced by topic detection and tracking system. Sentence utility and subsumption have also been applied to the evaluation of multiple document summaries.

2. System Architecture

Automatic summarization system developed in this study consists of five different modules. The architecture of that system has been explained in Figure 1 in detail. In the first module, the document has structurally been studied. The headlines, paragraphs and sentences were separated utilizing HTML tags. Then, the statistical results have been obtained of which data was collected from the documents. Third phase analyzed the key statements which were statistically learned by the system. In the fourth module, the text has been examined according to data received form the corpus. Finally, summary has been constituted by designating the related sentences. Summarization process is composed of 13 different steps comprehending five fundamental steps. Each phase and the output of each step is an input for the next phase,. Each module generally works as follows:

2 .1 Structural Analysis Module

This module comprehends the heading specification, paragraph description and word requirement processes. The document annotated with HTML tags is firstly separated according to titles of the context establishing the caption words. Since caption statements usually explain the subject of the text, it is convenient to search these words as the important indicators of content, and to emphasize the sentences including these words. Nevertheless stop words have to be removed. Since the solution area has been specialized according to the documents on economic subjects, we used heading word factor efficiently in this study. When resolution environment is not the studied special topic, heading word factor requires to be rearranged.

Moreover; the studied document is separated into the paragraphs, paragraphs are divided to the sentences, and sentences are split into words. During this procedure paragraph numbers, total of sentences and word accumulation are hold inside the program variables inclusive of the position of the sentences, location of the words.

2.2 Statistical Analysis Module

Basic operation of this module is to find the frequency of the words. The results are then saved in the database as temporarily or permanently according to user preference. After the determination of word frequencies, the positions of these numerals are obtained as the sentence situations. Most papers in Turkish as in other languages consist of introduction, development and conclusion sections to underline any idea, an interpretation or a research. In the introduction section, the writer makes the subject explicit with a few sentences, and tries to express his opinion in the article to the reader. The writer transfers his/her ideas in detail to the reader in the development section. The thought is gathered together in the conclusion section, and outcome is submitted to the reader. Therefore, introduction and conclusion divisions have more importance to comprehend the global subject of the text. The sentences in the introduction and conclusion paragraphs become important factors for the summary, and the weights of such sentences have to be increased with special evaluations.

