Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 6, No. 3, Januari 2012 **ISSN 0216 – 0544**

TRANSFORMING RHETORICAL DOCUMENT PROFILE INTO TAILORED SUMMARY OF SCIENTIFIC PAPER

[a]Masayu Leylia Khodra, [b]Mohammad Dimas,
[c]Dwi Hendratmo Widyantoro, [d]E. Aminudin Aziz, [e]Bambang Riyanto Trilaksono
[a,b,c,e]Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung
[d]Fakultas Seni dan Bahasa, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi 229 Bandung
E-Mail: [a] masayu@stei.itb.ac.id

Abstrak

Bagian abstrak makalah ilmiah yang bersifat *author-biased* sering membuat pembaca tidak dapat menemukan informasi yang dibutuhkannya. Konsep *tailored summary* dapat membantu pembaca mendapatkan ringkasan sesuai kebutuhan informasinya. Penelitian ini merupakan penelitian pertama yang mengimplementasikan sistem yang menghasilkan *tailored summary* dari makalah ilmiah. *Tailored summary* menstrukturkan suatu makalah ilmiah dalam representasi *Rhetorical Document Profile* berdasarkan skema retorik lima belas slot. Penelitian ini mengadaptasi *building plan* yang masih menggunakan skema retorik tujuh slot. Penelitian ini juga mengimplementasikan sistem yang membangkitkan *tailored summary*. Setelah mendapatkan ringkasan awal, proses *surface repair* dilakukan untuk meningkatkan keterbacaan ringkasan. Setiap kalimat pada ringkasan awal dikombinasikan dengan frase template berdasarkan metode kombinasi pohon sintaksis. Terdapat lima grup frase template yang digunakan dalam *surface repair*. Dengan mengkonstruksi standar evaluasi yang merupakan hasil penilaian oleh lima responden manusia dan menggunakannya sebagai data pengujian, subsistem pemilihan kalimat terbaik *Maximal Marginal Importance – Multi Sentence* dengan pembobotan TF.IDF memiliki kinerja dengan nilai *precision/recall* 0.61, dan subsistem *surface repair* memiliki kinerja dengan *acceptance* 0.91.

Kata kunci: *Tailored Summary, Rhetorical Document Profile, Building Plan, Maximal Marginal Importance – Multi Sentence, Surface Repair.*

Abstract

Since abstract of scientific paper is author biased, readers' required information may not be included in the abstract. Tailored summary may help them to get a summary based on their information needs. This research is the first one that implements tailored summary system for scientific paper. Tailored summary applies information extraction that transforms a scientific paper into Rhetorical Document Profile, a structured representation of paper content based on rhetorical scheme of fifteen slots. This research adapted building plan that used rhetorical scheme of seven slots. We also implement tailored summary system. After generating initial summary, surface repair is conducted to improve summary readability. Each sentence in initial summary is combined with template phrase based on syntax-tree combination method. There are five groups of template phrases provided in surface repair. We construct evaluation standards by asking five human raters. The best method for sentence selection subsystem that uses Maximal Marginal Importance-Multi Sentence is employing TF.IDF weighting system with precision/recall of 0.61. The surface repair subsystem has acceptance of 0.91.

Key words: Tailored Summary, Rhetorical Document Profile, Building Plan, Maximal Marginal Importance – Multi Sentence, Surface Repair.

INTRODUCTION

Abstract of scientific paper is often the first part to be read because it is brief and concise. However, it is author-biased [1]. The reader cannot find some information because author considered that information as unimportant to be included in the abstract. In assisting the reader to get his needed information from scientific paper, our research aims to investigate automatic summarization on scientific paper to produce tailored summary, which is a summary that takes into account user information needs.

Summarizing paper is the process of extracting important information from a scientific paper and transform it into a shorter text (summary). Automatic summarization is applied computational linguistics to develop intelligent system by acting humanly for natural language processing.

A paper summarization system generally produced generic summary. It is not much different from scientific abstract that is written by the authors. It means that generic summary cannot help user more than abstract. Our research will produce tailored summary based on user task and background knowledge.

