LAMP-TR-157
CS-TR-4988
UMIACS-TR-2011-14

August 2011

CITATION HANDLING FOR IMPROVED SUMMMARIZATION OF SCIENTIFIC DOCUMENTS

Michael Whidby, David Zajic, Bonnie Dorr

Computational Linguistics and Information Processing
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
mawhidby@umd.edu, dmzajic@umiacs.umd.edu, bonnie@umd.edu

Abstract

In this paper we present the first steps toward improving summarization of scientific documents through citation analysis and parsing. Prior work (Mohammad et al., 2009) argues that citation texts (sentences that cite other papers) play a crucial role in automatic summarization of a topical area, but did not take into account the noise introduced by the citations themselves. We demonstrate that it is possible to improve summarization output through careful handling of these citations. We base our experiments on the application of an improved trimming approach to summarization of citation texts extracted from Question-Answering and Dependency-Parsing documents. We demonstrate that confidence scores from the Stanford NLP Parser (Klein and Manning, 2003) are significantly improved, and that Trimmer (Zajic et al., 2007), a sentence-compression tool, is able to generate higher-quality candidates. Our summarization output is currently used as part of a larger system, Action Science Explorer (ASE) (Gove, 2011).

Keywords: summarization, scientific documents, citation handling

This work was supported, in part, by the National Science Foundation under Grant No. IIS-0705832 (iOPENER: Information Organization for PENning Expositions on esearch) and by the Center for Advanced Study of Language (CASL)

Citation Handling for Improved Summarization of Scientific Documents

Michael Whidby, David Zajic, Bonnie Dorr
Computational Linguistics and Information Processing Lab
University of Maryland Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA
mawhidby@umd.edu, {dmzajic, bonnie}@umiacs.umd.edu

Abstract

In this paper we present the first steps toward improving summarization of scientific documents through citation analysis and parsing. Prior work (Mohammad et al., 2009) argues that citation texts (sentences that cite other papers) play a crucial role in automatic summarization of a topical area, but did not take into account the noise introduced by the citations themselves. We demonstrate that it is possible to improve summarization output through careful handling of these citations. We base our experiments on the application of an improved trimming approach to summarization of citation texts extracted from Question-Answering and Dependency-Parsing documents. We demonstrate that confidence scores from the Stanford NLP Parser (Klein and Manning, 2003) are significantly improved, and that Trimmer (Zajic et al., 2007), a sentence-compression tool, is able to generate higher-quality candidates. Our summarization output is currently used as part of a larger system, Action Science Explorer (ASE) (Gove, 2011).

1 Introduction

It has become increasingly important to support the needs of users who seek to understand a wide range of scientific areas with which they are not currently familiar. For example, it has become common for interdisciplinary review panels to be called upon to review proposals in a wide range of areas, without access to the most up-to-date summaries (or surveys) of the relevant topical areas. NLP and visualization tools have been developed to accommodate this need (Gove et al., 2011) and steps have been taken to provide summaries for the purpose of survey creation, but citations that occur in the input texts introduce noise that leads to disfluent summarization output.

In this paper we present the first steps toward improving summarization of scientific documents through parsing of citation texts (sentences that cite other papers). Prior work (Mohammad et al., 2009) argues that citation texts play a crucial role in automatic summarization of a topical area, but did not take into account the noise introduced by the citations themselves. As a first step toward improving the fluency of summarization of citation texts, we apply two different approaches to citation handling and then examine the effects of these approaches on the parse trees produced by the Stanford Parser (Klein and Manning, 2003), as parsing is an intermediate step on the way to producing summarized output. We demonstrate that the quality of the parser's confidence scores are improved, and better parse trees are produced, with citation handling.

Finally, the improved parse trees serve as the basis of a parse-and-trim approach to summarization of citation texts. As such, we seek to demonstrate that the improved parsing output has a positive effect on Trimmer's (Zajic et al., 2007) sentence candidates for summarization of scientific articles. Our results indicate that the output summaries are significantly more fluent in comparison to those produced by a variant of the summarizer with unhandled citations. Our summarization output is currently used as part of a larger system, Action Science Explorer (ASE) (Gove, 2011).

