A Zipf-Like Distant Supervision Approach for Multi-document Summarization Using Wikinews Articles

Felipe Bravo-Marquez[1] and Manuel Manriquez[2]

[1] Department of Computer Science, University of Chile
[2] University of Santiago of Chile
fbravo@dcc.uchile.cl, manuel.manriquez@usach.cl

Abstract. This work presents a sentence ranking strategy based on distant supervision for the multi-document summarization problem. Due to the difficulty of obtaining large training datasets formed by document clusters and their respective human-made summaries, we propose building a training and a testing corpus from Wikinews. Wikinews articles are modeled as "distant" summaries of their cited sources, considering that first sentences of Wikinews articles tend to summarize the event covered in the news story. Sentences from cited sources are represented as tuples of numerical features and labeled according to a relationship with the given distant summary that is based on the Zipf law. Ranking functions are trained using linear regressions and ranking SVMs, which are also combined using Borda count. Top ranked sentences are concatenated and used to build summaries, which are compared with the first sentences of the distant summary using ROUGE evaluation measures. Experimental results obtained show the effectiveness of the proposed method and that the combination of different ranking techniques outperforms the quality of the generated summary.

1 Introduction

Automatic document summarization is the task of presenting the most important content of a text source in a condensed form to a final user [15]. The problem dates back to Luhn's work in 1958 [14] and has played an important role in information extraction systems where the amount of information presented to users is elevated. In environments such as document management systems and large scale search engines, summaries support users to identify which documents satisfy their information needs without the need to review all documents presented by the system. Moreover, the increasing amount of digital document collections and web pages available, makes the elaboration of human-made summaries expensive and unscalable. As stated in [6], the elaboration of human-quality summaries using a computational approach is difficult to achieve without natural language understanding. Nevertheless, for information retrieval purposes, the task can be reduced to a sentence ranking problem which is closely related to the more general information retrieval problem [6]. Extractive multi-document

L. Calderón-Benavides et al. (Eds.): SPIRE 2012, LNCS 7608, pp. 143–154, 2012.

summarization consists of selecting the most informative sentences from a cluster of event-related documents and use them to build a reduced summary describing the event. As stated in [22], a multi-document summary can be used to describe the information contained in a cluster of documents and facilitate users' understanding of the clusters. By the remainder of this work, we will refer to a document cluster as a set of documents covering the same event.

In this work, a multi-document summarization model based on the distant supervision paradigm is proposed. A supervised approach requires the existence of a dataset composed by event-related document clusters and their respective human-made summaries. The Document Understanding Conference (DUC)[1] which has moved to the Text Analysis Conference (TAC)[2] since 2007, both sponsored by NIST[3], provide training and evaluation datasets for summarization. Nevertheless, it is hard to find large scale training datasets for supervised multi-document summarization. The distant supervision paradigm [18] consists of using a weakly labeled training dataset based on a heuristic labeling function for supervised learning. We opted for using this approach for summarization by extracting knowledge from the Web using Wikinews[4] articles. Wikinews is a free-content news source that works through collaborative journalism. Around 1,860 news stories have been published in English in addition to other languages. The Wikinews style guide[5] suggests that authors summarize the whole story on the first paragraph. Furthermore, the inclusion of links to all references from other news sources is also suggested. These conventions make Wikinews articles and their news sources ideal for the distant supervision paradigm. We created a training dataset and a testing dataset using Wikinews articles as "distant" summaries of their respective news sources. The summaries are considered as "distant" because although they are not real summaries of their sources, there is empirical evidence that their first sentences summarize the event described in the sources. We extracted a training dataset and a testing dataset using articles together with their sources dated on 2010 and 2011 respectively.

A function that takes a cluster of source documents as argument and returns the sentences within the documents ranked by the likelihood of being part of a summary is trained from the training dataset. The learning procedure is performed by converting each sentence of the document cluster into a vector composed by numerical features and a noisy target variable. The features are extracted from statistical properties of the documents and are independent of the documents language. The target variable is a score computed according to a relationship between the source sentence and the corresponding article from Wikinews, which is calculated using the similarity between the sentence and the article. Due the fact that Wikinews articles are not real summaries of their sources, similarities between source sentences and first article sentences

[1] http://duc.nist.gov/
[2] http://www.nist.gov/tac/
[3] http://www.nist.gov/
[4] http://www.wikinews.org/
[5] http://en.wikinews.org/wiki/Wikinews:Style_guide

are weighted higher than the similarities between source sentences and the later sentences. This is achieved using weighting factors based on the Zipf law.

