

Development and Evaluation of a Multi-document Summarization Method Focusing on Research Concepts and Their Research Relationships

Shiyan Ou, Christopher S.G. Khoo, and Dion H. Goh

Division of Information Studies,
School of Communication & Information,
Nanyang Technological University,
Singapore, 637718
{pg00096125, assgkhoo, ashlgoh}@ntu.edu.sg

Abstract. This paper reports the design and evaluation of a method for summarizing a set of related research abstracts. This summarization method extracts research concepts and their research relationships from different abstracts, integrates the extracted information across abstracts, and presents the integrated information in a Web-based interface to generate a multi-document summary. This study focused on sociology dissertation abstracts, but can be extended to other research abstracts. The summarization method was evaluated in a user study to assess the quality and usefulness of the generated summaries in comparison to a sentence extraction method used in MEAD and a method that extracts only research objective sentences. The evaluation results indicated that the majority of sociology researchers preferred our variable-based summary generated with the use of a taxonomy.

1 Introduction

Multi-document summarization has begun to attract more and more attention in the last few years [6]. Different from single-document summarization, multi-document summarization is capable of condensing a set of related documents into one summary. It is more useful in digital libraries and Web search engines. A multi-document summary has several advantages over the single-document summary. It provides an overview of a topic indicating common information across many documents, unique information in each document, and cross-document relationships (relationships between pieces of information in different documents), and can allow users to zoom in for more details on aspects of interest.

The purpose of this study was to develop an automatic method to summarize a set of related sociology dissertation abstracts that may be retrieved by a digital library system or search engine in response to a user query. Recently, many digital libraries have begun to provide online dissertation abstract services, since dissertation abstracts contain a wealth of high-quality information by specifying research objectives, research methods and research results of dissertation projects. However, a dissertation abstract is relatively long, about 300~400 words, and browsing too many of such

abstracts can result in information overload. Therefore, it would be helpful to summarize a set of dissertation abstracts to assist users in grasping the main ideas on a specific topic.

The sociology domain was selected for the study because much of sociology research adopts the traditional quantitative research paradigm of looking for relationships between research concepts often operationalized as research variables. Sociology dissertation abstracts are also well-structured and have the classical research report structure with five standard sections - *background*, *research objectives*, *research methods*, *research results* and *concluding remarks* [9]. Many other domains such as psychology and medicine adopt this research paradigm and report structure.

The summarization method used in this study focuses on research concepts and their research relationships. Concepts are often operationalized as variables whose values vary. A relationship refers to the correspondence between two variables. A variable-based framework was developed to integrate research concepts and their research relationships extracted from different abstracts and thus summarize a set of dissertation abstracts on a specific topic [8]. The framework has a hierarchical structure in which the summarized information is in the top level and the more detailed information is found in lower levels. Based on the framework, an automatic summarization method for sociology dissertation abstracts was developed. The method extracts research concepts and their research relationships from different documents, integrates the extracted information across documents, and presents the integrated information using the variable-based framework. Although the summarization method was developed based on sociology dissertation abstracts, it also can be applied to other domains which adopt the same research paradigm of seeking to investigate research concepts and their relationships, and use a similar research report structure.

2 Review of Multi-document Summarization Approaches

In previous studies, the main approaches used for multi-document summarization include sentence extraction, template-based information extraction, and identification of similarities and differences between documents. With sentence extraction, documents or sentences across all the documents are clustered, following which, a small number of sentences are selected from each cluster and concatenated into a summary [1, 7, 13]. In order to generate more coherent summaries, lexical chains are sometimes considered for extracting internally linked sentences instead of separate sentences [2]. Some multi-document summarizers, such as SUMMONS [5], RIPTIDES [14] and GITEXTER [3], use information extraction techniques to extract pieces of information to fill in one or more pre-defined templates. Another important approach for multi-document summarization is to extract information that is common or repeated in several documents plus selected unique information in individual documents to generate the summaries [4]. In addition, cross-document rhetorical relationships are used to create multi-document summaries by extracting the sentences which have some specific rhetorical relations (e.g. *equivalence* or *contradiction*) among them [12, 15]. However, these existing summarization approaches focus more on physical granularities (words, phrases, sentences and paragraphs) and rhetorical relations based on shallow analysis, without paying much attention to higher-level semantic content

and semantic relations expressed within and across documents. Another problem is that different users have different information needs. Thus, an ideal multi-document summarization should provide different levels of detail for different aspects of the topic according to the user's interest. But these approaches usually construct fixed multi-document summaries.