The next section presents related work. We then present our motivations for introducing citation handling into our system. Following this, we present the tools and data used in our experiments: the Stanford parser (Klein and Manning, 2003), Trimmer (Zajic et al., 2007), our new citation handling techniques, and the ACL Anthology (Joseph and Radev, 2007). Finally, we evaluate the application of citation handling for both parsing and summarization. Our human inspection of the impact of citation handling on parsing indicates that the effect is indeed positive. Summarization is evaluated using both automatic (ROUGE) and human-mediated (nugget-based pyramid) measures. We demonstrate that properly handled citation texts yield more accurate parses and more fluent summaries.

2 Related Work

Previous work has focused on the analysis of citation and collaboration networks (Teufel et al., 2006; Newman, 2001) and scientific article summarization (Teufel and Moens, 2002). Bradshaw (2003) used citation texts to determine the content of articles and improve the results of a search engine. Citation texts have also been used to create summaries of single scientific articles in Qazvinian and Radev (2008) and Mei and Zhai (2008). Nanba and Okumura (1999) discuss citation categorization to support a system for writing a survey and Nanba et al. (2004) automatically categorize citation sentences into three groups using pre-defined phrase-based rules.

Elkiss et al. (2008) conducted several experiments on PubMed Central (PMC) articles and confirmed that the cohesion of a citation text of an article is consistently higher than that of its abstract. Mohammad et al. (2009) also demonstrated the usefulness of citation texts to produce a multi-document survey of scientific articles in comparison to other forms of input such as the abstracts or full texts of the source articles. As such, our experiments below adopt citation texts as input to parsing and summarization.

Our aim is not to determine the utility of citation texts for linguistic processing—as in the prior works cited above—but to determine the impact of proper citation handling within the citation texts for downstream processing. We examine the quality distinctions between the citation-handled input and citation-unhandled input both for parsing and for summarization. For the former, we examine the parser's confidence scores. For the latter, we compare the results to human-generated summaries using both automatic and nugget-based pyramid evaluation (Lin and Demner-Fushman, 2006; Nenkova and Passonneau, 2004; Lin, 2004).

3 Motivation

Citations introduce noise that causes issues in constituency parsers and summarization systems.

3.1 Parser Issues Caused By Citation Texts

Citation texts introduce noise into constituency parsers that may cause erroneous parse trees. Some citation sentences (e.g., "While the restriction to projctive analyses has a number of advantages, there is clear evidence that it cannot be maintained for real-world data *(Zeman, 2004; Nivre, 2006).*") contain citations that are not syntactically part of the sentence, and therefore add nothing in terms of sentence structure. A means for having the parser ignore the citations in these situations would improve the parse trees generated for the citation sentence. Improved parse trees would allow a summarization system to better apply syntactic rules to the citation sentence when generating sentence compressions.

3.2 Summarization Issues Caused by Citation Texts

We currently employ a system that applies syntactic rules to sentences to create sentence compressions for summarization. One syntactic rule that the system uses is a conjunction rule, which specifically creates two compressions from an *and* conjunction with two children. One candidate contains the first child, and the other the second child. Consider an example citing sentence, "The probability model may be either conditional (Duan et al., 2007) or generative (Titov and Henderson, 2007).". The citation "(Titov and Henderson, 2007)" contains a conjunction. When applying the conjunction rule, two sentence candidates are created that now contain erroneous citations:

1. "The probability model may be either conditional (Duan et al., 2007) or generative *(Titov,*

2007)."

2. "The probability model may be either conditional (Duan et al., 2007) or generative *(Henderson, 2007).*"

Note that in this case, the sentence candidates are no different from the source sentence in terms of actual content, but the application of the conjunction rule has made the original citations incorrect. A means for avoiding the application of the conjunction rule on *and* citations are necessary in order to maintain the integrity of the original citation.

4 Data and Methods

4.1 ACL Anthology

The *ACL Anthology* is a collection of papers from the Computational Linguistics journal, and proceedings of ACL conferences and workshops. It has almost $11,000$ papers. To produce the *ACL Anthology Network (AAN)*, Joseph and Radev (2007) manually parsed the references before automatically compiling the network metadata, and generating citation and author collaboration networks. The AAN includes all citation and collaboration data within the ACL papers, with the citation network consisting of $11,773$ nodes and $38,765$ directed edges.