Two ranking learning algorithms are used. The former is a linear regression and the latter is a ranking support vector machine (SVM). Furthermore, we propose using a third ranking function which combines the others using a simple ranking fusion technique called Borda count. The trained ranking functions are applied to the testing dataset and resulting top-ranked sentences are concatenated and presented as summaries. Resulting summaries obtained from each of the ranking functions are compared with the top sentences of the Wikinews articles using ROUGE evaluation measures [25]. Results obtained show that Borda count helps to improve the overall quality of summaries when different evaluation criteria are considered.

This paper is organized as follows, in section 2 related work in extractive summarization and ROUGE evaluation measures are presented. In section 3, the corpus extraction task is explained. Then, in section 4, proposed target score and selected features are described. In section 5, learning algorithms considered and the ranking fusion technique are explained. In section 6, main experiments and results are presented. Finally, in section 7, main conclusions and future work are discussed.

2 Related Work

2.1 Extractive Summarization

The document summarization problem has evolved over time from a single document summarization task to a multiple summarization task where the summary is created from a set of documents about a same subject. Another variation of the problem is the topic focused summarization, where the summary must consider information related with a given topic. All these tasks have been continuously supported by DUC and TAC conferences. A popular extractive summarization method is the centroid-based method proposed in [20]. In this method, sentences are scored based on sentence-level and inter-sentence features such as cluster centroids, position, tf-idf values, etc. MEAD[6] is a multi-document summarization system which implements this method. Graph-based centrality approaches TextRank and LexRank were proposed in [17,4]. The idea is to build a graph using sentences as vertexes and the relation between them as edges. Afterwards, a random walk is performed over the graph, obtaining therefore, a centrality score for each sentence within the cluster. While [17] uses weighted edges according to the number of words that sentences have in common, [4] uses unweighted edges according to a threshold to the cosine similarity between the sentences. The problem has also been addressed using supervised learning algorithms in [3,21], among other works. A support vector machine based ensemble approach is proposed in [3]. In that work, the problem is modeled using binary classification where each sentence is labeled as relevant or not relevant according to a

[6] `http://www.summarization.com/mead/`

score obtained from the reference summary. Then, an ensemble of support vector machines is trained using internal properties of sentences as features. In [12] a multilingual approach is proposed based on a linear combination of features using a genetic algorithm. The optimization problem consists of maximizing the ROUGE value between the generated and the reference summaries.

2.2 Summarization Evaluation Measures

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a text summarization evaluation package[7] that includes several metrics to determine the quality of a summary by comparing it to reference summaries created by humans [25]. The measures consider the number of overlapping units such as n-grams[8], word sequences, and word pairs. The effectiveness of ROUGE measures was assessed by comparing them with human evaluation judgments from the DUC competition. In this evaluation a high correlation between ROUGE measures and human judgments was observed [13]. The main ROUGE measures are presented below:

- **ROUGE-N**: is an n-gram recall measure of a generated summary sum and a list of given reference summaries REF computed as follows:

$$\text{ROUGE-N}(sum, REF) = \frac{\sum_{ref \in REF} \sum_{gram_n \in ref} Count_{match(sum,ref)}(gram_n)}{\sum_{ref \in REF} \sum_{gram_n \in ref} Count(gram_n)}$$

  In this expression, n is the length of $gram_n$, and $Count_{match}(gram_n)$ is the number of n-grams which co-occur in sum and ref.
- **ROUGE-S**: a Skip-bigram is a pair of words w_i, w_j, where w_i precedes w_j $(i < j)$ in the sentence they belong. **Skip-Bigram Co-Occurrence Statistics** measures the overlap of skip-bigrams between sum and reference summaries in REF. ROUGE-SN is a specification of ROUGE-S where the maximum skip distance is restricted to N $(j - i \leq N)$.

