Identifying Non-explicit Citing Sentences for Citation-based Summarization

Vahed Qazvinian
Department of EECS
University of Michigan
Ann Arbor, MI
vahed@umich.edu

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

Abstract

Identifying background (context) information in scientific articles can help scholars understand major contributions in their research area more easily. In this paper, we propose a general framework based on probabilistic inference to extract such context information from scientific papers. We model the sentences in an article and their lexical similarities as a *Markov Random Field* tuned to detect the patterns that context data create, and employ a *Belief Propagation* mechanism to detect likely context sentences. We also address the problem of generating surveys of scientific papers. Our experiments show greater pyramid scores for surveys generated using such context information rather than citation sentences alone.

1 Introduction

In scientific literature, scholars use citations to refer to external sources. These secondary sources are essential in comprehending the new research. Previous work has shown the importance of citations in scientific domains and indicated that citations include survey-worthy information (Siddharthan and Teufel, 2007; Elkiss et al., 2008; Qazvinian and Radev, 2008; Mohammad et al., 2009; Mei and Zhai, 2008).

A citation to a paper in a scientific article may contain explicit information about the cited research. The following example is an excerpt from a CoNLL paper[1] that contains information about Eisner's work on bottom-up parsers and the notion of span in parsing:

"*Another use of bottom-up is due to* **Eisner (1996)**, *who introduced the notion of a span.*"

[1]Buchholz and Marsi "CoNLL-X Shared Task On Multilingual Dependency Parsing", CoNLL 2006

However, the citation to a paper may not always include explicit information about the cited paper:

"*This approach is one of those described in* **Eisner (1996)**"

Although this sentence alone does not provide any information about the cited paper, it suggests that its surrounding sentences describe the proposed approach in Eisner's paper:

> "*... In an all pairs approach, every possible pair of two tokens in a sentence is considered and some score is assigned to the possibility of this pair having a (directed) dependency relation. Using that information as building blocks, the parser then searches for the best parse for the sentence. This approach is one of those described in* **Eisner (1996)**."

We refer to such *implicit citations* that contain information about a specific secondary source but do not explicitly cite it, as sentences with *context* information or *context sentences* for short. We look at the patterns that such sentences create and observe that context sentences occur withing a small neighborhood of explicit citations. We also discuss the problem of extracting context sentences for a source-reference article pair. We propose a general framework that looks at each sentence as a random variable whose value determines its state about the target paper. In summary, our proposed model is based on the probabilistic inference of these random variables using graphical models. Finally we give evidence on how such sentences can help us produce better surveys of research areas. The rest of this paper is organized as follows. Preceded by a review of prior work in Section 2, we explain the data collection and our annotation process in Section 3. Section 4 explains our methodology and is followed by experimental setup in Section 5.

ACL-ID	Author	Title	Year	#Refs all	#Refs AAN	# Sents
P08-2026	McClosky & Charniak	Self-Training for Biomedical Parsing	2008	12	8	102
N07-1025*	Mihalcea	Using Wikipedia for Automatic ...	2007	21	12	153
N07-3002	Wang	Learning Structured Classifiers ...	2007	22	14	74
P06-1101	Snow et, al.	Semantic Taxonomy Induction ...	2006	19	9	138
P06-1116	Abdalla & Teufel	A Bootstrapping Approach To ...	2006	24	10	231
W06-2933	Nivre et, al.	Labeled Pseudo-Projective Dependency ...	2006	27	5	84
P05-1044	Smith & Eisner	Contrastive Estimation: Training Log-Linear ...	2005	30	13	262
P05-1073	Toutanova et, al.	Joint Learning Improves Semantic Role Labeling	2005	14	10	185
N03-1003	Barzilay & Lee	Learning To Paraphrase: An Unsupervised ...	2003	26	13	203
N03-2016*	Kondrak et, al.	Cognates Can Improve Statistical Translation ...	2003	8	5	92

Table 1: Papers chosen from AAN as source papers for the evaluation corpus, together with their publication year, number of references (in AAN) and number of sentences. Papers marked with $*$ are used to calculate annotation inter-judge agreement.

