Coherent Citation-Based Summarization of Scientific Papers

Amjad Abu-Jbara
EECS Department
University of Michigan
Ann Arbor, MI, USA
amjbara@umich.edu

Dragomir Radev
EECS Department and
School of Information
University of Michigan
Ann Arbor, MI, USA
radev@umich.edu

Abstract

In citation-based summarization, text written by several researchers is leveraged to identify the important aspects of a target paper. Previous work on this problem focused almost exclusively on its extraction aspect (i.e. selecting a representative set of citation sentences that highlight the contribution of the target paper). Meanwhile, the fluency of the produced summaries has been mostly ignored. For example, diversity, readability, cohesion, and ordering of the sentences included in the summary have not been thoroughly considered. This resulted in noisy and confusing summaries. In this work, we present an approach for producing readable and cohesive citation-based summaries. Our experiments show that the proposed approach outperforms several baselines in terms of both extraction quality and fluency.

1 Introduction

Scientific research is a cumulative activity. The work of downstream researchers depends on access to upstream discoveries. The footnotes, end notes, or reference lists within research articles make this accumulation possible. When a reference appears in a scientific paper, it is often accompanied by a span of text describing the work being cited.

We name the sentence that contains an explicit reference to another paper *citation sentence*. Citation sentences usually highlight the most important aspects of the cited paper such as the research problem it addresses, the method it proposes, the good results it reports, and even its drawbacks and limitations.

By aggregating all the citation sentences that cite a paper, we have a rich source of information about it. This information is valuable because human experts have put their efforts to read the paper and summarize its important contributions.

One way to make use of these sentences is creating a summary of the target paper. This summary is different from the abstract or a summary generated from the paper itself. While the abstract represents the author's point of view, the citation summary is the summation of multiple scholars' viewpoints. The task of summarizing a scientific paper using its set of citation sentences is called citation-based summarization.

There has been previous work done on citation-based summarization (Nanba et al., 2000; Elkiss et al., 2008; Qazvinian and Radev, 2008; Mei and Zhai, 2008; Mohammad et al., 2009). Previous work focused on the extraction aspect; i.e. analyzing the collection of citation sentences and selecting a representative subset that covers the main aspects of the paper. The cohesion and the readability of the produced summaries have been mostly ignored. This resulted in noisy and confusing summaries.

In this work, we focus on the coherence and readability aspects of the problem. Our approach produces citation-based summaries in three stages: preprocessing, extraction, and postprocessing. Our experiments show that our approach produces better summaries than several baseline summarization systems.

The rest of this paper is organized as follows. After we examine previous work in Section 2, we outline the motivation of our approach in Section 3. Section 4 describes the three stages of our summarization system. The evaluation and the results are presented in Section 5. Section 6 concludes the paper.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 500–509,
Portland, Oregon, June 19-24, 2011.

2 Related Work

The idea of analyzing and utilizing citation information is far from new. The motivation for using information latent in citations has been explored tens of years back (Garfield et al., 1984; Hodges, 1972). Since then, there has been a large body of research done on citations.

Nanba and Okumura (2000) analyzed citation sentences and automatically categorized citations into three groups using 160 pre-defined phrase-based rules. They also used citation categorization to support a system for writing surveys (Nanba and Okumura, 1999). Newman (2001) analyzed the structure of the citation networks. Teufel et al. (2006) addressed the problem of classifying citations based on their function.

Siddharthan and Teufel (2007) proposed a method for determining the scientific attribution of an article by analyzing citation sentences. Teufel (2007) described a rhetorical classification task, in which sentences are labeled as one of Own, Other, Background, Textual, Aim, Basis, or Contrast according to their role in the authors argument. In parts of our approach, we were inspired by this work.

Elkiss et al. (2008) performed a study on citation summaries and their importance. They concluded that citation summaries are more focused and contain more information than abstracts. Mohammad et al. (2009) suggested using citation information to generate surveys of scientific paradigms.

