Automatic Text Summarization based on Betweenness Centrality

Gretel Liz De la Peña Sarracén
Center for Pattern Recognition and Data Mining, Cuba
gretel@cerpamid.co.cu

Paolo Rosso
PRHLT, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

ABSTRACT

Automatic text summary plays an important role in information retrieval. With a large volume of information, presenting the user only a summary greatly facilitates the search work of the most relevant. Therefore, this task can provide a solution to the problem of information overload. Automatic text summary is a process of automatically creating a compressed version of a certain text that provides useful information for users. This article presents an unsupervised extractive approach based on graphs. The method constructs an indirected weighted graph from the original text by adding a vertex for each sentence, and calculates a weighted edge between each pair of sentences that is based on a similarity/dissimilarity criterion. The main contribution of the work is that we do a study of the impact of a known algorithm for the social network analysis, which allows to analyze large graphs efficiently. As a measure to select the most relevant sentences, we use betweenness centrality. The method was evaluated in an open reference data set of DUC2002 with Rouge scores.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Natural language processing; Information extraction;

KEYWORDS

Automatic Text Summarization, Extractive Summary, Betweenness Centrality

ACM Reference Format:
Gretel Liz De la Peña Sarracén and Paolo Rosso. 2018. Automatic Text Summarization based on Betweenness Centrality. In *CERI 18: 5th Spanish Conference in Information Retrieval, June 26–27, 2018, Zaragoza, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3230599.3230611

CERI 18, June 26–27, 2018, Zaragoza, Spain

ACM ISBN 978-1-4503-6543-7/18/06…$15.00
https://doi.org/10.1145/3230599.3230611

1 INTRODUCTION

Automatic text summarization is the process of generating a brief and accurate representation of the input text such that the result covers the most important concepts of the source in a condensed way. It is an important object of research in particular, though not exclusively, in Natural Language Processing.

Summary approaches can be divided into two categories. *Abstraction* analyzes the original text in a deep linguistic way, semantically interprets the text in a formal representation, finds concise concepts to describe the text and generates a new abbreviated text with the same information content [3]. However, this approach is more challenging than the other. On the other hand, *extraction* identifies and recovers the most relevant parts at the source [10].

Extractive summarization involves assigning salience scores to some units of the document and extracting the sentences with highest scores. Usually, the scores are based on statistical analysis of individual or mixed features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The most important content can be treated as the most frequent or the most favorably positioned content. Thus, summarization based on sentence extraction can be seen as a sentence ranking problem. Most commonly, such ranking approaches use a sentence weighing strategy to rank them for the selection of those that will be included in the summary.

A lot of research mostly focuses on extractive summarization because of its applicability with today's models. Many methods use standard statistically based information with more or less shallow natural language processing and different heuristics [4, 7]. Some criteria for weighing sentences are the high frequency words, the presence or absence of certain cue words from a cue dictionary, and the position of sentences in the document.

Other kind of methods uses machine learning models [15]. Also, others such as the centroid-based and cluster-based models are very popular, which deal with the redundancy poblem that can affect the quality of the summaries [1, 13]. Moreover, models have been developed using Latent Semantic Analysis (LSA), which is based on mathematical technique which is named singular value decomposition to identify patterns in the relationships between the terms and sentences [8, 12].

A relevant group among these techniques is that of the graph-based algorithms. Graph-based ranking algorithms have been traditionally used in areas such as social networks,

and the analysis of the structure of the Web. These algorithms determine the importance of the vertices within a graph by taking into account the global information recursively calculated from such graph.

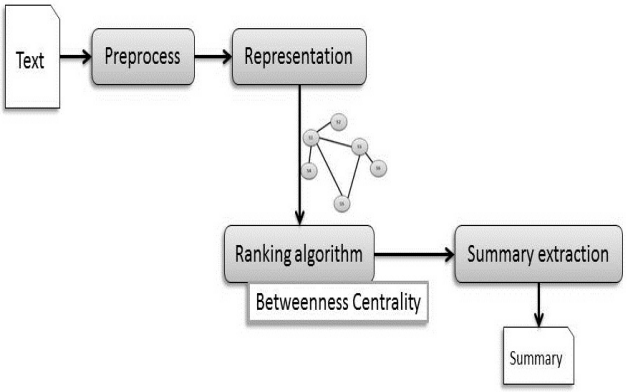
A similar idea can be applied to semantic graphs extracted from texts. TextRank [9] is a graph-based ranking model which has been used for tasks ranging from automated extraction of keyphrases to extractive summarization. LexRank [5] is another model for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In [14] the authors proposing a variation to the similarity function used to compute scores during sentence extraction for TextRank.

