Towards a New Summarization Approach for Search Engine Results: An Application for Turkish

Abstract- **With the drastic increase of available information sources on the Internet, people with different backgrounds share the same problem: locating useful information for their actual needs. Search engines make this task easier only in certain ways; people still have to do the sifting process by themselves. At this point, automatic summarization can complement the task of search engines. In this paper, we consider a new summarization approach for Web information retrieval; i.e. structure-preserving and query-biased summarization. We evaluate this approach on Turkish Web documents using TREC-like topics defined for Turkish. The results of the task-based evaluation show that this approach has significant improvement over Google snippets and unstructured query-biased summaries in terms of f-measure using the relevance prediction approach.**

I. Introduction

The drastic increase of documents on the World Wide Web in recent years has resulted in the wide-spread problem of information overload. That is, people have access to vast amounts of information sources especially with the aid of search engines; however it is getting more and more difficult and time-consuming for them to locate their actual information needs. A recent research shows that about 50% of documents viewed by users in the search engine results turn out to be irrelevant to their actual information need [1].

During information seeking, one aid of users is the short summaries (extracts) of documents listed under each link in the search results [2, 3]. Such summaries may direct users to relevant results and help them save time. The summaries may be especially helpful for specific and complex queries (such as *the effects of earthquakes on human*) rather than the ones with commonplace answers, such as *the date of a major past earthquake*. However, search engine summaries are not always adequate causing the users either to spend time with irrelevant documents or to miss relevant ones. Better methods for summarization can improve the effectiveness of Web search.

Traditional approaches of summarization have usually concentrated on generic summaries of documents. However, in an information retrieval paradigm, it has become important to adapt summaries to user's actual information need; i.e. the query. Also, most of the previous summarization approaches have ignored the structure of a document and have seen the document as a flat sequence of sentences. However, the document structure may be especially helpful in determining the relevancy of a document during information retrieval. First, it can be used to determine important sections and subsections of a document depending on the user query. Second, the structure can be provided as a part of the summary (i.e. headings and subheadings under which the important sentences are located) as clues about the document. In this paper, we provide the application of a query-biased approach utilizing document structure both during the summarization process and in the output summaries to Web search tasks in Turkish. To the best of our knowledge, the proposed approach combining these two aspects was not investigated before in Web search context.

The rest of this paper is organized as follows. First, the related work is given in Section II. This is followed by the structural processing and summarization method in sections III and IV, respectively. Then, the implementation and evaluation details are presented in sections V and VI. We present conclusions in Section VII.

II. Related Work

A. Web Document Analysis

Web document analysis is a younger field of research compared with the previous research on printed document analysis as in [4]. In a related work, the aim is to filter important content from Web documents by eliminating cluttered parts such as advertisements and navigation menus [5].

There is also some work on the identification of the hierarchical structure of Web documents; i.e. the parts and subparts of a document. This has several applications including display of Web document content on small-screen devices and summarization. For hierarchy identification, several methods have been tried in the literature, including [6] and [7]. None of these works use heading information during the identification of document hierarchy.

Another recent work aims at the identification of the main title (i.e. a single title) for Web documents [8]. The more general problem of finding all the headings of a Web document with the underlying hierarchy is defined in [9].

B. Automatic Summarization

Most of the related work in the literature focus on creating generic summaries without considering particular information needs of users, e.g. [10]. In [11] and [12], the effect of query-biased summarization is investigated.

From another perspective, most of the related work ignores document structure. As an attempt to break this limitation, some structure-based summarization approaches have been proposed [13, 14]. Alam et al. [13] propose an approach for summarizing Web documents by making use of the "table of content"-like hierarchy of a document, including sections and subsections. Also, in [14], the summarization method is based on document structure where a document is considered as consisting of multiple levels as chapters, sections, subsections, paragraphs, sentences and terms. In both of these works, the aim is to create general-purpose summaries.