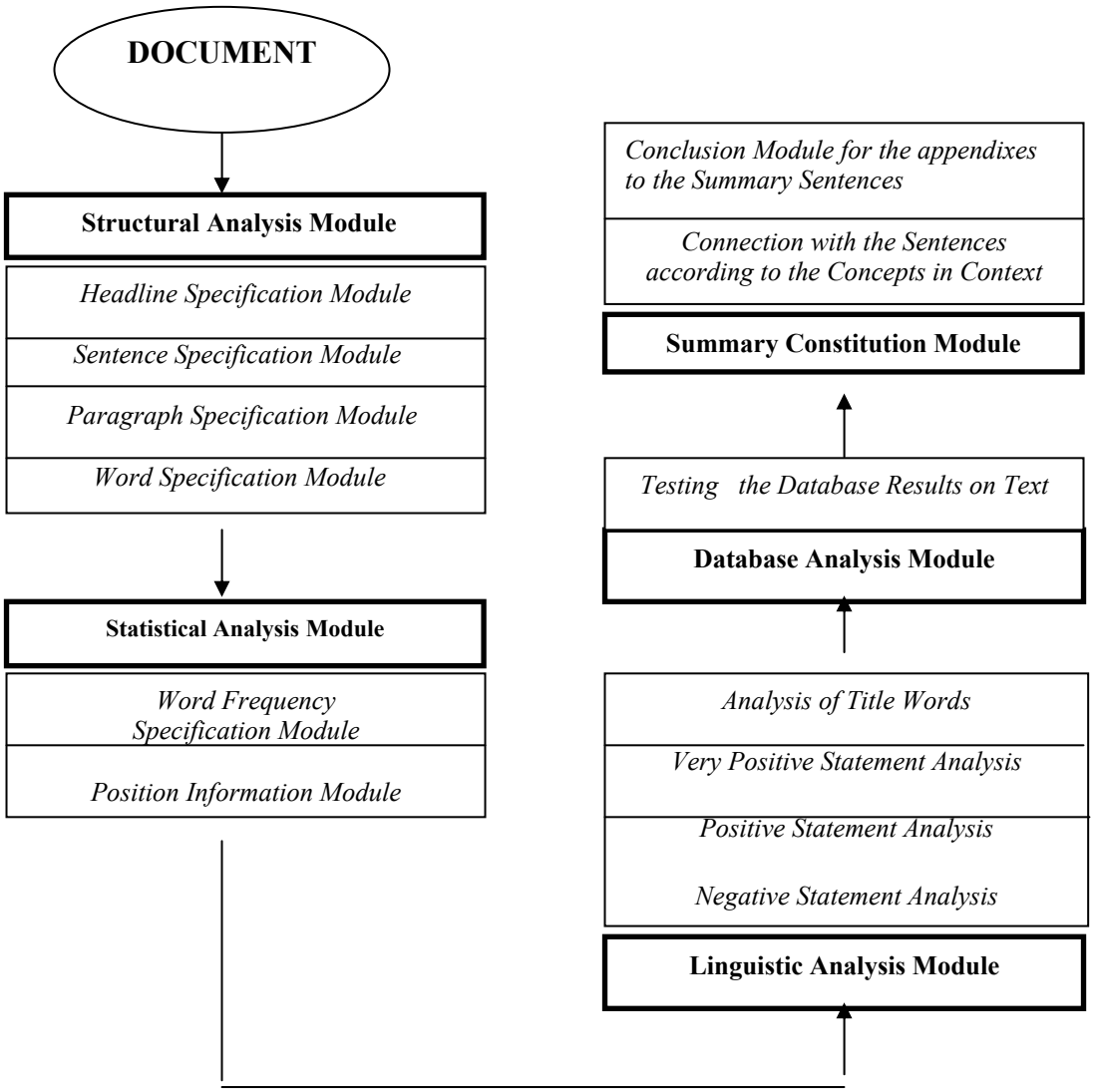

Figure 1: *The Architecture of Turkish Automatic Summarization System*

2.3 Linguistic Analysis Module

This module examines the predefined statements which would considerably affect the summary as affirmative or negative factors. These declarations are such words that describe the studied subject, and designate the meaning given from the sentence. The definition of those statements for each sentence in this stage contributes to the selection of summary sentences. For instance; any sentence beginning with the Turkish word "özetle" (as summary) will seriously be guaranteed to be included in the summary.

2.4 Database Analysis Module

We used 50 different papers about economics as corpus. If the analyzing text is intentionally added to the corpus, the statistical results of current text would be included in the conclusion. After completing this analysis step, a number of words with high frequency are searched in the document. This module has affirmative contribution both to the sentences including the words that are also in the corpus and to the summary.

2.5 Summary Constitution Phase

This is the final step which is decided to the summary sentences among the examined sentences. While determining the abstract sentences, the relationship between the sentences is established according to the concepts they comprehend. The sentences which will be appended to the summary are preferred between semantically related ones. The user can arbitrarily constitute one or two summary scenarios. The reflection of the general contextual values of the paragraphs to the sentences forms different scenarios.

3. Extracting Text for a Summary

Each summary has to be instantiated by one or more sections extracted from the original text. These can be paragraphs, paragraph parts, sentences, or even sentence fragments [9].

1. The specification of title as words and phrases is an important factor for the summarization, so one or more titles in the document have previously been specified. Title specification module finds these tags in HTML form. After all titles are labeled, they are saved to use in linguistic analysis phase.

2. Sentence specification module separates the sentences in the document. Firstly, punctuation marks are searched to obtain the estimated sentence number. Then different abbreviations used in Turkish are controlled to approve the sentence number. In such situations the researcher must be careful to decide whether the related punctuation mark ends the sentence or not. The abbreviations must separately be marked in the system.

3. Paragraph specification module utilizes tags to separate the paragraphs. They are numbered increasingly. During the execution of this module, user can learn the location of sentences according to the paragraph position.

