情 報 学 基 礎 69－6
（2002．11．6）

学術論文における用語間の意味関係抽出の一手法とその実験

石川大介† 宇陀則彦‡ 石塚英弘‡
図書館情報大学大学院情報メディア研究科†
図書館情報大学図書館情報学部‡⋆

機械可読な大規模な学術論文の情報として、NII-NACSISコレクションがある。これを用いて、論文の標題におけるキーワードの使われ方を調べ、その中の特定の構文パターンに注目して、用語間の意味関係の抽出を試みた。本研究では、品詞情報を必要としない手法を使い、三つ以上の用語間の意味関係を記述する構文パターンについて実験した。その結果、単純な意味関係を明確に記述する構文パターンを利用することにより、意味関係を抽出できることを見い出した。

A method of extracting semantic relationships among terms in science paper and its experimentation

Daisuke Ishikawa† Norihiko Uda‡ Hidehiro Ishizuka‡
Graduate School of Information and Media Studies, University of Liblary and Infomation Science†
Faculty of Liblary and Infomation Science, University of Liblaryand Infomation Science‡

NII-NACSIS collection is machine readable and large information about science papers. Using this information, we researched usage of keywords in article title and tried to extract for semantic relationships among terms using specific syntactic pattern in it. In this work, we experimented for syntactic pattern describing semantic relationships among three terms. The results shows that some syntactic pattern clearly describing simple semantic relationships are applicable extract some semantic relationships. Our method need not part-of-speech informations.

1 はじめに

人間の思考活動は、高度な知識の集合体によって支えられている。計算機が同等の能力を得るためには、同様に高度な知識の構築が必要であると考えられる。人間は何らかの媒体を介して言語として記述された情報を元にして解析し、そこから意味を抽出して自らの知識として蓄えていく。計算機も同じ手法により知識が得られないであろうかというのが本研究のテーマである。

機械可読な大規模な学術論文の情報として、NII-NACSISコレクション[10]がある。これは、情報検索の研究のために、日本の多くの学会から論文を電子的に収録したものである。我々はこのデータを使い、論文の標題において、キーワードがどのように使われているかを調べ、その中の特定の構文パターンに注目して、用語間の意味関係の抽出を試みた。

本著者の石川はこれまで、主に二つの用語間に

⋆2002年10月1日から大学統合により、筑波大学図書館情報学系となった。

のみ注目して意味関係を抽出してきた[8]。また、形態素解析を利用して品詞情報を元にして解析を行なってきた。今回は、品詞情報を必要としない手法を使い、三つ以上の用語間の意味関係を記述するような構文パターンについて実験した。

2 情報の資源化

我々は、専門的な領域における用語に注目し、その用語によって表される概念を、用語間の体系的な構造化によって関係を表現し、これを用いた応用を考えている。

現時点では、主に用語の意味関係に基づいた資源化を行なっている。この資源化方法は[1]、用語間の同値関係を自動抽出する方法や、専門用語の造語規則に基づいて階層関係と関連関係を自動抽出する方法が示されてあり、各方法についてそれぞれ研究がなされている。

SS-SANS法 (Semantically Specified Syntactic Analysis) [1] [2] は、特定用語に関する構文解析による方法である。これは、まず特定の用語を中心とする文章中から特定構文を利用して、概念間の関係を抽出する。次にその結果を用いて新しい特定構文を得る。これを再帰的に繰り返す方法である。SS-SANS法における詳しい内容は次節にて述べる。

シソーラスの構築において[3]、構造化すべき基本的な用語間の意味関係に、等価関係 (同値関係) の UF(use for) や USE、階層関係の上位をBT(broader term) と下位をNT(narrower term)、連想関係にある用語は関連語と呼ばれRT(related term) があり、これらを構造化している。SS-SANS法の役目は、階層関係や、関連関係 (連想関係) の抽出をすることである。関連関係には様々な関係があるが代表的なものとして、過程と道具 (速度測定と速度計)、行為と結果 (科学研究、科学的発展)、行為と受容体 (データ分析とデータ)、起源に関する概念 (情報と情報源)、因果関係 (傷害と事故)、などが挙げられている。このような関係を抽出するのが狙いである。

