Generic Text Summarization for Turkish

Celal Cığır
Department of Computer Science
Bilkent University
Ankara, Turkey
cigir@cs.bilkent.edu.tr

Mücahid Kutlu
Department of Computer Science
Bilkent University
Ankara, Turkey
mucahid@cs.bilkent.edu.tr

Ilyas Cicekli
Department of Computer Science
Bilkent University
Ankara,Turkey
ilyas@cs.bilkent.edu.tr

Abstract — **In this paper, we propose a generic text summarization method that generates summaries of Turkish texts by ranking sentences according to their scores calculated using their surface level features and extracting the highest ranked ones from the original documents. In order to extract sentences which form a summary with an extensive coverage of main content of the text and less redundancy, we use the features such as term frequency, key phrase, centrality, title similarity and position of the sentence in the original text. Sentence rank is computed using a score function that uses its feature values and the weights of the features. The best feature weights are learned using machine learning techniques with the help of human constructed summaries. Performance evaluation is conducted by comparing summarization outputs with manual summaries generated by 25 independent human evaluators. This paper presents one of the first Turkish summarization systems, and its results are promising.**

Keywords: *Text Summarization, Summary Extraction, Natural Language Processing*

I. INTRODUCTION

In past, retrieving any information about a subject was hard because of lack of resources. However, today, resources have increased in an uncontrolled manner by the exponential growth of world-wide web. Due to huge amount of information on the Internet, information retrieval technologies have become more popular for finding relevant information effectively. Text search engines return hundreds or even thousands of pages that people are overwhelmed to identify which page corresponds to their needs. These huge amount of retrieved pages make people to have a decision about which page contains more relevant to their needs. Therefore, there is an increasing need for new technologies that help users to access desired and relevant information quickly. Providing the documents with summary of each will smooth the progress of finding the desired documents. Text search and summarization are technologies to reduce the access time for information. Text search engines filter the pages according to user query and generate an initial set of relevant documents and text summarizers generates summary of documents that enables users quickly identify the content of the text to determine final set of relevant documents [5].

A document generally consists of several topics. Some topics are explained deeply by many sentences, and therefore form the main content of the document. Other topics may just be briefly described and supplied to make the whole story more complete. A good generic summary should cover the major topics of the text as much as possible while keeping redundancy to a minimum.

Automatic text summarization takes a document (or documents) as input and presents a well-formed summary by extracting the essence of the document(s) [4]. In text summarization, we can use sentence extraction or abstraction method. Abstraction is a method for novel phrasing describing the content of the text which requires heavy machinery from natural language processing, including grammars and lexicons for parsing and generation. Extraction is a method for determining salient text units (typically sentences) by looking at the text unit's lexical and statistical relevance or by matching phrasal patterns [9]. We cannot say any of the approaches are absolutely better than the other. Abstraction approaches provide sophisticated summaries and adapt well to high compression rates while extraction approaches are easy to adapt larger sources although the resulting summaries may be incoherent. The extraction method is used in the text summarization system presented in this paper.

In this paper, we propose generic text summarization method that creates summaries of Turkish texts by ranking and extracting valuable sentences from the original documents. This method uses information retrieval (IR) techniques like term frequency and Natural Language Processing (NLP) techniques like key phrase and centrality. In addition, position of sentences in the original document and existence of title keywords in sentences are some of heuristic approaches that are used to generate summaries. Our method aims to rank the sentences in the document and extract the higher ones in order to generate a summary with an extensive coverage of main content of the document. A score function of mentioned features is used to rank the sentences, and machine learning techniques are used to determine the optimal combination of coefficients of features. Performance evaluation is conducted by comparing summarization outputs with manual summaries generated by 25 independent human evaluators. ROUGE evaluation technique [15] is used to compare summarization outputs with human generated summaries.

The contribution of our study is the construction of a single document summarization system for Turkish texts by observing effects of features of sentences to form a good summary. The presented Turkish text summarization system is one of the first Turkish text summarization systems, and its results are promising. Moreover, summarization studies on Turkish texts are not sufficient, and there is no corpus for Turkish summarization systems. With 115 human generated summaries, this study contributes the researchers who want to study text summarization on Turkish texts.