Like most of previous multi-document summarization approaches, the method used in this study summarizes a set of related documents by identifying the similarities and differences among them. However, in the study, the identification of the similarities and differences is based more on semantic-level research concepts and their research relationships expressed in the text, instead of words, phrases or sentences themselves used in previous studies. To do that, the summarization method analyzes the macro-level (between sentences and segments) discourse structure peculiar to sociology dissertation abstracts to identify which segments of the text contain the more important information, and analyzes the micro-level (within sentences) to identify the specific kind of information to be extracted from specific segments, as well as analyze the cross-document structure to identify similar information, unique information, and relationships between pieces of information across documents and integrate them together using a variable-based framework.

3 The Summarization Method

The summarization method comprises five main steps: *data preprocessing*, *macro-level discourse parsing*, *information extraction*, *information integration*, and *information presentation*.

3.1 Data Preprocessing

The input files are a set of related dissertation abstracts on a specific topic retrieved from the Dissertation Abstracts International database indexed under *sociology* subject and *PhD* degree. Each file contains one dissertation abstract in HTML format. Each dissertation abstract was segmented into sentences using a simple algorithm. Then each sentence was parsed into a sequence of word tokens using the Conexor parser [11]. For each word token, its lemma (base form) and part-of-speech tag were indicated.

3.2 Macro-Level Discourse Parsing

In the macro-level discourse analysis, dissertation abstracts were parsed into five sections or categories - *background*, *research objectives*, *research methods*, *research results*, and *concluding remarks*. Each section comprises one or more sentences and contains a specific kind of information. *Research objectives* and *research results* sections are hypothesized to contain the more important information relating to the main ideas of the dissertation study. To parse the macro-level discourse structure automatically, a decision tree classifier was developed to assign each sentence in a dissertation abstract to one of the five categories or sections according to the sentence position in the document and the presence of indicator words in the sentence [9]. The categorization was improved using more reliable indicator phrases, such as "*The purpose of the*

study was to ...” found at the beginning of the sentences in the *research objectives* section, while “*The results indicate that ...*” found at the beginning of the sentences in the *research results* section.

3.3 Information Extraction from the Micro-Level Discourse Structure

In the micro-level discourse analysis, four kinds of information were extracted within sentences - *research concepts* and *their research relationships*, *research methods* and *contextual relations*.

At the linguistic level, research concepts, research methods and contextual relations appear as noun or noun phrases. A list of syntactic rules, specifying the possible sequences of part-of-speech tags in a noun phrase, was defined and used to identify sequences of contiguous words that are potential noun phrases. The terms relating to research methods and contextual relations were identified using indicator phrases. The terms relating to research concepts were selected from the *research objectives* and *research results* sections. To extract relationships between variables, linguistic patterns were constructed that a relationship pattern contains two or three slots and the concepts that match with the slots in the pattern represent the research variables connected by the relationship. Pattern matching was performed to identify the text segments in the sentences that match with each relationship pattern.

3.4 Information Integration Across Documents

In a set of related dissertation abstracts, the similarities and differences across different abstracts are mainly reflected through research concepts and their research relationships. Similar concepts were identified and clustered according to their syntactic structure. The terms of different word lengths which follow specific syntactic variation rules were considered term variants and represented similar concepts at different generalization levels, for example, “*abuse* -> *sexual abuse* -> *childhood sexual abuse* -> *survivor of childhood sexual abuse* -> *woman survivor of childhood sexual abuse* -> *adult woman survivor of childhood sexual abuse*”. An automatic integration method links shorter term variants to longer term variants from the single head word to a specific full term to form a hierarchical chain, and thus a group of similar concepts was obtained from the nodes of the chain. Concepts at the lower level can be generalized by the broader concepts at the higher level. The chains sharing the same root node are combined to form a hierarchical cluster tree which represented a cluster of similar concepts sharing the same cluster label. The concepts at the higher level in a cluster were selected and integrated together using a new sentence.

Research methods and contextual relations were identified using pre-defined indicator phrases and normalized using uniform terms. For example, “*qualitative design*” and “*qualitative study*” were normalized as “*qualitative research*”.

