The Journal of CIEES
Volume 2, Issue 1, July 2022, pp. 16 – 20, ISSN: 2738-7283 (Print), 2738-7291 (Online)
DOI: 10.48149/jciees.2022.2.1.3

An Extensible Model for Turkish Single Document Summarization

Metin Turan[1] and Mesut Pek[2*]
[1]Istanbul Commerce University/Computer Engineering, Istanbul, Turkey
[2]Istanbul Sisli Vocational School/Graphic Design, Istanbul, Turkey

*E-mail of corresponding author: mesut.pek@sisli.edu.tr

Abstract: The development of Automatic Text Summarization systems with the help of a computer is one of the most challenges in two decades. The amount of information currently available and the time problem of those who need information reveal the importance of such summaries. Generally, successful summaries are produced by people. However, it is obvious that human resources will not be enough to summarize the daily flow of information. Today, work is focused on designing and improving systems that automatically extract the abstract from the text. In this study, a summarization system based on a parametric method has been proposed and implemented. It summarizes a single Turkish text. The compression ratio of the summary can be determined by the user. The system also offers title and keyword support to the user. The evaluation reveals an average model however the hardness of the Turkish language considered is promising.

Received 20 May 2022
Accepted 6 July 2022
Published 22 July 2022

Keywords: Document Summarization, Sentence Scoring, Turkish NLP..

1. Introduction

The first study in this area is based on the word frequency in 1958 by Luhn [1]. Then Edmundson [2] in 1969 and the Salton [3] in 1989 found work on this topic. In Edmundson's work, he has proposed that we should consider the other facts in addition to vocabulary frequencies, for example expressions giving direct information on the subject (in conclusion, in summary, showing this article, etc.), the words containing the subject heading and the place where the sentence is held (usually the first sentences are more relevant) is also important .

Text summarization can be thought of as a two-stage job. In the first phase, source documents are analyzed and a set of sentences is extracted. However, the document set may contain repeated or conflicting information. Care should be taken when making summaries. More importantly, the documents may not be in a structure suitable to extract sentences directly. In cases where it is necessary to make a semantic or structural analysis, it is necessary to think of a preliminary process before the first stage. The second stage produces a readable summary of this sentence set.

Summarization is a complex task that requires detailed natural language processing capacity [4]. Current investigations for baseline reduction of the problem focus on extractive-summary [5]. Inferential summarization is a simple subset of the original text. This summary does not guarantee clarity, but provides enough content to make decisions about the text.

Erol [6] has worked on summarizing legal texts with keywords and Çetiner[7] has worked on finding similar cases in civil cases. Aksoy[8] worked to evaluate the answers to the questions asked in his master's thesis and Karaca[9] worked on creating titles for news texts. ** In his study, Uslu[10] also found studies on the classification of news texts. Noyan[11] has worked on understanding the natural language of the text written in his study.

[12] This application was developed as a course project within the course of natural language processing. In the study, a weighting method of characteristics was applied with a similar approach.

The rest of the article is organized as follows: text summarization procedure in section 2, method specifications in section 3, evaluation in section 4, conclusion in section 5, and future work in section 6.

2. Text summarization procedure

According to Lin and Hovy [13], text summarization consists of 3 stages. These are defined semantically as determining the subject (word frequencies, headings, information expressions, etc.), interpretation (detection of related sentences) and generation (selection of the most appropriate sentence from similar sentences, sentence elimination by summation rate).

Current studies on automatic summarization can be grouped under 3 main headings.

2.1. Statistical Methods

It deals with syntactic and statistical properties. Most of these methods are based on the assignment of a weight value to each sentence in the sentence. The weight value is determined by the position of the sentence in the document or the keyword content. The research is mainly focused on the following issues:

- Position of a sentence within the document or paragraph [14].

- The presence of hinted words and expressions are of the "important", "absolutely", "particularly", "ambiguous", "may be", "for example" in unfavourable sense.
- The existence of demonstrative structures, "the purpose of this research", "our research has shown that" [15] .
- The number of semantic relations between a sentence and its neighbors [16, 17].
- The presence of words within the title, sub-heading [14] or the source [18].