4. The words in the document are obtained by examining previously reserved sentences. All sentences are tested in a loop. The words are saved in an array. Word specification module does not control the string. Word frequency module calculates the frequency of the specified words. Stop words in Turkish like "gibi", "ve", "ki", "mi" are not accepted as words, and they are not saved as string by the word specification module. After completing the control of words, they are compared to specify the same words. This comparison occurs as the character analogy of two words. If the characters of the words match or one word overlaps the other one, they are accepted as the same words. We also tested to use our morphological analyzer developed in Prolog to extract the root words, but the result was not so successful. Because the necessity of reading at least 250-300 words from the previously built database for each document caused the parsing procedure to continue too long. Therefore the frequency of all words is hold in a table. After they are ordered, the words with high frequency are returned as one third of all words in the document. Since the upper limit for the word selection has been defined as %30 in many summarization systems, we preferred one third ratio for the word options from the document. Such words have important contribution to the sentences.

5. Negative statements, positive statements, very positive statements are classified. The analysis of these statements decreases or increases the contribution of related sentences to the summary. Title words have also positive contribution to the system.

6. Sentence relationship module is constituted according to the context. After completing the sentence weighting process, summary is normally constituted from the most weighted sentences. If the weighting factors are equal, one of the sentences must be chosen. In this case, we applied two different analyses to the sentences and finally the sentence of which its weight has too much increased is taken to constitute the summary.

4. The Summarization Algorithm

We present a 19-steps algorithm for generating summaries of news paper articles about economic topics in the following:

S1: Specify negative words list. "Çünkü" (because), "öyleyse" (in that case), "ancak" (only)…….

S2: Define very positive words list. "Özetle", "sonuçta" (result), "neticede" (in the end)……

S3: Define positive words list. "Piyasa" (market), "enflasyon" (inflation), "döviz" (foreign exchange)…….

S4: Describe abbreviation list. "Dr.", "Prof."……

S5: Remove unused HTML tags from the text.

S6: Find the titles in the document.

S7: Divide titles into words.

S8: Extract the titles from the text.

S9: Describe the sentences in the text, and count the sentence number.

S10: Separate the paragraps describe the position of sentences in terms of paragraphs.

S11: Separate all the words in the sentences.

S12: Count the word frequencies.

S13. Increase the weights of the sentences one addition (contribution) unit in the introduction and conclusion paragraphs

S14: Search negative statements in the sentences, and make equal to zero the weight of these sentences.

S15: Search the positive statements, and very positive statements in the sentences, and increase the weight of these sentences as different addition units.

S16: Search the title words in the sentences, and increase the weight of these sentences one (or another) addition unit(s).

S17: Return the words from database with maximum frequency as much as the document number, then search these words in the sentences, and increase the weight of these sentences one addition unit.

S18: If it is required to prefer the summary sentences among the equally weighted sentences, use different scenarios.
S19: Return maximum weighted sentences as summary with the user defined summary ratio.

5. Implementation and Evaluation

The summarizer has been implemented by a server-side scripting language PHP/MySQL as a demonstration system. PHP and MySQL are one of the best combinations for creating data-driven sites. PHP as a server-side scripting language is processed by the Web server. After the server plays with the PHP code, it returns plain old HTML back to the browser. MySQL is a small, compact database server ideal for small applications. In addition to supporting standard SQL (ANSI), it compiles on a number of platforms and has multithreading abilities on servers. MySQL can also be run on various Windows machines.

Figure 2 shows the interface to the interactive version of the summarizer. The interface allows the user to set a few critical parameters of the summarization process:

1. ***Choosing the summarized document*** *(Özetlenecek dosya* in Figure2*):* The user can either select the text as an HTML document or attach the document from the directory. He can also reach the document from any URL address (*Özetlenecek address* in Figure 2).
2. ***Selection of desired length of the summary*** (*Özetleme oranı* in Figure 2): The user can specify the length of the sentences as a percentage of the original document.
3. ***Is the summary added the database?****(Veri tabanına ekleyin* in Figure 2). According to the request of the user, the consequence can be added to the database, or can be obtained different solutions (*Farklı sonuçlar oluşturun* in Figure 2).

