An Aspect-driven Random Walk Model for Topic-Focused Multi-Document Summarization

Yllias Chali, Sadid A. Hasan, and Kaisar Imam

University of Lethbridge
Lethbridge, AB, Canada
{chali,hasan,imam}@cs.uleth.ca

Abstract. Recently, there has been increased interest in topic-focused multi-document summarization where the task is to produce automatic summaries in response to a given topic or specific information requested by the user. In this paper, we incorporate a deeper semantic analysis of the source documents to select important concepts by using a predefined list of important aspects that act as a guide for selecting the most relevant sentences into the summaries. We exploit these aspects and build a novel methodology for topic-focused multi-document summarization that operates on a Markov chain tuned to extract the most important sentences by following a random walk paradigm. Our evaluations suggest that the augmentation of important aspects with the random walk model can raise the summary quality over the random walk model up to 19.22%.

Keywords: Topic-focused summarization, multi-document summarization, aspects, random walk model

1 Introduction

The main goal of topic-focused multi-document summarization is to create a summary from the given documents that can answer the need for information expressed in the topic. We consider the problem of producing extraction-based[1] topic-focused multi-document summaries given a collection of relevant documents. To generate the summaries, we focus on a deeper semantic analysis of the source documents instead of relying only on document word frequencies to select important concepts. We use a predefined list of important aspects to direct our search for the most relevant sentences, and generate topic-focused summaries that cover all these aspects. For example, a topic about *Natural Disasters* might consider the aspects: *what happened; date; location; reasons for the disaster; casualties; damages; rescue efforts etc.* while generating the summary.

In this paper, we propose a novel topic-focused multi-document summarization framework that operates on a Markov chain model and follows a random

[1] Extract summaries contain original sentences extracted from the documents whereas abstract summaries can employ paraphrasing [8].

walk paradigm in order to generate possible summary sentences. We build three alternative systems for summary generation that are based on important aspects, random walk model, and a combination of both. We run our experiments on the TAC[2]-2010, and DUC[3]-2006 data and based on the evaluation results we argue that augmenting important aspects with a random walk model often outperforms the other two alternatives. Contributions of this work are: a) constructing an aspect-based summarization model that generates summaries based on given important aspects about the topics, b) building a novel summarization model based on a random walk paradigm that operates on a Markov chain exploiting topic signature [6] and Rhetorical Structure Theory (RST) [9] as node weights and WordNet[4]-based sentence similarities as edge weights, and c) generating a hybrid summarization model combining the aspect-based model with the random walk approach. Extensive automatic evaluations suggest that the combined model can raise the performance up to 19.22% while manual evaluations further confirm this improvement. The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 describes our three alternative summary generation models, Section 4 shows the evaluation results, and finally, Section 5 concludes the paper.

2 Related Works and Motivation

Although the task of topic-focused summarization has got a lot of attention recently (TAC-2010), the task is not new. A topic-sensitive LexRank is proposed in [12], where the set of sentences in a document cluster is represented as a graph. In this graph, the nodes are sentences and links between the nodes are induced by a similarity relation between the sentences. A substantial body of work on summarization using Information Extraction (IE) templates have been accomplished over the years in the Message Understanding Conferences (MUC[5]), DUC-2004 biography-related summarization task[6], as well as TREC[7]. In [15], they discuss the use of MUC templates for summarization. In [16], the authors define several biographical facts that should be included into a good biography. Filatova et al. [3] automatically create templates for several domains and use summarization-like task to evaluate the quality of the created templates. All the templates and facts are used in these researches to generate more focused summaries. Nastase [11] expands the query by using encyclopedic knowledge in Wikipedia and use the topic expanded words with activated nodes in the graph to produce an extractive summary. New features such as topic signature are used in the NeATS system by Lin and Hovy [7] to select important content from a set of documents about some topic to present them in coherent order.

[2] http://www.nist.gov/tac/2010/

[3] http://duc.nist.gov/

[4] http://wordnet.princeton.edu/

[5] http://www-nlpir.nist.gov/related_projects/muc/proceedings/ie_task.html

[6] http://duc.nist.gov/duc2004/

[7] http://trec.nist.gov/

An enhanced discourse-based summarization framework by rhetorical parsing tuning is proposed by Marcu [10]. In our work, we exploit topic signature and rhetorical structure theory [9] to weight the sentences. In [4], they introduced a paradigm for producing summary-length answers to complex questions. Their method operates on a Markov chain, by following a random walk with mixture model on a bipartite graph of relations established between concepts related to the topic of a complex question and subquestions derived from topic-relevant passages. Motivated by all these related researches, we propose to augment a predefined list of important aspects (that provides a better coverage of the topic on the entire document collection) into a random walk framework that no other study has used before to the best of our knowledge.