For our evaluation, we used a set of citation texts from papers in the research area of Question Answering (QA) and another set of papers on Dependency parsing (DP). The two sets of papers were compiled by selecting all the papers in AAN that had the words Question Answering and Dependency Parsing, respectively, in the title and the content. There were 10 papers in the QA set and 16 papers in the DP set.

4.2 Trimmer and Stanford Parser

Trimmer is a sentence-compression tool that extends the scope of an extractive summarization system by generating multiple alternative sentence compressions of the most important sentences in target documents (Zajic et al., 2007). Trimmer compressions are generated by applying linguistically-motivated rules to mask syntactic components of a parse of a source sentence. The rules can be applied iteratively to compress sentences below a configurable length threshold, or can be applied in all combinations to generate the full space of compressions.

Trimmer leverages the output of any constituency parser that uses the Penn Treebank conventions. At present, the Stanford Parser (Klein and Manning, 2003) is used. The set of compressions is ranked according to a set of features that may include metadata about the source sentences, details of the compression process that generated the compression, and externally calculated features of the compression.

Summaries are constructed from the highest scoring compressions, using the metadata and maximal marginal relevance (Carbonell and Goldstein, 1998) to avoid redundancy and over-representation of a single source.

4.3 Citation Handling

We now introduce our approach to citation handling, starting first with a description of the two types of citations encountered, and then a presentation of the approach we use for handling them.

4.3.1 Types of Citations

We argue that there are two types of citations in citation sentences: *syntactic* and *non-syntactic*. These two types of citations are used in semantically different ways, and such should be handled in different ways. Syntactic citations are citations that are grammatically part of the sentence; removing them would make the sentence ungrammatical. They typically function as nouns, or agents who did or claimed something. Some examples of syntactic citations include (citations italicized):

- "Moreover, the proof relies on lexico-semantic knowledge available from WordNet as well as rapidly formatted knowledge bases generated by mechanisms described in *(Chaudri et al, 2000).*"

- "Some Q&A systems, like *(Moldovan et al, 2000)* relied both on NE recognizers and some empirical indicators."

- "More details on the memory-based prediction can be found in *Nivre et al (2004)* and *Nivre and Scholz (2004).*"

Non-syntactic citations are citations that are not grammatically part of the sentence; removing them would not have any effect on the grammaticality of the sentence. They are typically used as an instance of some event or situation mentioned in the sentence. Some examples of non-syntactic citations include (citations italicized):

- "If the expected answer types are typical named entities, information extraction engines *(Bikel et al 1999, Srihari and Li 2000)* are used to extract candidate answers."
- "In English as well as in Japanese, dependency analysis has been studied *(Lafferty et al, 1992; Collins, 1996; Eisner, 1996)*."
- "That work extends the maximum spanning tree dependency parsing framework *(McDonald et al, 2005a; McDonald et al, 2005b)* to incorporate features over multiple edges in the dependency graph."

4.3.2 Citation Handling in Trimmer

We have made modifications to Trimmer for handling syntactic and non-syntactic citations. In the syntactic citation case, the entire citation is replaced with placeholder text "*CITATIONX*", where *X* is a unique number assigned to the citation. After all candidates for a sentence have been generated, we can easily place the original citation text back into the sentence. The placeholder text is seen as an out-of-vocabulary noun by the Stanford Parser. This is sensible, since the citation is grammatically part of the sentence and represents a single or multiple entities. Examples of handling syntactic citations:

- *Before:* "Moreover, the proof relies on lexico-semantic knowledge available from WordNet as well as rapidly formatted knowledge bases generated by mechanisms described in *(Chaudri et al, 2000)*."
  After: "Moreover, the proof relies on lexico-semantic knowledge available from WordNet as well as rapidly formatted knowledge bases generated by mechanisms described in *CITATION1*."
- *Before:* "Some Q&A systems, like *(Moldovan et al, 2000)* relied both on NE recognizers and some empirical indicators."
  After: "Some Q&A systems, like *CITATION2* relied both on NE recognizers and some empirical indicators."
- *Before:* "More details on the memory-based prediction can be found in *Nivre et al (2004)* and *Nivre and Scholz (2004)*."
  After: "More details on the memory-based prediction can be found in *CITATION3* and *CITATION4*."

