DocEng'19 Competition on Extractive Text Summarization

Rafael Dueire Lins
UFRPE/UFPE, Recife, Brazil
rdl.ufpe@gmail.com

Rafael Ferreira Mello
UFRPE, Recife, Brazil
rafael.mello@ufrpe.br

Steve Simske
Colorado State University, USA
Steve.Simske@colostate.edu

ABSTRACT

The DocEng'19 Competition on Extractive Text Summarization assessed the performance of two new and fourteen previously published extractive text sumarization methods. The competitors were evaluated using the CNN-Corpus, the largest test set available today for single document extractive summarization.

KEYWORDS

Text summarization, text documents, NLP, CNN Corpus

ACM Reference Format:
Rafael Dueire Lins, Rafael Ferreira Mello, and Steve Simske. 2019. DocEng'19 Competition on Extractive Text Summarization. In *ACM Symposium on Document Engineering 2019 (DocEng '19), September 23–26, 2019, Berlin, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3342558.3351874

1 INTRODUCTION

Automatic text summarization (ATS) is a computer method to create a shorter version of one or more text documents[2]. ATS may offer a way of finding relevant information in large text libraries or on the Internet. Text summarization techniques are classified as *extractive* or *abstractive*. The extractive approach, used in this competition, selects a set of the most significant sentences from a document, exactly as they appear.

This competition makes use of the the CNN-corpus [1] to evaluate the proposed systems in the single-document extractive summarization task. The current version of the CNN-corpus encompasses 3,000 texts in English, with abstractive summaries written by the original authors (the *highlights*) and extractive summaries (the *gold-standards*), which were carefully developed by a team of people following a rigorous methodology with a number of software tools specially designed for such a task. The starting point for the texts in the CNN-Corpus were the news articles extracted from the CNN website (www.cnn.com), which contains high-quality, grammatically correct texts reporting on subjects of general interest and use standard vocabulary.

2 COMPETITION PARTICIPANTS

Four teams from Europe and the Americas originally enrolled in this competition. Unfortunately, the summarization software platforms of two of the enrolled teams were not able to finish the task of summarizing the test set chosen for this competition, a 1,000 texts subset of the CNN-Corpus in English. Once contacted, they could not return the fixed version of their code in time to meet the response deadline, thus, they were excluded of this competition. Below, the competitors are presented in the order of enrollment, having as affiliation that of the first participant of the team.

Federal Institute of Espirito Santo, Brazil
(*Hilario Oliveira, Rinaldo Lima, Fred Freitas*): This system uses ILP to maximize the relevant concepts in the sentence selection task [5]. This approach addresses text summarization as a problem of maximum coverage, aiming to select the subset of the sentences of the original text that maximizes the coverage of the relevant concepts from the input document, taking into account the desired maximum size of the summary. It seeks to consider both the informativeness and an estimate of the local cohesion of the summary to be generated.

Fraunhofer Center for Machine Learning, Germany
(Eduardo Brito, Max Lubbering, David Biesner, Lars Patrick Hillebrand,and Christian Bauckhage): This approach adopts a recurrent neural network based model that learns to classify whether a sentence belongs to the corresponding extractive summary. More specifically, it is based on the SummaRuNNer model [4], with two modifications: (i) it operates directly on a sentence level; (ii) the system does not consider the absolute or relative position of each sentence.

Traditional algorithms: It is of paramount importance to analyze how the newly proposed systems perform in comparison with some other "classical" extractive summarization approaches [2]. Thus, the following methods are also part of the assessment reported here:

1. **Word Frequency:** The more frequently a words occurs in the text, the higher its score;
2. **TF/IDF:** It uses TF/IDF formula to score sentences;
3. **Word Co-occurrence (WC):** measures the probability of two terms in a text to appear alongside each other in a certain order;
4. **Lexical Similarity:** It is based on the assumption that the important sentences are identified by strong chains;
5. **Upper Case** This method assigns higher scores to words that contain one or more upper case letters;
6. **Proper Noun:** This method hypothesizes that sentences that contain a higher number of proper nouns are possibly more important than others.
7. **Cue-Phrases:** In general, the sentences started by "in summary", "in conclusion", etc. as well as domain-specific bonus phrases terms can be good indicators of significant content of a text document;
8. **Sentence Position:** The position of the sentence in the text is seen as an indicator of its importance;

DocEng '19, September 23–26, 2019, Berlin, Germany

ACM ISBN 978-1-4503-6887-2/19/09.
https://doi.org/10.1145/3342558.3351874

Table 1: Competition Results.