2.2. Speech Methods

These methods are based on speech analysis. It aims to determine a set of POS-tags according to the semantic properties of the sentence. These labels are derived from conjunctive words and superficial clues that link sentences, such as subject matter declaration, causality, and description. The significance of a sentence is obtained from the significance of the semantic label [19].

Other studies reveal the rhetorical structure to determine the sentence [20] and sentence associations [21], taking advantage of the statement representation of the original text. For example, in the method proposed by Teufel, theme units are defined as "past", "subject", "related work", "purpose", "solution", "result" and "product".

2.3 Template Extraction Methods

Another approach is to make a detailed semantic analysis of the source text in order to use artificial intelligence techniques, thus creating the meaning of the text with semantic representation. These methods are based on practical knowledge and the use of pre-defined summaries extracted from the original context.

Frequently used text model is vectorial [22]. After each text element is preprocessed - in terms of a sentence text summarization - it is considered to be an N dimensional vector. Some metrics can be used to measure similarity between two text elements. The most commonly used metric is the cosine measurement. The vectors of the x and y texts are subjected to scalar multiplication. If the cosine value is 0, it is completely incompatible. If it is 1, it is completely similar.

The evaluation of the quality of the summary produced is a key point in the summarization work. Classic information fetch precision and recall metrics can be used to evaluate the performance of automated summary generation routines. Precision is the ratio of selected true sentences to the total selected sentences and recall is the ratio of selected true sentences to the total true sentences.

The reference summaries are provided by the human jury and their performances are measured directly against the models. More important than this is the fact that the summaries the same human jury produced in different dates is similar only 55% [23], and the similarity of summaries produced in the jury is 46% [24]. This poses a problem in terms of the consistency of the evaluations.

3. Method specifications

3.1. Method Schema

Figure 1 and Figure 2 show the operation schema and implementation of the method respectively. Criteria weighting (default values are provided, the user can use these values), if any, keywords and summarization percentage are taken from the user. The user can determine which criteria should be used in sentence weighting. For this purpose, there is a check button next to each scoring criteria. Moreover, only one title (main title) is evaluated in the texts. User approval is required to determine this. The first sentence is regarded as the main title. No method has yet been applied to distinguish subheadings within the text. Later, the user selects the text file they want to summarize and starts the summarization. Up to this point, the process is interactive.

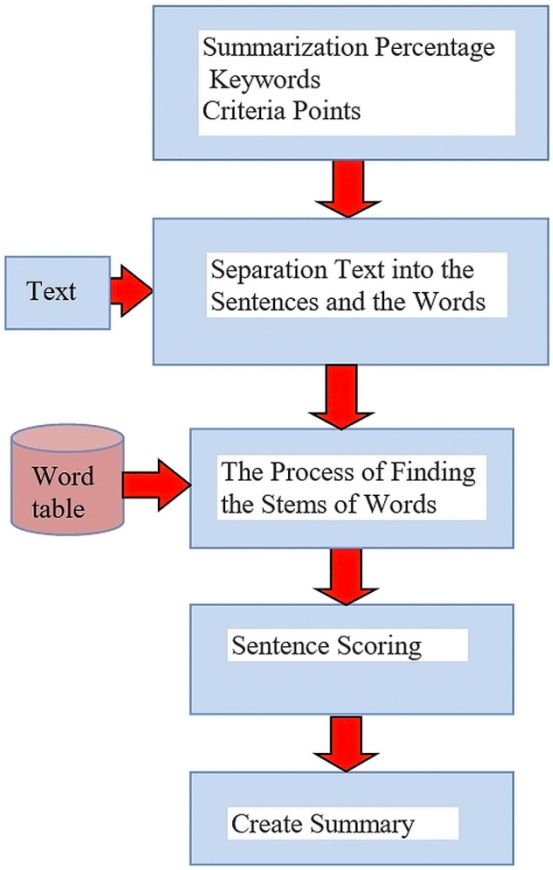

Fig. 1. Method operation schema

Then, the text is divided into sentences (in the direction of punctuation marks of the Turkish Language Society), and the words (tabs, spaces, characters are truncated) are obtained. In the words, the root is found by looking at the word root table (one letter at a time) otherwise the word is treated as root.

