Citances: Citation Sentences for Semantic Analysis of Bioscience Text

Preslav I. Nakov
CS Division
UC Berkeley
Berkeley, CA 94720
nakov@cs.berkeley.edu

Ariel S. Schwartz
CS Division
UC Berkeley
Berkeley, CA 94720
sariel@cs.berkeley.edu

Marti A. Hearst
SIMS
UC Berkeley
Berkeley, CA 94720
hearst@sims.berkeley.edu

ABSTRACT

We propose the use of the text of the sentences surrounding citations as an important tool for semantic interpretation of bioscience text. We hypothesize several different uses of citation sentences (which we call citances), including the creation of training and testing data for semantic analysis (especially for entity and relation recognition), synonym set creation, database curation, document summarization, and information retrieval generally. We illustrate some of these ideas, showing that citations to one document in particular align well with what a hand-built curator extracted. We also show preliminary results on the problem of normalizing the different ways that the same concepts are expressed within a set of citances, using and improving on existing techniques in automatic paraphrase generation.

1. INTRODUCTION

The scientific literature of biomedicine, genomics, and other biosciences is a rich, complex, and continually growing resource. With appropriate information extraction and retrieval tools, bioscience researchers can use the contents of the literature to further their research goals. In recent years the interest in automatic tools for information extraction and retrieval from bioscience literature has increased considerably. Evidence for that trend is the addition of the genomics track to the Text Retrieval Conference (TREC) [37], and the new BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) competition [34].

As part of the BioText project [35] we are interested in utilizing the large volume of available bioscience text when designing information extraction and retrieval tools. For example, instead of analyzing each document separately, multiple related documents can be analyzed together, thus increasing the accuracy of tools for tasks such as entity recognition, relation extraction, synonym disambiguation, and automatic summarization. So far, most of the Natural Language Processing (NLP) work in the bioscience domain has been done on MEDLINE abstracts. However, full text is becoming more available, providing new opportunities for automatic text processing. One such opportunity lies in the text around citations in full text papers.

In this paper we put forward a new vision for the path towards robust and large-coverage algorithms for semantic interpretation of bioscience articles. We suggest using the sentences that surround the citations to related work as the data from which to build semantic interpretation models. We also introduce a neologism, *citances*, to mean the sentence(s) surrounding the citation within a document.

Citations are used in every scientific literature, but they are particularly abundant in biosciences. Nearly every statement is backed up with at least one citation, and, conversely, it is quite common for papers in the bioscience domain to be cited by 30-100 other papers. The text around citations tends to state known biological facts with reference to the original papers that discovered them. The cited facts are typically stated in a more concise way in the citing papers than in the original papers. As the same facts are repeatedly stated in different ways in different papers, statistical models can be trained on existing citances to identify similar facts in unseen text.

With the availability of full text articles, and the nature of citation in bioscience literature, traditional citation analysis work can be greatly expanded. We believe that citations have great potential to be a valuable resource in mining the bioscience literature. In particular, we identify the following promising applications of citation analysis:

- **A source for unannotated comparable corpora**. *Comparable corpora*, which are typically generated from news articles on related events, are a useful resource for the development of NLP tools for question answering [20] and summarization [3]. Most domains outside of news do not contain many articles discussing the same events, but bioscience citances have some of the requisite characteristics in that they include redundancies that allow identification of comparable sentences. In the case of news articles, dates and named entities help link related sentences. In Section 4 we demonstrate the use of citances as comparable corpora for automatic paraphrase extraction.