Qazvinian and Radev (2008) proposed a method for summarizing scientific articles by building a similarity network of the citation sentences that cite the target paper, and then applying network analysis techniques to find a set of sentences that covers as much of the summarized paper facts as possible. We use this method as one of the baselines when we evaluate our approach. Qazvinian et al. (2010) proposed a citation-based summarization method that first extracts a number of important keyphrases from the set of citation sentences, and then finds the best subset of sentences that covers as many keyphrases as possible. Qazvinian and Radev (2010) addressed the problem of identifying the non-explicit citing sentences to aid citation-based summarization.

3 Motivation

The coherence and readability of citation-based summaries are impeded by several factors. First, many citation sentences cite multiple papers besides the target. For example, the following is a citation sentence that appeared in the NLP literature and talked about Resnik's (1999) work.

(1) Grefenstette and Nioche (2000) and Jones and Ghani (2000) use the web to generate corpora for languages where electronic resources are scarce, while ***Resnik (1999)*** *describes a method for mining the web for bilingual texts.*

The first fragment of this sentence describes different work other than Resnik's. The contribution of Resnik is mentioned in the underlined fragment. Including the irrelevant fragments in the summary causes several problems. First, the aim of the summarization task is to summarize the contribution of the target paper using minimal text. These fragments take space in the summary while being irrelevant and less important. Second, including these fragments in the summary breaks the context and, hence, degrades the readability and confuses the reader. Third, the existence of irrelevant fragments in a sentence makes the ranking algorithm assign a low weight to it although the relevant fragment may cover an aspect of the paper that no other sentence covers.

A second factor has to do with the ordering of the sentences included in the summary. For example, the following are two other citation sentences for Resnik (1999).

*(2) Mining the Web for bilingual text (****Resnik, 1999****) is not likely to provide sufficient quantities of high quality data.*

(3) ***Resnik (1999)*** *addressed the issue of language identification for finding Web pages in the languages of interest.*

If these two sentences are to be included in the summary, the reasonable ordering would be to put the second sentence first.

Thirdly, in some instances of citation sentences, the reference is not a syntactic constituent in the sen-

tence. It is added just to indicate the existence of citation. For example, in sentence (2) above, the reference could be safely removed from the sentence without hurting its grammaticality.

In other instances (e.g. sentence (3) above), the reference is a syntactic constituent of the sentence and removing it makes the sentence ungrammatical. However, in certain cases, the reference could be replaced with a suitable pronoun (i.e. he, she or they). This helps avoid the redundancy that results from repeating the author name(s) in every sentence.

Finally, a significant number of citation sentences are not suitable for summarization (Teufel et al., 2006) and should be filtered out. The following sentences are two examples.

(4) The two algorithms we employed in our dependency parsing model are the Eisner parsing ***(Eisner, 1996)*** *and Chu-Lius algorithm (Chu and Liu, 1965).*

(5) This type of model has been used by, among others, ***Eisner (1996)****.*

Sentence (4) appeared in a paper by Nguyen et al (2007). It does not describe any aspect of Eisner's work, rather it informs the reader that Nguyen et al. used Eisner's algorithm in their model. There is no value in adding this sentence to the summary of Eisner's paper. Teufel (2007) reported that a significant number of citation sentences (67% of the sentences in her dataset) were of this type.

Likewise, the comprehension of sentence (5) depends on knowing its context (i.e. its surrounding sentences). This sentence alone does not provide any valuable information about Eisner's paper and should not be added to the summary unless its context is extracted and included in the summary as well.

In our approach, we address these issues to achieve the goal of improving the coherence and the readability of citation-based summaries.

4 Approach

In this section we describe a system that takes a scientific paper and a set of citation sentences that cite it as input, and outputs a citation summary of the paper. Our system produces the summaries in three stages. In the first stage, the citation sentences are preprocessed to rule out the unsuitable sentences and the irrelevant fragments of sentences. In the second stage, a number of citation sentences that cover the various aspects of the paper are selected. In the last stage, the selected sentences are post-processed to enhance the readability of the summary. We describe the stages in the following three subsections.

