Variations of TextRank for Automated Summarization

Abstract. This article describes our proposal for new variants of the TextRank algorithm for automated summarization of texts. We describe the generalities of the TextRank algorithm on its original version and the different variations of the algorithm that we created. Some of these variants achieve a significative improvement over the original algorithm using the same metrics and dataset as the original publication.

Keywords: TextRank variations, automated summarization, Information Retrieval ranking functions

1 Introduction

We can describe the process of automated summarization as the extraction of the most important sentences in a document. Using different levels of compression a summarized version of the document of arbitrary length can be obtained. The TextRank algorithm is one of the most used methods for this task. TextRank builds a graph of sentences for a document and then applies PageRank to obtain a score for each sentence. In this article we describe several different proposals for the construction of the TextRank graph and report the results obtained with them.

The first section of this article describes previous work in the area and the TextRank algorithm in general. Then we report the results obtained for the different variations of the algorithm. Finally we describe the different metrics used for the evaluation of the results obtained from the proposed changes and the datasets used for these tests.

2 Previous work

The field of automated summarization has attracted interest since the late 50's [8]. Traditional methods for text summarization analyze the frequency of words or sentences in the first paragraphs of the text to identify the most important lexical elements. Several statistical models have been developed based on training corpora to combine different heuristics using keywords, position and length of sentences, word frequency and titles [7]. On a different approach, some algorithms analyze the semantic structure of the different textual units in a document with the goal of separating those that take a significative role in the representation of the document [10].

Other methods are based in the representation of the text as a graph: the most important sentences are the most connected ones in the graph and are used for

building a final summary [1]. These algorithms use different information retrieval techniques to identify similar sentences and determine the most important ones [15]. The TextRank algorithm developed by Mihalcea and Tarau [11] and the LexRank algorithm by Erkan and Radev [4] are based in ranking the lexical units of the text (sentences or words).

3 TextRank

3.1 Description

TextRank is an unsupervised algorithm for the automated summarization of texts that can also be used to obtain the most important keywords in a document. It was introduced in 2004 by Rada Mihalcea and Paul Tarau in [11].

The algorithm applies a variation of PageRank [12] over a graph constructed specifically for the task of summarization. This method provides an understanding of the structure of the document identifying its principal concepts without the need of previous training. Since the algorithm is based on PageRank it uses the idea of ranking of the elements in the graph, the most important elements are the ones that better describe the text. This approach allows TextRank to build summaries without the need of a training corpus or labeling and allows the use of the algorithm with different languages as long as there is a way to build the graph of sentences for it.

3.2 Text as a Graph

For the task of automated summarization, TextRank models any document as a graph using sentences as nodes [2]. A function to compute the similarity of sentences is needed to build edges in between. This function is used to weight the graph edges, the higher the similarity between sentences the more important the edge between them will be in the graph. In the domain of a Random Walker, as used frequently in PageRank, we can say that we are more likely to go from one sentence to another if those sentences are very similar.

The similarity function can use several different ideas. It can be based on the semantic of the sentences, on their proximity in the text, common words, and many other different metrics. The goal of this article is to experiment with these functions and report the results obtained when used along with TextRank.

The function featured in the original algorithm can be formalized as:

Definition 1 *Given S_i, S_j two sentences represented by a set of n words that in S_i are represented as $S_i = w_1^i, w_2^i, ..., w_n^i$. The similarity function for S_i, S_j can be defined as:*

$$Sim(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{1}$$

The result of this process is a dense graph representing the document. From this graph, PageRank is used to compute the importance of each vertex. The most significative sentences are selected and presented in the same order as they appear in the document as the summary.

4 Experiments

4.1 Our Variations

This section will describe the different variations that we propose over the original TextRank algorithm. These ideas are based in changing the way in which distances between sentences are computed to weight the edges of the graph used for PageRank. We found some of these variations to produce significative improvements over the original algorithm.

Longest Common Substring From two sentences we identify the longest common substring and report the similarity to be its length [5].

Cosine Distance The cosine similarity is a metric widely used to compare texts represented as vectors. We used a classical TF-IDF model to represent the documents as vectors and computed the cosine between vectors as a measure of similarity. Since the vectors are defined to be positive, the cosine results in values in the range [0,1] where a value of 1 represents identical vectors and 0 represents orthogonal vectors [16].

BM25 BM25 / Okapi-BM25 is a ranking function widely used as the state of the art for Information Retrieval tasks. BM25 is a variation of the TF-IDF model using a probabilistic model [14].