Although scientific paper is unstructured document, it has common structures, namely sections and paragraphs as explicit structure, and implicit structure of problem solving in the form of rhetorical information in each sentence. Rhetoric is the intention information to be conveyed to the reader by the author of the paper. Therefore, information in scientific paper can be structured in the form of Rhetorical Document Profile (RDP) [2] based on rhetorical structure.

RDP and tailored summary were proposed as extended concepts of argumentative zoning [2]. No previous study has implemented both concepts, and our study is the first one to implement them. In addition, our study also contributes in adapting the building plan (summary of the composition pattern) for the fifteen rhetorical categories.

There are two main stages to produce a tailored summary of scientific paper: (1) transforming scientific paper into RDP, and (2) transforming RDP to a tailored summary according to user information needs. The first phase of generating RDP is processed by taking all the sentences in the abstract and the main sentence of each paragraph in the other section [3], and then classifying the rhetorical category of each sentence [4]. The second phase uses the user needs to determine the building plan, takes the sentences of RDP, and surface repair. Our focus is to describe series of processes carried out in the second stage and some sample of various summary generated from a scientific paper.

The rest of the paper is organized as follows. The next section provides related work on automatic paper summarization and RDP. Section 3 discusses summarization using RDP, and section 4 describes the evaluation. In the last section, we discuss the conclusions and further research to be conducted.

AUTOMATIC SUMMARIZATION OF SCIENTIFIC PAPER

Automatic summarization has been studied since 1958 for producing scientific abstract in order to facilitate the reader in identifying the topic of scientific paper quickly and accurately [5]. Generating abstract automatically helps the author in making abstract or completing papers without abstract. However, summarizing scientific papers has less attention nowadays because the researchers focus to develop newspaper summarization.

Since summarization methods depends on the genre of the document, methods for news genre can not be applied directly for scientific articles.It is different in focus of identification, position of important information, context, and degree of compression [6].

Existing summarization system usually produced generic summary for all users. Some systems generates different summaries for different topic, and are known as topic-focused summarization [7]. Jiaming [8] identified topics in corpus and created summary for each topic. Moreover, user-focused summarization consideres user information needs [9][10][11] to generate tailored summary [2][12]. Personalized summarization adds personalization by keeping parameter values of user information needs in user profiles [13][14]. User does not need to input the parameter values if he want to use the system again [15].

Summarizing scientific paper generally uses extractive approach. This approach selects

important sentences and arrange the extraction into summary. Advantages of extractive approach is grammatical summary that is still easy to read by humans. The disadvantage is low coherence between the sentences. The main components of extractive summarization is important identification and summary generator. The first component will assess each sentence in the document and select some sentences with the highest score. The second component will put together a collection of sentences that are considered essential to form a summary.

Filho [16] modified GistSumm, newspaper summarizer, by adding structure detector and additional constraint that there is minimum one summary sentence representing each section. He compared this modified system and the original system, and concluded that performance of the modified system is better than the original.

Kupiec [17] developed trainable summarizer by using assumption that every sentence can be binary classified based on its relevance. This approach is popular and it used machine learning.

Qazvinian [18] summarized scientific paper based on clustering of all citations, and produced summary in the form of citation network. The summary is extraction of centroids of clusters. Agarwal [19] also developed summarization system of scientific paper by using clustering of co-citations based on user query.

RHETORICAL DOCUMENT PROFILE

Rhetorical Document Profile (RDP) is representation of extracted information of scientific paper. RDP is filled by argumentative zoning (AZ) [2] or rhetorical classification of topic sentences of paragraphs [20]. Summary of scientific paper is produced by using the information of filled RDP.

Teufel [21] defined fifteen rhetorical categories that state the intention information to be conveyed by the author of the paper. This scheme is also known as AZ-II [21] and it is improvement of AZ scheme that has seven categories [2]. Scheme AZ-II (see Table 1) is more informative because AZ-II can identify problem solving structure better than AZ. This paper uses scheme AZ-II.

In our previous research, filled RDP has been generated as shown by Figure 1. All sentences in each slot are extracts of a scientific paper.

