Generic Text Summarization for Turkish

Mücahid Kutlu[1], Celal Cığır[1] and Ilyas Cicekli[1*]

[1] *Department of Computer Engineering, Bilkent University, Ankara, Turkey*
[*] *Corresponding author: ilyas@cs.bilkent.edu.tr*

In this paper, we propose a generic text summarization method that generates summaries of Turkish texts by ranking sentences according to their scores. Sentence scores are calculated using their surface level features, and summaries are created by extracting the highest ranked sentences from the original documents. In order to extract sentences which form a summary with an extensive coverage of the main content of the text and less redundancy, we use features such as term frequency, key phrase, centrality, title similarity and sentence position. The 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 of two newly created Turkish data sets. This paper presents one of the first Turkish summarization systems, and its results are promising. We introduce the usage of key phrase as a surface level feature in text summarization, and we show the effectiveness of the centrality feature in text summarization. The effectiveness of the features in Turkish text summarization is also analyzed in detail.

Keywords: Text Summarization, Summary Extraction, Natural Language Processing

1. 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 as a result of which people are overwhelmed to identify which page corresponds to their needs. Therefore, there is an increasing need for new technologies that help users to access the desired and relevant information quickly. Presenting the documents with their summaries will smooth the process of finding the desired documents. Text search and summarization are the necessary technologies to reduce the access time for information. Text search engines filter the pages according to the user query and generate an initial set of relevant documents, and text summarizers generate the summaries of documents that enable users quickly identify the contents of documents in order to determine the final set of relevant documents [1].

A document generally consists of several topics. Some topics are explained deeply by many sentences, and therefore they 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.

An automatic text summarization system takes a document (or documents) as input and presents a well-formed summary by extracting the essence of the document(s) [2]. 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 [3]. We cannot say that one of the approaches is 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 text summarization system presented in this paper uses the extraction method.

Key phrases are considered as condensed versions of documents and short forms of their summaries. Therefore, it is possible to think of them as a set of phrases semantically covering most of the text. In spite of the obvious relations between key phrases and summaries, the usage of key phrases in text summarization has not been investigated. In our text summarization system, we use key phrases as sentence features that contribute to the scores of sentences, and we show that key phrase feature is one of the effective features.

In this paper, we propose a generic text summarization method that creates summaries of Turkish texts by ranking and extracting valuable sentences from the original documents. This method uses surface level sentence features such as term frequency, key phrase and centrality. In addition, the position of sentences in the original document and existence of title keywords in sentences are some of the 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 the main content of the document. A score function of the mentioned features is used to rank the sentences, and machine learning techniques are used to determine the optimal combination of coefficients of these features. Performance evaluation is conducted by comparing summarization outputs with manual summaries generated by human evaluators. ROUGE evaluation technique [4] is used to compare summarization outputs with human generated summaries in addition to the usage of an intrinsic evaluation technique with one of our data sets. The effectiveness of each feature in text summarization is also analyzed in detail.

The contribution of our study is the construction of a single document summarization system for Turkish texts by observing the effects of sentence features 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. The effects of different feature combinations in Turkish text summarization are evaluated in order to determine the effective features in Turkish text summarization. Moreover, summarization studies on Turkish texts are not sufficient, and there is no corpus for Turkish summarization systems. In this study, we used two data sets in order to test the performance of our summarization system. The first data set is a collection of 120 newspaper articles, and their summaries are created by human evaluators. The human evaluators picked important sentences from those newspaper articles in order to create their summaries. The second data set is a collection of 100 Turkish journal articles, and their summaries are created by the authors of those articles. With these two data sets, this study contributes the researchers who want to study text summarization on Turkish texts.

Although the most of the words of a document contributes the meaning of the document, nouns contribute more than any other word class. For this reason, our features mainly depend on nouns appearing in documents. Furthermore, Turkish is an agglutinative language, and a word can have many different forms. In order to treat the different forms of a word as a same word, we use a Turkish stemmer to determine the root words.

