# SciSumm: A Multi-Document Summarization System for Scientific Articles

**Nitin Agarwal**
Language Technologies Institute
Carnegie Mellon University
nitina@cs.cmu.edu

**Kiran Gvr**
Language Technologies Resource Center
IIIT-Hyderabad, India
kiran_gvr@students.iiit.ac.in

**Ravi Shankar Reddy**
Language Technologies Resource Center
IIIT-Hyderabad, India
krs_reddy@students.iiit.ac.in

**Carolyn Penstein Rosé**
Language Technologies Institute
Carnegie Mellon University
cprose@cs.cmu.edu

## Abstract

In this demo, we present SciSumm, an interactive multi-document summarization system for scientific articles. The document collection to be summarized is a list of papers cited together within the same source article, otherwise known as a co-citation. At the heart of the approach is a topic based clustering of fragments extracted from each article based on queries generated from the context surrounding the co-cited list of papers. This analysis enables the generation of an overview of common themes from the co-cited papers that relate to the context in which the co-citation was found. SciSumm is currently built over the 2008 ACL Anthology, however the generalizable nature of the summarization techniques and the extensible architecture makes it possible to use the system with other corpora where a citation network is available. Evaluation results on the same corpus demonstrate that our system performs better than an existing widely used multi-document summarization system (MEAD).

## 1 Introduction

We present an interactive multi-document summarization system called SciSumm that summarizes document collections that are composed of lists of papers cited together within the same source article, otherwise known as a co-citation. The interactive nature of the summarization approach makes this demo session ideal for its presentation.

When users interact with SciSumm, they request summaries in context as they read, and that context determines the focus of the summary generated for a set of related scientific articles. This behaviour is different from some other non-interactive summarization systems that might appear as a black box and might not tailor the result to the specific information needs of the users in context. SciSumm captures a user's contextual needs when a user clicks on a co-citation. Using the context of the co-citation in the source article, we generate a query that allows us to create a summary in a query-oriented fashion. The extracted portions of the co-cited articles are then assembled into clusters that represent the main themes of the articles that relate to the context in which they were cited. Our evaluation demonstrates that SciSumm achieves higher quality summaries than a state-of-the-art multidocument summarization system (Radev, 2004).

The rest of the paper is organized as follows. We first describe the design goals for SciSumm in 2 to motivate the need for the system and its usefulness. The end-to-end summarization pipeline has been described in Section 3. Section 4 presents an evaluation of summaries generated from the system. We present an overview of relevant literature in Section 5. We end the paper with conclusions and some interesting further research directions in Section 6.

## 2 Design Goals

Consider that as a researcher reads a scientific article, she/he encounters numerous citations, most of them citing the foundational and seminal work that is important in that scientific domain. The text surrounding these citations is a valuable resource as it allows the author to make a statement about her



*Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 115–120,
Portland, Oregon, USA, 21 June 2011. 

viewpoint towards the cited articles. However, to researchers who are new to the field, or sometimes just as a side-effect of not being completely up-to-date with related work in a domain, these citations may pose a challenge to readers. A system that could generate a small summary of the collection of cited articles that is constructed specifically to relate to the claims made by the author citing them would be incredibly useful. It would also help the researcher determine if the cited work is relevant for her own research.

As an example of such a co-citation consider the following citation sentence:

*Various machine learning approaches have been proposed for chunking (Ramshaw and Marcus, 1995; Tjong Kim Sang, 2000a; Tjong Kim Sang et al. , 2000; Tjong Kim Sang, 2000b; Sassano and Utsuro, 2000; van Halteren, 2000).*

Now imagine the reader trying to determine about widely used *machine learning* approaches for *noun phrase chunking*. He would probably be required to go through these cited papers to understand what is similar and different in the variety of chunking approaches. Instead of going through these individual papers, it would be quicker if the user could get the summary of the topics in all those papers that talk about the usage of *machine learning* methods in *chunking*. SciSumm aims to automatically discover these points of comparison between the co-cited papers by taking into consideration the contextual needs of a user. When the user clicks on a co-citation in context, the system uses the text surrounding that co-citation as evidence of the information need.