In [15], a structure-based and query-specific summarization approach is proposed. In that work, the structure is achieved by connecting related document fragments (e.g. paragraphs) and obtaining a document graph. However, that work is not based on the explicit structure of a document, i.e. the sectional hierarchy and heading structure.

There is not much work on document summarization for Turkish. In [16], a Turkish automatic text summarization system is developed. The system aims at creating general-purpose summaries of documents.

III. Structural Processing

Traditionally, Web documents are prepared in HTML (Hypertext Markup Language) format whose primary purpose is presentation of data, which brings limitations when a semantic interpretation of document content is desired. To eliminate this problem, semantic markup languages such as XML (Extensible Markup Language) have been developed. However, HTML documents still dominate the Web; our recent analysis on Google results with respect to document types showed that there are nearly 7.4 billion HTML pages but only 12 million XML pages indexed. Therefore, better methods for processing HTML documents are still needed.

In this section, we address the problem of finding the sectional hierarchy of a domain-independent HTML document which can consist of sections and subsections with corresponding headings and subheadings. The proposed structural processing method involves three steps:

(1) Document Object Model (DOM) tree processing
(2) Heading identification
(3) Hierarchy restructuring

In the first step, the DOM tree of a given document is converted to a simplified tree with only containment relationships of container tags; e.g. <*table*>. The format tags (e.g. <*font*>) are passed as features to tree nodes. Then, in the second step, the headings in the document are determined according to heuristics mostly based on content, HTML formatting and position.

In the final step, the tree from the previous steps is restructured bottom-up to obtain the final sectional hierarchy. For this purpose, headings in different levels of the hierarchy are identified based on feature-value pairs. As an example, a particular heading with features *{bold=true, font_size=2, allUperCase=true}* belongs to a different level than one with *{bold=false, font_size=1}*. The output of structural processing step is a tree where non-leaf nodes correspond to headings and subheadings and leaf nodes to underlying sentences as the example document hierarchy in Fig. 1.

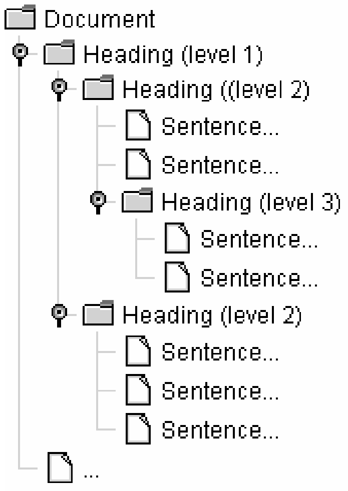


Fig. 1. Example output of structural processing.

IV. Summarization

In the proposed system, the aim of summarization is to create indicative summaries to direct users to relevant documents rather than informative summaries that can be used as a replacement of the original documents. We use the method of sentence extraction rather than sentence abstraction which involves rewriting. In this way, the structure of the original document and the context of the selected sentences can be preserved and thus the user can judge the relevancy of documents more precisely.

The summarization algorithm is run after the structural processing phase is completed. The algorithm utilizes the structural properties of documents both during the summarization process and in the output summaries. In addition, a query-biased approach is employed which is suitable to Web search. We generate the summaries of the documents using two levels of scoring: Sentence scoring and section scoring.

A. *Sentence Scoring*

The sentences are scored based on four different methods used in the summarization literature: Heading, location, term frequency and query methods. The heading and location methods are adapted to the system such that they utilize the output of the structural processing step. Also, stop words are eliminated and stemming is applied whenever relevant.

The intuition behind *heading method* is that headings in a document usually include key words related to the document content (e.g. [14], [17]). For this purpose, the sentences are assigned a heading score based on the number and frequency of the words appearing in a heading of the document.

According to *location method*, the sentences located at certain positions of a document, such as the beginning of the text, usually convey important information (e.g. [12], [14]). In our system, the first sentence of any section or subsection, as identified in the structural processing step, is given a positive score.