2 Prior Work

Analyzing the structure of scientific articles and their relations has received a lot of attention recently. The structure of citation and collaboration networks has been studied in (Teufel et al., 2006; Newman, 2001), and summarization of scientific documents is discussed in (Teufel and Moens, 2002). In addition, there is some previous work on the importance of citation sentences. (Elkiss et al., 2008) perform a large-scale study on citations in the free PubMed Central (PMC) and show that they contain information that may not be present in abstracts. In other work, (Nanba and Okumura, 1999; Nanba et al., 2004b; Nanba et al., 2004a) analyze citation sentences and automatically categorize them in order to build a tool for survey generation.

(Bradshaw, 2002; Bradshaw, 2003) uses citations to determine the content of articles. Similarly, (Qazvinian and Radev, 2008; Mei and Zhai, 2008; Mohammad et al., 2009) directly use the text of citation sentences to produce summaries of scientific papers. Determining the scientific attribution of an article has also been studied before. (Siddharthan and Teufel, 2007; Teufel, 2005) categorize sentences according to their role in the author's argument into predefined classes: Own, Other, Background, Textual, Aim, Basis, Contrast.

Little work has been done on automatic citation extraction from research papers. (Kaplan et al., 2009) introduces "citation-site" as a block of text in which the cited text is discussed. The mentioned work uses a machine learning method for extracting citations from research papers and evaluates the result using 4 annotated articles.

In our work we use graphical models to extract context sentences. Graphical models have a number of properties and corresponding techniques and have been used before on Information Retrieval tasks. (Romanello et al., 2009) use Conditional Random Fields (CRF) to extract references from unstructured text in digital libraries of classic texts. Similar work include term dependency extraction (Metzler and Croft, 2005), query expansion (Metzler and Croft, 2007), and automatic feature selection (Metzler, 2007).

3 Data

The ACL Anthology Network (AAN)[2] is a collection of papers from the ACL Anthology[3] published in the Computational Linguistics journal and proceedings from ACL conferences and workshops and includes more than $14,000$ papers over a period of four decades (Radev et al., 2009). AAN includes the citation network of the papers in the ACL Anthology. The papers in AAN are publicly available in text format retrieved by an OCR process from the original pdf files, and are segmented into sentences.

To build a corpus for our experiments we picked 10 recently published papers from various areas in NLP[4], each of which had references for a total of 203 candidate paper-reference pairs. Table 1 lists these papers together with their authors, titles, publication year, number of references, number of references within AAN, and the number of sentences.

[2] http://clair.si.umich.edu/clair/anthology/

[3] http://www.aclweb.org/anthology-new/

[4] Regardless of data selection, the methodology in this work is applicable to any of the papers in AAN.

L&P	S&al	Sentence
		. . .
$\mathcal{C}$	$\mathcal{C}$	Jacquemin (1999) and Barzilay and McKeown (2001) identify phrase level paraphrases, while **Lin and Pantel (2001)** and **Shinyama et al. (2002)** acquire structural paraphrases encoded as templates.
1	1	These latter are the most closely related to the sentence-level paraphrases we desire, and so we focus in this section on template-induction approaches.
$\mathcal{C}$	0	**Lin and Pantel (2001)** extract inference rules, which are related to paraphrases (for example, X wrote Y implies X is the author of Y), to improve question answering.
1	0	They assume that paths in dependency trees that take similar arguments (leaves) are close in meaning.
1	0	However, only two-argument templates are considered.
0	$\mathcal{C}$	**Shinyama et al. (2002)** also use dependency-tree information to extract templates of a limited form (in their case, determined by the underlying information extraction application).
1	1	Like us (and unlike Lin and Pantel, who employ a single large corpus), they use articles written about the same event in different newspapers as data.
1	1	Our approach shares two characteristics with the two methods just described: pattern comparison by analysis of the patterns respective arguments, and use of nonparallel corpora as a data source.
0	0	However, extraction methods are not easily extended to generation methods.
1	1	One problem is that their templates often only match small fragments of a sentence.
1	1	While this is appropriate for other applications, deciding whether to use a given template to generate a paraphrase requires information about the surrounding context provided by the entire sentence.
		. . .

Table 2: Part of the annotation for N03-1003 with respect to two of its references "Lin and Pantel (2001)" (the first column) "Shinyama et al. (2002)" (the second column). $\mathcal{C}$s indicate explicit citations, 1s indicate implicit citations and 0s are none.