Definition 2 *Given two sentences R, S, BM25 is defined as:*

$$BM25(R,S) = \sum_{i=1}^{n} IDF(s_i) \cdot \frac{f(s_i,R) \cdot (k_1+1)}{f(s_i,R) + k_1 \cdot (1 - b + b \cdot \frac{|R|}{avgDL})} \tag{2}$$

where k and b are parameters. We used $k = 1.2$ and $b = 0.75$. $avgDL$ is the average length of the sentences in our collection.

This function definition implies that if a word appears in more than half the documents of the collection, it will have a negative value. Since this can cause problems in the next stage of the algorithm, we used the following correction formula:

$$IDF(s_i) = \begin{cases} log(N - n(s_i) + 0.5) - log(n(s_i) + 0.5) & \text{if } n(s_i) > N/2 \\ \varepsilon \cdot avgIDF & \text{if } n(s_i) \leq N/2 \end{cases} \tag{3}$$

where ε takes a value between 0.5 and 0.30 and $avgIDF$ is the average IDF for all terms.

We also used BM25+, a variation of BM25 that changes the way long documents are penalized [9].

4.2 Evaluation

For testing the proposed variations, we used the database of the 2002 Document Understanding Conference (DUC) [3]. The corpus has 567 documents that are summarized to 20% of their size, and is the same corpus used in [11].

To evaluate results we used version 1.5.5 of the ROUGE package [6]. The configuration settings were the same as those in DUC, where ROUGE-1, ROUGE-2 and ROUGE-SU4 were used as metrics, using a confidence level of 95% and applying stemming. The final result is an average of these three scores.

To check the correct behaviour of our test suite we implemented the reference method used in [11], which extracts the first sentences of each document. We found the resulting scores of the original algorithm to be identical to those reported in [11]: a 2.3% improvement over the baseline.

4.3 Results

We tested LCS, Cosine Sim, BM25 and BM25+ as different ways to weight the edges for the TextRank graph. The best results were obtained using BM25 and BM25+. We achieved an improvement of 2.92% above the original TextRank result using BM25 and $\varepsilon = 0.25$. The following chart shows the results obtained for the different variations we proposed.

Table 1. Evaluation results for the proposed TextRank variations.

Method	ROUGE-1	ROUGE-2	ROUGE-SU4	Improvement
BM25 (Neg to epsilon)	0.4042	0.1831	0.2018	2.92%
BM25+ (Neg to epsilon)	0.404	0.1818	0.2008	2.60%
Cosine TF-IDF	0.4108	0.177	0.1984	2.54%
BM25+ (IDF = log(N/NI))	0.4022	0.1805	0.1997	2.05%
BM25 (IDF = log(N/NI))	0.4012	0.1808	0.1998	1.97%
Longest Common Substring	0.402	0.1783	0.1971	1.40%
BM25+ (Neg to zero)	0.3992	0.1803	0.1976	1.36%
BM25 (Neg to zero)	0.3991	0.1778	0.1966	0.89%
TextRank	**0.3983**	**0.1762**	**0.1948**	–
BM25	0.3916	0.1725	0.1906	-1.57%
BM25+	0.3903	0.1711	0.1894	-2.07%
DUC Baseline	0.39	0.1689	0.186	-2.84%

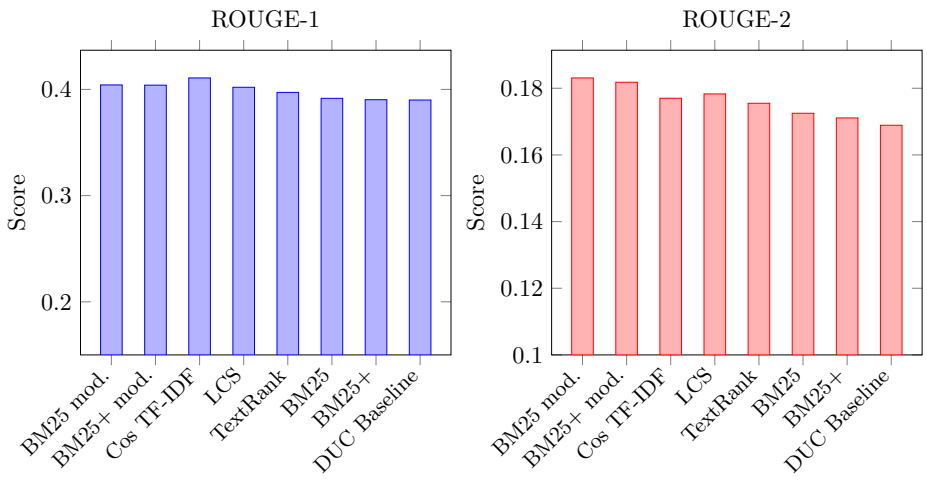

Fig. 1. ROUGE-1 and ROUGE-2 scores comparison.