The motivation of *term frequency method* originates from the idea that terms frequently occurring within text usually convey important information about the document contents (e.g. [11], [18]). Each sentence is given a term frequency score as the sum of term frequencies of the constituting words.

In the *query method*, the summaries are biased towards user queries (e.g. [11], [15]). Each sentence is given an additional query score as the number of query words it includes.

The overall sentence score is calculated as the weighted sum of each of the four scores where each score is normalized to one as in the following. In the experiments, we used the setting (w_1=w_2=w_3=1 and w_4=3) where the query score is given three times more weight than the others.

$$s_{sentence} = w_1 \times s_{heading} + w_2 \times s_{location} + w_3 \times s_{tf} + w_4 \times s_{query} \quad (1)$$

B. Section Scoring

In the system, each section and subsection of a document is given a section score as a measure of its importance. This score is calculated as the sum of sentence scores in that section. Also, in a hierarchical way, the score of a section is calculated as the sum of its subsection scores.

Each section or subsection is assigned a sentence quota based on the corresponding section score. This quota determines the number of sentences with which that section will be represented in the output summary. The initial quota for the whole document is selected as 25 which is the approximate number of sentences in the output summary. Then, hierarchically, this quota is divided between the sections and subsections as in (2):

$$quota_{subsection} \equiv quota_{section} \times \frac{score_{subsection}}{score_{section}} \quad (2)$$

When the quota of a section or subsection reaches a certain threshold or the section has no more subsections, the highest scored sentences are selected from that section one by one to be included in the summary together with the heading of that section. Also the predecessor headings in the hierarchy, all the way to the main heading, are selected as a part of the summary if not already included. The summarization continues until the summary quota for the whole document is reached.

V. Implementation

The system was developed using the GATE framework [19] for text engineering as the underlying development environment which is an open source project based on component-based technology in Java. In the proposed system, after an HTML document is loaded to the system, the following processing resources (modules) are applied to it in the indicated order and the final summary is generated:

- *Tokeniser* - splits the text into tokens, such as words, numbers and punctuation marks.
- *Sentence Splitter* - splits the text into sentences.
- *Stemmer* - applies stemming to individual words.
- *HTML Document Structure Analyzer* - applies the proposed structural processing algorithm on the document.
- *Summarization Engine* - runs the proposed summarization method on the document.

The tokeniser and sentence splitter were taken from ANNIE, a GATE implementation of an information extraction system. The stemmer used is the Turkish version of Porter's stemmer [20]. We implemented two new processing resources as plugins for GATE: HTML Document Structure Analyzer and Summarization Engine.

VI. Evaluation

The system was evaluated using two different experimental settings. In the first experiment, the accuracy of the structural processing step, i.e. heading-based sectional hierarchy identification, was measured. In the second experiment, the effectiveness of the summaries created by the proposed system was evaluated. In the experiments, five queries (see Table I) were used from a TREC-like test collection for Turkish [21]. For each query, ten documents were randomly collected from the top 50 results of Google in response to that query.

TABLE I
QUERIES USED IN THE EXPERIMENTS

Query Id	Query Keywords
1	*Tsunami* (Tsunami)
2	*Ekonomik kriz* (Economic crisis)
3	*Türkiye'de meydana gelen depremler* (Earthquakes in Turkey)
4	*Sanat ödülleri* (Art awards)
5	*Bilisim egitimi ve projeleri* (IT education and projects)

A. Hierarchy Identification Experiment

The documents in the test set are first manually investigated and their sectional hierarchies and headings are marked as the golden standard. Then, the accuracy of automatic hierarchy identification is calculated as the ratio of the number of correctly identified parent-child relationships (as compared with the golden standard) over the total number of parent-child relationships.

Table II shows the average DOM tree and hierarchy depths for the documents, and the average accuracy obtained for hierarchy identification. In Table III, the average number of headings in the documents and the performance of heading identification in terms of recall (R), precision (P) and f-measure (F) are given.

