Generating surveys of scientific paradigms

Saif Mohammad[†*], Bonnie Dorr[†‡*], Melissa Egan[†‡], Ahmed Hassan[ϕ], Pradeep Muthukrishan[ϕ], Vahed Qazvinian[ϕ], Dragomir Radev[ϕ§], David Zajic[†‡]

Laboratory for Computational Linguistics and Information Processing
Institute for Advanced Computer Studies[†] and Computer Science[‡], University of Maryland
JHU Human Language Technology Center of Excellence[*]
{saif,bonnie,mkegan,dmzajic}@umiacs.umd.edu

Department of Electrical Engineering and Computer Science[ϕ]
Department of Linguistics[§], University of Michigan
{hassanam,mpradeep,vahed,radev}@umich.edu

Abstract

The number of research publications in various disciplines is growing exponentially. Researchers and scientists are increasingly finding themselves in the position of having to quickly understand large amounts of technical material. In this paper we present the first steps in producing an automatically generated, readily consumable, technical survey. Specifically we explore the combination of citation information and summarization techniques. Even though prior work (Teufel et al., 2006) argues that citation text is unsuitable for summarization, we show that in the framework of multi-document survey creation, citation texts can play a crucial role.

1 Introduction

In today's rapidly expanding disciplines, scientists and scholars are constantly faced with the daunting task of keeping up with knowledge in their field. In addition, the increasingly interconnected nature of real-world tasks often requires experts in one discipline to rapidly learn about other areas in a short amount of time.

Cross-disciplinary research requires scientists in areas such as linguistics, biology, and sociology to learn about computational approaches and applications, e.g., computational linguistics, biological modeling, social networks. Authors of journal articles and books must write accurate surveys of previous work, ranging from short summaries of related research to in-depth historical notes.

Interdisciplinary review panels are often called upon to review proposals in a wide range of areas, some of which may be unfamiliar to panelists. Thus, they must learn about a new discipline "on the fly" in order to relate their own expertise to the proposal.

Our goal is to effectively serve these needs by combining two currently available technologies: (1) bibliometric lexical link mining that exploits the structure of citations and relations among citations; and (2) summarization techniques that exploit the content of the material in both the citing and cited papers.

It is generally agreed upon that manually written abstracts are good summaries of individual papers. More recently, Qazvinian and Radev (2008) argue that **citation text** is useful in creating a summary of the important contributions of a research paper. The citation text of a target paper is the set of sentences in other technical papers that explicitly refer to it Elkiss et al. (2008a). However, Teufel (2005) argues that using citation text directly is not suitable for document summarization.

In this paper, we compare and contrast the usefulness of abstracts and of citation text in automatically generating a technical survey on a given topic from multiple research papers. The next section provides the background for this work, including the primary features of a technical survey and also the types of input that are used in our study (full papers, abstracts, and citation texts). Following this, we describe related work and point out the advances of our work over previous work. We then describe how citation texts are used as a new input for multi-document summarization to produce surveys of a given technical area. We apply four different summarization techniques to data in the ACL Anthol-

ogy and evaluate our results using both automatic (ROUGE) and human-mediated (nugget-based pyramid) measures. We observe that, as expected, abstracts are useful in survey creation, but, notably, we also conclude that citation texts have crucial survey-worthy information not present in (or at least, not easily extractable from) abstracts. We further discover that abstracts are author-biased and thus complementary to the broader perspective inherent in citation texts; these differences enable the use of a range of different levels and types of information in the survey—the extent of which is subject to survey length restrictions (if any).

2 Background

Automatically creating technical surveys is significantly distinct from that of traditional multi-document summarization. Below we describe primary characteristics of a technical survey and we present three types of input texts that we used for the production of surveys.

2.1 Technical Survey

In the case of multi-document summarization, the goal is to produce a readable presentation of multiple documents, whereas in the case of technical survey creation, the goal is to convey the key features of a particular field, basic underpinnings of the field, early and late developments, important contributions and findings, contradicting positions that may reverse trends or start new sub-fields, and basic definitions and examples that enable rapid understanding of a field by non-experts.