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

2. RELATED WORK

Text summarization has been studied since 1950s [5] 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 focused on summary extraction, which selects the pieces such as keywords, sentences or even paragraphs from the source in order 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." [6]. Hovy & Lin distinguished summaries as indicative vs. informative; generic vs. query-based; single-document vs. multi-document [7]. Our proposed summarization system can be categorized as a generic single-document summarization system that uses sentence extraction.

Text summarization methods can also be categorized as supervised and unsupervised methods. The supervised methods require the data sets containing the documents and their human-generated summaries. They learn their summarization models from these data sets, and they use these models in the summarization of other documents. The unsupervised methods do not require any training data [1]. Our generic text summarization system can be

categorized as a supervised method since it learns the feature weights from our data sets.

Lin studied a selection function for extraction and used a machine learning algorithm to automatically learn good features coming from several heuristics [8]. On the other hand, Yeh & Ke used Latent Semantic Analysis (LSA) approach for extraction and compared the results with the feature extraction algorithm [2]. Gong and Liu also worked on summarization by using relevance measure and LSA [1]. Mihalcea proposed an unsupervised method named TextRank [9] which is a graph based ranking algorithm, and the sentences that are recommended by other sentences are selected into a summary [10]. Erkan and Radev also uses a graph based ranking algorithm in their summarization system [11].

Barzilay and Elhadad [12] describe a summarization system based on lexical chains of words. A lexical chain for a set of words is created if those words are semantically related. Therefore semantic relations among the words play a role in sentence extraction. Brunn et al. [13] and Doran et al. [14] also use semantic relations among the words in their summarization systems. In order to improve the sentence selection, Ercan and Cicekli [15] use the clusters of lexical chains instead of lexical chains alone. These text summarization methods use semantic features in text summarization.

Some summarization systems generate a single summary of multiple documents on the same subject. This is known as multi-document summarization [16]. Since the same sentence may appear in a slightly different form in other documents, the multi-document summarizers should eliminate the other forms of the sentence in order to get a concise summary.

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

3. GENERIC TEXT SUMMARIZATION SYSTEM

In this section, we propose a method that creates generic summaries by selecting valuable sentences with the help of a score function. This score function takes into account several kinds of document features, including term frequency, key phrase, and sentence positions in the original text, centrality and existence of title keywords in the sentences in order to generate summaries. First of all, a document is decomposed into individual sentences for further score computation. Later on, sentences are ranked to emphasize the significance of different sentences.

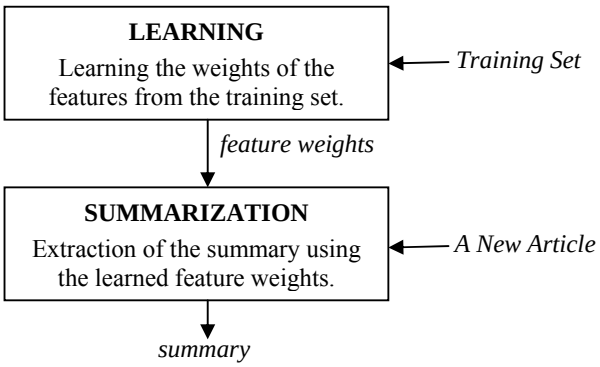

Figure 1. Structure of the Text Summarization System

Finally, the top scored sentences are selected in the order they appear in the original document in order to generate a well-formed summary. The weights of the features are calculated with the help of a training set. Figure 1 gives the main structure of our generic text summarization system. The details of features and summary generation are explained in following subsections.

3.1. Features and feature selection

In order to find the score of a sentence S, indicating the degree whether it belongs to the summary or not, five features are used in the score calculation. These features depend on the surface level clues of sentences. Before the score of a sentence is computed, these feature values must be computed for the sentence.

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.

