Automatic Tailored Multi-Paper Summarization based on Rhetorical Document Profile and Summary Specification

Masayu Leylia Khodra[1], Dwi Hendratmo Widyantoro[1], E. Aminudin Aziz[2] & Bambang Riyanto Trilaksono[1]

[1]School of Electrical Engineering and Informatics, Bandung Institute of Technology,
Jalan Ganesa No.10, Bandung 40132, Indonesia
[2]Faculty of Language and Arts Education, Indonesia University of Education,
Jalan Dr. Setiabudhi No. 229, Bandung 40154, Indonesia
Email: masayu@stei.itb.ac.id

Abstract. In order to assist researchers in addressing time constraint and low relevance in using scientific articles, an automatic tailored multi-paper summarization (TMPS) is proposed. In this paper, we extend Teufel's tailored summary to deal with multi-papers and more flexible representation of user information needs. Our TMPS extracts Rhetorical Document Profile (RDP) from each paper and presents a summary based on user information needs. Building Plan Language (BPLAN) is introduced as a formalization of Teufel's building plan and used to represent summary specification, which is more flexible representation of user information needs. Surface repair is embedded within the BPLAN for improving the readability of extractive summary. Our experiment shows that the average performance of RDP extraction module is 94.46%, which promises high quality of extracts for summary composition. Generality evaluation shows that our BPLAN is flexible enough in composing various forms of summary. Subjective evaluation provides evidence that surface repair operators can improve the resulting summary readability.

Keywords: *BPLAN; multi-paper summarization; Rhetorical Document Profile; summary specification; tailored summary; user information needs.*

1 Introduction

A large number of scientific articles lead to an increased effort in selecting the most relevant papers and reading them. Although reading the abstracts may help resolve the problems, readers prefer to use survey papers because reading these papers can be considered more effective than reading the abstracts [1]. A survey paper is a synthesis of critical analysis of some primary papers and its authors' perspectives in a domain [2], which provides a general understanding about the domain. To our knowledge, there is little work reported about generating survey paper automatically, but the resulting summary was still not readily consumable [3]. Existing researches on generating survey paper still focused on extracting main ideas of each papers, which is known as multi-paper summarization.

Received December 12th, 2012, Revised December 26th, 2012, Accepted for publication December 27th, 2012.
Copyright © 2012 Published by LPPM ITB, ISSN: 1978-3086, DOI: 10.5614/itbj.ict.2012.6.3.4

Summarizing multiple papers should be more challenging than that of only single paper because of three difficult activities [4]: (a) collecting the primary papers in a domain, (b) extracting useful information that describe the similarities and differences among the papers, and (c) generating new ideas-based sentences that cannot be extracted directly from the source papers. As a preliminary step in generating survey paper, this research focuses on activity (b). Activity (a) is replaced by manually inputting a set of related papers from users. Activity (c) is the most difficult task and still left for future research.

The majority of works in multi-paper summarization research area were focused on identifying important concepts in scientific abstracts [5-7] and identifying the abstract structures of rhetorical classification [6-7]. However, all existing works performed multi-paper summarization only on paper abstracts. Since it is obvious that full papers have more important contents than the abstracts, we employ summarization on full papers in this research.

A summarization system commonly produces a single version of summary for a particular reader's information needs; see for instance [3-7]. However, researchers who will use of this system may have various information needs due to different relevance judgments [8]. Relevance is a concept about users' judgments of quality of the relationship between information and information need at a certain point in time [9]. Teufel [10] has proposed a tailored summary, which is one that is created in accordance with the user information needs. To our knowledge, a tailored summary has been reported in the literature only for single paper summarization, and it cannot be directly applied to summarize multi-papers.

Our research aims to propose automatic tailored multi-paper summarization (TMPS) to assist researchers in addressing time constraint and low relevance. Specifically, TMPS combines multi-paper summarization and tailored summaryzation. This research shows how to adapt and develop tailored multi-paper summarizer to produce a summary from a set of full papers with more flexibility in describing user information needs as summary specification. Similar with tailored summary proposed by Teufel [10], our TMPS framework is based on Rhetorical Document Profile (RDP) [10], which is a rhetorical structured representation of a paper. As observed by Teufel [10], rhetoric information is the intention to be conveyed to the reader by an author of the paper. Compared to existing works in this area [5-7],[10], our main contributions are: (1) designing TMPS framework based on RDP and summary specification for multi-papers; (2) developing BPLAN (Building Plan Language) to provide summary specification. In BPLAN, some operators of surface repair are designed to allow a more readable summary.

The rest of the paper is organized as follows. In section 2, related works in this area are presented. Section 3 describes RDP structure based on rhetorical scheme originally proposed by Teufel [11]. In section 4, TMPS framework is described, and associated modules that build the framework are presented. In section 5, evaluation of RDP extraction module, generality evaluation, and subjective evaluation are discussed. Concluding remarks are presented in section 6.

2 Related Work

A summarization system transforms reductively a source text or collection of texts into a single summary through content condensation by selecting and integrating important contents in the sources [12]. Summarization systems are commonly classified as extractive and non-extractive, although there is no absolute distinction between them [12]. Extractive approach creates summary by selecting source sentences or its constituents, and focuses on how to identify important sentences in text [13],[14]. Non-extractive method, particularly abstractive method, creates summary without using extraction, and focuses on information extraction, information fusion, and compression [15].