4.1 Preprocessing

The aim of this stage is to determine which pieces of text (sentences or fragments of sentences) should be considered for selection in the next stage and which ones should be excluded. This stage involves three tasks: reference tagging, reference scope identification, and sentence filtering.

4.1.1 Reference Tagging

A citation sentence contains one or more references. At least one of these references corresponds to the target paper. When writing scientific articles, authors usually use standard patterns to include pointers to their references within the text. We use pattern matching to tag such references. The reference to the target is given a different tag than the references to other papers.

The following example shows a citation sentence with all the references tagged and the target reference given a different tag.

*In <**TREF**>**Resnik (1999)**</**TREF**>, <REF>Nie, Simard, and Foster (2001)</REF>, <REF>Ma and Liberman (1999)</REF>, and <REF>Resnik and Smith (2002)</REF>, the Web is harvested in search of pages that are available in two languages.*

4.1.2 Identifying the Reference Scope

In the previous section, we showed the importance of identifying the scope of the target reference; i.e. the fragment of the citation sentence that corresponds to the target paper. We define the scope of a reference as the shortest fragment of the citation sentence that contains the reference and could form a grammatical sentence if the rest of the sentence was removed.

To find such a fragment, we use a simple yet adequate heuristic. We start by parsing the sentence using the link grammar parser (Sleator and Temperley,

1991). Since the parser is not trained on citation sentences, we replace the references with placeholders before passing the sentence to the parser. Figure 1 shows a portion of the parse tree for Sentence (1) (from Section 1).

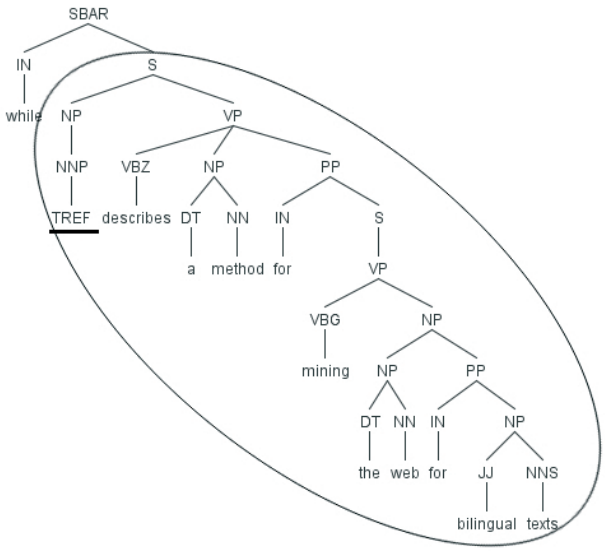

Figure 1: An example showing the scope of a target reference

We extract the scope of the reference from the parse tree as follows. We find the smallest subtree rooted at an S node (sentence clause node) and contains the target reference node. We extract the text that corresponds to this subtree if it is grammatical. Otherwise, we find the second smallest subtree rooted at an S node and so on. For example, the parse tree shown in Figure 1 suggests that the scope of the reference is:

Resnik (1999) *describes a method for mining the web for bilingual texts.*

4.1.3 Sentence Filtering

The task in this step is to detect and filter out unsuitable sentences; i.e., sentences that depend on their context (e.g. Sentence (5) above) or describe the own work of their authors, not the contribution of the target paper (e.g Sentence (4) above). Formally, we classify the citation sentences into two classes: suitable and unsuitable sentences. We use a machine learning technique for this purpose. We extract a number of features from each sentence and train a classification model using these features. The trained model is then used to classify the sentences. We use Support Vector Machines (SVM) with linear kernel as our classifier. The features that we use in this step and their descriptions are shown in Table 1.