Table 1. Rhetorical Categories of Scientific Paper [21].

Kategori	Deskripsi
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 for the paper)
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)

Title: Extractive Summarization Using Supervised and Semi-supervised Learning
Slot AIM:
In this paper, we propose a learning-based approach to combine various sentence features.
Slot CO_GRO:
Automatic text summarization involves condensing a document or a document set to produce a human comprehensible summary.
Slot OTHR:
Traditionally, features for summarization were studied separately.
...
Slot OWN_MTHD:
We investigate the effectiveness of different sentence features with supervised learning to decide which sentences are important for summarization.
...
Slot SUPPORT:
Recently, semi-structure events (<REF>; <REF>; <REF>) have been investigated by many researchers as they balanced document representation with words and structures.
Slot USE:
An automatic evaluation package, ie, ROUGE (<REF>) is employed to evaluate the summarization performance.

Figure 1. Example of Filled RDP Generated in The First Stage of Tailored Summary (Our Previous Research) [20].

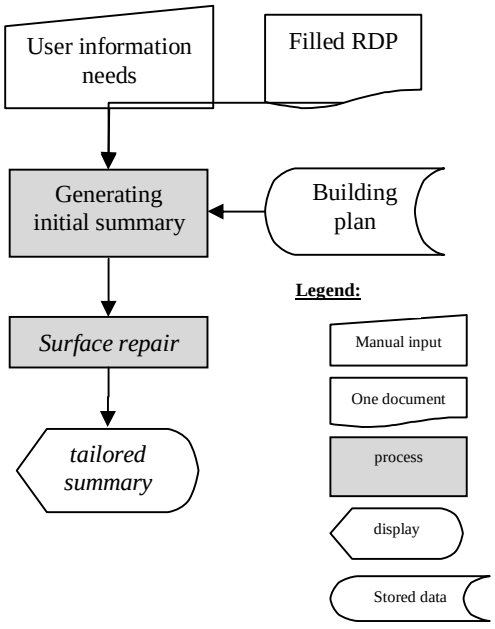

Figure 2. Second Stage: Transforming RDP into Tailored Summary.

Table 2. Building Plan for Short Summary.

	Length of summary = short	
Task \ User	Informed	Uninformed
General	2 Aim	1 Co_Gro + 1 Gap_Weak + 2 Aim
Contrastive	2 Aim + 1-2 Codi	1 Co_Gro + 1 Gap_Weak + 2 Aim +1-2 Codi
Ancestry	2 Aim + 1-2 Use	1 Co_Gro + 1 Gap_Weak + 2 Aim +1-2 Use

Table 3. Building plan for long summary

	Length of summary = long	
Task \ User	Informed	Uninformed
General	2-3 Aim + 1 Own_Mthd	1 Co_Gro + 1 Gap_Weak + 2-3 Aim +1 Own_Mthd
Contrastive	2-3 Aim +1-2 Codi +1 Own_Conc	1 Co_Gro + 1 Gap_Weak + 2-3 Aim +2 Codi +1 Own_Conc
Ancestry	2-3 Aim +1-2 Use +1 Fut	1 Co_Gro + 1 Gap_Weak + 2-3 Aim +1-2 Use +1 Fut

SUMMARIZING SCIENTIFIC PAPER BY PROCESSING RDP

If filled RDP has been generated as shown in Figure 1, the next step is processing RDP to produce tailored summary based on user information needs. This needs can be represented by three variables: task (general, ancestry, contrastive), background knowledge (informed, uninformed), and summary length (short, longer) [2]. Consequently, there are twelve user types based on user information needs.

For each user type, building plan is defined [2]. It contains summary composition based on RDP. For example, as shown in Table 2, summary composition for user type <general task, informed user, short summary> is 2 AIM sentences, but summary composition for user type <general task, uninformed user, short summary> is 2 CO_GRO sentences and 2 AIM sentences.

Since extractive summary is composed by important extracts that is loss of context, one of the problems is low coherence of summary. In this paper, improvement of summary coherence

is conducted by surface repair that makes summary is easier to be read.

