Extractive Based Text Summarization Using Sentence Features

Pelin Kocyigit
Hacettepe University
Ankara, Turkey
n19244371@cs.hacettepe.edu.tr

ABSTRACT

Text summarization is the process of generating a shorter version of the selected text that contains the most important information of the text. The summaries provide both making efficient decisions about documents and reaching required information in less time. There have been various studies for automatically text summarization task. They are mainly divided into two categories: extractive and abstractive based methods. Extractive based summarization focuses on detecting the most meaningful sentences to present the generel context of the text using various features of words and sentences. The first issue about this task is determining the features that can help to dedicate a sentence as valuable to be in the summary. In this study, a new extractive-based summarization approach is proposed based on sentence features to make contribution for selection and ranking process of the sentences.

KEYWORDS

extractive summarization, feature extraction, sentence ranking

1 INTRODUCTION

In the era of rapidly growing information, there is a remarkable demand for reaching a version of information that is reduced in size and conveying meaningful details of its concept. It is very difficult to evaluate each document in terms of conveying required information or creating their summaries manually. This challenge has brought the necessity in automatization of text summaries. A properly created summary is helpful for reader to undertand whether the selected document is a possible source to find demanded information in the fastest possible time. The main goal of this task is deciding which information or sentences should be included in the summary. There have been many studies for automatically text summarization process and can be categorized according to various aspects such as purpose, content type of the input and methodology [12].

Based on the aim of summarization, the generated summary can be Indicative, Informative or Critical. Indicative summaries are to merely show core idea of the document without including its contents [8]. Whereas Informative summaries present important information contained in the document in an objective way. In Critical summaries, an opinion is expressed about the document and so they primarly need analysis and evaluation of the content.

Another categorization of automatic summarization can be done according to content type of the document that are Generic and Query relevant summaries [4]. In Generic summaries, an overall content of the document is extracted and doesn't focus on necessity of users. Query relevant summaries consider questions of users and try to generate appropriate summary which includes answers for corresponding queries.

Acording to methodology, they can be categorized in two main types: extractive and abstractive. An extractive-based method selects relevant sentences of the text and compiles them without any changes. On the contrary, the process is more complicated in abstractive-based method since it requires high-level language processing, understanding meaning of sentences and paraphrasing of them [6]. Although both of them have their own challengings, extractive-based method is less complex thanks to preserving originality of sentences and lack of semantic consideration of documents. However, there are some issues about extractive summaries and one of them is complex structure of the texts that includes long sentences and excessive or inadequate adjectives/adverbs. Also, the major task of extractive summarization is detecting the most important sentences of the given text and features of the words/sentences that are adaptod to operate this task is crucial in terms of performance of the result. From this view, a new approach is proposed for extractive text summarization that focuses on features of sentences more than words for ranking process to facilitate the complexity of the task and try to understand human evaluation while extracting a summary from a given text.

For the rest of this report, similar works done in literature is presented in section 2. Then, methodology of the study is briefly explained in section 3. Section 4 shows experimental results of the proposed approach. Finally, Section 5 concludes the paper.

2 RELATED WORK

Extractive text summarization is done by underlyning most significant parts of the given document and generally consists of three steps that are preprocessing, processing and generation [6]. In preprocessing step, the given text is divided into sentences and then cleaned from all irrelated and unwanted data using methods such as stop-word removing, stemming and case folding. In the result of this step, defined features of all sentences are extracted and ready for the next step. In processing step, the sentences are scored based on the extracted features using a ranking method. Finally, generation step is to select the most important sentences according to their scores and create the summary with the selected sentences.

Ranking of sentences is important for picking most required sentences in the summary and some features have been defined to calculate ranking of sentences [1]. Some of them can be stated as [11]:

- `Keywords`: nouns mostly detected using tf x idf measure.
- `Title`: words contained in title of the given text.
- `Sentence Information`: location and length information of the selected sentence.
- `Sentence Similarity`: similarity between each sentence and centroid sentences.
- `Proper Nouns`: words such as name of a person or place.