3 Our Approaches

In this section, we give a detailed description of all the three models that we build for the task of topic-focused multi-document summarization. Our first model is solely based on aspect information, while the second follows a novel random walk framework, and the third model is the aspect-driven random walk approach that combines the intuitions of the first two models. We get a candidate summary from each of the model at the end of the summary generation procedure. Therefore, three models give us three candidate summaries for the same given topic. Figure 1 presents the overall architecture of our systems.

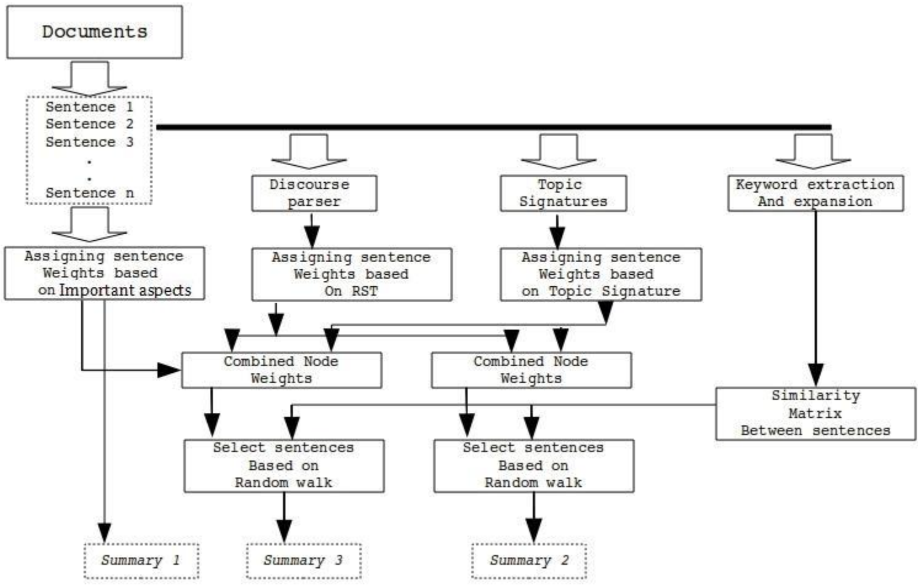
Documents
Sentence 1
Sentence 2
Sentence 3
.
.
Sentence n
Discourse parser
Topic Signatures
Keyword extraction And expansion
Assigning sentence Weights based on Important aspects
Assigning sentence Weights based On RST
Assigning sentence Weights based on Topic Signature
Combined Node Weights
Combined Node Weights
Similarity Matrix Between sentences
Select sentences Based on Random walk
Select sentences Based on Random walk
Summary 1
Summary 3
Summary 2

Fig. 1. The overall architecture of our approaches

3.1 Aspect–based Model

Our first approach exploits the predefined list of important aspects to find the most relevant sentences from the document collection for creating the summaries. For each question (i.e. aspect) of a topic, we did keyword expansion using WordNet[8] [2]. For example, the word "happen" being a keyword in the given *aspect*: "What happened?" returns the words: *occur, pass, fall out, come about, take place* from WordNet. On the other hand, for each document sentence in the collection we perform Named Entity (NE) tagging using the *OAK* system [13]. Named Entities (NE) are defined as terms that refer to a certain entity. For instance, *USA* refers to a certain *country*, and *$200* refers to a certain quantity of money. OAK system has 150 named entities (such as PERSON, LOCATION, ORGANIZATION, GPE (Geo-Political Entity), FACILITY, DATE, MONEY, PERCENT, TIME etc.) that can be tagged. They are included in a hierarchy. We weight each sentence based on the presence of one or more Named Entity classes. We rank the document sentences based on the following two criteria:

1. Similarity of each sentence with the expanded aspect (in terms of word matching), and
2. weight assigned to each sentence by the NE tagging procedure[9].

Finally, we select the top-ranked sentences to be included in the candidate summary (Summary 1 in Figure 1).