3.1 Annotation Process

We annotated the sentences in each paper from Table 1. Each *annotation instance* in our setting corresponds to a paper-reference pair, and is a vector in which each dimension corresponds to a sentence and is marked with a $\mathcal{C}$ if it explicitly cites the reference, and with a 1 if it implicitly talks about it. All other sentences are marked with 0s. Table 2 shows a portion of two separate annotation instances of N03-1003 corresponding to two of its references. Our annotation has resulted in 203 annotation instances each corresponding to one paper-reference pair. The goal of this work is to automatically identify all context sentences, which are marked as "1".

3.1.1 Inter-judge Agreement

We also asked a neutral annotator[5] to annotate two of our datasets that are marked with $*$ in Table 1. For each paper-reference pair, the annotator was provided with a vector in which explicit citations were already marked with $\mathcal{C}$s. The annotation guidelines instructed the annotator to look at each explicit citation sentence, and read up to 15 sentences before and after, then mark context sentences around that sentence with 1s. Next, the 29 annotation instances done by the external annotator were compared with the corresponding annotations that we did, and the Kappa coefficient (κ) was calculated. The κ statistic is formulated as

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the probability that annotators agree by chance if each annotator is randomly assigning categories. To calculate κ, we ignored all explicit citations (since they were provided to the external annotator) and used the binary categories (i.e., 1 for context sentences, and 0 otherwise) for all other sentences. Table 3 shows the annotation vector size (i.e., number of sentences), number of annotation instances (i.e., number of references), and average κ for each set. The average κ is above 0.85 in both cases, suggesting that the annotation process has a low degree of subjectivity and can be considered reliable.

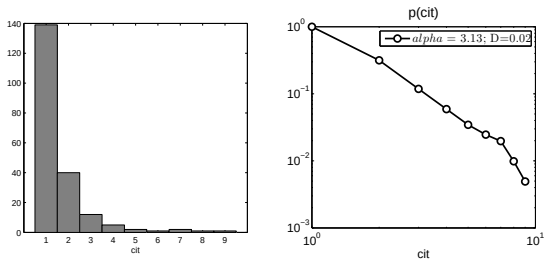
ACL-ID	vector size	# Annotations	$\overline{\kappa}$
N07-1025*	153	21	0.889 ± 0.30
N03-2016*	92	8	0.853 ± 0.35

Table 3: Average κ coefficient as inter-judge agreement for annotations of two sets

3.2 Analysis

In this section we describe our analysis. First, we look at the number of explicit citations each reference has received in a paper. Figure 1 (a) shows the histogram corresponding to this distribution. It indicates that the majority of references get cited in only 1 sentence in a scientific article, while the maximum being 9 in our collected dataset with only 1 instance (i.e., there is only 1 reference that gets cited 9 times in a paper). Moreover, the data exhibits a highly positive-skewed distribution. This is illustrated on a log-log scale in Figure 1 (b). This highly skewed distribution indicates that the majority of references get cited only once in a citing paper. The very small number of citing sentences can not make a full inventory of the contributions of the cited paper, and therefore, extracting explicit citations alone without context sentences may result in information loss about the contributions of the cited paper.

[5] Someone not involved in the paper but an expert in NLP.



a b

Figure 1: (a) Histogram of the number of different citations to each reference in a paper. (b) The distribution observed for the number of different citations on a log-log scale.

gap size	0	1	2	4	9	10	15	16
instance	273	14	2	1	2	1	1	1

Table 4: The distribution of gaps in the annotated data

Next, we investigate the distance between context sentences and the closest citations. For each context sentence, we find its distance to the closets context sentence or explicit citation. Formally, we define the *gap* to be the number of sentences between a context sentence (marked with 1) and the closest context sentence or explicit citation (marked with either C or 1) to it. For example, the second column of Table 2 shows that there is a gap of size 1 in the 9^{th} sentence in the set of context and citation sentences about Shinyama et al. (2002). Table 4 shows the distribution of gap sizes in the annotated data. This observation suggests that the majority of context sentences directly occur after or before a citation or another context sentence. However, it shows that gaps between sentences describing a cited paper actually exist, and a proposed method should have the capability to capture them.