- **Summarization of the target papers**. The set of citances that refer to a specific paper can be viewed as an indication of the important facts in the paper as seen by the scientific community in that field. This is an excellent resource for summarization. In fact, we believe that a paper that is cited enough times can be summarized using only the citances pointing to it. Instead of showing the user all the citances pointing to a paper (as is done in CiteSeer and in Nanba et al. [25]), we propose to first cluster related citances, and then display to the user only a summary of each cluster. The facts expressed by each cluster can be extracted and stored in a database in a normalized form. This could facilitate answering advanced queries on facts, such as "retrieve all documents that describe which genes upregulate gene G".
- **Synonym identification and disambiguation**. Bioscience literature is rife with abbreviations and synonyms. Citances referring to the same article may allow synonyms to be identified and recorded. On the flip side, in many cases the same terms have multiple meanings. Again, a collection of related citances can help disambiguate these meanings, since in some of the citances an unambiguous form of the term might be present.
- **Entity recognition and relation extraction**. Citances in bioscience literature are more likely to state biological facts than arbitrarily chosen sentences in an article. They also tend to be more concise, since the authors try to summarize previous related work, which has already been described in detail in the original paper. Language presents a myriad number of ways to express the same or similar concepts. Citances provide us a way to build a model of many of the different ways to express a relationship type R between entities of type A and B. We can seed learning algorithms with several examples using concepts that are semantically similar to A and similar to B, for which relation R is known to hold. Then we can train a model to recognize this kind of relation for situations for which the relation is not known. Since the results may extend to sentences that are not citances as well, citances-based corpora should provide a good collection for building NLP tools for recognizing entities and relations in unseen text.
- **Targets for curation**. We hypothesize that citances contain the most important information expressed in the cited document, and therefore contain the information that curators would want to make use of. We have found support for this hypothesis with two sample papers being used by a cancer researcher who is recording information about the process of apoptosis.
- **Improved citation indexes for information retrieval**. In addition to supporting advance queries over facts as just described, citation indexes can be improved by combining methods that use citances' *context* (e.g., Mercer and Di Marco [23]) with methods that use citances' *content* (e.g. [7]). For example, indexing terms can be taken from citances referring to a target paper, weighting them both by their relative frequency and the type of citations they appear in.

This section has defined and motivated the use of citances for semantic processing of bioscience text. In the next sections we first describe related work in the analysis of citation sentences, and then describe some of the challenges in processing such sentences. This is followed by a description of an algorithm for paraphrasing citances that discuss the same relationship between entities, its evaluation and relationship to related work on paraphrases extraction. Finally, we conclude with future work.

2. RELATED WORK

White [32] provides a good recent review of the field of citation analysis (for a more thorough but less recent review of the field see [22]). White describes three major lines of research in the field of citation analysis. All three focus on (mostly manual) analysis of citations based on the text around them.

First, *citation categorization* schemes date back to the 1960's [14, 21]. Citations are placed into categories such as *conceptual vs. operational, organic vs. perfunctory, evolutionary vs. juxtapositional, and confirmational vs. negative* [24]. The number of categories and their definitions vary between different classification schemes.

Second, *context analysis* is concerned with identifying recurring terms in citances, and potentially using them as subject headings for indexing purposes. Our proposal can be seen as an extension of this approach to the level of facts, such as relationships between entities, rather than being limited to keywords.

The third research area identified by White is the classification of *citer motivation*, identifying the reason authors cite earlier work, and the reasons some works are cited more often than others. This area is based mostly on sociological studies.

In addition to citation analysis, citations are used in *citation indexing* systems, which were first proposed in 1955 by Garfield [13], and are now in wide use in systems like ISI's SCI and CiteSeer [15]. A citation index aims to disambiguate the bibliographic references in scientific literature, making explicit the links between articles, which are formed by these references. It allows information retrieval tools to cluster the related articles, and to estimate the importance of a paper by counting the number of articles citing it. Citation indexes also allow users to navigate the scientific literature by following the links between articles going forwards and backwards in time. This feature is especially useful when looking for related work, or when learning about a new topic.

Recently, Mercer and Di Marco have described their work on using citances to improve indexing tools for biomedical literature [23]. They are mostly concerned with automatically classifying citations into a predefined classification scheme using cue phrases in citances. They propose to use these classifications to improve existing citation indexes that currently ignore the type of the citations in their algorithms. They do not use the terms in the citances directly to improve information retrieval.

Bradshaw [6, 7] proposes to improve information retrieval of scientific literature using a metric on citations called Reference Directed Indexing (RDI). RDI indexes articles based on the terms used in the citances citing them. RDI gives higher weight to terms that are more common in the citances to the target specific document compared to citances to other documents. When ranking retrieval results, RDI takes into account both the relevance of a document to the query terms, and the number of papers citing it. RDI treats all citations equivalently without using a classification scheme.

Teufel and Moens [31] identify and classify citations in scientific articles. They use the identified citances to improve summarization performance by using them as a feature when classifying candidate sentences in the citing paper, giving lower weight to citances as compared to other sentences. They do not use citances in the citing papers to summarize the cited paper, as we propose.