3.2 Random Walk Model

To include into our second candidate summary, we select the most relevant sentences by following a random walk on a graph where each node is a document sentence and the edges represent similarity between sentences. The whole procedure operates on a Markov chain (MC). A Markov chain is a process that consists of a finite number of states and some known probabilities p_{ij}, where p_{ij} is the probability of moving from state j to state i. For each node (i.e. sentence) and each edge in the graph, we calculate *"node weight"* and *"edge weight"*, respectively. Once we find all the node weights and edge weights, we perform a random walk on the graph following a Markov chain model in order to select the most important sentences. The initial sentence is chosen simply based on the node (sentence) weights using the following formula:

$$InitialSentence = \arg\max_{i=1}^{N} (weight\,(S_i)) \tag{1}$$

where N is the total number of nodes in the graph. After finding the initial best sentence, in each step of the random walk we calculate the probability (transition probability) of choosing the next relevant sentence based on the following equation:

[8] For simplicity, we consider the synsets up to level 1 in this research.

[9] For example, for an aspect like *"When did the accident happened?"*, we search for $< Time >$ tag in the NE tagged sentences and give them higher weights if found.

$$P(S_j|S_i) = \frac{1}{\alpha} \arg\max_{j=1}^{Z} \left(weight\left(S_j\right) * similarity\left(S_i, S_j\right) \right) \tag{2}$$

where S_i is the sentence chosen early, S_j is the next sentence to be chosen, Z is the set of sentence indexes that does not contain i, the $similarity(S_i, S_j)$ function returns a similarity score between the already selected sentence and a new sentence under consideration, and α is the normalization factor that is determined as follows:

$$\alpha = \sum_{j=1}^{Z} \left(weight\left(S_j\right) * similarity\left(S_i, S_j\right) \right) \tag{3}$$

Node Weight We associate each node (sentence) in the graph a weight that indicates the importance of the node with respect to the document collection. Node weights are calculated based on a Topic Signature (TS) model [6] and Rhetorical Structure Theory (RST) [9]. We combine the weights of TS and RST, and normalize it to get the final weights of the sentences/nodes.

Using Topic Signature Topic signatures are typically used to identify the presence of a complex concept–a concept that consists of several related components in fixed relationships [6]. Inspired by the idea presented in [6], for each topic present in the data set, we calculate its topic signature defined as below:

$$\begin{aligned} TS &= \{topic, signature\} \\ &= \{topic, \langle (t_1, w_1), \cdots, (t_n, w_n) \rangle \} \end{aligned} \tag{4}$$

where *topic* is the target concept and signature is a vector of related terms. Each t_i is a term highly correlated to the *topic* with association weight, w_i. We use the following log-likelihood ratio to calculate the weights associated with each term (i.e. word) of a sentence:

$$w_i = log \frac{occurrences\, of\, t_i\, in\, topic\, j\ sentences}{occurrences\, of\, t_i\, in\, all\, topics'\ sentences} \tag{5}$$

To calculate the topic signature weight for each sentence, we sum up the weights of the words in that sentence and then, normalized the weights. Thus, a sentence gets a high score if it has a set of terms that are highly correlated with a target concept (topic).

Exploiting Rhetorical Structure Theory (RST) Rhetorical Structure Theory provides a framework to analyze text coherence by defining a set of structural relations to composing units ("spans") of text. The most frequent structural pattern in RST is that two spans of text are related such that one of them has a specific role relative to the other. A paradigm case is a claim followed by evidence for the claim. RST assumes an "Evidence" relation between the two spans that is denoted by calling the claim span a *nucleus* and the evidence span a *satellite*[10]. In this paper, we parse each document sentence within the framework of Rhetorical Structure Theory (RST) using a Support Vector Machine (SVM)-based discourse parser described in [1] that was shown 5% to 12% more accurate than current state-of-the-art parsers. We observe that in a relation the nucleus often contains the main information while the satellite provides some additional information. Therefore, we assign a weight to each sentence that is a nucleus of a relation and normalize the weights at the end.

Edge Weight Edge weight is determined by measuring similarity between the sentences. Initially, we remove the stopwords from the sentences using a stopword list. Then, we use the *OAK* system [13] to get the stemmed words of a sentence. We expand the remaining keywords of the sentence using WordNet. Finally, we find the similar words between each pair of sentences that denotes the edge weight between the two sentences. We build a similarity matrix by populating into it the edge weights between sentences.