4 Proposed Method

In this section we propose our methodology that enables us to identify the context information of a cited paper. Particularly, the task is to assign a binary label X_C to each sentence S_i from a paper S, where $X_C = 1$ shows a context sentence related to a given cited paper, C. To solve this problem we propose a systematic way to model the network level relationship between consecutive sentences. In summary, each sentence is represented with a node and is given two scores (context, non-context), and we update these scores to be in harmony with the neighbors' scores.

A particular class of graphical models known as *Markov Random Fields* (MRFs) are suited for solving inference problems with uncertainty in observed data. The data is modeled as an undirected graph with two types of nodes: hidden and observed. Observed nodes represent values that are known from the data. Each hidden node x_u, corresponding to an observed node y_u, represents the true state underlying the observed value. The state of a hidden node is related to the value of its corresponding observed node as well as the states of its neighboring hidden nodes.

The *local Markov property* of an MRF indicates that a variable is conditionally independent on all other variables given its neighbors: $x_v \perp\!\!\!\perp x_{V \setminus cl(v)} | x_{ne(v)}$, where $ne(v)$ is the set of neighbors of v, and $cl(v) = \{v\} \cup ne(v)$ is the closed neighborhood of v. Thus, the state of a node is assumed to statistically depend only upon its hidden node and each of its neighbors, and independent of any other node in the graph given its neighbors.

Dependencies in an MRF are represented using two functions: *Compatibility function (ψ)* and *Potential function (ϕ)*. $\psi_{uv}(x_c, x_d)$ shows the edge potential of an edge between two nodes u, v of classes x_c and x_d. Large values of ψ_{uv} would indicate a strong association between x_c and x_d at nodes u, v. The Potential function, $\phi_i(x_c, y_c)$, shows the statistical dependency between x_c and y_c at each node i assumed by the MRF model.

In order to find the marginal probabilities of x_is in a MRF we can use *Belief Propagation (BP)* (Yedidia et al., 2003). If we assume the y_is are fixed and show $\phi_i(x_i, y_i)$ by $\phi_i(x_i)$, we can find the joint probability distribution for unknown variables x_i as

$$p(\{x\}) = \frac{1}{Z} \prod_{ij} \psi_{ij}(x_i, x_j) \prod_i \phi_i(x_i)$$

In the BP algorithm a set of new variables m is introduced where $m_{ij}(x_j)$ is the message passed from i to j about what state x_j should be in. Each message, $m_{ij}(x_j)$, is a vector with the same dimensionality of x_j in which each dimension shows i's opinion about j being in the corresponding class. Therefore each message could be considered as a probability distribution and its components should sum up to 1. The final belief at a node i, in the BP algorithm, is also a vector with

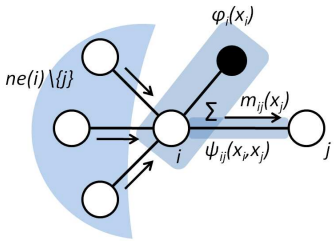


Figure 2: The illustration of the message updating rule. Elements that make up the message from a node i to another node j: messages from i's neighbors, local evidence at i, and propagation function between i, j summed over all possible states of node i.

the same dimensionality of messages, and is proportional to the local evidence as well as all messages from the node's neighbors:

$$b_i(x_i) \leftarrow k\phi_i(x_i) \prod_{j \in ne(i)} m_{ji}(x_i) \quad (1)$$

where k is the normalization factor of the beliefs about different classes. The message passed from i to j is proportional to the propagation function between i, j, the local evidence at i, and all messages sent to i from its neighbors except j:

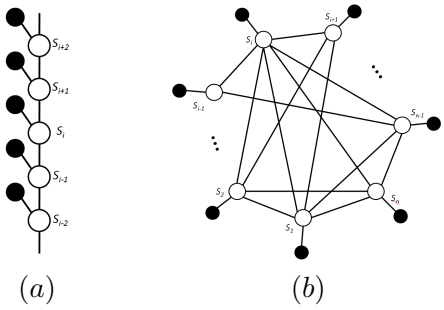
$$m_{ij}(x_j) \leftarrow \sum_{x_i} \phi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in ne(i)\setminus j} m_{ki}(x_i) \quad (2)$$

Figure 2 illustrates the message update rule.

