Language Independent Extractive Summarization

Rada Mihalcea
Department of Computer Science and Engineering
University of North Texas
rada@cs.unt.edu

Abstract

TextRank is a system for unsupervised extractive summarization that relies on an innovative application of iterative graph-based ranking algorithms to graphs encoding the cohesive structure of texts. An important characteristic of the system is that it does not rely on any language-specific knowledge resources or any manually constructed training data, and thus it is highly portable to new languages or domains.

Introduction

Given the overwhelming amounts of information available today, on the Web and elsewhere, techniques for efficient automatic text summarization are essential to improve the access to such information. Algorithms for extractive summarization are typically based on techniques for sentence extraction, and attempt to identify the set of sentences that are most important for the understanding of a given document. Some of the most successful approaches to extractive summarization consist of supervised algorithms that attempt to learn what makes a good summary by training on collections of summaries built for a relatively large number of training documents, e.g. (Hirao *et al.* 2002), (Teufel & Moens 1997). However, the price paid for the high performance of such supervised algorithms is their inability to easily adapt to new languages or domains, as new training data are required for each new type of data. *TextRank* is specifically designed to address this problem, by using an extractive summarization technique that does not require any training data or any language-specific knowledge resources. *TextRank* can be effectively applied to the summarization of documents in different languages without any modifications in the algorithm and without any requirements for additional data. Moreover, results from experiments performed on standard data sets have demonstrated that the performance of *TextRank* is competitive with that of some of the best summarization systems available today.

Extractive Summarization

Ranking algorithms, such as Kleinberg's HITS algorithm (Kleinberg 1999) or Google's PageRank (Brin & Page 1998) have been traditionally and successfully used in Web-link analysis, social networks, and more recently in text processing applications. In short, a graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, by taking into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. The basic idea implemented by the ranking model is that of *voting* or *recommendation*. When one vertex links to another one, it is basically casting a vote for that other vertex. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex.

These graph ranking algorithms are based on a random walk model, where a walker takes random steps on the graph, with the walk being modeled as a Markov process – that is, the decision on what edge to follow is solely based on the vertex where the walker is currently located. Under certain conditions, this model converges to a stationary distribution of probabilities associated with vertices in the graph, representing the probability of finding the walker at a certain vertex in the graph. Based on the Ergodic theorem for Markov chains (Grimmett & Stirzaker 1989), the algorithms are guaranteed to converge if the graph is both aperiodic and irreducible. The first condition is achieved for any graph that is a non-bipartite graph, while the second condition holds for any strongly connected graph. Both these conditions are achieved in the graphs constructed for the extractive summarization application implemented in *TextRank*.

While there are several graph-based ranking algorithms previously proposed in the literature, $PageRank$ (Brin & Page 1998) is perhaps one of the most popular. Given a directed graph $G = (V, E)$, with $In(V_i)$ denoting the set of vertices that point to a vertex V_i (predecessors), and $Out(V_i)$ denoting the set of vertices that vertex V_i points to (successors), the $PageRank$ score associated with each vertex in the graph is recursively defined as:

$$PR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \tag{1}$$

where d is a parameter that is set between 0 and 1, and has the role of integrating random jumps into the random walking model. Starting from arbitrary values assigned to each node in the graph, the computation iterates until convergence below a given threshold is achieved. After running the algorithm, a score is associated with each vertex, which represents the *importance* of that vertex within the graph. Note that the final scores are not affected by the choice of initial vertex values, only the number of iterations to convergence may be different.

