

Automatic Multi-document Summarization of Research Abstracts: Design and User Evaluation

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

Division of Information Studies, School of Communication & Information, Nanyang Technological University, Singapore, 637718

E-mail: ou_shiyan@pmail.ntu.edu.sg, {assgkhoo, ashlgoh}@ntu.edu.sg

Telephone: 65-67906971

Fax: 65-67915214

Corresponding author: Shiyan Ou (ou_shiyan@pmail.ntu.edu.sg)

Abstract

The purpose of this study was to develop a method for automatic construction of multi-document summaries of sets of research abstracts that may be retrieved by a digital library or search engine in response to a user query. Sociology dissertation abstracts were selected as the sample domain in this study. A variable-based framework was proposed for integrating and organizing *research concepts* and *relationships* as well as *research methods* and *contextual relations* extracted from different dissertation abstracts. Based on the framework, a new summarization method was developed, which parses the discourse structure of abstracts, extracts research concepts and relationships, integrates the information across different abstracts, and organizes and presents them in a Web-based interface. A user evaluation was performed to assess the overall quality and usefulness of the summaries. Two types of variable-based summaries generated using the summarization method – with or without the use of a taxonomy – were compared against a sentence-based summary that only lists the research objective sentences extracted from each abstract and another sentence-based summary generated using the MEAD system that extracts important sentences. The evaluation results indicated that the majority of sociological researchers (70%) and general user (64%) preferred the variable-based summaries generated with the use of the taxonomy.

Introduction

Automatic summarization has attracted attention both in the research community and commercially as a solution for reducing information overload and helping users to scan a large number of documents to identify documents of interest (Mani & Maybury, 1999). It is an important function that should be available in large digital libraries and information retrieval systems (e.g. search engines), where the retrieval of too many documents and the resulting information overload is a major problem for users. While single-document summarization is a well-developed field, especially in the use of sentence extraction techniques, multi-document summarization has begun to attract attention only in the last few years (DUC, 2002). Multi-document summarization is capable of condensing a set of related documents, rather than a single document, into one summary. Thus a multi-document summary has several features different from the single-document summary and sometimes is more useful than the latter in digital libraries and search engines. It provides a domain overview of a topic indicating what is similar and different across different documents and relationships between pieces of information in different documents, and allows users to zoom in for more details on aspects of interest. Furthermore, in academic digital libraries, it can be used for knowledge discovery to identify connections between research results that are not obvious and gaps in the field for future research.

The purpose of this study was to develop a method for automatic construction of multi-document summaries of sets of research abstracts that may be retrieved by a digital library or search engine in response to a user query. Dissertation abstracts in sociology domain were selected as sample documents. There is increasing interest in constructing digital libraries of dissertations (Moxley, 2001), because there is a ready supply of student dissertations in universalities. However, the access to full-text dissertations is often restricted by institutional policies and copyright concerns, whereas dissertation abstracts are often freely available (e.g. ProQuest Digital Dissertations¹). A dissertation abstract is a high-quality informative research abstract providing substantial information on the research objectives, research methods and results of dissertation projects. It is relatively long (about 300-400 words), and browsing too many of such abstracts result in information overload. Therefore, it is helpful to summarize a set of dissertation abstracts to assist users in grasping the central ideas in the groups of

¹ The website of ProQuest Digital Dissertations is at <http://wwwlib.umi.com/dissertations/gateway>.

research studies and the relations between the different studies. Dissertation abstracts have a relatively clear and standard structure. The language is more formal and standardized than in other corpora, e.g. news articles, and is thus easier to process using current natural language processing techniques. The sociology domain was selected because of its clear micro-level discourse structure focusing on research concepts and relationships. Much of sociology research adopts the traditional quantitative research paradigm of looking for relationships between research concepts usually operationalized as variables. Although some studies adopt a qualitative research paradigm, many of them also seek to identify relationships between concepts representing events, behaviors, attributes, and situations.

In this study, we did not use traditional sentence extraction approaches. Instead, a hybrid summarization method involving both extractive and abstractive techniques was developed. This method focused on extracting and integrating similarities and differences across different documents to summarize a set of related documents. However, the identification of similarities and differences was based more on semantic contents and semantic relations (i.e. meaningful research concepts and relationships) expressed in the text, rather than words, phrases or sentences and their rhetorical relations used in previous studies (e.g. Mani & Bloedorn, 1999; Zhang, Blair-Goldensohn, & Radev, 2002). To do that, the macro-level discourse structure (between sentences and segments) peculiar to sociology dissertation abstracts was analyzed to identify which segments of the text contain more important research information. Then the micro-level discourse structure (within sentences) was analyzed to identify which kinds of information could be extracted from specific segments. Moreover, the cross-document structure was analyzed to identify similar information, unique information, and relationships between pieces of information across different dissertation abstracts. A variable-based framework was proposed to integrate and organize similarities and differences among different dissertation abstracts, focusing on research concepts and relationships investigated in different dissertation studies. Based on the variable-based framework, similar information was integrated across different dissertation abstracts, and different kinds of information were combined, organized, and presented in a Web-based interface to generate an interactive summary viewable through a Web browser.

A user evaluation involving 20 sociological researchers and 40 general users was carried out to assess the overall quality and usefulness of the summaries. This is the focus of this paper. 20 research topics, obtained from the researchers in the field of sociology, were used in the evaluation. For each topic, two types of variable-based summaries – generated with and without the use of a taxonomy – were compared against a sentence-based summary that lists only the research objective sentences extracted from each dissertation abstract and another sentence-based summary generated using a state-of-the-art system MEAD that extracts important sentences based on a variety of criteria.

Review of Multi-document Summarization and Evaluation Approaches

Multi-document summarization approaches can be broadly divided into extractive and abstractive approaches. The most extensively used extractive approach is the statistics-based sentence extraction. Since the important sentences are extracted from different documents, the resulting multi-document summaries often lack cohesion and coherence and contain more redundant information. To reduce the redundancy, the summarization system MEAD (Radev, Jing & Budzikowska, 2000), XDoX (Hardy, Shimizu, Strzalkowski, Ting, Wise, & Zhang, 2002), and MultiGen (McKeown, Klavans, Hatzivassiloglou, Barzilay, & Eskin, 1999) clustered similar documents or sentences and selected representative sentences from each cluster to form a summary. Carbonell and Goldstein (1998) used Maximal Marginal Relevance (MMR) to minimize redundancy and maximize diversity among extracted sentences. To produce more readable and coherent multi-document summaries, cohesive links, such as lexical chain (Brunn, Chali & Dufour, 2002; Chali & Kolla, 2004) and co-reference (Bergler et al., 2004), were used to identify internally related sentences across documents.

In contrast to extractive approaches, abstractive approaches are more complicated to implement because they require extensive domain knowledge to interpret source texts and generate new texts. But they can produce more coherent summaries and obtain a higher compression rate. Thus abstractive approaches seem more appropriate for multi-document summarization (Afantenos, Karkaletsis, & Stamatopoulos, 2005). However, real abstractive approaches that completely imitate human abstracting behavior are difficult to achieve with current natural language processing techniques (Goldstein, Kantrowitz, Mittal, & Carbonell, 1999). Current abstractive approaches are in reality hybrid approaches involving both extractive and abstractive techniques, i.e. extracting important information based on statistical and linguistic features or pre-defined templates and integrating similarities and differences among the extracted information. Multi-document summarization mainly focuses on the similarities and differences across documents. The MultiGen summarizer (McKeown et al., 1999) identified

the common phrases among a set of similar sentences and used them to reformulate a new sentence as the component of the summary. Mani and Bloedorn (1999) extracted a set of sentences or fragments (pieces of sentences) containing similar words or phrases and a set of sentences or fragments containing different words or phrases from different documents and concatenated them into a summary covering both similarities and differences. Some multi-document summarizers, such as SUMMONS (McKeown & Radev, 1995), RIPTIDES (White, Korelsky, Cardie, Ng, Pierce, & Wagstaff, 2001) and GITEXTER (Harabagiu & Lacatusu, 2002), used template-based information extraction techniques to extract pieces of information to fill in one or more pre-defined templates and combined the instantiated slots based on the similarities and differences between them to generate fluent sentences. In addition, some studies made use of cross-document rhetorical relations to identify the similarities and differences between text units in different documents and synthesized them to generate the summary. Zhang, Blair-Goldensohn, and Radev (2002) extracted the sentences that have *contradiction* and *equivalence* relations to improve the quality of the sentence-based summary. Afantenos, Doura, Kapellou, and Karkaletsis (2004) connected the topic-specific templates using rhetoric relations such as *identity*, *elaboration*, *contradiction*, *equivalence*, to create the summary focusing on the similarities and differences between news sources.

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. Current approaches usually construct fixed multi-document summaries.

After building a summarization system, it is important to evaluate its effectiveness and usefulness. Evaluation methods can be divided into two types – *intrinsic* evaluation and *extrinsic* evaluation (Jones & Galliers, 1996). In previous studies, intrinsic evaluation was used extensively. A few intrinsic evaluations were performed by asking human assessors to judge the quality of summaries directly according to some criteria, such as *grammaticality*, *cohesion*, *coherence*, *organization*, and *coverage of key ideas* (e.g. Minel, Nugier, & Piat, 1997; Saggion, Radev, Teufel, Lam, & Strassel, 2002). However, most intrinsic evaluations were carried out by asking human assessors to compare a machine-generated summary against one or more human reference summaries (e.g. Edmunson 1969; Kupiec Pedersen, & Chen, 1995). In comparison to single-document summaries, the evaluation of multi-document summaries is more difficult. Currently, there is no widely accepted procedure or methodology for evaluating multi-document summaries (Schlesinger, Conroy, Okurowski, & O'Leary, 2003; Radev, Jing, & Budzikowska, 2004). The existing types of tasks for extrinsic evaluations (e.g. relevance assessment task, categorization task) are mainly for evaluating single-document summaries, and not appropriate for evaluating multi-document summaries. Moreover, it is more difficult to obtain uniform reference summaries, since human differed greatly from each other in summarizing multiple documents (Schlesinger et al., 2003).