Existing multi-paper summary can be composed of concepts [5],[7], sentences [3],[6], or text fragments [16]. Fiszman and Rindflesch [5] developed a semantic abstraction approach to identify important concepts and generate a semantic network as a multi-paper summary from a set of scientific abstracts in the biomedical domain. Shiyan [7] proposed variable-based approach to generate concept-based summary from dissertation abstracts in sociology domain. Macrostructure and microstructure-based summarization was built by Jiaming [6] to process a set of abstracts of engineering technical reports. Agarwal [16] developed clustering-based summarization from fragments of co-cited papers. The majority of works processed set of abstracts or fragments, and produced one version of multi-paper summary without considering user information needs. In contrast to those works, our research focuses on summarizing a set of full-papers resulting a sentence-based summary, and produced summary based on user information needs.

3 Rhetorical Document Profile

Rhetorical Document Profile (RDP) is an instantiated template consisting of rhetorical slots where each slot contains sentences with specific rhetorical category. This structured representation is combined with building plan to provide the flexibility of summary contents.

Rhetorical scheme was originally introduced with 7 categories [10] and recently was refined into 15 categories since the refined set is more informative, better at recognizing the structure of problem solving, and more subtle in describing a difference [11]. This research employs the refined version. Table 1 provides a short description of each category.

Table 1 Rhetorical scheme with 15 categories [11].

Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant)
OTHR	Significant knowledge claim held by somebody else. Neutral description
PREV_OWN	Significant knowledge claim held by authors in a previous paper. Neutral description.
OWN_MTHD	New Knowledge claim, own work: methods
OWN_FAIL	A solution/method/experiment in the paper that did not work
OWN_RES	Measurable/objective outcome of own work
OWN_CONC	Findings, conclusions (non-measurable) of own work
CODI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)

4 Tailored Multi-Paper Summarization System

In order to generate a multi-paper summary, our summarizer accepts a summary specification and a set of input papers on one related topic. User summary specification is written using our new building plan language, which is called BPLAN (Building Plan Language). Unlike Teufel's building plan, our BPLAN is more dynamic and designed for multi-paper summarization.

Our summarizer consists of three main modules: preprocessing, RDP extraction, and summary presentation, as shown in Figure 1. The preprocessing module reads each input paper (pdf) and saves its contents and structures into xml. The extraction module evaluates each sentence to determine its rhetorical category, and produces the corresponding RDP of each paper. The summary presentation module processes all filled RDPs to generate multi-paper summary with respect to a BPLAN-based summary specification.

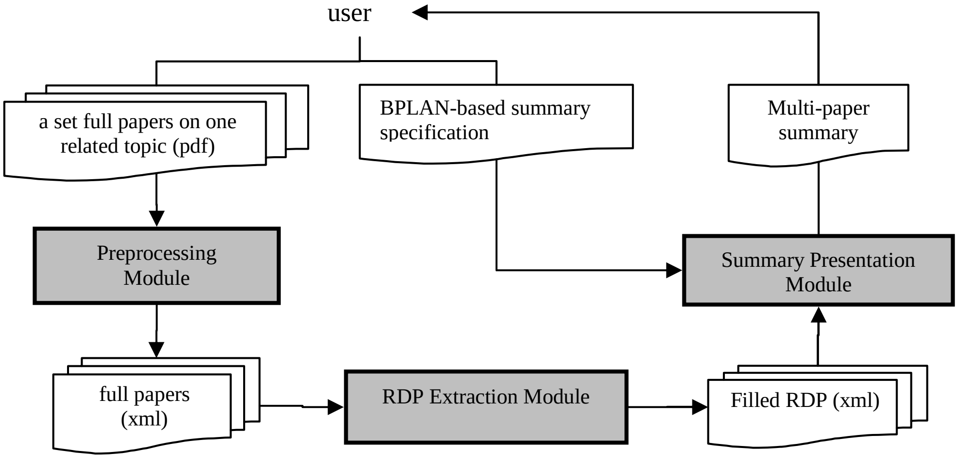

user
a set full papers on one related topic (pdf)
BPLAN-based summary specification
Multi-paper summary
Preprocessing Module
Summary Presentation Module
full papers (xml)
RDP Extraction Module
Filled RDP (xml)

Figure 1 Diagram of tailored multi-paper summarization system architecture.

4.1 Preprocessing Module

The input data are a set of full papers on one related topic. In our research, the papers are retrieved from ACL-Anthology Reference Corpus (ACL-ARC), which is a corpus of scholarly publications about Computational Linguistics [17]. The preprocessing module transforms each input paper from pdf format into xml format as follows:

1. Text of pdf document is extracted by using PDFBox 1.6.0 [18].
2. The text is divided into sections using the bookmarks in the pdf file. If no bookmark information is supplied, user has to define the section list.
3. Each section is divided into paragraphs by using a common paragraph delimiters, which are carriage return and line feed [19].
4. A list of sentences in a paragraph is extracted by employing Maximum Entropy-based sentence detector from OpenNLP 1.5.0 [20]. This sentence detector estimates joint probability of a potential punctuation character and its surrounding context [21].
5. The parsed sentences are arranged in its original hierarchical structures (section, paragraph, sentence) and saved in xml format.

4.2 RDP Extraction Module

RDP extraction module returns filled RDP slots with rhetorical sentence classification for each input paper. Given a set of rhetorical categories as shown in Table 1, each sentence is classified to determine its rhetorical category. Rhetorical classifier is automatically constructed from a training set by using a supervised learning algorithm. In this paper, Table 2 shows our feature set adapted from Teufel's features [10] for rhetorical classification, and * indicates features that are considered to provide better performance. The last column of Table 2 shows the range of values of each feature.

