Extraction of Semantic Relationships from Academic Papers using Syntactic Patterns

Akihiro Kameda, Kiyoko Uchiyama, Hideaki Takeda, Akiko Aizawa
National Institute of Informatics
Tokyo, Japan.
{*kameda, kiyoko, takeda, aizawa*}*@nii.ac.jp*

***Abstract*—Integrating concept and citation networks on a specific research subject can help researchers focus their own work or use methods described in prior works. In this paper, we propose a method to extract semantic relations from concepts and citation in the descriptions of related work. Specifically, we examined (i) topic-paper relations between research topics and reference papers and (ii) method-purpose relations between research topics. We also defined 15 lexico-syntactic patterns for the relation extraction. Results of experiments using a manually annotated dataset of 15 papers demonstrated the effectiveness of using the proposed lexico-syntactic patterns.**

Keywords*-*Relation extraction; Citation context; Knowledge extraction.

I. INTRODUCTION

Most researchers locate, read, and analyze relevant papers to investigate prior studies related to their research fields when they are about to narrow their subject of research or publish their findings in a paper. The objective of such prior study searches is to position their own work among problems in related works or their surroundings and to clarify their own predominance. This type of background work is crucial in terms of qualifying a paper for publication.

In this paper, we examine related work descriptions in academic papers to support such research activities. Our approach is based on a knowledge representation that integrates the citation structure and research concept network of related topics. The proposed representation contains two kinds of node – a Paper and a Concept nodes – and links that connect these nodes. In order to extract comprehensible mutual relations, we define a concept as a text span that is associated with a specific paper in the citation context. These text spans may include different linguistic units such as technical terms, verb phrases, and clauses. For example, to a research the theme "modeling text and citation together in a similar corpus", technical terms such as "PHITS" and "PLSA", noun phrases such as "influence propagation model", and verb phrases such as "explain various phenomena related to linked structure of the corpus" are recognized as Concept Nodes (Fig. 1).

Related work descriptions include information such as methods used in prior studies, other themes addressing the use of these methods, and the original papers in which the methods were proposed. As such, the generated network of preceding studies can help researchers position their own studies properly into surrounding fields and related works. In the following, we propose our method of constructing a paper knowledge network and report the results of experiments conducted with a dataset of academic papers written in English.

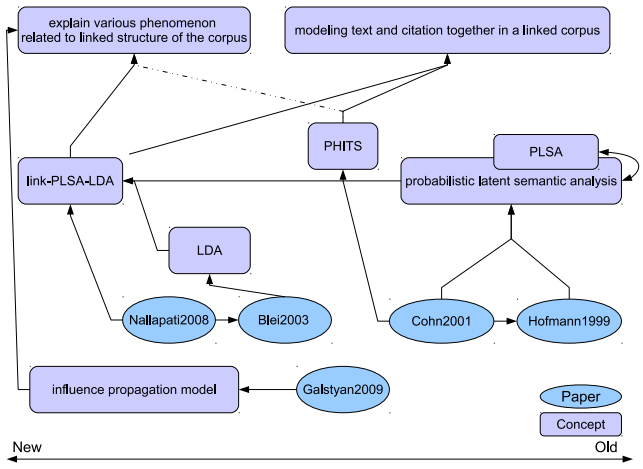

Figure 1. Example of a network representation of papers and concepts.

II. RELATED WORK

Zhang et al. [1] proposed a method to extract the relations of the key concepts of academic papers based on the clustering results. In their method, papers were first clustered and then keywords representative to each cluster were extracted. Then, a hierarchical structure of the keywords was constructed. While their study considered only the existence of a keyword in a target paper, our study looks into more detailed relations, such as whether a certain keyword is mentioned as a method or a research target.

Regarding the refinement of relations among papers, Nanba et al. [2] analyzed sentences that contain references to other papers and classified them into three categories: those citing other papers as their basis, those pointing out relative differences, and otherwise.

