4K1-OS-2-4

構文パターンを用いた論文の引用文脈からの関係情報抽出

Relation Extraction of Citation Context Sentences Using Lexico-Syntactic Patterns

亀田 尭宙[*1] 内山 清子[*2] 武田 英明[*2] 相澤 彰子[*1*2]
KAMEDA Akihiro UCHIYAMA Kiyoko TAKEDA Hideaki AIZAWA Akiko

[*1]東京大学 [*2]国立情報学研究所
The University of Tokyo National Institute of Infomatics

To construct knowledge network among papers can help researchers direct their own work and utilize existing methods in previous related studies. In this paper, we propose a method for extracting semantic relationships between academic papers and concepts based on lexico-syntactic patterns from related work descriptions. We focused on the relationships among specific research topics, the methods and the reference papers, and built 15 lexico-syntactic patterns from the analysis. The extraction experiment was performed using a parser. The experimental result demonstrated the effectiveness of the proposed method.

1. はじめに

研究者が自分の研究テーマを絞り込んだり，研究成果を論文に執筆したりする際に，自分の研究に関連した先行研究を調査するために，関連研究論文を検索し，読んで分析することは重要な研究活動の一つである．先行研究調査の目的は，関連研究やその周囲にある問題の中に自分の研究を位置付け，優位性を明確化することである．これにより，論文で公表することが可能になる．本研究では，こうした研究活動を支援するために，関連研究章 (related work) を対象として，その中に記述されている引用文献情報，研究で用いられている手法等の情報を集約し，リンクで関連付けた論文知識ネットワーク構造を提示することを目指している．

関連研究章は，自分の研究が取り組んでいるテーマの関連研究を列挙し，その特徴と具体的な解決手法が記述されている．たとえば,「コーパスにおけるテキストと引用関係のモデル化」というテーマに対して，先行研究で用いられた手法，その手法を使って取り組んだ他のテーマ，その手法が提案されたオリジナルの論文，などの関係記述である．つまり，関連研究章には引用元の論文と引用先の論文及び引用先の論文同士の関係を把握するための情報が詰まっている．そこで本論文では，アルゴリズムや手法などの専門用語（e.g. “PHITS”，“PLSA”），名詞句 (e.g. “テキストのリンク情報”)，動詞句（e.g. “語の共起行列を生成する”），節など，論文知識を表現するに当たって有用な文以下の単位を纏めて**概念**と呼称して抽出し，これらと引用されている論文を結び付けることで，論文と概念の２種類のノードによるネットワークを構築する．

同じテーマを扱っている論文を収集し，論文内から抽出される論文―概念間，概念―概念間の関係をまとめることで，その分野全体の先行研究のネットワークが概観できる．また，同じ論文を引用している論文にある関係を統合することで，先行研究の様々な取り組みについて理解を深められる．さらに，分野を概観することで，自身の研究を周辺分野および関連研究の中に的確に位置づけやすくなる．本論文では，このような形で研究者の研究活動を支援するための論文知識ネットワークの構築手法を提案し，英語論文のデータセットを用意して実験を行った．

連絡先: 亀田 尭宙, 東京大学大学院情報理工学系研究科, 東京都文京区本郷 7-3-1, kameda@nii.ac.jp

2. 関連研究

論文からの意味情報抽出には多くの研究がある．その中でも，概念同士の関連を抽出しているものとしては，例えば，Zhang らの研究 [Zhang 08] が挙げられる．彼らは，論文のクラスタリングを行い，各クラスタを代表するようなキーワードを抽出し，さらにキーワード間の階層構造を抽出して論文検索のナビゲーションとして用いることを提案している．これを本論文の課題に則して考えてみると，論文内に現れるキーワードという関係と，キーワードの階層的関係を抽出する研究として捉えることができる．しかし，本論文では，キーワード間がどのような関係となっているのか，たとえばあるキーワードが別のキーワードおよび論文に対して手法の関係にあるのか，長所短所であるのかなどさらに詳細な関係の分析であり，このような情報を獲得するには，論文内の文章を詳細に解析し，統合する手法が必要である．