## 3 System Overview

A high level overview of our system's architecture is presented in Figure 1. The system provides a web based interface for viewing and summarizing research articles in the ACL Anthology corpus, 2008. The summarization proceeds in three main stages as follows:

- A user may retrieve a collection of articles of interest by entering a query. SciSumm responds by returning a list of relevant articles, including the title and a snippet based summary. For this SciSumm uses standard retrieval from a Lucene index.

- A user can use the title, snippet summary and author information to find an article of interest. The actual article is rendered in HTML after the user clicks on one of the search results. The co-citations in the article are highlighted in bold and italics to mark them as points of interest for the user.

- If a user clicks on one, SciSumm responds by generating a query from the local context of the co-citation. That query is then used to select relevant portions of the co-cited articles, which are then used to generate the summary.

An example of a summary for a particular topic is displayed in Figure 2. This figure shows one of the clusters generated for the citation sentence "Various machine learning approaches have been proposed for chunking (Ramshaw and Marcus, 1995; Tjong Kim Sang, 2000a; Tjong Kim Sang et al. , 2000; Tjong Kim Sang, 2000b; Sassano and Utsuro, 2000; van Halteren, 2000)". The cluster has a label *Chunk, Tag, Word* and contains fragments from two of the papers discussing this topic. A ranked list of such clusters is generated, which allows for swift navigation between topics of interest for a user (Figure 3). This summary is tremendously useful as it informs the user of the different perspectives of co-cited authors towards a shared problem (in this case "Chunking"). More specifically, it informs the user as to how different or similar approaches are that were used for this research problem (which is "Chunking").

### 3.1 System Description

SciSumm has four primary modules that are central to the functionality of the system, as displayed in Figure 1. First, the Text Tiling module takes care of obtaining tiles of text relevant to the citation context. Next, the clustering module is used to generate labelled clusters using the text tiles extracted from the co-cited papers. The clusters are ordered according to relevance with respect to the generated query. This is accomplished by the Ranking Module.

In the following sections, we discuss each of the main modules in detail.

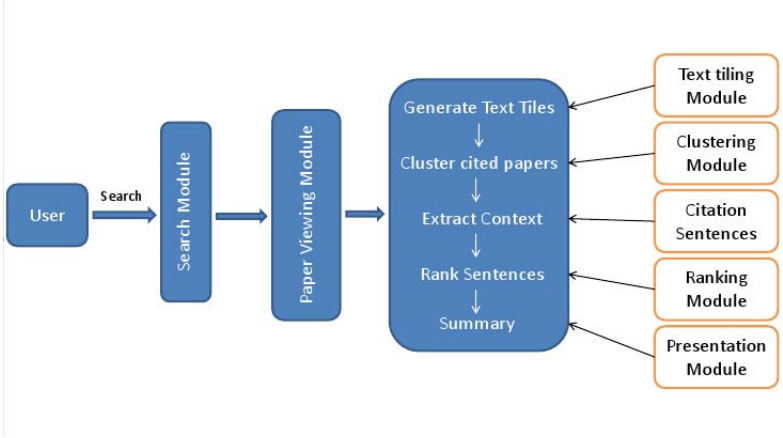


Figure 1: SciSumm summarization pipeline

### 3.2 Texttiling

The Text Tiling module uses the TextTiling algorithm (Hearst, 1997) for segmenting the text of each article. We have used text tiles as the basic unit for our summary since individual sentences are too short to stand on their own. This happens as a side-effect of the length of scientific articles. Sentences picked from different parts of several articles assembled together would make an incoherent summary. Once computed, text tiles are used to expand on the content viewed within the context associated with a co-citation. The intuition is that an embedded co-citation in a text tile is connected with the topic distribution of its context. Thus, we can use a computation of similarity between tiles and the context of the co-citation to rank clusters generated using Frequent Term based text clustering.