Table 2 Feature pool for rhetorical classifier adapted from Teufel's features [10].

Type	Name	Description	Values
Content	Cont-1	Incidence of significant terms of document	Boolean
	Cont-2	Incidence of words occurring in document title or section title	Boolean
	Cont-3*	Incidence of significant terms of abstract	Boolean
Absolute location	Loc	Sentence position within document relation to 10 segments	A-J
Explicit structure	Struct-1	Sentence position within section	7 values
	Struct-2	Sentence position within paragraph	Initial, medial, final
	Struct-3	Prototypical type of section title	17 prototypical titles or Non-Prototypical
Sentence length	Length	Is the sentence longer than 15 words?	Boolean
Syntax	Syn	Is the 1st finite verb modified by modal auxiliary?	Boolean
	Adj*	Incidence of qualifying adjective	Boolean
Citations	Cit-1	Citation or self citation incidence	Citation, self citation, none
	Cit-2	Citation location in sentence	Beginning, middle, end, none
Formulaic expression	$\text{Formu}_{1..21}$*	Incidence of each formulaic expression in sentence	Boolean
Agentivity	$\text{Ag-1}_{1..16}$*	Incidence of each agent type	Boolean
	$\text{Ag-2}_{1..9}$*	Incidence of each action type	Boolean
	Negation	Incidence of negation in sentence	Boolean

Similar to Teufel [10], the features are grouped in eight feature types as follows. The first group is content features to indicate whether a sentence has significant terms in its document (Cont-1), its abstract (Cont-3), and its titles (Cont-2).

Absolute location and explicit structures are expected to show the usual location of particular rhetorical sentences in a paper. Absolute location (Loc) defines 10 differently-sized segments that represent the structure of ideal documents [10]. While Loc represents global locational structure, features of explicit structures represent the internal locational structure of section (seven values for Struct-1) and paragraph (three values for Struct-2). Moreover, prototypical titles (Struct-3) have fixed seventeen prototypical titles or NonPrototypical [10]. For example, sentences about future research are commonly found at the end of papers. Its Loc value is J, Struct-3 value is Conclusion, and values for Struct-1 and Struct-2 depend on sentence position in its section and paragraph.

Sentence length is used to show sentence complexity that indicates the characteristics of some particular rhetorical sentences. For example, sentences about method commonly describe the details of the solution, and tend to be lengthy and less complex than other rhetorical sentences [10].

Syntax features are expected to be the indicators of rhetorical structures. Modality feature (Syn) correlates for hedging that is used by authors to discuss the results of their research [10]. Qualifying adjective feature (Adj) is commonly used to indicate a particular rhetorical category, for example to conclude experiment results.

Citations are indicators of other researcher's work. Sentences of some rhetorical categories such as *use*, *support*, or *antisupport* can be recognized by incidences of citations and its citation location, but sentences of other categories such as *aim* or *own_res* do not use citations.

Lastly, formulaic expression and agentivity are the most important indicators of rhetorical categories. These features are known as meta-discourse features. Hyland [22] pointed out that metadiscourse is more generally seen as the author's linguistic and rhetorical manifestation in the text in order to bracket the discourse organization and the expressive implications of what is being said. These features were used to capture a profile of "who-does-what" sentence structure with some syntactic variations. It means that the agent types capture the person involved, action types capture what actions are conducted, and formulaic expressions (FE) capture the structure. For this purpose, Teufel [10] used three metadiscourse features, i.e. formulaic expression, the type of agent, and type of action. Teufel [10] restricted one value per sentence for simplification. Since a sentence can have more than one value, each possible value of formulaic expression/agent type/action type is modified to be one boolean feature. Finally, there are 21 formulaic expression features, 16 agent features, and 9 action features. As an illustration, Figure 2 shows a sentence and its metadiscourse features extracted by using some syntactical patterns. For example, pattern `@self_nom @presentation_act` matches with `we concern`, where `@self_nom` is replaced by `we`, and `@presentation_act` is replaced by `concern`.

In this paper, Support Vector Machine (SVM) [23] is employed to build our rhetorical classifier. SVM has been proved to be superior in various text classification tasks [24-27]. SVM was originally designed for binary classification problem [28-29], but rhetorical classification is a multiclass problem. Fortunately, multiclass problem can be reduced to multiple binary classification problems [30]. It means that there is a set of different binary classifiers to handle this multiclass problem.

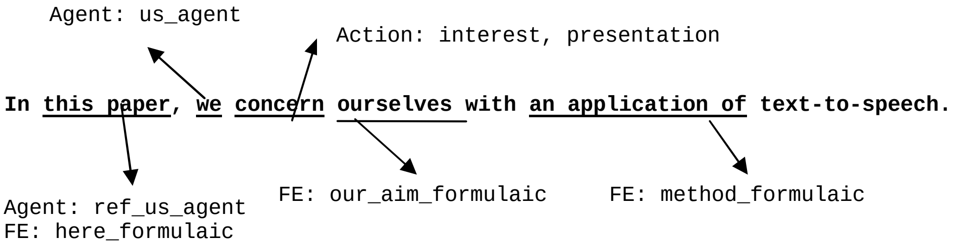

Agent: us_agent
Action: interest, presentation
In this paper, we concern ourselves with an application of text-to-speech.
Agent: ref_us_agent
FE: here_formulaic
FE: our_aim_formulaic
FE: method_formulaic

Figure 2 Example of meta discourse features extracted from a sentence.