4.2 Extraction

In the first stage, the sentences and sentence fragments that are not useful for our summarization task are ruled out. The input to this stage is a set of citation sentences that are believed to be suitable for the summary. From these sentences, we need to select a representative subset. The sentences are selected based on these three main properties:

First, they should cover diverse aspects of the paper. Second, the sentences that cover the same aspect should not contain redundant information. For example, if two sentences talk about the drawbacks of the target paper, one sentence can mention the computation inefficiency, while the other criticize the assumptions the paper makes. Third, the sentences should cover as many important facts about the target paper as possible using minimal text.

In this stage, the summary sentences are selected in three steps. In the first step, the sentences are classified into five functional categories: *Background, Problem Statement, Method, Results*, and *Limitations*. In the second step, we cluster the sentences within each category into clusters of similar sentences. In the third step, we compute the LexRank (Erkan and Radev, 2004) values for the sentences within each cluster. The summary sentences are selected based on the classification, the clustering, and the LexRank values.

4.2.1 Functional Category Classification

We classify the citation sentences into the five categories mentioned above using a machine learning technique. A classification model is trained on a number of features (Table 2) extracted from a labeled set of citation sentences. We use SVM with linear kernel as our classifier.

4.2.2 Sentence Clustering

In the previous step we determined the category of each citation sentence. It is very likely that sentences from the same category contain similar or overlapping information. For example, Sentences (6), (7), and (8) below appear in the set of citation

Feature	Description
Similarity to the target paper	The value of the cosine similarity (using TF-IDF vectors) between the citation sentence and the target paper.
Headlines	The section in which the citation sentence appeared in the citing paper. We recognize 10 section types such as *Introduction, Related Work, Approach, etc*.
Relative position	The relative position of the sentence in the section and the paragraph in which it appears
First person pronouns	This feature takes a value of 1 if the sentence contains a first person pronoun (I, we, our, us, etc.), and 0 otherwise.
Tense of the first verb	A sentence that contains a past tense verb near its beginning is more likely to be describing previous work.
Determiners	Demonstrative Determiners (this, that, these, those, and which) and Alternative Determiners (another, other). The value of this feature is the relative position of the first determiner (if one exists) in the sentence.

Table 1: The features used for sentence filtering

Feature	Description
Similarity to the sections of the target paper	The sections of the target paper are categorized into five categories: 1) *Introduction, Motivation, Problem Statement*. 2) *Background, Prior Work, Previous Work,* and *Related Work*. 3) *Experiments, Results,* and *Evaluation*. 4) *Discussion, Conclusion,* and *Future work*. 5) All other headlines. The value of this feature is the cosine similarity (using TF-IDF vectors) between the sentence and the text of the sections of each of the five section categories.
Headlines	This is the same feature that we used for sentence filtering in Section 4.1.3.
Number of references in the sentence	Sentences that contain multiple references are more likely to be *Background* sentences.
Verbs	We use all the verbs that their lemmatized form appears in at least three sentences that belong to the same category in the training set. Auxiliary verbs are excluded. In our annotated dataset, for example, the verb *propose* appeared in 67 sentences from the *Methodology* category, while the verbs *outperform* and *achieve* appeared in 33 *Result* sentences.

Table 2: The features used for sentence classification

sentences that cite Goldwater and Griffiths' (2007). These sentences belong to the same category (i.e *Method*). Both Sentences (6) and (7) convey the same information about Goldwater and Griffiths (2007) contribution. Sentence (8), however, describes a different aspect of the paper methodology.

(6) ***Goldwater and Griffiths (2007)*** *proposed an information-theoretic measure known as the Variation of Information (VI)*

(7) ***Goldwater and Griffiths (2007)*** *propose using the Variation of Information (VI) metric*

(8) A fully-Bayesian approach to unsupervised POS tagging has been developed by ***Goldwater and Griffiths (2007)*** *as a viable alternative to the traditional maximum likelihood-based HMM approach.*

Clustering divides the sentences of each category into groups of similar sentences. Following Qazvinian and Radev (2008), we build a cosine similarity graph out of the sentences of each category. This is an undirected graph in which nodes are sentences and edges represent similarity relations. Each edge is weighted by the value of the cosine similarity (using TF-IDF vectors) between the two sentences the edge connects. Once we have the similarity network constructed, we partition it into clusters using a community finding technique. We use the Clauset algorithm (Clauset et al., 2004), a hierarchical agglomerative community finding algorithm that runs in linear time.