- `Upper-case Letters`: words including upper-case letters.
- `Cue Phrases`: words affecting the corresponding sentence positively or negatively.
- `Font`: words typed as bold, italic or underlined.
- `Pronouns`: words such as she or they.
- `Numerical Data`: words including numbers.

These features are significant to highlight required or core sentences of the given text. However, evaluating them and creating a robust algorithm for sentence ranking are another crucial steps of the summarization process. There have been various techniques for extractive summarization. In [7], bag of words model is adapted based on weighted term-frequency and inverse sentence frequency measure. Then, summary is generated according to the highest scored sentences. In [22], term frequency-inverse document frequency (TF-IDF) is used for devoting weigths to words based on each document and the whole corpus. Based on graph theoretic approach, sentences in the given document are considered as nodes of a graph after they are prepared using stop-words removal and stemming methods [14]. In [3], an approach is presented that merges syntactic and semantic features for summarization. [9] presents a method using Latent Semantic Analysis to reach highly scored sentences for both single and multi-document summarization. A Bayesian sentence based model is proposed that considers term-sentence relationships by Wang et al. [20]. In [18], features of the sentences are accepted as input for a fuzzy system and some rules are defined for summarization. The ouput is produced by assigning an importance degree to each sentence and selecting the most important ones among them to display in the summary.

In [2], a graph based summarization tool is presented which is based on term correlations. The graph-based ranking algorithm is adapted to define weights for extracted keywords based on lexical association by Krishna and Reddy [13].

The methods mentioned up to now can be considered as traditional ways of the summarization task. In recent years, deep learning based approaches have been adapted to achive challenges of this task. Verma and Nidhi [19] have presented a deep learning based summarization model by enhancing features of the sentences. Extractive based summarizaztion for single document is provided using auto-encoder by learning features of the sentences and scoring them [21]. In [17], a feedforward neural network model is utilized for single document summarization without considering any linguistic information. In [5], both neural networks and genetic algorithms are adapted based pre-defined six properties of the sentences.

3 METHODOLOGY

The proposed method requires substeps for preprocessing and processing phases due to the developed approach. Hence, the steps of the summarization task are divided as below:

i Preproccessing-1
ii Sentence Scoring-1
iii Preprocessing-2
iv Sentence Scoring-2
v Generation

3.1 Preprocessing-1

The preposeccing step is mostly for cleaning the sentences by removing punctuations, numerics or returning the words to their root forms. However, in this study some of these removed features are utilized for sentence scoring and so the originality of the sentences are preserved until sentence scoring phase is completed. Therefore, this step is only for splitting the given text into the sentences. Before sentence division, the sentences which are located as a paragraph start is recorded to involve sentence scoring calculations in the next steps.

3.2 Sentence Scoring-1

In this paper, a new approach is proposed based on extractive summaries using single document. The idea behind the approach is focusing on features of the sentences more than frequencies or semantic meaning of the words. To this aim, two types of sentence scoring is calculated and then combination of them is used for selecting the sentences which are suitable to include in the summary.

The purpose of the Sentence Scoring-1 (SS1) is to evalute the sentences in terms of pre-defined features. The first challenging step of the task is defining features of the sentences can be used in ranking process. In other words, it is the core point that understanding which properties make one sentence more valuable than others. The approach for feature selection is developed considering human evaluation for text summarization. When a reader look over a given text, some features can be attention-grabbing at the first glance. They are defined as:

- `Paragraph Starts`: sentences which are the first one in a paragraph.
- `Punctuations`: which are used for explanations, giving examples or details and emphasising. These are colons, quotation marks, parantheses, exclamation marks, question marks and backslash signs.
- `Capital Letter`: words containing upper-case letters except the first word of the selected sentence.
- `Linking Words`: words which are used for explanation, conclusion or linking two consecutive sentences such as "although", "that is", "namely", "likewise", "finally" and "in addition".
- `Bonus Words`: words which can convey meaning of comparison, giving example or counting such as "most", "quarter", "such as", "etc", "giant" and "third"
- `Numeric Values`: words containing numbers.

How many times these features are contained in a given sentence is not counted since if one of them is included once time, it is enough to highlight the selected sentence in terms of corresponding feature. Thus, if a sentence including one of the features, its value is assigned to 1.