The user evaluation used in the study was mainly intrinsic. The summarization method was evaluated at two levels: (1) accuracy and usefulness of each important summarization step; and (2) overall quality and usefulness of the summaries. The evaluation of each summarization step has been reported by Ou et al. (2005b). In the evaluation, the quality of the system-generated summaries was judged directly by human subjects, rather than compared against "ideal" summaries. This is because it is time-consuming and mentally strenuous work for human abstractors to create reference multi-document summaries. It is like writing a literature review. The usefulness of the summaries was also judged directly in this study, rather than based on a specific task. It is difficult to design a reasonable task which can reflect real-world applications to evaluate the usefulness of the summary accurately.

A Variable-based Framework for Multi-document Summarization

With the macro-level discourse analysis, most of dissertation abstracts (almost 85%) were found to have a clear structure – their sentences could be subsumed under five sections – *background*, *research objectives*, *research methods*, *research results* and *concluding remarks* (Ou, Khoo & Goh, 2002). Each section comprises one or more sentences and contains a specific kind of information. Although some abstracts do not contain all the five sections and a few sentences do not belong to any section, the *research objectives* and *research results* sections are clearly discernable in most dissertation abstracts. These two sections were hypothesized to contain more important research information. Although a small percentage of abstracts (about 15%) are hard to segment into the five sections, the *research objectives* section is nevertheless clearly discernable and also hypothesized to contain more important research information.

With the micro-level discourse analysis, it was found that four kinds of information, i.e. *research concepts* and *relationships*, *contextual relations*, *research methods*, can be extracted from each dissertation abstract. In sociological research, concepts are used to represent elements of society and human social behavior (Macionis, 2000). Much of sociological research focuses on research concepts and relationships (Macionis, 2000). A research relationship refers to the correspondence between two research variables that are investigated in a sociological study (Trochim, 1999). Quantitative research usually seeks to investigate relationships between research concepts often operationalized as research variables. In contrast, qualitative research usually focuses on describing and explaining human behaviors or social phenomena and thus does not operationalize concepts as variables. However, many of qualitative studies also seek to identify relationships between concepts representing events, behaviors, attributes, and situations. Thus, sociology research can be divided into three types (Trochim, 1999):

- **Descriptive research:** one or more target concepts are investigated to identify attributes of interest.
- **Relational research:** two or more variables are investigated at the same time to see if there is any relationship between them;
- **Causal research:** one or more variables are manipulated by the researcher to see how they cause or affect one or more outcome variables;

In causal research, one or more variables which the researchers are interested in explaining or predicting are designated as the *dependent variables* (DVs), whereas another group of variables that affect or are used to predict the dependent variables are designated as the *independent variables* (IVs). In relational research however, variables are not distinguished as dependent and independent variables. Sometimes, a third variable, known as *mediator variable* or *moderator variable* (Baron & Kenny, 1986), comes in between the two variables.

While many studies aim to explore relationships directly, some studies explore relationships in the context of a *framework*, *model*, *theory*, *hypothesis* etc., or in the *perception* or *attitude* of a target population. For example, “The purpose of this qualitative, descriptive study was to examine *mothers’ perception* of how their children are affected by exposure to domestic violence.” We call this a *contextual relation*. In a research study, certain concepts and relationships are investigated using one or more research methods. With conceptual analysis, three types of information, *research design*, *sampling*, and *data measurement & analysis*, were usually found in the *research methods* section of the dissertation abstracts. In addition, research methods are sometimes mentioned casually in the *research objectives* and *research results* sections.

In a set of dissertation abstracts on a specific topic, the same or similar concepts are often investigated in different dissertation projects – maybe focusing on relationships with different concepts, using different research methods, and in different contexts or from different perspectives. These similar concepts can be generalized using a common broader concept. To analyze the cross-document discourse structure of a set of dissertation abstracts, we focused on research concepts and relationships to identify what is similar information and unique information, and how the similar and unique information is linked in different dissertation abstracts. Thus a variable-based framework was proposed to present research concepts and relationships as well as contextual relations and research methods in a set of related dissertation abstracts, and thus to integrate and organize similar and different information across different dissertation abstracts to summarize a set of dissertation abstracts. The framework contains four kinds of information as follows:

- **Main concepts:** The common research concepts, often operationalized as research variables, investigated by most of the dissertation abstracts in a document set.
- **Research relationships between concepts:** For each main concept, the descriptive attribute values or relationships with other concepts (e.g. correlations and cause-effect relationships) investigated in different dissertation abstracts.
- **Contextual relations:** Concepts and relationships in the perception, attitude, insight, etc. of a target population, or in the context, framework, model, theory, etc.
- **Research methods:** One or more research methods used to explore the attributes of concepts and relationships, including research design, sampling, and data measurement & analysis method.

In the framework, the central elements are the *main concepts* that were investigated by most of the dissertation abstracts in a document set. Each kind of information is integrated across different dissertation abstracts, and then the four kinds of information are combined and organized based on the *main concepts*. It has a hierarchical structure in which the summarized information is given at the top level and the more detailed information is at the lower levels. Similar concepts extracted from different dissertation abstracts are clustered together and summarized by a broader concept called *main concept*. For a specific concept, its attribute values or research relationship with other concept(s) are given together with the contextual relations and research methods used in the dissertation studies. All the relationships involving the same *main concept* are combined together and

summarized using a simple, standard expression. The contextual relations and research methods are summarized using simple, uniform terms. Figure 1 shows some of the information which was extracted from 10 dissertation abstracts on the topic of “school crime” and integrated and organized using the variable-based framework.

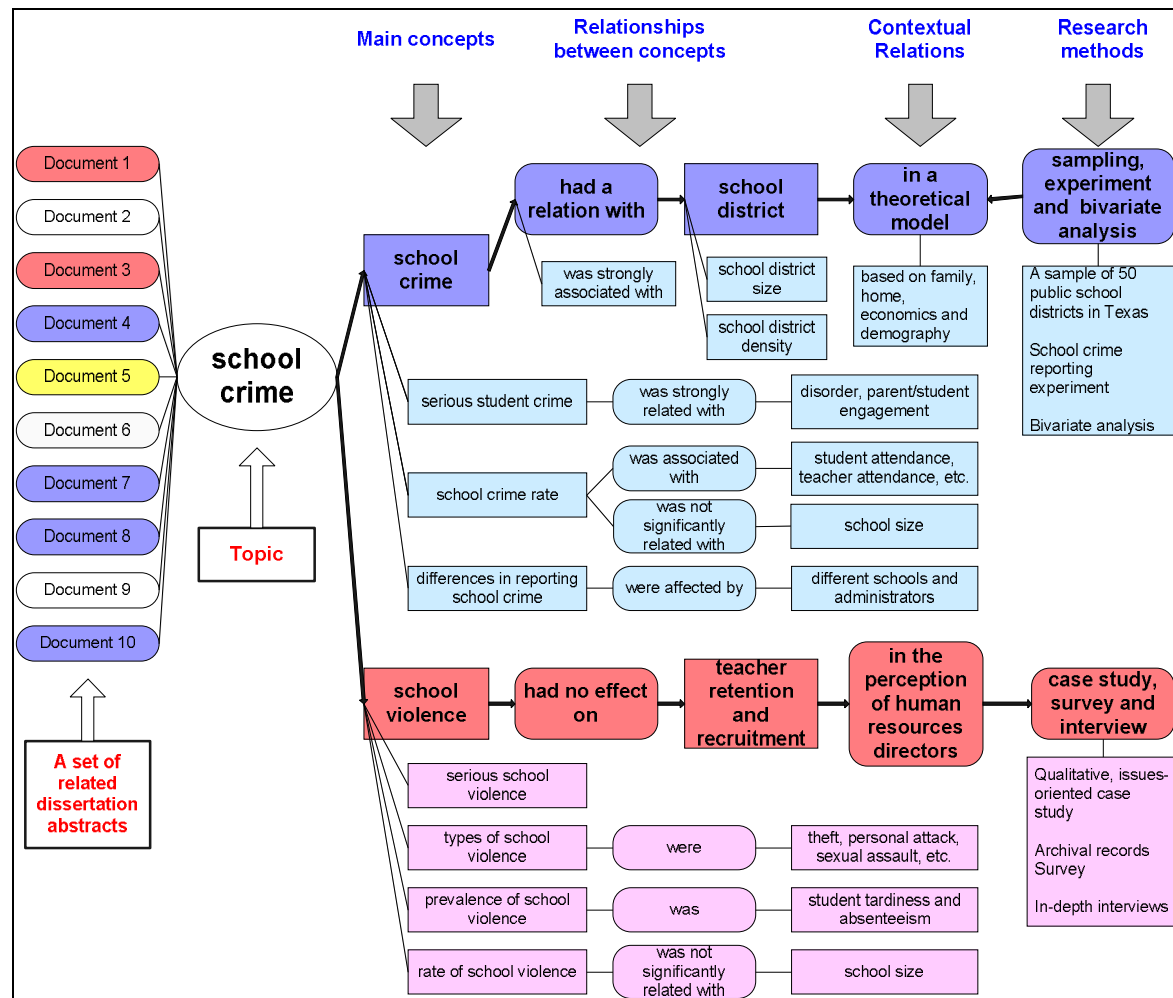


FIG. 1. Integrated and organized information across 10 dissertation abstracts on the topic of “school crime” using the variable-based framework.

In Figure 1, two main concepts – “school crime” and “school violence” – are presented. “School crime” was investigated in five dissertation abstracts. It includes three sub-level concepts, which represent different subclasses of the parent concept or specify different aspects or characteristics of the parent concept – (1) *serious student crime*; (2) *school crime rate*; and (3) *differences in reporting school crime*.

- For “school crime”, the following relationships were investigated:
 - *School crime* was strongly associated with school district including school district size and density.
 - *Differences in reporting school crime* were affected by different schools and administrators.
 - *School size* was not significantly related with *rate of school crime*.
 - *Serious student crime* was strongly related with disorder, parent/student engagement.
 - *School crime rate* was associated with student attendance, teacher attendance, etc.
- The above five relationships associated with “school crime” could be integrated as follows:

-
- “School crime” was related with school district, disorder, parent/student engagement, student attendance, teacher attendance etc.
 - “School crime” was not related with school size.
 - “School crime” was affected by different schools and administrators.

