Generating Impact-Based Summaries for Scientific Literature

Qiaozhu Mei
University of Illinois at Urbana-Champaign
qmei2@uiuc.edu

ChengXiang Zhai
University of Illinois at Urbana-Champaign
czhai@cs.uiuc.edu

Abstract

In this paper, we present a study of a novel summarization problem, i.e., summarizing the impact of a scientific publication. Given a paper and its citation context, we study how to extract sentences that can represent the most influential content of the paper. We propose language modeling methods for solving this problem, and study how to incorporate features such as authority and proximity to accurately estimate the impact language model. Experiment results on a SIGIR publication collection show that the proposed methods are effective for generating impact-based summaries.

1 Introduction

The volume of scientific literature has been growing rapidly. From recent statistics, each year 400,000 new citations are added to MEDLINE, the major biomedical literature database [1]. This fast growth of literature makes it difficult for researchers, especially beginning researchers, to keep track of the research trends and find high impact papers on unfamiliar topics.

Impact factors (Kaplan and Nelson, 2000) are useful, but they are just numerical values, so they cannot tell researchers which aspects of a paper are influential. On the other hand, a regular content-based summary (e.g., the abstract or conclusion section of a paper or an automatically generated topical summary (Giles et al., 1998)) can help a user know about the main content of a paper, but not necessarily the most influential content of the paper. Indeed, the abstract of a paper mostly reflects the expected impact of the paper as perceived by the author(s), which could significantly deviate from the actual impact of the paper in the research community. Moreover, the impact of a paper changes over time due to the evolution and progress of research in a field. For example, an algorithm published a decade ago may be no longer the state of the art, but the problem definition in the same paper can be still well accepted.

Although much work has been done on text summarization (See Section 6 for a detailed survey), to the best of our knowledge, the problem of impact summarization has not been studied before. In this paper, we study this novel summarization problem and propose language modeling-based approaches to solving the problem. By definition, the impact of a paper has to be judged based on the consent of research community, especially by people who cited it. Thus in order to generate an impact-based summary, we must use not only the original content, but also the descriptions of that paper provided in papers which cited it, making it a challenging task and different from a regular summarization setup such as news summarization. Indeed, unlike a regular summarization system which identifies and interprets the *topic* of a document, an impact summarization system should identify and interpret the *impact* of a paper.

We define the impact summarization problem in the framework of extraction-based text summarization (Luhn, 1958; McKeown and Radev, 1995), and cast the problem as an impact sentence retrieval

[1] http://www.nlm.nih.gov/bsd/history/tsld024.htm

Proceedings of ACL-08: HLT, pages 816–824,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

problem. We propose language models to exploit both the citation context and original content of a paper to generate an impact-based summary. We study how to incorporate features such as authority and proximity into the estimation of language models. We propose and evaluate several different strategies for estimating the impact language model, which is key to impact summarization. No existing test collection is available for evaluating impact summarization. We construct a test collection using 28 years of ACM SIGIR papers (1978 - 2005) to evaluate the proposed methods. Experiment results on this collection show that the proposed approaches are effective for generating impact-based summaries. The results also show that using both the original document content and the citation contexts is important and incorporating citation authority and proximity is beneficial.

An impact-based summary is not only useful for facilitating the exploration of literature, but also helpful for suggesting query terms for literature retrieval, understanding the evolution of research trends, and identifying the interactions of different research fields. The proposed methods are also applicable to summarizing the impact of documents in other domains where citation context exists, such as emails and weblogs.

The rest of the paper is organized as follows. In Section 2 and 3, we define the impact-based summarization problem and propose the general language modeling approach. In Section 4, we present different strategies and features for estimating an impact language model, a key challenge in impact summarization. We discuss our experiments and results in Section 5. Finally, the related work and conclusions are discussed in Section 6 and Section 7.

2 Impact Summarization

Following the existing work on topical summarization of scientific literature (Paice, 1981; Paice and Jones, 1993), we define an impact-based summary of a paper as a set of sentences extracted from a paper that can reflect the impact of the paper, where "impact" is roughly defined as the influence of the paper on research of similar or related topics as reflected in the citations of the paper. Such an extraction-based definition of summarization has also been quite common in most existing general summarization work (Radev et al., 2002).