A prototypical example of a technical survey is that of "chapter notes," i.e., short (50–500 word) descriptions of sub-areas found at the end of chapters of textbook, such as Jurafsky and Martin (2008). One might imagine producing such descriptions automatically, then hand-editing them and refining them for use in an actual textbook.

We conducted a human analysis of these chapter notes that revealed a set of conventions, an outline of which is provided here (with example sentences in italics):

1. Introductory/opening statement: *The earliest computational use of X was in Y, considered by many to be the foundational work in this area.*
2. Definitional follow up: *X is defined as Y.*
3. Elaboration of definition (e.g., with an example): *Most early algorithms were based on Z.*
4. Deeper elaboration, e.g., pointing out issues with initial approaches: *Unfortunately, this model seems to be wrong.*
5. Contrasting definition: *Most algorithms since then...*
6. Introduction of additional specific instances / historical background with citations: *Two classic approaches are described in Q.*
7. References to other summarization work: *R provides a comprehensive guide to the details behind X.*

The notion of *text level categories* or *zoning* of technical papers—related to the survey components enumerated above—has been investigated previously in the work of Nanba and Kan (2004b) and Teufel (2002). These earlier works focused on the *analysis* of scientific papers based on their rhetorical structure and on determining the portions of papers that contain new results, comparisons to earlier work, etc. The work described in this paper focuses on the *synthesis* of technical surveys based on knowledge gleaned from rhetorical structure not unlike that of the work of these earlier researchers, but perhaps guided by structural patterns along the lines of the conventions listed above.

Although our current approach to survey creation does not yet incorporate a fully pattern-based component, our ultimate objective is to apply these patterns to guide the creation and refinement of the final output. As a first step toward this goal, we use citation texts (closest in structure to the patterns identified by convention 7 above) to pick out the most important content for survey creation.

2.2 Full papers, abstracts, and citation texts

Published research on a particular topic can be summarized from two different kinds of sources: (1) where an author describes her own work and (2) where others describe an author's work (usually in relation to their own work). The author's description of her own work can be found in her paper. How

others perceive her work is spread across other papers that cite her work. We will refer to the set of sentences that explicitly mention a target paper Y as the citation text of Y.

Traditionally, technical survey generation has been tackled by summarizing a set of research papers pertaining to the topic. However, individual research papers usually come with manually-created "summaries"—their abstracts. The abstract of a paper may have sentences that set the context, state the problem statement, mention how the problem is approached, and the bottom-line results—all in 200 to 500 words. Thus, using only the abstracts (instead of full papers) as input to a summarization system is worth exploring.

Whereas the abstract of a paper presents what the authors think to be the important contributions of a paper, the citation text of a paper captures what others in the field perceive as the contributions of the paper. The two perspectives are expected to have some overlap in their content, but the citation text also contains additional information not found in abstracts (Elkiss et al., 2008a). For example, how a particular methodology (described in one paper) was combined with another (described in a different paper) to overcome some of the drawbacks of each. A citation text is also an indicator of what contributions described in a paper were more influential over time. Another distinguishing feature of citation texts in contrast to abstracts is that a citation text tends to have a certain amount of redundant information. This is because multiple papers may describe the same contributions of a target paper. This redundancy can be exploited to determine the important contributions of the target paper.

Our goal is to test the hypothesis that an effective technical survey will reflect information on research not only from the perspective of its authors but also from the perspective of others who use/commend/discredit/add to it. Before describing our experiments with technical papers, abstracts, and citation texts, we first summarize relevant prior work that used these sources of information as input.

3 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) has 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; Mei and Zhai, 2008). However, to the knowledge of the authors, there is no previous work that uses the text of the citation texts to produce a multi-document survey of scientific articles. Furthermore, there has been no study contrasting the quality of surveys generated from citation summaries—both automatically and manually produced—to surveys generated from other forms of input such as the abstracts or full texts of the source articles.

Nanba and Okumura (1999) discuss citation categorization to support a system for writing a survey. Nanba et al. (2004a) automatically categorize citation sentences into three groups using pre-defined phrase-based rules. Based on this categorization a survey generation tool is introduced in Nanba and Kan (2004b). Nanba et al. (2004b) report that co-citation implies similarity by showing that the textual similarity of co-cited papers is proportional to the proximity of their citations in the citing article.