Some relationships were investigated based on a contextual relation and using some research methods. For example, the relationship between “school crime” and “school district” was investigated *in a theoretical model based on family, home, economics and demography*, using the following research methods – (1) *a sample of public school districts in Texas, summarized as “sampling”*; (2) *school crime reporting experiment, summarized as “experiment”*; and (3) *Bivariate analysis*.

The framework presents a full map of a specific topic by integrating research concepts and relationships as well as contextual relations and research methods extracted from different dissertation abstracts using a hierarchical structure and organizing them based on the main concepts. It has two advantages: giving an overview of a subject area by presenting the summarized information at the top level; and also allowing users to zoom in to more details of interest by exploring the specific information at the lower levels. The framework provides a way to summarize a set of dissertation abstracts that is different from the traditional sentence extraction methods.

The Summarization Process

Based on the variable-based framework, a summarization method was developed for constructing multi-document summaries of sets of sociology dissertation abstracts. This method include four major summarization steps: (1) parsed each dissertation abstracts to identify which sections contain more important research information; (2) extracted four kinds of important information from each dissertation abstract; (3) integrated each kind of information extracted from different dissertation abstracts; and (4) combined and organized the four kinds of information, and presented them in an interactive Web-based interface.

The input files are a set of related dissertation abstracts on a specific topic retrieved from the Dissertation Abstracts International database indexed under *Sociology* and *PhD degree*. Each file contains one dissertation abstract in HTML format and then was transformed into a uniform XML representation. The abstract text was segmented into sentences using a short list of end-of-sentence punctuation marks (e.g. “.”, “?”, “!”). Each sentence was parsed into a sequence of word tokens using the Conexor Parser (Pasi & Timo, 1997), indicating each token’s document number, sentence number, token number (word position in the sentence), word form (the real form used in the text), base form (aka lemma), and part-of-speech tag for the subsequent summarization steps.

Macro-level Discourse Parsing

In discourse parsing, each dissertation abstract was segmented into the five or some of the five sections. In the study, we treated discourse parsing as a text categorization problem – assigning each sentence in a dissertation abstract to one of the five sections or categories. A decision tree classifier that made use of sentence position and indicator words present in the sentence was developed and applied to sentence categorization (Ou, Khoo, & Goh, 2004). Furthermore, cue phrases found at the beginning of some sentences in the *research objectives* and *research results* sections were used to improve the decision tree classification of these two categories or sections. For example, “*The purpose of this study was to investigate*” often occurs at the beginning of the first sentence in the *research objectives* section and is more reliable to help identify a research objective sentence than the single indicator words and sentence position used in the decision tree classifier. In the evaluation, a low average accuracy of 63.4% was obtained in categorizing all the sentences, but a high accuracy of 90.8% obtained in identifying *research objectives* and *research results* sentences.

Information Extraction from the micro-level discourse structure

In information extraction, four kinds of important information – *research concepts* and *relationships*, *contextual relations*, and *research methods* – were extracted from the micro-level structure (i.e. sentence level) of each dissertation abstract. At the linguistic level, research concepts, contextual relations and research methods appear as nouns 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 were potential noun phrases. Research concept terms were selected from the *research objectives* and *research results* sections, since these two sections are most likely to contain more important research information. Research method and contextual relation terms were identified using indicator phrases.

To extract research relationships between concepts, linguistic patterns indicating various kinds of relationships were constructed manually based on a sample of 300 dissertation abstracts. The linguistic patterns

used are regular expression patterns, each comprising a sequence of tokens with two or three slots. Each token which is not a slot is constrained with a part-of-speech tag. For example, “[slot: independent variable] have DET (ADJ) effect/influence/impact on/in [slot: dependent variable]” is a pattern describing one way in which cause-effect relationship can be expressed in the text. A pattern matching program was developed to identify the text segments in the sentences that matched with each relationship pattern. The terms in the text that matched with the slots in a pattern represent the research variables connected by the relationship.

Information Integration across documents

In information integration, similar concepts and relationships extracted from different dissertation abstracts were integrated using concept generalization and relationship conflation.

Similar concepts were identified and clustered according to their syntactic variations. 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,

- *student* → *undergraduate student* → *black undergraduate student* → *adjustment of black undergraduate student* → *college adjustment of black undergraduate student*
- *student* → *behavior of student* → *delinquent behavior of student* → *prevention of delinquent behavior of student*

These hierarchical term chains were formed by linking shorter term variants to longer term variants, and thus a group of similar concepts from the nodes of the chain was derived. Concepts at the lower level can be generalized by the broader concepts at the higher level. Chains sharing the same root node were combined to form a hierarchical cluster tree which represented a cluster of similar concepts sharing the same cluster label. The concepts at the higher levels in a cluster were selected and integrated together using a new sentence. For example, all the concepts relating to “*student*” can be integrated in the following sentence by selecting the main concept at the top level and the concepts at the second level:

- *Student, including college student, undergraduate student, American student, and so on,
Its different aspects are investigated, including characteristics of student, behavior of student, and so on.*

The second-level concepts are divided into two types – subclass concepts and facet concepts. *Subclass concepts* represent a subclass of parent concept, whereas *facet concepts* represent an aspect or characteristic of the parent concept. Thus the sentence is divided into two parts: the first part for subclass concepts (“*including*”) and the second part (“*its different aspects are investigated, including*”).

For a cluster of similar concepts, their relationships with other concepts were integrated together to provide an overview of all associated variables connected by various types of relationships. Each type of relationship (e.g. *correlation* and *cause-effect relation*) was identified using a group of patterns. For the same type of relationships, linguistic normalization was carried out to normalize the different surface expressions using a standard expression and to conflate them. For example, the following two relationships are associated with the concept “*student*”:

- *Expected economic returns affected the college students’ future career choices.*
- *School socioeconomic composition has effect on Latino students’ academic achievement.*

They can be normalized and conflated into a simple sentence as follows:

- *Some facets of students were affected by expected economic returns and school socioeconomic composition.*

Summary Presentation

In summary presentation, the four kinds of information are combined and organized to generate the final summary. The summary is presented in an interactive Web-based interface rather than in traditional plain text so that it not only provides an overview of the topic but also allows users to zoom in and explore more details of interest. A compression rate in terms of the number of the words is specified to determine the length of the generated summary.

The four kinds of information are presented at three levels of generality:

- The top level – the summarized information;
- The second level – the specific information extracted from each dissertation abstract;
- The third level – the original dissertation abstracts.

The three levels are linked through hyperlinks. The user can click on the hyperlinks to access the more detailed information at the lower levels.

The summarized information at the top level is presented in the main window. It is viewed as the main summary. There are two types of main summaries: (1) *SYSTEM 1* generated without the aid of the taxonomy (see

Figure 2); and (2) *SYSTEM 2* generated with the aid of the taxonomy (see Figure 3). The taxonomy was constructed based on a sample of 3214 sociology dissertation abstracts using a semi-automatic method (Ou et al., 2005a). It contains lists of important n-word concepts (n=1, 2, 3, 4, and 5) in sociology domain, and the subjects of the 1-word concepts. Its function is to filter out non-concept terms, specify the important concepts in the domain, and categorize concepts into different subjects. The specific information extracted from the original dissertation abstracts is presented in separate pop-up windows of the main window. The original abstracts are presented in separate pop-up windows at the lower level.

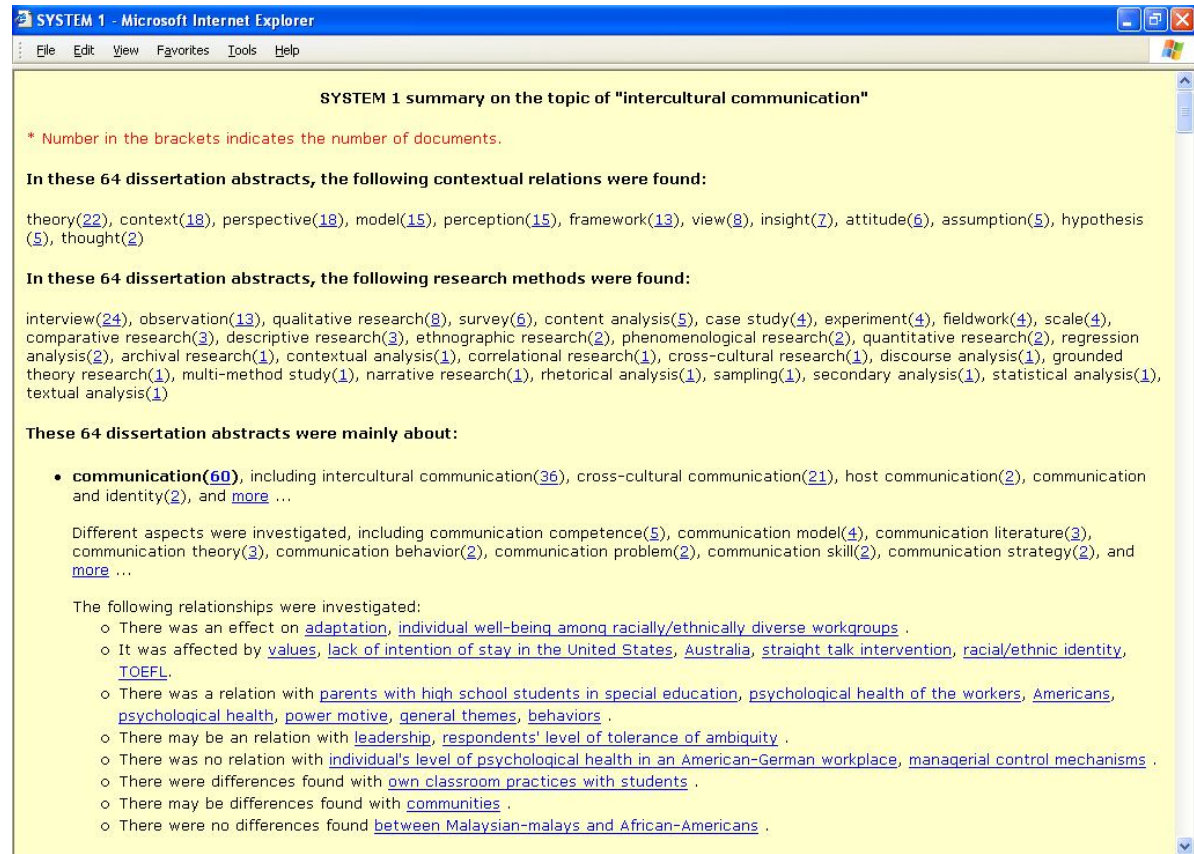


FIG. 2. A SYSTEM 1 summary generated *without* the aid of the taxonomy on the topic of “intercultural communication”.