In our case, only nouns are considered as terms. We use the stems of nouns, and two different nouns are treated as the same noun if their stems are equal. For example, since the Turkish nouns *masada* (*table* in the locative form) and *masanın* (*table* in the genitive form) have the same stem *masa* (*table*), they are treated as same. In order to find the nouns in a sentence and their stems, Zemberek, which is an open source ongoing study on Turkish Language, is used in this paper [17].

After the term frequencies of all nouns in the sentence are found, the term frequency score of a sentence S is computed as follows:

$$Score_{f1}(S) = \frac{\sum_{i=1}^{m} tf_i}{m}$$

where *m* is the number of nouns in the sentence S. 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 the number of nouns in the sentence.

The important sentences of a document contain high frequency nouns. So, this feature increases the likelihood of a sentence containing high frequency nouns to be included in the summary. Of course, the effect of this feature is learned from the training data.

f_2 - Title Similarity (TS): Titles contain groups of words that give important clues about the 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 follows:

$$Score_{f2}(S) = \frac{\left| words\ in\ S \ \cap \ words\ in\ the\ title \right|}{\left| words\ in\ S \ \cup \ words\ in\ the\ title \right|}$$

In order to find the title similarity score of a sentence S, the number of words in the intersection of the sentence words and the title words is divided by the number of words in their union. If a title word appears more than once in the sentence, its occurrence is assumed to be 1. In the computation of this feature, we consider all title words (not only nouns) as words. We again use the stems of the words, and the stems of the words are found using Zemberek stemmer.

f_3 - Key Phrases (KP): Since words are the essential elements of a sentence, the more content-covering keywords a sentence has, the more important it is. Key phrases are short noun phrases that capture the main topics discussed in a given document. Using this feature, it is expected to select the sentences which capture the main topics of the document to be included in the summary.

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 [18]. 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. All key phrases of a document can have same weight or they can have different weights. We decided to use the second approach in order to reward the sentences containing the most important key phrases. The key phrase extraction system in [18] returns the key phrases with their values. The higher value means that the phrase is a more important key phrase. The key phrase extraction system returns the phrases with the top values as the key phrases. The values of the top ten key phrases are used in the computation of the key phrase score of a sentence.

For a sentence S, the key phrase score is defined as follows:

$$Score_{f3}(S) = \frac{summation\ of\ the\ values\ of\ keyphrases\ in\ S}{number\ of\ nouns\ in\ S}$$

In order to compute the key phrase score of a sentence, the summation of the values of the key phrases appearing in the sentence is found. Then, the result is divided by the number of nouns in S in order to avoid the bias of the sentence length.

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 give the general information of the document which are suitable to form a summary. The sentences in the middle of the documents are details about the document 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 the begining positions of the document [2]. In fact, a simple baseline summarization system always selects the first sentences of the document in order to form its summary, and the performance of this simple summarization system is quite satisfactory.

In order to formalize the sentence position, we give a position value P_i (P_i is equal to i, and the position of the first sentence is 0) to each sentence. Then, in order to give higher scores to the first sentences, we use the following formula which gives the position score of a sentence S.

$$Score_{f4}(S) = \frac{R - P_i}{R}$$

where R is the total number of sentences in the corresponding document. Thus, the position score of the first sentence is 1, and the position score of the last sentence in the document is 1/R. The position values of the other sentences are between these two values.

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.

In order to find the centrality score of a sentence S, the following formula is used.

$$Score_{f2}(S) = \frac{|\ words\ in\ S\ \cap\ words\ in\ other\ sentences\ |}{|\ words\ in\ S\ \cup\ words\ in\ other\ sentences\ |}$$

The centrality score of a sentence is the division of the number of words in the intersection of the words of S and the words of other sentences with the number of words in their union. Again, we use the stemmed nouns as words in the computation of the value of this feature.