SS-SANS法によって得れらた意味関係は、前述の資源化手法によって得られた同値、階層関係との統合が可能である。これによって、より多くの意味関係が相互作用することにより、高度な意味関係の処理ができる[4]。こうして、意味関係の統合化を進め[5]、この構造化された意味関係を利用して演繹推論、帰納推論、仮説推論など[6]高度な思考機能の実現を目指している。

3 SS-SANS法

以下に、具体的な処理方法と、従来までの研究について述べる。

3.1 具体的な処理方法

SS-SANS法の処理を具体的に表すと以下のようになる。

1. 構文パターンを S_n とし、用語を T_n とする
2. S_1 を「～を用いて～を～」とした場合、

$$S_1(T_1,T_2,T_3) = T_1\text{を用いて}\ T_2\text{を}\ T_3$$

とする
3. i=1 から n において、
S_iにより T_1,T_2,T_3 を抽出する
4. T_1,T_2,T_3 を使って、新たな S_{i+1} を抽出する
5. 以下、3,4 を繰り返えす。

3.2 従来の研究

SS-SANS法は、最初は分析化学における標題の意味解析を行なうために考案された[2]。分析化学の標題では主に、分析試料、分析成分、分析方法の用語から構成されている。これらの分析に関連する用語を、それらの用語を結ぶ構文のパターンを解析することにより、構文パターンをまとめ、自動抽出を行なった。

この手法は標題だけでなく本文にも適用可能であり、日本語の論文の要旨において、簡単な構文パターンを当てはめて、用語間の意味関係の自動抽出を行なった研究が挙げられる[7] [8]。この研究では、処理方法の項目3,4,5の処理を行なった。これらの処理過程は、形態素解析ツールを利用して品詞情報を手掛かりにテキストを照合している。また、二つの用語間に限定して、二用語間の意味関係のみを中心に抽出している。この結果は、関連関係データ ARTS [9] として収録し、まとめてある。

なお、本研究では、処理方法の項目3について、従来の方法[8]と違った処理方法を提案し、実験した。

4 NII-NACSIS コレクション

NII-NACSIS コレクションは、主に情報検索のために利用されている大規模なテストコレクションである [10]。日本国内の様々な学会の発表論文などが収録されている。この NII-NACSIS コレクションは、論文一件毎に、

- 整理番号
- 標題
- 著者
- 発表様式
- 日時
- 要旨本文
- キーワード
- 学会名

の項目が記述されている。ここで、キーワードとは発表者が自分の研究に対して自由に付与できる最大 6 個までの語である。

本研究では、NII-NACSIS コレクションの 1999 年度版を使って、日本語論文の標題について議論していく。さらに、この収録された論文のうち、人工知能学会関連[1] の論文 2031 件を対象とした。なお、コレクションの中でも語分割データ (分かち書きされているもの) を使用した。

4.1 論文の標題で使われる語の集計

各論文の標題は分かち書きがなされている。その最小単位をここでは語と呼ぶ。各本文から語のみを切り出して集計したものを表 1 に示す。

上位には助詞や句読点、標題でよく使用される記号「-」や「:」が並んでいる。20 位以下から「開発」や「応用」、「提案」、「評価」など、研究の種類を具体的に示す名詞が現れた。

4.2 各論文に付与されたキーワードの集計

前節と同様に、各論文に付与されているキーワードのタグの中身を集計した。表 2 に示す。

キーワードは全部で約 4500 種類である。論文件数が約 2000 件であるから、そのほぼ倍の数であった。上位に位置するキーワードは、どれも人工知能を連想させるキーワードとなっている。