 ISBN: 978-1-61208-254-7

There is another similar framework in a study by Teufel et al.[3], where relations among papers were extracted based on the classification of the citation context. Dunne [4] also proposed a method to generate a targeted field overview by clustering the citation network. A devised interface that displays the summary of a citation context simultaneously enables users to obtain a detailed understanding of the topic. However, none of these studies investigated the semantic relations among different citation contexts. It remains unclear which portion of the sentence corresponds specifically to the concept such as a name of method or a expression of purpose appearing in each sentence.

III. Network Representation

A. Paper and Concept Nodes

We prepared two types of node to represent a network structure: *Paper* and *Concept* nodes. Concept nodes are labeled by character strings such as technical terms and noun and verb phrases, while paper nodes are identified by their own URIs or DOIs in a digital library. These nodes are extracted on the basis of lexico-syntactic patterns explained in section IV-B. Both types of node can connect with each other and represent various relations, as follows.

B. Method-Purpose Relation

We examined the relation between a method and its purpose and application in building a knowledge network from papers, hereafter referred to as *the Method-Purpose relation*. This relation consists of Concept nodes. An example sentence is

> *Similar observations have also been made in [38] where Probabilistic Latent Semantic Indexing (PLSI) was used to learn a lower dimension representation of text in terms of probabilistic topics.*

The text provides a relation in that "Probabilistic Latent Semantic Indexing" contributes "to learn a lower dimension representation of text in terms of probabilistic topics."

C. Paper-Topic Relation

A relation to connect a concept with the paper in which it is mentioned is referred to as *the Paper-Topic relation*. For instance, in the above example sentence, the relation with which the concept of "Similar observations" is addressed in reference "[38]" is extracted as the Topic-Paper relation. In this case, coreference resolution should be done for extracting truly semantic relation because the word "Similar observations" by itself is not sufficient for semantics. However, that relation is regarded as correct relation in this paper and we would like to consider coreference resolution for our future work.

D. Other Relation Types

In addition to the two types of relations described thus far, others that play important roles include *the ¬ Method-Purpose relation*, that is, the negation of the Method-Purpose relation, *the Citing-Cited relation* between a citing paper and a cited paper, *the Same-As relation* representing a synonym, and *the Super-Sub relation* expressing a hierarchical or whole-part relation.

IV. Relation Extraction Method

A. Definition and Notation of Extraction Patterns

In the proposed method, we use lexico-syntactic patterns to extract relations. First, we split related work sections into sentences, and then, we apply a syntactic parser to obtain a syntax tree. Syntactic tags such as "Noun Phrase (NP)" or "Verb Phrase (VP)" are labeled for each span and our system can exploit them. For example, in the sentence *"DRAGO [10] specifically examines a distributed reasoning based on the P2P-like architecture"* , while the expression "based on" acts as a key to extract the Method-Purpose relation that "the P2P-like architecture" is applicable to "a distributed reasoning", excessive spans such as "DRAGO [10] ..." cannot be excluded with the simple regular expression "*-based on-*". Syntactic tags can help system determine the boundaries of such spans for the targeted concepts.

The system uses the following extraction rule as a notation: "({NP}) based on ({NP}) = <mp> /2 /1", where <mp> represents the Method-Purpose relation and "/2 /1" denotes that a noun phrase appearing first expresses a purpose and the one appearing next is a method. A general regular expression can also be used in a rule.

B. Corpus Analysis for Extraction Patterns

In our study, we focused on two relations – the Topic-Paper (hereafter designated as <tp>) relation and the Method-Purpose (hereinafter designated as <mp>) relation – and analyzed the related work sections to identify frequent lexico-syntactic patterns.

In our analysis, we first chose 18 papers from the proceedings of the Association for the Advancement of Artificial Intelligence (AAAI2010). We manually annotated all the relations in the sentences of related work section and established 14 lexico-syntactic patterns for automatic relation extraction of <mp>. Additionaly, only the pattern appearing most frequently with a noun phrase immediately before a cited reference sign was used as lexico-syntactic patterns of the <tp> relation. The lexico-syntactic patterns obtained by the analysis are listed in the results section (Table I, Table II).