3.3 Aspect-driven Random Walk Model

The third model that we construct to generate a candidate summary is based on augmentation of the predefined important aspects into the random walk framework. Motivated by Harabagiu et al. [4], where they describe how a random walk can be used to populate a network with potential decompositions of a complex question, we propose to use the list of aspects (given in TAC-2010) in the random walk model as a guided way to provide a better coverage to satisfy a wide range of information need on a given topic. Through out the rest of the paper, we term this model as a *Combined Model* since it combines the important aspects with the random walk paradigm. The whole procedure can be again formulated according to a Markov Chain principle described in Section 3.2 except the fact that the node(sentence) weights will also include the weights obtained by using the list of aspects' information as defined in Section 3.1. Figure 2 shows a part of the graph with node and edge weights (after applying the combined model) for the top ranking sentences that were chosen by the random walk. This is an example of a DUC-2006 topic outlined below.

```
<topic id = "D0626H"
category = "2">
<title> bombing of US embassies
in Africa  </title>
```

S1: Among them is Saudi dissident Osama bin Laden, who allegedly runs al Qaida, a radical Islamic network accused of planning the bombings.

[10] http://www.sfu.ca/rst/01intro/intro.html

S2: In an interview Tuesday, Home Affairs Minister Ali Ameir Mohamed likened Ahmed to a chameleon.

S3: It said Khalid, who can not speak English or Kiswahili but only Arabic, was identified by a guard and a civilian worker at the embassy and a third witness.

S4: Although no details were released in court, local media said traces of chemicals that could have been used to make the bomb had been found in Saleh's home and car.

S5: The action contrasted markedly to a decision by Kenya, where the American Embassy was bombed on the same day.

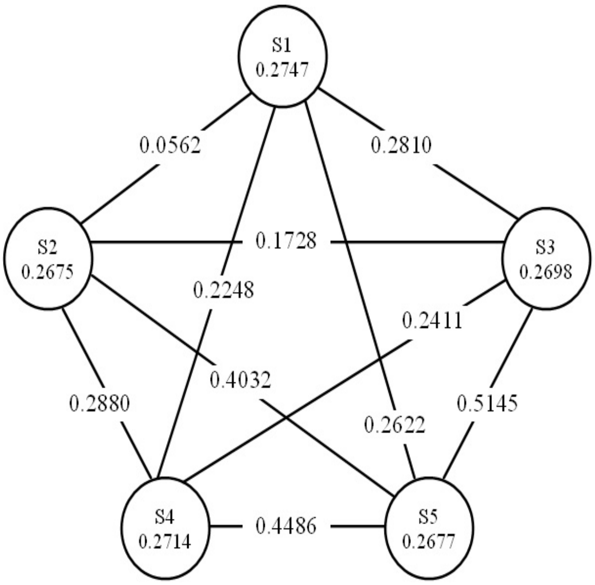

Fig. 2. Important aspects with random walk model

From Figure 2, we get to the fact that initially, sentence *S1* is chosen into the candidate summary as it has the highest node (sentence) weight, then, by performing a random walk based on the transition probabilities of the Markov chain model, we find *S2* as the next candidate sentence, then, *S3*, *S4*, *S5* and so on. The random walk stops after the k steps which is related to reaching the summary-length of 250 words.

4 Evaluation Results and Analyses

4.1 Task Description

TAC-2010 provides a new research direction for multi-document summarization by the means of predefined supervision or guide (the category and its aspects) that defines what information the reader is actually looking for. The task of DUC-2006 models the real-world complex question answering in terms of multi-document summarization. That is: *"Given a complex question (topic description) and a collection of relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic"*. In this paper, we consider a modified task description that induces the guided concept of TAC-2010 in order to automatically generate 250-word summaries like DUC-2006. Our summarization task can be defined as:

"To write a 250-word summary of a set of given newswire articles for a given topic, where the topic falls into a predefined category."

4.2 Data

In this research, we run our experiments using the TAC-2010 and DUC-2006 data applying three different models to generate three candidate summaries for each topic. The test dataset in TAC-2010 is composed of 44 topics, divided into five categories: Accidents and Natural Disasters, Attacks, Health and Safety, Endangered Resources, Investigations and Trials. In this paper, we consider only the first two categories[11]. As DUC-2006 data were not categorized, we manually categorize them to put into our chosen categories: Accidents and Natural Disasters, and Attacks. Since human-generated abstract summaries are not publicly available, we perform an extensive manual evaluation on the TAC-2010 data to report comparisons based on linguistic quality and responsiveness of the summaries. For DUC-2006 data, we obtain both an automatic[12] and a manual evaluation.

