Multi-Document Summarization of Scientific Corpora

Ozge Yeloglu
Faculty of Computer Science
Dalhousie University
Halifax, Canada
yeloglu@cs.dal.ca

Evangelos Milios
Faculty of Computer Science
Dalhousie University
Halifax, Canada
eem@cs.dal.ca

Nur Zincir-Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada
zincir@cs.dal.ca

ABSTRACT

In this paper, we investigated four approaches for scientific corpora summarization when only gold-standard keyterms available. MEAD with built-in default vocabulary, MEAD with corpus specific vocabulary extracted by Keyphrase Extraction Algorithm (KEA), LexRank (a state-of-the-art summarization algorithm based on random walk) and W3SS (summarization algorithm based on keyword density) are tested on two Computer Science research paper collections. We use a content evaluation method, pyramid method, instead of the well-known ROUGE metrics since there are no gold-standard summaries available for our data. Evaluations with pyramid method indicates that including a corpus specific vocabulary to the traditional summarization methods improves the performance but not significantly. On the other hand, visual inspection shows us that current content evaluation methods, which use only the gold-standard keyterm information, are not intuitive and focus must turn into better evaluation techniques especially for the multi-document summarization problem. Even though the pyramid method looks for important keyterms in the resulting summaries, it cannot distinguish between a general introductory sentence about the area and a specific sentence on the core idea, if they both contain the same keyterm. Also, our results show that the state of the art summarization method LexRank is not feasible for scientific corpus summarization because of its high computational cost.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*abstracting methods*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

Keywords

Multi-Document Summarization, Scientific Corpora, Keyterm Extraction, Content Evaluation

1. INTRODUCTION

The increasing availability of digital information raises the need of automatic text summarization. Specifically, the study of multi-document summarization has drawn attention in recent years. Multi-document summaries can be used to describe the information contained in a corpus and help the users in getting an overview of the corpus.

Several summarization systems have been developed for this purpose; a typical example is MEAD [18]. There are other systems both freely and commercially available. However, systems are typically evaluated with short documents, such as newspaper articles. The main reason behind this is the lack of publicly available corpora accompanied by ideal summaries, for document collections other than news. The most commonly used datasets are from the Document Understanding Conferences (DUC)[1], which consist of online newspaper articles. ROUGE metrics [11] are used on the DUC datasets for evaluation where the gold standard summaries already exist. However, if one wants to use a different dataset with no ideal summaries available, then either user studies or other methods for content analysis must be applied in the evaluation step.

In our work, we are interested in scientific multi-document summarization. To investigate the current multi-document summarization methods on scientific topic summarization, we performed experiments using MEAD [18] and W3SS [22] frameworks. We employed MEAD original, MEAD with corpus specific vocabulary, LexRank and W3SS methods on an ACM computer science research paper corpus consisting of manually downloaded 238 scientific papers in 6 categories and a Computer Science Publications dataset (DUCSP) consisting of 584 research publications in 23 groups. We automatically extracted the keyterms from the corpus using a Keyphrase Extraction Algorithm (KEA) [21]. Then, we included this vocabulary to the original MEAD framework.

Our results show that including a corpus specific vocabulary to the MEAD summarization process improves the performance of the centroid method but not significantly. Also, the-state-of-the-art summarization method LexRank is proved to be impracticable for multi-document summarization of the full text of scientific documents due to its computational requirements.

[1]http://www-nlpir.nist.gov/projects/duc/intro.html

2. RELATED WORK

Approaches for text summarization typically branch out in several dimensions. We can roughly distinguish: *single-document* vs. *multi-document* depending on the source text, *extractive* vs. *abstractive* depending on the output and *news* vs. *technical* depending on the text genre.

2.1 Scientific Document Summarization

The interest in automatically shortening the texts started when the scientific papers and books were to be digitally stored. Earliest instances of research on scientific (single-)document summarization proposed using term frequency and distribution to compute the significance of the words and the sentences [12] and cue words, title and headings, and sentence location [5].