Nanba et al. [25] use citances as features for classifying papers into topics. They also propose to use citances as part of a support system for writing review articles on specific topics. Given a document, their system displays the citances originating from other papers. However, when many citances are uncovered for the same document the summary will be too large.

A related field to citation indexing is the use of link structure and anchor text of Web pages. Anchor text is used in search engines such as Google [8] for indexing and retrieval of Web pages. Applications of anchor text include identification of home pages of people and companies [11], classification of Web pages [10, 12], Web crawlers [27], and improved ranking of search results [28]. Amitay and Paris [1] present a system for Web page summarization using sentences around links to the target Web page. Their system picks a single representing snippet as a description of the target Web page, helping users to follow the best search result. For a more extensive review on the use of anchor text see [7].

3. ISSUES FOR PROCESSING CITANCES

Several issues must be addressed in order to effectively use citances in various applications.

- **Text span.** The span or scope of the text that should be included with the citation must be determined. The appropriate span can be a phrase, a clause, a sentence, or several sentences or their fragments. Furthermore, citations themselves must be parsed, as they can be shown as lists or groups (e.g., “[22-25]”).
- **Identifying the different topics.** The different reasons a given paper is cited must be determined, and citances that cite a document for a similar reason must be grouped together.
- **Normalizing or paraphrasing citances.** Once the citances with the same meaning are grouped together, they will convey essentially the same information in different ways, or express different subsets of the same information. Thus it is important to be able to “normalize” or paraphrase the citances for many applications, including indexing in a database or an IR system, document summarization [4, 3], learning synonyms [16, 17], building a model of the different expressions of the same relationship for IE [30, 29], extracting patterns for question answering [20], machine translation [26].

In the next section we describe our early experiments in addressing the normalization problem, as well as sketching a preliminary attempt at the topic grouping problem.

4. PARAPHRASING CITANCES

4.1 Example Paraphrases

Our strategy is to extract paraphrases expressing roughly the same relation between two named entities, such as gene / protein names or MeSH terms. For the sample sentences below, the target entities are *BIM* and *NGF* (nerve growth factor). We also need to identify a “gold standard” or target sentence to which we want to convert the citances via paraphrase. We begin with the target sentence *Bim is induced after NGF withdrawal.*

Now consider the following citances which refer to the target paper [33], ordered according to how well they reflect the meaning of the target sentence, where the part that matches the target relation is underlined:

1. *NGF* withdrawal from sympathetic neurons induces *Bim*, which then contributes to death.
2. *Nerve growth factor* withdrawal induces the expression of *Bim* and mediates Bax dependent cytochrome c release and apoptosis.
3. Recently, *Bim* has been shown to be upregulated following both *nerve growth factor withdrawal* from primary sympathetic neurons, and serum and potassium withdrawal from granule neurons.
4. The proapoptotic Bcl-2 family member *Bim* is strongly induced in sympathetic neurons in response to *NGF* withdrawal.
5. In neurons, the BH3 only Bcl2 member, *Bim*, and JNK are both implicated in apoptosis caused by *nerve growth factor* deprivation.

Below are shown the paraphrases that should be extracted from these sentences:

1. *NGF* withdrawal induces *Bim*.
2. *Nerve growth factor* withdrawal induces the expression of *Bim*.
3. *Bim* has been shown to be upregulated following *nerve growth factor* withdrawal.
4. *Bim* is induced in sympathetic neurons in response to *NGF* withdrawal.
5. *Bim* implicated in apoptosis caused by *nerve growth factor* deprivation.

Here we adopt a “liberal” definition of *good* paraphrase, which does not require an exact meaning equivalence, but allows for minor variations, provided that no elements are removed or added (as in Paraphrases 1-3). We consider a candidate paraphrase *acceptable*, if it adds more details than expected, such as modifiers and prepositional phrases, but otherwise expresses roughly the same meaning (as seen in Paraphrase 4). In all other cases, we consider the candidate paraphrase to be *bad* (as in Paraphrase 5, which does not talk about induction or upregulation but instead is rather vague about the role of Bim). In the next subsections we describe how these paraphrases are extracted from the citances and present a preliminary evaluation of the results for one set of citances.