4.3 Automatic Evaluation

For the DUC-2006 data, we carried out automatic evaluation of our candidate summaries using ROUGE [5] toolkit, which has been widely adopted for automatic summarization evaluation. For all our systems, we report the widely accepted important metrics: ROUGE-2 and ROUGE-SU. We also present the ROUGE-1 scores since they provide a better correlation with the human judgement. We show the 95% confidence intervals of the important evaluation metrics for our systems to report significance for doing meaningful comparison. ROUGE uses a randomized method named bootstrap resampling to compute the confidence interval. We used 1000 sampling points in the bootstrap resampling. Table 1 to Table 3 show the ROUGE-1, ROUGE-2, and ROUGE-SU scores of our three different summary generation models.

[11] TAC provides already categorized data.

[12] Abstract summaries are available for comparisons.

Scores	Aspects	Random Walk	Combined
Recall	0.3488	0.3344	0.3624
Precision	0.3415	0.3604	0.3556
F-score	0.3444	0.3460	0.3587

Table 1. ROUGE-1 measures

Scores	Aspects	Random Walk	Combined
Recall	0.0711	0.0500	0.0633
Precision	0.0693	0.0545	0.0609
F-score	0.0701	0.0520	0.0620

Table 2. ROUGE-2 measures

For all the three systems, Table 4 shows the F-scores of the reported ROUGE measures while Table 5 reports the 95% confidence intervals of the important ROUGE measures. Table 4 clearly shows that the **Combined** system improves the ROUGE-1, ROUGE-2, and ROUGE-SU scores over the **Random walk** system by **3.67%**, **19.22%**, and **8.21%**, respectively, whereas, it could not beat the ROUGE-2 score of **Aspect**–based system but improves the ROUGE-1, and ROUGE-SU scores by **4.15%**, and **4.97%**, respectively. These results suggest that augmenting important aspects with the random walk model provides a better content coverage to satisfy the information need. The proposed methods are also compared with a *Baseline* system. The *Baseline* is the official baseline system established in DUC-2006 that generated the summaries by returning all the leading sentences (up to 250 words) in the $\langle TEXT \rangle$ field of the most recent document(s). We also list the average ROUGE scores of all the participating systems of DUC-2006 (i.e. *DUC-Average*). Table 6 presents this comparison which denotes that the **Combined** system improves the ROUGE-1, and ROUGE-2 scores over the **Baseline** system by **11.77%**, and **17.78%**, respectively, whereas, it performs closely to the average DUC-2006 systems.

4.4 Manual Evaluation

Even if the ROUGE scores come up promising, it might be possible to generate bad summaries that get state-of-the-art ROUGE scores [14]. So, we conduct an

Scores	Aspects	Random Walk	Combined
Recall	0.1159	0.1029	0.1211
Precision	0.1109	0.1182	0.1156
F-score	0.1123	0.1090	0.1179

Table 3. ROUGE-SU measures

Systems	ROUGE-1	ROUGE-2	ROUGE-SU
Aspects	0.3444	0.0701	0.1123
Random walk	0.3460	0.0520	0.1090
Combined	0.3587	0.0620	0.1179

Table 4. ROUGE F-scores for different systems

Systems	ROUGE-2	ROUGE-SU
Aspects	0.0569 - 0.0844	0.1053 - 0.1190
Random walk	0.0373 - 0.0682	0.0894 - 0.1262
Combined	0.0364 - 0.0879	0.0989 - 0.1363

Table 5. 95% confidence intervals for different systems

extensive manual evaluation in order to analyze the effectiveness of our approach. We judged the summaries for linguistic quality and overall responsiveness according to the DUC evaluation guidelines[13]. Table 7 and Table 8 presents the average linguistic quality and overall responsive scores of all the systems on TAC-2010 data and DUC-2006 data, respectively. To compare the proposed models' performance with the state-of-the-art systems, in Table 8 we also list the scores of *LCC's GISTexter*[14] system that participated in the DUC-2006 competition and was ranked as one of the best systems. Analyzing these results yields the fact that augmenting important aspects with the random walk model often outperforms the random walk model alone in terms of linguistic quality and responsiveness scores. Table 8 shows that the proposed aspect-driven random walk model (i.e. *Combined*) performs very close to LCC's system in terms of linguistic quality while considerably outperforming it in terms of overall responsiveness scores. This confirms that the use of the aspect information enhances the coverage of the information that is necessary to satisfy the quest of the users.