As an example, the original sentence is:

"The minimum for each purchase is £100 and Bonds are sold in multiples of £10 (more than last year)" .

Extracted features of the sample sentence are colored as red:

"The minimum for each purchase is £100 and Bonds are sold in multiples of £10 (more than last year)".

According to the sample sentence, it contains bonus word ("minimum"), numeric value, capital letter and puntuation. Since reoccurence of them is not important, first mathcing feature is marked and others are ignored such as a bonus word "last" or numeric value "£10".

Another point is calculating SS1 based on the score gained from these features. If the sentence includes more than one feature, then SS1 could be sum of them. However, if the text is about for instance mathematics, then containing a numeric value can not make sense as expected since most of the sentences may contain it. Thus, a weight for each feature is calculated by considering total count of the selected feature and entire features. And the equation is as follows.

$$w_i = \frac{total_i}{total_f} \tag{1}$$

where, i is the selected feature and f belongs to the entire features.

Then SS1 is calculated for a sentence as sum of all weighted features which are hold by the selected sentence. If it doesn't have one of the feature, its value is assigned to 0 for the correspondig feature. SS1 score is given by:

$$w_I = w_s + w_p + w_c + w_l + w_b + w_n \tag{2}$$

Here, I is assigned to SS1, s to paragraph starts, p to punctuations, c to capital letters, l to lingking words, b to bonus words and n to numeric values.

3.3 Preprocessing-2

After SS1 value is assigned to each sentence, all sentences of the text are ready for traditional preprocessing step. They are firstly converted to lower-case form. Then, all punctuations and numerical values are removed from the sentences. The words which can be considered as not valuable for the sentences are removed by using a pre-defined stop-word list. Finally, the remaining words of the sentences are stemmed to their base form using Porter stemming algorithm [16]. In the result, the text is ready for the next calculation which deals with word weights.

3.4 Sentence Scoring-2

When the first basic features are extracted from the sentences, it is observed that a sentence that doesn't have any of these features may also be required for summary due to the contained words. Thus, a scroring to evaluate weights of words for each sentence is developed that is Sentence Scoring-2 (SS2). The purpose of SS2 is to understand what the contribution is having a selected word for sentences. Repetition of words in the same sentence is not considered because having it once time is enoguh for the corresponding sentece to involve its contribution. From this view, the features needed for SS2 are stated as:

- `Sentence Length`: count of the remaining words in the selected sentence after Preprocessing-2 step is operated.
- `Word-Sentence Count`: count of the sentences which contain the selected word.
- `Total-Sentence Count`: sum of the word-sentence count values of all words.

It is the equation for weights of words as follows.

$$w_w = \frac{c1}{c2} \tag{3}$$

where, w is the selected word, $c1$ is word-sentence count and $c2$ is total-sentence count. By this way, all words are assigned to a weight which states its importance to the corresponding sentence. Then, SS2 is calculated using the equation below:

$$w_{II} = \sum_{i=1}^{n} \frac{w_i}{n} \tag{4}$$

where, n refers to sentence length and i stands for weight of the selected word.

After SS1 and SS2 values are defined to each sentence, it is observed that a sentence which has the highest SS1 can has the lowest SS2 at the same time. It is needed to find the optimum scoring for sentences. Therefore, a final scoring for each sentence is computed as follows.

$$w_s = w_I * w_{II} \tag{5}$$

Here, I shows SS1 and II refers to SS2 scores.

Figure 1 presents scoring results of a sample text to show differences in ranking of the sentences when they are ordered based on each SS1, SS2 and final scoring. It can be understood that SS1 and SS2 complete each other. For example, the sentence whose id 12 is expected to be in the summary. While this sentence isn't even discovered by SS1, it gets the second highest order given by SS2. When final scoring is computed, this sentence is involved in the summary. It is needed to be stated here that the reason of computing structural features of the sentences (SS1) and weight of words (SS2) as seperately is to provide a balanced importance between them. In other words, each scoring result can be in different range according to some factors such as length of the text or context of the document. If one of them is involved to other one, its effect could be minor due to their numerical range. Hence, final scoring is computed to devote a balanced ranking to each sentence.