4.2 Paraphrase Extraction Algorithm

Our paraphrase extraction algorithm is a variation of that proposed by Lin and Pantel [20], at whose core is a dependency parse, and Shinyama et al. [30], who extend this idea to use specific named entities to anchor the paraphrase:

Algorithm:

```
Given a target paper cited by multiple other papers:

1. Extract the sentences that cite the target.
2. Mark the NEs of interest (genes/proteins and MeSH terms)
   and normalize.
3. Parse using a dependency parser.
4. For each sentence
       For each pair of NEs of interest
           i. Extract the path between them.
          ii. Extract a paraphrase from the path.
5. Rank the candidates for a given pair of NEs.
6. Select only the ones above a threshold.
7. Generalize.
```

We discuss steps 1-4 only.[1] The dependency parse was produced by *Minipar* [19], which builds a rooted directed tree where each node (not just the leaves, as in a constituency parse) is associated with a word from the sentence and annotated with its part of speech. The directed links represent grammatical relations between nodes. Figure 1 shows a sample dependency parse for the sentence "*NGF withdrawal from sympathetic neurons induces Bim.*". The grammatical relations (e.g., subject, determiner, modifier etc.) are not shown as, unlike [20], we do not use them as constraints. We extract a paraphrase from the simple dependency path between the two target named entities (there is exactly one path, since this is a tree). In Figure 1, the path is shown in bold and omits the unnecessary prepositional phrase "*from sympathetic neurons*". As the arrows show, the path starts at *NGF*, goes up to the root (*induces*), and then descends to *Bim*. Here the root is a verb, but this is not always the case: sometimes it is a noun or an artificial non-word entity, which is created by *Minipar* when it tries to perform co-reference resolution.

The extraction of a paraphrase from the dependency path, involves two steps:

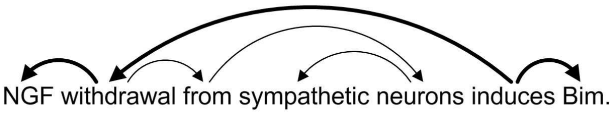

Figure 1: Sample dependency parse. The dependency path between the target NEs is shown in bold.

1. Arranging the words from the dependency path linearly, according to the original order in the sentence;

2. Including additional words to improve the grammaticality.

The second step is needed because the simple dependency path often omits words that then render it ungrammatical. In particular, there is a problem with the complex verb forms (e.g., passive, infinitive, past tense etc.). Unlike Lin and Pantel [20] and Ibrahim et al. [18], who manipulate the parser's output to account for some of these before proceeding to path extraction, we use the following "2-word" heuristic:

If the path extracted from the dependency parse skips over either one or two words, those one or two words are inserted back into the paraphrase, unless those words are adverbs.

The number 2 was chosen because most of the verb forms contain up to 3 words. This heuristic appears to work well in practice and accounts for a variety of other cases, e.g., omission of prepositions, determiners, etc. However, it needs to be refined further as sometimes it includes more details than necessary, mainly in the form of additional adjectives. For the five example citances above, we obtained the following paraphrases (words in square brackets [] have been added by the 2-word heuristic):

1. NGF withdrawal induces Bim

2. Nerve growth factor withdrawal induces [the] expression of Bim

3. Bim [has] [been] shown [to] be [upregulated] following nerve growth factor withdrawal

4. Bim [is] induced in [sympathetic] neurons in response to NGF withdrawal

5. member Bim implicated in apoptosis caused by nerve growth factor deprivation

All are grammatical except for the last one, which contains a redundant starting word: *member* (we will consider this issue below). Note the third example, where the 2-word heuristic added all missing verb forms, including the important passive *upregulated*.

The gene/protein *Bim* and the MeSH term *neural growth factor* are marked prior to parsing, using BioText [35] group tools for the TREC 2003 Genomics track [5].

[1] We have done some experiments with rank ordering the paraphrases according to their similarity to the target sentence, but space limitations prevent discussion of those results here.

4.3 Evaluation

For our experiments, we chose an influential journal paper from *Neuron* [33] and collected 99 journal papers that cited it, where we were able to identify 203 citances in total. An expert in apoptosis identified 36 different types of important biological factoids that could be extracted from the target paper. A person with a background in genomics then examined the 203 citances and identified which of the 36 categories each citance could be associated with, and the degree to which it "covered" the factoid. Additionally, a citance could support more than one factoid.