4.3 BPLAN Summary Specification

In tailored summarization, user information needs is an important factor in addition to input documents. Teufel's tailored summarizer is proposed by identifying user types as the combinations of parameter values of user information needs. Since scientific paper is represented as Rhetorical Document Profile (RDP), building plan allows flexibility of summary contents by specifying the number and type of RDP slot fillers in the summary. The building plan allows user to easily describe summary composition as his information needs, and let the summarizer to easily understand the information needs. In Teufel's building plan, user needs to list the summary composition for each input paper, and thus it is applicable for a fixed number of input papers. Consequently, a separate building plan needs to be defined for the same user type when the system summarizes from a different number of input papers.

Unlike Teufel's building plan, our system accepts a more flexible building plan that describes not only summary composition, but also group of sentences in paragraphs as well as rule set for surface repair. Specific language, BPLAN (Building Plan LANguage), is designed for encoding this summary specification. Figure 3 shows BPLAN global syntax.

BPLAN needs collection_identifier as a pointer to a set of input papers, and it must be declared in `collection_declaration`. BPLAN provides two main blocks of summary specification, i.e. `composition_block` to specify summary composition, and `surface_repair_block` to specify operations that can make the summary more readable. The next section describes the detailed syntax for both blocks.

```
<building_plan>           ::= <collection_declaration>
                              <composition_block> <surface_repair_block>
<collection_declaration>  ::= GIVEN <collection_identifier>

<collection_identifier>   ::= <identifier>
<identifier>              ::= @[<letter>|_] [<letter>|<digit>|_]*
```

Figure 3 BPLAN global syntax.

4.3.1 BPLAN for Summary Composition

FOREACH construct is introduced to compose a multi-paper summary. FOREACH construct adds the same composition into summary from each input paper in collection_identifier. Figure 4 shows BPLAN syntax of summary composition specification.

```
<composition_block>      ::= [<create_paragraph>][<foreach_block>]+
<create_paragraph>       ::= CREATE PARAGRAPH
<foreach_block>          ::= FOREACH <for_identifier> IN
                             <collection_identifier> DO
                             [<simple_statement>]+ END
<simple_statement>       ::= <create_paragraph>|<add_statement>
<add_statement>          ::= ADD(<for_identifier>.
                             <rhetorical_category>,<number>)

<for_identifier>         ::= <identifier>
<rhetorical_category>    ::= AIM|ANTISUPP|CODI|CO_GRO|FUT|GAP_WEAK|
                             NOV_ADV|OTHR|OWN_CONC|OWN_FAIL|OWN_MTHD|
                             OWN_RES|PREV_OWN|SUPPORT| USE
<number>                 ::= [<digit>]+
```

Figure 4 BPLAN syntax of summary composition block.

This summary composition specification (composition_block) consists of two parts. The first part is creating paragraph explicitly (CREATE PARAGRAPH) to allow users to arrange sentences in summary. This statement can be written before or in the FOREACH block. If it is written before a FOREACH block, one new paragraph will be created and all sentences that are added by the FOREACH block will be arranged in the same paragraph. If statement CREATE PARAGRAPH is written in a FOREACH block, a new paragraph will be created for each iteration i.e. each paragraph is composed of sentences from the same input paper. The second part is adding a number of rhetorical sentences (add_statement) to summary. The rhetorical category and number of sentences are the parameters of this operator ADD.

As an illustration, suppose that there are n topic-related papers in the collection, and user needs indicative summary to determine relevance of each paper to the topic. The summary is designed to consist of short summaries of each paper. Figure 5 shows the corresponding summary composition. This summary specification produces a summary consisting of n paragraphs.

```
//short summaries:  each paragraph consists of research goal and research
method from a paper.
foreach @p in @paper_set do
    create paragraph
    add(@p.aim, 1),
    add(@p.own_mthd, 1),
end
```

Figure 5 Example of summary composition in summary specification

4.3.2 BPLAN for Surface Repair

Our surface repair is designed to be flexible, i.e. the applicable rules can be specified in the user summary specification. As shown by Figure 6, surface repair specification (`surface_repair_block`) consists of three parts. The first part is `create_ruleset` statement. The second part declares identifiers using `declare_statement`. Identifiers define vocabularies in original sentences (using `define_identifier`) and in target summary (using `template_identifier`).

```
<surface_repair_block>      ::= <create_ruleset><declare_statement>
                                <surface_repair_statement>
<create_ruleset>            ::= CREATE RULESET
<declare_statement>         ::= DEFINE <define_identifier>=<define_list>
                                |TEMPLATE<template_identifier>=
                                <template_list>
<surface_repair_statement>  ::= <substituteSubject>|<removePhrase>|
                                <toActive>|<npSubjectToVp>
<substituteSubject>         ::= SUBSTITUTESUBJECT(<define_identifier>,
                                <template_identifier>)
<removePhrase>              ::= REMOVEPHRASE(<define_identifier>)
<toActive>                  ::= TOACTIVE(<rhetorical_category>)
<npSubjectToVp>             ::= NPSUBJECTTOVP(<rhetorical_category>)
```

```
<define_identifier>         ::= <identifier>
<template_identifier>       ::= <identifier>
<template_list>             ::= {[<template_element>][,
                                <template_element>]}
<template_element>          ::= <element>
<define_list>               ::= {[<define_element>][, <define_element>]}
<define_element>            ::= <element>|<penn_treebank_tag>
<penn_treebank_tag>         ::= CC|CD|DT|EX|FW|IN|JJ|JJR|JJS|LS|...|WRB
<element>                   ::= <string>
<string>                    ::= “[<letter|digit>]*”
```

Figure 6 BPLAN syntax of surface repair block.