TABLE II
HIERARCHY IDENTIFICATION RESULTS

DOM Tree Depth	Hierarchy Depth	Hierarchy Accuracy
17.2	6.1	0.70

TABLE III
HEADING IDENTIFICATION RESULTS

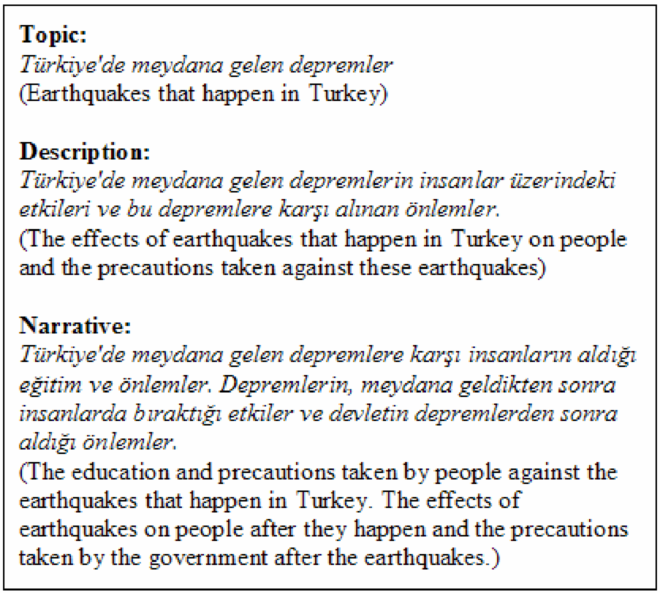
Average Number of Headings	R	P	F
5.40	0.79	0.57	0.65

B. *Task-based Evaluation*

The effectiveness of the proposed system summaries were tested on a task-based evaluation. For comparison, four different types of summaries are used in the experiment:

- *Google*: Query-biased extracts of Google
- *Unstructured*: Query-biased summaries without the use of structural information
- *Structured1*: Query-biased summaries using output of the structural processing step
- *Structured2*: Query-biased summaries using manually identified structure.

In the experiment, we used a repeated measures design in order to reduce the differences among subjects [22]. Three subjects were used and each subject was required to complete all the queries in Table I on a Web-based interface. The subjects are given a description and a narrative for each query, as in Fig. 2, and were requested to make a judgment on the relevancy of the original document given the summary. The summaries are displayed in a random order to reduce carryover effects. Also, the recently proposed relevance prediction approach was used in the experiments [23]. That is, subject's judgment on a summary is compared with his/her own judgment on the original document instead of a gold standard.



Topic:
Türkiye'de meydana gelen depremler
(Earthquakes that happen in Turkey)

Description:
Türkiye'de meydana gelen depremlerin insanlar üzerindeki etkileri ve bu depremlere karşı alınan önlemler.
(The effects of earthquakes that happen in Turkey on people and the precautions taken against these earthquakes)

Narrative:
Türkiye'de meydana gelen depremlere karşı insanların aldığı eğitim ve önlemler. Depremlerin, meydana geldikten sonra insanlarda bıraktığı etkiler ve devletin depremlerden sonra aldığı önlemler.
(The education and precautions taken by people against the earthquakes that happen in Turkey. The effects of earthquakes on people after they happen and the precautions taken by the government after the earthquakes.)

Fig. 2. Details of the example query used in the experiments.

The summaries except Google extracts are longer summaries and have the same size (around 25 sentences) to make them comparable. An example structured summary output by the proposed system (for the third query in Table I) is given in Fig. 3. As seen, the summaries are displayed in a hierarchical way in accordance with the sectional hierarchy obtained in the structural processing step. Also, headings and subheadings are given as bold and query keywords are highlighted.

For each summary type, four different results were identified by comparing the relevancy judgments for the summary and the original document: TP (true positive), FP (false positive), FN (false negative), and TN (true negative) as in Table IV. Based on these values, accuracy (A), recall (R), precision (P), and f-measure (F) values for the summarization experiment were calculated.