Elkiss et al. (2008b) conducted several experiments on a set of 2,497 articles from the free PubMed Central (PMC) repository.[1] Results from this experiment confirmed that the cohesion of a citation text of an article is consistently higher than the that of its abstract. Elkiss et al. (2008b) also concluded that citation texts contain additional information are more focused than abstracts.

Kan et al. (2002) use annotated bibliographies to cover certain aspects of summarization and suggest using metadata and critical document features as well as the prominent content-based features to summarize documents. Kupiec et al. (1995) use a statistical method and show how extracts can be used to create a summaries but use no annotated metadata in summarization. Kan et al. (2002) use annotated bibliographies for summarization and suggest that summaries should also include metadata and critical document features as well as the prominent content-based features.

Siddharthan and Teufel (2007) describe a new ref-

[1] http://www.pubmedcentral.gov

erence task and show high human agreement as well as an improvement in the performance of **argumentative zoning** (Teufel, 2005). In argumentative zoning, which is a rhetorical classification task, seven classes (Own, Other, Background, Textual, Aim, Basis, and Contrast) are used to label sentences according to their role in the author's argument.

Our aim is not only to determine the utility of citation texts for survey creation, but also to examine the quality distinctions between this form of input and other forms of input, including abstracts and full texts—comparing the results to human-generated surveys using both automatic and nugget-based pyramid evaluation (Lin and Demner-Fushman, 2006; Nenkova and Passonneau, 2004; Lin, 2004).

4 Summarization systems

We used four different summarization systems for our survey-creation approach: **Trimmer, LexRank, C-LexRank,** and **C-RR**. Trimmer is a syntactically-motivated parse-and-trim approach. LexRank is a graph-based similarity approach. C-LexRank and C-RR both use graph clustering (the 'C' in their name stands for clustering). We describe each of these, in turn, below.

4.1 Trimmer

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 can leverage 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.

4.2 LexRank

We also used LexRank (Erkan and Radev, 2004), a state-of-the-art multidocument summarization system, to generate summaries. LexRank first builds a graph of all the candidate sentences. Two candidate sentences are connected with an edge if the similarity between them is above a threshold. We used cosine as the similarity metric with a threshold of 0.15. Once the network is built, the system finds the most central sentences by performing a random walk on the graph. The salience of a node is recursively defined on the salience of adjacent nodes. This is similar to the concept of prestige in social networks, where the prestige of a person is dependent on the prestige of the people he/she knows. However, since random walk may get caught in cycles or in disconnected components, we reserve a low probability to jump to random nodes instead of neighbors (a technique suggested by Langville and Meyer (2006)). Note also that unlike the original PageRank method, the graph of sentences is undirected. This updated measure of sentence salience is called as LexRank. The sentences with the highest LexRank scores form the summary.

4.3 Clustering Summarizers: C-LexRank and C-RR

To create summaries, we also use two clustering methods that are proposed in Qazvinian and Radev (2008), which are called C-RR and C-LexRank. They both create a full connected network in which nodes are sentences and edges are cosine similarities. Next, a cutoff value of 0.1 is applied to prune the graph and make a binary network. Then the largest connected component of the network is extracted and clustered. Both of the mentioned summarizers cluster the network similarly but use different approaches to select sentences from different communities. In C-RR sentences are picked from different clusters in a round robin (RR) fashion. C-LexRank first calculates LexRank within each cluster to find the most salient sentences of each community. Then it picks the most salient sentence of

Most of work in QA and paraphrasing focused on folding paraphrasing knowledge into question analyzer or answer locator Rinaldi et al, 2003; Tomuro, 2003. In addition, number of researchers have built systems to take reading comprehension examinations designed to evaluate children's reading levels Charniak et al, 2000; Hirschman et al, 1999; Ng et al, 2000; Riloff and Thelen, 2000; Wang et al, 2000. so-called " definition " or " other " questions at recent TREC evalua - tions Voorhees, 2005 serve as good examples. To better facilitate user information needs, recent trends in QA research have shifted towards complex, context-based, and interactive question answering Voorhees, 2001; Small et al, 2003; Harabagiu et al, 2005. [And so on.]