Figure 7 shows some examples how to declare these indentifiers. The contents of element of define_identifier can be words or tags. Two examples of define_identifier represent list of words (`@example_words`) or word tags (`@example_tags` using Penn Treebank tags) in original sentences. The next examples are template_identifier (`@exampleT1` and `@exampleT2`) for summary sentences. `@exampleT1` contains selections of words to substitute main verbs. Meanwhile, `@exampleT2` represents writing format in target summary.

```
define @example_words = {"we","this paper"}
define @example_tags = {"JJR","JJS","RBR","RBS"}
template @exampleT1 = {"aim to","intend to"}
template @exampleT2 = {"@author (@year)"}
```

Figure 7 Some examples of identifier declaration.

The last part of surface repair specification is `surface_repair_statement`. BPLAN provides four operators to improve sentence readability in the summary. The following provides details of each operator.

`substituteSubject` Operator

Some extracted sentences of scientific articles often used subjects "`we`" or "`this paper`" as a reference to its authors. Although it can be left unmodified in a single-paper summary, it can cause problem in multi-paper summary. As shown below, the words "`we`" are ambiguous because it could refer to different authors.

```
In this paper, we propose a learning-based approach to combine various
sentence features. We present a study that explores the summary space of
each domain via an exhaustive search strategy.
```

Operator substituteSubject is provided to replace the subject words of a sentence with its authors' names. Figure 8 shows the syntax and usage examples of this operator. The first parameter defines vocabulary for subject terms in the original sentence, and the second parameter specifies replacement terms and format in target summary. The words `we` in previous example are replaced by `Wong et al. (2008)` in the first sentence, and `Ceylan et al. (2010)` in the second sentence. If a paragraph consists of sentences from the same scientific paper, the authors' names are applied only in the first sentence in that paragraph, and use terms the author[s] in the rest of sentences in the paragraph.

```
Syntax: <substituteSubject> ::= SUBSTITUTESUBJECT(<define_identifier>,
                                <template_identifier>)

Summary specification:
substituteSubject(@subject_author,@author_surface)
define @subject_author = {"we","this paper"}
template @author_surface={"@author (@year)"}

Sentences with substituteSubject operator:
In this paper, Wong et al. (2008) propose a learning-based approach to combine
various sentence features. Ceylan et al. (2010) present a study that explores
the summary space of each domain via an exhaustive search strategy.
```

Figure 8 Syntax and examples using `substituteSubject`

`removePhrase` Operator

Another problem in multi-paper summarization is the use of references to other objects in a paper, for example term ”`in this paper`”. As shown in Fig. 8, the term is no longer valid in multi-paper summary. Operator `removePhrase` is designed to remove these invalid references (defined by `@remove_phrase`) from the original sentences (see Figure 9). The only parameter for this operator is list of terms to remove in target summary.

```
Syntax: <removePhrase> ::=  REMOVEPHRASE(<define_identifier>)

Summary specification:
removePhrase(@remove_phrase)
define @remove_phrase = {"in this paper", "in this work", " (see CREF )",
" in CREF"}

Sentences without removePhrase operator:
In this paper,  Wong et al. (2008) propose a learning-based approach to
combine various sentence features.

Sentences with removePhrase operator:
Wong et al. (2008) propose a learning-based approach to combine various
sentence features.
```

Figure 9 Syntax and examples using `removePhrase`.

`toActive` Operator

Since an active voice sentence is more readable than a passive voice sentence [31], operator toActive is provided to transform passive voice into active voice sentence. The sentences that will be transformed are those whose rhetorical category is specified in the operator parameter. Figure 10 shows the usage example of this operator.

```
Syntax: <toActive> ::= TOACTIVE(<rhetorical_category>)

Summary specification:toActive(aim)

Sentences without toActive operator:
A two-step talker-location algorithm is introduced

Sentences with toActive operator:
We introduce a two-step talker-location algorithm.
```

Figure 10 Syntax and examples using `toActive`.

`npSubjectToVP` Operator

Scientific writing often puts emphasis on the experiment or process being described [31]. Some authors write nominalization of main verb to give emphasis. For example the sentence in Figure 11, the subject word investigation is a result of nominalization of the verb investigate. Different from scientific writing, the writing of multi-paper summary puts emphasis on who-do-what. Therefore, in this paper, operator npSubjectToVP provides sentence transformation where subject nominalization word is changed with more natural sentence (Subject-Verb-Object) by verbalisation. We use morpho-semantic WordNet database [32] to get verb-noun pairs, such as investigate (verb) – investigation (noun), invent (verb) -invention (noun) pair. This verb-noun pairs connect similar-meaning words from different classes [33]. Similar to toActive, this operator only needs one parameter, which defines applicable sentence in respect to its rhetorical category, as shown by Figure 11.

```
Syntax: <npSubjectToVp> ::= NPSUBJECTTOVP(<rhetorical_category>)

Summary specification: npSubjectToVp(aim)

Sentences without npSubjectToVp operator:
Our investigations  involve problems which are not currently well
understood.

Sentences with toActive operator:
We investigate problems which are not currently well understood
```

Figure 11 Syntax and examples using `npSubjectToVP`.

4.4 Summary Presentation Module

The summary presentation module accepts a user summary specification and a set of filled RDP from topic-related papers, and returns a summary. Based on the summary specification, generating summary requires two sequential processes, i.e., selecting sentences based on summary composition, and repairing sentences with respect to rules of surface repair in the specification.