By definition, in order to generate an impact summary of a paper, we must look at how other papers cite the paper, use this information to infer the impact of the paper, and select sentences from the original paper that can reflect the inferred impact. Note that we do not directly use the sentences from the citation context to form a summary. This is because in citations, the discussion of the paper cited is usually mixed with the content of the paper citing it, and sometimes also with discussion about other papers cited (Siddharthan and Teufel, 2007).

Formally, let $d = (s_0, s_1, ..., s_n)$ be a paper to be summarized, where s_i is a sentence. We refer to a sentence (in another paper) in which there is an explicit citation of d as a *citing sentence* of d. When a paper is cited, it is often discussed consecutively in more than one sentence near the citation, thus intuitively we would like to consider a window of sentences centered at a citing sentence; the window size would be a parameter to set. We call such a window of sentences a *citation context*, and use C to denote the union of all the citation contexts of d in a collection of research papers. Thus C itself is a set (more precisely bag) of sentences. The task of **impact-based summarization** is thus to 1) construct a representation of the impact of d, I, based on d and C; 2) design a scoring function $Score(.)$ to rank sentences in d based on how well a sentence reflects I. A user-defined number of top-ranked sentences can then be selected as the impact summary for d.

The formulation above immediately suggests that we can cast the impact summarization problem as a retrieval problem where each candidate sentence in d is regarded as a "document," the impact of the paper (i.e., I) as a "query," and our goal is to "retrieve" sentences that can reflect the impact of the paper as indicated by the citation context. Looking at the problem in this way, we see that there are two main challenges in impact summarization: first, we must be able to *infer* the impact based on both the citation contexts and the original document; second, we should measure how well a sentence reflects this inferred impact. To solve these challenges, in the next section, we propose to model impact with unigram language models and score sentences using

Kullback-Leibler divergence. We further propose methods for estimating the impact language model based on several features including the authority of citations, and the citation proximity.

3 Language Models for Impact Summarization

3.1 Impact language models

From the retrieval perspective, our collection is the paper to be summarized, and each sentence is a "document" to be retrieved. However, unlike in the case of ad hoc retrieval, we do not really have a query describing the impact of the paper; instead, we have a lot of citation contexts that can be used to infer information about the query. Thus the main challenge in impact summarization is to effectively construct a "virtual impact query" based on the citation contexts.

What should such a virtual impact query look like? Intuitively, it should model the impact-reflecting content of the paper. We thus propose to represent such a virtual impact query with a unigram language model. Such a model is expected to assign high probabilities to those words that can describe the impact of paper d, just as we expect a query language model in ad hoc retrieval to assign high probabilities to words that tend to occur in relevant documents (Ponte and Croft, 1998). We call such a language model the *impact language model* of paper d (denoted as θ_I); it can be estimated based on both d and its citation context C as will be discussed in Section 4.

3.2 KL-divergence scoring

With the impact language model in place, we can then adopt many existing probabilistic retrieval models such as the classical probabilistic retrieval models (Robertson and Sparck Jones, 1976) and the Kullback-Leibler (KL) divergence retrieval model (Lafferty and Zhai, 2001; Zhai and Lafferty, 2001a), to solve the problem of impact summarization by scoring sentences based on the estimated impact language model. In our study, we choose to use the KL-divergence scoring method to score sentences as this method has performed well for regular ad hoc retrieval tasks (Zhai and Lafferty, 2001a) and has an information theoretic interpretation.

To apply the KL-divergence scoring method, we assume that a candidate sentence s is generated from a sentence language model θ_s. Given s in d and the citation context C, we would first estimate θ_s based on s and estimate θ_I based on C, and then score s with the negative KL divergence of θ_s and θ_I. That is,

$$Score(s) = -D(\theta_I || \theta_s)$$

$$= \sum_{w \in V} p(w|\theta_I) \log p(w|\theta_s) - \sum_{w \in V} p(w|\theta_I) \log p(w|\theta_I)$$

where V is the set of words in our vocabulary and w denotes a word.

From the information theoretic perspective, the KL-divergence of θ_s and θ_I can be interpreted as measuring the average number of bits wasted in compressing messages generated according to θ_I (i.e., impact descriptions) with coding non-optimally designed based on θ_s. If θ_s and θ_I are very close, the KL-divergence would be small and $Score(s)$ would be high, which intuitively makes sense. Note that the second term (entropy of θ_I) is independent of s, so it can be ignored for ranking s.