When the graphs are built starting with natural language texts, it may be useful to integrate into the graph model the *strength* of the connection between two vertices V_i and V_j indicated as a weight w_{ij} added to the corresponding edge. The ranking algorithm was thus adapted to include edge weights:

$$PR^W(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR^W(V_j)}{\sum_{V_k \in Out(V_j)} w_{kj}} \quad (2)$$

For the task of single-document extractive summarization, the goal is to rank the sentences in a given text with respect to their importance for the overall understanding of the text. A graph is therefore constructed by adding a vertex for each sentence in the text, and edges between vertices are established using sentence inter-connections. These connections are defined using a similarity relation, where "similarity" is measured as a function of content overlap. Such a relation between two sentences can be seen as a process of *recommendation*: a sentence that addresses certain concepts in a text gives the reader a *recommendation* to refer to other sentences in the text that address the same concepts, and therefore a link can be drawn between any two such sentences that share common content.

The overlap of two sentences can be determined simply as the number of common tokens between the lexical representations of the two sentences, or optionally it can be run through filters that e.g. eliminate stopwords, count only words of a certain category, etc. Moreover, to avoid promoting long sentences, we are using a normalization factor, and divide the content overlap of two sentences with the length of each sentence.

The resulting graph is highly connected, with a weight associated with each edge, indicating the strength of the connections between various sentence pairs in the text. The graph can be represented as: (a) simple *undirected* graph; (b) directed weighted graph with the orientation of edges set from a sentence to sentences that follow in the text (*directed forward*); or (c) directed weighted graph with the orientation of edges set from a sentence to previous sentences in the text (*directed backward*).

After the ranking algorithm is run on the graph, sentences are sorted in reversed order of their score, and the top ranked sentences are selected for inclusion in the extractive summary.

Evaluation

TextRank was evaluated on two standard summarization data sets consisting of newspaper articles in English and Portuguese. For English, the evaluation was conducted on 567 English news articles made available during the Document Understanding Evaluations 2002 (DUC 2002). For Portuguese, we used the TeMário data set, consisting of 100 documents from Brazilian newspapers (Pardo & Rino 2003). Regardless of the language, the evaluation was performed using the ROUGE evaluation toolkit (Lin & Hovy 2003). Table 1 shows the results obtained on both data sets for different graph settings. The table also lists baseline results, obtained on summaries generated by taking the first sentences in each document. By ways of comparison, the best participating system in DUC 2002 was a *supervised* system that led to a ROUGE score of 0.5011.

Data set	Graph			Baseline
	Undirected	Forward	Backward	
English	0.4904	0.4202	**0.5008**	0.4799
Portuguese	0.4939	0.4574	**0.5121**	0.4963

Table 1: Evaluation of *TextRank* extractive summaries.

The results are encouraging: for both data sets, *TextRank* applied on a *directed backward* graph structure exceeds by a large margin the performance achieved through a simple (but competitive) baseline. These results prove that graph-based ranking algorithms, previously found successful in Web link analysis and social networks, can be turned into a state-of-the-art tool for extractive summarization when applied to graphs extracted from texts. Moreover, due to its unsupervised nature, the algorithm was also shown to be language independent, leading to similar results and similar improvements over baselines when applied on documents in different languages. More extensive experimental results with the *TextRank* system are reported in (Mihalcea & Tarau 2004), (Mihalcea & Tarau 2005).

References

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7).

DUC. 2002. Document understanding conference 2002. http://www-nlpir.nist.gov/projects/duc/.

Grimmett, G., and Stirzaker, D. 1989. *Probability and Random Processes*. Oxford University Press.

Hirao, T.; Sasaki, Y.; Isozaki, H.; and Maeda, E. 2002. Ntt's text summarization system for duc-2002. In *Proceedings of the Document Understanding Conference 2002*.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.

Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*.

Mihalcea, R., and Tarau, P. 2004. TextRank – bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Mihalcea, R., and Tarau, P. 2005. Iterative graph-based algorithms for language independent extractive summarization. (submitted).

Pardo, T., and Rino, L. 2003. TeMario: a corpus for automatic text summarization. Technical report, NILC-TR-03-09.

Teufel, S., and Moens, M. 1997. Sentence extraction as a classification task. In *ACL/EACL workshop on "Intelligent and scalable Text summarization"*, 58–65.