The remaining parts of the paper consist of four sections: Section 2 describes related works, Section 3 describes the proposed technique and system, Section 4 presents the performance evaluations and Section 5 concludes the paper by summarizing the study and gives some future work.

II. Related Work

Text summarization has been studied since 1950s [10] and a variety of summarization methods has been proposed and evaluated. There are two ways of summarization: abstraction and sentence extraction. In fact majority of researches have been focused on summary extraction, which selects the pieces such as keywords, sentences or even paragraph from the source to generate a summary. Abstraction can be described as "reading and understanding the text to recognize its content which is then compiled in a concise text." [2]. Hovy & Lin distinguished summaries as indicative vs. informative; generic vs. query-based; single-document vs. multi-document [1]. Our proposed summarization system can be categorized as a generic single-document summarization system that uses sentence extraction.

Lin studied on a selection function for extraction and used a machine learning algorithm to automatically learn good features coming from several heuristics [3]. On the other hand, Yeh & Ke used Latent Semantic Analysis (LSA) approach for extraction and compared the results with the feature extraction algorithm [4]. Gong and Liu are also worked on summarization by using relevance measure and LSA [5].

Barzilay and Elhadad [11] describe a summarization system based on lexical chains of words. Since a lexical chain for a set of words is created if those words are semantically related, the semantic relations among the words play a role in the sentence extraction. Brunn et al. [12] and Doran et al. [13] also use semantic relations among the words in their summarization systems. In order to improve the sentence selection, Ercan and Cicekli [14] use the clusters of lexical chains instead of lexical chains alone.

The summarization system described in this paper extracts the sentences depending on their surface level features. An original feature that is tried for this Turkish summarization is the key phrase feature. The weights of features are determined with the help of machine learning techniques.

III. System Description

In this section, we propose a method that creates generic summaries by selecting valuable sentences with the help of a score function. This formula takes into account several kinds of document features, including term frequency, key phrase, and positions of the sentences in the original text, centrality and existence of title keywords in the sentences to generate summaries. First of all, document is decomposed into individual sentences for further score computation. Later on, sentences are ranked to emphasize the significance of different sentences. Finally, the top scored sentences are selected in the order how they are in the original document to generate a well-formed summary. The details of features and summary generation part are explained in following subsections.

A. Feature Selection

In order to find the score of a sentence S, indicating the degree whether it belongs to the summary or not, the following features are used in the score calculation.

1) f_1: Term Frequency (TF): The term frequency of a term in a given document is simply the number of occurrences of the term in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term tf_i within the particular document. Term frequency of a term is calculated as follows:

$$tf_i = \frac{n_i}{\sum_k n_k} \qquad (1)$$

where n_i is the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms in document (1). In our case, only nouns are considered as terms.

The term frequency score of a sentence S is defined as (2).

$$Score_{f_1}(S) = \frac{\sum_k tf_k}{number\ of\ nouns\ in\ S} \qquad (2)$$

To avoid the bias of the sentence length, the summation of TF scores of the nouns in the sentence is normalized by the length of the sentence, which is number of nouns in the sentence.

In order to find the nouns in a sentence, Zemberek which is an open source ongoing study on Turkish Language is used in this study [6].

2) f_2: Title Similarity (TS): Titles contain groups of words that give important clues about subjects of documents. Therefore, if a sentence S has higher intersection with the title words than others, then we can assume that S is more important than others. Hence, we can formalize TS score of a sentence S as (3).

$$Score_{f_2}(S) = \frac{|words\ in\ s \cap words\ in\ the\ title|}{|words\ in\ s \cup words\ in\ the\ title|} \qquad (3)$$

3) f_3: Key Phrases (KP): Since words are the essential elements of a sentence, the more content-covering keywords that a sentence has, the more important it is. Key phrases are short noun phrases that capture the main topics discussed in a given document. In order to evaluate key phrase scores of sentences, the key phrases of the document are required. The key phrases of a document are found by the key phrase extraction system described in [7]. However, the number of key phrases should be limited for summarization because a high amount of key phrases decreases the importance of key phrases. Therefore, we limited our key phrase number to 10. For a sentences S, the key phrase score is defined as follows:

$$Score_{f3}(S) = \frac{number\ of\ key\ phrases}{number\ of\ nouns\ in\ S} \tag{4}$$

The number of key phrases in the sentence is divided by the number of nouns in S in order to avoid the bias of the sentence length.