In the non-syntactic citation case, the citation is removed entirely from the sentence. This also makes sense, since the citation in this case is not grammatically part of the sentence. After all sentence compression candidates have been generated, we currently place the citations at the end of the sentence. We leave a better means of replacing non-syntactic citations as future work. Examples of handling non-syntactic citations:

- *Before:* "If the expected answer types are typical named entities, information extraction engines *(Bikel et al 1999, Srihari and Li 2000)* are used to extract candidate answers."
  After: "If the expected answer types are typical named entities, information extraction engines are used to extract candidate answers."
- *Before:* "In English as well as in Japanese, dependency analysis has been studied *(Lafferty et al, 1992; Collins, 1996; Eisner, 1996)*."
  After: "In English as well as in Japanese, dependency analysis has been studied."
- *Before:* "That work extends the maximum spanning tree dependency parsing framework *(McDonald et al, 2005a; McDonald et al, 2005b)* to incorporate features over multiple edges in the dependency graph."
  After: "That work extends the maximum spanning tree dependency parsing framework to incorporate features over multiple edges in the dependency graph."

4.4 Mechanical Turk Tasks

We used Mechanical Turk to clean citation sentences and annotate citations in the DP and QA datasets as being syntactic or non-syntactic. These annotations are used for citation handling in our summarization system. We conducted five different Turk tasks: a pilot study, a study to identify garbage sentences, another study to identify incorrect citation text spans, a study to correct the erroneous citation text spans, and a final study to annotate all citations.

4.4.1 Pilot Study

Before continuing with any other MTurk tasks, we conducted a pilot study to determine whether humans could agree on the citation annotation task. In the citation annotation task, Turkers were presented with a citation sentence, with the citation highlighted. They were then asked to classify the citation as "syntactic", "non-syntactic", or "ambiguous/incorrect citation". The "ambiguous/incorrect" choice was used in case our citation detection was erroneous, or if the Turker was unable determine which category the citation belonged to.

Turkers annotated 50 citations in 50 different randomly selected citation sentences from the citation texts from QA and DP. Four Turkers were allowed to annotate each citation. 9 different Turkers participated in the pilot study, annotating an average of 22.2 citations each. The Krippendorff (Passonneau et al., 2006) agreement score was 0.785578, which we found to be sufficient to continue with the remaining tasks, and sufficient for the main task of annotating all citations in the QA and DP sets.

4.4.2 Identify Garbled Sentences Task

After the pilot study, we had Turkers identify any garbled sentences. We define a garbled sentence as any sentence that contained special symbols/characters from LaTeX (e.g., $\sum$, $\sqrt{x}$, $\prod$), or any other wording or phrasing that wasn't coherent. These sentences cause the Stanford Parser to fail in generating a parse tree, and as such should not be included in the pool of citation sentences. In the task, Turkers were presented with a citation sentence, and asked to label it as "clean" or "garbled/garbage". Again, each sentence was annotated by 3 different Turkers.

We removed a sentence from our system if at least 2 Turkers annotated the sentence as being garbled. 29 different Turkers participated in the task, annotating an average of 50.1 sentences each. Out of the 484 total citation sentences in the QA and DP sets, 52 were garbled/garbage (10.74%). Turkers found this task hardest to agree upon, with a Krippendorff agreement score of 0.468806. We attribute this to the task being more open-ended than some of the other tasks, and perhaps there were not enough examples in quantity or quality provided to help Turkers with the task. In addition, it could also be due to the confusing content and style of ACL papers for a non-specialist reader. However, this annotation task was used as a filter to ensure we studied sentences in which the interference was caused by citations, and not due to other features of the AAN sentences (or sentences taken from LaTeX papers). Despite the low agreement score, we were liberal in accepting what Turkers labeled as garbled, because we wanted to be safe in excluding those sentences.

4.4.3 Identify Incorrect Citation Text Spans Task

We also had Turkers identify incorrect citation text spans that our algorithms may have mislabeled or missed entirely. In this task, Turkers were presented with a citation sentence, with a possible citation highlighted. They were then asked to identify whether or not the highlighted citation was a correct citation text span. Several examples of correct and incorrect citation text spans were provided for the Turkers to reference. Again, each citation text span was annotated by 3 different Turkers.