3 The Wikinews Corpus

In this section we describe how we extracted a corpus of document clusters related upon a common subject and their respective "distant" summaries from Wikinews. News events lead to the publication of many articles in different Web sources. Wikinews journalists gather these documents and write articles whose first sentences summarize the event covered on them. The Wikimedia Foundation provides public dumps[9] of its projects. In order to obtain the main text of the articles' sources, we used Boilerpipe Article Extractor[10], which allows

[7] http://berouge.com/
[8] A n-gram is a sequence of n contiguous terms from a string.
[9] http://dumps.wikimedia.org/
[10] http://code.google.com/p/boilerpipe/

the extraction of the main text from within an HTML environment. Main algorithms provided by the library are detailed in [10]. Using Wikinews dumps and Boilerpipe we created a training corpus from articles dated in 2010 and their sources. Likewise, we created a testing corpus from articles dated in 2011[11]. The documents were split into sentences using the sentence detector provided by the OpenNLP[12] library. In order to reduce the noise in our data, all sources written in languages other than English were discarded. Moreover, articles of which more than the 30% of their sources were discarded, were not included in the datasets. The main characteristics of the datasets are detailed in Table 1.

Table 1. Corpus properties

	Train	Test
Number of Articles	886	546
Number of Sources	2,523	1,840
Article Sentences	13,644	9,570
Source Sentences	62,675	59,941

4 Distant Supervision and Features

In this section, we explain how the extracted corpus $\mathcal{C}$ is converted into a dataset to be used for supervised learning algorithms. Each entry $e_i \in \mathcal{C}$, is a tuple with the form $\langle art_i, src_{(i,1)}, src_{(i,2)}, \ldots, src_{(i,N)} \rangle$, where art_i is an article from Wikinews and $src_{(i,j)}$ is the j-th source cited by the i-th article from the corpus. Furthermore, each article art_i is composed by a sequence of sentences denoted by $\langle s_i^1, \ldots, s_i^k, \ldots, s_i^n \rangle$ where k represents the position of the sentence within the article. Likewise, source documents $src_{(i,j)}$ sentences are denoted as $s_{(i,j)}^z$, z being the position of the sentence within the document.

The main idea, is to extract for all source sentences $s_{(i,j)}^z$ in the corpus a vector of features which are independent of the content provided by art_i on one side and a label value dependent of the article on the other. Feature values will be used as independent variables in order to predict the label value for unseen sentences from document clusters where a reference summary is not given. Proposed label and features are described in the following subsections.

4.1 Zipf-Like Distant Label

The label of a source sentence $s_{(i,j)}^z$ is a score that represents the likelihood of the sentence as being part of a manual summary. The idea behind this is that

[11] The training and testing corpus can be downloaded from: `http://lahuen.dcc.uchile.cl/ mmanriquez/papers/Multi-DocumentSummarization/corpus.tar.gz`

[12] `http://incubator.apache.org/opennlp/`

high scored sentences are more adequate to be included in the target summary, because they are strongly related with a human-made summary. In [21] the score was computed as the average ROUGE-1 value between the source sentence and all sentences of the given reference summary. Nevertheless, in our case, we have no real summaries of the document clusters. In fact, we just know that the article and the sources cover the same event, and that first sentences of art_i summarize it. By the remainder of this work, according to the empirical properties of summaries described in [6] and the properties of Wikinews articles, we will consider that the first 5 sentences of the Wikinews article summarize the content of their sources. Nevertheless, we also hypothesize that all sentences of the article provide "relevant" information for summarization, but not in the same manner.

The Zipf law [26] has been used to describe the distribution of term frequencies within document collections. If f denotes the frequency of a word within a corpus r denotes its relative rank, then f and r are related as $f = \frac{c}{r^\alpha}$, where c is a constant and $\alpha > 0$. The α parameter controls how the frequency declines with ranking. If $\alpha = 1$, then f follows exactly the Zipf law, otherwise, is it said to be Zipf-like. Zipf-like distributions have been used to model a broad variety of phenomena, including the number of links in a web page [16]. Moreover, in [2], Zipf-Like factors were used in a metasearch engine ranking function. That shows that the Zipf law can also be included as part of a scoring function, and hence we strongly believe that this law can also be used in our "distant" labeling problem. Considering that first sentences of Wikinews articles summarize the content of their sources with more probability than the latter sentences, we state the following hypothesis: The probability that a sentence of a Wikinews article summarizes the content of the sources cited by the article, follows a Zipf-like distribution of the sentence position. Then, let be $sent(art_i)$ the number of sentences of art_i, we propose to score source sentences using their ROUGE similarities with the article sentences together with a Zipf-like weighting factor that considers the position of the article sentence:

$$score(s^z_{(i,j)}) = \frac{1}{sent(art_i)} \sum_k \frac{1}{k^\alpha} \times \text{ROUGE-1}(s^z_{(i,j)}, s^k_i) \tag{1}$$

The α value regulates how the relevance of the ROUGE similarity between the article sentence and the source sentence declines with the position of the article sentence, where its optimal value has to be found experimentally.