4) f_4: Sentence Position (SP): Locations of sentences are important in order to get a well-formed, easy-understandable document. Sentences at the beginning of the texts, especially of news documents, give the general information of the document which are suitable to form a summary. The sentences at the middle of the documents are details about the news which we need them less in a summary. Therefore we can say that important sentences, which should be included in the summary, are usually located at some particular positions [4]. In order to formulize the sentence position, we give each sentence a position value P_i (P_i is equal to i). Then to give higher score to the first sentences, we use the formula in (5) which gives the position score of a sentence S.

$$Score_{f_4}(S) = \frac{R - P_i}{R} \tag{5}$$

where R is the total number of sentences in the corresponding document.

5) f_5: Centrality (C): The centrality of a sentence implies its similarity to others, which can be measured as the degree of vocabulary overlapping between the sentence and other sentences. If a sentence has high centrality, then we can use that sentence in the summary to introduce many topics of the document. Therefore, high centrality sentences are more preferable in summary than low centrality sentences. To find score of centrality of a sentence S, the formula defined in (6) is used.

$$Score_{f_5}(S) = \frac{|words\ in\ s \cap words\ in\ other\ sentences|}{|words\ in\ s \cup words\ in\ other\ sentences|} \tag{6}$$

B. Summary Generation

For a sentence *S*, the weighted score function in (7) is used to combine all the feature scores of the sentence, where w_i indicates the weight of score of feature f_i.

$$Score(S) = \sum_i w_i \times Score_{f_i}(S) \tag{7}$$

After finding scores of all sentences of a document, sentences are ranked according to their scores and top-ranked sentences are selected and those sentences are sorted according to their order in the document in order to have a well-formed summary.

Score function is trained by using machine learning techniques in order to obtain a suitable combination for feature weights. For this aim, a training set which consists of documents and their human generated summaries. For this training set, all possible weight combinations between 0 and 1 where interval between two features are at least 0.01 are experimented. Then, the one that generates highest average accuracy result for the training set is selected. The found weights of features are presented in the evaluation part.

IV. EVALUATION

In this section, we describe our corpus and evaluation techniques, and the results of our summarization system.

A. Data Corpus

25 independent human annotators who are senior and graduate students helped us to construct our corpus. Each of annotators selected news articles independently without any restriction on subject and sources of the news. We obtained articles in the domain of politics, sports, economy, entertainment and etc. They are collected from the online Turkish newspapers such as Milliyet News, Hurriyet News, Zaman News and some news portals. Also there was no restriction on the size of summaries. Hence, we get a chance to observe the size of a summary that humans think considerable. Table 1 shows statistics of the data corpus.

TABLE I. STATISTICS OF THE CORPUS

Property	Value
Sentences per document	22.5
Sentences per manual summary	5.2
Words per document	348.7
Words per manual summary	95.7
Words per document sentence	15.5
Words per manual summary sentence	18.3

This data corpus is one of the main contributions of this study since there is no enough study on Turkish and there is no reference summary data corpus for Turkish. It is available to ones who want to study in this field and need data corpus.