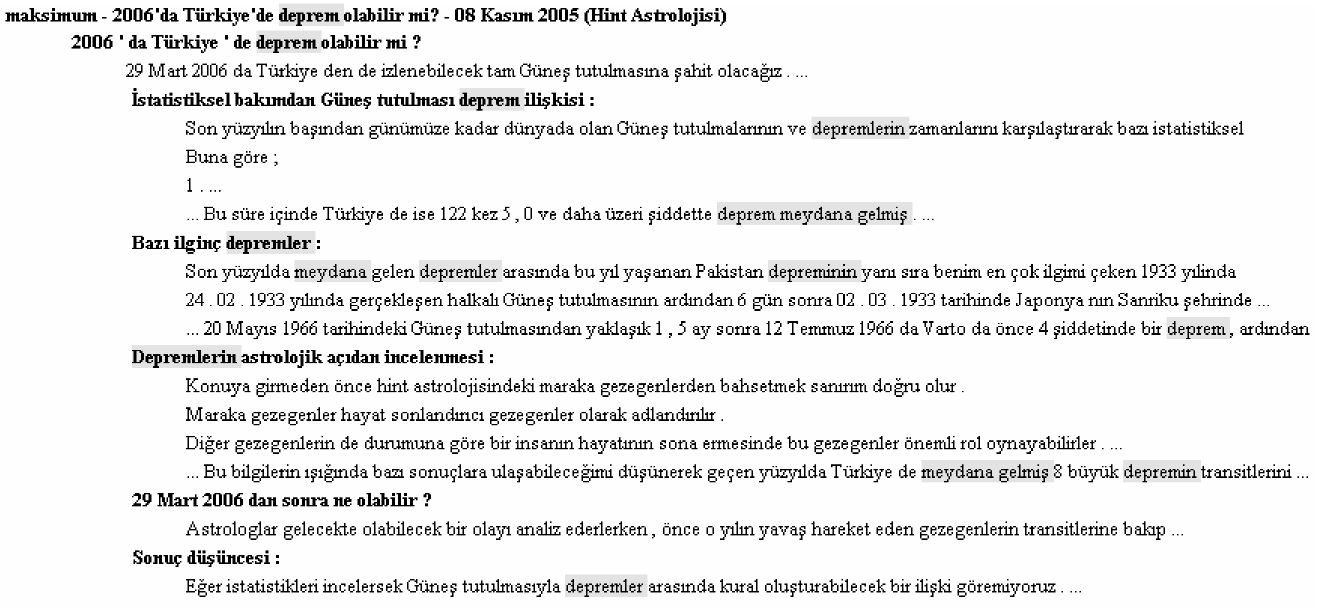
TABLE IV
CONTINGENCY TABLE FOR THE SUMMARIZATION EXPERIMENT

		Original document judgment	
		relevant	irrelevant
Summary	relevant	TP	FP
	irrelevant	FN	TN

The effectiveness of each method is given in Table V. The results show that structured summaries (*Structured1* and *Structured2*) are superior to unstructured ones and Google snippets. A structured summary provides an overview of the document and makes it for the user much easier to focus on the relevant parts by ignoring the irrelevant details. Table V also shows the average judgment times of the users for each method. Although *Structured1*, *Structured2* and *Unstructured* methods provide summaries much longer than Google snippets, we see that there is only about two times increase in response time. This may indicate that people just look at the related parts on the summarized text without delving into the details. We can conclude that the response times of the proposed method are acceptable. Another justification in favor of this argument is the extra time spent in case of irrelevant documents. When the user clicks on the link of an irrelevant document, he/she will spend some time during page loading and to understand that the document is in fact irrelevant. Thus it will take more time to find the desired documents. Due to the lower performance ratios, this situation occurs more frequently in the case of short extracts and unstructured summaries.