Figure 2 shows processes of second stage that is transforming filled RDP into tailored summary. It consists of two processes: generating initial summary and surface repair.

GENERATING INITIAL SUMMARY ACCORDING TO BUILDING PLAN

Before generating initial summary, we have to define building plan first. Teufel [2] has defined building plan, but it was built for AZ with 7 rhetoric classes. That's why adaption of building plan for AZ-II with 15 rhetoric classes needs to be done.

Modification of building plan will be done by following AZ to AZ-II conversion scheme [21], for example background problem and background aim those are exist on AZ will be converted into CO_GRO. Besides the modification of existing building plan, there is also necessity to define user's needs combination that is not exist yet, it is summary with requirement long for task contrastive and ancestry. Result of this modification and improvement are shown in Table 2 and Table 3, it shows all combination of needed information from three defined variables.

Based on the user information needs, the matched building plan will be selected. If the required sentence in building plan is less than the number of sentences in the slot, sentence selection must be conducted. This research used Maximal Marginal Importance – Multi Sentence (MMI-MS) [22]. This method is a selection method that will choose some most important sentences and have least similarity among the other sentences. It was derived from Maximal Marginal Relevance [23] that is already used widely in generating summary.

SURFACE REPAIR

Surface repair is used to enhance the readability of initial summary and also to make sentences in the summary more related each other. This step will increase the coherency between sentences in the final summary.

This research used syntax tree based combination method [24]. After preparing some template phrases, sentence combining will be conducted based on the collected phrases [2]. Five groups of template phrases will be explained later.

Group I: "This paper's goal is to", "This paper's topic is to"

This phrase group is only able to be combined for sentence which type is AIM and the sentence is the first sentence in summary paragraph. This phrase is not able to be used for passive sentence, AIM sentence with "we/i" as subject, and sentence that already has word "to". Those restrictions are there in order to prevent changing the meaning and grammatical errors on the result of surface repair. For example, K2 is a surface repair from K1.

K1: `This paper describes recent work on developing an integrated heuristic scheme for selecting the parse that is deemed ``best`` from such a collection.`

K2: **`This paper 's goal is to`** `describes recent work on developing an integrated heuristic scheme for selecting the parse that is deemed `` best `` from such a collection.`

Group II: "This paper's specific goal is to", "Another goal is to", "The goal is to", "This approach"

This phrase groups is only able to be used when the combined sentence's type is aim and not the first sentence of the summary. Similar with the previous group, phrase can't be used for passive sentence, sentence that contains word "to", and sentence with subject "we/i". If K1 on the previous sentence's type is AIM and not the first sentence, then K3 is the surface repair for K1.

K3: **`This approach`** `describes recent work on developing an integrated heuristic scheme for selecting the parse that is deemed `` best `` from such a collection.`

Group III: "More specifically"

Template phrase "more specifically" will be used when the combined sentence is not first sentence and preceded with sentence which type is AIM too. Because this phrase is a conjuction between two AIM sentences, so the form of the sentence is free. In the following example, K5 combined with the template phrase and become clause of K4 as shown in K6.

K4-K5: `I describe a compiler and development environment for feature-augmented two-level morphology rules`

integrated into a full NLP system. The compiler is optimized for a class of languages including many or most European ones, and for rapid development and debugging of descriptions of new languages.

K6: **The authors** describe a compiler and development environment for feature-augmented two-level morphology rules integrated into a full NLP system, **more specifically** the compiler is optimized for a class of languages including many or most European ones , and for rapid development and debugging of descriptions of new languages.

Group IV: "The authors"

Template phrase "*the authors*" is used as word substitute in summary's sentences those refer to the writer, for instance we or i. For this phrase the combination method is different with the other groups, it is by replacing "i" or 'we" from the original sentence directly. In the previous example, K6 is result of surface repair for K4.