For a cluster of similar concepts, their relationships with other concepts were integrated together to provide an overview of all associated concepts connected by various types of relationships. Each type of relationship (e.g. correlation or cause-effect relationship) was identified using a group of patterns. For the same type of relationships, linguistic normalization was carried out to normalize different surface expressions using a standard expression and to conflate them. For example, “*school size is not significantly related with school crime rate*” and “*there is no relationship*

not significantly related with school crime rate” and *“there is no relationship between school size and school dropout rate”* were transformed and conflated into *“school size is not related with school crime rate and school dropout rate”*.

3.5 Information Presentation

The four kinds of information - *research concepts and their research relationships, research methods and contextual relations*, were combined and reformulated for presentation in a Web-based interface to generate an interactive summary viewable through a Web browser. The interface presents the combined information at three hierarchical levels which are connected through hyperlinks: (1) the summarized information at the top level; (2) the specific information extracted from individual dissertation abstracts at the second level; and (3) the original dissertation abstract at the third level. The hierarchical structure of the interface allows users to explore details of interest by clicking on hyperlinks rather than viewing traditional plain text summaries.

The screenshot shows a web browser window titled "SYSTEM 1 - Microsoft Internet Explorer" displaying a summary page for "intercultural communication". The page content includes:

- A legend: "* Number in the brackets indicates the number of documents."
- Section 1: "In these 64 dissertation abstracts, the following contextual relations were found:" followed by a list of terms and counts (e.g., theory(22), context(18), perspective(18), model(15), perception(15), framework(13), view(8), insight(7), attitude(5), assumption(5), hypothesis(5), thought(2)).
- Section 2: "In these 64 dissertation abstracts, the following research methods were found:" followed by a list of methods and counts (e.g., interview(24), observation(13), qualitative research(5), survey(6), content analysis(5), case study(4), experiment(4), fieldwork(4), scale(4), comparative research(3), descriptive research(3), ethnographic research(2), phenomenological research(2), quantitative research(2), regression analysis(2), archival research(1), correlational research(1), cross-cultural research(1), discourse analysis(1), grounded theory research(1), rhetorical analysis(1), statistical analysis(1), textual analysis(1)).
- Section 3: "These 64 dissertation abstracts were mainly about:" followed by a list of main concepts:
 - communication(60)**, including intercultural communication(36), cross-cultural communication(21), host communication(2), communication and identity(2), and more ...
 - Different aspects were investigated, including communication competence(5), communication model(4), communication literature(3), communication theory(3), communication behavior(2), communication problem(1), communication skill(2), communication strategy(2), and more ...
 - The following relationships were investigated:
 - There was an effect on adaptation, individual well-being among racially/ethnic groups.
 - It was affected by values, lack of intention of stay in the United States, Australia, attempt to assimilate, bicultural identity, TOEFL.
 - There was a relation with parents with high school students in special education, psychological health of the workers, Americans, psychological health, power motive, general themes, behaviors.
 - There may be an relation with leadership, respondents' level of tolerance of ambiguity.
 - There was no relation with individual's level of psychological health in an American-German workplace, managerial control mechanisms.
 - There were differences found with own classroom practices with students.
 - There may be differences found with communities.
 - There were no differences found between Malaysian-malays and African-Americans.
 - culture(31)**, including language and culture(2), American culture(1), and more ...

Overlaid on the page is a hierarchical diagram with three levels:

- A main concept**: A box containing "communication(60)".
- A subclass concept**: A box containing "cross-cultural communication(21)".
- A facet concept**: A box containing "communication competence(5)".

Arrows indicate the relationships: "A main concept" points to "A subclass concept" and "A facet concept". "A subclass concept" points to "A facet concept".

Fig. 1. The SYSTEM 1 summary generated without the use of taxonomy on “intercultural communication”

The summarized information is displayed in the main window as the main summary while the other two hierarchies are displayed separately in pop-up windows. In the main window, the clustered and summarized research methods, contextual relations, research concepts and their research relationships extracted from different

documents, are combined based on the variable-based framework. There are two types of main summaries – (1) SYSTEM 1 generated without the use of the taxonomy (see Figure 1); and (2) SYSTEM 2 generated with the use of the taxonomy (see Figure 2). The function of the taxonomy is to remove non-concept terms, highlight the important sociology concepts, and categorize main concepts into different subjects.