Convergence can be determined based on a variety of criteria. It can occur when the maximum change of any message between iteration steps is less than some threshold. Convergence is guaranteed for trees but not for general graphs. However, it typically occurs in practice (McGlohon et al., 2009). Upon convergence, belief scores are determined by equation 1.

4.1 MRF construction

To find the sentences from a paper that form the context information of a given cited paper, we build an MRF in which a hidden node x_i and an observed node y_i correspond to each sentence. The structure of the graph associated with the MRF is dependent upon the validity of a basic assumption. This assumption indicates that the generation of a sentence (in form of its words) only depends on its surrounding sentences. Said differently, each sentence is written independently of



(a) (b)

Figure 3: The structure of the MRF constructed based on the independence of non-adjacent sentences; (a) left, each sentence is independent on all other sentences given its immediate neighbors. (b) right, sentences have dependency relationship with each other regardless of their position.

all other sentences given a number of its neighbors. This local dependence assumption can result in a number of different MRFs, each built assuming a dependency between a sentence and all sentences within a particular distance. Figure 3 shows the structure of the two MRFs at either extreme of the local dependence assumption. In Figure 3 a, each sentence only depends on one following and one preceding sentence, while Figure 3 b shows an MRF in which sentences are dependent on each other regardless of their position. We refer to the former by $\mathbf{BP_1}$, and to the latter by $\mathbf{BP_n}$. Generally, we use $\mathbf{BP_i}$ to denote an MRF in which each sentence is connected to i sentences before and after.

$\psi_{ij}(x_c, x_d)$	$x_d = 0$	$x_d = 1$
$x_c = 0$	0.5	0.5
$x_c = 1$	$1 - S_{ij}$	S_{ij}

Table 5: The compatibility function ψ between any two nodes in the MRFs from the sentences in scientific papers

4.2 Compatibility Function

The compatibility function of an MRF represents the association between the hidden node classes. A node's belief to be in class 1 is its probability to be included in the context. The belief of a node i, about its neighbor j to be in either classes is assumed to be 0.5 if i is in class 0. In other words, if a node is not part of the context itself, we assume it has no effect on its neighbors' classes. In contrast, if i is in class 1 its belief about its neighbor j is determined by their mutual lexical similarity.

If this similarity is close to 1 it indicates a stronger tie between i, j. However, if i, j are not similar, i's probability of being in class 1, should not affect that of j's. To formalize this assumption we use the sigmoid of the cosine similarity of two sentences to build ψ. More formally, we define S to be

$$S_{ij} = \frac{1}{1 + e^{-cosine(i,j)}}$$

The sigmoid function obtains a value of 0.5 for a cosine of 0 indicating that there is no bias in the association of the two sentences. The matrix in Table 5 shows the compatibility function built based on the above arguments.

4.3 Potential Function

The node potential function of an MRF can incorporate some other features observable from data. Here, the goal is to find all sentences that are about a specific cited paper, without having explicit citations. To build the node potential function of the observed nodes, we use some sentence level features. First, we use the explicit citation as an important feature of a sentence. This feature can affect the belief of the corresponding hidden node, which can in turn affect its neighbors' beliefs. For a given paper-reference pair, we flag (with a 1) each sentence that has an explicit citation to the reference.

The second set of features that we are interested in are discourse-based features. In particular we match each sentence with specific patterns and flag those that match. The first pattern is a bigram in which the first term matches any of "*this; that; those; these; his; her; their; such; previous*", and the second term matches any of "*work; approach; system; method; technique; result; example*". The second pattern includes all sentences that start with "*this; such*".

Finally, the similarity of each sentence to the reference is observable from the data and can be used as a sentence-level feature. Intuitively, if a sentence has higher similarity with the reference paper, it should have a higher potential of being in class 1 or $\mathcal{C}$. The flag of each sentence here is a value between 0 and 1 and is determined by its cosine similarity to the reference. Once the flags for each sentence, S_i are determined, we calculate normalized f_i as the unweighted linear combination of individual features. Based on f_is, we compute the potential function, ϕ, as shown in Table 6.

$\phi_i(x_c, y_c)$	$x_c = 0$	$x_c = 1$
	$1 - f_i$	f_i

Table 6: The node potential function ϕ for each node in the MRFs from the sentences in scientific papers is built using the sentences' flags computed using sentence level features.