Group V: "It uses"

Template phrase *"it uses"* is used for sentence that has continuation characteristic, it is identified by the type of sentence that is "use". For the combination of this phrase, the method is a little different with the other phrase combination method. In this combination, the method will look for NP (noun phrase) from the sentence's POS (part of speech) which belongs to the first VP (verb phrase) of the sentence. The objective is to obtain sentence's object. After that POS is found, words those are appeared before VP and NP will be removed. And the last step is to add phrase "it uses" in the beginning of sentence.

K7: This is basically the entropy used in Quinlan, 1986.

K8 : **It uses** the entropy used in Quinlan,1986.

Example of Tailored Summary

Based on the building plan that is shown in Table 2 and Table 3, there will be 12 summary combinations that can be generated from a scientific paper. The sentence's order in the summary follows the sentence's order in the original source.

<?xml version="1.0" ?>
- <root><slot><slotName>**aim**</slotName>
- <sentences> <sentence>**In this paper we presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences.** </sentence>
- <sentence>**In this work we propose a method for estimating the probability of such previously unseen word combinations using available information on ``most similar`` words.** </sentence>
- <sentence>**We focus here on a particular kind of configuration, word cooccurrence.** </sentence>
</sentences>
</slot>
- <slot> <slotName>**co_gro**</slotName>
- <sentences><sentence>**For example, a speech recognizer may need to determine which of the two word combinations ``eat a peach`` and ``eat a beach`` is more likely.** </sentence>
- <sentence>**In many applications of natural language processing it is necessary to determine the likelihood of a given word combination.** </sentence>
- <sentence>**The MLE for the probability of a bigram (w1,w2) is simply: where is the frequency of (w1,w2) in the training corpus and N is the total number of bigrams.**
</sentence> </sentences> </slot>
- <slot> <slotName>**gap_weak**</slotName>
- <sentences><sentence>**However, the nature of language is such that many word combinations are infrequent and do not occur in a given corpus.** </sentence>
- <sentence>**Because of data sparseness, we cannot reliably use a maximum likelihood estimator (MLE) for bigram probabilities.** </sentence>
- <sentence>**However, this estimates the probability of any unseen bigram to be zero, which is clearly undesirable.**
</sentence> </sentences> </slot>
<slot><slotName>**own_mthd**</slotName>
- <sentences> <sentence>**Equation ( ) modifies slightly Katz`s presentation to include the placeholder Pr for alternative models of the distribution of unseen bigrams.**
</sentence>
- <sentence>**Arc scores in those lattices are sums of an acoustic score (negative log likelihood) and a language-model score, in this case the negative log probability provided by the baseline bigram model.**
</sentence>
- <sentence>**We describe a probabilistic word association model based on distributional word similarity, and apply it to improving probability estimates for unseen word bigrams in a variant of Katz`s back-off model.**
</sentence> </sentences> </slot>
- <slot> <slotName>**use**</slotName>
- <sentences> <sentence>**Following Pereira,Tishby,Lee,1993 , we measure word similarity by the relative entropy, or Kullback-Leibler (KL) distance, between the corresponding conditional distributions .**</sentence>
- <sentence>**We evaluated our method by comparing its perplexity and effect on speech-recognition accuracy with the baseline bigram back-off model developed by MIT Lincoln Laboratories for the Wall Street Journal (WSJ) text and dictation corpora provided by ARPA`s HLT program Paul,1991 .** </sentence>
- <sentence>**For perplexity evaluation, we tuned the similarity model parameters by minimizing perplexity on an additional sample of 57.5 thousand words of WSJ text, drawn from the ARPA HLT development test set.**
</sentence></sentences> </slot>

Figure 3. Example of Dilled RDP

In this work the authors propose a method for estimating the probability of such previously unseen word combinations using available information on `` most similar `` words , more specifically in this paper the authors presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences.

Figure 4. Example for Short Summary for User with General Task and Informed User Background.

In this work the authors propose a method for estimating the probability of such previously unseen word combinations using available information on `` most similar `` words , more specifically the authors focus here on a particular kind of configuration , word cooccurrence. Arc scores in those lattices are sums of an acoustic score (negative log likelihood) and a language-model score, in this case the negative log probability provided by the baseline bigram model. In this paper the authors presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences.