論文間の関係の分析については，難波らが論文内における他論文を引用している文を詳細に分析して，他の論文を基礎として引用しているもの, 差異を指摘しているもの，それ以外，の 3 種に分別する研究を行っている [Nanba 99]．また，近年では，Angrosh らが同様のタスクにおいて，背景，関連研究，関連研究の問題点，当該論文の成果に関する記述など 13 種類の記述を CRF を用いて分類し，96.51%という高い精度を達成している [Angrosh 10]．

これらの研究は，分類ラベルを付して論文間や論文と概念の関係を抽出するという同じタスクを解いていると読み変えることができる．例えば,「関連研究における欠点」と分類されたものは，その分類を関係のラベルとして，その文に現れる論文と概念の関係を表していると言える．しかし，どの論文を指しているかは括弧で囲まれているといった記述の様式から比較的明らかであるものの，概念については具体的に文のどの部分が相当するかということが明らかではない．

枠組みが近いものには，引用文脈の分類問題を論文間の関係の抽出として解いている Teufel らの研究 [Teufel 06] があるが，これもまた言葉による概念の表現を論文ネットワークのなかに直接位置付けてはいない．Dunne らの研究 [Dunne 12] では，論文間のネットワークに着目し，さらにネットワークをクラスタリングすることで分野の俯瞰を行うとともに，引用文脈の要約を同時に横に表示させるというインタフェース面での

工夫によって，テキスト情報による詳細な理解を可能にしている．我々はより直接的に，テキスト情報による詳細な情報を論文間のネットワークに位置づけたいと考え，引用文脈から概念間の関係抽出を行っている．

3. ネットワークのモデル

3.1 2 種類のノード

1. 章で述べた通り**概念**と**論文**の 2 種類のノードが論文知識ネットワークを構成する．概念はその文字列で同定され，論文はデジタルライブラリ上の URI で同定されている．

3.2 6 種類の関係

我々は，論文からの知識ネットワークを構築するにあたって，手法と目的や応用先の関係に着目し，**Method-Purpose 関係**と呼ぶことにする．ある手法が様々な問題の解決のために応用され，さらにその問題解決を応用することでさらに大きな問題が解けるといったネットワークは，工学的知識のネットワークにおいて重要な意味を持つためである．以下に例を示す．

> 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.

という文からは “Probabilistic Latent Semantic Indexing” が “to learn a lower dimension representation of text in terms of probabilistic topics”に寄与しているという関係が取得できる．さらに後者に含まれる “a lower dimension representation” というフレーズを手掛かりとして，類似の目的を持った手法に主成分分析があることや，さらなる応用に効率的な画像分類があるといった Method-Purpose 関係を繋ぎ合わせることができる．

> We can find **a lower dimensional representation** for the words using techniques like Principal Component Analysis (PCA) [5].

> Instead of solving the OCA optimization in the original image space, we limit the search to the span of the training images using **a lower dimensional representation**.

また，このような知識ネットワークが論文のネットワークと結び付くためには，それぞれのフレーズで表された概念がどの論文で扱われているかといった関係が必要である．これを **Paper-Topic 関係** として抽出する．

例えば，上に示された文のように，引用される文献は記号を用いて明示的に示される．本研究では，“[5]” といった文献の引用を示す符号を**引用文献符号**と呼び，これを手がかりに [5] の文献の中で，“Principal Component Analysis (PCA)” という概念が扱われていう関係を Topic-Paper 関係として抽出する．

ここまで述べた 2 種類の関係に加え，Method-Purpose 関係の否定となる **¬ Method-Purpose 関係**，デジタルライブラリから取得することのできる論文同士の引用-被引用関係である **Citing-Cited 関係** が主要な役割を果たす．また，同義語を表す **SameAs 関係** と上位下位もしくは全体部分関係を表す **Super-Sub 関係** も共に知識ネットワークを構成することで疎になりがちな概念間の関係を密にする．