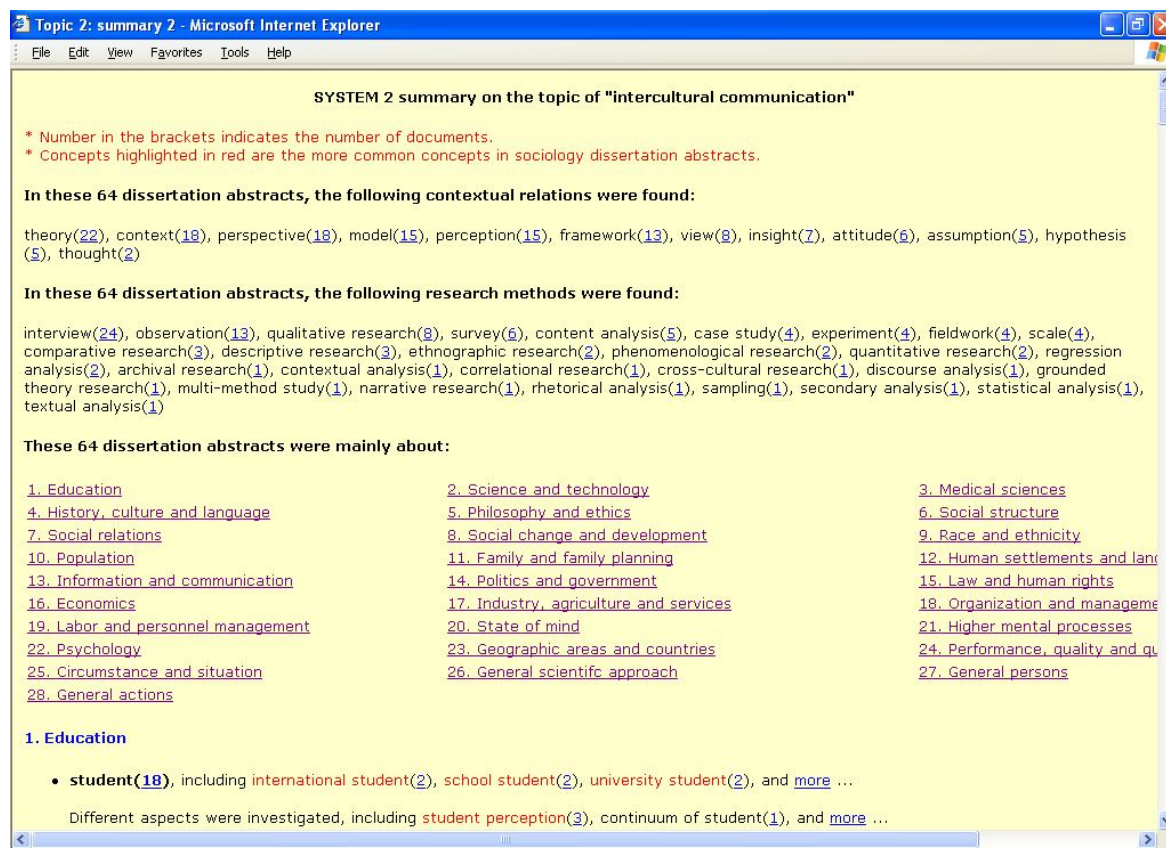


FIG. 3. A SYSTEM 2 summary generated with the aid of the taxonomy on the topic of "intercultural communication".

In the main window, the four kinds of summarized information was combined and organized. *Contextual relations* and *research methods* are presented first followed by *research concepts* and *relationships*. Only concept clusters with high document frequency (df) are presented in the main window, and the others were truncated to satisfy the compression rate of the summary. In the SYSTEM 1 summary illustrated in Figure 4, the concept clusters are arranged in descending order of the document frequency of their main concepts. In the SYSTEM 2 summary in Figure 6, the non-concept terms not found in the taxonomy (e.g. *level*, *rate* and *size*) are removed, the important sociology concepts found in the taxonomy are highlighted in red, and the concept clusters are categorized into different subjects based on the taxonomy. These features provided by the taxonomy can give users an initial overview of the covered subjects in the summary and help them to locate the subjects of interest quickly, and thus facilitate user browsing.

For each concept cluster, only the main concept and some of the second level concepts in the cluster tree are displayed. The main concept is displayed first followed by the second-level concepts occurring in at least two documents. These second-level concepts are divided into two subgroups – subclass *concepts* and facet *concepts*. For each subgroup, the 2-word, 3-word, 4-word and 5-word concepts whose document frequencies are above or equal to two are displayed in sequence. If the number of the concepts in each subgroup is more than twenty, the longer concepts are cut off. If the number of the concepts in each subgroup is less than two, two concepts occurring in one document are displayed instead. Moreover, all the concepts in the cluster are displayed in two separate pop-up windows – one for subclass concepts and another for facet concepts. The pop-up windows can be accessed from the main window through hyperlinks.

For each concept cluster, the number of documents is given in parenthesis. This is clickable and is linked to a list of documents sharing a given concept in a separate pop-up window. For each document, the title, research concepts, contextual relations and research methods are displayed (see Figure 4). The title of the document is also clickable and is linked to the original dissertation abstract in another pop-up window at the lower level.

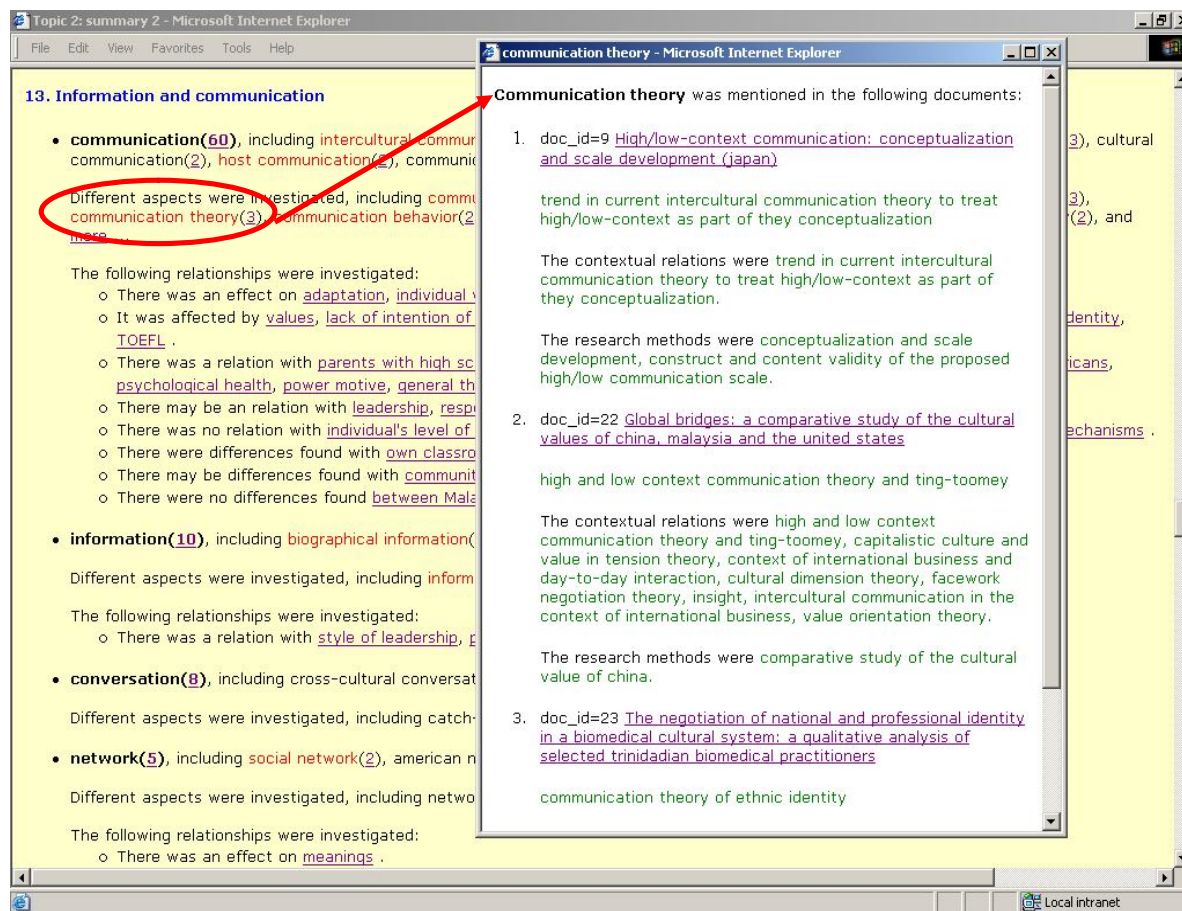


FIG. 4. A list of summarized single documents sharing a given concept in a pop-up window.

For each relationship, each variable concept is attached to a pop-up box which displays an original expression of the relationship involving the concept. The concept is also clickable and is linked to the document describing the relationship in a separate pop-up window. For the document, the title, the original sentence describing the relationship (highlighted in green), and adjacent sentences are displayed.

Evaluation of the Summarization Method

The accuracy of each step in the summarization process has been evaluated and was reported by Ou et al. (2005b). This paper reports the subsequent user evaluation for assessing the overall quality and usefulness of the summaries. The used evaluation was mainly intrinsic, in which the users were asked to subjectively judge the quality of the summaries and their usefulness for specific purposes. Two types of the variable-based summaries generated using our summarization method – with and without the use of the taxonomy – were compared against two types of sentence-based summaries – one generated by a state-of-the-art system, MEAD, using a sentence extraction method, and another generated by extracting research objective sentences only from each dissertation abstract.