Table 1: First few sentences of the QA citation texts survey generated by Trimmer.

each cluster, and then the second most salient and so forth until the summary length limit is reached.

5 Data

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 the purpose of evaluation, we chose create technical surveys from a set of papers in the research area of Question Answering (QA) and another set of technical surveys from a 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 *D*ependency 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.

6 Experiments

We automatically generated surveys for both QA and DP from three different types of documents: (1) full papers from the QA and DP sets—**QA and DP full papers (PA)**, (2) only the abstracts of the QA and DP papers—**QA and DP abstracts (AB)**, and (3) the citation texts corresponding to the QA and DP papers—**QA and DP citations texts (CT)**.

We generated twenty four (4x3x2) surveys, each of length 250 words, by applying Trimmer, LexRank, C-LexRank and C-RR on the three data types (citation texts, abstracts, and full papers) for both QA and DP. (Table 1 shows a fragment of one of the surveys automatically generated from QA citation texts.) We created six (3x2) additional 250-word surveys by randomly choosing sentences from the citation texts, abstracts, and full papers of QA and DP. We will refer to them as **random surveys**.

6.1 Evaluation

Our goal was to determine if citation texts do indeed have useful information that one will want to put in a survey and if so, how much of this information is NOT available in the original papers and their abstracts. For this we evaluated each of the automatically generated surveys using two separate approaches: nugget-based pyramid evaluation and ROUGE (described in the two subsections below). Two sets of gold standard data were manually created from the QA and DP citation texts and the abstracts, respectively:[2] (1) We asked two impartial judges to identify important nuggets of information worth including in a survey. (2) We asked four fluent speakers of English to create 250-word surveys of the datasets. Then we determined how well the different automatically generated surveys perform against these gold standards. If the citation texts essentially have only redundant information with respect to the abstracts and original papers, then the surveys of citation texts will not perform better than others.

6.1.1 Nugget-Based Pyramid Evaluation

For our first approach we use a nugget-based evaluation methodology (Lin and Demner-Fushman, 2006; Nenkova and Passonneau, 2004; Hildebrandt

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

et al., 2004; Voorhees, 2003). We asked three impartial 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 surveys 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 had two distinct tasks for the QA set, and one for the DP set: (1) extract nuggets for each of the 10 QA papers, based only on the citation texts for those papers; (2) extract nuggets for each of the 10 QA papers, based only on the abstracts of those papers; and (3) extract nuggets for each of the 16 DP papers, based only on the citation texts for those papers.[3]

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.[4]

To evaluate a given survey, 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 survey by the number of sentences constituting that survey, 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 surveys that include highly weighted (important) facts, rather than just a great number of facts.

[3]We first experimented using only the QA set. Then to show that the results apply to other datasets, we asked human annotators for gold standard data on the DP citation texts. Additional experiments on DP abstracts were not pursued because this would have required additional human annotation effort to establish a point we had already made with the QA set, i.e., that abstracts are useful for survey creation.

[4]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 for the selected weighting scheme, but the latter was a stronger distinguisher among the four systems."

	Human Performance: Pyramid F-measure				
	Human1	Human2	Human3	Human4	Average
Input: QA citation surveys					
QA–CT nuggets	0.524	0.711	0.468	0.695	0.599
QA–AB nuggets	0.495	0.606	0.423	0.608	0.533
Input: QA abstract surveys					
QA–CT nuggets	0.542	0.675	0.581	0.669	0.617
QA–AB nuggets	0.646	0.841	0.673	0.790	0.738
Input: DP citation surveys					
DP–CT nuggets	0.245	0.475	0.378	0.555	0.413

Table 2: Pyramid F-measure scores of human-created surveys of QA and DP data. The surveys are evaluated using nuggets drawn from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT).

Table 2 gives the F-measure values of the 250-word surveys manually generated by humans. The surveys 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.

Table 3 gives the F-measure values of the surveys generated by the four automatic summarizers, evaluated using nuggets drawn from the QA citation texts, QA abstracts, and DP citation texts. The table also includes results for the baseline random summaries.