Summary generating screen in *Figure 3* demonstrates the performance of the conclusion according to information received from the user in the first screen. As we can see from the figure, the document consists of 32 sentences. %10 summarizing ratio constitutes the summary from four sentences (sentence zero, and 23rd, 27th, 29th sentences are defined as the summary sentences). The table above the summary consequence shows the words with high frequency in the document.

6. Discussion and Conclusion

The purpose of the developed system in this study is to constitute such a summary that reflects the subject perfectly, underlying the contextual and locational relations of the sentences very carefully. The system generally utilizes statistical approximations. Key words which were previously defined by the system, most frequent words in the text, and the constituted corpus have important influence on summarization process. Finally, summary has been implemented according to summarization ratio given by the user.

In respect to the experiments and corpus, we can explain that the most frequent words do not always reflect the subject of the document correctly. Although some words with high frequency like “çok” (very), “gün” (daytime), “bir” (certain) do not have much contribution to the summary in Turkish, they have positive impression on the result. These words have occasionally negative effects to the performance of the system, but may increase the reliability of the system. In spite of the corpus has been composed from the articles on economic topics, the documents about different topics can satisfactorily be summarized. In this connection, we have to emphasize the word frequency approximation.

It has been observed that the titles reflect the subject appropriately. The key words have an addition to the examined text as in many other summarization studies. While processing the weighting factors of sentences, the greatest addition values are established with respect to the analysis of key words. On the contrary of key words, the sentence which has definitely been excluded from the summary is designated by negative statements. The summary sentences must reflect the subject thoroughly; so the relation of the sentences with each other has carefully been analyzed. In other words, semantic integrity of the selected sentences is very important.

Since the existence of various automated text summarization systems developed so far, we can obtain more than one correct and sufficient summaries for any document. Because of the diversity of the results, it could not be created one best method or solution. As the summary can be established according to the importance of the sentences, we can also use methods changing the structure of the selected sentences. But, it is required a comprehensive database to apply the second type methods. Moreover, the database must distinguish the various word types (adjective, noun etc.) and include the relationships between different types. In this case, it is possible to learn the linguistic properties of the sentence, and to make changes if it is required. For example, WordNet [10] is one of the widespread electronic dictionaries including very different classifications. The richness of Turkish language makes difficult to complete an electronic dictionary such as WordNet. The lack of such a tool in Turkish affects negatively not only the

summarization area but other important research disciplines in computational linguistics.

Moreover, ontology construction is such a difficult and complex process that our initialization will barely be a small prototype of the subject. The problem of getting knowledge based applications both as data and software system deprived us the building of this concept. After the first experiment results were obtained, we decided to construct a draft on economy describing the fundamental concepts, their attributes and relationships among them. Since no project including ontology exists in Turkish, it is required to investigate thoroughly previous ontologies in English. Ontolingua [11] can be represented as a core system on ontological development domain. Our conceptual approximation will be defined similar to frame ontology which is the basic structure of Ontolinqua.

References

[1] Mani I., and Maybury M.T. (eds.) *Advances in Automatic Text Summarization,* Section 1 (The MIT Press, 1999).

[2] Kupiec J., Pedersen J.O. and Chen F., A Trainable Document Summarizer, *Research and Development in Information Retrieval*, 1995, 68-73.

[3] McKeown K., Robin J. & Kukich K., Generating Concise Natural Language Summaries. *Information Processing and Management*, 31(5), 1995, 43-48 .

[4] Grishman R., Information Extraction and Speech Recognition", *In DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, 1998.

[5] Hovy E. and Lin C., Automated text summarization in SUMMARIST",In Mani I and Maybury M.T. (eds) *Advance in Automatic Text Summarization*, (The MIT Press, 1999, 81-97).

[6] Elhadad N., McKeowwn R., Towards Generation Patient Specific Summaries of Medical Articles, *Workshop on Automatic Summarization*, Carnegie Mellon University, 2001.

[7] Mani I. and Bloedorn E., Summarizing similarities and differences among related documents, *Information Retrieval*, 1999, 35-67.