[1] 人工知能学会において、全国大会やシンポジウム、各種研究会の報告など。約半分の 1000 件が全国大会。

順位	個数	語
1	2313	の
2	1267	に
3	508	を
4	486	と
5	380	おける
6	322	よる
7	245	-
8	237	基づく
9	190	用いた
10	170	ついて
11	162	ため
12	116	から
13	103	へ
14	100	な
15	95	その
16	89	関する
17	81	する
18	77	利用
19	71	:
20	66	学習
21	66	)
22	64	(
23	55	開発
24	54	構築
25	54	研究
26	44	実現
27	44	応用
28	38	提案
29	36	知識
30	34	評価

表 1: 分かち書きされた標題中の語の集計

5 本研究の手法と実験

本研究では、SS-SANS 法による [8] の研究と異なり、形態素解析ツールによる品詞情報を必要としない手法を提案する。前回までの結果は、扱うべき用語 (抽出の対象となる用語) は限られており、最終的にはこれらの用語間以外は不要であること、また用語以外の助詞や動詞などは表現の自由さから特定が難しいこと挙げられる。そのため、予め抽出しうる用語を特定しておき、それ以外は全て同等として扱うことを考えた。

また、この手法では三つ以上の用語間の意味関係を中心に抽出を試みた。以下、その具体的手法を述べる。なお、本実験は NII-NACSIS コレクションの人工知能学会関連の全論文を対象とし、標題に限定して実験を行なった。

個数	キーワード名
98	知識_獲得
79	人工_知能
77	エキスパートシステム
73	機械_学習
69	知的_CAI
65	知識_表現
63	学習
60	遺伝的_アルゴリズム
55	マルチエージェント
52	事例_ベース_推論

表 2: キーワードの集計

5.1 用語の選定

我々の狙いは、用語間の意味関係の抽出である。そのため、抽出すべき必要のある用語は用語集に登録しておく。具体的には、NII-NACSIS コレクションの人工知能学会の論文を使うので、これらのそれぞれの論文に付与されているキーワードを全て登録した。しかし、これだけでは不十分なため、分かち書きによって語基数が二個以上になる語は複合語とみなし、用語集に登録されている語と同等の扱いをするような処理をしてある。

これ以外の全ての語は、用語でないものと見なす。この中には助詞や助動詞、一般的な名詞や形容詞、さらに記号なども含まれる。

5.2 用語集を使ったテキストの変換

前述の用語集を使ってテキストを変換する。テキスト (ここでは各論文の標題) の語を一つづつ調べ、用語集に登録されている語であるならば「T」に変換する。それ以外は「w」に変換する。なお、「w」は連続していても「w」と見なすことにする。つまり「ww」や「www」も全て「w」に変換される。この手順により、標題のテキストを「T」と「w」に変換をした。以下に、変換前と後の処理の出力を示す。

```
TwT  ←  直交型_推論 に 基づく 問題_解決_機構
TwT  ←  類推 の ため の 抽象化
wTw  ←  証明 の 一般化_技法 に ついて
 Tw  ←  学習_アルゴリズム の 変換 に ついて
TwT  ←  ノイズ を 含んだ 例 から の 学習_可能性
 Tw  ←  アルゴリズム論的_教示 の 理論
 ...  ←  ...
```

5.3 特定の語「生成」を用いた標題の出力

特定の用語を調べることについて考える。ここでは「生成」という語に注目する。標題中のテキストに「生成」が存在すれば「T」ではなく、一番目に注目しているという意味で「1」に変換する。こうして、「1」が含まれる標題だけを出力した。この処理によって、21 個の標題が出力された。この一部を表 3 に示す。

この表から、「生成」という語は主に標題の終りの方に使用され、「〜の生成」というパターンで出現することが多いことが分かる。そのさらに前の部分はいろいろとあり、「〜における〜の生成」や「〜による〜の生成」などがあった。