A citation was labelled incorrect if at least 2 Turkers annotated the citation text span as being incorrect. 30 different Turkers participated in the task, annotating an average of 69 citations each. Of the 690 citations from non-garbled sentences, 429 were labelled as correct, and 261 as incorrect 37.8%. The majority of these incorrect citations were of the form "name (date)", e.g. "Slughorn (1957)". Turkers were easily able to agree in this task, with a Krippendorff agreement score of 0.924808.

4.4.4 Correct Erroneous Citation Text Spans Task

With the incorrect citation text spans identified, we then created a task for Turkers to fix the text spans. In this task, Turkers were presented with the citation sentence, and the incorrect citation text span highlighted. They were then asked to copy and paste what they believed to be the correct citation text span. For this task, we had 2 Turkers annotate each incorrect citation text span. If the Turkers were not in agreement, then we had another Turker annotate the text span as a tie-breaker.

In this task, Turkers agreed on the correct citation text spans; however, they did not format the citations the same way, so it was difficult to run metrics on the results. For example, one Turker might label a citation text span as *Johnson (2008)*, whereas another labeled it as *"Johnson (2008)"*. In other instances, instead of copy/pasting the text from the source citation sentence, some Turkers typed in their answers and made either typographical errors or formatted the citation in a different way from the source sentence (e.g., "(Johnson, 2008)" versus "(Johnson 2008)"). These sorts of errors can be expected when using an open-ended text input answer format.

4.4.5 Annotate Citations Task

The final Turk task we conducted was similar to the pilot study, but using the entire set of citation sentences from DP and QA that were identified as being clean sentences from the *Identify Garbled/Garbage Sentences Task.* With all erroneous citation text spans corrected, and garbled sentences identified, we presented Turkers with a citation sentence, with the citation text span highlighted. The Turkers were then asked to classify the citation as "syntactic" or "non-syntactic". Each citation was annotated by 3 different Turkers.

A citation was labelled as "syntactic" or "non-syntactic" if at least 2 Turkers agreed on a labeling. In the task, 30 different Turkers participated, annotating an average of 69 citations each. Out of the 690 citations from the non-garbled sentences, 370 were labeled as "non-syntactic" (53.62%), and 320 were labeled as "syntactic" (46.38%). Similar to our pilot study, the Krippendorff agreement score was 0.752202.

5 Experiments and Results

Our evaluation experiments are on a set of papers in the research area of Question Answering (QA) and another set of papers on Dependency parsing (DP). The two sets of papers were compiled by selecting all the papers in AAN that had the words *Question Answering* and *Dependency Parsing*, respectively, in the title and the content. There were 10 papers in the QA set and 16 papers in the DP set. We also compiled the citation texts for the 10 QA papers and the citation texts for the 16 DP papers.

We automatically parsed and generated summaries for both QA and DP from the citation texts corresponding to the QA and DP papers. We generated 2 parse outputs and 2 corresponding summaries, each of length 250 words, by applying Trimmer to citation texts for both QA and DP, using two different methods of citation handling (citation handling and no citation handling). We created two additional 250-word summaries by randomly choosing sentences from the citation texts of QA and DP. We will refer to them as *random summaries*.

Our goal was to determine the impact of proper citation handling on both parsing and summarization, as described below.

5.1 Evaluation of Parser Confidence Scores on Citation Sentences

We evaluated the confidence scores of the Stanford Parser in parsing citation sentences with and without citation handling. Figure 1 shows the distribution of the confidence scores of the citation handling and non-citation handling cases. We observe that the data appeared to be normal and bimodal, and with a set of outliers that were much lower in scores. We set threshold of -750, in which a score below this threshold was considered an outlier. In the non-citation handling case 1.17% of the scores were outliers and 2.8% of the scores were outliers in the citation handling case. We ran a Chi-squared test with Yates' continuity correction and found that there was not a significant difference in the number of outliers between the conditions.

In our t-test we only included sentences whose scores were above the threshold in both cases. The number of sentences where neither condition produced an outlier was 412 (96.26%). We ran a paired T-test on the sentences in which neither condition produced an outlier and found citation handling to have a significant effect, with $t = 10.254, df = 411, p < 0.01$.