As described above, the number of sentences to add into summary is determined using parameter add statement. If the available sentences are more than the required, then our system will select the necessary sentences using Maximal Marginal Relevance (MMR) [34]. MMR is applied to reduce the redundancy because this method balances the centrality of a sentence and its novelty compared to the sentences that have been selected [6]. The MMR is based on Eq. (1).

$$MR(s_i) = sim(s_i, C) - \max_{s_j \in S} sim(s_i, s_j) \tag{1}$$

where s_i is the current sentence, C is candidate sentences to be selected, s_i is a member of C, and S is the set of selected sentences. The similarity between sentences is measured using cosine similarity. For each sentence in C, MR score is calculated, and the sentence with maximum MR score is selected.

All selected sentences are checked whether they fulfill the preconditions of surface repair operators. If a sentence satisfies more than one operator, sequence of surface repair operators will be applied.

5 Evaluation

In this section, we evaluate the performance of our summarizer with regard to three aspects, i.e. accuracy of RDP extraction module, generality of the BPLAN-based summary specification, and effectiveness of surface repair to improve summary readability.

5.1 Evaluation of RDP extraction module

One of the important factors in generating multi-paper summary is the summarizer's ability to classify the rhetorical category of sentences. Therefore, we need to evaluate the performance of the RDP extraction module. The evaluation necessitates the use of a rhetorical corpus. Due to the unavailability of such corpus, we constructed one based on an ACL-ARC paper collection. A rhetorical category is assigned for each sentence of the 75 papers in the collection. The result is an annotated corpus of 10877 rhetorically labeled sentences [35]. This corpus is then randomly split into a training set and a test set. The training set consists of sentences from two-thirds of the total number of papers in the corpus (50 papers), while sentences from the remaining 25 papers are used as the test set. We built a binary classifier for each rhetorical category. The performance of each rhetorical binary classifier of the system is shown in the last column of Table 3

Table 3 Description of dataset and SVM classifier performances.

Category	Training set		Test set		Testing Accuracy Rate
	Positive	Negative	Positive	Negative	
AIM	136	7103	77	3561	98.46%
NOV_ADV	179	7060	68	3570	97.97%
CO_GRO	271	6968	113	3525	97.22%
OTHR	528	6711	444	3194	87.30%
PREV_OWN	471	6768	150	3488	96.04%
OWN_MTHD	3608	3631	1717	1921	65.97%
OWN_FAIL	46	7193	24	3614	99.34%
OWN_RES	264	6975	155	3483	95.66%
OWN_CONC	385	6854	193	3445	94.26%
CODI	69	7170	42	3596	98.85%
GAP_WEAK	241	6998	124	3514	96.67%
ANTISUPP	36	7203	24	3614	99.37%
SUPPORT	284	6955	109	3529	96.12%
USE	244	6995	196	3442	94.61%
FUT	113	7126	38	1.04%	99.01%

As shown in Table 3, the performance of the binary classifiers in the majority of categories is quite high. In eleven categories, the accuracy exceeds 95%, while in three others it ranges between 87.30% and 94.61%. Due to false positives, however, the classifier for OWN_MTHD has a rather low accuracy of 65.97%. OWN_MTHD sentences have more various patterns to be identified by its classifier, for example, to describe research activities (e.g.`“We employ a bilingual thesaurus.”`), to describe a concept (e.g.`“Each noun has similar words in the corpus-based thesaurus.”`), to explain a formula or case (e.g.“Suppose w is a word to be translated.”), and to describe another objects in paper (e.g.“Step 1 is shown by Figure 2.”).

Although there is still room for improvement, with an average accuracy rate of 94.46% for all categories, we believe that these classifiers can be justifiably used in extracting RDP to generate good summaries.

5.2 Generality Evaluation of Summarizer

The purpose of generality evaluation is to test whether a sufficient level of generality is present in the evaluated summarizer. It can be considered sufficient if this summarizer can produce some existing forms of paper summaries. As described before, Teufel’s tailored summary [10] proposed a single paper summary for each user type, while Jiaming’s summary [6] produced general and unique information for each paper. We will show the summary specifications for these summaries, and the results given by our system.

For single-paper tailored summary, Table 4 shows some summaries produced automatically from the same paper “Event Based Extractive Summarization” based on the summary specifications for different user types. These examples show that our BPLAN provides a flexible structure to define different summary compositions. In addition to the summary compositions, the same rules of surface repair are written for these summaries.

Table 4 Summaries and their summary specifications for different user types.

User type	Summary specification	Summary of paper “Event Based Extractive Summarization”

General purpose, short, informed reader	given @paper_set
	create paragraph
	foreach @p
	in @paper_set
	do
	add(@p.aim,2)
	end
		Filatova and Hatzivassiloglou (2004) investigate the effect this new feature has on extractive summarization, compared with a baseline feature set consisting of the words in the input documents, and with state-of-the-art summarization systems. **The authors** discuss a general model which treats summarization as a three component problem, ...

General purpose, short, uninformed reader	given @paper_set
	create paragraph
	foreach @p
	in @paper_set
	do
	add(@p.co_gro,1)
	
	add(@p.gap_weak,1)
	add(@p.aim,2)
	end
		The main goal of extractive summarization can be concisely formulated as extracting from the input pieces of text which contain the information about the most important concepts mentioned in the input text or texts. **Filatova and Hatzivassiloglou (2004)** investigate the effect this new feature has on extractive summarization, compared with a baseline feature set consisting of the words in the input documents, and with state-of-the-art summarization systems. **The authors** discuss a general model which treats summarization as a three component problem, ...