5 Experiments

The intrinsic evaluation of our methodology means to directly compare the output of our method with the gold standards obtained from the annotated data. Our methodology finds the sentences that cite a reference implicitly. Therefore the output of the inference method is a vector, υ, of 1's and 0's, whereby a 1 at element i means that sentence i in the source document is a context sentence about the reference while a 0 means an explicit citation or neither. The gold standard for each paper-reference pair, ω (obtained from the annotated vectors in Section 3.1 by changing all $\mathcal{C}$s to 0s), is also a vector of the same format and dimensionality.

Precision, recall, and F_β for this task can be defined as

$$p = \frac{\upsilon \cdot \omega}{\upsilon \cdot \mathbf{1}}; \quad r = \frac{\upsilon \cdot \omega}{\omega \cdot \mathbf{1}}; \quad F_\beta = \frac{(1 + \beta^2)p \cdot r}{\beta^2 p + r} \qquad (3)$$

where $\mathbf{1}$ is a vector of 1's with the same dimensionality and β is a non-negative real number.

5.1 Baseline Methods

The first baseline that we use is an IR-based method. This baseline, $\mathbf{B_1}$, takes explicit citations as an input but use them to find context sentences. Given a paper-reference pair, for each explicit citation sentence, marked with $\mathcal{C}$, $\mathbf{B_1}$ picks its preceding and following sentences if their similarities to that sentence is greater than a cutoff (the median of all such similarities), and repeats this for neighboring sentences of newly marked sentences. Intuitively, $\mathbf{B_1}$ tries to find the best chain (window) around citing sentences.

As the second baseline, we use the hand-crafted discourse based features used in MRF's potential function. Particularly, this baseline, $\mathbf{B_2}$, marks each sentence that is within a particular distance (4 in our experiments) of an explicit citation and matches one of the two patterns mentioned in Section 4.3. After marking all such sentences, $\mathbf{B_2}$ also marks all sentences between them and the

paper	$\mathbf{B_1}$	$\mathbf{B_2}$	**SVM**	$\mathbf{BP_1}$	$\mathbf{BP_4}$	$\mathbf{BP_n}$
P08-2026	0.441	0.237	0.249	0.470	**0.613**	0.285
N07-1025	0.388	0.102	0.124	0.313	**0.466**	0.138
N07-3002	0.521	0.339	0.232	**0.742**	0.627	0.315
P06-1101	0.125	0.388	0.127	0.649	**0.889**	0.193
P06-1116	0.283	0.104	0.100	0.307	**0.341**	0.130
W06-2933	0.313	0.100	0.176	0.338	**0.413**	0.160
P05-1044	0.225	0.100	0.060	0.172	**0.586**	0.094
P05-1073	0.144	0.100	0.144	0.433	**0.518**	0.171
N03-1003	0.245	0.249	0.126	**0.523**	0.466	0.125
N03-2016	0.100	0.181	0.224	0.439	**0.482**	0.185

Table 7: Average $F_{\beta=3}$ for similarity based baseline ($\mathbf{B_1}$), discourse-based baseline ($\mathbf{B_2}$), a supervised method (**SVM**) and three MRF-based methods.

closest explicit citation, which is no farther than 4 sentences away. This baseline helps us understand how effectively this sentence level feature can work in the absence of other features and the network structure.

Finally, we use a supervised method, **SVM**, to classify sentences as context/non-context. We use 4 features to train the **SVM** model. These 4 features comprise the 3 sentence level features used in MRF's potential function (i.e., similarity to reference, explicit citation, matching certain regular-expressions) and a network level feature: distance to the closes explicit citation. For each source paper, P, we use all other source papers and their source-reference annotation instances to train a model. We then use this model to classify all instances in P. Although the number of references and thus source-reference pairs are different for different papers, this can be considered similar to a 10-fold cross validation scheme, since for each source paper the model is built using all source-reference pairs of all other 9 papers.

We compare these baselines with 3 MRF-based systems each with a different assumption about independence of sentences. $\mathbf{BP_1}$ denotes an MRF in which each sentence is only connected to 1 sentence before and after. In $\mathbf{BP_4}$ locality is more relaxed and each sentence is connected to 4 sentences on each sides. $\mathbf{BP_n}$ denotes an MRF in which all sentences are connected to each other regardless of their position in the paper.