4.2.3 Ranking

Although the sentences that belong to the same cluster are similar, they are not necessarily equally important. We rank the sentences within each cluster by computing their LexRank (Erkan and Radev, 2004). Sentences with higher rank are more important.

4.2.4 Sentence Selection

At this point we have determined (Figure 2), for each sentence, its category, its cluster, and its relative importance. Sentences are added to the summary in order based on their category, the size of their clusters, then their LexRank values. The categories are

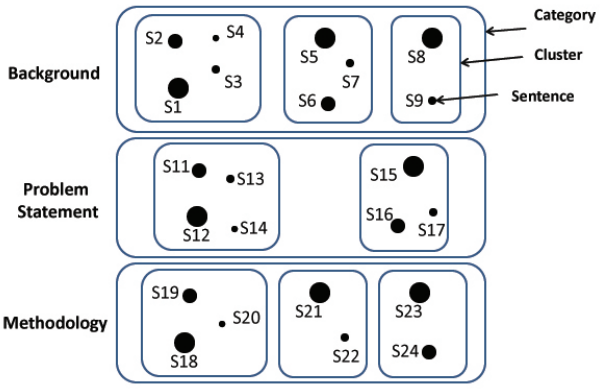

Figure 2: Example illustrating sentence selection

ordered as *Background, Problem, Method, Results*, then *Limitations*. Clusters within each category are ordered by the number of sentences in them whereas the sentences of each cluster are ordered by their LexRank values.

In the example shown in Figure 2, we have three categories. Each category contains several clusters. Each cluster contains several sentences with different LexRank values (illustrated by the sizes of the dots in the figure.) If the desired length of the summary is 3 sentences, the selected sentences will be in order S1, S12, then S18. If the desired length is 5, the selected sentences will be S1, S5, S12, S15, then S18.

4.3 Postprocessing

In this stage, we refine the sentences that we extracted in the previous stage. Each citation sentence will have the target reference (the author's names and the publication year) mentioned at least once. The reference could be either syntactically and semantically part of the sentence (e.g. Sentence (3) above) or not (e.g. Sentence (2)). The aim of this refinement step is to avoid repeating the author's names and the publication year in every sentence. We keep the author's names and the publication year only in the first sentence of the summary. In the following sentences, we either replace the reference with a suitable personal pronoun or remove it. The reference is replaced with a pronoun if it is part of the sentence and this replacement does not make the sentence ungrammatical. The reference is removed if it is not part of the sentence. If the sentence contains references for other papers, they are removed if this doesn't hurt the grammaticality of the sentence.

To determine whether a reference is part of the sentence or not, we again use a machine learning approach. We train a model on a set of labeled sentences. The features used in this step are listed in Table 3. The trained model is then used to classify the references that appear in a sentence into three classes: *keep*, *remove*, *replace*. If a reference is to be replaced, and the paper has one author, we use "*he/she*" (we do not know if the author is male or female). If the paper has two or more authors, we use "*they*".

5 Evaluation

We provide three levels of evaluation. First, we evaluate each of the components in our system separately. Then we evaluate the summaries that our system generate in terms of extraction quality. Finally, we evaluate the coherence and readability of the summaries.

5.1 Data

We use the ACL Anthology Network (AAN) (Radev et al., 2009) in our evaluation. AAN is a collection of more than 16000 papers from the Computational Linguistics journal, and the proceedings of the ACL conferences and workshops. AAN provides all citation information from within the network including the citation network, the citation sentences, and the citation context for each paper.