Jiaming [6] identified topics from a set of scientific papers, and produced summaries for each topic. A topic summary consists of two parts (general information and unique information from each paper). For example, Figure 12 and Figure 13 show summary specifications and two-part summaries generated by our TMPS system. It is shown that our summarizer is flexible in composing various forms of summary.

```
Summary specification for general information:
create paragraph
foreach @p in @paper_set do
   add(@p.co_gro, 1, @author_surface),
end

create ruleset
template @author_surface={"(@author, @year)"}

Summary specification for unique information from each paper:
foreach @p in @paper_set do
    create paragraph
    add(@p.aim, 1),
end
```

Figure 12 Two summary specifications based on Jiaming's summary [6].

General information:

Automatic text summarization involves condensing a document or a document set to produce a human comprehensible summary (Wong et al., 2008). Extractive summarization selects sentences which contain the most salient concepts in documents (Li et al., 2006).

Unique information from each paper:

Wong et al. (2008) propose a learning-based approach to combine various sentence features.

Li et al. (2006) define an event as one or more event terms along with the named entities associated, and present a novel approach to derive intra- and inter- event relevance using the information of internal association, semantic relatedness, distributional similarity and named entity clustering.

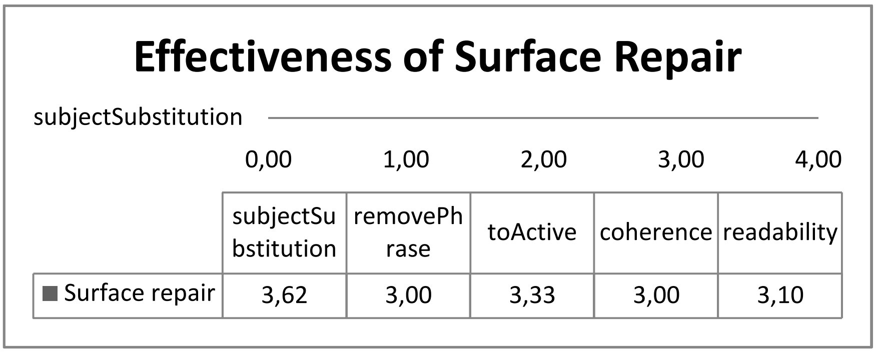
Figure 13 Two part summaries based on Jiaming's summary [6].

5.3 Subjective Evaluation of Surface Repair Effectiveness

This research also evaluates the effectiveness of surface repair operators in improving summary readability. We conduct questionnaire-based evaluation that is commonly employed in information science domain [36]. In this evaluation, questionnaires are sent to 50 respondent candidates who posses background knowledge on natural language processing or frequently use survey papers. The questionnaire is designed to compare some summary pairs with and without surface repair operators. Table 5 shows the example of summary pairs for evaluating subject substitution operator. 21 respondents reply our questionnaires by giving scores for the each operators: 1 if the operator decreases readability, 2 for no improvement, 3 for little improvement, or 4 for significant improvement. Figure 14 shows the scores of the effectiveness.

Table 5 Example of summary pair in questionnaire for subjective evaluation.

Summary without subject substitution	**Summary with subject substitution**
`Automatic text summarization involves condensing a document or a document set to produce  a human comprehensible summary.`	`Automatic text summarization involves condensing a document or a document set to produce a human comprehensible summary (Wong et al., 2008).`



	subjectSubstitution	removePhrase	toActive	coherence	readability
■ Surface repair	3,62	3,00	3,33	3,00	3,10

Figure 14 Scores for effectiveness of surface repair processes.

Subject substitution operator gets the best score because it clarifies the authorship of the work (average score of 3.62). Remove phrase operator and coherence gives little improvement (average score of 3.0). Sentence transformation from passive to active voice gives better improvement (average score of 3.33). For readability, the average score is 3.1, which indicates surface repair operator gives little improvement. Furthermore, the study shows that respondents have different preferences in using surface repair. It confirms our design approach which specifies flexible surface repair in our summarizer.

6 Conclusions

In this paper, we have described an automatic tailored multi-paper summarization system which has the ability to generate a multi-paper summary from a set of scientific papers based on user summary specification. The system adapted Teufel's tailored summary which employs Rhetorical Document Profiles (RDPs) and building plan to achieve flexibility in composing multi-paper summaries. The summary specification of the system has a more flexible representation than user type in the original tailored summary as proposed by Teufel. Additionally, the system offers surface repair for better readability of the resulting summary.

By providing a new BPLAN (Building Plan Language), which is a language to write summary specification, our tailored multi-paper summarization (TMPS) system has the ability to produce flexible summaries. Our system also provides some operators for surface repair to make more readable summaries. In generality evaluation, we show that our summarizer is flexible in composing various forms of summary. Effectiveness evaluation of surface repair operators point out that subject substitution and passive-to-active sentence transformation are the most effective operators.

Acknowledgment

Thanks to Renny Pradina and Ridho Reinanda for the proofreading.

References

[1] Maxie, G., *Critical Writing and Reading of Review Articles*, The Canadian Veterinary Journal, **31**(6), pp.413-414, 1990.

[2] Torraco, R.J., *Writing Integrative Literature Reviews: Guidelines and Examples*, Human Resource Development Review, **4**(3), 356-367, 2005.

[3] Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D. & Zajic, D., *Using Citations to Generate Surveys of Scientific Paradigms*, in Proc. of HLT/ NAACL 2009.