However, the interest of the summarization research community switched from scientific papers to news articles until the former became active in nineties. A corpus-based approach was introduced in [10] with a trainable summarizer where a set of documents and corresponding manually created abstracts are used as training set. A simple Bayesian classification algorithm decides if a sentence should be included in the summary based on the sentence length and location, cue phrases and acronyms in the sentence. Thereafter, a strategy which concentrates on the rhetorical status of statements in the scientific article was proposed in [20] to summarize single Computational Linguistic articles. However, this approach is heavily dependent on a manually annotated training set which creates an important drawback. Latest work in this area is conducted in [3] with a strategy for summarizing single Organic Chemistry documents. This technique combines domain-specific document pre-processing with a sentence scoring method relying on the statistical properties of documents.

2.2 Multi-Document Summarization

Multi-document summarization became active in the mid-1990s with the focus being mostly on news articles. The NLP group at Columbia University has pioneered the field with SUMMONS [13] system, being the first to suggest the potential of combining information extraction with natural language generation in a summarization system. SUMMONS is an abstractive system (puts strong emphasis on the form, aiming to produce a grammatical summary) and designed to work on the news domain, specifically news articles about terrorism. It heavily relies on advanced natural language generation techniques as any abstractive system.

As an alternative, a number of rather impressive extractive summarization systems, specifically on news focus, have emerged such as the SUMMARIST [9], the NewsBlaster [7] and the MEAD [18]. Extractive systems analyze the source texts using information retrieval techniques (e.g. keyword identification, frequency analysis) to determine and extract the most significant sentences. Approaches vary on how the sentence similarities are used for extraction: [14] identifies common topics through clustering and then selects one sentence to represent each cluster while [2] generates a composite sentence from each cluster. Another approach is proposed in [19] where the single documents are summarized first and summaries are grouped in clusters, then representative passages from the clusters are selected. Even though there are many approaches to multi-document summarization, the number of systems available is not large. Among the publicly available systems, MEAD [18] is the only one intended to be domain-independent.

In our work, we are interested in scientific multi-document (topic) summarization. There is a limited amount of work in this research area [10, 20, 3, 16]. However, all these methods are developed for single-document summarization except [16]. Even though [16] has promising results, intensive manual work is required to create the training set for the discourse parsing classifier. Moreover, this system is heavily dependent on Natural Language Processing and tailored for a specific dataset only. Therefore the problem of summarizing collections of scientific articles is still open.

3. SUMMARIZATION METHODS

We used four summarization methods for our experiments: MEAD, MEAD with corpus specific vocabulary, LexRank and W3SS.

3.1 MEAD

MEAD [18] is an open-source toolkit for summarization. MEAD has been typically defined as "centroid-based" summarizer. The "Centroid", in this context, is a pseudo-document which consists of words defining the topic of the group of documents. MEAD also has a built-in general purpose English vocabulary with corresponding IDF values of every word in the vocabulary. These IDF values are used during the centroid feature. MEAD algorithm can be summarized as:

- Create the topic representative pseudo-document of the group (centroid)
- Calculate three different features for each sentence: Centroid, Position and Length
- Sort sentences according to the linear combination of three features with default weights (centroid: 1, position: 1) and the default threshold for length: 9.
- Check redundancy on candidate sentences starting from the second sentence on the ranked list

3.2 MEAD with Vocabulary

MEAD has a general purpose English vocabulary built-in for its "centroid" calculation. We conjecture that an addition of a corpus specific vocabulary may have a positive impact on the resulting summaries in the case of scientific summarization. Therefore, we created a corpus specific vocabulary (keyword–IDF pairs).

Keywords are extracted with the Keyphrase Extraction Algorithm [21] (KEA) and the corresponding IDF values are calculated from the corpus. Finally, these keyword–IDF values are incorporated into the original MEAD framework. MEAD's built-in general purpose English vocabulary is replaced by the new corpus specific vocabulary. MEAD now uses this vocabulary when creating its topic representative pseudo-document (centroid).