3.2. Summary Generation

In order to extract the summary of a given document, first the sentences of the document are identified. In the determination of the sentences, the simple heuristics are used. For example, one of the heuristics says that a dot marks the end of a sentence unless it cannot be a part of a token. The dot in a real number is a part of that real number.

After the sentences of the document are identified, the score of each sentence is computed. The feature scores of sentences are normalized to have a value between 0 and 1. For a sentence *S*, the following weighted score function is used to combine all the feature scores of the sentence.

$$Score(S) = \sum_{i=1}^{5} w_i * Score_{fi}(S)$$

In this formula, w_i indicates the weight of the feature f_i, and each weight is a real number between 0 and 1. The weight of a feature indicates the contribution of that feature in the computation of the sentence score.

The 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 are used. For this training set, all possible weight combinations between 0 and 1 with increments of 0.01 are experimented. Then, the one that generates highest average recall result for the training set is selected. The weights of features are presented in the evaluation part.

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

4. EVALUATION

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

4.1. Data corpus

We prepared two data sets for the evaluation of our Turkish summarization system. The first data set is a collection of 120 newspaper articles. Independent human annotators, who are senior and graduate students, helped us to construct this corpus. Each annotator selected news articles independently without any restriction on the subject and sources of the news. We obtained articles in the domain of politics, sports, economy, entertainment, etc. They are collected from online Turkish newspapers such as Milliyet News, Hurriyet News, Zaman News and some other 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. Human annotators created summaries by selecting the sentences from newspaper articles. The second column of Table 1 shows statistics of the newspaper data set.

The second data set is a collection of 100 Turkish journal articles. Most of the articles are selected from Turkish humanities journals. We used the abstracts of the articles that are given by the authors as their human-generated summaries. Of course, the sentences in these abstracts are the authors' sentences, and they may not appear in the articles. The third column of Table 1 shows statistics of the journal data set.

TABLE 1. Statistics of the corpus

Property	*Newspaper Data Set*	*Journal Data Set*
Sentences per document	19.83	147.28
Sentences per manual summary	4.86	6.65
Words per document	329.49	2939.22
Words per manual summary	132.64	133.77
Words per document sentence	16.61	19.95
Words per manual summary sentence	27.25	20.11

The number of words in an average manual summary is almost same in both data sets although the number of sentences in them is different. In the newspaper data set, the size of an average summary is 40% of the original newspaper article. On the other hand, the size of an

average summary in the journal data set is 5% of the original article.

The data corpus is one of the main contributions of this study since there is not enough study on Turkish and there is no reference summary data corpus for Turkish. Our corpus is available to the ones who want to study in this field and need data corpus.

4.2. Performance evaluation

We performed a set of evaluations in order to test the performance of our system. One of the evaluations is done to determine the effectiveness of each feature when it is used alone in text summarization. This evaluation is done for both data sets. The second evaluation is done to measure the overall effectiveness of the system when all features are used in the summarization process. Of course, the weights of the features are learned from training sets. Since we used two-fold cross validation, half of each data set is used in the training phase, and the other half is used for testing.

We used two different evaluation methods in our experiments. ROUGE evaluation techniques are used for both of our data sets. Since the sentences of the manual summaries in the newspaper data set are picked from the original text by human annotators, we also used the intrinsic evaluation method.

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, recall, and f-measure to judge the coverage between manual and machine generated summaries. If we assume that T is the set of sentences of a manual summary and S is the set of sentences of a machine generated summary, *precision*, *recall* and *f-measure* can be defined as follows:

$$precision = \frac{|S \cap T|}{|S|}$$

$$recall = \frac{|S \cap T|}{|T|}$$

$$f\text{–}measure = \frac{2 * precision * recall}{precision + recall}$$

Precision is the fraction of the number of correctly selected sentences divided by the number of all sentences in the machine generated summary. *Recall* is the fraction of the number of correctly selected sentences divided by the number of all sentences in the human generated summary. *F-measure*, harmonic mean of precision and recall, provides a method for combining precision and recall scores into a single value.