4.2 Features

The idea of extracting features from the sentences of a document is to identify properties of the sentences that can be used to predict the target variable described above. We are assuming therefore, that sentences provide some information that is independent of a reference summary which could be used to discriminate between relevant and non relevant sentences for inclusion into a summary. In this work, in order to achieve a multilingual summarizer, features

do not rely on the document language. The proposed features are presented in the following:

- **Position**: the position of the sentence within the document.
- **AvgFreq**: is the average term frequency of all sentence words.
- **AvgInvSentFreq**: is the average inverted sentence frequency (isf) of the words in the sentence [19] defined by following expression:

$$1 - \frac{log(sentnum(term))}{log(\sum_j sent(src_{(i,j)}))} \tag{2}$$

  where $sentnum(term)$ is the number of sentences containing the $term$.
- **CosDocSim**: represents the cosine similarity from the vector space model between the sentence and the document cluster.
- **ROUGENcomp**: is the average n-gram overlap between the sentence and rest of the cluster sentences. We used unigram and bigram overlaps as features.
- **ROUGEScomp**: is the average skip bigram overlap between s and the rest of the sentences. The maximum skip distance was set to 3.
- **LevenshteinComp**: is the average Levenshtein distance [11] value between the sentence and the rest of the sentences in the cluster. The Levensthein distance between two strings str_1 and str_2 is the minimum number of edits needed to transform str_1 into str_2. This value is normalized by the average length of both strings.

5 Learning to Rank Sentences

In this section we describe the learning algorithms selected for this work. It is important to consider, that in our problem we are more interested in learning to rank sentences according to a score, than in predicting the real value of the target variable. The problem of predicting the ordinal scale of a target variable is known as **Learning to Rank** in information retrieval problems [7]. The first method selected is a linear regression, which is a point-wise approach. In this approach, the features from the training dataset are expressed by a matrix X and the target variable is expressed by a column vector y. The target function is fitted using ordinary least squares by minimizing the residuals of the data ($w = (X^T X)^{-1} X^T y$). Resulting weight vector w can be used as a score function for ranking. The second one, is a ranking SVM, which is a pair-wise approach formulated in [8]. The idea is to learn a function $h(x)$ so that for any pair of examples (x_i, y_i) and (x_j, y_j) $h(x_i) > h(x_j) \Leftrightarrow y_i > y_j$. Ranking SVM algorithms were used for summarization purposes in [23]. Moreover, an efficient algorithm for training ranking SVMs was proposed in [9], whose implementation is included in the *SVM-light* package[13]. In addition to the linear regression and the ranking SVM functions, we developed a third ranking function based on the combination

[13] `http://svmlight.joachims.org/`

of the others using a data fusion approach. The idea of combining different multi-summarization methods was proposed in [24], were the overall quality of the summaries was improved. In this work we opted for Borda Count. Borda count is a ranking fusion method discussed in [1]. The model is based on democratic voting, where sentences of the cluster are considered as candidates and ranking functions as voters. For each ranking function a sentence is given $c - (r - 1)$ points where r is the local ranking and c is maximum numbers of sentences for a document within the collection. Afterwards, points from sentences are added. Finally, sentences are ranked in order of total points.

6 Experiments and Evaluation

6.1 Finding the Optimal Zipf-Like Distant Label

Before performing the learning procedure, we needed to find the optimal α value from the distant label (section 4.1). We applied equation 1 to all source sentences within the testing dataset using different values of α. For each value of α, the top 5 scored sentences were concatenated and matched against the first 5 sentences of the respective article from Wikinews. The extracted summary created using information provided by the Wikinews article, gives us an idea of the best possible summary that could be generated using a sentence extraction approach. The idea is to find the α value which maximizes the quality of the summary generated by extraction. We used ROUGE-1, ROUGE-2 and ROUGE-S as evaluation measures.