We see that according to the KL-divergence scoring method, our main tasks are to estimate θ_s and θ_I. Since s can be regarded as a short document, we can use any standard method to estimate θ_s. In this work, we use Dirichlet prior smoothing (Zhai and Lafferty, 2001b) to estimate θ_s as follows:

$$p(w|\theta_s) = \frac{c(w,s) + \mu_s * P(w|D)}{|s| + \mu_s} \qquad (1)$$

where $|s|$ is the length of s, $c(w,s)$ is the count of word w in s, $p(w|D)$ is a background model estimated using $\frac{c(w,D)}{\sum_{w' \in V} c(w',D)}$ (D can be the set of all the papers available to us) and μ_s is a smoothing parameter to be empirically set. Note that as the length of a sentence is very short, smoothing is critical for addressing the data sparseness problem.

The remaining challenge is to estimate θ_I accurately based on d and its citation contexts.

4 Estimation of Impact Language Models

Intuitively, the impact of a paper is mostly reflected in the citation context. Thus the estimation of the impact language model should be primarily based on the citation context C. However, we would like our impact model to be able to help us select impact-reflecting sentences from d, thus it is important for the impact model to explain well the paper content in general. To achieve this balance, we treat the citation context C as prior information and the current document d as the observed data, and use Bayesian estimation to estimate the impact language model.

Specifically, let $p(w|C)$ be a citation context language model estimated based on the citation context C. We define Dirichlet prior with parameters $\{\mu_C p(w|C)\}_{w \in V}$ for the impact model, where μ_C encodes our confidence on this prior and effectively serves as a weighting parameter for balancing the contribution of C and d for estimating the impact model. Given the observed document d, the posterior mean estimate of the impact model would be (MacKay and Peto, 1995; Zhai and Lafferty, 2001b)

$$P(w|\theta_I) = \frac{c(w,d) + \mu_c p(w|C)}{|d| + \mu_c} \qquad (2)$$

μ_c can be interpreted as the equivalent sample size of our prior. Thus setting $\mu_c = |d|$ means that we put equal weights on the citation context and the document itself. $\mu_c = 0$ yields $p(w|\theta_I) = p(w|d)$, which is to say that the impact is entirely captured by the paper itself, and our impact summarization problem would then become the standard single document (topical) summarization. Intuitively though, we would want to set μ_c to a relatively large number to exploit the citation context in our estimation, which is confirmed in our experiments.

An alternative way is to simply interpolate $p(w|d)$ and $p(w|C)$ with a constant coefficient:

$$p(w|\theta_I) = (1-\delta)p(w|d) + \delta p(w|C) \qquad (3)$$

We will compare the two strategies in Section 5.

How do we estimate $p(w|C)$? Intuitively, words occurring in C frequently should have high probabilities. A simple way is to pool together all the sentences in C and use the maximum likelihood estimator,

$$p(w|C) = \frac{\sum_{s \in C} c(w,s)}{\sum_{w' \in V} \sum_{s' \in C} c(w', s')} \qquad (4)$$

where $c(w,s)$ is the count of w in s.

One deficiency of this simple estimate is that we treat all the (extended) citation sentences equally. However, there are at least two reasons why we want to assign unequal weights to different citation sentences: (1) A sentence closer to the citation label should contribute more than one far away. (2) A sentence occurring in a highly authorative paper should contribute more than that in a less authorative paper. To capture these two heuristics, we define a weight coefficient α_s for a sentence s in C as follows:

$$\alpha_s = pg(s)pr(s)$$

where $pg(s)$ is an authority score of the paper containing s and $pr(s)$ is a proximity score that rewards a sentence close to the citation label.

For example, $pg(s)$ can be the PageRank value (Brin and Page, 1998) of the document with s, which measures the authority of the document based on a citation graph, and is computed as follows: We construct a directed graph from the collection of scientific literature with each paper as a vertex and each citation as a directed edge pointing from the citing paper to the cited paper. We can then use the standard PageRank algorithm (Brin and Page, 1998) to compute a PageRank value for each document. We used this approach in our experiments.

We define $pr(s)$ as $pr(s) = \frac{1}{\alpha^k}$, where k is the distance (counted in terms of the number of sentences) between sentence s and the center sentence of the window containing s; by "center sentence", we mean the citing sentence containing the citation label. Thus the sentence with the citation label will have a proximity of 1 (because $k = 0$), while the sentences away from the citation label will have a decaying weight controlled by parameter α.

