3H2-OS3-7

関連研究に関する記述の分析による論文間の意味的関係の抽出

Extraction of Semantic Relationships among Academic Papers from Their Related Work Sections

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

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

In accademic communication, it is a very important activity to grasp the relationships among academic papers. For aiding this activity, we tried to extract semantic relationships automatically from related work sections in those papers. Specifically, we extracted 9 types of relationships by applying heuristic rules. We describe about our framework and results of some preliminary experiments.

1. はじめに

各分野の研究者にとって，論文を探し，それを読んで理解し，自分の研究に活かし，自分の研究を論文コミュニティの中に位置づけ，論文を投稿する，という一連の作業は欠かせない活動となっている．この活動の中で，論文とその中にある概念，概念間の関係を把握することが研究者には要求される．たとえば,「リンク構造のあるテキストの解析」というトピックについて研究している場合，その中でどのような手法があり，どれがよく使われており，それぞれの手法の短所長所はどこであり，その短所を解決する手法はどれなのか，その手法を提案している論文はどれなのか，といった関係である．

我々は，図 1 のような知識ネットワーク情報を提供することでこれらの関係を把握する手助けをしたい．図の中の語彙は表 1 のものであり，まさに先ほどの問いに対する答えの知識ネットワーク表現になっている．楕円形で表されたノードは論文を表し，角丸方形で表されたノードは概念を表している．ここで,「機械学習」といった専門用語的なものや,「リンク構造のあるテキストの解析」のように多くの論文で現れながらも名詞句の専門用語になるに至っていない概念を纏めて「概念」と呼称している．

同じトピックを扱っている論文内の記述を統合することで，その分野全体を見渡すことができ，同じ論文に言及している論文内の記述を統合することで，一つの論文に対する多様な捉え方を知ることができる．本研究では，このような論文と概念のネットワークを複数論文から抽出し，統合するための手法を提案する．

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，意味同士の関連を抽出しているものとしては，例えば，Zhang ら [Zhang 08] の研究が挙げられる．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．これを我々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，我々がやりたいのは，手法の長所短所などさらに詳細な関係の分析であり，こういった情報を獲得するには，論文内の文章を詳細に見ていくような手法が必要である．

論文間の関係の分析については，早くに，難波ら [Nanba 99] が論文内における他論文を引用している文を，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．また，近年では，Angrosh ら [Angrosh 10] が同様のタスクにおいて 13 種類のカテゴリを設定し，96.51%という高精度の分類を達成している．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 [] 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．

本研究では引用文のレトリカルな分類によって詳細な関係情報を取得しつつ，その関係情報に対応した概念の抽出まで行うことで，前章で述べたような問題の解決に取り組んでいる．

3. モデルと抽出手法

前述の課題を解決するために，我々は，論文間の関係を積極的に記述している関連研究の章から，論文と概念をノードとし，9 種の関係をエッジとして持つようなネットワークを抽出することに取り組んだ（表 1）．

また，ここまでネットワークのノードとして扱ってきた論文や概念は，その中に構造を持つ．論文は，文末の参考文献の記述の分析から，著者名やタイトルなど，BibTex に用いられるような情報をまとめ，それを論文ノードとしている．概念のノードに関しても，それは同義語や説明文の集合になっている．こうすることによって，語句の一致などを測ることによって，論文内や論文間における同論文や同義語の判定をおこなうことができる．

さらに，ノード間の関係を表す情報やノード内の語句表現について，それぞれ来歴情報を付与している．これによって，それぞれの論文や概念がどのように説明されているかが分かり，元の情報を辿ることができる．また，来歴情報は意味ネットワークを構築する際に信頼性を測るために重要な情報である [Carroll 05]. 関係や説明が記述されている文と，文が言及