B. Performance Evaluation

There are different methods to evaluate the performance of a text summarization system. In this study, we have chosen *intrinsic evaluation*. Intrinsic evaluation judges the quality of a machine generated summary based on the correspondence between the generated summary and the human generated summary. We have used precision (P) and recall (R) to judge the coverage between manual and machine generated summaries. Assume *T* is a manual summary and *S* is a machine generated summary, the precision can be defined as (8) and recall can be defined as (9).

$$Precision: P = \frac{|S \cap T|}{|S|} \tag{8}$$

$$Recall: R = \frac{|S \cap T|}{|T|} \tag{9}$$

We separated the corpus into a training set consisting of 50 articles and a test set consisting of 65 articles. First, we generated summaries of 65 articles of corpus, which are not used for training, by using each feature individually in order to see the effect of each feature. In Table 2, the recall and precision results of each feature tested individually can be found. When only TF is considered, the recall result is 0,197. TS, KP, POS and CN produce the results of 0,262, 0,282, 0,339 and 0,241 respectively. The highest recall and precision results are produced by the sentence position feature. The reason for this is that our data corpus consists of news articles and first sentences of news articles usually mention the main idea of the article, thus, first sentences are suitable to form a summary. The feature that gives the second highest recall is key phrase feature which is an expected result since key phrases are noun phrases that covers the content more than other nouns. Title similarity is also useful to form generic summaries. This is because titles should represent the topic of the article since people select news articles to read by looking their titles.

TABLE II. RECALL AND PRECISION RESULTS OF EACH FEATURE TESTED INDIVIDUALLY

Features	Recall results	Precision Results	F-measure
TF	0,197	0,252	0,221
TS	0,262	0,306	0,282
KP	0,282	0,313	0,297
SP	0,339	0,402	0,368
C	0,241	0,271	0,255

To generate a well-formed and most content covering summary, we can use all features and each can have different weights according to their importance. In order to find the optimal weight of the features, all possible combinations in range 0 to 1 were tested with 50 articles of the data corpus. The one which maximizes the *f-value* is selected. The compression rate is set to %21 which is average compression rate of training part of the corpus and this rate is also enough for a well-formed summary. Table 3 represents the optimal weights of each feature that gives maximum f-value, which is 0,436, for the training set.

TABLE III. OPTIMAL WEIGHTS OF EACH FEATURE

Features	Optimal weights
Term frequency	0,02
Title Similarity	0,18
Key Phrase	0,08
Position	0,48
Centrality	0,24

When we look at Table 3, term frequency seems to have lowest effect on summary and position has the highest. However, we cannot say this without performing a test for recall and precision values on a different corpus since heuristics used for scoring sentences may affect the coefficients highly. That is to say, for example, if we use five key phrases instead of ten, the coefficient of key phrase would increase, vice versa. Therefore, finding optimal coefficients eliminates the anomalies come from heuristics.

After finding optimal weights of each feature, we have tested the system with the remaining 65 articles. In Table 4, recall and precision results of applying all features and all quadruple combinations of features with their weights in Table 3 can be found. When all features are taken into account, recall result is 0,324 and precision is 0,354. However, when we used all features except title similarity or except centrality, recall and precision results interestingly increases which means they had negative effect for test part of corpus. On the other hand, when we ignore sentence position or term frequency, recall and precision decreases which means that they both have a positive effect for generating a well-formed summary. Key phrase has a little negligible effect on recall and negative effect on precision. In addition, we can say that sentence position is the most important feature since the decrease is the highest when sentence position is ignored.

TABLE IV. RECALL AND PRECISION RESULTS OF APPLYING ALL FEATURES AND ALL QUADRUPLE COMBINATIONS OF FEATURES

Features	Recall results	Precision Results	F-measure
All Features	0,324	0,354	0,338
Without TF	0,321	0,350	0,335
Without TS	0,350	0,399	0,373
Without KP	0,324	0,361	0,341
Without SP	0,284	0,314	0,298
Without C	0,325	0,364	0,338

Our evaluation is performed by intrinsic method which checks exact match of sentences. That is to say, if we have two different sentences that have same (or similar) meaning and only one of them is in summary. Then if we choose the other one, we will get a zero point because of this mismatch. However, it is obvious that using different sentences that have same (or similar) meaning will not decrease the quality of summary so much. Therefore, we have also evaluated our system with ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluation technique [15]. In Table 5, ROUGE results for recall and precision values of each feature tested individually can be found. The reason for high ROUGE scores is that the human annotators also selected the sentences from texts to create their summaries.