ID	System/Method	Results						
		ROUGE-1			ROUGE-2			
		Precision	Recall	F-measure	Precision	Recall	F-measure	Direct Matching
1	Word Frequency	0.41(0.20)	0.50(0.21)	0.44(0.19)	0.28(0.23)	0.35(0.26)	0.31(0.23)	0.28
2	TF/IDF	0.41(0.20)	0.50(0.21)	0.44(0.20)	0.28(0.23)	0.35(0.27)	0.31(0.24)	0.28
3	Word Co-occurrence	0.33(0.19)	0.39(0.19)	0.35(0.18)	0.19(0.22)	0.22(0.23)	0.20(0.22)	0.16
4	Lexical Similarity	0.28(0.15)	0.29(0.16)	0.28(0.14)	0.12(0.16)	0.13(0.18)	0.12(0.16)	0.11
5	Upper Case	0.36(0.19)	0.44(0.20)	0.39(0.19)	0.23(0.21)	0.28(0.25)	0.25(0.22)	0.23
6	Proper Noun	0.36(0.18)	0.44(0.20)	0.39(0.18)	0.23(0.21)	0.28(0.25)	0.25(0.22)	0.23
7	Cue-Phrases	0.32(0.16)	0.34(0.17)	0.32(0.16)	0.17(0.19)	0.18(0.20)	0.17(0.19)	0.16
8	Sentence Position	0.38(0.18)	0.41(0.20)	0.38(0.18)	0.24(0.21)	0.26(0.23)	0.24(0.21)	0.24
9	Resemblance to the Title	0.40(0.19)	0.46(0.21)	0.42(0.19)	0.27(0.22)	0.32(0.26)	0.29(0.23)	0.27
10	Sentence Centrality	0.26(0.13)	0.21(0.13)	0.22(0.12)	0.08(0.14)	0.07(0.13)	0.07(0.13)	0.08
11	Sentence Length	0.35(0.18)	0.44(0.20)	0.38(0.18)	0.21(0.21)	0.27(0.25)	0.23(0.22)	0.20
12	Inclusion of Numerical Data	0.36(0.18)	0.41(0.19)	0.37(0.18)	0.22(0.20)	0.25(0.24)	0.23(0.21)	0.22
13	Bushy Path	0.32(0.15)	0.38(0.17)	0.34(0.15)	0.17(0.17)	0.20(0.20)	0.18(0.18)	0.17
14	Aggregate Similarity	0.32(0.15)	0.37(0.17)	0.34(0.15)	0.16(0.17)	0.19(0.20)	0.17(0.18)	0.16
15	Oliveira *et al.*	0.44(0.21)	0.57(0.19)	0.49(0.19)	0.34(0.23)	0.45(0.25)	0.37(0.22)	0.37
16	Brito *et al.*	0.46(0.21)	0.46(0.19)	0.46(0.19)	0.33(0.23)	0.34(0.25)	0.32(0.22)	0.33

9. Resemblance to the Title: If the vocabulary in the sentence resembles the title, it is regarded as important.
10. Sentence Centrality: there is a vocabulary overlap between a sentence and other sentences in the document;
11. Sentence Length: This feature is employed to penalize sentences that are either too short or long;
12. Inclusion of Numerical Data: sentences with numerical data are seen as important.
13. Bushy Path: of a sentence on a map is the number of links connecting it to other sentences on the map;
14. Aggregate Similarity: Instead of counting the number of links connecting a sentence to other Bushy Path, it sums up the weights of the links.

3 EVALUATION MEASURES

Two quantitative methodologies were used to compare the automatic generated summaries to the gold standard.

ROUGE: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [3] is widely used in the assessment of the quality of summaries. This fully automated evaluator measures the degree of "content" similarity between a system-developed summary and another summary taken as a reference. Two evaluation measures are taken here the ROUGE-1 and ROUGE-2, which compute the number unigrams and bigrams overlaps, respectively.

Direct Matching: The direct matching evaluation is done by counting the numbers of sentences selected by the system that match the human developed gold standard.

4 RESULTS

This competition made publicly available 50 texts with their highlights (abstractive) and gold standard (extractive) summaries from the CNN-Corpus in English. The text compression rate set was of 10% of the original text document, limited to a minimum of 3 sentences. Table 1 shows the final result of the assessing the competing and classical summarizers with a test set of 1,000 texts randomly chosen from the 3,000 texts in the CNN-Corpus, having as reference the gold-standard summaries for each text. The figures shown within parentheses stand for the standard deviation of the measure shown to their left. A number of conclusions may be drawn:

(1) The percent of sentences of the summaries that directly match the gold standard summaries is almost exactly the Precision figures found for ROUGE-2.
(2) The extractive summaries generated using the "classical" techniques of Word Frequency and TF/IDF in isolation, yield quantitative results closest to the newly proposed techniques assessed here, which were the top 2 of 16.
(3) The technique presented by Oliveira and his colleagues performed slightly better than the one by Brito et al.
(4) There is still much room for improving the results in ATS.

The results of ROUGE-1 and ROUGE-2 obtained with the highlights as a reference, were very close to the ones presented in Table 1.

REFERENCES

[1] Rafael Dueire Lins, Hilário Tomaz, Rafael Ferreira, Bruno Avila, Luciano Cabral, Jamilson Batista, Gabriel Silva, and Steven Lima, Rinaldoand Simske. 2019. The CNN-Corpus: A Large Textual Corpus for Single-Document Extractive Summarization. In *DocEng*. ACM, 1–10.
[2] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J Simske, and Luciano Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications* 40, 14 (2013), 5755–5764.
[3] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL-04 Workshop*, Marie-Francine Moens and Stan Szpakowicz (Eds.). Association for Computational Linguistics, Barcelona, Spain, 74–81.
[4] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI Conference on Artificial Intelligence*.
[5] Hilário Oliveira, Rinaldo Lima, Rafael Dueire Lins, Fred Freitas, Marcelo Riss, and Steven J Simske. 2016. A concept-based integer linear programming approach for single-document summarization. In *BRACIS*. IEEE, 403–408.