Figure 5. Example for Long Summary for User With General Task and Informed User Background.

In this work the authors propose a method for estimating the probability of such previously unseen word combinations using available information on `` most similar `` words . It uses their method by comparing its perplexity and effect on speech-recognition accuracy with the baseline bigram back-off model developed by MIT Lincoln Laboratories for the Wall Street Journal -LRB- WSJ -RRB- text and dictation corpora provided by ARPA ` s HLT program Paul ,1991 . It uses the similarity model parameters by minimizing perplexity on an additional sample of 57.5 thousand words of WSJ text , drawn from the ARPA HLT development test set . In this paper the authors presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences.

Figure 6. Example for short summary for ancestry user task and informed user background.

For example, Figure 4 until Figure 7 show 4 summaries, with various length parameter (Figure 4 and Figure 5), task parameter (Figure 4 and Figure 6), and also user's background knowledge (Figure 6 dan Figure 7). The summaries are composed by sentences that are obtained from Aim, Own_Mthd, and Use in XML form (Figure 3).

Figure 4 shows short summary with general user task and informed user background, besides that Figure 5 shows longer version of summary for the same user. Based on building plan, short summary will be consisted of 2 Aim sentences those are connected with template phrase "more specifically". Long summary will be consisted 2-3 Aim sentences and 1 OWN_MTHD Sentence. In Figure 5, third sentence's type is OWN_MTHD.

Difference between user tasks in Figure 4 and Figure 6 for short summary and informed user background is 1-2 additional sentences with use type. In Figure 6, the second and third sentence's type is USE. In Figure 4, there is a merger process between two AIM sentences with template phrase, but this does not happen in the summary in Figure 6 because the AIM sentences are the first and the fourth sentence.

However, the nature of language is such that many word combinations are infrequent and do not occur in a given corpus. In this work the authors propose a method for estimating the probability of such previously unseen word combinations using available information on `` most similar `` words .The MLE for the probability of a bigram (w1,w2) is simply: where is the frequency of (w1,w2) in the training corpus and N is the total number of bigrams. It uses their method by comparing its perplexity and effect on speech-recognition accuracy with the baseline bigram back-off model developed by MIT Lincoln Laboratories for the Wall Street Journal -LRB- WSJ -RRB- text and dictation corpora provided by ARPA ` s HLT program Paul ,1991 . It uses the similarity model parameters by minimizing perplexity on an additional sample of 57.5 thousand words of WSJ text , drawn from the ARPA HLT development test set . In this paper the authors presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences.

Figure 7. Example for Short Summary with Ancestry User Task and Uninformed User Background.

Instruction:
You'll be given some sentences those are already grouped. Please choose some number of sentences based on the instruction for each group by giving circle on the number besides the sentence. Choose sentence that is most representative or the most different compared to the other sentences.

Paper's Title: AN INTEGRATED HEURISTIC SCHEME FOR PARTIAL PARSE EVALUATION
Group of sentences that contain the purpose of written paper:

Sentences (choose 2)
1. This paper describes recent work on developing an integrated heuristic scheme for selecting the parse that is deemed ``best`` from such a collection.
2. Preliminary results from experiments conducted on parsing speech recognized spontaneous speech are also reported.
3. First, we wanted to compare the parsing capability of the GLR* parser with that of the original GLR parser.

Figure 8. Example of Grouped Sentences in The Feedback form for Evaluating Sentence Selection Process.

Petunjuk :
Akan diberikan 2 buah kalimat untuk setiap poin. Kalimat tersebut merupakan kalimat asli dari makalah dan kalimat setelah dilakukan *surface repair*. Anda diminta untuk menentukan apakah *surface repair* untuk poin yang bersangkutan dapat diterima atau tidak.
Kalimat asli: Following Pereira,Tishby, Lee,1993 , we measure word similarity by the relative entropy, or Kullback-Leibler (KL) distance, between the corresponding conditional distributions .
Kalimat *surface repair* : It uses word similarity by the relative entropy , or Kullback-Leibler -LRB- KL -RRB- distance , between the corresponding conditional distributions .
Terima : (ya/tidak)

Figure 9. Example of Paired Sentence in Feedback form for Evaluating Result of Surface Repair.