When we used the nuggets from the abstracts set for evaluation, the surveys created from abstracts scored higher than the corresponding surveys created from citation texts and papers. Further, the best surveys generated from citation texts outscored the best surveys generated from papers. **When we used the nuggets from citation sets for evaluation,** the best automatic surveys generated from citation texts outperform those generated from abstracts and full papers. Taken as a whole, these pyramid results demonstrate that citation texts can contain useful information that is not available in the abstracts or the original papers, and that abstracts can contain useful information that is not available in the citation texts or full papers.

Among the various automatic summarizers, Trimmer performed best at this task, in two cases ex-

	System Performance: Pyramid F-measure				
	Random	C-LexRank	C-RR	LexRank	Trimmer
Input: QA citation surveys					
QA–CT nuggets	0.321	0.434	0.268	0.295	**0.616**
QA–AB nuggets	0.305	0.388	0.349	0.320	**0.543**
Input: QA abstract surveys					
QA–CT nuggets	0.452	0.383	**0.480**	0.441	0.404
QA–AB nuggets	**0.623**	0.484	0.574	0.606	0.622
Input: QA full paper surveys					
QA–CT nuggets	0.239	**0.446**	0.299	0.190	0.199
QA–AB nuggets	0.294	**0.520**	0.387	0.301	0.290
Input: DP citation surveys					
DP–CT nuggets	0.219	0.231	0.170	**0.372**	0.136
Input: DP abstract surveys					
DP–CT nuggets	**0.321**	0.301	0.263	0.311	0.312
Input: DP full paper surveys					
DP–CT nuggets	0.032	0.000	0.144	*	**0.280**

Table 3: Pyramid F-measure scores of automatic surveys of QA and DP data. The surveys are evaluated using nuggets drawn from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT).
* LexRank is computationally intensive and so was not run on the DP-PA dataset (about 4000 sentences).

	Human Performance: ROUGE-2				
	human1	human2	human3	human4	average
Input: QA citation surveys					
QA–CT refs.	0.1807	0.1956	0.0756	0.2019	0.1635
QA–AB refs.	0.1116	0.1399	0.0711	0.1576	0.1201
Input: QA abstract surveys					
QA–CT refs.	0.1315	0.1104	0.1216	0.1151	0.1197
QA-AB refs.	0.2648	0.1977	0.1802	0.2544	0.2243
Input: DP citation surveys					
DP–CT refs.	0.1550	0.1259	0.1200	0.1654	0.1416

Table 4: ROUGE-2 scores obtained for each of the manually created surveys by using the other three as reference. Results for ROUGE-1 and ROUGE-L followed similar patterns.

ceeding the average human performance. Note also that the random summarizer outscored the automatic summarizers in cases where the nuggets were taken from a source different from that used to generate the survey. However, one or two summarizers still tended to do well. This indicates a difficulty in extracting the overlapping survey-worthy information across the two sources.

6.1.2 ROUGE evaluation

Table 4 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 four summarization systems against the references. Table 5 lists the ROUGE scores of various surveys when the manually created 250-word survey of the QA citation texts, survey of the QA abstracts, and the survey of the DP citation texts, are used as gold standard.

When we use manually created citation text surveys as reference, then the surveys generated from citation texts obtained significantly better ROUGE scores than the surveys generated from abstracts and full papers ($p < 0.05$) [RESULT 1]. This shows that crucial survey-worthy information present in citation texts is not available, or hard to extract, from abstracts and papers alone. Further, the surveys generated from abstracts performed significantly better than those generated from the full papers ($p < 0.05$) [RESULT 2]. This shows that abstracts and citation texts are generally denser in survey worthy information than full papers.