At the end, each word and sentence based calculations are made considering the criterion scores. If the user does not want to include some criterion for scoring, he can specify them on the criterion selection screen.

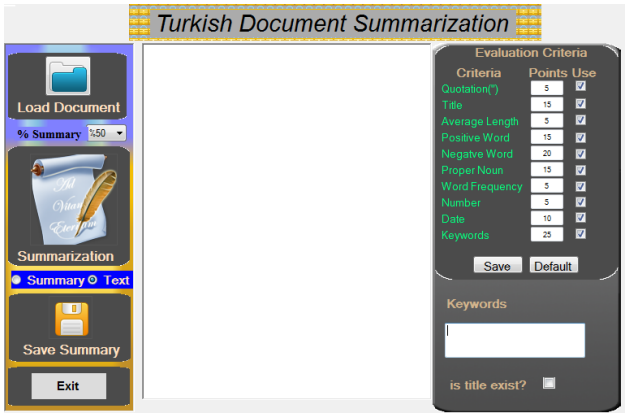

Fig. 2. Model Implementation

Sentences are determined according to the score obtained and the percentage of summarization. In this study, extractive summarization method is used. In other words, there is no semantic analysis module exists.

3.2. Method Criteria

Two types of criterion groups are used as in other studies in the system. The sentence-based ones are as follows:

Quotation Symbol – The sentences contain this symbol are accepted to be important in the texts.

Title - Review whether the phrase contains the words in the title and in the subtitles if any. It is assumed that such sentences are closely related to the subject.

Average Length - The average length of the sentences in the document is calculated as words. Assumptions with an average length of ± 1 are assumed to be significant. This approach is favored as a long-term counterpoint in some studies [25].

The criteria applied on a word-by-word basis are as follows:

Positive Word – The sentences are examining whether they contain such resurgent words as a result, in summary, in conclusion and ultimately.

Negative Word - The sentences containing the words because and however include detailed information on the subject and do not need to be included in the summary. It is assumed that such details contain unnecessary information [26].

Special Names - Special names (starting with capital letters) in the texts about the content, place of the event, person etc. provides important information. So the model should give importance to the specifics of private names.

Word Frequency - The frequency of each word in the text is calculated. The words and, or, with, for, et al. are not included (stop words). The words in the first 10% of the list are taken into account in the sentence rating. It examines whether a sentence includes the highest frequency words.

Numbers - The sentences containing the numbers are presumed to contain important information indicating the relevant size.

Dates - The sentence containing the name of the day or month is examined. Such information is important because it contains historical information.

Keywords - The presence of key words given by the user as a clue is perceived as a necessity for those phrases within the desired summary, and is often given a high score.

3.3. Word Table

A dictionary containing 49,000 stem words in Turkish is used. Every word in the dictionary also contains information about whether it is a name, an adjective, or a verb. These labels are not used in the model at the moment, but as a sub-structure in terms of subsequent linguistic studies.

Once the words in the text are separated, they are searched in the dictionary. If no word is found while searching the dictionary, the search is continued by subtracting one character from the end. If a stem word can't be found in the dictionary, the original word is regarded as the stem. Thus, the root is found by removing the derivational affixes and inflections.

Due to the linguistic nature of Turkish, if a word ends with one of the letters b, c, d, g, these are replaced by letters p, ç, t, k respectively. The root words are stored in accordance with this linguistic standard.

3.4. Parsing

Punctuation marks question mark, point, exclamation and three points are used in order to determine sentence boundary in Turkish. In addition, punctuation marks, single and double quotation, brackets, commas and other symbols (except for numbers and date information) are removed from the text. Words whose initials begin with capital letters are treated as private names.

Paragraph decomposition is not applied because sentence-based extraction is preferred in the model.