5.2 Evaluation of Trimmer Output

We also evaluated the quality of the sentence candidates output by Trimmer by running the sentence candidates back through the Stanford Parser, and ex-

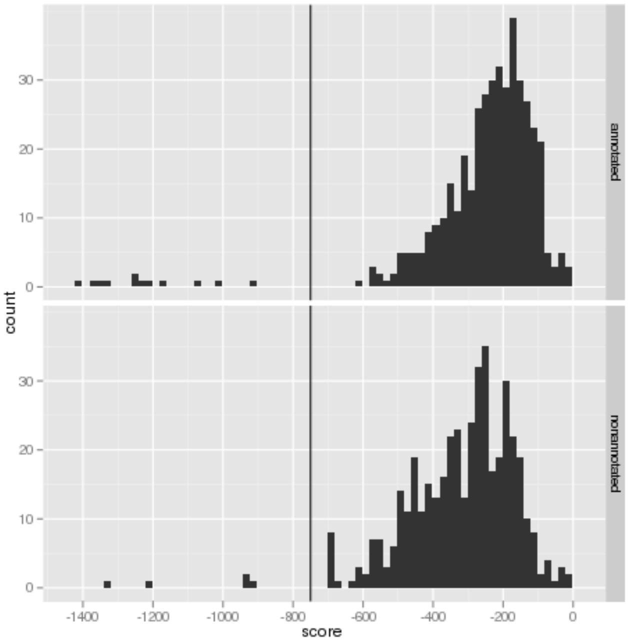

Figure 1: Distribution of Stanford Parser confidence scores on citation texts. "annotated" denotes scores with citation handling, "nonannotated" denotes scores without citation handling

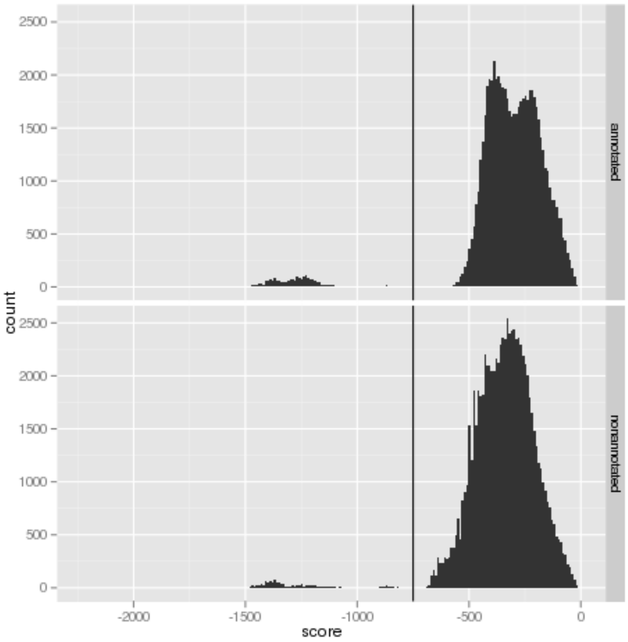

Figure 2: Distribution of Stanford Parser confidence scores on Trimmer output candidates. "annotated" denotes scores of sentences with citation handling, "nonannotated" denotes scores of sentences without citation handling

amining the confidence scores. Figure 2 shows the distribution of the confidence scores of the citation handling and non-citation handling cases, respectively, with bin size of 200. The data appeared to be normal and bimodal, with a set of outliers again that score much lower than average. We used the same threshold score of -750 as before. In this data set, the non-citation handling condition had 1.43% outliers, while citation handling had 3.28%. We found these differences in the percentage of outliers to be significant in a Chi-squared test, but the number of outliers was small enough to continue with the analysis. We used a Welch Two Sample t-test because Trimmer generates different sets of compressed sentences for the citation handling and non-citation handling cases. We only included sentences whose scores were above the threshold, $62,836$ sentences for the citation handling case, and $79,594$ sentences for the non-citation handling case. We found citation handling to have a significant effect, with $t = 72.8097, df = 140018.5, p < 0.001$.