With α_s, we can then use the following "weighted" maximum likelihood estimate for the impact language model:

$$p(w|C) = \frac{\sum_{s \in C} \alpha_s c(w,s)}{\sum_{w' \in V} \sum_{s' \in C} \alpha_{s'} c(w', s')} \qquad (5)$$

As we will show in Section 5, this weighted maximum likelihood estimate performs better than the simple maximum likelihood estimate, and both $pg(s)$ and $pr(s)$ are useful.

5 Experiments and Results

5.1 Experiment Design

5.1.1 Test set construction

Because no existing test set is available for evaluating impact summarization, we opt to create a test set based on 28 years of ACM SIGIR papers (1978 - 2005) available through the ACM Digital Library[2] and the SIGIR membership. Leveraging the explicit citation information provided by ACM Digital Library, for each of the 1303 papers, we recorded all other papers that cited the paper and extracted the citation context from these citing papers. Each citation context contains 5 sentences with 2 sentences before and after the citing sentence.

Since a low-impact paper would not be useful for evaluating impact summarization, we took all the 14 papers from the SIGIR collection that have no less than 20 citations by papers in the same collection as candidate papers for evaluation. An expert in Information Retrieval field read each paper and its citation context, and manually created an impact-based summary by selecting all the "impact-capturing" sentences from the paper. Specifically, the expert first attempted to understand the most influential content of a paper by reading the citation contexts. The expert then read each sentence of the paper and made a decision whether the sentence covers some "influential content" as indicated in the citation contexts. The sentences that were decided as covering some influential content were then collected as the gold standard impact summary for the paper.

We assume that the title of a paper will always be included in the summary, so we excluded the title both when constructing the gold standard and when generating a summary. The gold standard summaries have a minimum length of 5 sentences and a maximum length of 18 sentences; the median length is 9 sentences. These 14 impact-based summaries are used as gold standards for our experiments, based on which all summaries generated by the system are evaluated. This data set is available at http://timan.cs.uiuc.edu/data/impact.html. We must admit that using only 14 papers and only one expert for evaluation is a limitation of our work. However, going beyond the 14 papers would risk reducing the reliability of impact judgment due to the sparseness of citations. How to develop a better test collection is an important future direction.

5.1.2 Evaluation Metrics

Following the current practice in evaluating summarization, particularly DUC[3], we use the ROUGE evaluation package (Lin and Hovy, 2003). Among ROUGE metrics, ROUGE-N (models n-gram co-occurrence, N = 1, 2) and ROUGE-L (models longest common sequence) generally perform well in evaluating both single-document summarization and multi-document summarization (Lin and Hovy, 2003). Since they are general evaluation measures for summarization, they are also applicable to evaluating the MEAD-Doc+Cite baseline method to be described below. Thus although we evaluated our methods with all the metrics provided by ROUGE, we only report ROUGE-1 and ROUGE-L in this paper (other metrics give very similar results).

5.1.3 Baseline methods

Since impact summarization has not been previously studied, there is no natural baseline method to compare with. We thus adapt some state-of-the-art conventional summarization methods implemented in the MEAD toolkit (Radev et al., 2003)[4] to obtain three baseline methods: (1) **LEAD:** It simply extracts sentences from the beginning of a paper, i.e., sentences in the abstract or beginning of the introduction section; we include **LEAD** to see if such "leading sentences" reflect the impact of a paper as authors presumably would expect to summarize a paper's contributions in the abstract. (2) **MEAD-Doc:** It uses the single-document summarizer in MEAD to generate a summary based solely on the original paper; comparison with this baseline can tell us how much better we can do than a conventional topic-based summarizer that does not consider the citation context. (3) **MEAD-Doc+Cite:** Here we concatenate all the citation contexts in a paper to form a "citation document" and then use the MEAD multidocument summarizer to generate a summary from the original paper plus all its citation documents; this baseline represents a reasonable way

[2] http://www.acm.org/dl

[3] http://duc.nist.gov/

[4] "http://www.summarization.com/mead/"