We used 55 papers from AAN as our data. The papers have a variable number of citation sentences, ranging from 15 to 348. The total number of citation sentences in the dataset is 4,335. We split the data randomly into two different sets; one for evaluating the components of the system, and the other for evaluating the extraction quality and the readability of the generated summaries. The first set (*dataset1*, henceforth) contained 2,284 sentences coming from 25 papers. We asked humans with good background in NLP (the area of the annotated papers) to provide two annotations for each sentence in this set: 1) label the sentence as *Background, Problem, Method, Result, Limitation,* or *Unsuitable*, 2) for each reference in the sentence, determine whether it could be *replaced* with a pronoun, *removed*, or should be *kept*.

Feature	Description
Part-of-speech (POS) tag	We consider the POS tags of the reference, the word before, and the word after. Before passing the sentence to the POS tagger, all the references in the sentence are replaced by placeholders.
Style of the reference	It is common practice in writing scientific papers to put the whole citation between parenthesis when the authors are not a constitutive part of the enclosing sentence, and to enclose just the year between parenthesis when the author's name is a syntactic constituent in the sentence.
Relative position of the reference	This feature takes one of three values: *first*, *last*, and *inside*.
Grammaticality	Grammaticality of the sentence if the reference is removed/replaced. Again, we use the Link Grammar parser (Sleator and Temperley, 1991) to check the grammaticality

Table 3: The features used for author name replacement

Each sentence was given to 3 different annotators. We used the majority vote labels.

We use Kappa coefficient (Krippendorff, 2003) to measure the inter-annotator agreement. Kappa coefficient is defined as:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

where $P(A)$ is the relative observed agreement among raters and $P(E)$ is the hypothetical probability of chance agreement.

The agreement among the three annotators on distinguishing the *unsuitable* sentences from the other five categories is 0.85. On Landis and Kochs(1977) scale, this value indicates an *almost perfect* agreement. The agreement on classifying the sentences into the five functional categories is 0.68. On the same scale this value indicates *substantial agreement*.

The second set (*dataset2*, henceforth) contained 30 papers (2051 sentences). We asked humans with a good background in NLP (the papers topic) to generate a readable, coherent summary for each paper in the set using its citation sentences as the source text. We asked them to fix the length of the summaries to 5 sentences. Each paper was assigned to two humans to summarize.

5.2 Component Evaluation

Reference Tagging and Reference Scope Identification Evaluation: We ran our reference tagging and scope identification components on the 2,284 sentences in *dataset1*. Then, we went through the tagged sentences and the extracted scopes, and counted the number of correctly/incorrectly tagged (extracted)/missed references (scopes). Our tagging component achieved 98.2% precision and 94.4% recall. The reference to the target paper was tagged correctly in all the sentences.

-	Bkgrnd	Prob	Method	Results	Limit.
Precision	64.62%	60.01%	88.66%	76.05%	33.53%
Recall	72.47%	59.30%	75.03%	82.29%	59.36%
F1	68.32%	59.65%	81.27%	79.04%	42.85%

Table 4: Precision and recall results achieved by our citation sentence classifier

Our scope identification component extracted the scope of target references with good precision (86.4%) but low recall (35.2%). In fact, extracting a useful scope for a reference requires more than just finding a grammatical substring. In future work, we plan to employ text regeneration techniques to improve the recall by generating grammatical sentences from ungrammatical fragments.

Sentence Filtering Evaluation: We used Support Vector Machines (SVM) with linear kernel as our classifier. We performed 10-fold cross validation on the labeled sentences (*unsuitable* vs *all other categories*) in *dataset1*. Our classifier achieved 80.3% accuracy.

Sentence Classification Evaluation: We used SVM in this step as well. We also performed 10-fold cross validation on the labeled sentences (the five functional categories). This classifier achieved 70.1% accuracy. The precision and recall for each category are given in Table 4

Author Name Replacement Evaluation: The classifier used in this task is also SVM. We performed 10-fold cross validation on the labeled sentences of *dataset1*. Our classifier achieved 77.41% accuracy.