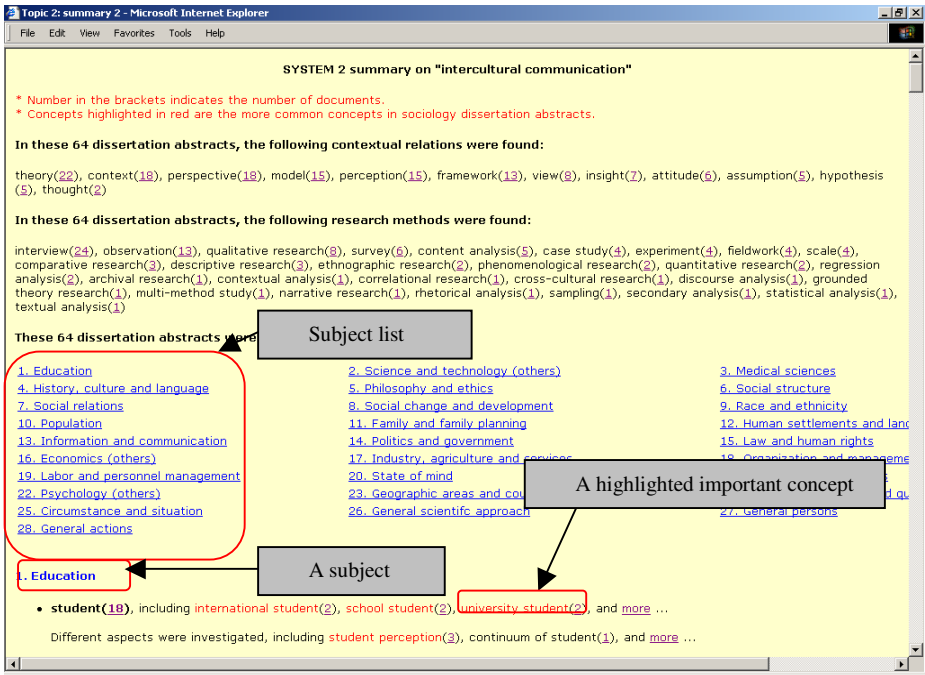


Fig. 2. The SYSTEM 2 summary generated with the use of taxonomy on “intercultural communication”

For each concept, the number of documents is given in parenthesis. This is clickable and links to a list of summarized single documents sharing the given concept in a pop-up window. For each document, the title, research concepts, research methods and contextual relations are displayed. The title of the document is also clickable and links to the original dissertation abstract in a separate pop-up window.

4 Evaluation of the Summarization Method

The overall quality and usefulness of the final summaries were assessed intrinsically in a user study. The users were asked to subjectively judge the quality of the summaries and their usefulness for research-related purposes, by comparing the summaries generated using our summarization method with or without the use of the taxonomy

against a summary generated by MEAD using a sentence extraction method and a summary generated by extracting research objective sentences only.

4.1 Evaluation Design

20 research topics were obtained from 20 researchers in the field of sociology, who were Master's or PhD research students or faculty members at Nanyang Technological University, Singapore, and National University of Singapore. Each researcher was asked to submit one research topic that he/she was working on or had worked on. For each topic, a set of PhD sociology dissertation abstracts were retrieved from the Dissertation Abstracts International database using the topic as the search query, but at most 200 abstracts were retained. The set of dissertation abstracts retrieved for each topic was condensed into a summary. Four different summaries were provided for each topic with two kinds of structures – (1) variable-based summaries, and (2) sentence-based summaries. The four types of summaries were:

- *A variable-based summary generated without the use of the taxonomy.* It focuses on research concepts and their research relationships, as well as research methods and context relations. This type of summary was labeled *SYSTEM 1* (see Figure 1).
- *A variable-based summary generated with the use of the taxonomy.* It also focuses on research concepts and their research relationships, as well as research methods and contextual relations. Furthermore, based on the taxonomy, non-concept terms were filtered out, important sociology concepts were highlighted in red, and concept clusters were categorized into different subjects. This type of summary was labeled *SYSTEM 2* (see Figure 2).
- *A sentence-based summary generated by extracting research objectives of each abstract.* It consists of sentences that are research objectives extracted from each dissertation abstract. The type of summary was labeled *OBJECTIVES*.
- *A sentence-based summary generated by a peer system.* It consists of sentences that were ranked as important, according to certain sentence features, in the set of dissertation abstracts. It was created by a multi-document summarization system MEAD 3.08, which uses a centroid-based sentence extraction method [13]. This type of summary was labeled *MEAD*.

