Text Summerization Techniques

Anjana Tiha
Department of Computer Science
University of Memphis, Memphis, TN
Email: atiha@memphis.edu

Abstract—This paper is focused with text summarization and currently popular text summarization techniques. This paper surveys on recent research on text summarization.

Index Terms—Text Summerization, Machine Learning, Trending Text Summarization Techniques, Deep Learning.

I. Introduction

Automatic text summarization is the process of generating concise and condensed, representation from one or more text documents data that fluently captures the core meaning and concept of the original text. There has been exponential growth of the world-wide web and social media usage causing dramatical increase in speed and the scaling of information dissemination. With such vast amount of accessible text documents on the Internet, it has become imperative to use text summarization in order to save time and effort in finding the right information. Summarization can be used for text, document, image and video. In image summarization the summarization system finds the most representative and important images. For videos, system tries to extract the important events from the long-time frame and uneventful context. Text summarization is a sub field of machine learning and data mining. There are mainly two approaches for generating automatic text or document summarization. Most common one is using extractive style, which extracts sentences essential to preserve the core idea of the text and combines them together to generate comprehensive text summery. This technique often use machine learning to score word, sentence or paragraph. Abstractive approach focuses more on the semantic meaning of the text and generates summery using natural language processing techniques.

II. Preprocessing

For extractive summarization, preprocessing phase creates a representation of the original document. Usually, it identifies text boundaries and splits the text into paragraphs, sentences, and tokens. Sometimes, preprocessing requires special character, lower case and stop word removal along with stemming.

III. Extraction-based summarization

After creating the intermediate representation of the original text using sentence boundary identification, special character, lower case and stop words removal and stemming, summarization system scores objects like words, sentences, paragraphs based on selected features. Sentences with highest scores are selected for summary. Text summarization can be reduced to 3 steps:

1) Create an intermediate representation of the original text.
2) Sentence scoring.
3) Selecting high scoring sentences for generating summary.

A. Word Based Scoring

There are different techniques for generating summery using word based scoring. In word based scoring, each sentence in a document is scored based on some selected features. After scoring based on features, cumulative score for each sentence is calculated for ranking. Highest scored sentences are selected to be in summary. Some of the common features for word based scoring includes word frequency, TF-IDF, upper case letter, proper noun and numerical data inclusive sentence.

1) Word Frequency: Assumes that words that occur frequently through the document are most important (Lloret & Palomar, 2009; Gupta , 2011) and contain key information necessary for developing summarization. Each word of the document is scored based on weight derived from calculating frequency of it in the full text. Then cumulative score for each sentence is generated for all sentences present in the document. Finally, the sentences with higher scores are selected for document summary.

$$TF/IDF = DN \times log(\frac{(1+tf)}{log(df)})$$

2) TF/IDF: Text is first preprocessed by removing, hyperlinks, special characters, digits and stop words. Then words are stemmed to reduce to root words. Finally, TF-IDF is calculated for each sentence. Here, each sentence is treated as a document for term frequency calculation. Document frequency is calculated for each document. Sentences with highest cumulative scores are selected for summarization.

3) Lexical Similarity: Scores sentences based on idea that important sentences have strong chains words or lexical similarity (Gupta 2011; Barrera & Verma).

4) Upper Case: Sentences with more importance are assumed to have more uppercase (Prasad et al., 2012) letters containing key information with proper which can be name, initials, highlighted words, proper nouns. The sentences with two or more upper case letters are assigned higher score and summarization is generated based on the highest ranked sentences.

$$CPTW = \frac{NCW(j)}{NPW(j)}$$

where:
CPTW = Ratio of total first letter capital words present in the sentence to the total number of words present in the sentence.
NCW = Number of first letter capital words.
NTW = Total number of words present in sentence.

$$UCf = \frac{CPTW(j)}{MAX(CPT(j))}$$

where, UCf = Uppercase feature value.

5) Proper Noun: Hypotheses on the concept that sentences with higher number of proper nouns are more important and hence, assigned higher scores (Fattah & Ren, 2009).

B. Sentence-based Scoring

This approach is based on the features of the sentence itself.

1) Cue-phrases: Several cue phrases can identify most informative sentences in a document. For example, the sentences started with "In summary", "In conclusion" contain the concise summery. Also, emphasis is given by phrases such as, "Most importantly", "Most significantly", "Particularly". Phrases like "according to the study", "in our finding" contains factual data. Also, domain-specific phrases or terms can be good indicators of significant content of a text (Gupta et al., 2011).

$$CP = \frac{CPS}{CPD}$$

where,
CP = Cue-phrase score.
CPS = Number of cue-phrases in a sentence.
CPD = Total number of cue-phrases in a single document.