In the current experiments, we used as our model (target) sentence the factoid for which the most citances were identified (*Bim is induced after NGF withdrawal*). The corresponding set of 67 citances defines our *Set 1*. We plan in future to examine paraphrases for all of the remaining categories.

We defined another subset of the 203 citances, *Set 2*, which is limited to the ones that contain both *Bim* and *NGF* (or their variant(s), identified using Biotext group tools). Finally, we built *Set 3* by extracting 102 sentences that contained both *Bim* and *NGF* and were not necessarily citances, resulting in the following three sets of sentences:

- **Set 1:** 67 citances pointing to the target paper and manually found to contain a good or acceptable paraphrase of *"Bim is induced after NGF removal."* (but do not necessarily contain *Bim* or *NGF*);
- **Set 2:** 65 citances pointing to the target paper and containing both *Bim* and *NGF*;
- **Set 3:** 102 sentences extracted from the 99 texts and containing both *Bim* and *NGF*.

Set 1 was introduced to assess the system under ideal conditions, which is an upper bound on the performance of a filtering limited to citances. The system's performance on *Set 2* vs. *Set 3* allows us to test our hypothesis that citances will produce more accurate paraphrases than general sentences.

We ran the paraphrase extraction algorithm on the 3 sets. Some longer sentences produced more than one paraphrase, as either *Bim* or *NGF* (or both) had been mentioned multiple times, but we kept only one paraphrase from each sentence[2], obtaining 55, 65 and 102 paraphrases, accordingly. The results of the annotations are shown in Table 1.

4.3.1 Correctness

All paraphrases were manually investigated and judged on their *correctness* (good: 1; acceptable: .5; bad: 0) and *grammaticality* (grammatical: 1; almost grammatical: .5; non-grammatical: 0). A paraphrase was judged as *bad* under the following conditions: (1) different relation between *Bim* and *NGF* than in the model (often *phosphorylation* aspect); (2) opposite meaning; or (3) vagueness (wording not clear enough for conclusion).

	correctness				**grammaticality**			
set	**1.0**	**0.5**	**0.0**	%	**1.0**	**0.5**	**0.0**	%
1	20	25	10	81.82	22	19	14	74.55
2	20	25	20	69.23	26	22	17	73.58
3	25	38	39	61.76	42	22	38	62.75
cluster	16	15	11	73.81	20	12	10	76.19

Table 1: Correctness and grammaticality of the paraphrases extracted from the 3 sets and from the clustering. % indicates percent good or acceptable.

A paraphrase was judged *acceptable*, if it was not *bad* and: (1) it contained additional terms (e.g., *DP5 protein*) or topics (e.g., prepositional phrases like *in sympathetic neurons*); or (2) the relation was suggested in the statement but not definitely. The correctness judgements were done by the same person with biological background as the citations assignment.

Table 1 shows that 81.82% are labeled good or acceptable when using citances known to have paraphrases, and 69.23% are labeled good or acceptable from the *Set 2* citances. This is compared to 61.76% when sentences at large are used. Table 1 shows there is a significant drop in correctness when going from *Set 2* to *Set 3* (see the first column labeled '%'), which supports our hypothesis that citances help focus the paraphrases.

The system's recall is easiest to calculate on *Set 1*, where 60 paraphrases have been extracted out of the 67 citances. Five sentences produced two different paraphrases, which yields a sentence-level recall of 55/67, i.e. 82.09%. Most of the misses were judged *acceptable* rather than *bad*. All 12 misses were due to unrecognized variants of the term *NGF*: mostly contextual hyponyms (e.g. *neurotrophin*), hypernyms (e.g. *growth factor*, *factor*, *serum*) or related terms (e.g. *survival factors*). In fact, missing a target term is the only possible reason for not producing a candidate paraphrase as the system always generates a dependency path when the target NEs are discovered and the sentence is parsed correctly. An example of such missed sentence is (no *NGF* or its variants have been found):

Growth factor withdrawal was shown to cause increased Bim expression in various populations of neuronal cell types.