3.3 LexRank

LexRank [6] is a state-of-the-art multi-document summarization system which works based on a random walk on the cosine similarity of sentences . LexRank first builds a graph of all candidate sentences where nodes are the sentences and the edges are the cosine similarity values. Two candidate sentences are connected with an edge if the similarity between them is above a threshold. The system finds the most central sentences of the graph by performing a random walk on it.

Sentences vote for each other just by virtue of being adjacent to each other. This is similar to the concept of prestige in social networks where is possible to find the most prestigious, or popular member of a network by analyzing the relationships among network members. As a result, the sentences with the highest scores are considered to contain the gist of the document and form the summary.

3.4 W3SS

In W3SS [22], the cluster summarization relies on the extraction of the most significant sentences from the target cluster based on the density of a list of key phrases that best describe the entire cluster. This method was originally designed for web page summarization. However the fact that it uses the list of keyphrases for the summarization purpose makes this method attractive for our research.

The first component of W3SS is the narrative text classifier. Web pages often contain bullets or short sentences, instead of a narrative structure. This also applies to the scientific documents. There are often figures, tables and bullets in scientific documents which would not be meaningful in a summary. First, a classifier is trained for determining if a paragraph is long enough to be considered in narrative paragraph classification. Then, a second classifier is trained to classify long paragraphs into narrative or non-narrative. C5.0 [1] decision tree classifiers were used for both classification tasks. However, these classifiers were trained for Web documents where the training set was manually created. We used the same decision trees in our work.

The second component of W3SS is the keyphrase extractor where a keyphrase can be either keyword or keyterm. Top N keyphrases from the narrative paragraphs are extracted using the CNC method [8] which applies both linguistic (part-of-speech tagging [4] and linguistic filter) and statistical analysis (frequency analysis, C-value, NC-value).

The last component of W3SS is the key sentence extractor where the top N most significant sentences are retrieved from all narrative paragraphs based on the presence density of keyphrases.

3.5 Data

Two datasets are used for our experiments: The ACM Dataset and the DUCSP, a dataset consisting of the publications of the faculty members of a Computer Science unit.

The ***ACM Dataset*** is a Computer Science publications corpus manually downloaded from Association for Computing Machinery (ACM) digital library[2]. There are 238 papers with full text including ACM category terms and author assigned keyterms. These papers are grouped by ACM under 6 subcategories where the number of documents per subcategory is between 28 and 57. A summary of the corpus can be seen in Table 1.

In the ***DUCSP Dataset*** there are 584 papers in 23 groups with full text including category terms and author assigned keyterms. Each group belongs to a specific author. However, one publication might be under more than one author if the document has co-authors. A summary of the corpus can be seen in Table 2.

[2]http://portal.acm.org/dl.cfm

Table 1: All the subcategories in ACM collection with the number of documents and the number of ACM and author assigned keyterms (gold-standard keyterms) in them.

Subcategory Name	**Number of Documents**	**Number of Gold-standard Keyterms**
H.2.7	35	57
H.2.8	28	46
H.3.1	54	103
H.5.4	52	70
I.2.6	35	52
I.2.7	57	88

Table 2: Groups of DUCSP collection with the number of documents and the number of author assigned keyterms (gold-standard keyterms).

Group Name	**Number of Documents**	**Number of Gold-standard Keyterms**
Author1	69	71
Author2	52	64
Author3	50	107
Author4	47	77
Author5	47	44
Author6	45	49
Author7	34	46
Author8	28	13
Author9	23	32
Author10	22	31
Author11	20	25
Author12	16	3
Author13	15	15
Author14	15	29
Author15	15	27
Author16	15	7
Author17	14	14
Author18	11	20
Author19	11	16
Author20	10	10
Author21	10	14
Author22	9	22
Author23	6	18

4. EXPERIMENTAL SETUP

We seperated abstract and body (text between abstract and references) of every document in both datasets and created 4 different datasets as a result. Then, we automatically generated summaries of 25 sentences for every group in every dataset using 4 summarization methods to evaluate with pyramid method.