Produced using our system
There has been a large number of studies in tagging and morphological disambiguation using various techniques such as statistical techniques, e.g. constraint-based techniques and transformation-based techniques. A thorough removal of ambiguity requires a syntactic process. A rule-based tagger described in Voutilainen (1995) was equipped with a set of guessing rules that had been hand-crafted using knowledge of English morphology and intuitions. The precision of rule-based taggers may exceed that of the probabilistic ones. The construction of a linguistic rule-based tagger, however, has been considered a difficult and time-consuming task.
Produced using Qazvinian and Radev (2008) system
Another approach is the rule-based or constraint-based approach, recently most prominently exemplified by the Constraint Grammar work (Karlsson et al. , 1995; Voutilainen, 1995b; Voutilainen et al. , 1992; Voutilainen and Tapanainen, 1993), where a large number of hand-crafted linguistic constraints are used to eliminate impossible tags or morphological parses for a given word in a given context. Some systems even perform the POS tagging as part of a syntactic analysis process (Voutilainen, 1995). A rule-based tagger described in (Voutilainen, 1995) is equipped with a set of guessing rules which has been hand-crafted using knowledge of English morphology and intuition. Older versions of EngCG (using about 1,150 constraints) are reported (butilainen et al. 1992; Voutilainen and HeikkiUi 1994; Tapanainen and Voutilainen 1994; Voutilainen 1995) to assign a correct analysis to about 99.7% of all words while each word in the output retains 1.04-1.09 alternative analyses on an average, i.e. some of the ambiguities remait unresolved. We evaluate the resulting disambiguated text by a number of metrics defined as follows (Voutilainen, 1995a).

Table 5: Sample Output

5.3 Extraction Evaluation

To evaluate the extraction quality, we use *dataset2* (that has never been used for training or tuning any of the system components). We use our system to generate summaries for each of the 30 papers in *dataset2*. We also generate summaries for the papers using a number of baseline systems (described in Section 5.3.1). All the generated summaries were 5 sentences long. We use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) based on the longest common substrings (ROUGE-L) as our evaluation metric.

5.3.1 Baselines

We evaluate the extraction quality of our system (FL) against 7 different baselines. In the first baseline, the sentences are selected randomly from the set of citation sentences and added to the summary. The second baseline is the MEAD summarizer (Radev et al., 2004) with all its settings set to default. The third baseline is LexRank (Erkan and Radev, 2004) run on the entire set of citation sentences of the target paper. The forth baseline is Qazvinian and Radev (2008) citation-based summarizer (QR08) in which the citation sentences are first clustered then the sentences within each cluster are ranked using LexRank. The remaining baselines are variations of our system produced by removing one component from the pipeline at a time. In one variation (FL-1), we remove the *sentence filtering* component. In another variation (FL-2), we remove the *sentence classification* component; so, all the sentences are assumed to come from one category in the subsequent components. In a third variation (FL-3), the clustering component is removed. To make the comparison of the extraction quality to those baselines fair, we remove the *author name replacement* component from our system and all its variations.

5.3.2 Results

Table 6 shows the average ROUGE-L scores (with 95% confidence interval) for the summaries of the 30 papers in *dataset2* generated using our system and the different baselines. The two human summaries were used as models for comparison. The *Human* score reported in the table is the result of comparing the two human summaries to each others. Statistical significance was tested using a 2-tailed paired t-test. The results are statistically significant at the 0.05 level.

The results show that our approach outperforms all the baseline techniques. It achieves higher ROUGE-L score for most of the papers in our testing set. Comparing the score of FL-1 to the score of FL shows that sentence filtering has a significant impact on the results. It also shows that the classification and clustering components both improve the extraction quality.