Our second evaluation methodology is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [4] which is one of the most popular summarization evaluation methodologies. ROUGE calculates the recall of text units using N-grams, LCS (longest common subsequences) and weighted longest common subsequences. All of these metrics are aimed to find the percentage of overlap between system-generated summaries and human-generated summaries. ROUGE-N score is the percentage of overlap calculated using N-grams. ROUGE-L score is calculated using LCS and ROUGE-W score is calculated using weighted LCS. In ROUGE evaluations, we give our results in terms of ROUGE-1, ROUGE-2, ROGUE-L, and ROUGE-W.

4.2.1. Effectiveness of each feature

Our first evaluation is the measurement of the effectiveness of each feature when it is the only feature that is used in the summarization process. This evaluation corresponds to setting the weight of that feature to 1 and the weights of other features to 0. The results of this evaluation indicate which simple generic text summarization system based on a single feature performs better.

In this evaluation, there is no training phase and the results are obtained from all articles in the data sets. We obtained ROUGE scores for both data sets, and we also obtained intrinsic evaluation results for the newspaper data set.

TABLE 2. Intrinsic evaluation results of each feature tested individually for the newspaper data set.

Feature	*precision*	*recall*	*f-measure*
Term Frequency (TF)	0.363	0.284	0.292
Title Similarity (TS)	0.346	0.282	0.286
Key Phrase (KP)	0.313	0.239	0.248
Sentence Position (SP)	0.393	0.298	0.314
Centrality (C)	0.274	0.206	0.214

Table 2 gives intrinsic evaluation results when each feature is tested separately. According to f-measure results in Table 2, sentence position is the best feature when a single feature is used in the summarization process. This result is consistent with the summarization literature, and a simple baseline generic text summarization uses only sentence position feature by selecting the first sentences into the summary. Next best features are term frequency

and title similarity. Although centrality feature is not a good feature according to the results in Table 2, it is a very good feature according to ROUGE evaluation results. The reason for this can be that the sentences in human-generated summaries are also selected from the texts and a sentence in system-generated summary contributes fully if it matches to a sentence in a human-generated summary. On the other hand, its contribution is 0, if they do not match. In other words, there are no credits for partial matches. In ROUGE evaluation, the partial matches also contribute to the score. ROUGE scores indicate that centrality feature helps to select sentences similar to the sentences in human-generated summaries.

TABLE 3. ROUGE evaluation results of each feature tested individually for the newspaper data set.

Feature	*rouge-1*	*rouge-2*	*rouge-L*	*rouge-W*
Term Frequency (TF)	0.368	0.314	0.327	0.109
Title Similarity (TS)	0.410	0.351	0.363	0.124
Key Phrase (KP)	0.369	0.303	0.323	0.107
Sentence Position (SP)	0.515	0.484	0.492	0.177
Centrality (C)	0.564	0.487	0.507	0.164

TABLE 4. ROUGE evaluation results of each feature tested individually for the journal data set.

Feature	*rouge-1*	*rouge-2*	*rouge-L*	*rouge-W*
Term Frequency (TF)	0.151	0.015	0.116	0.037
Title Similarity (TS)	0.209	0.022	0.154	0.047
Key Phrase (KP)	0.209	0.022	0.154	0.047
Sentence Position (SP)	0.222	0.025	0.165	0.051
Centrality (C)	0.329	0.044	0.247	0.073

Table 3 and Table 4 give ROUGE evaluation results of the newspaper data set and the journal data set respectively when each feature is tested separately. According to ROUGE-1, ROUGE-2 and ROUGE-L scores in both data sets, centrality feature is the best feature, and sentence position feature is the second best feature when each feature is used individually. Although ROUGE-W score of sentence position is higher than the score of centrality feature in the newspaper data set, the centrality score is still the best score in the journal data set. Since the differences between ROUGE scores of these two features are much higher in the journal data set, the centrality feature is much effective in the summarization of longer documents. These results are not surprising because humans put the important sentences at the beginning of the documents, and the sentences that have more common parts with other sentences in the documents are also important sentences.