Sum. Length	Metric	Random	LEAD	MEAD-Doc	MEAD-Doc+Cite	KL-Divergence
3	ROUGE-1	0.163	0.167	0.301*	0.248	**0.323**
3	ROUGE-L	0.144	0.158	0.265	0.217	**0.299**
5	ROUGE-1	0.230	0.301	0.401	0.333	**0.467**
5	ROUGE-L	0.214	0.292	0.362	0.298	**0.444**
10	ROUGE-1	0.430	0.514	0.575	0.472	**0.649**
10	ROUGE-L	0.396	0.494	0.535	0.428	**0.622**
15	ROUGE-1	0.538	0.610	0.685	0.552	**0.730**
15	ROUGE-L	0.499	0.586	0.650	0.503	**0.705**

Table 1: Performance Comparison of Summarizers

of applying an existing summarization method to generate an impact-based summary. Note that this method may extract sentences in the citation contexts but not in the original paper.

5.2 Basic Results

We first show some basic results of impact summarization in Table 1. They are generated using constant coefficient interpolation for the impact language model (i.e., Equation 3) with $\delta = 0.8$, weighted maximum likelihood estimate for the citation context model (i.e., Equation 5) with $\alpha = 3$, and $\mu_s = 1,000$ for candidate sentence smoothing (Equation 1). These results are not necessarily optimal as will be seen when we examine parameter and method variations.

From Table 1, we see clearly that our method consistently outperforms all the baselines. Among the baselines, MEAD-Doc is consistently better than both LEAD and MEAD-Doc+Cite. While MEAD-Doc's outperforming LEAD is not surprising, it is a bit surprising that MEAD-Doc also outperforms MEAD-Doc+Cite as the latter uses both the citation context and the original document. One possible explanation may be that MEAD is not designed for impact summarization and it has been trapped by the distracting content in the citation context [5]. Indeed, this can also explain why MEAD-Doc+Cite tends to perform worse than LEAD by ROUGE-L since if MEAD-Doc+Cite picks up sentences from the citation context rather than the original papers, it would not match as well with the gold standard as LEAD which selects sentences from the original papers. These results thus show that conventional summarization techniques are inadequate for impact summarization, and the proposed language modeling methods are more effective for generating impact-based summaries.

In Table 2, we show a sample impact-based summary and the corresponding MEAD-Doc regular summary. We see that the regular summary tends to have general sentences about the problem, background and techniques, not very informative in conveying specific contributions of the paper. None of these sentences was selected by the human expert. In contrast, the sentences in the impact summary cover several details of the impact of the paper (i.e., specific smoothing methods especially Dirichlet prior, sensitivity of performance to smoothing, and dual role of smoothing), and sentences 4 and 6 are also among the 8 sentences picked by the human expert. Interestingly, neither sentence is in the abstract of the original paper, suggesting a deviation of the actual impact of a paper and that perceived by the author(s).

5.3 Component analysis

We now turn to examine the effectiveness of each component in the proposed methods and different strategies for estimating θ_I.

Effectiveness of interpolation: We hypothesized that we need to use both the original document and the citation context to estimate θ_I. To test this hypothesis, we compare the results of using only d, only the citation context, and interpolation of them in Table 3. We show two different strategies of interpolation (i.e., constant coefficient with $\delta = 0.8$ and Dirichlet with $\mu_c = 20,000$) as described in Section 4.

From Table 3, we see that both strategies of interpolation indeed outperform using either the origi-

[5] One anonymous reviewer suggested an interesting improvement to the MEAD-Doc+Cite baseline, in which we would first extract sentences from the citation context and then for each extracted sentence find a similar one in the original paper. Unfortunately, we did not have time to test this approach before the deadline for the camera-ready version of this paper.