Figure 7 shows summary for user with informed background that is different with Figure 6. Building plan for this summary is 1 CO_GRO sentence, 1 GAP_WEAK sentence, 2 AIM sentences, and 1-2 USE sentence(s). Sentence's rhetorical category in summary as shown in Figure 7 are GAP_WEAK, AIM, CO_GRO, USE, and AIM.

RESULT AND DISCUSSION

Evaluation was conducted in each subsystem. There are evaluation of initial summary step and evaluation of surface repair step. There were five respondents which were fourth year students of Informatics Engineering. Each respondent was asked to fill feedback form related to sentence selection and surface repair.

Evaluation in Sentence Selection

In order to evaluate sentence selection, it is required to have a standard sentence selection that is done by human. This standard will be compared to the result of sentence selection that is done by the system.

In sentence selection process for generating initial summary, system chooses some sentences in certain slot. Number of selected sentences depends on building plan.

All respondents were given feedback form that contains 18 groups of sentences which total was 51 sentences from one paper. Figure 8 shows example of sentence group within the feedback form. Sentences in feedback form were grouped based on its slot in RDP. The result was 18 sentences from human respondent.

Measurement of the evaluation uses precision and recall, Equation (1) and (2) is counted based on the number of sentences those are choosen by both respondent and system. A sentence will be included to the evaluation measurement if that sentence is choosen by at least 3 respondents from total 5 respondents. If a sentence is choosen by both system and respondent, then that sentence will be counted as true positive (TP). If a sentence is choosen by system but not choosen by respondent, then that sentence will be counted as false positive (FP). And the other way, sentence will be counted as false negative (FN).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The weighting process for counting MMI-ms uses three type combinations. The first combination is idf, tf, normalization, and IGR. The second combination is tf, and idf. And the third combination is tf, idf, and normalization.

For the first and third weighting combination, it is given value 0.56 for recall and precision of system. And for second weighting combination, the value of system's recall and precision is 0.61 which is also the best value among other combinations. Precision and recall have the same value because the value of FP and FN are same. Complete result of selection is shown in Table 4.

Table 4. Result of sentence selection compared to sentence selection standard that is done by human respondent.

Combination	TP	FP	FN	P	R
Tf, idf, normalization, IGR	10	8	8	0.56	0.56
Tf, idf	11	7	7	0.61	0.61
Tf, idf, normalization	10	8	8	0.56	0.56

Evaluation in Surface Repair

Just like the evaluation for sentence selection, evaluation for surface repair was done based on standard that was the result of feedback form to human respondent. In that feedback gorm, there are some sentences that were taken from the system's summary. For one RDP, there will be some variation of summary based on parameters for customizable user's needs. Basically, this evaluation is done to evaluate the surface repair in the final summary. Therefore, evaluation is not done on each template in surface repair, but it is done on each sentence that is being surface repaired in the final summary. Respondent answer for each sentence is a yes or no , yes if respondent accepts and no if respondent doesn't accept. Figure 9 shows example of paired sentence in feedback form.

From 15 pair of sentences those are given to respondents, 14 pairs are accepted so the acceptance value is 0.91. Sentence is assumed as accepted if that sentence is accepted by at least 3 out of 5 respondents. Acceptance value is ratio between accepted sentences and total sentences. Sentence can't be accepted by respondent basically caused by the changing of meaning, or resulting grammatical error but does not give meaning change.

CONCLUSION

This paper has discussed the phase to transform RDP to tailored summary which is the second step of summary generation after generating RDP. This research is aim to implement tailored summary that has already proposed by Teufel [2].

There are two main processes, generating initial summary and surface repair. Based on the information needs, initial summary is created by selecting sentence from RDP based on the slot listed in building plan. Surface repair combines sentence in initial summary with template phrase based on syntax tree combination method.

Further works will integrate each subsystem that is already developed into one integrated system for summarizing scientific paper. Besides that, this research is the initial research towards research in summarizing multiple scientific paper.