4.1 Evaluation

Our goal was to determine if an addition of the corpus specific vocabulary to the summarization process makes an improvement over the traditional summarization methods. To do this, we evaluated each of the automatically generated summaries using the Pyramid Method [15].

To be able to use the pyramid method on our resulting summaries, we need gold-standard dictionaries. In both datasets, there are ACM and author assigned keyterms for each group. These keyterms are extracted from every document in the groups and the frequencies of the keyterms in the summary sentences are calculated for pyramid method.

4.1.1 Pyramid Method

We employ the pyramid evaluation method [15] at the sentence level to evaluate the automatically generated summaries. To our knowledge, the pyramid method at the sentence level is first used in [17] to evaluate the summaries of citation sentences. The analysis of summary content in pyramid method is based on Summarization Content Units (SCUs). SCUs emerge from annotation of a corpus of summaries and are not bigger than a clause [15]. We benefit from the list of ACM and author assigned keyterms, which are equivalent to the *SCUs* in [15], as the ground truth of summarization evaluation.

The pyramid score is calculated as follows: Assume a pyramid that has n tiers. Tier T_i contains keyterms appearing in i sentences of the input documents, thus has weight i. If a keyterm appears in more sentences, then it falls in a higher tier. Let $|T_i|$ denote the number of keyterms in tier T_i, and D_i is the size of a subset of T_i whose members are in the automatically generated summary. Then the total SCU weight D is:

$$D = \sum_{i=j+1}^{n} i \times D_i \qquad (1)$$

The maximum possible pyramid score for a summary with X keyterms is,

$$Max = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|) \qquad (2)$$

where $j = max_i(\sum_{t=i}^{n} |T_t| \leq X)$.

Finally, the pyramid score P is the ratio of D to Max,

$$P = \frac{D}{Max} \qquad (3)$$

which ranges from 0 to 1. The higher score shows that the summary contains more keyterms or more heavily weighted (more topic focused) keyterms.

4.2 Results and Discussion

Figure 1 shows the pyramid scores of every summarization method on ACM dataset when abstracts only (Figure 1(a)) and bodies only (Figure 1(b)) are used as input text. It is seen that there is no clear winner among the summarization methods.

Figure 2 shows the pyramid scores of every summarization method on DUCSP dataset when abstracts only (Figure 2(a)) and bodies only (Figure 2(b)) are used as input text. These scores indicate that including a corpus specific vocabulary to the MEAD summarization process improves the performance of the centroid method but not statistically significantly, based on two-tail t-tests. On the other hand, it is clear on Figure 1(b) and Figure 2(b) that W3SS's performance gets competitive with the other methods when the bodies of the documents are used as input text. This result is expected since W3SS works better with longer texts because of its C/NC keyterm extraction algorithm.

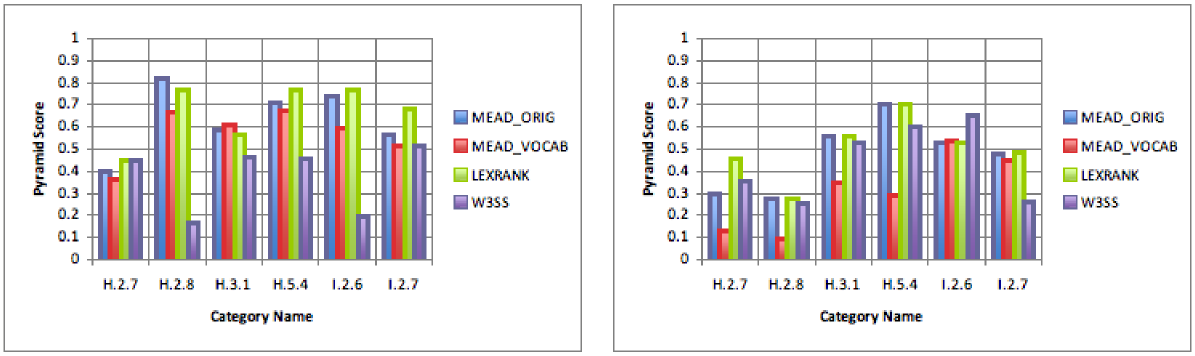
Running times of algorithms are seen in Table 3 and Table 4 for abstract and body input texts, respectively. LexRank's computational cost is very high compared to other methods but its pyramid scores are not higher than MEAD or W3SS. It is also clear that W3SS is the fastest summarization method when the keyterm extraction part of the algorithm is excluded. Therefore, it is clear that Lexrank is not practical for multidocument summarization of scientific papers.