Impact-based summary:
1. Figure 5: Interpolation versus backoff for Jelinek-Mercer (top), Dirichlet smoothing (middle), and absolute discounting (bottom).
2. Second, one can de-couple the two different roles of smoothing by adopting a two stage smoothing strategy in which Dirichlet smoothing is first applied to implement the estimation role and Jelinek-Mercer smoothing is then applied to implement the role of query modeling
3. We find that the backoff performance is more sensitive to the smoothing parameter than that of interpolation, especially in Jelinek-Mercer and Dirichlet prior.
4. We then examined three popular interpolation-based smoothing methods (Jelinek-Mercer method, Dirichlet priors, and absolute discounting), as well as their backoff versions, and evaluated them using several large and small TREC retrieval testing collections.
summary 5. By rewriting the query-likelihood retrieval model using a smoothed document language model, we derived a general retrieval formula where the smoothing of the document language model can be interpreted in terms of several heuristics used intraditional models, including TF-IDF weighting and document length normalization.
6. We find that the retrieval performance is generally sensitive to the smoothing parameters, suggesting that an understanding and appropriate setting of smoothing parameters is very important in the language modeling approach.
Regular summary (generated using MEAD-Doc):
1. Language modeling approaches to information retrieval are attractive and promising because they connect the problem of retrieval with that of language model estimation, which has been studied extensively in other application areas such as speech recognition.
2. The basic idea of these approaches is to estimate a language model for each document, and then rank documents by the likelihood of the query according to the estimated language model.
3. On the one hand, theoretical studies of an underlying model have been developed; this direction is, for example, represented by the various kinds of logic models and probabilistic models (e.g., [14, 3, 15, 22]).
4. After applying the Bayes' formula and dropping a document-independent constant (since we are only interested in ranking documents), we have $p(d\|q) \propto (q\|d)p(d)$.
5. As discussed in [1], the righthand side of the above equation has an interesting interpretation, where, $p(d)$ is our prior belief that d is relevant to any query and $p(q\|d)$ is the query likelihood given the document, which captures how well the document "fits" the particular query q.
6. The probability of an unseen word is typically taken as being proportional to the general frequency of the word, e.g., as computed using the document collection.

Table 2: Impact-based summary vs. regular summary for the paper "A study of smoothing methods for language models applied to ad hoc information retrieval".

nal document model ($p(w|d)$) or the citation context model ($p(w|C)$) alone, which confirms that both the original paper and the citation context are important for estimating θ_I. We also see that using the citation context alone is better than using the original paper alone, which is expected. Between the two strategies, Dirichlet dynamic coefficient is slightly better than constant coefficient (CC), after optimizing the interpolation parameter for both strategy.

			Interpolation	
Measure	$P(w\|d)$	$P(w\|C)$	ConstCoef	Dirichlet
ROUGE-1	0.529	0.635	0.643	**0.647**
ROUGE-L	0.501	0.607	0.619	**0.623**

Table 3: Effectiveness of interpolation

Citation authority and proximity: These heuristics are very interesting to study as they are unique to impact summarization and not well studied in the existing summarization work.

pg(s)		**pr(s)=1/α^{k}**		
	pr(s) off	$\alpha = \mathbf{2}$	$\alpha = \mathbf{3}$	$\alpha = \mathbf{4}$
Off	0.685	0.711	0.714	0.700
On	0.708	0.712	0.706	0.703

Table 4: Authority (pg(s)) and proximity (pr(s))

In Table 4, we show the ROUGE-L values for various combinations of these two heuristics (summary length is 15). We turn off either $pg(s)$ or $pr(s)$ by setting it to a constant; when both are turned off, we have the unweighted MLE of $p(w|C)$ (Equation 4). Clearly, using weighted MLE with any of the two heuristics is better than the unweighted MLE, indicating that both heuristics are effective. However, combining the two heuristics does not always improve over using a single one. Since intuitively these two heuristics are orthogonal, this may suggest that our way of combining the two scores (i.e., taking a product of them) may not be optimal; further study is needed to better understand this. The ROUGE-1 results are similar.

Tuning of other parameters: There are three other parameters which need to be tuned: (1) μ_s for candidate sentence smoothing (Equation 1); (2) μ_c in Dirichlet interpolation for impact model estimation (Equation 2); and (3) δ in constant coefficient interpolation (Equation 3). We have examined the sensitivity of performance to these parameters. In general, for a wide range of values of these parameters, the performance is relatively stable and near optimal. Specifically, the performance is near optimal as long as μ_s and μ_c are sufficiently large ($\mu_s \geq 1000$, $\mu_c \geq 20,000$), and the interpolation parameter δ is between 0.4 and 0.9.

6 Related Work

General text summarization, including single document summarization (Luhn, 1958; Goldstein et al., 1999) and multi-document summarization (Kraaij et al., 2001; Radev et al., 2003) has been well studied; our work is under the framework of extractive summarization (Luhn, 1958; McKeown and Radev, 1995; Goldstein et al., 1999; Kraaij et al., 2001), but our problem formulation differs from any existing formulation of the summarization problem. It differs from regular single-document summarization because we utilize extra information (i.e. citation contexts) to summarize the impact of a paper. It also differs from regular multi-document summarization because the roles of original documents and citation contexts are not equivalent. Specifically, citation contexts serve as an indicator of the impact of the paper, but the summary is generated by extracting the sentences from the original paper.