### 3.3 Frequent Term Based Clustering

The clustering module employs Frequent Term Based Clustering (Beil et al., 2002). For each co-citation, we use this clustering technique to cluster all the of the extracted text tiles generated by segmenting each of the co-cited papers. We settled on this clustering approach for the following reasons:

- Text tile contents coming from different papers constitute a sparse vector space, and thus the centroid based approaches would not work very well for integrating content across papers.
- Frequent Term based clustering is extremely fast in execution time as well as and relatively efficient in terms of space requirements.
- A frequent term set is generated for each cluster, which gives a comprehensible description that can be used to label the cluster.

Frequent Term Based text clustering uses a group of frequently co-occurring terms called a frequent term set. We use a measure of entropy to rank these frequent term sets. Frequent term sets provide a clean clustering that is determined by specifying the number of overlapping documents containing more than one frequent term set. The algorithm uses the first $k$ term sets if all the documents in the document collection are clustered. To discover all the possible candidates for clustering, i.e., term sets, we used the *Apriori* algorithm (Agrawal et al., 1994), which identifies the sets of terms that are both relatively frequent and highly correlated with one another.

### 3.4 Cluster Ranking

The ranking module uses cosine similarity between the query and the centroid of each cluster to rank all the clusters generated by the clustering module. The context of a co-citation is restricted to the text of the segment in which the co-citation is found. In this way we attempt to leverage the expert knowledge of the author as it is encoded in the local context of the co-citation.

## 4 Evaluation

We have taken great care in the design of the evaluation for the SciSumm summarization system. In a

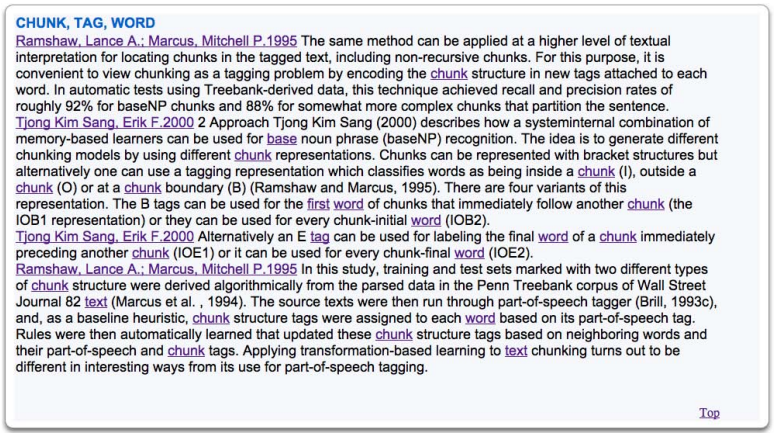


Figure 2: Example of a summary generated by our system. We can see that the clusters are cross cutting across different papers, thus giving the user a multi-document summary.

typical evaluation of a multi-document summarization system, gold standard summaries are created by hand and then compared against fixed length generated summaries. It was necessary to prepare our own evaluation corpus, consisting of gold standard summaries created for a randomly selected set of co-citations because such an evaluation corpus does not exist for this task.

### 4.1 Experimental Setup

An important target user population for multi-document summarization of scientific articles is graduate students. Hence to get a measure of how well the summarization system is performing, we asked 2 graduate students who have been working in the computational linguistics community to create gold standard summaries of a fixed length (8 sentences $\sim$ 200 words) for 10 randomly selected co-citations. We obtained two different gold standard summaries for each co-citation (i.e., 20 gold standard summaries total). Our evaluation is designed to measure the quality of the content selection. In future work, we will evaluate the usability of the SciSumm system using a task based evaluation.

In the absence of any other multi-document summarization system in the domain of scientific article summarization, we used a widely used and freely available multi-document summarization system called MEAD (Radev, 2004) as our baseline. MEAD uses centroid based summarization to create informative clusters of topics. We use the default configuration of MEAD in which MEAD uses length, position and centroid for ranking each sentence. We did not use query focussed summarization with MEAD. We evaluate its performance with the same gold standard summaries we use to evaluate SciSumm. For generating a summary from our system we used sentences from the tiles that are clustered in the top ranked cluster. Once all of the extracts included in that entire cluster are exhausted, we move on to the next highly ranked cluster. In this way we prepare a summary comprising of 8 highly relevant sentences.