5.4 Coherence and Readability Evaluation

We asked human judges (not including the authors) to rate the coherence and readability of a number of summaries for each of *dataset2* papers. For each paper we evaluated 3 summaries. The sum-

-	**Human**	**Random**	**MEAD**	**LexRank**	**QR08**
ROUGE-L	0.733	0.398	0.410	0.408	0.435
-	**FL-1**	**FL-2**	**FL-3**	**FL**	-
ROUGE-L	0.475	0.511	0.525	**0.539**	-

Table 6: Extraction Evaluation

Average Coherence Rating	Number of summaries		
	Human	FL	QV08
1≤ coherence <2	0	9	17
2≤ coherence <3	3	11	12
3≤ coherence <4	16	9	1
4≤ coherence ≤5	11	1	0

Table 7: Coherence Evaluation

mary that our system produced, the human summary, and a summary produced by Qazvinian and Radev (2008) summarizer (the best baseline - after our system and its variations - in terms of extraction quality as shown in the previous subsection.) The summaries were randomized and given to the judges without telling them how each summary was produced. The judges were not given access to the source text. They were asked to use a five point-scale to rate how coherent and readable the summaries are, where 1 means that the summary is totally incoherent and needs significant modifications to improve its readability, and 5 means that the summary is coherent and no modifications are needed to improve its readability. We gave each summary to 5 different judges and took the average of their ratings for each summary. We used Weighted Kappa with linear weights (Cohen, 1968) to measure the inter-rater agreement. The Weighted Kappa measure between the five groups of ratings was 0.72.

Table 7 shows the number of summaries in each rating range. The results show that our approach significantly improves the coherence of citation-based summarization. Table 5 shows two sample summaries (each 5 sentences long) for the Voutilainen (1995) paper. One summary was produced using our system and the other was produced using Qazvinian and Radev (2008) system.

6 Conclusions

In this paper, we presented a new approach for citation-based summarization of scientific papers that produces readable summaries. Our approach involves three stages. The first stage preprocesses the set of citation sentences to filter out the irrelevant sentences or fragments of sentences. In the second stage, a representative set of sentences are extracted and added to the summary in a reasonable order. In the last stage, the summary sentences are refined to improve their readability. The results of our experiments confirmed that our system outperforms several baseline systems.

Acknowledgments

This work is in part supported by the National Science Foundation grant "iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains", jointly awarded to University of Michigan and University of Maryland as IIS 0705832, and in part by the NIH Grant U54 DA021519 to the National Center for Integrative Biomedical Informatics.

Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the supporters.

References

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Dec.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213 – 220.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *J. Am. Soc. Inf. Sci. Technol.*, 59(1):51–62.

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

E. Garfield, Irving H. Sher, and R. J. Torpie. 1984. *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc., Philadelphia, Pennsylvania, USA.

T. L. Hodges. 1972. Citation indexing-its theory and application in science, technology, and humanities. *Ph.D. thesis, University of California at Berkeley.Ph.D. thesis, University of California at Berkeley.*

Klaus H. Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, 2nd edition, December.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio, June. Association for Computational Linguistics.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592, Boulder, Colorado, June. Association for Computational Linguistics.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 926–931, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hidetsugu Nanba, Noriko Kando, Manabu Okumura, and Of Information Science. 2000. Classification of research papers using citation links and citation types: Towards automatic review article generation.

M. E. J. Newman. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, January.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696, Manchester, UK, August.

Vahed Qazvinian and Dragomir R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 555–564, Uppsala, Sweden, July. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Ozgur. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 895–903, Beijing, China, August. Coling 2010 Organizing Committee.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The acl anthology network corpus. In *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61, Morristown, NJ, USA. Association for Computational Linguistics.

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? attributing scientific work to citations. In *In Proceedings of NAACL/HLT-07*.

Daniel D. K. Sleator and Davy Temperley. 1991. Parsing english with a link grammar. In *In Third International Workshop on Parsing Technologies*.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *In Proc. of EMNLP-06*.

Simone Teufel. 2007. Argumentative zoning for improved citation indexing. computing attitude and affect in text. In *Theory and Applications, pages 159170*.