It is interesting to note though that 10 of the 67 relevant citances in *Set 1* were initially missed by the human annotator and were added only later (8 of these were subsequently judged *good*, 2 were *acceptable*). Thus the human recall on this set is 57/67, i.e., about 85.07%, which compares very favorably to the 82.09% for the system. The missed sentences were generally quite complex and the Bim-NGF relation was not central and thus they were easy to overlook. An example of such a missed sentence is:

The precise targets of c-Jun necessary for the induction of apoptosis have been the subject of intense interest and re-

[2] We made the best choice according to a score, which takes into account whether the root of the dependency path is a verb, and the value of a gap-weighted POS sequence kernel [9] comparison of the dependency path to that of the model sentence.

cently, Bim and Dp5, both "BH3-domain only" family members, have been identified as pro-apoptotic genes induced in a c-Jun-dependent manner in both sympathetic neurons subjected to NGF withdrawal and in cerebellar granule cells deprived of KCl.

While containing an *acceptable* paraphrase, it is also an example of a case where no *good* paraphrase can be possibly extracted: there is simply no subsequence of words that can be judged *good*. Even if we can figure out how to automatically remove the prepositional phrase (PP) "*in sympathetic neurons*", we still cannot get a good paraphrase without damaging the grammaticality, as this PP is needed for the PP "*to NGF withdrawal*" to attach itself to.

The correctness of the paraphrases for *Set 1* is much better than that for the *Set 2*. This is due to the fact that citances to the same paper do not necessarily express the same facts even if they include the target entities. Since manual annotation of every citance, as was done for *Set 1*, is impractical for larger collections, we would like to automatically group citances by their semantics. The *cluster* row in Table 1 shows the results of an initial attempt in this direction. We can see that the correctness is improved as compared to *Set 2*.

To cluster the 203 citances (we cluster all of them, not just those from *Set 1*, *Set 2* or *Set 3*) we use an in-house tool that identifies gene names in text and maps them to their ID in LocusLink [36]. The affinity between two citances $\hat{k}(c_1, c_2)$ is then defined using a polynomial kernel:

$$\hat{k}(c_1, c_2) = (k(c_1, c_2) + R)^d,$$

where $k(c_1, c_2)$ is the number of identical genes in the two citances, R is non-negative, and d is a positive integer. We use a *spectral clustering* algorithm [2] to cluster the citances. The *cluster* dataset is obtained by selecting the citances in clusters for which more than 80% of the citances include both *NGF* and *Bim*.

4.3.2 Grammaticality

The grammaticality was judged by a native speaker. The last column of Table 1 shows that *Set 1* and *Set 2* were judged more grammatical than *Set 3*. This is partially due to the better sentence extraction for the first two sets, where more conservative regular expressions were used, while *Set 3* included truncated or merged sentences, or sentences coming from titles, which produced ungrammatical paraphrases.

We discovered multiple repeating sources of bad grammaticality. For example, the *Minipar* parser does not include the coordinating *and* in the parse and so *and* cannot be placed into the path. This led to errors in which two noun phrases are run together, as in: "*Hrk/DP5 Bim [have] [been] found [to] be upregulated after NGF withdrawal*". Correcting the parser's output would fix this problem.

Other problems are caused by the fact that a verb-rooted path between the target terms includes exactly two arguments for that verb. While most of the time ignoring any additional arguments is desirable, there are cases when it leads to ungrammatical sentences in which the subject or object of a verb can be missing, for example: "*caused by NGF role for Bim*". Repairing this problem would require knowledge about the possible sub-categorization frames of the target verb (as is done in [29]).

Another common problem is the inclusion of extra subject words, e.g., *member* in "*member Bim implicated in apoptosis caused by NGF deprivation*", due to its dependency on Bim in the original sentence: "*In neurons, the BH3-only Bcl2 member, Bim, and JNK are both implicated in apoptosis caused by NGF deprivation.*".

4.4 Previous Paraphrase Work

Most work on automatic paraphrasing is relatively recent. We identify four classes of related work: word-level, phrase-level, template-based, and sentence-level paraphrases.

Word-level paraphrases. Grefenstette uses a semantic parser to explore the local context surrounding a word and to compare the distributional similarity of such contexts to learn word synonyms ([16], [17]). The assumption is that similar contexts tend to contain similar words. These are not necessarily synonyms though, e.g. *cat* and *dog*.