When we compare Table 2 and Table 5, we interestingly see the huge increase in recall value of centrality. Centrality gives the highest recall value when we used ROUGE and lowest value in intrinsic method. Therefore, we can say that centrality is also a good measure to generate a good summary. However, in terms of precision, centrality becomes the 4th of the features while sentence position gives the highest score. While title similarity and term frequency gives similar results as in intrinsic method, key phrase becomes the worst feature.

TABLE V. ROUGE RESULTS FOR RECALL AND PRECISION VALUES OF EACH FEATURE TESTED INDIVIDUALLY

Features	Recall results	Precision Results	F-measure
TF	0,369	0,718	0,487
TS	0,405	0,744	0,524
KP	0,250	0,660	0,363
SP	0,490	0,779	0,602
C	0,513	0,692	0,589

In Table 6, there are ROUGE results for recall and precision values of applying all features and all quadruple combinations of features with their weights in Table 3. In terms of recall, term frequency and sentence position have no effect and key phrase has insignificant negative effect. Centrality and title similarity have positive effect on recall values which is opposite of evaluation by intrinsic method. This indicates that we can form good summaries without using sentence position feature which is known as most significant feature. That is to say, our rest of the features are good enough to select the middle-positioned sentences giving the main idea of the article. In terms of precision, while term frequency has no effect, centrality seems to have negative effect and the rest has positive effect.

TABLE VI. ROUGE RESULTS FOR RECALL AND PRECISION VALUES OF APPLYING ALL FEATURES AND ALL QUADRUPLE COMBINATIONS OF FEATURES

Features	Recall results	Precision Results	F-measure
All Features	0,540	0,809	0,648
Without TF	0,540	0,809	0,648
Without TS	0,534	0,798	0,640
Without KP	0,543	0,805	0,649
Without SP	0,540	0,770	0,635
Without C	0,540	0,809	0,648

When we look at the big picture, we can say that sentence position is the most important feature to form a good summary. However, not using sentence position does not decrease ROUGE recall value which indicates that other features are also good enough. Centrality is another important feature. By using centrality feature, although we may not select the exact sentences that users selected (intrinsic method), the system-selected sentences are similar to the user-selected sentences. This explains the reason of having big difference between intrinsic results and ROUGE results. Title similarity also acts like centrality giving low results in intrinsic method and higher in ROUGE results. Term frequency has little effect on summary since term frequency is also used for finding key phrases. Therefore, key phrase may decrease its effect. Although key phrases have small effect on summary, this can be improved by developing the better extraction method for key phrases. The number of key phrases may also affect the result. We can claim that key phrase feature is a good feature that can be used in generation of summaries.

V. CONCLUSION AND FUTURE WORK

In this study, a generic text summarization system for Turkish texts is developed using sentences extraction method. The surface-level document features such as term frequency, title similarity, key phrases, position of the sentence in the document, and centrality of the sentence are used to determine significance of the sentence. These document features are combined by a scoring function in which each feature has a different weight. The most suitable combination of feature weights is obtained by using a training corpus. This scoring function aims to rank sentences and summary generation is performed by selecting the top-ranked sentences. In order to test the system, we used a test corpus having more documents than training corpus. We have obtained 0,54 recall and 0,809 precision values with ROUGE evaluation method when compression rate is the average compression rate for test set of corpus.

Besides building a summarization system for Turkish articles working with high precision value, data corpus is another main contribution of this study. Since there is not enough study on Turkish text summarization, the data corpus is available for ones who study on Turkish language.

We have just made an introduction to Turkish text summarization by this study. Therefore, we have many future works to do after this study. We plan to focus on observing effect of key phrases by changing scoring function of key phrase and number of key phrases while trying different compression rates.