3.5 Generation

This final step is to generate the summary based on the results of sentence scoring approach. According to the final scores of the sentences, they are ranked in descending order. Sentences whose scores are highest are picked up to include in the summary. And their original positions in the given text are preserved to display the summary. An example of original text, reference summary and proposed summary is shown in Figure 2.

4 EXPERIMENTAL RESULT

This section is to describe aboout the results of the proposed approach and evaluation of them. During the experiments a dataset which contains news articles of BBC and their pair summaries are used [10] to make a comparison with the generated summaries. The context of the selected articles are about business, entertainment, politics, sport and technology.

To evaluate performance of the proposed approach, the ROGUE (Recall-Oriented Understudy for Gisting Evaluation) evaluation metrics are adapted [15]. It is to measue quality of machine-generated

sentenceId	sentenceWeight1	sentenceWeight2	totalWeight
0	0.931	0.014	0.013034
7	0.93	0.014	0.01302
11	0.803	0.008	0.006424
10	0.73	0.013	0.00949
2	0.729	0.017	0.012393
6	0.729	0.013	0.009477

(a) Scoring results in descending order based on SS1

sentenceId	sentenceWeight1	sentenceWeight2	totalWeight
2	0.729	0.017	0.012393
12	0.565	0.015	0.008475
0	0.931	0.014	0.013034
7	0.93	0.014	0.01302
3	0.528	0.013	0.006864
6	0.729	0.013	0.009477

(b) Scoring results in descending order based on SS2

sentenceId	sentenceWeight1	sentenceWeight2	totalWeight
0	0.931	0.014	0.013034
7	0.93	0.014	0.01302
2	0.729	0.017	0.012393
10	0.73	0.013	0.00949
6	0.729	0.013	0.009477
12	0.565	0.015	0.008475

(c) Scoring results in descending order based on final scoring

Figure 1: An example of scoring results

summaries by comparing them with reference summaries which are mostly created by humans. Precision is to measure how much of the generated summary is actually required. Recall shows how much of the generated summary and reference summary overlapping. Finally, f-measure is to measure accuray of the output by using the results computed by Precision and Recall metrics. The Table 1 shows total word count contained in the reference summary and evaluation results of ten randomly selected articles based on Precision, Recall and F-measure metrics.

It is observed that the proposed approach always tend to select shorter sentences. If the text is comprised of long sentences, it can miss few sentences which are in the reference summary. However, it is good at extracting short but valuable sentences. The reason is that SS1 takes into account structural features of the sentence and increases the importance of the sentence even if it doesn't have a high score based on word weights. Besides, when the sentences are ordered only based on SS1 scores, the result is quite closed to the reference summary.

Also, the proposed method doesn't benefit from keywords, titles or similarity of sentences. Since the main purpose is simplifying the summarization task, it could present summaries without deeply considering about words. Besides, some words which are generally

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

(a) The original text

TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn.Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. Time Warner's fourth quarter profits were slightly better than analysts' expectations.

(b) Reference summary

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year earlier. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Time Warner's fourth quarter profits were slightly better than analysts' expectations. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. For the full year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

(c) Proposed summary

Figure 2: An example output of the proposed method

Table 1: The evaluation results of the proposed summaries

Id	Word Count	Precision (%)	Recall (%)	F-measure (%)
1	163	82.14	70.55	75.9
2	163	94.57	84.47	89.2
3	189	71.98	69.31	70.6
4	217	80.21	69.12	74.3
5	173	77.63	68.20	72.6
6	238	84.54	73.53	78.7
7	274	75.10	67.15	70.9
8	176	87.88	82.39	85.0
9	212	80.66	68.13	73.9
10	181	87.91	88.39	88.2

accepted as a stop-word are involved to bonus-word or linking-word lists in this study. Some of them are "firstly", "second", "three", "most", "suh as", "half", "however", "but" and "likewise".

5 CONCLUSION

Extractive text summarization is an important tool to have overall idea about a document. It consists of mostly threes steps which are extracting features, scoring sentences and generating the summary. The result of each step affects accuracy of the next step. Thus, the first step which is feature extraction is significant for generating efficient summaries and to facilitate the complexity of the task. This study is to evaluate structural features of the sentences instead of only focusing on word weights or semantic consideration of the texts.