C. Rule Application and Extraction

Given one sentence in related work section and an extraction pattern described above, we apply the pattern to the sentence in following method:

 ISBN: 978-1-61208-254-7

1) Delete symbols that a parser cannot process properly: About parentheses and blackets, we delete them and words inside them. If there is any citation mark, their positions are recorded for later use. Words between quotations are concatenated with hyphens and quotation marks are deleted.
2) Berkeley Parser http://code.google.com/p/berkeleyparser/ is used for analysing syntactic tree.
3) Syntactic tag such as "{NP}" or "{VP}" is replaced by wildcard of regular expression ".∗" and sentences which matches that expression are extracted as candidates.
4) For each candidate sentence, the span of words corresponding to wildcard are examined using syntactic tag of the rule and the parsed tree. If those syntactic information matches, the span is extracted as a Concept node.

V. Experiment

We performed two experiments. In the first experiment, we checked the recall and precision of our method in a small but clean data set. Error analysis was also performed. In the second experiment, we evaluated the precision on a knowledge network extracted from a large data set. Owing to the large amount of data, many meaningful knowledge relationships were extracted. We describe some examples and discuss their implications in the next section.

A. Experiment 1

As we mentioned in Section IV, we need sentences in the "Related Work" section to use as an input. This means that various pre-processing steps are needed to extract a desirable format of the input.

1) PDFs of papers were converted to texts using a conversion tool (pdftotext http://www.foolabs.com/xpdf/).
2) Related work chapters were extracted from papers.
3) Reference sections were extracted from papers and divided for each paper.
4) Cited reference signs were extracted from the related work descriptions and matched with those in 3.
5) Related works descriptions were divided into sentences with a sentence division tool (GENIA Sentence Splitter http://www-tsujii.is.s.u-tokyo.ac.jp/y-matsu/geniass/).

Assuming that a set of co-citing papers contain closely related concepts, we selected 15 papers that cited either of the two papers: "Probabilistic latent semantic analysis"[5] or "Probabilistic latent semantic indexing"[6]. Because the test data are in the same research field as the data for lexico-syntactic pattern generation, their writing styles are expected to be similar.

These papers were processed in the same manner described in the previous section, and the correct relation was annotated using an annotation tool called brat http://brat.nlplab.org/.

Comparing those, recall and precision were calculated.

B. Experiment 2

In experiment 1, the data set was quite clean and small and therefore appropriate for basic statistics. However, we could not extract many meaningful knowledge networks because of the smallness, and it would not be feasible to increase the size of the dataset because some processes rely heavily on manual effort.

In experiment 2, we used a large dataset from Microsoft Academic Search http://academic.research.microsoft.com/ to evaluate the precision of each rule and extract knowledge networks.

The data set consisted of 906,788 cited papers and 8,388,909 citation context sentences. The domain was confined to computer science. From those, the number of papers whose citation count was no less than 100 was 9,252, and their citation context sentences numbered 1,952,112 in total. We used 18 paper of them and its 3099 citation context sentences.

VI. Results and Discussion

Table.I and Table.II show a part of the result of Experiment 1 and 2. Each row corresponds to the evaluation result extracted by each rule. Our method obtained an overall accuracy of 76.9% for Experiment 1 and 71.7% for Experiment 2.

Some rules lowered the accuracy and others were very accurate. The errors resulting from failure in syntactic analysis are unavoidable since the reported accuracy of the parser is about 90% without domain dependency problem [7]. Accuracy of parsing is low when the sentence contains a present or past participle because of intrinsic ambiguity. However, the influence of this type of errors (e.g. "close sense clusters" is extracted for correct span "finding close sense clusters") on the knowledge network may be limited if these extracted Concept nodes can be properly unified.

On the other hand, recall of Experiment 1 is 12.2%. This is quite low and thus we need to construct meta heuristics of making more rules. Besides, our goal is not to extract all the relationship pieces from each paper, but to describe whole image with enough semantic details. So, in the future, we plan to examine intra-paper and inter-paper redundancy of relationships and comprehensibility of result knowledge network representation.

From the result of Experiment 2, we were able to construct meaningful knowledge graph. For example, sentences and extracted relations from it are shown as follows.

Original Sentences

The Gen 2 MAC protocol is based on Framed Slotted Aloha [19].

At the MAC layer, readers and tags use a variation on slotted Aloha [14] to solve the multi-access problem in a setting where readers can hear tags but tags cannot hear each other.