TABLE V
SUMMARIZATION EXPERIMENT RESULTS

System	A	P	R	F	Time
Google	0.80	0.72	0.76	0.74	11.04
Structured 1	0.91	0.87	0.91	0.89	19.96
Structured 2	0.88	0.86	0.87	0.86	19.71
Unstructured	0.83	0.73	0.72	0.72	19.96



maksimum - 2006'da Türkiye'de deprem olabilir mi? - 08 Kasım 2005 (Hint Astrolojisi)

2006 ' da Türkiye ' de deprem olabilir mi ?

29 Mart 2006 da Türkiye den de izlenebilecek tam Güneş tutulmasına şahit olacağız . ...

İstatistiksel bakımdan Güneş tutulması deprem ilişkisi :

Son yüzyılın başından günümüze kadar dünyada olan Güneş tutulmalarının ve depremlerin zamanlarını karşılaştırarak bazı istatistiksel

Buna göre ;

1 . ...

... Bu süre içinde Türkiye de ise 122 kez 5 , 0 ve daha üzeri şiddette deprem meydana gelmiş . ...

Bazı ilginç depremler :

Son yüzyılda meydana gelen depremler arasında bu yıl yaşanan Pakistan depreminin yanı sıra benim en çok ilgimi çeken 1933 yılında

24 . 02 . 1933 yılında gerçekleşen halkalı Güneş tutulmasının ardından 6 gün sonra 02 . 03 . 1933 tarihinde Japonya nın Sanriku şehrinde ...

... 20 Mayıs 1966 tarihindeki Güneş tutulmasından yaklaşık 1 , 5 ay sonra 12 Temmuz 1966 da Varto da önce 4 şiddetinde bir deprem , ardından

Depremlerin astrolojik açıdan incelenmesi :

Konuya girmeden önce hint astrolojisindeki maraka gezegenlerden bahsetmek sanırım doğru olur .

Maraka gezegenler hayat sonlandırıcı gezegenler olarak adlandırılır .

Diğer gezegenlerin de durumuna göre bir insanın hayatının sona ermesinde bu gezegenler önemli rol oynayabilirler . ...

... Bu bilgilerin ışığında bazı sonuçlara ulaşabileceğimi düşünerek geçen yüzyılda Türkiye de meydana gelmiş 8 büyük depremin transitlerini ...

29 Mart 2006 dan sonra ne olabilir ?

Astrologlar gelecekte olabilecek bir olayı analiz ederlerken , önce o yılın yavaş hareket eden gezegenlerin transitlerine bakıp ...

Sonuç düşüncesi :

Eğer istatistikleri incelersek Güneş tutulmasıyla depremler arasında kural oluşturabilecek bir ilişki göremiyoruz . ...

Fig. 3. An example structure-preserving and query-biased summary.

Table VI shows the performance improvement provided by the proposed method (*Structured1*) over *Google* and *Unstructured* summaries. The proposed system has 20.3% improvement over Google and 23.6% improvement over unstructured summaries in terms of f-measure. The statistical tests (repeated measures ANOVA) we performed on the performance ratios verify that Structured1 method yields significantly better results than both Google and unstructured summaries with $p<0.05$ for f-measure.

TABLE VI
IMPROVEMENT OF PROPOSED SYSTEM OVER OTHER METHODS

	Google	Unstructured
A	+13.8%	+9.6%
P	+20.8%	+19.2%
R	+19.7%	+26.4%
F	+20.3%	+23.6%

VII. Conclusion

In this paper, we investigated a new approach to summarization of Web documents. It is a query-biased method utilizing document structure both during the summarization process and in the output summaries where it is distinguished from traditional summarization approaches. The method contains two stages: automatic analysis of document structure and summarization.