We see the advantage of including corpus specific vocabulary to the summarization process with MEAD with vocabulary and W3SS methods especially in text body summarization. Abstracts are already focused and concentrated documents. Therefore, the resulting summaries will be focused as well. However it is harder to pick informative and focused sentences from the body which includes many details of the paper with tables, figures and formulas. Assuming abstract of a scientific document is not always easy to extract, when not directly available in the document metadata, we can conclude that vocabulary brings a potential advantage to the summarization.

Even though these results from pyramid score evaluation show that there is no significant difference between the methods not using corpus specific vocabulary and methods using it, visual inspection shows us different. As seen in Table 5, MEAD Original and LexRank tend to extract long sentences from the beginning of the documents, which are the introduction sentences. Also, it is surprising that Lexrank, a sophisticated and computationally expensive method, extracts almost the same sentences with the baseline MEAD Original method. However, MEAD with vocabulary and W3SS extract more specific and detailed sentences about the topic of the documents. To validate these intuitive observations, a user study is required, a common requirement for evaluating summarization methods when gold standard summaries are not available.

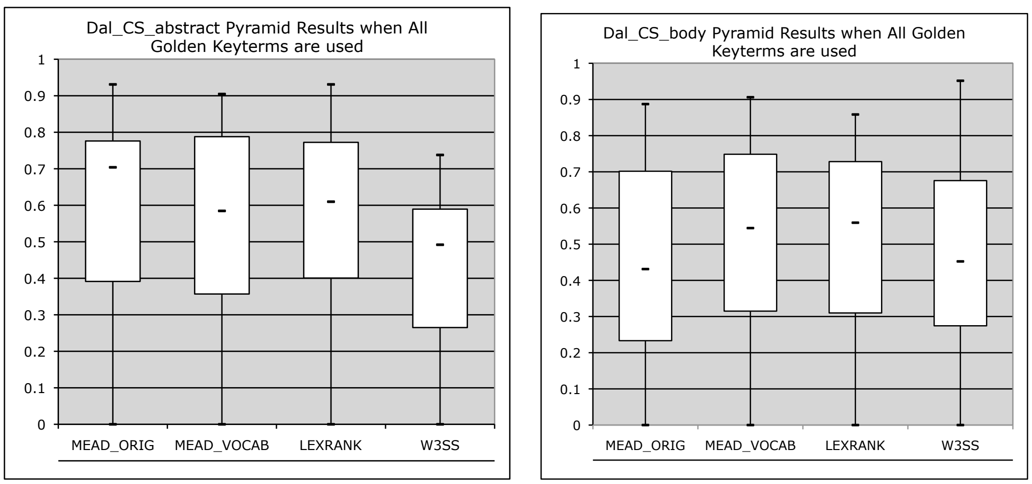
5. CONCLUSION AND FUTURE WORK

We have investigated four summarization methods, MEAD with default vocabulary, MEAD with corpus specific vocabulary extracted by KEA, LexRank, and W3SS with a CNC extracted vocabulary and tested them on Computer Science publications. Evaluations of the resulting summaries are performed with pyramid method. Results show that including a corpus specific vocabulary to the MEAD summarization process improves the performance of traditional summarization methods but not significantly. Also, MEAD with vocabulary and W3SS perform better according to the



(a) (b)

Figure 1: Pyramid scores of ACM Dataset



(a) (b)