3.5. Summary Extraction

The summary is extracted using the summary percentage that the user has specified and the weight value assigned to each sentence. First, the sentences are sorted in decreasing order according to the weight values they have obtained from the criteria. Then, by applying the simple formula in (1), it is decided how many sentences will be in the summary.

$$ES = (TS \cdot SP)/100 \quad (1)$$

Here, ES is the number of sentences in the summary, TS is the total number of sentences in the text, and SP is the summary percentage. The result is rounded up.

4. Evaluation

The values in Table 1 were provided as default values in the method. These values were given completely intuitively.

Due to the characters in the text, it has been observed that some of the texts can't be properly separated. For the experiments, five texts containing different topics and sentences were used. The properties of the texts used in experiments are given in Table 2.

The experiments are summarized at Table 3. Two types of summaries (20%, 50%) are produced for each sample text. P and R used in the Table 3 are abbreviations of Precision and Recall respectively.

Table 1: Criteria default value

Criterion	Weight Value
Quotation Symbol	5
Title	15
Average Length	5
Positive Word	15
Negative Word	20
Special Names	15
Word Frequency	5
Numbers	5
Dates	10
Keywords	25

Table 2. Corpus properties.

Sample	Topic	Title	Sentence Count	Word Count
Text 1	twilight	no	48	516
Text 2	two ships	yes	27	335
Text 3	İstanbul	yes	21	440
Text 4	fashion	no	14	202
Text 5	the system	no	4	67

It can be seen that, whenever the text size gets smaller, the percentage of both recall and precision increases. As the summary percentage decreases, then the percentage of both recall and precision decrease significantly, as a result of increased number of sentences. It is observed that the 50% summarization achieves acceptable success, regardless of the text size.

Table 3. Summary percentage and ımportance of text sıze

Sample	20%		50%	
	P	R	P	R
Text 1	33%	20%	60%	40%
Text 2	40%	13%	48%	48%
Text 3	50%	25%	40%	50%
Text 4	100%	30%	84%	84%
Text 5	100%	50%	100%	100%

The results are obtained with the default parameters. The effects of the parameters on the model will be completed during the remainder of the study.

Furthermore, because of no Turkish text corpus exists for summary evaluation, confirming the validity of the experiments is impossible, however the results obtained are promising for future works.

5. Conclusion

In this study, an experimental application was developed which will form a summary for a single document written in Turkish. This system is designed with parameter weighting method and has the flexibility to expand with different text features. This study provides an experimental platform in order to train the learners who will meet for the first time with the natural language processing (NLP). In this model based on the weighting method, we have observed how important the document size and the summary percentage to be generated are in the summary. The effects of selected features and weight values on the model are not addressed in this study.

This system can be further expanded to create an application infrastructure for many NLP applications such as multi-document summarization, document classification, and so on. Similar studies on this subject have been mentioned in the literature section. But none of these are open source applications. The current application was developed with C # and is intended to be translated into Python and to be used by researchers as open source code in the near future.

6. Future work

A limited number of criteria were applied in the study. The effects of selected features and weight values on the model are not also addressed in this study. Moreover, the new criterion may be suggested and evaluation of the successes of criterion can be modeled easily.

One other issue is to determine the effects of the criteria for different text groups. Moreover, adaptation and modeling of criterion values can be achieved by machine learning algorithms.

At the moment, semantic analysis is not used for the summarization study. Semantic analysis can be applied so that the sentences follow a semantic sequence with each other and the summary forms integrity. As a simple method for this, a metric can be used that takes into account the distance between the sentences in the summary candidates or the sentence alignment.

The study is developed for a single document summarization. For multi-document summarization in the future, a summarization model can be modelled that uses the results of the each single summarization

References

[1] Luhn, H. P. (1958) The automatic creation of literature abstracts, *IBM Journal*, pp: 159-165

[2] Edmundson H. P. (1069) New methods in automatic abstracting", *Journal of the ACM*, **16**(2), pp: 264-285

[3] Salton, G., Allan, J., Buckey, C. & Singhal, A. (1989) Automatic analysis, theme generation and summarization of machine-readable texts, *Science*, pp:1421–1426