The proposed approach provides two types of sentence scoring and the final score of the sentences is calculated as combination of these scoring types. BBC news article dataset is adapted to experiment the proposed method. It is observed that this method is able to detect short important sentences whose word weights are quite low. However, it can omit long sentences which are expected to be in the output. It is needed to improve performance of the method for exact matching with the reference summaries and the start point could be considering about SS2 in terms of sentence lengths. Also, it could be interesting to apply the proposed approach for summarizaiton of multiple documents.

REFERENCES

[1] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys J. Kochut. 2017. Text Summarization Techniques: A Brief Survey. *ArXiv* abs/1707.02268 (2017).

[2] Elena Baralis, Luca Cagliero, Naeem A. Mahoto, and Alessandro Fiori. 2013. GraphSum: Discovering correlations among multiple terms for graph-based summarization. *Inf. Sci.* 249 (2013), 96–109.

[3] Araly Barrera and Rakesh Verma. 2012. Combining Syntax and Semantics for Automatic Extractive Single-Document Summarization. 366–377.

[4] Yllias Chali. 2002. Generic and Query-Based Text Summarization Using Lexical Cohesion, Vol. 2338. 293–302.

[5] Niladri Sekhar Chatterjee, Ashna Mittal, and S. Goyal. 2012. Single document extractive text summarization using Genetic Algorithms. *2012 Third International Conference on Emerging Applications of Information Technology* (2012), 19–23.

[6] Ahmed El-Refaiy, A.R. Abas, and I. Elhenawy. 2018. Review of recent techniques for extractive text summarization. *Journal of Theoretical and Applied Information Technology* 96 (12 2018), 7739–7759.

[7] René Arnulfo García-Hernández and Yulia Ledeneva. 2009. Word Sequence Models for Single Text Summarization. *2009 Second International Conferences on Advances in Computer-Human Interactions* (2009), 44–48.

[8] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, and Bahareh Gholamzadeh. 2009. A Comprehensive Survey on Text Summarization Systems. *2009 2nd International Conference on Computer Science and its Applications* (2009), 1–6.

[9] Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 19–25. https://doi.org/10.1145/383952.383955

[10] Derek Greene and Pádraig Cunningham. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, 377–384.

[11] Vishal Gupta and Gurpreet Singh Lehal. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence* 2 (2010), 258–268.

[12] U. Hahn and I. Mani. 2000. The challenges of automatic summarization. *Computer* 33, 11 (2000), 29–36.

[13] R. V. V. Murali Krishna and Ch. Satyananda Reddy. 2016. Extractive Text Summarization Using Lexical Association and Graph Based Text Analysis.

[14] Canasai Kruengkrai and Chuleerat Jaruskulchai. 2003. Generic text summarization using local and global properties of sentences. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)* (2003), 201–206.

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*, 10.

[16] Martin Porter. 2001. Snowball: A language for stemming algorithms.

[17] Aakash Sinha, Abhishek Yadav, and Akshay Gahlot. 2018. Extractive Text Summarization using Neural Networks. *ArXiv* abs/1802.10137 (2018).

[18] Ladda Suanmali, Mohammed Salem Binwahlan, and Naomie Salim. 2009. Sentence Features Fusion for Text Summarization Using Fuzzy Logic. *2009 Ninth International Conference on Hybrid Intelligent Systems* 1 (2009), 142–146.

[19] Sukriti Verma and Vagisha Nidhi. 2017. Extractive Summarization using Deep Learning. *ArXiv* abs/1708.04439 (2017).

[20] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-Document Summarization using Sentence-based Topic Models. *ACL-IJCNLP*, 297–300. https://doi.org/10.3115/1667583.1667675

[21] Mahmood Yousefi-Azar and Len Hamey. 2017. Text summarization using unsupervised deep learning. *Expert Syst. Appl.* 68 (2017), 93–105.

[22] Yongzheng Zhang, A. Nur Zincir-Heywood, and Evangelos E. Milios. 2005. Narrative text classification for automatic key phrase extraction in web document corpora. In *WIDM '05*.