Table 7 shows $F_{\beta=3}$ for our experiments and shows how $\mathbf{BP_4}$ outperforms the other methods on average. The value 4 may suggest the fact that although sentences might be independent of distant sentences, they depend on more than one sentence on each side.

The final experiment we do to intrinsically evaluate the MRF-base method is to compare different sentence-level features. The first feature used to build the potential function is explicit citations. This feature does not directly affect context sentences (i.e., it affects the marginal probability of context sentences through the MRF network connections). Therefore, we do not alter this feature in comparing different features. However, we look at the effect of the second and the third features: hand-crafted regular expression-based features and similarity to the reference. For each paper, we use $\mathbf{BP_4}$ to perform 3 experiments: two in absence of each feature and one including all features. Figure 4 shows the average $F_{\beta=3}$ for each experiment. This plot shows that the features lead to better results when used together.

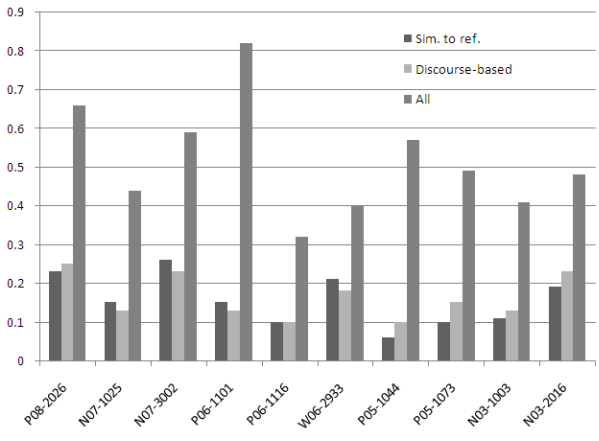


Figure 4: Average $F_{\beta=3}$ for $\mathbf{BP_4}$ employing different features.

... Naturally, our current work on question answering for the reading comprehension task is most related to those of (Hirschman et al. , 1999; Charniak et al. , 2000; Riloffand Thelen, 2000 ; Wang et al. , 2000). **In fact, all of this body of work as well as ours are evaluated on the same set of test stories, and are developed (or trained) on the same development set of stories.** The work of (Hirschman et al. , 1999) initiated this series of work, and it reported an accuracy of 36.3% on answering the questions in the test stories. **Subsequently, the work of (Riloffand Thelen , 2000) and (Chaxniak et al. , 2000) improved the accuracy further to 39.7% and 41%, respectively. However, all of these three systems used handcrafted, deterministic rules and algorithms**...
...**The cross-model comparison showed that the performance ranking of these models was: U-SVM $>$ PatternM $>$ S-SVM $>$ Retrieval-M.** Compared with retrieval-based [Yang et al. 2003], pattern-based [Ravichandran et al. 2002 and Soubbotin et al. 2002], and deep NLP-based [Moldovan et al. 2002, Hovy et al. 2001; and Pasca et al. 2001] answer selection, machine learning techniques are more effective in constructing QA components from scratch. **These techniques suffer, however, from the problem of requiring an adequate number of handtagged question-answer training pairs. It is too expensive and labor intensive to collect such training pairs for supervised machine learning techniques** ...
... **As expected, the definition and person-bio answer types are covered well by these resources.** The web has been employed for pattern acquisition (Ravichandran et al. , 2003), document retrieval (Dumais et al. , 2002), query expansion (Yang et al. , 2003), structured information extraction, and answer validation (Magnini et al. , 2002). **Some of these approaches enhance existing QA systems, while others simplify the question answering task, allowing a less complex approach to find correct answers** ...

Table 8: A portion of the QA survey generated by LexRank using the context information.

	citation survey	context survey
	QA	
CT nuggets	0.416	**0.634**
AB nuggets	0.397	**0.594**
	DP	
CT nuggets	0.324	**0.379**

Table 9: Pyramid $F_{\beta=3}$ scores of automatic surveys of QA and DP data. The QA surveys are evaluated using nuggets drawn from citation texts (CT), or abstracts (AB), and DP surveys are evaluated using nuggets from citation texts (CT).

6 Impact on Survey Generation

We also performed an extrinsic evaluation of our context extraction methodology. Here we show how context sentences add important survey-worthy information to explicit citations. Previous work that generate surveys of scientific topics use the text of citation sentences alone (Mohammad et al., 2009; Qazvinian and Radev, 2008). Here, we show how the surveys generated using citations and their context sentences are better than those generated using citation sentences alone.