Evaluation Design

20 research topics (e.g. *social support*, *inter-cultural communication*, *public health care*) 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. Although the number of dissertation abstracts in the 20 topics was various from 12 to 200, 200 abstracts were retained at most for each topic by removing the rest. The following four types of summaries were generated for the set of dissertation abstracts on each topic:

- *A variable-based summary generated without the aid of a taxonomy:* It focuses on research concepts and relationships, as well as research methods and contextual relations. This type of summary was labeled *SYSTEM 1* (see Figure 2).
- *A variable-based summary generated with the aid of a taxonomy:* It also focuses on research concepts and relationships. Furthermore, based on the taxonomy, non-concept terms were filtered out, important sociology concepts were highlighted in red, and concepts were categorized into different subjects. This type of summary was labeled *SYSTEM 2* (see Figure 3).
- *A sentence-based summary generated by extracting only the research objectives sentences from each abstract:* It consists of the sentences that are from the *research objectives* section of each dissertation abstract. The type of summary was labeled *OBJECTIVES* (see Figure 5).
- *A sentence-based summary generated by a system MEAD:* It consists of the sentences that were ranked as important, according to centroid words, sentence position and first sentence overlap, in the set of dissertation abstracts. It was created by a state-of-the-art multi-document summarization system MEAD 3.08, which adopts a centroid-based cross-document sentence extraction method (Radev, Blitzer, Winkel, Allison, & Topper, 2003). This type of summary was labeled *MEAD* (see Figure 6).

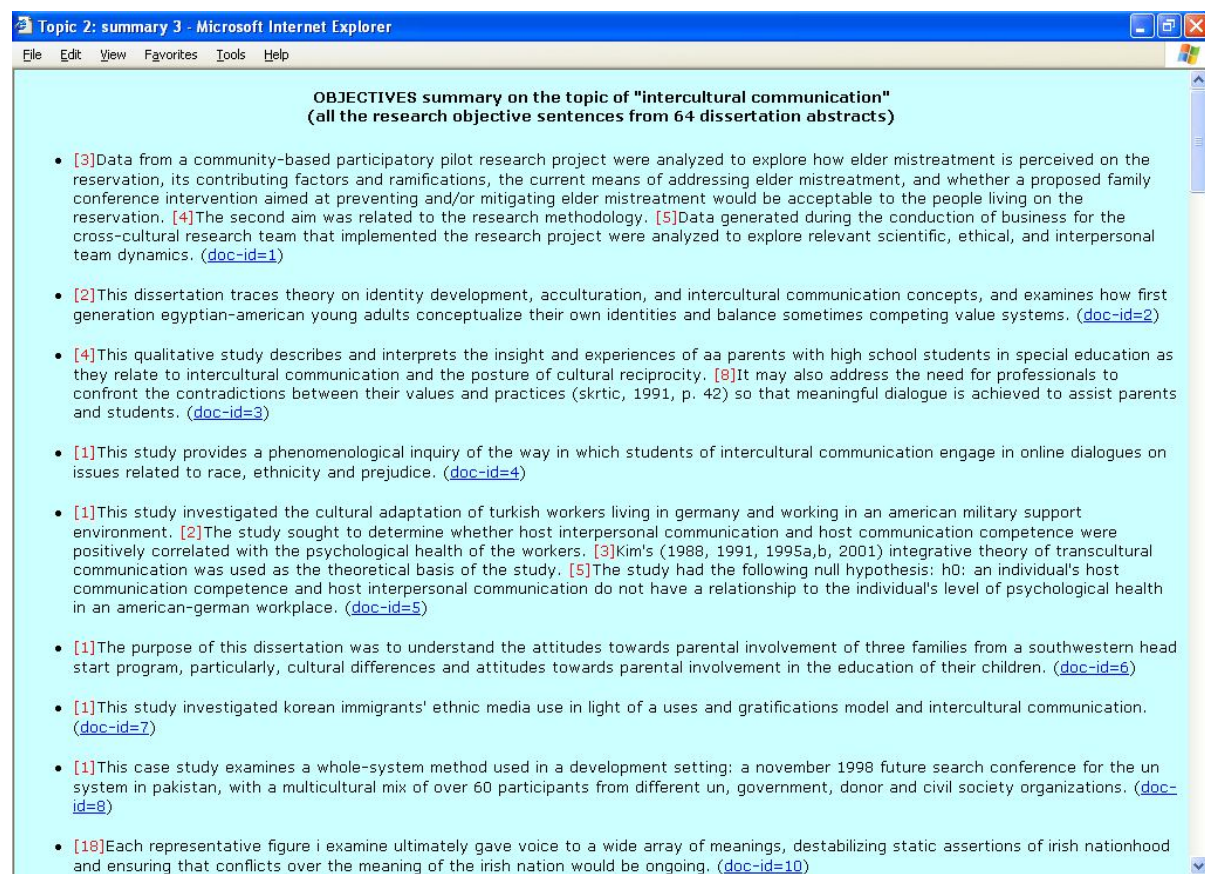


FIG. 5. A OBJECTIVES summary generated by extracting only the research objectives sentences on the topic of “intercultural communication”.

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. Both quality and usefulness were assessed subjectively by the human subjects. Each subject was asked to score and rank the summaries according to several pre-defined criteria (e.g. readability, comprehensibility and usefulness) and answer some open-ended questions. A questionnaire was used to record the subjects' evaluation (see Appendix A). There are two groups of human subjects – sociological researchers and general users. The first group of subjects was the 20 sociological researchers who submitted their research topics for generating the summaries for the evaluation and used the summaries for their research-related purposes. Each researcher only read and evaluated a set of summaries generated for this/her topic. The second group of subjects was 40 general users who were graduate students in the MSc (Information Studies and Knowledge Management) programs at Nanaynag Technological University, Singapore. They were not familiar with the research topics and read the summaries to obtain an overview or general information on the topic. Each general user also read and evaluated a set of summaries for a topic that was assigned to him randomly. Each topic was evaluated by one researcher and two general users.

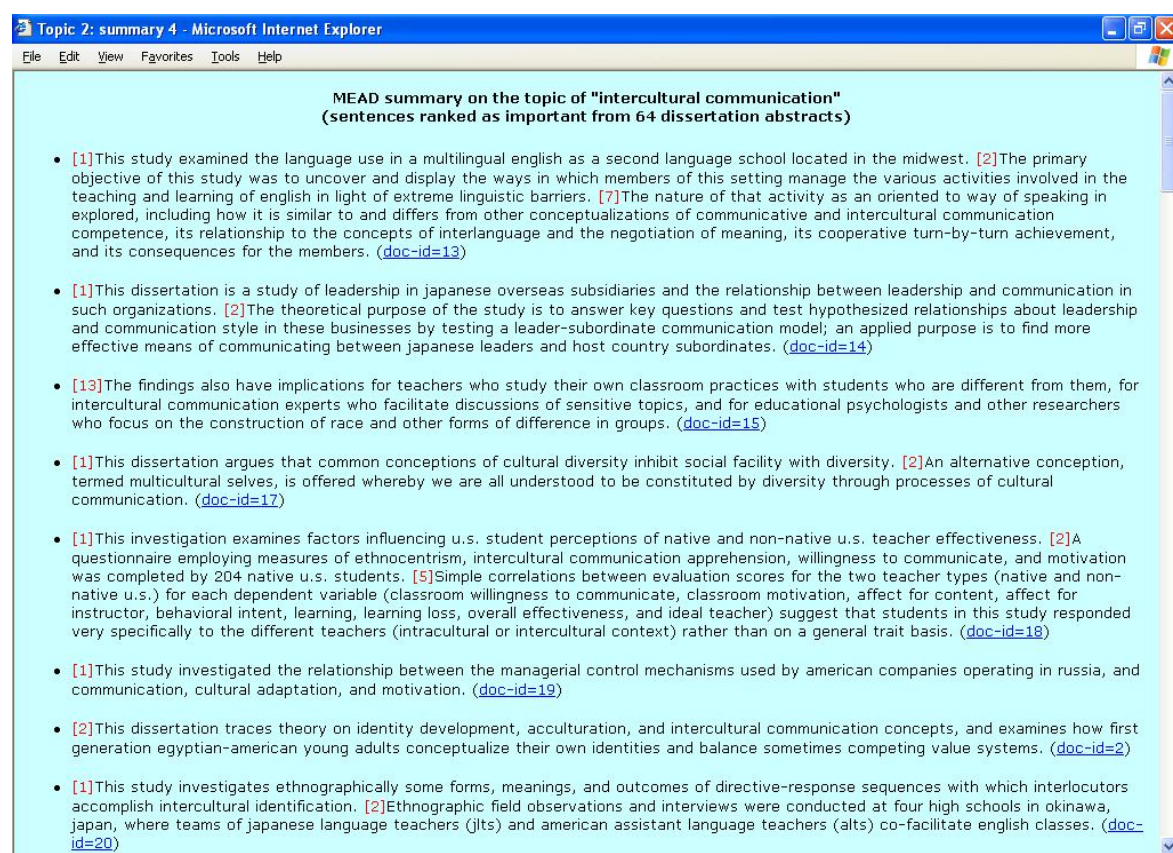


FIG. 6. A MEAD summary generated by a system MEAD that extracts important sentences based on various criteria on the topic of “intercultural communication”.

Since the four summaries were generated from the same source documents and the subjects had to examine all four summaries, there may be carry-over effects from the summaries read earlier. After the subject has read the first summary, familiarity with the first summary and its content may influence the subject's reading and assessment of the subsequent summaries. Also the subject may not read the subsequent summaries as thoroughly. To compensate for this, the four types of summaries were presented to different subjects in four presentation orders:

- (1) SYSTEM 1 → SYSTEM 2 → OBJECTIVES → MEAD
- (2) SYSTEM 1 → SYSTEM 2 → MEAD → OBJECTIVES
- (3) OBJECTIVES → MEAD → SYSTEM 1 → SYSTEM 2
- (4) MEAD → OBJECTIVES → SYSTEM 1 → SYSTEM 2

Evaluation Results for the Researchers

(1) Average Scores

The overall readability, comprehensibility, and usefulness were scored on a 7-point scale, from 1 indicating unreadable to 7 indicating very fluent. The average scores for the four types of summaries from the 20 researchers are shown in Table 1.