2) Sentence Position: The position of a sentence can be crucial indicator of its relative importance. For example, the most important sentences tend to come at the beginning or end of a document or near the title text. There are couple of approaches for position based sentence scoring. One is giving weight 1 to the first N sentences (Satoshi et al., 2001) and the rest of the sentences are given 0 weight. In paper (Barrera & Verma, 2012), sentences have been scored based on three positions, sentences closer to the title, sentences at the beginning of the paragraph and sentences at the end of paragraph. Sentences closer to these 3 positions were given more weight and hence more likely to be selected for summary. Also, for domain specific document summarization, domain knowledge can be used for automated summary.

3) Sentence Resemblance To The Title: Sentence resemblance to the title is the vocabulary overlap between this sentence and the document title (Satoshi et al., 2001). Words in the title is considered most important and contains key terms that convey the central concept of the document. Sentence with high resemblance with the title is have an overlapping vocabulary with title. The concept centrality embodied in title vocabulary and thus can help retain important sentences.

$$Score = Ntw/T$$

where,
Ntw = Number of title words in sentence.
T = Number of words in the title.

4) Sentence Length: This scoring method penalizes sentences that are or too long or too short, as they are not considered as an ideal selection. Length is counted by number of words in a sentence.

$$Score = Length(s) \times (AverageSentenceLength)$$

5) Sentence Inclusion of Numerical Data: Sentences containing numerical data is considered more important and containing crucial information for informative summary. Hence, sentences with numerical data is likely to be included (Abuobieda et al., 2012;) in the text summary.

C. Graph-based Scoring:

In graph-based methods the score is generated by the analyzing relationship among sentences in a document. When a sentence refers to another it generates a link with an associated weight between them. The weights are used to generate the scores of a sentence.

Text is preprocessed by removing digits, special characters, stop words and converting them to lower case. After initial preprocessing, stemming is done to reduce to root words. In the graph based approach, all the sentences are considered as a node in a graph. Based on shared or similar information like overlapping vocabulary, two nodes are connected. The nodes or sentences with higher cardinality is considered more important for summarization.

Some graph based techniques are:

1) Text Rank: It extracts the important keywords from a text document and determines the weight of the "important" of words within the entire document by using a graph-based model. Sentences with higher rank are selected.

2) Cluster Based Method: Text documents can have an organizational pattern and key information can be extracted by identifying locations/sections of important segment of the document. In this approach, each document is divided into multiple segments or clusters and objects (word, sentence) in each segment are given weights based on term importance (TF-IDF or other features). Next, each segment is ranked by calculating the overall importance score for all the objects present in each segment. Segments with highest scores are then selected for final summary.

3) Chalanges Of Extractive Summarization: Problem with extractive summarization is that it can lead to lack of coherency. Also, summery especially unit sentences can be too long to understand the concept accurately. Often document can contain counter information. Presenting them in summery can lead to ambiguity and confusion beating the original motivation behind summarization. Also, key information could get ignored due to selection of scoring method.

IV. Dataset

From Ferreira [2] experiment:

1) CNN Dataset: The CNN corpus developed by Lins and colleagues (Lins et al., 2012) 400 texts assigned to 11 categories: Africa, Asia, business, Europe, Latin America, Middle East, US, sports, tech, travel, and world news.

2) Blog summarization dataset: Hu and colleagues (Hu, Sun, & Lim, 2007, 2008) collected data from two blogs, Cosmic Variance (http://cosmicvariance.com) and Internet Explorer Blog (http://blogs.msdn.com/i.e./). From those blogs, 100 posts, 50 from each blog, were randomly chosen to form the evaluation dataset.

3) SUMMAC dataset: The SUMMAC Corpus was elaborated with University of Edinburgh, as part of the SUMMAC conference organizer group.183 papers on Computation and Language, obtained from the repository LANL (Los Alamos National Laboratory) maintained by Cornell University Library.

V. Results

From Ferreira experiment:

Table 2
Algorithms.

alg01	Word frequency
alg02	TF/IDF
alg03	Upper case
alg04	Proper noun
alg05	Word co-occurrence
alg06	Lexical similarity
alg07	Cue-phrase
alg08	Inclusion of numerical data
alg09	Sentence length
alg10	Sentence position 1
alg11	Sentence position 2
alg12	Sentence centrality 1
alg13	Sentence centrality 2
alg14	Resemblance to the title
alg15	Aggregate similarity
alg16	TextRank score
alg17	Bushy path

Table 3
Results of ROUGE having CNN dataset as gold standard applied to the proposed algorithms.