5.4 構文パターン「〜からの〜の〜」

表 3 から、一つ例として「〜からの〜の〜」という構文パターンを取り上げることとする。本研究では、この構文パターンを

$$T_M \text{ から の } T_O \text{ の } T_P$$

- T_M：媒体を示す語
- T_O：対象を示す語
- T_P：処理を示す語

と考える。この構文パターンを利用して出力した結果の一部を表 4 に示す。全部で 12 個の標題が出力された。

この表において、下線が入っている語は、そのカテゴリーに属するとは言えない、もしくは疑わしい語である。それ以外は、多少の議論があるかもしれないが、おおむね媒体、対象、処理の意味として捉えられる語であった。

実際には、標題は「〜からの〜の〜」の構文パターンの前後に様々な語が用語を伴って結合している。「〜と〜からの〜の〜」や、「〜による〜からの〜の〜」などの構文パターンがあった。用語の役割を推測してみると、媒体を示す用語 1 を T_{M1}、媒体を示す用語 2 を T_{M2} とすると、

「〜と〜からの〜の〜」

↓

「T_{M1} と T_{M2} から の T_O の T_P」

TwTwlwT	←	日本語_文章 に おける 主題_表現 の 生成 に 関する 基礎的_考察
TwTwTwl	←	知識 の 並列_探索 に よる 機械_設計_手順 の 生成
TwTwTwl	←	組立_説明図 から の 情報 の 融合 に よる 組み立て_プラン の 生成
TwTwTwl	←	説明 に 基づく 物語_生成_システム に おける プロット の 生成
TwlwTwT	←	言語_理解 ・ 生成 に おける 処理 の 階層 と 統合
TwTwl	←	遺伝的_アルゴリズム に よる パレート 最適 な 決定木_集合 の 生成
TwTwTwl	←	力学_問題 に おける 補助_問題 の 分類 ・ 生成
TwTwl	←	対象_知識 から の 状態_空間_モデル の 生成
TwTwlwT	←	遺伝的_アルゴリズム を 使った 解析 、 生成 および 自然_言語_処理_課程
TwTwl	←	機能_連鎖_構造 に 基づく ヘルプ_応答 の 生成
TwTwTwl	←	遺伝的_プログラミング を 用いた ゲーム の 局面_評価_関数 の 生成

表 3: 「生成」という語が使用される標題

媒体 T_M (から の)	対象 T_O (の)	処理 T_P
トレース	問題_分割_戦略	獲得
確率_データ	帰納_学習	アルゴリズム
定量_情報	定性_物理_モデル	帰納_学習
二言語_対訳_コーパス	動詞	格フレーム_獲得
データベース	構造的_従属_関係	発見
画像_情報	相対_概念	獲得
ハードウェア_記述_言語_ VHDL	対象_モデル	自動_生成法
対話_過程	構文_規則	自動_生成
二言語_対訳_コーパス	翻訳_知識	自動_獲得
対象_知識	状態_空間_モデル	生成
知識	階層_構造	獲得
数値_データ	微分_方程式	学習

表 4: 構文パターン「T_M から の T_O の T_P」の出力結果

となりそうである。また、いろいろと使われ方によって捉え方が違うかもしれないが、ここでは T_P を操作または過程として考えた場合、T_T を行為者または道具として考えるならば、

「〜による〜からの〜の〜」

↓

「T_T による T_M から の T_O の T_P」

となるだろう。

5.5 構文パターン「〜を用いた〜の〜」

今度は、前節と同様に「〜を用いた〜の〜」という構文パターンについて調べることにする。この構文のパターンに当てはまって処理された標題は 35 個であった。

ここで、標題が完全に「〜を用いた〜の〜」という構文パターンに当てはまったのは 17 個であった。この結果を表 5 にまとめる。逆に、標題の一部がこの構文パターンによって一致したものを、表 6 にまとめた。