The approach was tested in two steps using Turkish Web documents collected from the results of Google. In the first step, the accuracy of the structural processing is evaluated where acceptable performance is obtained. In the second step, the effectiveness of summarization is evaluated on an information retrieval task using TREC-like queries developed for Turkish. The results show that the proposed system has significant improvement over both Google (20.3%) and unstructured summaries of the same size (23.6%) in terms of f-measure.

As a future work, the structural processing stage will be improved using machine learning techniques. Also the summarization method will be improved with other natural language processing methods, including incorporation of syntactic phrases and WordNet information.

Acknowledgments

References

[1] B. J. Jansen and A. Spink, "An analysis of web searching by European Alltheweb.com users," Information Processing and Management, vol. 41-2, 2005, pp. 361-381.

[2] Google, 2008. http://www.google.com.

[3] AltaVista, 2008. http://www.altavista.com.

[4] D. Niyogi and N. Srihari, "Knowledge-based derivation of document logical structure," Proceedings of the Third International Conference on Document Analysis and Recognition, 1995.

[5] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren, "Automating content extraction of HTML documents," World Wide Web, vol. 8, 2005, pp. 179 – 224.

[6] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, "Accordion summarization for end-game browsing on PDAs and cellular phones," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2001.

[7] S. Vadrevu, F. Gelgi, and H. Davulcu, "Information extraction from web pages using presentation regularities and domain knowledge," World Wide Web, 10-2, 2007, pp. 157–179.

[8] Y. Xue, Y. Hu, G. Xin, R. Song, S. Shi, Y. Cao, C.-Y. Lin, and H. Li., “Web page title extraction and its application,” Information Processing and Management, vol. 43, 2007, pp. 1332-1347.

[9] F. C. Pembe and T. Güngör, ”Heading-based sectional hierarchy identification for HTML documents,” Proceedings of the 22nd International Symposium on Computer and Information Sciences, IEEE Xplore, 2007.

[10] S. F. Liang, S. Devlin, and J. Tait, ”Investigating sentence weighting components for automatic summarisation,” Information Processing and Management, vol. 43-1, 2007, pp. 146-153.

[11] A. Tombros and M. Sanderson, ”Advantages of query biased summaries in information retrieval,” Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[12] R. W. White, J. M. Jose, and I. Ruthven, “A task-oriented study on the influencing effects of query-biased summarization in Web searching,” Information Processing and Management, vol. 39, 2003, pp. 707-733.

[13] H. Alam, A. Kumar, M. Nakamura, A. F. R. Rahman, Y. Tarnikova, and C. Wilcox, ”Structured and unstructured document summarization: design of a commercial summarizer using lexical chains,” Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003.

[14] C.C. Yang and F.L. Wang, ”Fractal summarization for mobile devices to access large documents on the web,” Proceedings of the International12th WWW Conference, 2003.

[15] R. Varadarajan and V. Hristidis, ”Structure-based query-specific document summarization,” Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005.

[16] Z. Altan, ”A Turkish automatic text summarization system,” Proceedings of IASTED International Conference on Artificial Intellegence and Applications, 2004.

[17] H. P. Edmundsun, “New methods in automatic extracting,” Journal of the ACM, vol. 16-2, 1969, pp. 265-285.

[18] H. P. Luhn, “The automatic creation of literature abstracts,” IBM Journal of Research and Development, vol. 2, 1958, pp. 159-165.

[19] GATE, A General Architecture for Text Engineering, 2007. http://gate.ac.uk/.

[20] M. Porter, “An algorithm for suffix stripping,” Program, 14(3), 1980, pp. 130–137.

[21] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, and O. M. Vursavas, "Information retrieval on Turkish texts," Journal of the American Society for Information Science and Technology, vol. 59-3, 2008, pp. 1-15.

[22] D. C. Montgomery, Design and Analysis of Experiments, John Wiley, New York, USA, 2001.

[23] S. P. Hobson, B. J. Dorr, C. Monz, and R. Schwartz, “Task-based evaluation of text summarization using relevance prediction,” Information Processing and Management, 43-6, 2007, pp. 1482-1499.