	Average_R	Average_P	Average_F
alg01	0.71(0.19)	0.35(0.13)	0.46(0,15)
alg02	0.73(0.17)	0.35(0.12)	0.46(0.15)
alg03	0.64(0.19)	0.35(0.12)	0.44(0.12)
alg04	0.64(0.20)	0.35(0.13)	0.45(0.15)
alg05	0.59(0.20)	0.33(0.13)	0.42(0.15)
alg06	0.69(0.19)	0.35(0.13)	0.46(0.14)
alg07	0.50(0.22)	0.35(0.13)	0.40(0.14)
alg08	0.56(0.21)	0.36(0.13)	0.43(0.14)
alg09	0.70(0.18)	0.33(0.12)	0.44(0.15)
alg10	0.61(0.22)	0.40(0.13)	0.47(0.15)
alg11	0.52(0.22)	0.36(0.13)	0.41(0.12)
alg12	0.46(0.25)	0.37(0.16)	0.38(0.15)
alg13	0.33(0.21)	0.31(0.13)	0.30(0.15)
alg14	0.67(0.20)	0.36(0.12)	0.46(0.14)
alg15	0.57(0.20)	0.34(0.12)	0.42(0.14)
alg16	0.62(0.20)	0.34(0.12)	0.43(0.14)
alg17	0.56(0.20)	0.35(0.13)	0.42(0.14)

Table 5
Results of ROUGE having blog summarization dataset as the gold standard applied to the proposed algorithms.

	Average_R	Average_P	Average_F
alg01	0.72(0.13)	0.63(0.15)	0.67(0.14)
alg02	0.75(0.11)	0.63(0.15)	0.68(0.13)
alg03	0.58(0.16)	0.61(0.15)	0.59(0.15)
alg04	0.57(0.17)	0.63(0.14)	0.59(0.14)
alg05	0.65(0.14)	0.63(0.14)	0.63(0.13)
alg06	0.71(0.14)	0.63(0.14)	0.66(0.14)
alg07	0.52(0.18)	0.64(0.14)	0.57(0.15)
alg08	0.54(0.18)	0.63(0.15)	0.58(0.16)
alg09	0.76(0.11)	0.62(0.14)	0.68(0.13)
alg10	0.46(0.19)	0.60(0.13)	0.51(0.17)
alg11	0.52(0.18)	0.63(0.14)	0.56(0.16)
alg12	0.50(0.18)	0.65(0.14)	0.56(0.16)
alg13	0.46(0.20)	0.60(0.15)	0.51(0.18)
alg14	0.60(0.18)	0.64(0.13)	0.61(0.16)
alg15	0.58(0.18)	0.62(0.13)	0.59(0.16)
alg16	0.68(0.14)	0.63(0.14)	0.65(0.13)
alg17	0.58(0.17)	0.63(0.13)	0.60(0.15)

Table 7
Results of ROUGE having SUMMAC dataset as gold standard applied to the proposed algorithms.

	Average_R	Average_P	Average_F
Alg01	0.48(0.10)	0.19(0.10)	0.26(0.10)
Alg02	0.47(0.11)	0.19(0.10)	0.26(0.10)
Alg03	0.25(0.11)	0.17(0.07)	0.19(0.06)
Alg04	0.22(0.11)	0.17(0.07)	0.18(0.06)
Alg05	0.23(0.16)	0.16(0.10)	0.17(0.10)
Alg06	0.46(0.11)	0.19(0.10)	0.26(0.10)
Alg07	0.33(0.11)	0.24(0.10)	0.26(0.07)
Alg08	0.25(0.11)	0.20(0.08)	0.21(0.07)
Alg09	0.49(0.09)	0.16(0.10)	0.23(0.09)
Alg10	0.31(0.10)	0.28(0.10)	0.28(0.06)
Alg11	0.24(0.11)	0.24(0.10)	0.22(0.08)
Alg12	0.07(0.11)	0.17(0.17)	0.07(0.08)
Alg13	0.22(0.11)	0.23(0.10)	0.21(0.08)
Alg14	0.36(0.14)	0.28(0.10)	0.29(0.08)
Alg15	0.22(0.08)	0.22(0.07)	0.21(0.05)
Alg16	0.46(0.10)	0.22(0.10)	0.28(0.09)
Alg17	0.23(0.10)	0.22(0.08)	0.21(0.06)

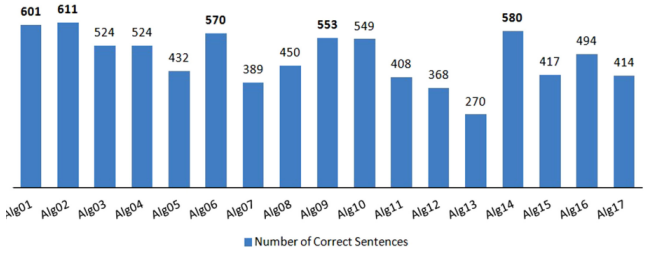

Fig. 1. Number of correct sentences *x* algorithms – using CNN dataset.