5.3 Evaluation of Summarization Output

We evaluated each of the automatically generated summaries using two separate approaches: nugget-based pyramid evaluation and ROUGE (described in the two subsections below).

Gold standard data was manually created from the QA and DP citation texts using three techniques:[1] (1) We asked two impartial judges to identify important nuggets of information worth including in a summary; (2) We asked four fluent speakers of English to create 250-word summaries of the datasets. Then we determined how well Trimmer performed both with and without proper citation handling against these gold standards.

5.3.1 Nugget-Based Pyramid Evaluation

For our first approach we used a nugget-based evaluation methodology (Lin and Demner-Fushman, 2006; Nenkova and Passonneau, 2004; Hildebrandt et al., 2004; Voorhees, 2003). We asked three impar-

[1] Creating gold standard data from complete papers is fairly arduous, and was not pursued.

tial annotators (knowledgeable in NLP but not affiliated with the project) to review the citation texts and/or abstract sets for each of the papers in the QA and DP sets and manually extract prioritized lists of 2–8 "nuggets," or main contributions, supplied by each paper. Each nugget was assigned a weight based on the frequency with which it was listed by annotators as well as the priority it was assigned in each case. Our automatically generated summaries were then scored based on the number and weight of the nuggets that they covered. This evaluation approach is similar to the one adopted by Qazvinian and Radev (2008), but adapted here for use in the multi-document case.

The annotators were instructed to extract nuggets for each of the 10 QA and 16 DP papers, based only on the citation texts for those papers. We obtained a weight for each nugget by reversing its priority out of 8 (e.g., a nugget listed with priority 1 was assigned a weight of 8) and summing the weights over each listing of that nugget.[2]

To evaluate a given summary, we counted the number and weight of nuggets that it covered. Nuggets were detected via the combined use of annotator-provided regular expressions and careful human review. Recall was calculated by dividing the combined weight of covered nuggets by the combined weight of all nuggets in the nugget set. Precision was calculated by dividing the number of distinct nuggets covered in a summary by the number of sentences constituting that summary, with a cap of 1. F-measure, the weighted harmonic mean of precision and recall, was calculated with a beta value of 3 in order to assign the greatest weight to recall. Recall is favored because it rewards summaries that include highly weighted (important) facts, rather than just a great number of facts.

Table 1 gives the F-measure values of the 250-word summaries manually generated by humans[3]. The summaries were evaluated using the nuggets drawn from the QA citation texts, QA abstracts, and DP citation texts. The average of their scores (listed in the rightmost column) may be considered a good score to aim for by the automatic summarization methods.

Human Performance: Pyramid F-measure					
Input	Hum1	Hum2	Hum3	Hum4	Avg
QA	0.350	0.458	0.403	0.577	0.447
DP	0.179	0.467	0.362	0.513	0.380

Table 1: Pyramid F-measure scores of human-created summaries of QA and DP data.

System Performance: Pyramid F-measure			
Input	Random	Trimmer1	Trimmer2
QA	0.321	0.442	0.410
DP	0.219	0.241	0.298

Table 2: Pyramid F-measure scores of automatic summaries of QA and DP data. The summaries are evaluated using nuggets drawn from QA and DB citation texts. Trimmer1 is the original Trimmer1 system without citation handling; Trimmer2 is the version of Trimmer with citation handling.

Table 2 gives the F-measure values of the surveys generated by the random summarizer and three variants of automatic summarizers, evaluated using nuggets drawn from the QA and DP citation texts. Among the various automatic summarizers, neither Trimmer1 or Trimmer2 performed significantly better than the other at this task.

5.3.2 ROUGE evaluation

Table 3 presents ROUGE scores (Lin, 2004) of each of human-generated 250-word surveys against each other. The average (last column) is what the automatic surveys can aim for. We then evaluated each of the random surveys and those generated by the three variants of Trimmer against the references. Table 4 lists ROUGE scores of surveys when the manually created 250-word survey of the QA and DP citation texts were used as gold standard. Among the automatic summarizers, Trimmer2, our version of Trimmer with citation handling, performs best.