The result of Cosine Similarity was also satisfactory with a 2.54% improvement over the original method. The LCS variation also improved the original TextRank algorithm with 1.40% total improvement.

The performance in time was also improved. We could process the 567 documents from the DUC2002 database in 84% of the time needed in the original version.

5 Example

A sample document from the 2002 DUC (Document Understanding Conference) dataset [3] is shown in Figure 2. The modified TextRank algorithm builds the graph shown in Figure 3 and produces the summary shown in Figure 4. The summary generated by the original method is shown in Figure 5.

6 Reference Implementations and Gensim Contribution

A reference implementation of our proposals was coded as a Python module. It can be obtained for testing and to reproduce results from an URL that will be provided in the non-anonymous version.

We also contributed the BM25-TextRank algorithm to the Gensim project [13].

7 Conclusions

This work presented three different variations to the TextRank algorithm for automatic summarization. The three variations presented improved significantly

Fig. 2. Sample document from the DUC 2002 corpus.

1. Super Bowl Was Lowest-Rated in 21 Years
2. By RONALD BLUM AP Sports Writer NEW YORK (AP)
3. The San Francisco 49ers' 55-10 rout of the Denver Broncos was the lowest-rated Super Bowl in 21 years and the third-lowest ever.
4. The game on CBS averaged a 39.0 rating and a 63 share, the lowest Super Bowl rating since 1969, when the New York Jets' 16-7 victory over Baltimore got a 36.0 on NBC for the worst rating ever, A.C. Nielsen Co. said today.
5. The rating is a percentage of the nation's televisions; each point represents 921,000 homes.
6. The share is the percentage of the televisions on at the time.
7. Despite the low rating, Sunday's game was seen by about 108.5 million people, making it the ninth most-watched TV show ever in the United States behind eight Super Bowls and the final episode of "MASH".
8. The higher viewership was made possible by the annual increase in the number of homes with television.
9. "Given the expected blowout, the numbers are completely understandable and we're happy to have a 39," said Susan Kerr, director of programming for CBS Sports.
10. "It's still a remarkable rating for prime time".
11. The 49ers won by the biggest margin in Super Bowl history.
12. Only the 1969 and 1968 Super Bowls had lower ratings; Green Bay's 33-14 victory over Oakland in 1968 got a 36.8.
13. The first Super Bowl was in 1967.
14. The highest-rated Super Bowl was in 1982, when the 49ers' 26-21 victory over Cincinnati got a 49.1 on CBS.
15. Last year's San Francisco-Cincinnati Super Bowl, won by the 49ers 20-16, got a 43.5 rating and was seen in 39.3 million homes, according to Nielsen estimates.
16. This year's game was seen in about 35.9 million homes.
17. Viewership of the 5 p.m. game this year peaked at a 41.6 rating and a 69 share from 6 p.m. to 6:30 p.m., dropping off as it became clear that Denver would not rally.
18. The pregame show got a 39.0 rating and a 63 share and the postgame show got a 22.5 rating and a 35 share.
19. The musical "Annie," which was on ABC opposite the game, got a 9.4 rating.
20. "Life Goes On" on NBC got a 6.0 rating and "Love With a Twist" on NBC a 7.6 .