[8] Radev D., Jing H. & Budzikowska M., Centroid-based summarization of multiple documents: sentence extraction and user studies, ANLP/*NAACL 2000 Workshop*, 2002, 21-29.

[9] Strzalkowski T., Stein G., Wang J. and Wise B., A Robust Practical Text Summarizer, In Mani I. and Maybury M.T.(eds) *Advance in Automatic Text Summarization*, (The MIT Press, 1999 , 137-154) .

[10] Fellbaum C., *WordNet: An Electronic Lexical Database,* (The MIT Press, 1999).

[11] Fikes, R., Farquhar, A. and Rice, J. Tools for Assembling Modular Ontologies in Ontolingua, *Technical Report KSL-97-03*, Stanford Un., 1997.

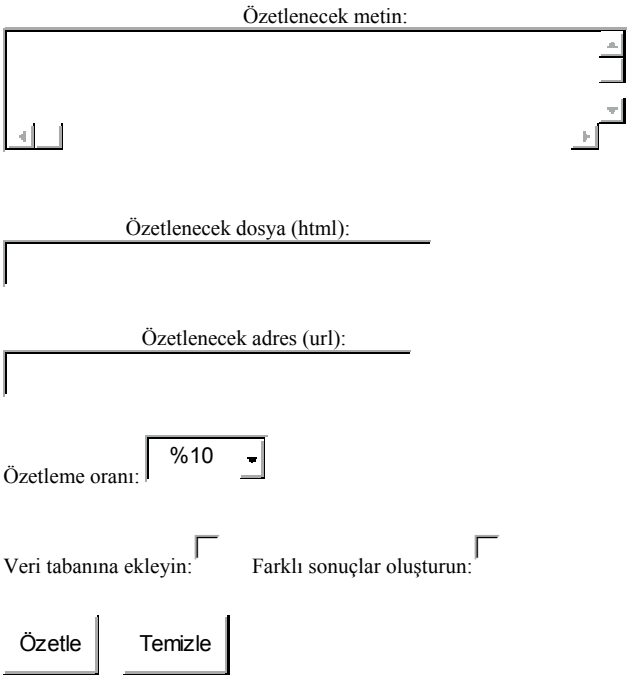

Figure 2: *Automated Summarization System Interface*

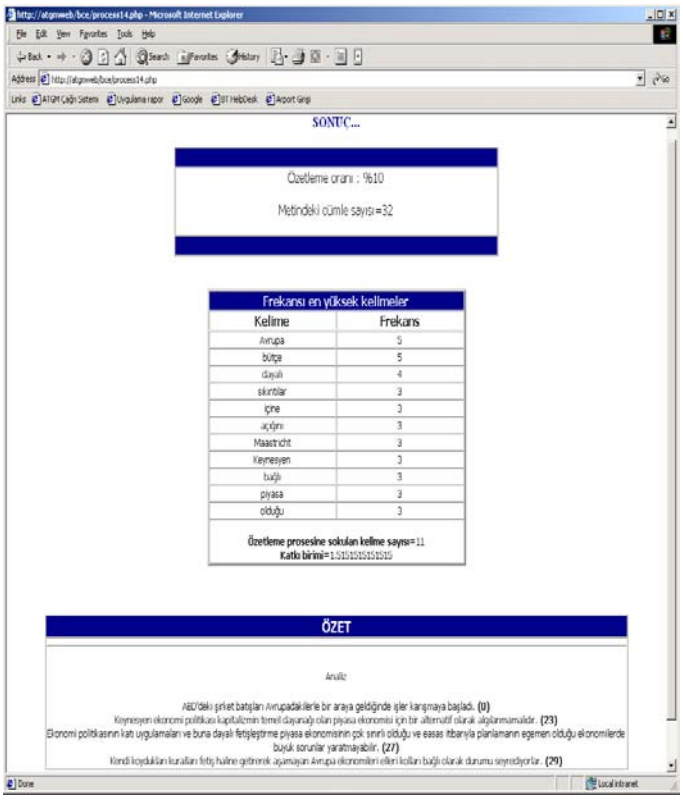

Figure 3*: Summary Generating Screen*