TABLE 1. Average scores for the four types of summaries from the 20 researchers.

Criterion	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
Readability	4.40	5.20	5.70	5.00
Comprehensibility	4.75	5.10	5.60	4.95
Usefulness	5.0	5.70	5.65	4.9

SYSTEM 2 obtained the second highest readability (5.2) and comprehensibility (5.1) scores among the four types of summaries. It was much better than SYSTEM 1's scores (4.4 and 4.75). This indicates that with the use of a taxonomy for information filtering and organization, the readability and comprehensibility of the variable-based summary can be substantially improved, and was better than the set of important sentences generated by MEAD (5.0 and 4.95), but still worse than the research objective sentences in OBJECTIVES (5.7 and 5.6). For research-related work, SYSTEM 2 and OBJECTIVES were more useful than SYSTEM 1 and MEAD. SYSTEM 2 and OBJECTIVES obtained similar usefulness scores (5.7 and 5.65) whereas SYSTEM 1 and MEAD obtained similar scores (5.0 and 4.9). This indicates that with the use of the taxonomy for information organization, the usefulness of the variable-based summary can be improved, and was as useful as the research objective sentences in OBJECTIVES and much more useful than the set of important sentences generated by MEAD. In addition, this indicates that the researchers were more concerned about research objectives than other kinds of information in a dissertation.

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 present important concepts and simple relationship sentences whereas OBJECTIVES and MEAD present 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 multiple clusters from difference perspectives. Moreover, separate concepts are less fluent than complete sentences. Although OBJECTIVES and MEAD both contained important sentences, the researchers indicated that the research objective sentences in OBJECTIVES were more concise and easier to read than the set of important sentences generated by MEAD.

For comprehensibility, the researchers indicated that OBJECTIVES and MEAD were a little easier to understand than SYSTEM 1 and 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 to a greater extent than the set of important sentences generated by MEAD. With the use of the taxonomy for information filtering and organization, SYSTEM 2 was easier to understand and can indicate the main ideas of the topic to a greater extent than SYSTEM 1.

For usefulness, the researchers were asked to select in which aspects each summary was useful for their purpose. The evaluation scores for the nine aspects of usefulness are given in Table 2.

SYSTEM 1 & 2 were selected by more researchers than OBJECTIVES and MEAD on the three aspects of usefulness – (5), (7) and (9). OBJECTIVES and MEAD were also useful to a lesser extent in the (7) aspect. However, OBJECTIVES and MEAD were almost useless in the (5) and (9) aspects. OBJECTIVES and MEAD were selected by more researchers than SYSTEM 1 and 2 on the two aspects of usefulness – (3) and (8). SYSTEM 1 and 2 were also useful to a lesser extent in the (3) aspect. However, SYSTEM 1 and 2 were almost useless in the (8) aspect. The four types of summaries were selected by similar numbers of researchers on the three aspects of usefulness – (2), (4) and (6). The four types of summaries were useful in the (6) aspect. But all were almost useless in the (2) aspect. SYSTEM 2 was selected by the majority of researchers (70%) among the four types of summaries on the (1) aspect. However, the remaining three types of summaries were also useful to a lesser extent in this aspect.

TABLE 2. Frequency (percentage) of researchers selecting the different aspects of usefulness for the four types of summaries.

Aspect of usefulness	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
(1). Give you an overview of the research area	11 (55%)	14 (75%)	11 (55%)	10 (50%)
(2). Help you identify research gaps easily	5 (25%)	6 (30%)	4 (20%)	3 (15%)
(3). Help you identify documents of interest easily	10 (50%)	10 (50%)	13 (65%)	13 (65%)
(4). Indicate research trends in the area	10 (50%)	9 (45%)	8 (40%)	8 (40%)
(5). Indicate similarities among previous studies	10 (50%)	12 (60%)	3 (15%)	2 (10%)
(6). Indicate differences among previous studies	3 (15%)	5 (25%)	5 (25%)	2 (10%)
(7). Indicate important concepts in the area	14 (70%)	15 (75%)	9 (45%)	8 (40%)
(8). Indicate important theories, views, or ideas	5 (25%)	7 (35%)	10 (50%)	10 (50%)
(9). Indicate important research methods used	14 (70%)	11 (55%)	6 (30%)	5 (25%)

- *Bold figures indicate the higher frequency (percentage) of the summaries for each criterion.*

(2) Significant Tests

Repeated measures analyses of variance (ANOVA) were performed using the SPSS statistical software to investigate whether there were significant differences in the average scores of readability, comprehensibility, and usefulness of the four types of summaries. The results are shown in Table 3.

There was a significant difference in the average scores of readability among the four types of summaries ($p=0.012$). However, only the difference between SYSTEM 2 and SYSTEM 1 was found significant ($p=0.008$). The differences between SYSTEM 2 and the other two summaries (OBJECTIVES and MEAD) were not significant. There was no significant difference in the average scores of comprehensibility ($p=0.19$) and usefulness ($p=0.30$) among the four types of summaries.

TABLE 3. Repeated measures ANOVA to test for significant differences in the average scores among the four types of summaries from the 20 researchers.

Criterion	Source of variation	Sum of squares	Degree of freedom	Mean square	F-ratio	Significance (p)
Readability	SUMMARY	17.873	3	5.958	4.036	0.012*
	SUMMARY * ORDER	39.796	9	4.422	2.996	0.006*
	Error (SUMMARY)	70.854	48	1.476		
Comprehensibility	SUMMARY	8.540	1.985	4.301	1.748	0.191
	SUMMARY * ORDER	36.412	5.956	6.114	2.484	0.044*
	Error (SUMMARY)	78.188	31.764	2.461		
Usefulness	SUMMARY	8.290	3	2.763	1.271	0.295
	SUMMARY * ORDER	26.800	9	2.978	1.370	0.228
	Error (SUMMARY)	104.313	48	2.173		

- *SUMMARY variable has four levels: SYSTEM1, SYSTEM2, OBJECTIVES, and MEAD;*
- *ORDER variable has four levels: $S1 \rightarrow S2 \rightarrow O \rightarrow M$, $S1 \rightarrow S2 \rightarrow M \rightarrow O$, $O \rightarrow M \rightarrow S1 \rightarrow S2$, $M \rightarrow O \rightarrow S1 \rightarrow S2$;*
- *"*" indicates significance at the 5% level;*

Considering the presentation order of the four types of summaries, there was a significant difference in the average scores of readability ($p=0.006$) and a marginally significant differences in the average scores of comprehensibility ($p=0.04$) for the four presentation orders ($p=0.006$), but no significant difference was found in

the average scores of usefulness scores ($p=0.23$). As shown in Table 4, when SYSTEM 2 was presented earlier than MEAD (order 1 & 2), SYSTEM 2 obtained a lower readability score (4.3) and comprehensibility score (4.4) than MEAD (5.6 and 5.7). In contrast, when SYSTEM 2 was presented later than MEAD (order 3 & 4), SYSTEM 2 obtained a higher readability score (5.9) and comprehensibility score (5.6) than MEAD (4.5 and 4.3). It appears that the summary that was read earlier obtained a lower readability and comprehensibility score. However, the readability and comprehensibility scores of OBJECTIVES did not change too much among the four presentation orders. This may be because research objectives sentences were easy to read and understand.

(3) Overall Evaluation

The 20 researchers were asked to rank the four types of summaries. A weighted rank score was calculated for each summary. A weight value of 4 was assigned to the first rank, 3 to the second rank, 2 to the third rank, and 1 to the fourth rank. The researchers were also asked to select one or more summaries that they preferred to use for their research-related work. The ranks and the researchers' preferences are shown in Table 5.

TABLE 4. Average scores for the four presentation orders from the 20 researchers.

Criterion	Order	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
Readability	1. S1→S2→O→M	4.00	4.67	5.67	5.67
	2. S1→S2→M→O	2.75	4.00	5.00	5.50
	Average for 1 and 2	3.38	4.34	5.34	5.59
	3. O→M→S1→S2	4.75	5.25	5.75	5.25
	4. M→O→S1→S2	5.67	6.50	6.17	3.83
	Average for 3 and 4	5.21	5.88	5.96	4.54
Comprehensibility	1. S1→S2→O→M	4.67	5.00	5.83	5.67
	2. S1→S2→M→O	3.25	3.75	5.25	5.75
	Average for 1 and 2	3.96	4.38	5.54	5.71
	3. O→M→S1→S2	4.50	5.25	5.25	4.75
	4. M→O→S1→S2	6.00	6.00	5.83	3.83
	Average for 3 and 4	5.25	5.63	5.54	4.29

TABLE 5. Ranking and preference for the four types of summaries from the 20 researchers.

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%)

The overall ranking for the four types of summaries based on the weighted rank score was: **1. SYSTEM 2** → **2. OBJECTIVES** → **3. SYSTEM 1** → **4. MEAD**. 70% of the researchers indicated preference for SYSTEM 2 for their research-related work. 55% of the researchers indicated preference for OBJECTIVES. Only 25% indicated preference for MEAD. SYSTEM 1 & 2 used the variable-based structure whereas OBJECTIVES and MEAD used a sentence-based structure. The comments given by the researchers for the two kinds of structures (variable-based summaries and sentence-based summaries) are listed in Table 6.

Evaluation Results for the General Users

The readability, comprehensibility and usefulness were scored using the 7-point scale with the researchers, and the average scores for the four types of summaries from the 40 general users are shown in Table 7.

TABLE 6. Comments by the researchers on the variable-based structure and sentence-based structure.

Researchers' comments	Variable-based structure	Sentence-based structure
Positive points	<ul style="list-style-type: none"> It is more efficient to give an overview of a topic. It can help researchers find what has been done easily; It is well-organized and concise. It makes easier for researchers to find similar information. It is useful for information scanning For quantitative studies which focus on relationships between variables, it is more useful. 	<ul style="list-style-type: none"> It provides more direct information and is easy to understand; It is useful to provide more specific and detailed information; It is useful for understanding research questions and key findings;
Negative points	<ul style="list-style-type: none"> It is too brief to provide accurate information on the topic. The simple terms in the variable-based structure are easy to make users confused and lost. 	<ul style="list-style-type: none"> Researchers have to read all the sentences in the summary to know what it is about; Only parts of sentences are presented in the summary, which can not cover all important aspects of a topic; It is not very well-organized and sometimes confusing; It is time-consuming to read the complete sentences; It is too broad to find out the right information.