The proposed model in this paper is an extraction based summarization. In this model a graph was constructed based on the input sentences and the similarity/dissimilarity between each two sentences was calculated. A ranking algorithm is applied and the most important sentences based on their corresponding rank are identified.

2 PROPOSED APPROACH

The proposed approach in this paper is a graph-based extractive summarization such as Figure 1 shows. We will refer to the method with the name of BetweennesCent.

Figure 1: General scheme of BetweennesCent



The first step is a preprocessing, where sentences are split based on the punctuation marks that represent the end of a sentence. Also stop-words are removed to be excluded in the procedure and the rest of words are lemmatized. Second, the processed text is represented into an indirected weighted graph by adding a vertex for each sentence. Each sentence is represented as a bag of words. A similarity/dissimilarity criterion is used to represent the semantic relation between nodes as the weight of the edges. Only edges with a value of similarity/dissimilarity above/below a threshold are represented as part of the structure of the graph. Two criteria of similarity (Correlation and Cosine similarity) and one criteriion of dissimilarity (Euclidean distance) were used in this work:

- *Cosine Similarity*: Cosine distance between vectors u and v:
$$1 - \frac{u \cdot v}{||u||_2 \cdot ||v||_2}$$
- *Correlation*: Correlation distance between vectors u and v:
$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{||(u - \bar{u})||_2 \cdot ||(v - \bar{v})||_2}$$
Where $\bar{u}$ is the mean of the elements of vector v, $x \cdot y$ is the dot product x of and y and $||x||_2$ is the 2-norm of its argument x.
- *Euclidean distance*: The ordinary straight-line distance between two points in Euclidean space.

Then, a ranking algorithm is used as it is explained below. Finally, the summary is obtained by concatenation of these important sentences according to the original order of the sentences in the source. The number of sentences depends on the desired size for the summary.

2.1 Ranking Algorithm

In order to find which sentences are more relevant, a ranking algorithm is applied to the generated graph. Sentences are ranked on the basis of the betweenness centrality measure. This measure is determined by the algorithm proposed in [2]. The algorithm is mostly used in the social network analysis.

The main contribution of our work lies precisely in evaluating the impact of this algorithm on the task of extractive summarization, since the proposed methodology allows the analysis of large graphs efficiently that is convenient for processing large documents.

In the analysis of social networks, centrality indices defined on the vertices of the graph are quite important. They are designed to rank the vertices according to their position in the network and interpreted as the prominence of vertices embedded in a social structure.

The betweeness centrality is a measure of centrality that is presented as a metric to quantify responsibility within a network. Therefore, a vertex has a high betweeness centrality when it has a high probability of appearing in the shortest way between two vertices chosen uniformly at random. Thanks to this metric we can globally determine the most important sentences.

The betweenness centrality index of a vertex v ($C_B(v)$), can be calculated as the sum of the pair-dependencies of all pairs on that vertex.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v) \quad (1)$$

The pair-dependencies $\delta_{st}(v)$ of a pair $s, t \in V$ on an intermediary $v \in V$, where σ_{st} denotes the number of shortest paths from $s \in V$ to $t \in V$ with $\sigma_{ss} = 1$, and $\sigma_{st}(v)$ denotes the number of shortest paths from s to t that some $v \in V$ lies on, is given by:

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

$\sigma_{st}(v) = 0$ if $d_G(s,t) < d_G(s,v) + d_G(v,t)$
and $\sigma_{st}(v) = \sigma_{sv} \cdot \sigma_{vt}$ otherwise

Where G is a graph and V its set of verteces. $d_G(s,t)$ denotes the distance between vertices s and t. By definition, $d_G(s,s) = 0$ for every $s \in V$, and $d_G(s,t) = d_G(t,s)$ for $s,t \in V$.