The third best feature is title similarity feature according to ROUGE scores in Table 3 and Table 4. Since the key phrase feature scores catch up the title similarity feature results in the journal data set, the key phrase feature is also an effective feature in longer documents. The results indicate that term frequency feature is the least effective feature when each feature is used separately.

4.2.2. Overall evaluation of the summarization system

In order to evaluate the effectiveness of our summarization system when all features are used together, we tested our system on both data sets. In our evaluations, we used two-fold cross validation method. In two-fold cross validation, the data set is separated into two sets. The system is trained with the first set to learn the optimum weights for the features, and it is tested with the second test set using these optimum weights. Then the system is trained with the second set, and it is tested with the first set in the two-fold cross validation. The averages of the results of these two training-test phases are presented as the evaluation results.

With the newspaper data set, we used both intrinsic evaluation and ROUGE evaluation methodology. We learned the optimum weights for the features during the training phase of the intrinsic evaluation, and we used these weights in both the intrinsic and ROUGE evaluation of the test set. In order to find the optimal weights for the features, all possible combinations in range 0 to 1 were tested with the training set. The weights that maximize *recall* value were selected as optimum values. During the training, the size of the system-generated summaries was set to the average length of the human-generated summaries in the training set.

With the journal data set, we obtained our results using two-fold cross validation again, and we use a similar approach. In this case, we selected the weights that maximize ROUGE-L value for the training set as optimum values. These optimum values were used in the ROUGE evaluation of the test set.

Table 5 gives the optimum weights for the newspaper data set and the journal data set. These values are the average values of the two training phases during two-fold cross validations. The optimum weights for the newspaper data set are obtained during the intrinsic evaluation by maximizing the *recall* value, and the average value for maximum *recall* values is 0.438 for the training sets. The optimum weights for the journal data set are obtained during the ROUGE evaluation by maximizing ROUGE-L

value, and the average value for maximum ROUGE-L values is 0.387 for the training sets. For the newspaper data set, term frequency, title similarity and sentence position are the main contributors. On the other hand centrality feature is the main contributor for the journal data set.

TABLE 5. Optimum feature weights for the data sets.

Feature	*newspaper data set optimum weights*	*journal data set optimum weights*
Term Frequency (TF)	0.330	0.075
Title Similarity (TS)	0.280	0.150
Key Phrase (KP)	0.070	0.050
Sentence Position (SP)	0.250	0.125
Centrality (C)	0.070	0.600

We obtained the intrinsic and ROUGE evaluation results for the newspaper data set using two-fold cross validation, and the results are given in Table 6 and Table 7, respectively. We also obtained ROUGE evaluation results for the journal data set, and they are given in Table 8. The values given in the tables are the averages of the results that are obtained during the evaluation of the test sets using the optimum feature weights obtained during the training phases.

Each table gives the results for six different systems. The first one, *allFeatures*, uses all features in the creation of the summaries using the optimum feature weights. In other words, we use the score function in Section 3.2 with the optimum feature weights in order find the scores of sentences. The other five (*withoutTF*, *withoutTS*, *withoutKP*, *withoutSP* and *withoutC*) use only four features by dropping one feature in the score calculation of the sentences. For example, *withoutTF* uses 0 for the weight of term frequency feature, and the optimum weights for the remaining four features. With these tests, we see whether the effectiveness of the dropped feature can be captured by the remaining four features.