Technical paper summarization has also been studied (Paice, 1981; Paice and Jones, 1993; Saggion and Lapalme, 2002; Teufel and Moens, 2002), but the previous work did not explore citation context to emphasize the impact of papers.

Citation context has been explored in several studies (Nakov et al., 2004; Ritchie et al., 2006; Schwartz et al., 2007; Siddharthan and Teufel, 2007). However, none of the previous studies has used citation context in the same way as we did, though the potential of *directly* using citation sentences (called *citances*) to summarize a paper was pointed out in (Nakov et al., 2004).

Recently, people have explored various types of auxiliary knowledge such as hyperlinks (Delort et al., 2003) and clickthrough data (Sun et al., 2005), to summarize a webpage; such work is related to ours as anchor text is similar to citation context, but it is based on a standard formulation of multi-document summarization and would contain only sentences from anchor text.

Our work is also related to work on using language models for retrieval (Ponte and Croft, 1998; Zhai and Lafferty, 2001b; Lafferty and Zhai, 2001) and summarization (Kraaij et al., 2001). However, we do not have an explicit query and constructing the impact model is a novel exploration. We also proposed new language models to capture the impact.

7 Conclusions

We have defined and studied the novel problem of summarizing the impact of a research paper. We cast the problem as an impact sentence retrieval problem, and proposed new language models to model the impact of a paper based on both the original content of the paper and its citation contexts in a literature collection with consideration of citation autority and proximity.

To evaluate impact summarization, we created a test set based on ACM SIGIR papers. Experiment results on this test set show that the proposed impact summarization methods are effective and outperform several baselines that represent the existing summarization methods.

An important future work is to construct larger test sets (e.g., of biomedical literature) to facilitate evaluation of impact summarization. Our formulation of the impact summarization problem can be further improved by going beyond sentence retrieval and considering factors such as redundancy and coherency to better organize an impact summary. Finally, automatically generating impact-based summaries can not only help users access and digest influential research publications, but also facilitate other literature mining tasks such as milestone mining and research trend monitoring. It would be interesting to explore all these applications.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive comments. This work is in part supported by a Yahoo! Graduate Fellowship and NSF grants under award numbers 0713571, 0347933, and 0428472.

References

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web*, pages 107–117.

J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. 2003. Enhanced web document summarization using hyperlinks. In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pages 208–215.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: an automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, pages 89–98.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of ACM SIGIR 99*, pages 121–128.

Nancy R. Kaplan and Michael L. Nelson. 2000. Determining the publication impact of a digital library. *J. Am. Soc. Inf. Sci.*, 51(4):324–339.

W. Kraaij, M. Spitters, and M. van der Heijden. 2001. Combining a mixture language model and naive bayes for multi-document summarisation. In *Proceedings of the DUC2001 workshop*.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of ACM SIGIR 2001*, pages 111–119.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

D. MacKay and L. Peto. 1995. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):289–307.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82.

P. Nakov, A. Schwartz, and M. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of ACM SIGIR'04 Workshop on Search and Discovery in Bioinformatics*.

Chris D. Paice and Paul A. Jones. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78.

C. D. Paice. 1981. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, pages 172–191.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Comput. Linguist.*, 28(4):399–408.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization: the mead project. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 375–382.

A. Ritchie, S. Teufel, and S. Robertson. 2006. Creating a test collection for citation-based ir experiments. In *Proceedings of the HLT-NAACL 2006*, pages 391–398.

S. Robertson and K. Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Hpracop Saggion and Guy Lapalme. 2002. Generating indicative-informative summaries with sumUM. *Computational Linguistics*, 28(4):497–526.

A. S. Schwartz, A. Divoli, and M. A. Hearst. 2007. Multiple alignment of citation sentences with conditional random fields and posterior decoding. In *Proceedings of the 2007 EMNLP-CoNLL*, pages 847–857.

A. Siddharthan and S. Teufel. 2007. Whose idea was this, and why does it matter? attributing scientific work to citations. In *Proceedings of NAACL/HLT-07*, pages 316–323.

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 194–201.

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

ChengXiang Zhai and John Lafferty. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pages 403–410.

Chengxiang Zhai and John Lafferty. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342.