[4] Mitra, M., Singhal, A. & Buckley, C. (1997) Automatic Text Summarization by Paragraph Extraction", *In Proceedings of the ACL.97/EACL.97 Workshop on Intelligent Scalable Text Summarization*, Madrid

[5] Sparck-Jones, K. (1999) Automatic summarizing: factors and directions. In Mani, *MIT Press*, pp: 1-12

[6] Gödur, E. (2021) Automatic Summarization with Keyword for Legal Texts. *Rahva Journal of Technical and Social Studies*, 1.1: 24-36

[7] Cetiner, M. & Akgül. Y. S. (2021) Automatic Precedent Legal Document Detection, *Düzce University Journal Of Science & Technology* 9.6: 83-94

[8] Aksoy N. (2021) Automatic Assessment Of Open-Ended Exams In Turkish Language By Natural Language Processing, *Balikesir University Institute of Science Electrical and Electronics Engineering*

[9] Karaca A. (2021) Generating Turkish News Headlines with Deep Learning, *Trakya University Institute of Natural Sciences*, Department of Computer Engineering

[10] Uslu O. & Akyol S. (2021) Turkish News Articles Classification Using Machine Learning Techniques, *Eskişehir*

Turkic World Application and Research Center (ESTUDAM) Journal of Informatics, p:15-20

[11] Noyan, T. , Kuncan, F. , Tekin, R. & Kaya, Y. (2022) A new content-free approach to identification of document language: Angle patterns, *Journal of the Faculty of Engineering and Architecture of Gazi University* , 37.3, 1277-1292

[12] Salih, B. A. L. (2021) Günal, efnan şora. "A new model on automatic text summarization for Turkish", *Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering*, 22.2: 189-198,2021

[13] Hovy E. & Lin, C.Y. (1998) Automated Text Summarization in SUMMARIST, *Annual Meeting of the ACL Proceeding of a Workshop*

[14] Edmundson, H. P. (1969) New methods in automatic extracting, *Newspaper of ACM tea*, pp: 264–85

[15] Paice, C. D. & Al. (1993) The Identification of Important Concepts in Highly Structures Technical Papers, *Proceeding of Sixteenth Annual International* ACM SIGIR Conference, 69–78

[16] Boguraev, B. & Kennedy, C. (1997) Salience-based Content Characterization of Text Documents, *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, ACL/EACLConference Spain, pp: 2–9

[17] Gerard, S., Allan, J. & Singhal, A. (1996) Automatic text decomposition and structuring, *Information Processing & Management*, pp: 127–138

[18] Kan, M. Y., Klavans, J., & McKeown, J. (2002) Using the Annotated Bibliography as a Resource for Indicative Summarization, *Proc LREC* 2002 Spain

[19] Conroy, J. & Leary, D.P.O. (2001) Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition, *Technical Report, Dept.Comp.Sci*. CS-TR-221 Univ. Maryland

[20] Marcu, D. (2001) Discourse-based Summarization in DUC-2000, *Proceedings of the Document Understanding, Conference* DUC'01

[21] Teufel, S. & Moens, M. (1997) Sentence Extraction as a Classification Task, *Proceedings of the Workshop on Intelligent Scalable Summarization* ACL/EACL Conference Spain, 58–65

[22] Salton. G. & Buckley, C. (1997) Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, p:513-523, 1988. Reprinted in: K. Sparck-Jones, P. Willet, Readings in Retrieval, Morgan Kaufmann, p: 323-328

[23] Rath, J.G. , Resnick, A. & Savvage, R. (1961) The formation of abstracts by the selection of sentences, *American Documentation*, p:139-141

[24] Mitra, M., Singhal, A. & Buckley, C. (1997) Automatic Text Summarization by Paragraph Extraction, *In Proceedings of the ACL.97/EACL.97*, Workshop on Intelligent Scalable Text Summarization

[25] Kondratyev, M (2005) *Web Sites Automatic Summarization*, SYRCoDIS

[26] Larocca Neto J., Freitas, A. A. & Kaestner, C. A. A. (2002) Automatic Text Summarization Using Machine Learning Approach, *SBIA*, pp: 205-215