### 4.2 Results

For measuring performance of the two summarization systems (SciSumm and MEAD), we compute the ROUGE metric based on the 2 * 10 gold standard summaries that were manually created. ROUGE has been traditionally used to compute the performance based on the N-gram overlap (ROUGE-N) between the summaries generated by the system and the target gold standard summaries. For our evaluation we used two different versions of the ROUGE metric, namely ROUGE-1 and ROUGE-2, which correspond to measures of the unigram and bigram overlap respectively. We computed four metrics in order to get a complete picture of how SciSumm performs in relation to the baseline, namely ROUGE-1 F-measure, ROUGE-1 Recall, ROUGE-2 F-measure, and ROUGE-2 Recall.

From the results presented in Figure 4 and 5, we can see that our system performs well on average in comparison to the baseline. Especially important is

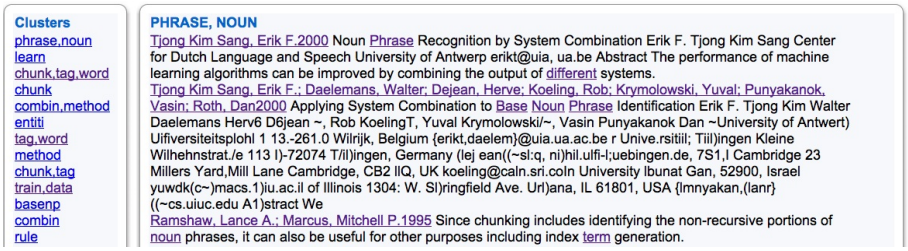


Figure 3: Clusters generated in response to a user click on the co-citation. The list of clusters in the left pane gives a bird-eye view of the topics which are present in the co-cited papers

Table 1: Average ROUGE results. * represents improvement significant at $p < .05$, † at $p < .01$.

| Metric | MEAD | SciSumm |
|---|---|---|
| ROUGE-1 F-measure | 0.3680 | 0.5123 † |
| ROUGE-1 Recall | 0.4168 | 0.5018 |
| ROUGE-1 Precision | 0.3424 | 0.5349 † |
| ROUGE-2 F-measure | 0.1598 | 0.3303 * |
| ROUGE-2 Recall | 0.1786 | 0.3227 * |
| ROUGE-2 Precision | 0.1481 | 0.3450 † |

the performance of the system on recall measures, which shows the most dramatic advantage over the baseline. To measure the statistical significance of this result, we carried out a Student T-Test, the results of which are presented in the results section in Table 1. It is apparent from the p-values generated by T-Test that our system performs significantly better than MEAD on three of the metrics on which both the systems were evaluated using $(p < 0.05)$ as the criterion for statistical significance. This supports the view that summaries perceived as higher in value are generated using a query focused technique, where the query is generated automatically from the context of the co-citation.

## 5 Previous Work

Surprisingly, not many approaches to the problem of summarization of scientific articles have been proposed in the past. Qazvinian et al. (2008) present a summarization approach that can be seen as the converse of what we are working to achieve. Rather than summarizing multiple papers cited in the same source article, they summarize different viewpoints expressed towards the same paper from different papers that cite it. Nanba et al. (1999) argue in their work that a co-citation frequently implies a consistent viewpoint towards the cited articles. Another approach that uses bibliographic coupling has been used for gathering different viewpoints from which to summarize a document (Kaplan et al., 2008). In our work we make use of this insight by generating a query to focus our multi-document summary from the text closest to the citation.

## 6 Conclusion And Future Work

In this demo, we present SciSumm, which is an interactive multi-document summarization system for scientific articles. Our evaluation shows that the SciSumm approach to content selection outperforms another widely used multi-document summarization system for this summarization task.

Our long term goal is to expand the capabilities of SciSumm to generate literature surveys of larger document collections from less focused queries. This more challenging task would require more control over filtering and ranking in order to avoid generating summaries that lack focus. To this end, a future improvement that we plan to use is a variant on MMR (Maximum Marginal Relevance) (Carbonell et al., 1998), which can be used to optimize the diversity of selected text tiles as well as the relevance based ordering of clusters, i.e., so that more diverse sets of extracts from the co-cited articles will be placed at the ready fingertips of users.