Table 2. Average ROUGE scores obtained using different α values

α	ROUGE-1	ROUGE-2	ROUGE-S
0.0	0.5278	0.1771	0.1137
0.25	0.5379	0.1842	0.1179
0.5	0.5472	0.1917	0.1222
1.0	**0.5525**	**0.1972**	0.1276
1.5	0.5444	0.1921	**0.1288**
2.0	0.5347	0.1841	0.1254

From Table 2, we can observe that when $\alpha = 1.0$, the best overall quality of the summary created by extraction is obtained. Therefore, we can conclude that the level of "useful information for summarization" provided by a sentence of a Wikinews article, follows a Zipf law of the sentence position.

6.2 Feature Selection and Learning

Using the proposed features (section 4.2) and the optimal distant label, we created a matrix of data from the training corpus. This learning matrix has the

sentences from all document clusters as rows and has the extracted features together with the distant label as columns. Moreover, feature values are scaled using a min-max normalization. A second matrix was obtained by applying the same process to the testing corpus. The idea is to train learning algorithms described in section 5 over the training matrix and apply resulting ranking functions to the testing data in order to create summaries of the clusters.

In several supervised learning algorithms factors like the presence of features which are strongly related to each other, or which not provide relevant information to predict the target variable, can affect the performance of the learned model. Feature selection is the task of identifying the best subset of variables within the training dataset for the learning purpose. In this work, the nature of the learning problem is ordinal. Therefore, selection techniques to be considered differ from methods used for binary classification or regression. Inspired by the work in [5], we performed a greedy feature selection algorithm in order to obtain different feature subsets. In the algorithm, two measures are used. The first is an importance score for each feature, and the second is a similarity measure between each feature pair. We used the absolute value of Kendall's τ coefficient between the features and the target variable as relevance scores, and the absolute value of Spearman's correlation coefficient between feature pairs as the similarity measure. Both coefficients are carried out on the ranks of the data. While Kendall's τ compares the number of concordant pairs and discordant pairs between two vectors, Spearman ρ is the Pearson's correlation between the ranking of two variables. The greedy selection algorithm proceeds to extract continuously the feature with the highest importance score available unless it has a high similarity to one of the previously selected. This task is repeat until there are no more high-scored features to select. We initialized the algorithm with different starting high-scored features and created several feature subsets. For each feature subset, we trained a linear regression and a ranking SVM over the training dataset. The optimal C parameter for the ranking SVM was found using a grid search, evaluating Kendall's τ value between the distant label and the predicted value over the testing dataset.

6.3 Summary Generation and Evaluation

Once several ranking functions were trained over the training dataset using different subsets of features, we proceeded to create summaries for all document clusters within the testing corpus. The summaries were created by the concatenation of top 5 ranked sentences obtained from each ranking approach. The quality of obtained summaries was evaluated using the same measures as the ones used for finding the optimal Zipf-value. Afterwards, we selected the best two resulting functions, which were a linear regression and a ranking SVM. Then, from these selected functions we created new summaries using Borda count. As baseline we implemented the LexRank algorithm [4] using the PageRank implementation provided by the Jung Framework[14]. The edges were created using the cosine similarity between the sentences with a threshold of 0.5.

[14] `http://jung.sourceforge.net/`

Table 3. Summarization Performance

Ranking Approach	ROUGE-1	ROUGE-2	ROUGE-S
LexRank	0.3521	0.1061	0.1103
Best Linear Reg.	0.4255	0.1277	0.1215
Best Ranking SVM	0.4076	0.1253	0.1195
Borda	**0.4310**	**0.1333**	**0.1248**

Average ROUGE measures obtained from the LexRank algorithm, the selected linear regression, the selected ranking SVM, and Borda count are shown in Table 3. From the table we can see that learned models greatly exceed the performance of the non-supervised LexRank algorithm. Furthermore, learned models produce summaries with very acceptable ROUGE scores in comparison to those obtained using the information provided by the Wikinews article in Table 2. Another interesting observation is that the linear regression achieves a better overall performance than the ranking SVM. In addition, Borda count achieves the best performance for all evaluation criteria. Thus, we can conclude that the combination of different sentence ranking techniques enhances the overall quality of the generated summaries.