TABLE 7. Average scores for the four types of summaries from the 40 general users.

Criterion	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
Readability	3.90	5.00	3.78	3.80
Comprehensibility	3.88	4.58	3.85	3.73
Usefulness	3.73	4.57	3.73	3.76

SYSTEM 2 obtained the highest readability score (5.0), the highest comprehensibility score (4.6), and the highest usefulness score (4.6) among the four summaries. There were significant differences in the average scores of readability between SYSTEM 2 and each of the other three summaries ($p=2.21E-7$ for SYSTEM 1, $p=0.01$ for OBJECTIVES, and $p=0.01$ for MEAD). There was no significant difference in the average scores of comprehensibility ($p=0.10$) and usefulness ($p=0.13$) among the four types of summaries.

Considering the presentation order of the four types of summaries, there was no significant difference in the average scores of readability ($p=0.24$) and usefulness ($p=0.3$) for the four presentation orders. However, there was a significant difference in the average scores of comprehensibility ($p=0.04$) between SYSTEM 2 and MEAD. As shown in Table 8, when MEAD was presented later than SYSTEM 2, it obtained a lower score of comprehensibility (4.1) than SYSTEM 2 (4.4). In contrast, when MEAD was presented earlier than SYSTEM 2, its comprehensibility score (3.6) became worse. It appears that the summary which was read earlier obtained a lower comprehensibility score for the general users, which was the same with the researchers (see section 5.2).

TABLE 8. Average scores for the four presentation orders from the 40 general users.

Criterion	Order	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
	1. S1→S2→O→M	3.21	4.21	3.79	3.71

Comprehensibility	2. S1→S2→M→O	3.50	4.50	4.50	4.38
	Average for 1 and 2	3.36	4.36	4.15	4.05
	3. O→M→S1→S2	4.14	4.00	3.71	4.29
	4. M→O→S1→S2	4.82	5.45	3.55	2.91
	Average for 3 and 4	4.48	4.73	3.63	3.6

The 40 general users were asked to rank the four types of summaries. A weighted rank score was calculated for each summary using the same method for the 20 researchers. The general users were also asked to select one or more summaries that they preferred to use for obtaining the general information on a topic. The ranks and the general users' preferences are shown in Table 9.

The overall ranking for the four types of summaries based on the weighted rank scores was: **1. SYSTEM 2** → **2. SYSTEM 1** → **3. OBJECTIVES** → **4. MEAD**. 64% of the general users indicated preference for SYSTEM 2 for obtaining general information on the topic. But only 31% of the general users indicated preference for OBJECTIVES.

Comparison between Researchers and General Users

Both researchers and general users ranked SYSTEM 2 first and MEAD last among the four types of summaries. 70% of the researchers indicated their preference to use SYSTEM 2 for their research-related work, whereas 64% of the general users indicated their preference to use SYSTEM 2 for obtaining general information on a topic. In addition, 30% of the researchers and 18% of the general users indicated preference for SYSTEM 1. A higher percentage of researchers appear to prefer the variable-based summaries than general users.

The researchers ranked the OBJECTIVES second, whereas the general users ranked it third. Moreover, more researchers (55%) preferred to use OBJECTIVES than the general users (31%). For the general users, approximately the same percentage indicated preference for using MEAD (33%) as for OBJECTIVES (31%). However, for the researchers, only 25% indicated preference for using MEAD. This suggests that the researchers found the research objectives sentences more useful than the general important sentences extracted by MEAD, whereas the general users did not find the research objective sentences more useful.

TABLE 9. Ranking and preference for the four types of summaries from the 40 general users

Rank	SYSTEM 1	SYSTEM 2	OBJECTIVES	MEAD
No.1 (weight=4)	4(10%)	22(55.0%)	6(15.0%)	8(20.0%)
No.2 (weight=3)	19(47.5%)	6(15.0%)	10(25.0%)	5(12.5%)
No.3 (weight=2)	9(22.5%)	4(10.0%)	15(37.5%)	12(30.0%)
No.4 (weight=1)	8(20.0%)	8(20.0%)	9(22.5%)	15(37.5%)
Weighted rank score	2.475	3.05	2.325	2.15
Preference	7(17.9%)	25(64.1%)	12(30.8%)	13(33.3%)

Discussion and Conclusion

This study developed a method for automatic construction of multi-document summaries of sociology dissertation abstracts, focusing on research concepts and their research relationships. A variable-based framework was proposed for integrating and organizing different kinds of information extracted from different dissertation abstracts. Based on the variable-based framework, a summarization method was developed. It focused on extracting *research concepts* and *their research relationships* as well as *contextual relations* and *research methods* from different dissertation abstracts, integrating them across dissertation abstracts using concept generalization and relationship normalization and conflation, combining and organizing the four kinds of information, and presenting them in a Web-based interface. The summarization method developed in this study is just one way of operationalizing the variable-based framework. In particular, different presentation formats can be used to organize and present the summary. Two presentation formats were investigated in this study. One presentation format made use of a taxonomy to filter out non-concept terms, highlight important concepts in the domain, and categorize concepts into different subjects. The other presentation format did not use a taxonomy for information filtering, highlighting and categorization. A user evaluation was carried out to compare the two types of variable-

based summaries against two types of sentence-based summaries: one generated by displaying research objective sentences only and another generated by the MEAD system which identified important sentences using a variety of general features (e.g. centroid words, sentence position and first-sentence overlap).

In the user evaluation, 70% of the researchers and 64% of the general users indicated their preference for the variable-based summaries generated with the use of the taxonomy, 55% of the researchers and 31% of the general users indicated their preference for the research objective summary, and only 25% of the researchers and 31% of the general users indicate their preference for the MEAD summary. The researchers indicated that the variable-based summaries were efficient in giving an overview of the topic and useful for information scanning. Comparing the two types of variable-based summaries, the summary generated with the use of the taxonomy obtained the highest rank score from the researchers, whereas the one that did not make use of the taxonomy obtained lower scores. This demonstrates that using a taxonomy for filtering out non-concept terms, highlighting important concepts in the domain and categorizing concepts into different subjects can substantially improve the quality and usefulness of the variable-based summaries. On the other hand, 55% of the researchers indicated their preference for the research objective sentences, since the sentence-based summaries could provide more direct information and were easy to understand. A higher percentage of researchers (55%) than general users (31%) preferred the research objective summary, indicating that the researchers had a great interest in the research objectives.

Interestingly, in the user evaluation, it was found that the presentation order of the different summaries influenced the assessment of the users. The summaries presented later were more likely to be assessed favorably and be given a better score. This is because of the carry-over effect from the summaries read earlier. After a user had read the previous summaries, familiarity with the content may make the subsequent summaries easier to understand.

The user evaluation indicated that summarizing a set of sociology dissertation abstracts by focusing on research concepts and their research relationships was a promising approach. In future, other presentation methods for operationalizing the variable-based framework can be investigated. In the current variable-based summaries, contextual relations and research methods were displayed separately from the research concepts and relationships. However, contextual relations and research methods complement the information on research concepts and relationships, giving more details of how the research concepts and relationships are studied. Future presentation methods should combine the research concepts and relationships with the contextual relations and research methods used, to provide more complete information of the research.

The researchers also showed an interest in the research objective sentences. Thus, the research objective sentences can be added to the variable-based summaries to provide alternative information presentation. The users can select to browse research concepts or the sentences that are likely to contain the research concepts and relationships. Moreover, the research objective sentences can be ranked according to the criteria used in MEAD and only the most important research objective sentences need be displayed. Furthermore, the research concepts can be extracted only from the research objective sentences or the more important research objective sentences rather than from the research objectives + research results sections as used in this study.

References

- Afantenos, S., Doura, I., Kapellou, E., & Karkaletsis, V. (2004). Exploiting cross-document relations for multi-document evolving summarization. In G. A. Vouros, & T. Panayiotopoulos (Eds.), *Methods and Applications of Artificial Intelligence in Volume 3025 of Lecture Notes in Computer Science: Proceedings of the 3rd Hellenic Conference on Artificial Intelligence* (pp.410-419). Berlin: Springer-Verlag.
- Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: A survey. *Journal of Artificial Intelligence in Medicine*, in press.
- Alonso, L. (2005). Representing discourse for automatic text summarization via shallow NLP techniques. *PhD Thesis*. Department de Lingüística General, Universitat de Barcelona. Retrieved April 10, 2005, from <http://lalonso.sdf-eu.org/tesi.pdf>
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Bergler, S., Witte, R., Li, Z., Khalife, M., Chen, Y., Doandes, M., & Andreevskaia, A. (2004). Multi-ERSS and ERSS 2004. In *Proceedings of the Document Understanding Conference 2004*. Retrieved April 10, 2005, from <http://www-nlpir.nist.gov/projects/duc/pubs.html>