The intrinsic results in Table 6 indicate that the sentence position is an important feature, and the centrality feature is not a very important feature. But ROUGE results in Table 7 and Table 8 indicate that the centrality feature is a very important feature. The reason for this inconsistency between the results is that intrinsic method checks exact match of sentences. If there are two different sentences that have the same (or similar) meaning, and only one of them is in the human-generated summary, we will get a zero point for this mismatch when the other one is chosen by the text summarization system. However, it is obvious that using different sentences that have the same (or similar) meaning will not decrease the quality of the summary so much. In fact, this fact can be observed in ROUGE results in Table 7 and Table 8.

TABLE 6. Intrinsic evaluation results using all features and all quadruple combinations of features for the newspaper data set.

Features	*precision*	*recall*	*f-measure*
allFeatures	0.482	0.418	0.412
withoutTF	0.423	0.360	0.359
withoutTS	0.420	0.355	0.353
withoutKP	0.481	0.419	0.413
withoutSP	0.422	0.357	0.356
withoutC	0.480	0.418	0.411

TABLE 7. ROUGE evaluation results using all features and all quadruple combinations of features for the newspaper data set.

Features	*rouge-1*	*rouge-2*	*rouge-L*	*rouge-W*
allFeatures	0.580	0.550	0.561	0.190
withoutTF	0.547	0.515	0.530	0.181
withoutTS	0.541	0.513	0.524	0.180
withoutKP	0.559	0.534	0.543	0.187
withoutSP	0.492	0.453	0.465	0.159
withoutC	0.523	0.497	0.506	0.176

TABLE 8. ROUGE evaluation results using all features and all quadruple combinations of features for the journal data set.

Features	*rouge-1*	*rouge-2*	*rouge-L*	*rouge-W*
allFeatures	0.506	0.194	0.368	0.112
withoutTF	0.506	0.193	0.369	0.113
withoutTS	0.503	0.186	0.381	0.111
withoutKP	0.504	0.190	0.367	0.112
withoutSP	0.510	0.192	0.371	0.112
withoutC	0.301	0.118	0.229	0.072

According to the results in Table 7, the biggest drops in ROUGE values occur when the sentence position feature is not used in the score calculation. The second biggest drops occur for the centrality feature. This means that other four features cannot capture the effectiveness of the sentence position feature for the newspaper data set.

The same thing is true for the centrality feature. Since the drops for other three features are not significant, we may conclude that the remaining features can capture the effectiveness of the missing feature. The results in Table 8 indicate that the centrality feature is the most important feature for the journal data set. If the centrality feature is not used, ROUGE results decrease significantly.

When we look at the results in Table 7 and Table 8, ROUGE results for the newspaper data set are higher than the results for the journal data set. The reason for this is that the sentences in the human-generated sentences are also selected from the documents, and a correctly selected sentence in a system-generated summary contributes perfectly to ROUGE scores. But there is no chance to have a sentence of a human-generated summary selected into a system-generated summary because that sentence may not appear in the document of that summary.

When we look at the big picture, we can say that the sentence position and centrality features are the most important features to form a good summary. However, ignoring the sentence position does not decrease ROUGE values, and this means that other features are also good enough. Despite the fact that the sentences selected by the users (intrinsic method) may not be selected by the system, the system-selected sentences become similar to the user-selected sentences by using the centrality feature. This explains the reason of having a big difference between the intrinsic results and ROUGE results for the centrality feature. The title similarity feature also acts like the centrality feature giving low results in the intrinsic method and higher values in ROUGE results. The term frequency feature has little effect on summary since the term frequency is also used for finding key phrases. Therefore, the key phrase feature may decrease its effect. Although key phrases have small effects on summary, this can be improved by developing a 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 the generation of summaries.

5. 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 the 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 the training corpus. This scoring function aims to rank sentences and the summary generation is performed by selecting the top-ranked sentences. In order to test the system, we used two data sets and two-fold cross validation methodology. We have obtained 0.561 and 0.368 ROUGE-L scores for the newspaper and journal data sets, respectively when compression rate is the average compression rate for the test sets.