Phrase-level paraphrases. Barzilay and McKeown[4] use a corpus of multiple translations of the same text and part of speech (POS) information from the local context surrounding the target words to extract word- and contiguous phrase-level paraphrases, e.g. (*countless*, *lots of*). They use co-training (fix the known paraphrases and collect their contexts, then fix the contexts and try to find more paraphrases, then repeat again etc.) to train a classifier that uses the local surrounding context to decide whether two phrases are paraphrases or not.

Template paraphrases. Lin and Pantel [20] use an idea similar to that of Grefenstette, that words used in the same contexts tend to have a similar meaning, but apply it to dependency tree paths (extracted with the *Minipar* parser). The paths are generalized by converting their ends to slots and are considered similar if these slots tend to contain similar sets of words. A single large text corpus is used to extract the template rules, e.g. "Y is resolved by X", "X resolves Y", "X finds a solution to Y" and "X tries to solve Y" are example paraphrases for "X solves Y". Many limitations are imposed on the kinds of paths considered.

Shinyama et al. [30] produce similar kinds of templates for Japanese. They use an IR system to find newspaper texts on a given topic and then find pairs of articles describing the same event. Then the named entities are tagged and used to align sentences, from which paraphrases are extracted using a dependency parse with the approximately matched NEs as anchors. Lastly, the NEs are generalized as variables like LOCATION, ORGANIZATION etc., as defined by their NE recognition system.

The same idea is further refined in a later work by Shinyama and Sekine [29] where a very limited form of coreference resolution is added and some structural restrictions on the possible portions of expressions to be extracted are applied. In addition, more than two anchors per sentence pair are allowed during matching.

A variation of the same approach is described by Ibrahim et al. [18], who use multiple translations of the same text and extract paraphrases from sentence pairs. The sentences are parsed and paraphrases are extracted from pairs of aligned sentences, when the paths have compatible slots and depending on the distribution of the slot values.

Sentence-level paraphrases. Barzilay and Lee [3] use comparable texts (news from Reuters and AFP) and multiple sequence alignment algorithms to learn paraphrases represented as word lattices. Cross-corpus sentence pairs written on the same day and on the same topic are compared in terms of word overlap. The approach however tends to produce ungrammatical sentences. Pang et al. [26] use multiple human translations of Chinese documents into English. They perform syntactic parsing and try to merge parse trees into a single finite-state transducer similar to the lattices built by Barzilay and Lee [3]. While no slots are generated, the lattices thus produced capture several different generalizations and can be used to generate a large number of sentences given only a small number of training examples.

4.5 Relationship to Previous Paraphrase Work

In this work we focused on the extraction of *grammatical* template-level paraphrases similar to those described by Lin and Pantel [20], and further refined by [18], [29] and [30]. The main differences from this earlier work is:

Citances. We focus the extraction by using multiple citations of the same fixed target document. The assumption is that there are a limited number of important facts in the target that it could be cited for.

Extracting complex paths. Unlike the above mentioned methods we do not impose any limitations on the path, which can get quite complex.

Focus on grammatical paraphrases. We extract *grammatical* paraphrases rather than just templates that can be matched against text. Thus, we have a postprocessing step, which adds some additional words that were not in the dependency path.

Use of lexical resources. This allows for the identification of the entities of interest such as genes/proteins and MeSH terms. We believe these are much better as anchors than simply compatible nouns or noun phrases. In addition, they are the natural candidates for slots. Further, using a lexical hierarchy allows for different levels of generalization – the appropriate one can be chosen to be consistent with e.g. all observations in the text.

Biomedical domain. Finally, we work in the biomedical domain; most other work used newswire text.

Overall, the results show the grammaticality of the extracted paraphrases is high and almost uniform across the three sets, which is not surprising. At the same time, the paraphrases extracted from citances have a correctness of over 20% higher as compared to *Set 3*. However, this is a small preliminary study; in future we need to evaluate the algorithm on more examples.

5. CONCLUSIONS

We have motivated and discussed the potentially enormous role that the use of sentences surrounding citations, or *citances*, can have for automated analysis of bioscience literature. In work not yet reported, we have found that citances align very well with rich information being curated by hand by a molecular biologist, and suspect they will be equally useful for other curation tasks. We also hypothesize that it will be a gold mine of data for training algorithms to perform semantic analysis of bioscience text, and will improve the results of querying the bioscience literature.