Figure 2: Pyramid scores of DUCSP Dataset

Table 3: Running times of 4 algorithms when abstracts are used as input

Number of Documents	**KEA**	**MEAD**	**LexRank**	**CNC**	**W3SS**
6 - 15	4SECs - 7SECs	2SECs - 5SECs	10MINs - 3HOURs	1.5SECs - 5.5SECs	0.5SEC - 0.5SEC
16 - 28	7SECs - 7SECs	5SECs - 7SECs	4HOURs - 15HOURs	8.5 SECs - 8.5SECs	0.5SEC - 0.5SEC
34 - 45	8SECs - 9SECs	7SECs - 10SECs	16HOURs - 1DAY	13SECs - 17SECs	1SEC - 1SEC
47 - 69	9SECs - 10SECs	10SECs - 20SECs	1DAY - 5DAYs	18SECs - 21SECs	1SEC - 1SEC

Table 4: Running times of 4 algorithms when bodies are used as input

Number of Documents	**KEA**	**MEAD**	**LexRank**	**CNC**	**W3SS**
6 - 15	7SECs - 7SECs	2SECs - 5SECs	10MINs - 3HOURs	31SECs - 3MINs	1SEC - 1SEC
16 - 28	7SECs - 8SECs	5SECs - 7SECs	4HOURs - 15HOURs	3MINs - 4MINs	1SEC - 1SEC
34 - 45	8SECs - 9SECs	7SECs - 10SECs	16HOURs - 1DAY	5MINs - 6MINs	1SEC - 1SEC
47 - 69	9SECs - 11SECs	10SECs - 20SECs	1DAY - 5DAYs	7MINs - 13MINs	1SEC - 1SEC

Table 5: Summary examples from 4 algorithms

MEAD Original	[1] Clinical decision-making involves an active interplay between various medical knowledge modalities– the spectrum of medical knowledge modalities spanning from tacit knowledge to experiential knowledge to explicit knowledge to data-induced knowledge 1-4. [2] Our proposed knowledge morphing framework, attempts to alleviate the above knowledge retrieval problems by assisting practitioners seeking case-specific knowledge to formulate a single semantically rich knowledge query that is applied to multiple knowledge resources. [3] Evidence-based healthcare is a prevalent practice amongst both medical practitioners and management as it provides a sound basis for quality and consistent healthcare delivery
LexRank	[1] Clinical decision-making involves an active interplay between various medical knowledge modalities– the spectrum of medical knowledge modalities spanning from tacit knowledge to experiential knowledge to explicit knowledge to data-induced knowledge 1-4 . [2] Evidence-based healthcare is a prevalent practice amongst both medical practitioners and management as it provides a sound basis for quality and consistent healthcare delivery. [3] HEALTH CARE is facing exceptional challenges to keep pace with demands for "actionable" healthcare knowledge in the face of new treatments, procedures, guidelines, and delivery practices vis- -vis more stringent service-quality and outcome-measurement criterion 1 , 2 .
MEAD with Vocabulary	[1] The volume of information over the Internet is increasing at a tremendous rate, and as a consequence the search for 'relevant' and 'useful' information is becoming proportionally difficult. [2] The training set is used to identify the N N 10 recommended items while the test set is used to measure the quality of the recommendation in terms of the F1-metric F1 and the appropriateness degree AD . [3] Next, in the second stage, the compositional adaptation method takes into account the degree of relevance of the retrieved information items and the weighted frequency of the recurring constituent information components in order to select the most appropriate information components.
W3SS	[1] Such a provocation is to be achieved by repetitively presenting domain experts 'hypothetical' Scenarios [Che00] pertaining to novel or atypical problems and then observe and analyse the domain expert's tacit knowledgebased problem-solving methodology and procedures. [2] Tacit Knowledge Acquisition Process: The systematic acquisition of tacit health-care knowledge from domain experts takes place in three stages. [3] In view that the objective of this project is the ubiquitous acquisition and sharing of tacit knowledge from domain experts situated at different locations, there is a need to increase interoperability between the client-side and server-based applications through the Internet.

pyramid method when the full bodies of the documents are used as input than the abstracts only used as input. This gives an advantage to the user when there are no abstracts to use for summarization. Finally, we demonstrate that the state of the art summarization method, LexRank, does not appear suitable for multi-document summarization of scientific corpora since it is computationally expensive. MEAD and W3SS generate the summaries of a scientific corpus in minutes whereas it takes hours (even days for corpora with over 50 documents) for LexRank.