[4] Wang, M., Tanaka, H. & Zhong, Y., *Generating Summaries of Multiple Technical Articles*, in Proc. of Sino-Japan Symposium on IIN, 2000.

[5] Fiszman, M. & Rindflesch, T.C., *Abstraction Summarization for Managing the Biomedical Research Literature*, in Proc. of HLT/NAACL 2004.

[6] Jiaming, Z. *Exploiting Textual Structures of Technical Papers for Automatic Multi-Document Summarization*, PhD Thesis, NUS, 2008

[7] Shiyan, O., Khoo, C.S.G. & Goh, D.H., *Design and Development of A Concept-based Multi Document Summarization System for Research Abstracts*, Journal of Information Science, **34**, pp.308-326, 2008.

[8] Saracevic, T., *Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance*, Journal of The American Society For Information Science And Technology, **58**, pp. 2126-2144, 2007.

[9] Borlund, P., *The Concept of Relevance*, Journal of The American Society for Information Science and Technology, **54**, pp. 913-925, 2003.

[10] Teufel, S., *Argumentative Zoning: Information Extraction from Scientific Text*, PhD Dissertation, University of Edinburgh, 1999.

[11] Teufel, S., Siddhartan, A. & Batchelor, C., *Towards Discipline-Independent Argumentative zoning Evidence from Chemistry and Computational linguistics*, in Proc. of Conference on Empirical Methods in NLP 2009.

[12] Jones, K. S., *Automatic Summarising: The State of The Art*, Information Processing and Management, **43**, pp.1449-1481, 2007.

[13] Marcu, D., *Automatic Abstracting*, Encyclopedia of Library and Information Science, pp.245-256, 2003.

[14] Gupta, V. & Lehal, G.S., *A Survey of Text Summarization Extractive Techniques*, Journal of Emerging Technologies In Web Intelligence, **2** (3), 2010.

[15] Radev, D. R., Hovy, E. & McKeown, K., *Introduction to the Special Issue on Summarization*, Journal Computational Linguistics-Summarization, **28**(4), 2002.

[16] Agarwal, N. & Gvr, K., *Towards Multi-Document Summarization of Scientific Articles:Making Interesting Comparisons with SciSumm*, in Proc. of the Workshop on Automatic Summarization ACL 2011.

[17] Bird, S., Dale, R., Dorr, B.J. & Gibson, B., *The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational linguistics*, in Proc. of Language Resources and Evaluation Conference 2008.

[18] PDFBox, http://pdfbox.apache.org/, (November 2011).

[19] Luebbert, D. L., *Method and System for Handling Text That Includes Paragraph Delimiters of Differing Formats*, US Patent June 1996.

[20] OpenNLP, http://opennlp.sourceforge.net, (April 2011).

[21] Reynar, J.C. & Ratnaparkhi, A., *A Maximum Entropy Approach to Identifying Sentence Boundaries*, in Proc. of the 5th Conference on Applied Natural Language Processing 1997.

[22] Hyland, K. & Tse, P., *Metadiscourse in Academic Writing: A Reappraisal*, Applied Linguistics, **25**, pp.156-177, 2004.

[23] Hsu, C.W., Chang, C.C. & Lin, C.J., *A Practical Guide to Support Vector Classification*, www.csie.ntu.edu.tw/~cjlin/papers/guide/, (December 2009).

[24] Sun, A., Lim, E.P. & Liu, Y., *On Strategies for Imbalanced Text Classification Using SVM: A Comparative Study*, DSS Elsevier 2009.

[25] Chau M., Chen, H., *A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis*, DSS Elsevier, **44**, pp. 482-494, 2008.

[26] Sun, A., Lim, E.P. & Ng, W.K., *Performance Measurement Framework for Hierarchical Text Classification*, Journal of The American Society For Information Science And Technology, **54**, pp. 1014-1028, 2003.

[27] Zhang, Y., Dang, Y., Chen, H., Thurmond, M. & Larson, C., *Automatic Online News Monitoring and Classification for Syndromic Surveillance*, DSS Elsevier 2009.

[28] Hsu, C.W. & Lin, C.J., *A Comparison of Methods for Multiclass Support Vector Machines*, IEEE Transactions On Neural Networks, **13**, pp.415-425, 2002.

[29] Rifkin, R. & Klautau, A., *In Defense of One-Vs-All Classication*, Journal of Machine Learning Research, **5**, pp.101-141, 2004.

[30] Allwein, E.L., Schapire, R.E. & Singer, Y., *Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers*, Journal of Machine Learning Research, **1**, pp. 113-141, 2000.

[31] Hinkel, E., *Tense, Aspect and The Passive Voice in L1 And L2 Academic Texts*, Language Teaching Research, **8**, pp. 5-29, 2004.

[32] WordNet: Standoff Files, http://wordnet.princeton.edu, (April 2012).

[33] Fellbaum, C., Osherson, A. & Clark, P.E., *Putting Semantics into WordNet's "Morphosemantic" Links*, Springer Lecture Notes in Informatics, **5603**, pp. 350-358, 2009.

[34] Carbonell, J., Goldstein, J., *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*, in Proc. of SIGIR 1998.

[35] Khodra, M.L., Widyantoro, D.H., Aziz, E.A. & Trilaksono, B.R., *Information Extraction for Scientific Paper Using Rhetorical Classifier*, in Proc. of ICEEI 2011.

[36] Gorrell, G., Ford, N., Madden, A., Holdridge, P. & Eadlestone, B., *Countering Method Bias in Questionnaire-Based User Studies*, Journal of Documentation, **67**, pp.507-524, 2011.