The four types of summaries were constructed using the same compression rate of 20% in terms of the number of the words. For each topic, the four types of summaries were compared by human subjects on two aspects: (1) quality of the summaries including readability and comprehensibility; (2) usefulness of the summaries for research-related purposes.

4.2 Evaluation Results

The overall quality (readability and comprehensibility) and usefulness of the four types of summaries were scored by the human subjects on a 7-point scale. The average scores for the four types of summaries from 20 researchers are shown in Table 1.

(1) *Quality (readability and comprehensibility)*

SYSTEM 2 obtained the second highest readability and comprehensibility score (5.2 and 5.1) among the four types of summaries. It was rated better than SYSTEM 1 (4.4

Table 1. Average scores for the overall quality (readability and comprehensibility) and usefulness of the four types of summaries

		SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
Quality	Readability	4.40	5.20	5.70	5.00
	Comprehensibility	4.75	5.10	5.60	4.95
Usefulness		5.00	5.70	5.65	4.9

and 4.75), indicating that with the use of a taxonomy for information filtering and organization, the quality of the variable-based summary was substantially improved. SYSTEM 2 was rated better than the set of important sentences in MEAD (5.0 and 4.95), but still worse than the research objective sentences in OBJECTIVES (5.7 and 5.6). There was a significant difference in the average readability score between SYSTEM 1 and SYSTEM 2 ($p=0.008$).

For readability, the researchers indicated that SYSTEM 1 & 2 were more concise and contain less vacuous or general information than OBJECTIVES and MEAD. This is because SYSTEM 1 & 2 consist of important concepts and simple relationship sentences whereas OBJECTIVES and MEAD consist of complete sentences. On the other hand, the researchers indicated that SYSTEM 1 & 2 contain more duplicate information and dangling anaphors, and are less fluent than OBJECTIVES and MEAD. This is because a concept can be assigned to more than one cluster from difference perspectives. Moreover, separate concepts are less fluent than complete sentences.

For comprehensibility, the researchers indicated that OBJECTIVES and MEAD are a little easier to understand than SYSTEM 1 & 2. This is because complete sentences are easier to understand than separate concepts. Furthermore, the researchers indicated that the research objective sentences in OBJECTIVES can indicate the main ideas of the topic very well.

(2) Usefulness

For research-related purposes, SYSTEM 2 obtained similar usefulness score (5.7) as the research objective sentences in OBJECTIVES (5.65), but was rated much better than the set of the general important sentences in MEAD (4.9) and SYSTEM 1 (5.0). This indicates that with the use of the taxonomy for information organization, the usefulness of the variable-based summary was improved. In addition, the research objective sentences was rated much better than the set of the general importance sentences, indicating that the researchers were more concerned about research objectives than other kinds of information in a dissertation.

The researchers indicated that SYSTEM 1 & 2 were more useful in indicating similarities among previous studies, important concepts and research methods used in the area. OBJECTIVES and MEAD were more useful in identifying the documents of interest easily and indicating important theories, views or ideas in the area.

The four types of summaries were ranked by 20 researchers. A weighted rank score was calculated for each summary: a weight of 4 was assigned to the first rank, 3 for the second rank, 2 for the third rank, and 1 for the fourth rank. The researchers were also asked to select one or more summaries that they preferred to use for their

Table 2. Ranking and preference for the four types of summaries

Rank	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
No.1 (weight=4)	3 (15%)	11 (55%)	6 (30%)	0
No.2 (weight=3)	5 (25%)	2 (10%)	7 (35%)	6 (30%)
No.3 (weight=2)	5 (25%)	6 (30%)	5 (25%)	4 (20%)
No.4 (weight=1)	7 (35%)	1 (5%)	2 (10%)	10 (50%)
Weighted rank score	2.15	3.15	2.85	1.8
Preference	6 (30%)	14 (70%)	11 (55%)	5 (25%)

research-related work. The ranking and the researchers' preference for the four types of summaries are summarized in Table 2.

According to the weighted rank scores, the final ranking of the four types of summaries is: (1) SYSTEM 2, (2) OBJECTIVES, (3) SYSTEM 1 and (4) MEAD. SYSTEM 2 obtained the first rank among the four types of summaries. The highest percentage of the researchers (70%) indicated preference for SYSTEM 2 for their research-related work, and 55% of the researchers indicated preference for OBJECTIVES.