6 Conclusion

In this paper, we investigated the impact and effectiveness of citation handling for parsing and summarization of citation texts (sentences that cite other papers). We parsed and summarized a set of Question

[2] Results obtained with other weighting schemes that ignored priority ratings and multiple mentions of a nugget by a single annotator showed the same trends as the ones shown by the selected weighting scheme, but the latter was a stronger distinguisher among the four systems.

[3] These numbers do not match those reported in (Mohammad et al., 2009). We are using a different weighting scheme in our pyramid evaluations.

Human Performance: ROUGE-2					
Input	Hum1	Hum2	Hum3	Hum4	Avg
QA	0.1807	0.1956	0.0756	0.2019	0.1635
DP	0.1550	0.1259	0.1200	0.1654	0.1416

Table 3: ROUGE-2 scores of human-created summaries of QA and DP data. ROUGE-1 and ROUGE-L followed similar patterns.

System Performance: ROUGE-2			
Input	Random	Trimmer1	Trimmer2
QA	0.116	0.169	0.173
DP	0.107	0.101	0.139

Table 4: Pyramid F-measure scores of automatic summaries of QA and DP data. The summaries are evaluated using nuggets drawn from QA and DB citation texts. Trimmer1 is the original Trimmer1 system without citation handling; Trimmer2 is the version of Trimmer with citation handling.

Answering (QA) and Dependency Parsing (DP) citation texts both with and without citation handling. We then evaluated the parse output and also applied two separate summarization-evaluations to determine the degree of effectiveness of citation handling. The results indicate the importance of proper citation handling prior to parsing and summarization of citation texts.

In the future, we would like to implement a better means of inserting non-syntactic citations back in to the sentence candidates. Currently, the citations are appended to the end of the sentence rather than in their original location in the sentence. In addition, we would like to examine the outliers in the confidence scores for the Parser and determine what features of citations may be causing these catastrophic errors with the Parser. We would also like to carry out additional Turk tasks to determine the effectiveness of citation handling in generating summaries. These tasks would involve Turkers rating various characteristics of sentence candidates, such as fluency. We would create tasks for sentence candidates that used citation handling, ones that did not use citation handling, and sentences generated using bag of words. Finally, we would also like to develop a system that automatically determines whether a citation is syntactic or non-syntactic, as currently we have used Turkers to annotate our work.

Acknowledgments

This work was supported, in part, by the National Science Foundation under Grant No. IIS-0705832 (iOPENER: Information Organization for PENning Expositions on Research) and by the Center for Advanced Study of Language (CASL). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. We would like thank Melissa Egan, Jacob Devlin, Saif Mohammad, Dragomir Radev, Ahmed Hassan, Pradeep Muthukirshan, Vahed Qazvinian, and Scott Jackson for provision of data and valuable suggestions.

References

Shannon Bradshaw. 2003. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir R. Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Robert Gove, Cody Dunne, Ben Shneiderman, Judith Klavans, and Bonnie Dorr. 2011. Evaluating visual and statistical exploring of scientific literature networks. In *VL/HCC'11*.

Robert Gove. 2011. Understanding scientific literature networks: Case study evaluations of integrating visualizations and statistics. Master's thesis, University of Maryland, College Park, College Park, MD.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Overview of the trec 2003 question-answering track. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*.

Mark Joseph and Dragomir Radev. 2007. Citation analysis, centrality, and the ACL Anthology. Technical Report CSE-TR-535-07, University of Michigan. Dept. of Electrical Engineering and Computer Science.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of ACL*, pages 423–430.

Jimmy J. Lin and Dina Demner-Fushman. 2006. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop on Text Summarization Branches Out*.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL '08*, pages 816–824.

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 NAACL-HLT 2009*.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *IJCAI1999*, pages 926–931.

Hidetsugu Nanba, Takeshi Abekawa, Manabu Okumura, and Suguru Saito. 2004. Bilingual presri: Integration of multiple research paper databases. In *Proceedings of RIAO 2004*, pages 195–211, Avignon, France.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *Proceedings of the HLT-NAACL conference*.

Mark E. J. Newman. 2001. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409.

Rebecca Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *In Proceedings of LREC*.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *COLING 2008*, Manchester, UK.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of EMNLP*, pages 103–110, Australia.

Ellen M. Voorhees. 2003. Overview of the trec 2003 question answering track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*.

David M. Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management (Special Issue on Summarization)*.