REFERENCES

[1] Mohammad S, Dorr B, Egan M, Hassan A, Muthukrishan P, Qazvinian V, Radev D, and Zajic D. Using Citations to Generate Surveys of Scientific Paradigms. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*. 584–592. 2009.

[2] Teufel S. Argumentative zoning: Information Extraction from Scientific Text. PhD Dissertation. University of Edinburgh. 1999.

[3] Khodra ML, Widyantoro DH, Aziz EA, and Trilaksono BR. Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences. *ITB Journal of Information and Communication Technology*. 5:17-34. 2011.

[4] Khodra ML, Widyantoro DH, Aziz EA, and Trilaksono BR. Information Extraction for Scientific Paper Using Rhetorical Classifier. *Proceedings of International Conference on Electrical Engineering and Informatics*. 2011.

[5] Luhn HP. The automatic creation of literature abstracts. *IBM Journal of Research and Development*. 2:159-165. 1958.

[6] Teufel S and Moens M. Summarizing Scientific Articles – Experiments with Relevance and Rhetorical Status.

Computational Linguistics. 28:409-445. 2002.

[7] Conroy JM, Schlesinger JD, and O'Leary DP. Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. *Proceedings of the COLING/ACL*. 2006.

[8] Jiaming Z. Exploiting Textual Structures of Technical Papers for Automatic Multi-Document Summarization. PhD Thesis. Department of Mechanical Engineering, National University of Singapore. 2008.

[9] Mani I, Bloedorn E. Machine Learning of Generic and User-Focused Summarization. *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. 1998.

[10] Hahn U, Mani I. The Challenges of Automatic Summarization. *IEEE Computer*. 33:29-36. 2000.

[11] Mani I. Recent Developments in Text Summarization. *Proceedings of CIKM'01*. 2001.

[12] Paris C, Wan S. Capturing the user's reading context for tailoring summaries. *Proceedings of International Conference on User Modelling, Adaptation and Presentation (UMAP)*. 2009.

[13] Berkovsky S, Baldwin T, and Zukerman I. Aspect-Based Personalized Text Summarization. *Proceedings of 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. 2008.

[14] Agnihotri L, Kender JR, Dimitrova N, and Zimmerman J. User Study for Generating Personalized Summary Profiles. *IEEE International Conference on Multimedia and Expo*. 2005.

[15] Radev DR., Fan W., Zhang Z. WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System. *NAACL Workshop on Automatic Summarization*. Pittsburgh. 2001.

[16] Filho PPB, Pardo TAS. Summarizing Scientific Texts: Experiments with Extractive Summarizers. *Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications, IEEE Computer Society*. 2007.

[17] Kupiec J, Pedersen J, and Chen E. A Trainable Document Summarizer. *Proceedings of 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 1995.

[18] Qazvinian V and Radev DR. Scientific Paper Summarization Using Citation Summary Networks. COLING 2008. Manchester. 2008.

[19] Agarwal N, Gvr K, Reddy RS, and Rose CP. Towards Multi-Document Summarization of Scientific Articles: Making Interesting Comparisons with SciSumm. *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Language*. 2011.

[20] Khodra ML, Widyantoro DH, Aziz EA, and Trilaksono BR. Automatic Tailored Multi-Paper Summarization. *Proceedings of The Asia Oceania Top University League on Engineering*. 2011.

[21] Teufel S, Siddhartan A, and Batchelor C. Towards Discipline-Independent Argumentative zoning Evidence from Chemistry and Computational linguistics. *Proceedings Of the 2009 Conference on Empirical Methods in Natural Language Processing*. 2009.

[22] Mori T and Sasaki T. Information Gain Ratio meets Maximal Marginal Relevance. *Proceedings of the Third NTCIR Workshop*. 2003.

[23] Carbonell J and Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998.

[24] Pang B, Knight K, and Marcu D. Syntax based Alignment of Multiple Translations: Extraction Paraphrases and Generating New Sentences. *Proceedings of HLT-NAACL*. 2003.