When we use manually created abstract surveys as reference, then the surveys generated from abstracts obtained significantly better ROUGE scores than the surveys generated from citation texts and full papers ($p < 0.05$) [RESULT 3]. Further, and more importantly, the surveys generated from citation texts performed significantly better than those generated from the full papers ($p < 0.05$) [RESULT 4]. Again, this shows that abstracts and citation texts are richer in survey-worthy information. These results also show that abstracts of papers and citation

	System Performance: ROUGE-2				
	Random	C-LexRank	C-RR	LexRank	Trimmer
Input: QA citation surveys					
QA–CT refs.	0.11561	**0.17013**	0.09522	0.13501	0.16984
QA–AB refs.	0.08264	**0.11653**	0.07600	0.07013	0.10336
Input: QA abstract surveys					
QA–CT refs.	0.04516	0.05892	**0.06149**	0.05369	0.04114
QA–AB refs.	0.12085	0.13634	0.12190	**0.20311**	0.13357
Input: QA full paper surveys					
QA–CT refs.	0.03042	0.03606	0.03599	**0.28244**	0.03986
QA–AB refs.	0.04621	0.05901	0.04976	**0.10540**	0.07505
Input: DP citation surveys					
DP–CT refs.	0.10690	**0.13164**	0.08748	0.04901	0.10052
Input: DP abstract surveys					
DP–CT refs.	0.07027	0.07321	0.05318	**0.20311**	0.07176
Input: DP full paper surveys					
DP–CT refs.	0.03770	0.02511	0.03433	*	**0.04554**

Table 5: ROUGE-2 scores of automatic surveys of QA and DP data. The surveys are evaluated by using human references created from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT). These results are obtained after Jack-knifing the human references so that the values can be compared to those in Table 4.
* LexRank is computationally intensive and so was not run on the DP full papers set (about 4000 sentences).

texts have some overlapping information (RESULT 2 and RESULT 4), but they also have a significant amount of unique survey-worthy information (RESULT 1 and RESULT 3).

Among the various automatic summarizers, C-LexRank and LexRank perform best. This is unlike the results found through the nugget-evaluation method, where Trimmer performed best. This suggests that Trimmer is better at identifying more useful nuggets of information, but C-LexRank and LexRank are better are producing unigrams and bigrams expected in a survey. To some extent this may be due to the fact that Trimmer uses smaller (trimmed) fragments of sources sentence in its summaries.

7 Conclusion

In this paper, we investigated the usefulness of directly summarizing citation texts (sentences that cite other papers) in the automatic creation of technical surveys. We generated surveys of a set of Question Answering (QA) and Dependency Parsing (DP) papers, their abstracts, and their citation texts using four state-of-the-art summarization systems (C-LexRank, C-RR, LexRank, and Trimmer). We then used two separate approaches, nugget-based pyramid and ROUGE, to evaluate the surveys. The results from both approaches and all four summarization systems show that both citation texts and abstracts have unique survey-worthy information. These results also demonstrate that, unlike single document summarization (where citing sentences have been suggested to be inappropriate (Teufel et al., 2006)), multidocument summarization—especially technical survey creation—benefits considerably from citation texts.

We next plan to generate surveys using both citation texts and abstracts together as input. Creating readily consumable surveys is a hard task, especially when using only raw text and simple summarization techniques. Therefore we intend to combine these summarization and bibliometric techniques with suitable visualization methods towards the creation of iterative technical survey tools—systems that present surveys and bibliometric links in a visually convenient manner and which incorporate user feedback to produce even better surveys.

Acknowledgments

This work was supported, in part, by the National Science Foundation under Grant No. IIS-0705832, and in part, by the Human Language Technology Center of Excellence. 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 sponsor.

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. 2008a. 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.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir R. Radev. 2008b. 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.

Güneş Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

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

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics (2nd edition)*. Prentice-Hall.

Min-Yen Kan, Judith L. Klavans, and Kathleen R. McKeown. 2002. Using the Annotated Bibliography as a Resource for Indicative Summarization. In *Proceedings of LREC 2002*, Las Palmas, Spain.

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

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *SIGIR '95*, pages 68–73, New York, NY, USA. ACM.

Amy Langville and Carl Meyer. 2006. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

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.

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. 2004a. Bilingual presri: Integration of multiple research paper databases. In *Proceedings of RIAO 2004*, pages 195–211, Avignon, France.

Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. 2004b. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the 11th SIG Classification Research Workshop*, pages 117–134, Chicago, USA.

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.

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

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

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 the EMNLP*, pages 103–110, Sydney, Australia, July.

Simone Teufel. 2005. Argumentative Zoning for Improved Citation Indexing. *Computing Attitude and Affect in Text: Theory and Applications*, pages 159–170.

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)*.