We also conclude that the evaluation of summarization methods continues to be a difficult problem. Visual inspection shows us that the addition of the vocabulary to the summarization process actually creates more topic oriented summaries. The results of the pyramid method do not appear to agree with visual inspection, which suggests that the current content evaluation methods are not intuitive. The ROUGE metrics can only be used when there is a manually created summary, which is very expensive to create for multi-document summarization of scientific text. Therefore, one must go through an extensive amount of work with user studies for multi-document summarization evaluation.

Visual inspections also suggest that sentences by themselves may not be the best way to summarize scientific document corpora since the extracted keysentences might be about a very specific topic, e.g. "*...[10] We subsequently examined different ways to embed the resulting translation models in a cross-language information retrieval system. [11] They use the resulting parallel corpora to induce a probabilistic translation dictionary that is then embedded into a cross-language information retrieval system. ...*". In contrast to news articles which have facts in a focused short document that would make it easier to understand the extracted sentences, scientific documents are discourse structured long documents with a flow which makes it harder to understand the concept based on a few extracted sentences. It is also shown in [16] that the most preferred summary style is a concept based summary where the keyterms (which define the concepts) are presented together with the full sentences. Therefore, we conjecture that including the keyterms/keyphrases into the summaries would result in more intuitive summaries.

6. ACKNOWLEDGMENTS

This work was supported by Natural Sciences and Engineering Research Council of Canada, and the MITACS NCE.

7. REFERENCES

[1] R. R. 2002. C5.0: An informal tutorial. http://www.rulequest.com/see5-unix.html, last accessed on April 1, 2010.

[2] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[3] F. Boudin, J. M. T. Moreno, and P. Velázquez-Morales. An efficient statistical approach for automatic organic chemistry summarization. In *GoTAL*, pages 89–99, 2008.

[4] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[5] H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, 1969.

[6] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, 2004.

[7] D. K. Evans, J. L. Klavans, and K. R. McKeown. Columbia newsblaster: multilingual news summarization on the web. In *HLT-NAACL '04: Demonstration Papers at HLT-NAACL 2004 on XX*, pages 1–4, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[8] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, August 2000.

[9] E. Hovy and C.-Y. Lin. Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 197–214, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[10] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73, New York, NY, USA, 1995. ACM.

[11] C.-Y. Lin. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25 - 26 2004.

[12] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, 1958.

[13] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, New York, NY, USA, 1995. ACM.

[14] K. R. McKeown, J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards multidocument summarization by reformulation: progress and prospects. In *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, pages 453–460, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.

[15] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[16] S. Ou, C. S.-G. Khoo, and D. H. Goh. Design and development of a concept-based multi-document summarization system for research abstracts. *J. Inf. Sci.*, 34(3):308–326, 2008.

[17] V. Qazvinian, D. R. Radev, and A. Özgür. Citation summarization through keyphrase extraction. In *The 23rd International Conference on Computational Linguistics, COLING 2010*, August 23 0 August 27 2010.

[18] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May 2004.

[19] G. C. Stein, A. Bagga, and G. B. Wise. Multi-document summarization: Methodologies and evaluations. In *In Proceedings of the 7th Conference on Automatic Natural Language Processing (TALN Õ00*, pages 337–346, 2000.

[20] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445, 2002.

[21] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: practical automatic keyphrase extraction. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255, New York, NY, USA, 1999. ACM.

[22] Y. Zhang, N. Zincir-Heywood, and E. Milios. World wide web site summarization. *Web Intelli. and Agent Sys.*, 2(1):39–53, 2004.