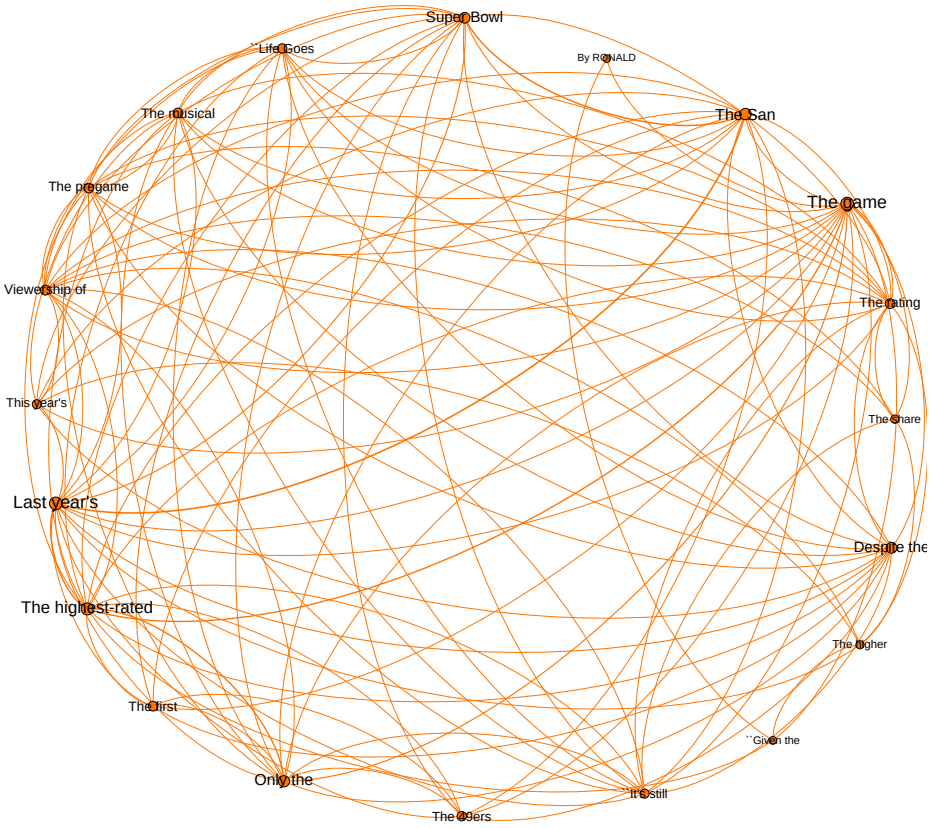

Fig. 3. Graph created from the sample text.

the results of the algorithm using the same metric and database used in the original paper. Given that TextRank performs 2.84% over the baseline, our improvement of 2.92% over the TextRank result is an important result.

The combination of TextRank with modern Information Retrieval ranking functions such as BM25 and BM25+ results in the creation of a robust method for automatic summarization that performs better than the standard techniques used previously.

Based on this results we suggest the use of BM25 along with TextRank for the task of unsupervised automatic summarization of texts. The results obtained and the examples analyzed show that this variation is better than the original TextRank algorithm without a performance penalty.

Fig. 4. Summary generated by modified TextRank with BM25.

The San Francisco 49ers' 55-10 rout of the Denver Broncos was the lowest-rated Super Bowl in 21 years and the third-lowest ever. The game on CBS averaged a 39.0 rating and a 63 share, the lowest Super Bowl rating since 1969, when the New York Jets' 16-7 victory over Baltimore got a 36.0 on NBC for the worst rating ever, A.C. Nielsen Co. said today. The highest-rated Super Bowl was in 1982, when the 49ers' 26-21 victory over Cincinnati got a 49.1 on CBS. Last year's San Francisco-Cincinnati Super Bowl, won by the 49ers 20-16, got a 43.5 rating and was seen in 39.3 million homes, according to Nielsen estimates.

Fig. 5. Summary generated by original TextRank from the sample text.

Super Bowl Was Lowest-Rated in 21 Years The game on CBS averaged a 39.0 rating and a 63 share, the lowest Super Bowl rating since 1969, when the New York Jets' 16-7 victory over Baltimore got a 36.0 on NBC for the worst rating ever, A.C. Nielsen Co. said today. The highest-rated Super Bowl was in 1982, when the 49ers' 26-21 victory over Cincinnati got a 49.1 on CBS. Last year's San Francisco-Cincinnati Super Bowl, won by the 49ers 20-16, got a 43.5 rating and was seen in 39.3 million homes, according to Nielsen estimates.

References

1. Barzilay, R., McKeown, K.: Sentence fusion for multidocument news summarization. Computational Linguistics 31(3), 297–328 (2005), http://dblp.uni-trier.de/db/journals/coling/coling31.html#BarzilayM05
2. Christopher D. Manning, Prabhakar Raghavan, H.S.: Introduction to Information Retrieval. Cambridge University Press (2008)
3. Document Understanding Conference: Duc 2002 guidelines (July 2002), http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html
4. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. (JAIR) 22, 457–479 (2004), http://dblp.uni-trier.de/db/journals/jair/jair22.html#ErkanR04
5. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY, USA (1997)
6. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain (2004)
7. Lin, C.Y., Hovy, E.H.: Identifying topics by position. In: Proceedings of 5th Conference on Applied Natural Language Processing. Washington D.C. (March 1997)
8. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (Apr 1958), http://dx.doi.org/10.1147/rd.22.0159
9. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011. pp. 7–16 (2011)
10. Marcu, D.: The theory and practice of discourse parsing and summarization. Computational Linguistics 28(1), 81–83 (2000), http://dblp.uni-trier.de/db/journals/coling/coling28.html#Hahn02
11. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (July 2004)

12. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference. pp. 161–172. Brisbane, Australia (1998), citeseer.nj.nec.com/page98pagerank.html
13. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
14. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994. pp. 109–126 (1994), http://trec.nist.gov/pubs/trec3/papers/city.ps.gz
15. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing and Management 33(2), 193 – 207 (1997)
16. Singhal, A.: Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24(4), 35–43 (2001), http://singhal.info/ieee2001.pdf