The researchers indicated that the variable-based summaries were more effective in providing an overview of a topic and can help researchers find similar information easily. However, they were too brief to provide accurate information and sometimes confused the users. On the other hand, the sentence-based summaries can provide more direct information and were easy to understand. But it was time-consuming to read them and hard to locate the relevant information.

5 Conclusion and Future Work

This paper has reported the development and evaluation of an automatic method for summarizing a set of sociology dissertation abstracts. Our system focuses on extracting research concepts and their research relationships from each document, integrating the extracted information across documents, and presenting the integrated information in an interactive Web interface.

A user study was carried out to evaluate the quality (readability and comprehensibility) and usefulness of the summaries using a questionnaire. In the variable-based summary generated with the use of a taxonomy, non-concept terms were filtered out, concepts were categorized into different subjects, and important sociology concepts were highlighted. The evaluation results demonstrated that this kind of summary was more readable, comprehensible, and useful than the one generated without the use of the taxonomy. It ranked higher than the research objective summary and the MEAD summary. The majority of the sociology researchers in the study (70%) indicated preference for the variable-based summary generated with the use of a taxonomy, and 55% indicated preference for the research objective summary.

References

1. Boros, E., Kanto, P.B., & Neu, D.J.: A clustering based approach to creating multi-document summaries. *Document Understanding Conferences* (2002). Available at http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/rutgers_final.pdf
2. Brunn, M., Chali, Y., & Dufour, B. (2002). The University of Lethbridge text summarizer at DUC 2002. *Document Understanding Conferences*. Retrieved May 19, 2003 from http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/lethbridge_chali.pdf.
3. Harabagiu, S.M., & Lacatusu, F.: Generating single and multi-document summaries with GISTEXTER. *Document Understanding Conferences* (2002). Available at http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/utdallas_sanda.pdf
4. Mani, I., & Bloedorn, E.: Summarization similarities and differences among related documents. *Information Retrieval*, 1(1) (1999) 1-23
5. Mckeown, K., & Radev, D.: Generating summaries of multiple news articles. In *Proceedings of the 13th Annual International ACM Conference on Research and Development in Information Retrieval (ACM SIGIR)*. Seattle, WA (1995) 74-82
6. National Institute of Standards and Technology. : *Document Understanding Conferences* (2002). Available at <http://www-nlpir.nist.gov/projects/duc/index.html>
7. Otterbacher, J.C., Winkel, A.J., & Radev, D.R.: The Michigan single and multi-document summarizer for DUC 2002. *Document Understanding Conferences* (2002) Available at http://www-nlpir.nist.gov/projects/duc/pubs/2002papers/umich_otter.pdf
8. Ou, S., Khoo, C., & Goh, D.: Multi-document summarization of dissertation abstracts using a variable-based framework. In *Proceedings of the 66th Annual Meeting of the American Society for Information Science and Technology (ASIST)*. Long Beach, CA, 19-23 October (2003) 230-239
9. Ou, S., Khoo, C., Goh, D., & Heng, Hui-Hing. : Automatic discourse parsing of sociology dissertation abstracts as sentence categorization. In *Proceedings of the 8th International ISKO Conference*. London, UK, 13-16 July (2004) 345-350
10. Ou, S., Khoo, C., & Goh, D.: A multi-document summarization system for sociology dissertation abstracts: design, implementation and evaluation. In *Proceedings of the 9th Conference on Research and Advanced Technology for Digital Libraries (ECDL)*. Vienna, Austria, 18-23 September (2005) in press
11. Pasi, J. & Timo, J.: A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington, DD: Association for Computational Linguistics(1997) 64-71
12. Radev, D.: A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue* (2000). Available at <http://www.sigdial.org/sigdialworkshop/proceedings/radev.pdf>
13. Radev, D., Jing, H., & Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Workshop held with Applied Natural Language Processing Conference / Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/ANNCL)* (2000) 21-29
14. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., & Wagstaff, K.: Multi-document summarization via information extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT-01)* (2001)
15. Zhang, Z., Blair-Goldensohn, S., & Radev, D.: Towards CST-enhanced summarization. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002)*. Edmonton, Canada, August (2002)