構文パターン「〜を用いた〜の〜」のそれぞれに当てはまった用語について、標題においてそれぞれの用語は以下のような役割を担っていると考えてみる。

T_T を 用いた T_O の T_P

- T_T : 道具を示す語
- T_O : 対象を示す語
- T_P : 処理を示す語

表 5 は、この構文パターンに完全に一致した標題である。そのため、下線を引いた用語以外、ほぼこの役割を示す語として扱われていた。下線部の語「停止性_検証_システム」は、実際には対象を示す語、つまり T_{O2} と考えられる。

表 6 は、部分一致により構文パターンが含まれていた標題である。そのため、あまりこの役割

道具 T_T (を用いた)	対象 T_O (の)	処理 T_P
文_タイプ_情報	話題_構造	認識
自然_言語_インタフェース	検索_結果	視覚化
遺伝的_アルゴリズム	外国_為替_市場	シミュレーション
二分_決定_グラフ	書換え_型_プログラム	停止性_検証_システム
市場_モデル	アプリケーション_QoS	制御

表 5: 構文パターン「T_T を 用いた T_O の T_P」に完全一致

-	道具 T_T (を用いた)	対象 T_O (の)	処理 T_P	-
可能性 と 必然性 の	関係	故障_診断_対象	知識_表現_方法	の 強化 と その 推論_手法
	定性的_距離	類似度_関数	重み付け	の 学習
	ベイジアンネットワーク	対話_相手	知識	の 推定法
	オントロジー	情報	自動_収集	と 分類 へ の アプローチ
	遺伝的_プログラミング	ゲーム	局面_評価_関数	の 生成
ベイジアンネット に よる	情報_統合	自律_移動_ロボット	自己_位置_同_定	

表 6: 構文パターン「T_T を 用いた T_O の T_P」に部分一致

通りの用語の使われ方にはなっていない。特に目立った結果として、「〜を用いた〜の〜の〜」というパターンが多い。これは、対象を示す語 1、2 を T_{O1}、T_{O2} として考えた場合、

「〜を用いた〜の〜の〜」
↓
「T_T を 用いた T_{O1} の T_{O2} の T_P」

となり、これらの標題では実際にはこの役割で語が使われていたと考えられる。

6 考察

ここまでの結果について考察する。まず、構文のパターンによっては、比較的簡単に設定した役割通りに解析できるパターンと、それだけではほとんど正しく解析できないパターンがあることが分かった。また、どちらにおいても、パターンが与えた役割に準じない語が現れることも分かった。

さらに、構文パターンと完全一致のみ許した場合、部分一致を含むものと比べると、与えた役割とは違う結果になってしまう標題が減少した。これにより、より厳しくパターンを照合すれば、抽出結果は減少するが、解析の精度が上がると考えられる。

一般に、名詞句が何をするのかという場合、役割という意味で格 (case) という言葉が使われる。これは、Fillmore が提唱した格文法と呼ばれる文法理論である [11]。このような格を使って名詞句を解析し、応用する方法は人工知能の研究において古くから使われてきた [12]。また、計算機システムが自然言語の意味を抽出しようとするとき、求める情報だけを抽出する手段としてしばしば応用されてきた [13]。

格文法の理論に従って得られた結果について考えると、今回の実験の結果は、処理の役割を担う語 T_P は動詞句にあたり、他の格 (役割を表す語) の役割を支配していることになる。そのため、T_P を「生成」に限定したために現れた構文パターン「〜からの〜の〜」において、それぞれに設定した格には適した語が当てはまったものと考えられる。逆に、「〜を用いた〜の〜の〜」では、T_P にいろいろな語が当てはまる可能性があるため、それぞれの格に曖昧性が増したものと考えられる。

7 おわりに

今回の研究によって、標題の用語の使われ方に注目し、ごく単純な意味関係を明確に記述するような構文パターンを利用することにより、簡単な意味関係を抽出できることを示した。また、構文パターンによっては、適切な意味関係を抽出するためには、さらに多少の工夫が必要なことも分かった

本研究において、重要な点は、限定した意味関係のみの抽出であったが、三つ以上の用語間の意味関係を扱ったことである。今後は、二用語間の意味関係との比較や、さらなる構文パターンの適