3 EXPERIMENTAL RESULTS

The used dataset to evaluate was the task 1 of DUC 2002 for evaluation. The task aimed to evaluate generic summaries with a length of approximately 100 words or less. DUC 2002 provided 567 English news articles for single-document summarization task.

The *ROUGE-N* [6] evaluation was used which measures summary quality by counting overlapping between the candidate summary and reference summaries. *BLEU* [11] is also used which is a modified form of precision, widely used in the evaluation of machine translation.

BetweennessCent was evaluated with the three different distance measures mentioned earlier and a threshold to decide the inclusion of edges was 0.5. The size of the generated summaries was limited by 100 words according to the summaries used from the dataset.

Table 1: The ROUGE-1 and BLEU results for the proposed method with different distance measures

Method	$ROUGE-1$	BLEU
Baseline	0.3157	0.4995
LSA	0.3337	0.4865
TexRank	0.3400	0.5317
BetweenesCent(Correlation)	0.3774	0.5391
BetweenesCent(Cosine)	0.3816	0.5525
BetweenesCent(Euclidean)	**0.3833**	**0.5581**

Table 1 illustrates the evaluation results for each measure. As it is showed in the table, the variant using the Euclidean distance obtains better results according to the measure used. In addition, the results are compared with other algorithms. One of them, a common method that selects sentences randomly throughout the original text until the required size limit is obtained (Baseline). Other methods used in the comparison was LSA for Text Summarization and TextRank. The implementations was taken from Gensim. In both cases, our proposal outperforms the results.

4 CONCLUSION

Automatic text summarization is an important area of NLP research is becoming more common in digital library environment. Extractive summarization is a kind of text summarization where parts of text are selected automatically based on some criteria to produce summary. This paper presents an unsupervised extractive approach based on graphs. The method constructs an indirected weighted graph from the original text by adding a vertex for each sentence, and compute a weighted edge between each pair of sentences with a similarity/dissimilarity criterion. The principal contribution of the work is the study of the impact of a known algorithm for the social network analysis, which allows to analyze large graphs efficiently. In this way, this could be a good strategy for extractive summarization based on graphs when the documents are long, because in the experimentation, the obtained results were acceptable. It must be taken into account that the evaluation in this task is subjective, since there is no perfect summary. However, when analyzing the results obtained, using known measures in the area, it can be said that the summaries were suitable.

ACKNOWLEDGMENTS

The work of the second author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

REFERENCES

[1] Rasim Magamed Alguliev and Ramiz Magamed Alyguliev. 2008. Automatic Text Documents Summarization through Sentences Clustering. *Journal of Automation and Information Sciences* 40, 9 (2008).
[2] Ulrik Brandes. 2001. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* 25, 2 (2001), 163–177.
[3] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 93–98.
[4] Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)* 16, 2 (1969), 264–285.
[5] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.
[6] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries using n-gram Co-occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 71–78.
[7] Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165.
[8] IV Mashechkin, MI Petrovskiy, DS Popov, and Dmitry V Tsarev. 2011. Automatic Text Summarization using Latent Semantic Analysis. *Programming and Computer Software* 37, 6 (2011), 299–305.
[9] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.
[10] N Moratanch and S Chitrakala. 2017. A Survey on Extractive Text Summarization. In *Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on*. IEEE, 1–6.
[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.
[12] Pallavi D Patil and PM Mane. 2015. Improving the Performance for Single and Multi-document Text Summarization via LSA & FL. *IJCST* 2, 4 (2015).
[13] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based Text Summarization through Compositionality of Word Embeddings. In *Proceedings of the MultiLing 2017*

Workshop on Summarization and Summary Evaluation Across Source Types and Genres. 12–21.

[14] Sunchit Sehgal, Badal Kumar, Maheshwar, Lakshay Rampal, and Ankit Chaliya. 2018. A Modification to Graph Based Approach for Extraction Based Automatic Text Summarization. *Progress in Advanced Computing and Intelligent Engineering* 564 (2018), 373–378.

[15] Hamey L Yousefi Azar M. 2017. Text Summarization Using Unsupervised Deep Learning. *Expert Systems with Applications* 68 (2017), 93–105.