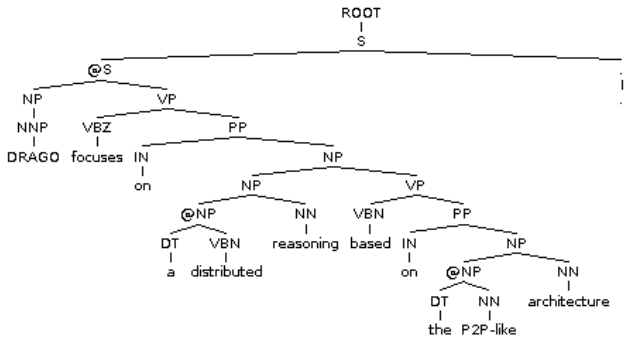

図 1: 構文解析の結果例

4. 論文からの関係抽出手法

本章では，1 文ごとに区切られた関連研究章の文を入力とし，前章で定義した関係を抽出する方法について述べる．

例えば，“DRAGO [10] focuses on a distributed reasoning based on the P2P-like architecture.”という文から “the P2P-like architecture” が “a distributed reasoning” に応用可能であるという Method-Purpose 関係を取り出すのには，“based on” というフレーズがその手掛かりになる．但し，“based on” の前後という形で正規表現を用いたマッチングでは，“. . . focuses on” といった余計な部分を省くことができない．そこで我々は文構造情報を用いた構文パターンによって関係を抽出することにした．

構文解析ツールによって 図 1 のような構文情報が得られる．これに対し名詞句 (NP) の情報を活かして “({NP}) based on ({NP})” となっている部分を見つけ出すことで目的の関係が抽出できる．このような抽出ルールを “= ({NP}) based on ({NP}) = <mp> \2 \1 ”のような記法でシステムは保持している．<mp> はこれが Method-Purpose 関係を抽出するルールであることを表しており，“\2 \1” という部分は先に出てきた名詞句が目的で次に出てきた名詞句の方が手段であるという順序を表している．また，ルール内には一般的な正規表現も同時に用いることができるようになっている．

さらに同じ文章において，引用文献符号直前の名詞句は当該の引用文献と Topic-Paper 関係にあるという構文パターン（システム内の構文パターン文法では “= ({NP}) ({CITE}) = <tp> \1 \2 ”のように表現されている）から，“DRAGO”というシステムは引用文献符号 [10] で指されている文献の中の主題であるという関係が抽出されてくる．実際は，引用文献符号については構文解析器が適切に扱えないため削除して構文解析をしている．そのため，元の文を用いて引用文献符号の直前にある単語を特定し，その単語を含む名詞句を取得している．

5. 関係抽出の実験

本論文では 3 章で述べた関係のうち Topic-Paper (以下 <tp>) 関係と Method-Purpose (以下 <mp>) 関係の抽出実験を以下の手順で行った．

まず構文パターン作成用のトレーニングデータは国際会議 Association for the Advancement of Artificial Intelligence (AAAI) 2010 の論文から，関連研究章があるものを 18 論文用いた．その 18 論文の関連研究章の文について，人手で論文・概念間関係のアノテーションを行った．その結果を用いて，関係を自動抽出するための構文パターン文法を 15 個作成した．

そのうち 14 個は <mp> 関係の構文パターンで，<tp> 関係は最も出現の多い，引用文献符号の直前に名詞句があるものだけ構文パターンとして用いた．

また，次節で述べるように別の論文からテスト用のデータセットを作成し，アノテーションを行った．

手法の適用に際して，構文解析ツールには Berkeley Parser[*1]を用いた．ただし，構文解析器が適切に扱えないシンボルなどについては前処理で削除した．