Besides building a summarization system for Turkish articles working with high precision value, our 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 researchers 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 the effect of key phrases by changing the scoring function of the key phrase feature and the number of key phrases while trying different compression rates. We believe that the contribution of the key phrase feature can be as good as the centrality feature.

Improving the existing document features and adding some new features such as cue phrases, conjunctions and answers of 5W1H (Who, Where, Why, When, What and How) are also future work. Some words make the sentences more important than the others which can be seen as cue phrases. For instance, Turkish word “özetle” (as a conclusion) summarizes and concludes the document content and therefore its occurrences in the summary makes the summary more content-bearing. Moreover, news articles include the answers of questions “Who, Where, Why, When, What, How”. As a characteristic of news articles, sentences that answer one of these questions are more important than others. Finding cue phrases, conjunctions and answers of 5W1H will make our study more specific work on Turkish language.

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 the construction of our corpus.

REFERENCES

[1] Gong, Y. and Liu, X. (2001) Generic text summarization using relevance measure and latent semantic analysis. Proceedings of 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'01), New Orleans, LA, USA, 9-12 September, pp. 19-25, ACM, New York, NY, USA.

[2] Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I.H. (2005) Text summarization using a trainable summarizer and latent semantic analysis. Information Processing & Management, 41, 75-95.

[3] Hahn, U. and Mani, I. (2000) The challenges of automatic summarization. Computer, 33, 29–36.

[4] Lin, C.Y. and Hovy, E. (2003) Automatic evaluation of summaries using N-gram co-occurrence statistics. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL-2003), Edmenton, Canada, May 27 - June 1, pp. 71-78, Association for Computational Linguistics, Morristown, NJ, USA.

[5] Luhn, H.P. (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159-165.

[6] Saggion, H. and Lapalme, G. (2002) Generating informative-indicative summaries with SumUM, Computational Linguistics, 28, 497-526.

[7] Hovy, E. and Lin, C.Y. (1999) Automated text summarization in SUMMARIST. In Maybury, M. and Mani, I. (eds), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, USA.

[8] Lin, C.Y. (1999) Training a selection function for extraction. Proceedings of the 8th international conference on information and knowledge management (CIKM'99), Kansas City, Missouri, USA, 2-6 November, pp. 55-62, ACM, New York, NY, USA

[9] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, pp. 404-411, 25-26 July, Association for Computational Linguistics, Morristown, NJ, USA.

[10] Mihalcea, R. (2005) Language independent extractive summarization. Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, Ann Arbor, Michigan, USA, 25-30 June, pp. 49-52, Association for Computational Linguistics, Morristown, NJ, USA.

[11] Erkan, G. and Radev, D.R. (2004) LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479.

[12] Barzilay, R. and Elhadad, M. (1999) Using Lexical Chains for Text Summarization, In Maybury, M. and Mani, I. (eds), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, USA.

[13] Brunn, M., Chali, Y. and Pinchak, C.J. (2001) Text summarization using lexical chains, Proceedings of the Document Understanding Conference (DUC01), New Orleans, LA, USA, 13-14 September, pp. 135-140.

[14] Doran, W.P., Stokes, N., Carthy, J., and Dunnion, J. (2004) Assessing the impact of lexical chain scoring methods and sentence extraction schemes on summarization, Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2004), Seoul, Korea, 15-21 February, pp. 627–635, Lecture Notes in Computer Science 2945, Springer, Berlin.

[15] Ercan, G. and Cicekli, I. (2008) Lexical cohesion based topic modeling for summarization. Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008), Haifa, Israel, 17-23 February, pp. 582-592, Lecture Notes in Computer Science 4919, Springer, Berlin.

[16] Radev, D.R., & McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24 (3), 469-500.

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

[18] Kalaycılar, F. and Cicekli, I. (2008) TurKeyX: Turkish keyphrase extractor. Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS 2008), Istanbul, Turkey.