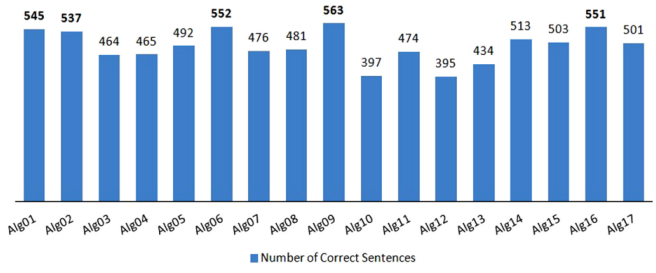

Fig. 2. Number of correct sentences *x* algorithms – using blog summarization dataset.

Table 8
Execution time using SUMMAC dataset.

Alg	Execution time (s)
Alg01	17.287
Alg02	2,499.092
Alg03	9.606
Alg04	7.083
Alg05	73.948
Alg06	549.284
Alg07	29.954
Alg08	8.187
Alg09	9.670
Alg10	7.822
Alg11	7.750
Alg12	107.439
Alg13	2,602.518
Alg14	8.175
Alg15	38.316
Alg16	84.970
Alg17	39.903

Table 4
Execution time using CNN dataset.

Alg	Execution time (s)
alg01	13.986
alg02	196.269
alg03	5.724
alg04	25.723
alg05	20.490
alg06	419.609
alg07	8.133
alg08	4.029
alg09	4.820
alg10	4.122
alg11	4.292
alg12	8.999
alg13	47.267
alg14	5.617
alg15	7.708
alg16	322.045
alg17	8.309

Table 6
Execution time using blog summarization dataset.

Alg	Execution time (s)
Alg01	2.508
Alg02	14.810
Alg03	1.841
Alg04	7.083
Alg05	2.943
Alg06	87.496
Alg07	2.982
Alg08	1.641
Alg09	2.391
Alg10	1.722
Alg11	1.799
Alg12	2.374
Alg13	5.225
Alg14	1.706
Alg15	2.194
Alg16	67.136
Alg17	2.508

VI. Comarative Performance From Paper

Among all the algorithm, word frequency, TF-IDF, sentence length seemed to perform better(Assessing sentence scoring techniques, Rafael).

VII. Abstraction-based summarization

In abstraction based approach, summarization system paraphrases segments of the source document. Abstraction based techniques can generate more condensed text than extraction techniques. Abstractive summaries try to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. But, this requires usage of natural language generation technology, which is still in growth stage. While there has been some recent work in abstractive summarization, the majority of summarization systems are using extractive methods.

VIII. Conclusion

For automatic text summarization, extraction based techniques are mostly used. Some recent work has been done on abstractive text summarization. Semantic based approach has also been experimented. But as it requires better NLP techniques, it is still work in progress. Recurrent neural network has also been applied for summarization recently.

References

[1] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." Journal of emerging technologies in web intelligence 2.3 (2010): 258-268.

[2] Ferreira, Rafael, et al. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40.14 (2013): 5755-5764.

[3] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).

[4] Gupta P, Pendluri VS, Vats I. Summarizing text by ranking text units according to shallow linguistic features. InAdvanced Communication Technology (ICACT), 2011 13th International Conference on 2011 Feb 13 (pp. 1620-1625). IEEE.

[5] Balahur A, Lloret E, Boldrini E, Montoyo A, Palomar M, Martínez-Barco P. Summarizing threads in blogs using opinion polarity. InProceedings of the workshop on events in emerging text types 2009 Sep 17 (pp. 23-31). Association for Computational Linguistics.

[6] Barrera A, Verma R. Combining syntax and semantics for automatic extractive single-document summarization. Computational Linguistics and Intelligent Text Processing. 2012:366-77.

[7] Fattah MA, Ren F. GA, MR, FFNN, PNN and GMM based models for automatic text summarization.

Computer Speech & Language. 2009 Jan 31;23(1):126-44.

[8] Shardan R, Kulkarni U. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization.

[9] Lloret E, Palomar M. Text summarisation in progress: a literature review. Artificial Intelligence Review. 2012 Jan 1;37(1):1-41.