Extracted Relations

Table I
TOTAL NO. OF <MP> RELATIONSHIPS EXTRACTED AND ACCURACY ABOUT EACH RULE – EXPERIMENT 1

No.	Rule	Total	Accuracy(%)
1.	= ({NP}) {be} based on ({NP}) = <mp> \2 \1	6	100.0
2.	= ({NP}) based on ({NP}) = <mp> \2 \1	15	73.3
3.	= ({VP}) using ({NP}) = <mp> \2 \1	16	50.0
8.	= use(?:s|d)? ({NP}) to ({VP}) = <mp> \1 \2	5	100.0

Table II
TOTAL NO. OF <MP> RELATIONSHIPS EXTRACTED AND ACCURACY ABOUT EACH RULE – EXPERIMENT 2

No.	Rule	Total	Accuracy(%)
1	= ({NP}) {be} based on ({NP}) = <mp> \2 \1	48	93.8
2	= ({NP}) based on ({NP}) = <mp> \2 \1	106	65.1
3	= ({VP}) using ({NP}) = <mp> \2 \1	154	52.6
8	= use(?:s|d)? ({NP}) to ({VP}) = <mp> \1 \2	97	95.9
13	= ({NP}) {be} used to ({VP}) = <mp> \1 \2	29	93.1
14	= ({NP}) {be} proposed to ({VP}) = <mp> \1 \2	14	100.0

- <tp> relation of “Framed Slotted Aloha” and “[19]”
- <mp> relation of “Framed Slotted Aloha” and “The Gen 2 MAC protocol”
- <tp> relation of “a variation on slotted Aloha” and “[14]”
- <mp> relation of “a variation on slotted Aloha” and “solve the multi-access problem in a setting where readers can hear tags but tags cannot hear each other”

The paper represented as “[19]” or “[14]” is identified by URL http://academic.research.microsoft.com/Publication/1242802/ and its title is “ALOHA packet system with and without slots and capture”.

Two types representation of “Framed Slotted Aloha” contribution – descriptive explanation and the name of succession protocol – helps us understand the position of that research.

VII. CONCLUSION AND FUTURE WORK

As described in this paper, we have proposed an approach for extracting relations among papers and concepts to construct a paper knowledge network. A sentence citing another paper is extracted from a related work chapter, and a lexico-syntactic pattern is established for extracting semantic relations between the papers from the quoted sentence. We then performed extraction experiments using academic papers written in English. Analysis of failure examples in the experiment revealed that analytical failures can be attributed to the parser that was used and to a limited number of ambiguous lexico-syntactic patterns. We expect to improve the accuracy by performing post-processing for the acquired set of relations and the addition of lexico-syntactic patterns. In the future, we hope to express a paper knowledge network for a whole field on the basis of relations among papers and concepts by addressing the integration of knowledge extracted from multiple papers.

REFERENCES

[1] C. Zhang and D. Wu, “Concept extraction and clustering for topic digital library construction,” *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, pp. 299–302, 2008.

[2] H. Nanba and M. Okumura, “Towards multi-paper summarization reference information,” in *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 926–931. [Online]. Available: http://portal.acm.org/citation.cfm?id=1624312.1624351

[3] S. Teufel, A. Siddharthan, and D. Tidhar, “Automatic classification of citation function,” in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 103–110. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610091

[4] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr, “Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization,” *JASIST: Journal of the American Society for Information Science and Technology*, 2012. [Online]. Available: http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2011-16

[5] T. Hofmann, “Probabilistic Latent Semantic Analysis,” in *Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[6] T. Hofimann, “Probabilistic latent semantic indexing,” in *Research and Development in Information Retrieval*, 1999.

[7] S. Petrov and D. Klein, “Improved inference for unlexicalized parsing,” in *HLT-NAACL*, C. L. Sidner, T. Schultz, M. Stone, and C. Zhai, Eds. The Association for Computational Linguistics, 2007, pp. 404–411. [Online]. Available: http://dblp.uni-trier.de/db/conf/naacl/naacl2007.html#PetrovK07

 ISBN: 978-1-61208-254-7