Improvement of existing document features and adding some new features such as cue phrases, conjunctions and answers of 5W1H (Who, Where, Why, When, What and How) is also a future work. Some words make the sentences more important than the others which can be seen as cue phrases. For instance, “Özetle”, “Neticede” (as a conclusion) in Turkish, summarizes and concludes the document content and therefore their occurrences in the summary makes the summary more content-bearing. In addition to that, Turkish has some conjunction words such as “Çünkü” (because), “Ancak” (but) etc. but their usage in the sentences makes the sentences longer than the other and these sentences mostly includes detailed information. Since, our aim is to make the summary as concise as possible; these sentences need to be eliminated. On the other hand, the sentences with these conjunction words sometimes carry the content of the text. In order to make a distinction, the weight of this feature will determine the importance of these sentences. Moreover, news include answers of questions “Who, Where, Why, When, What, How” which are known as 5W1H. As a characteristic of news, sentences which answer one of these questions get more important than others. Finding cue phrases, conjunctions and answers of 5W1H will make our study to more specific work on Turkish language which is important because of existence of less study on Turkish.

Besides these, we plan to apply Latent Semantic Analysis (LSA) combined with document features to Turkish text summarization. We also plan to use other semantic based approaches to text summarization in Turkish text summarization systems.

ACKNOWLEDGMENT

This work is partially supported by The Scientific and Technical Council of Turkey Grant ``TUBITAK EEEAG-107E151".

We would like to thank all students of Information Retrieval course of Bilkent University in 2008 and course instructor Prof. Dr. Fazlı Can for their contribution in construction of our corpus.

REFERENCES

[1] Hovy, E. & Lin, C.Y. (1997). “Automatic text summarization in SUMMARIST”. In proceedings of the ACL’97/EACL’97 workshop on intelligent scalable text summarization (pp. 18-24) Madrid, Spain.

[2] H. Saggion, “Automatic abstracting; towards a text based generation”. Universite de Montreal.

[3] Lin, C.Y. (1999). “Training a selection function for extraction.”. In proceedings of the 8th international conference on information and knowledge management (CIKM’99) (pp. 55-62), Kansas City, MO. USA.

[4] Yeh, J. Y., Ke, H.R.& Yang, W.P. “Text summarization using a trainable summarizer and Latent Semantic Analysis” . In Proceedings of Information processing and management 2005 (pp.75-95)

[5] Gong, Y. & Liu, X. (2001). “Generic text summarization using a trainable summarizer and Latent Semantic Analysis”. In Proceedings of 24th annual international ACM SIGIR conference on research and development in information retrieval (SGIR’01) (pp. 19-25), New Orleans, LA, USA.

[6] Zemberek website : http://code.google.com/p/zemberek/

[7] Fırat Kalaycılar & Ilyas Cicekli, TurKeyX: Turkish Keyphrase Extractor, in: Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS 2008), Istanbul, Turkey, 2008.

[8] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. Information Retrieval, 1(1–2), 35–67.

[9] Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. IEEE-Computer, 33(11), 29–36.

[10] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research andDevelopment, 2(2), 159–165.

[11] R. Barzilay, M. Elhadad, “Using Lexical Chains for Text Summarization”, In Mani, I., Maybury, M.T., eds.: Advances in Automatic Text Summarization. The MIT Press (1999) 111–121

[12] M. Brunn, Y. Chali, C.J. Pinchak, “Text summarization using lexical chains”, In: Proceedings of the Document Understanding Conference (DUC01), New Orleans, LA (2001)

[13] W.P. Doran, N. Stokes, J. Carthy, J. Dunnion, “Assessing the impact of lexical chain scoring methods and sentence extraction schemes on summarization”, In: Proceedings of the 5th International Conferences on Intelligent Text Processing and Computational Linguistics (CICLing-2004), (2004) 627–635

[14] G. Ercan, and I. Cicekli, “Lexical Cohesion Based Topic Modeling for Summarization”, Lecture Notes in Computer Science 4919, Springer Verlag, (2008), pp: 582-592.

[15] Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of HLT-NAACL-2003, Edmenton, Canada (2003)