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

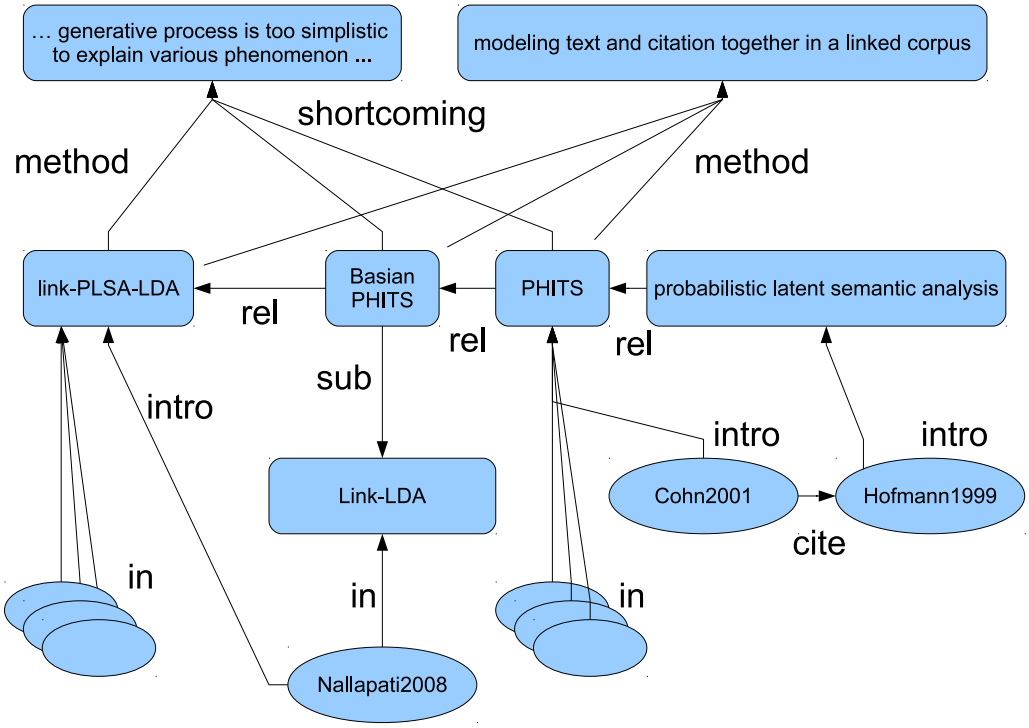

図 1: 抽出したいネットーワークの例

表 1: **Types of Relationships**

No.	Label	Description
Relationships between concept - paper		
1.	rel	A is somehow related to the paper B.
2.	introduced	A is introduced in the paper B.
3.	method	the paper B uses the concept A as means.
4.	shortcoming	the paper B has a shortcoming B.
Relationships between concept - concept		
5.	rel	A is somehow related to B.
6.	sub	A is part of B or subclass of B.
7.	method	A is used as a method for solving B. In other words, A provides a solution of B.
8.	shortcoming	A is a shortcoming of B.
Relationships between paper - paper		
9.	base	A is based on B.

表 2: **An Example of Concept Node**

- semi-supervised bootstrap learning methods
  - Sentence: Our approach employs semi-supervised bootstrap learning methods, which begin with a small set of labeled data, train a model, then use that model to label more data.
  - In: (Andrew 2010)
- bootstrap learning
- bootstrap learning methods

している論文，その文を含んでいる論文が来歴情報に含まれている．例えば表 2 のように 3 つの同義語で一つの概念ノードを構成しており，それぞれの表現が，どの論文のどの文の中に現れたか，という来歴情報を伴っている．

本研究では，ネットワークの抽出のために，まずそれぞれの論文について情報を抽出を行い，複数の論文で抽出した情報を統合する．今回は特にそれぞれの論文からの抽出について述べる．

Angrosh らの研究 [Angrosh 10] では CRF を用いて分類した 13 のカテゴリのうち，(1) 背景に関する記述 (2) 関連研究の説明記述 (3) 関連研究の問題点に関する記述 (4) 当該論文の成果に関する記述，の 4 つが突出して多かった．よって，本研

表 3: **Procedure**

1. 論文内の関連研究 (Related Work) 章を取り出す.
2. 論文の引用になっている文字列 (e.g. [1], [Kameda 2011]) を抽出する.
3. GeniaSS[*1]を用いて一文単位に切り取る
4. NLTK[*2]を用いて POS タグの付与や，名詞句節のタグの付与を行い，データを生成する.
5. ルールを用いて概念とその関係を抽出する.
6. Conditional Random Fields (CRF) を用いて各文のレトリカルな役割を推定し，それを利用して，文型からは抽出できない暗黙的な関係を抽出する.
7. "A , B and C"など纏めて抽出した概念を分割する.

究ではこの 4 つのカテゴリに文を分類している.

元データとしては PDF で発行されている英語論文を用い，関連研究章の抽出まではルールでの抽出の後，人手で修正を加えている．論文の引用になっている文字列の抽出はルールで行っている.

予備実験として，国際会議 Association for the Advancement of Artificial Intelligence (AAAI) の論文から，関連研究章があるものを用い，論文と概念間の一般的な関係（表 1 における"rel"）の抽出を試みた．まず，論文の引用になっている文字列をそのまま含んだままで処理を行うと，表 3 にある前処理を 5 論文 82 文に対してかけると，一文単位に切り取る部分で 2 文失敗し，POS タグの付与の部分では 31 文で精度が落ちた．結果，１つのルールで取得可能な 12 の関係のうち，5 つが取得できなかった．よって，論文の引用になっている文字列の分離を行った後に他の前処理をしている.

4. おわりに

本論文では，関連研究の章から論文間の意味的関係を抽出するためのフレームワークを提案した．今後，複数の論文に対して手法を適用し，個々の論文からの抽出精度について検証するとともに，複数論文から抽出した知識の統合にも取り組むことで，分野全体としての知識ネットワークを表現したい．複数の論文からの情報の統合においては，同義語や同論文の判定の技術が別途必要である．また，抽出した情報を RDF レポジトリとして提供することで，SPARQL[*3]で検索できるようにしたいと考えている.

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

[Carroll 05] Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P.: Named graphs, provenance and trust, in *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pp. 613–622, New York, NY, USA (2005), ACM

[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.

[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)

∗3 http://www.w3.org/TR/rdf-sparql-query/