5.1 データセット

テスト用のデータには，論文 “Probabilistic latent semantic analysis”[*2]と “Probabilistic latent semantic indexing”[*3]のどちらかを引用している論文から，関連研究章があるもの 15 論文を用いた．テストデータについては，論文知識ネットワークを作る際に，同じ概念や同じ論文を言及している論文がデータセットの中に密に含まれることが必要なので，このように同じ論文を引用している共引用論文集合という形で用意した．また，構文パターン作成用のデータと同じ情報分野であることから，関連研究の書き方や現れる内容はある程度類似していることが期待できるとともに，情報分野の中でも様々な分野に応用されている技術を提案した論文を引用していることから多面的な引用文脈によって興味深いネットワークが抽出できることが期待できる．

テストデータは下のような手順で，1 文ごとに区切られた関連研究章の文とそこから取り出すべき関係，引用文献のリストのセットとして用意された．なお，今回は “This”や “Our approach”のように，論文知識ネットワークの構築に使うためには照応解析が必要なものもそのまま正解としてアノテーションした．

1. 論文 PDF を変換ツール (pdftotext[*4]) を用いてテキスト化
2. 論文内の関連研究 (Related Work) 章を抽出
3. 論文内の引用文献 (References) 章を抽出，各論文ごとに切り出し
4. 関連研究章内から引用文献符号を抽出し，2 と対応付け
5. 文区切りツール (GENIA Sentence Splitter[*5]) を用いて一文単位に切り取り
6. 各文についてアノテーションツール (brat [*6]) を用いて正解となる関係をアノテーション

5.2 結果と考察

<mp> 関係と <tp> 関係のそれぞれについて，表 1 に実験の結果を示した．また，実験の結果から，失敗した事例を収集し詳細に分析を行った．その結果，パーサが構文の解析に失敗しているか構文パターンが適切でないかの 2 つの要因で失敗が生じていた．Berkeley Parser は [Petrov 07] に基づいて実装されており，論文によると英語では 9 割程度の精度である．よって，構文解析の失敗に起因する間違いも避けられない．

<mp> 関係

一方，構文パターンに原因があって抽出できない例としては，“Our approach, however, is based on a model significantly different from these papers.”に対して “= ({NP}) {be} based on ({NP}) = <mp> \2 \1”を適用した場合のように副詞句が挿入されて抽出できないものもあった．

以上の考察を踏まえて，表 1 を解釈すると，using と based on を分詞として用いたもののみが 100%ではなく，他はすべて 100%の精度で抽出することができたのは，分詞が用法に関して強い曖昧性を持っており，意味の上から推測するしかないためであると考えられる．よって，パーサも解釈を間違えやすく，構文パターンも誤ったものを抽出しやすい．パーサについては前述のようにもともとの精度が 9 割程度であり，さらに分詞を用いた例が多かったことで今回は 84.6%の精度になっていた．但し，論文知識ネットワークを構築するという目的を考えると，“finding close sense clusters”と取ってくるべきところを “close sense clusters”と取ってきた場合などは，“cluster senses”といった類似の概念とまとめる段階で問題なく処理できる可能性もある．

<tp> 関係

前述のように，<tp> 関係については引用文献符号の直前に名詞句があるものを対象に抽出を行った．その結果，表 1 の通り 59 個の関係が取り出せたが，精度は 69.5%と十分ではなく，構文解析の失敗に起因するものが 9 個 (15.3%)，構文パターンに問題があるパターンが 11 個 (18.6%) あった．

構文パターンに問題があった例としては，“Documentcentric approaches are claimed to be much more effective than profile-centric (Balog2006), ...” という文から，“profile-centric”を抽出してしまったというものがあった．文や節の末尾にある引用文献符号は文全体にかかっているのか直前の句にかかっているのかが曖昧であり，suggest ／ claim ／ propose のような動詞が使われた場合はその目的語と意味上の主語である文献の間に <tp> 関係が成り立つということが多いということが分かった．今後はこれらの構文パターンを加え，構文パターン間の優先順位を設定できるようにすることで対応したい．また，今回は名詞句内に著者の姓名もしくは “et al.”が含まれていた場合は著者名として除外しているが，このケースは非常に多く，suggest ／ claim ／ propose といった動詞に対し主語が著者になっている場合も見られたため，著者名も構文パターンの拡張として加えるべきだと考えられる．