Much work must be done before citances can be put to full use. We have demonstrated some initial results in paraphrasing citances that discuss the same topic, but more work remains to be done to improve results, and to group similar citances together. In future work, we plan to thoroughly explore the possibilities surrounding the analysis and use of citances for bioscience text analysis.

6. ACKNOWLEDGEMENTS

We would like to thank Janice Hamerja for performing the sentence annotations and to Dr. Tingting Zhang for supplying the factoids for [33]. This research was supported by NSF grant DBI-0317510 and a gift from Genentech.

7. REFERENCES

[1] E. Amitay and C. Paris. Automatically summarising web sites: is there a way around it? In *Proceedings of the ninth international conference on Information and knowledge management*, pages 173–179. ACM Press, 2000.

[2] F. R. Bach and M. I. Jordan. Learning spectral clustering. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[3] R. Barzilay and L. Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL.*, pages 16–23, 2003.

[4] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL.*, pages 50–57, 2001.

[5] G. Bhalotia, P. Nakov, A. Schwartz, and M. Hearst. Biotext team report for the trec 2003 genomics track. In *Proceedings of TREC*, 2003.

[6] S. Bradshaw. Reference directed indexing: Indexing scientific literature in the context of its use. ph.d. dissertation. In *Northwestern University (Tech Report NWU-CS-02-7)*, 2002.

[7] S. Bradshaw. 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*, 2003.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *COMPUTER NETWORKS AND ISDN SYSTEMS*, 1–7:107–117, 1998.

[9] N. Cancedda, Éric Gaussier, C. Goutte, and J.-M. Renders. Word-sequence kernels. *Journal of Machine Learning Research*, 3:1059–1082, 2003.

[10] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 65–74. Elsevier Science Publishers B. V., 1998.

[11] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–257. ACM Press, 2001.

[12] J. Fürnkranz. Exploiting structural information for text classification on the www. In *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, pages 487–498. Springer-Verlag, 1999.

[13] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

[14] E. Garfield. Can citation indexing be automated? *National Bureau of Standards Miscellaneous Publication*, 269:189–192, 1965.

[15] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM Press, 1998.

[16] G. Grefenstette. Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of ACL*, pages 324–326, 1992.

[17] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, 1994.

[18] A. Ibrahim, B. Katz, and J. Lin. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of Second International Workshop on Paraphrasing (IWP 2003)*, pages 57–64, 2003.

[19] D. Lin. Dependency-based evaluation of minipar. In *Proceedings of Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation.*, 1998.

[20] D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360, 2001.

[21] B. A. Lipetz. Improvements of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16:81–90, 1965.

[22] M. Liu. Progress in documentation. the complexities of citation practice: A review of citation studies. *Journal of Documentation*, 49(4):370–408, 1993.

[23] R. E. Mercer and C. D. Marco. A design methodology for a biomedical literature indexing tool using the rhetoric of science. In *BioLink workshop in conjunction with NAACL/HLT*, pages 77–84, 2004.

[24] M. J. Moravcsik and P. Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92, 1975.

[25] H. Nanba, N. Kando, and M. Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *American Society for Information Science SIG Classification Research Workshop: Classification for User Support and Learning*, pages 117–134, 2000.

[26] B. Pang, K. Knight, and D. Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT-NAACL*, pages 181–188, 2003.

[27] J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 335–343. Morgan Kaufmann Publishers Inc., 1999.

[28] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

[29] Y. Shinyama and S. Sekine. Paraphrase acquisition for information extraction. In *Proceedings of Second International Workshop on Paraphrasing (IWP2003)*, 2003.

[30] Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, pages 40–46, 2002.

[31] S. Teufel and M. Moens. Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[32] H. D. White. Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25(1):89–116, 2004.

[33] J. Whitfield, S. Neame, L. Paquet, O. Bernard, and J. Ham. Dominantnegative c-jun promotes neuronal survival by reducing bim expression and inhibiting mitochondrial cytochrome c release. *Neuron*, 29:629–643, 2001.

[34] http://www.mitre.org/public/biocreative/.

[35] http://biotext.berkeley.edu/.

[36] http://www.ncbi.nlm.nih.gov/locuslink/.

[37] http://trec.nist.gov/.