5 Conclusion

In this paper, we present a novel methodology to solve the topic-focused multi-document summarization task that uses a predefined list of important aspects in a random walk framework by performing a deeper semantic analysis of the source documents instead of relying only on document word frequencies to select important concepts. Experiments on the DUC-2006 and TAC-2010 data indicate that augmenting the important aspects into the random walk model considerably outperforms the random walk model alone. This suggests the fact that the aspects can provide a certain amount of supervision to cover all the relevant perspectives of a topic and hence, the use of it with any sophisticated model such as

[13] http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

[14] http://duc.nist.gov/pubs/2006papers/lcc2006.pdf

Systems	ROUGE-1	ROUGE-2
Aspects	0.3444	0.0701
Random walk	0.3460	0.0520
Combined	0.3587	0.0620
Baseline	0.3209	0.0526
DUC-Average	0.3778	0.0748

Table 6. Comparison with DUC-2006 systems

Systems	Lin. Quality	Responsiveness
Aspects	4.00	4.00
Random walk	3.60	3.00
Combined	4.00	3.00

Table 7. Linguistic quality and responsiveness scores (TAC-2010 data)

Systems	Lin. Quality	Responsiveness
Aspects	3.72	3.00
Random walk	3.52	3.00
Combined	3.76	3.20
LCC	4.10	2.84

Table 8. Linguistic quality and responsiveness scores (DUC-2006 data)

random walk can enhance the model's performance substantially in comparison to the model if used alone.

Acknowledgments

The research reported in this paper was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada – discovery grant and the University of Lethbridge.

References

1. duVerle, D.A., Prendinger, H.: A Novel Discourse Parser Based on Support Vector Machine Classification. 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 (ACL '09): Volume 2. pp. 665–673 (2009)
2. Fellbaum, C.: WordNet - An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
3. Filatova, E., Hatzivassiloglou, V., McKeown, K.: Automatic Creation of Domain Templates. In: Proceedings of the COLING/ACL on Main conference poster sessions. pp. 207–214. COLING-ACL '06 (2006)

4. Harabagiu, S., Lacatusu, F., Hickl, A.: Answering Complex Questions with Random Walk Models. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 220 – 227. ACM (2006)
5. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of Association for Computational Linguistics. pp. 74–81. Barcelona, Spain (2004)
6. Lin, C.Y., Hovy, E.H.: The Automated Acquisition of Topic Signatures for Text Summarization. In: Proceedings of the 18th conference on Computational linguistics. pp. 495–501 (2000)
7. Lin, C.Y., Hovy, E.H.: From Single to Multi-Document Summarization: A Prototype System and Its Evaluation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 457–464. Philadelphia (2002)
8. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press (1999)
9. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. In: Text. pp. 8(3): 243–281 (1988)
10. Marcu, D.: Improving Summarization Through Rhetorical Parsing Tuning. In: The Sixth Workshop on Very Large Corpora. pp. 206–215. Montreal, Canada (1998)
11. Nastase, V.: Topic-Driven Multi-Document Summarization with Encyclopedic Knowledge and Spreading Activation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08). pp. 763–772 (2008)
12. Otterbacher, J., Erkan, G., Radev, D.R.: Using Random Walks for Question-focused Sentence Retrieval. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 915–922. Vancouver, Canada (2005)
13. Sekine, S.: Proteus Project OAK System (English Sentence Analyzer), http://nlp.nyu.edu/oak. (2002)
14. Sjöbergh, J.: Older Versions of the ROUGEeval Summarization Evaluation System Were Easier to Fool. Information Processing and Management 43, 1500–1505 (2007)
15. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., Wagstaff, K.: Multidocument Summarization via Information Extraction. In: Proceedings of the First International Conference on Human Language Technology Research. pp. 1–7. HLT '01 (2001)
16. Zhou, L., Ticrea, M., Hovy, E.H.: Multi-document Biography Summarization. CoRR abs/cs/0501078 (2005)