6. おわりに

本論文では，論文知識ネットワーク構築のために論文・概念間の関係を抽出する手法を提案した．関連研究の章から他の論文を引用している引用文を抽出し，その引用文から論文間の意味的関係を抽出するための構文パターンを作成し，英語の論文を対象として抽出実験を行った．実験結果の失敗例を分析し，使用したパーサの解析失敗や作成した構文パターンの不適切さが失敗の原因と分かった．これらの問題については，取得した関係集合に対して後処理することや構文パターンを追加することによって精度が向上するのではないかと考えている．今後，複数論文から抽出した知識の統合に取り組むことで，論文・概念間の関係に基づいて分野全体としての論文知識ネットワークを表現したいと考えている．

∗1 berkeleyparser http://code.google.com/p/berkeleyparser/
∗2 Thomas Hofmann: Probabilistic latent semantic analysis, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI’99), 1999.
∗3 Thomas Hofmann: Probabilistic latent semantic indexing, Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR’99), 1999.
∗4 http://www.foolabs.com/xpdf/
∗5 http://www-tsujii.is.s.u-tokyo.ac.jp/ y-matsu/geniass/
∗6 http://brat.nlplab.org/

表 1: **Total No. of relationships extracted and accuracy about each rule**

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
4.	= attempt(?:s|ed)? to ({VP}) using ({NP}) = <mp> \2 \1	1	100.0
5.	= ({NP}) attempt(?:s|ed)? to ({VP}) = <mp> \1 \2	1	100.0
6.	= attempt(?:s|ed)? to ({VP}) by ({VP}) = <mp> \2 \1	1	100.0
7.	= the use of ({NP}) to ({VP}) = <mp> \1 \2	2	100.0
8.	= use(?:s|d)? ({NP}) to ({VP}) = <mp> \1 \2	5	100.0
9.	= ({NP}) {be} introduced to ({VP}) = <mp> \1 \2	1	100.0
10.	= introduce(?:s|d)? ({NP}) for ({VP}) = <mp> \1 \2	1	100.0
11.	= ({NP}) {be} (?:.+\s)?extension of ({NP}) (?:to)? = <mp> \2 \1	1	100.0
12.	= approach(?:es|ed)? to ({VP}) {be} ({NP}) = <mp> \2 \1	0	-
13.	= ({NP}) {be} used to ({VP}) = <mp> \1 \2	0	-
14.	= ({NP}) {be} proposed to ({VP}) = <mp> \1 \2	2	100.0
15.	= ({NP}) ({CITE}) = <tp> \1 \2	59	69.5

参考文献

[Angrosh 10] Angrosh, M. A., Cranefield, S., and Stanger, N.: Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries, in *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, pp. 293–302, New York, NY, USA (2010), ACM

[Dunne 12] Dunne, C., Shneiderman, B., Gove, R., Klavans, J., and Dorr, B.: Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization, *JASIST: Journal of the American Society for Information Science and Technology* (2012)

[Nanba 99] Nanba, H. and Okumura, M.: Towards multi-paper summarization reference information, in *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*, pp. 926–931, San Francisco, CA, USA (1999), Morgan Kaufmann Publishers Inc.

[Petrov 07] Petrov, S. and Klein, D.: Improved Inference for Unlexicalized Parsing, in Sidner, C. L., Schultz, T., Stone, M., and Zhai, C. eds., *HLT-NAACL*, pp. 404–411, The Association for Computational Linguistics (2007)

[Teufel 06] Teufel, S., Siddharthan, A., and Tidhar, D.: Automatic classification of citation function, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pp. 103–110, Stroudsburg, PA, USA (2006), Association for Computational Linguistics

[Zhang 08] Zhang, C. and Wu, D.: 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)