-
- Brunn, M., Chali, Y., & Dufour, B. (2002). The University of Lethbridge text summarizer at DUC 2002. In *Proceedings of the Document Understanding Conference 2002*. Retrieved April 10, 2005, from <http://www-nlpir.nist.gov/projects/duc/pubs.html>
- Carbonell, J. G., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-336). New York: ACM Press.
- Chali, Y., & Kolla, M. (2004). Summarization techniques in DUC 2004. In *Proceedings of the Document Understanding Conference 2004*. Retrieved July 10, 2005, from <http://www-nlpir.nist.gov/projects/duc/pubs.html>
- Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM*, 16(2), 264-285.
- Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval* (pp. 121-128). New York: ACM Press.
- Hand, T. F. (1997). A proposal for task-based evaluation of text summarization systems. In I. Mani, & M. T. Maybury (Eds.): *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (pp.31-38). Morristown, NJ: Association for Computational Linguistics.
- Harabagiu, M. S. & Lacatusu, F. (2002). Generating single and multi-document summaries with GISTEXER. In *Proceedings of the Document Understanding Conference 2002*. Retrieved April 20, 2005, from <http://www-nlpir.nist.gov/projects/duc/pubs.html>
- Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Wise, G., & Zhang, X. (2002). Cross-document summarization by concept classification. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.121-128). New York: ACM Press.
- Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: Experiment and analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization* (pp.60-68). Retrieved April 15, 2005, from http://www.cs.cornell.edu/~regina/my_papers/evaluation.ps.gz
- Jones, K.S., & Galliers, J.R. (1996). *Evaluating natural language processing systems: An analysis and review*. In J.G. Carbonell, & J. Siekmann (Eds.), *Volume 1083 of Lecturer Notes in Artificial Intelligence*. Berlin: Springer Verlag.
- Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In EA Fox, P. Ingwersen, & R. Fidel (Eds.), *SIGIR-95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68-73). New York: ACM Press.
- Macionis, John, J. (2000). *Sociology* (8th ed.). Prentice Hall.
- Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67. Reprinted in I. Mani, I., & M. T. Maybury (Eds.): *Advances in automatic text summarization* (pp.357-389). Cambridge, MA: MIT Press.
- Mani, I. & Maybury, M. T. (Eds.). (1999). *Advances in automatic text summarization*. Cambridge, MA: MIT press.
- Mani, I. (2001). Summarization evaluation: an overview. In *Proceedings of the 2nd NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization*. Tokyo: National Institute of Informatics.
- McKeown, K., & Radev, D. (1995). Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.74-82). New York: ACM Press.
- McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In R. Dechter, M.Kearns, & R. Sutton (Eds.): *Proceedings of the 16th National Conference on Artificial Intelligence* (pp. 453-460). Menlo Park, CA: American Association for Artificial Intelligence.
- Minel, J. -L., Nugier, S., & Piat, G. (1997). How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. In I. Mani, & M. T. Maybury (Eds.): *Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (pp.25-33). Morristown, NJ: Association for Computational Linguistics.
- Moxley, Joseph M. (2001). Universities should require electronic theses and dissertations. *Educause Quarterly*, 3, 61-63.
- Document Understanding Conference (DUC) (2002). Retrieved April 24, 2005, from <http://www-nlpir.nist.gov/projects/duc/index.html>
- Ou, S., Khoo, S.G., & Goh, H.L. (2002). A hierarchical framework for multi-document summarization of dissertation abstracts. In E.-P. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, & T. Costantino (Eds.), *Volume*

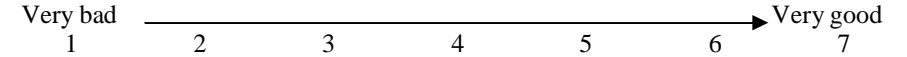
-
- 2555 of *Lecturer Notes in Computer Science: Proceedings of the 5th International Conference on Asian Digital Libraries* (pp. 99-110). Berlin: Springer Verlag.
- Ou, Shiyang, Khoo, S. G. Christopher, & Goh, Dion, & Heng, Hui-Hing. (2004). Automatic parsing discourse structure of sociology dissertation abstract as sentence categorization. In *Proceedings of the 8th International Conference of the International Society for Knowledge Organization* (pp. 345-350). Würzburg: Ergon Verlag.
- Ou, Shiyang, Khoo, S. G. Christopher, & Goh, Dion. (2005a). Constructing a taxonomy to support multi-document summarization of dissertation abstracts. *Journal of Zhejiang University SCIENCE*, 6A (11), 1258-1267.
- Ou, Shiyang, Khoo, S. G. Christopher, & Goh, Dion. (2005b). A multi-document summarization system for sociology dissertation abstracts: design, implementation and evaluation. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries 2005* (pp.450-461). Berlin: Springer Verlag.
- Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26(1), 171-186.
- Radev, R. D., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation and user studies. In *ANLP/NAACL 2000 Workshop on Automatic Summarization* (pp.21-29). Retrieved May 30, 2005, from <http://tangra.si.umich.edu/~radev/papers/centroid.pdf>
- Radev, D. R., Blitzer, J., Winkel, A., Allison, T., & Topper, M. (2003). MEAD Documentation, Version 3.08. Retrieved May 24, 2005, from <http://www.summarization.com/mead/>
- Radev, D. R., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 40, 919-938.
- Rath, G.J., Resnick, A., & Savage, T.R. (1961). The formation of abstracts by the selection of sentences. *American Documentation*, 12 (2), 139-141.
- Saggion, H., Radev, D., Teufel, S., Lam, W., & Strassel, S. M. (2002). Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In *Proceedings of Language Resources and Evaluation Conference 2002* (pp. 29-31). Retrieved June, 5, 2005, from http://www.cl.cam.ac.uk/users/sht25/papers/jhu_workshop_lrec2002.pdf
- Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management*, 33(2), 193-207.
- Schlesinger, J. D. & Conroy, J. M. Okurowski, M. E. & O'Leary, D. P. (2003). Machine and human performance for single and multidocument summarization. *IEEE Intelligent Systems*, 18(1), 46-54.
- Trochim, William M. K. (1999). *The research methods knowledge base*. Cincinnati, OH: Atomic Dog Publishing.
- White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., & Wagstaff, K. (2001). Multi-document summarization via information extraction. In *Proceedings of the 1st International Conference on Human Language Technology Research*. Retrieved June 3, 2005, from <http://www.cse.buffalo.edu/faculty/drpierce/papers/hlt2001.html>
- Zhang, Z., Blair-Goldensohn, S., & Radev, D. R. (2002). Towards CST-enhanced summarization. In R. Dechter, M. Kearns, & R. Sutton (Eds.): *Proceedings of the 18th National Conference on Artificial Intelligence* (pp. 439-445). Menlo Park, CA: American Association for Artificial Intelligence.

Appendix: Questionnaire for User Evaluation

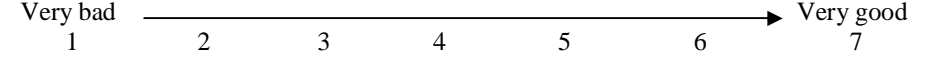
The purpose of this evaluation is to assess the quality and usefulness of the multi-document summaries of sociology dissertation abstracts. The set of dissertation abstracts retrieved using a research topic is condensed into a summary. Four different summaries have been constructed with two kinds of structures: (1) focusing on research concepts and relationships; and (2) focusing on important sentences. Please read the four summaries using Internet Explorer, and judge the quality and usefulness of each of the summaries using the questionnaire below.

A. Evaluation of Readability

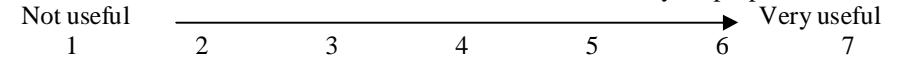
1. Does each of the summaries contain the vacuous or very general information?
☐ None ☐ A Little ☐ Some ☐ A lot
2. Does each of the summaries contain duplicate information?
☐ None ☐ A little ☐ Some ☐ A lot
3. Does each of the summaries contain dangling anaphora (e.g. unresolved pronouns)?
☐ None ☐ A little ☐ Some ☐ A lot

-
4. Is each of the summaries fluent?
☐ Not at all ☐ A little ☐ Quite fluent ☐ Very fluent
5. Is each of the summaries concise?
☐ Not at all ☐ A little ☐ Quite concise ☐ Very concise
6. Is each of the summaries coherent?
☐ Not at all ☐ A little ☐ Quite coherent ☐ Very coherent
7. Please rate the overall readability of each of the summaries in terms of the above aspects.
Very bad  Very good
1 2 3 4 5 6 7

B. Evaluation of comprehensibility

8. Is each of the summaries easy or hard to understand?
☐ Very hard ☐ Quite hard ☐ Quite easy ☐ Very easy
9. Does each of the summaries give a clear indication of the main ideas of the topic?
☐ Not at all ☐ A little ☐ Some ☐ A lot
10. Please rate the overall comprehensibility of each of the summaries in terms of the above aspects.
Very bad  Very good
1 2 3 4 5 6 7

C. Evaluation of usefulness

11. Please select your purpose for reading the dissertation abstracts. If your purpose is not found in the follows, please write down.
☐ For your PhD project ☐ For your Master project ☐ For your coursework ☐ For teaching
12. In which ways is each of the summaries useful for your purpose? You can select one or more. Please write down, if you have any other ways.
☐ Give you an overview of the research area;
☐ Help you identify research gaps in the area easily;
☐ Help you identify the documents of interest easily;
☐ Indicate research trends in the area;
☐ Indicate similarities among previous studies;
☐ Indicate differences among previous studies;
☐ Indicate important concepts in the area;
☐ Indicate important theories, views, or ideas in the area;
☐ Indicate important research methods used in the area;
13. Please rate the overall usefulness of each of the summaries for your purpose.
Not useful  Very useful
1 2 3 4 5 6 7

D. Overall Evaluation

14. Please rank the four summaries in order of your preference. Please explain.
• No. 1 : _____ No. 2 : _____ No. 3 : _____ No. 4 : _____
•

-
15. If the four summaries are readily available for you, which summary (or summaries) are you likely to use for your purpose? You can select one or more.

☐ Summary 1 ☐ Summary 2 ☐ Summary 3 ☐ Summary 4 ☐ None

E. In-depth Questions

16. Two types of summary structure are used in the summaries. *Summary 1* & 2 focus on research variables, whereas *Summary 3* & 4 focus on important sentences. Which type of structure do you prefer? Please explain.
☐ Variable-based ☐ Sentence-based
17. *Summary 1* & 2 both focus on research variables and relationships. However, in *Summary 2*, the concepts are categorized, and important concepts are highlighted. Do you find concept categorization in *Summary 2* useful? Please explain.
☐ Not useful at all ☐ A little useful ☐ Quite useful ☐ Very useful
18. Do you find concept highlighting in *Summary 2* useful? Please explain.
☐ Not useful at all ☐ A little useful ☐ Quite useful ☐ Very useful
19. *Summary 3* & 4 both focus on important sentences. However, *Summary 3* mainly contains research objectives, whereas *Summary 4* contains sentences that are ranked as important. Which one do you prefer? Please explain.
☐ Mainly research objective sentences ☐ Sentences ranked as important

F. Suggestions

20. What improvements would you suggest for each of the summaries, if any?