7 Conclusions and Future Work

The main contribution of this work is a new distant supervision method for multi-document summarization where the knowledge required for the learning process is obtained directly from Web sources. The Zipf-like distant label proposed, was experimentally proved to be appropriate for news summarization, showing that the "useful information for summarization" of a Wikinews article sentence follows a Zipf law of the position. Furthermore, due the fact that proposed features are independent of the language, our model could be easily extended to other languages. We also showed that Borda count helps to enhance the quality of the generated summaries.

As future work, in order to improve the quality of the model and obtain a better generalization, a larger training dataset should be used in the training task. This dataset could be obtained using more articles from Wikinews or other sources having similar conditions. Furthermore, more features and learning to rank algorithms could also be considered. The model could be extended to a topic oriented summarization approach by using a query-dependent sentence ranking method and including it in the proposed ranking fusion framework.

Acknowledgment. Authors would like to thank Alejandro Figueroa for his valuable comments and suggestions. This work has been partially supported by FONDEF-CONICYT project D09I1185. The first author is granted by CONICYT Master's Scholarship.

References

1. Aslam, J.A., Montague, M.: Models for metasearch. In: SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 276–284. ACM, New York (2001)
2. Bravo-Marquez, F., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Hypergeometric Language Model and Zipf-Like Scoring Function for Web Document Similarity Retrieval. In: Chavez, E., Lonardi, S. (eds.) SPIRE 2010. LNCS, vol. 6393, pp. 303–308. Springer, Heidelberg (2010)
3. Chali, Y., Hasan, S.A., Joty, S.R.: A SVM-Based Ensemble Approach to Multi-Document Summarization. In: Gao, Y., Japkowicz, N. (eds.) AI 2009. LNCS, vol. 5549, pp. 199–202. Springer, Heidelberg (2009)
4. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. 22(1), 457–479 (2004)
5. Geng, X., Liu, T.-Y., Qin, T., Li, H.: Feature selection for ranking. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 407–414. ACM, New York (2007)
6. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: Summarizing text documents: sentence selection and evaluation metrics. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 121–128. ACM, New York (1999)
7. He, C., Wang, C., Zhong, Y.-X., Li, R.-F.: A survey on learning to rank. In: 2008 International Conference on Machine Learning and Cybernetics, pp. 1734–1739. IEEE (July 2008)
8. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. MIT Press, Cambridge (2000)
9. Joachims, T.: Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 217–226. ACM, New York (2006)
10. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010, pp. 441–450. ACM, New York (2010)
11. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707 (1966)
12. Litvak, M., Last, M., Friedman, M.: A new approach to improving multilingual summarization using a genetic algorithm. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 927–936. Association for Computational Linguistics, Stroudsburg (2010)
13. Liu, F., Liu, Y.: Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. In: Proceedings of ACL 2008: HLT, Short Papers, pp. 201–204. Association for Computational Linguistics, Columbus (2008)
14. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165 (1958)
15. Mani, I.: Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
16. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
17. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Proceedings of EMNLP 2004 and the 2004 Conference on Empirical Methods in Natural Language Processing (July 2004)

18. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2, ACL 2009, pp. 1003–1011. Association for Computational Linguistics, Stroudsburg (2009)
19. Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Generating Text Summaries through the Relative Importance of Topics. In: Monard, M.C., Sichman, J.S. (eds.) SBIA 2000 and IBERAMIA 2000. LNCS (LNAI), vol. 1952, pp. 300–309. Springer, Heidelberg (2000)
20. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, NAACL-ANLP-AutoSum 2000, pp. 21–30. Association for Computational Linguistics, Stroudsburg (2000)
21. Fisher, S., Roark, B.: Feature expansion for query-focused supervised sentence ranking. In: Document Understanding (DUC 2007) Workshop Papers and Agenda (2007)
22. Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 299–306. ACM, New York (2008)
23. Wang, C., Jing, F., Zhang, L., Zhang, H.-J.: Learning query-biased web page summarization. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, pp. 555–562. ACM, New York (2007)
24. Wang, D., Li, T.: Many are better than one: improving multi-document summarization via weighted consensus. In: SIGIR, pp. 809–810 (2010)
25. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Workshop in Text Summarization, ACL, pp. 25–26 (2004)
26. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley (1949)