We use the data from (Mohammad et al., 2009) that contains two sets of cited papers and corresponding citing sentences, one on Question Answering (QA) with 10 papers and the other on Dependency Parsing (DP) with 16 papers. The QA set contains two different sets of nuggets extracted by experts respectively from paper abstracts and citation sentences. The DP set includes nuggets extracted only from citation sentences. We use these nugget sets, which are provided in form of regular expressions, to evaluate automatically generated summaries. To perform this experiment we needed to build a new corpus that includes context sentences. For each citation sentence, $\mathbf{BP_4}$ is used on the citing paper to extract the proper context. Here, we limit the context size to be 4 on each side. That is, we attach to a citing sentence any of its 4 preceding and following sentences if BP4 marks them as context sentences. Therefore, we build a new corpus in which each explicit citation sentence is replaced with the same sentence attached to at most 4 sentence on each side.

After building the context corpus, we use LexRank (Erkan and Radev, 2004) to generate 2 QA and 2 DP surveys using the citation sentences only, and the new context corpus explained above. LexRank is a multidocument summarization system, which first builds a cosine similarity graph of all the candidate sentences. Once the network is built, the system finds the most central sentences by performing a random walk on the graph. We limit these surveys to be of a maximum length of 1000 words. Table 8 shows a portion of the survey generated from the QA context corpus. This example shows how context sentences add meaningful and survey-worthy information along with citation sentences. Table 9 shows the Pyramid $F_{\beta=3}$ score of automatic surveys of QA and DP data. The QA surveys are evaluated using nuggets drawn from citation texts (CT), or abstracts (AB), and DP surveys are evaluated using nuggets from citation texts (CT). In all evaluation instances the

surveys generated with the context corpora excel at covering nuggets drawn from abstracts or citation sentences.

7 Conclusion

In this paper we proposed a framework based on probabilistic inference to extract sentences that appear in the scientific literature, and which are about a secondary source, but which do not contain explicit citations to that secondary source. Our methodology is based on inference in an MRF built using the similarity of sentences and their lexical features. We show, by numerical experiments, that an MRF in which each sentence is connected to only a few adjacent sentences properly fits this problem. We also investigate the usefulness of such sentences in generating surveys of scientific literature. Our experiments on generating surveys for Question Answering and Dependency Parsing show how surveys generated using such context information along with citation sentences have higher quality than those built using citations alone.

Generating fluent scientific surveys is difficult in absence of sufficient background information. Our future goal is to combine summarization and bibliometric techniques towards building automatic surveys that employ context information as an important part of the generated surveys.

8 Acknowledgments

The authors would like to thank Arzucan Özgür from University of Michigan for annotations.

This paper is based upon work supported by the National Science Foundation grant "iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains", jointly awarded to U. of Michigan and U. of Maryland as IIS 0705832. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Shannon Bradshaw. 2002. *Reference Directed Indexing: Indexing Scientific Literature in the Context of Its Use*. Ph.D. thesis, Northwestern University.

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*.

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.

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

Dain Kaplan, Ryu Iida, and Takenobu Tokunaga. 2009. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95, Suntec City, Singapore, August. Association for Computational Linguistics.

Mary McGlohon, Stephen Bay, Markus G. Anderle, David M. Steier, and Christos Faloutsos. 2009. Snare: a link analytic system for graph labeling and risk detection. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1265–1274.

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

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479.

Donald Metzler and W. Bruce Croft. 2007. Latent concept expansion using markov random fields. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–318.

Donald A. Metzler. 2007. Automatic feature selection in the markov random field model for information retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 253–262.

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

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *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.

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.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL anthology network corpus. In *ACL workshop on Natural Language Processing and Information Retrieval for Digital Libraries*.

Matteo Romanello, Federico Boschetti, and Gregory Crane. 2009. Citations in the digital library of classics: Extracting canonical references by using conditional random fields. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 80–87, Suntec City, Singapore, August. Association for Computational Linguistics.

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? attributing scientific work to citations. In *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*, Sydney, Australia, July.

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

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2003. Understanding belief propagation and its generalizations. pages 239–269.