Another important direction is to refine the interaction design through task-based user studies. As we collect more feedback from students and researchers through this process, we will used the insights gained to achieve a more robust and effective implementation.

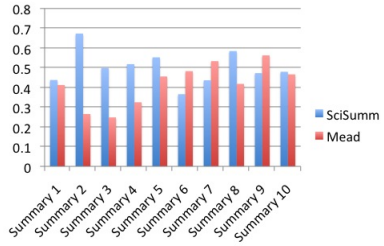


Figure 4: ROUGE-1 Recall

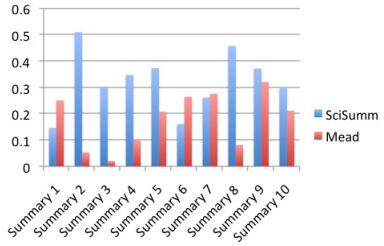


Figure 5: ROUGE-2 Recall

## 7 Acknowledgements

This research was supported in part by NSF grant EEC-064848 and ONR grant N00014-10-1-0277.

## References

Agrawal R. and Srikant R. 1994. Fast Algorithm for Mining Association Rules In *Proceedings of the 20th VLDB Conference* Santiago, Chile, 1994

Baxendale, P. 1958. Machine-made index for technical literature - an experiment. *IBM Journal of Research and Development*

Beil F., Ester M. and Xu X 2002. Frequent-Term based Text Clustering In *Proceedings of SIGKDD '02* Edmonton, Alberta, Canada

Carbonell J. and Goldstein J. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries In *Research and Development in Information Retrieval*, pages 335–336

Councill I. G. , Giles C. L. and Kan M. 2008. ParsCit: An open-source CRF reference string parsing package *INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION European Language Resources Association*

Edmundson, H.P. 1969. New methods in automatic extracting. *Journal of ACM.*

Hearst M.A. 1997 TextTiling: Segmenting text into multi-paragraph subtopic passages *In proceedings of LREC 2004, Lisbon, Portugal, May 2004*

Joseph M. T. and Radev D. R. 2007. Citation analysis, centrality, and the ACL Anthology

Kupiec J. , Pedersen J. , Chen F. 1995. A training document summarizer. In *Proceedings SIGIR '95*, pages 68-73, New York, NY, USA. 28(1):114–133.

Luhn, H. P. 1958. *IBM Journal of Research Development.*

Mani I. , Bloedorn E. 1997. Multi-Document Summarization by graph search and matching In *AAAI/IAAI*, pages 622-628. [15, 16].

Nanba H. , Okumura M. 1999. Towards Multi-paper Summarization Using Reference Information In *Proceedings of IJCAI-99*, pages 926–931 .

Paice CD. 1990. Constructing Literature Abstracts by Computer: Techniques and Prospects *Information Processing and Management* Vol. 26, No.1, pp, 171-186, 1990

Qazvinian V. , Radev D.R 2008. Scientific Paper summarization using Citation Summary Networks In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 689–696 Manchester, August 2008

Radev D. R . , Jing H. and Budzikowska M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility based evaluation, and user studies In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 21-30, Morristown, NJ, USA. [12, 16, 17].

Radev, Dragomir. 2004. *MEAD - a platform for multidocument multilingual text summarization.* In proceedings of LREC 2004, Lisbon, Portugal, May 2004.

Teufel S. , Moens M. 2002. Summarizing Scientific Articles - Experiments with Relevance and Rhetorical Status In *Journal of Computational Linguistics*, MIT Press.

Hal Daume III , Marcu D. 2006. Bayesian query-focussed summarization. In *Proceedings of the Conference of the Association for Computational Linguistics*, ACL.

Eisenstein J , Barzilay R. 2008. Bayesian unsupervised topic segmentation In *EMNLP-SIGDAT.*

Barzilay R , Lee L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization In *Proceedings of 3rd Asian Semantic Web Conference (ASWC 2008), pp.182-188,.*

Kaplan D , Tokunaga T. 2008. Sighting citation sights: A collective-intelligence approach for automatic summarization of research papers using C-sites In *HLT-NAACL.*