

テストコレクションを用いた用語の意味関係の抽出

4H-7

石川大介† 近藤雄裕† 森本貴之† 藤原譲★

神奈川大学大学院理学研究科† 神奈川大学理学部† 独立行政法人工業所有権総合情報館★

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。

論文などの専門用語を多く扱ったテキストデータに NII-NACSIS コレクションがある。本研究ではこの中の日本語要旨に記述された情報を抽出し、意味関係の抽出を試みている。今回、ある特定の構文によって得られた用語は何を意味するのか、また構文同士の関係性などを検討した。

2 SS-SANS 法

用語間には同値、階層など様々な関連関係があり、特に因果関係を論文などの文章から自動的に抽出する方法として SS-SANS (Semantically Specified Syntactic Analysis of Sentences) 法がある [1]。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。

3 関連関係

本研究では、NII-NACSIS コレクション (1999 年度版) [6] の NTCIR1 (語分割データ) を使用し、各学会の日本語の論文要旨 (約 33 万件) を入力データとしている。これを形態素解析ツール JUMAN [7] により品詞情報を得て、語分割がなされている用語を複合名詞としている。これらの情報を要旨の文章に添付してデータ化し、これを SS-SANS 法によって処理した。この時、初期条件として構文ファイルに「を 行う」という構文から処理を行った。ここで、使用したテンプレートは「複合名詞 助詞 動詞 複合名詞」である。

Extraction of semantic relationships in Test Collection
Daisuke ISHIKAWA†, Takahiro KONDO†, Takayuki MORIMOTO†, Yuzuru FUJIWARA★
Graduate School of Science, Kanagawa University†
Faculty of Science, Kanagawa University†
National Center for Industrial Property Information★

これにより約 15 万種類の用語間を結ぶ関連関係が得られ [4]、用語の主語、目的語の関係について検討された [5]。その他、この関連関係データを用いた研究として、情報検索システム [2] や知識の構造化 [3] などがあり、現在も検討中である。

4 用語間の関係の解析方法

関連関係データを、「情報処理用語大辞典」を元に C-TRAN 法や SS-KWEIC 法を用いて同値関係や階層関係にしたデータベース [3] を使って解析を行う。その詳細は以下の通りである。

- 同値関係：1510 集合、全用語数 3405
- 階層関係：26708 関係、全用語数 53428
- 兄弟関係：2597 集合、全用語数 22961

これらのデータの一部を以下に示す。

- 同値:[コンピュータネットワーク コンピュータ網 計算機網], [L S I 大規模集積回路]...
- 階層:[アルゴリズム 学習アルゴリズム],[検索 ファジィ検索]...
- 兄弟:[階層型パーセプトロン 実数パーセプトロン 多層パーセプトロン], [分散型データベース管理システム オブジェクト指向データベース管理システム]...

5 解析結果と考察

関連関係データのうち、同値関係を示す構文は一つも得られなかった。一方、階層関係を表す文章は 52 個と全体から見ればわずかだが得ることができた。階層関係を表す構文を以下に示す。また、兄弟関係を示す文章は 336 個と階層関係に比べて若干多く得られた。兄弟関係を表す構文を以下に示す。

どちらも「と して」という構文が多数であった。ここで、兄弟関係を表すデータには、造語規則に基づいた用語の階層化をしているため、実際には更に広義な用語も同じ階層になってしまうケースがある。例として、「学習アルゴリズム と して EM アルゴリズム」という結果が挙げられ、これは兄弟用語として保持しているが、実際の意味では学習アルゴリズムが上位概念であろう。よって、ここで表れる兄弟関係の構文は実際には階層関係を意味する場合があり、更に検討が必要である。

個数	構文名	個数	構文名
28	として	81	として
7	である	49	である
2	を用いた	34	による
1	を組み合わせた,	19	を用いた
	を含む,	19	における
	に対して,	14	に対する
	において,	12	を用いて
	に属する,	12	に基づく
	からなる,	9	において
-	-	-	-
階層関係		兄弟関係	

6 構文の分類について

構文の中には、他の構文で言い替えられるものがある。これをまとめることにより、構文の分類を行った。方法として、ある構文が結ぶ関連関係を示す用語が、他の構文でも同様に結ばれているかどうかを調べている。以下に結果の一部を示す。

対象構文	関連関係	構文数	構文名	個数
による	25866	447	を用いた	163
			を用いて	56
			によって	50
を用いた	12422	370	による	163
			を用いて	96
			に基づく	24
における	21871	201	において	94
			による	30
			に対する	18
を持つ	2255	94	をもつ	33
			を有する	29
として	8240	84	である	32
			とした	14
			とする	10

この結果から、漢字やひらがなや送り仮名の違いという表記の差違、それとほぼ同じ意味の言い回しの違いなど、意味的に同じ構文の分類が可能と思われる結果が得られた。

7 複数の用語に関する考察

今後、複数の用語からなるテンプレートを処理する上で、様々な問題がある。そこで、まずはあるテンプレートと同一の文章の総数を調べることにした。以下に、テンプレートとその総数を示す。

これにより、用語数を増やすと指数的にテンプレート数が減ることが分かる。

用語数	テンプレートの種類	総数
2	NSVN	174839
3	NSNSVN	23595
3	NSVNSN	27149
4	NSNSVNSN	3180
5	NSNSNSVNSN	282
5	NSNSVNSNSN	348

N: 複合名詞、S: 助詞、V: 動詞

8 まとめ

今回の解析により階層関係を示す代表的な構文が得られた。また、構文の多様性はある程度の範囲では吸収できそうと言える。しかし、意味関係を抽出する上で、決定的と言えるほどの結論は得られなかった。今後、更に構文における意味解析を進めると共に他のテンプレートによる意味抽出などを考慮するなど、多角的な検討が必要であろう。

謝辞

本研究において、国立情報学研究所で作成されたNII-NACSIS コレクションのNTCIR1を使用しました。深く感謝いたします。

参考文献

- [1] Hikomaro Sano, Yuzuru Fujiwara: Syntactic and semantic structure analysis of article titles in analytical chemistry, Journal of Information Science 19, 119-124, 1993
- [2] 森本貴之 近藤雄裕 杉田勝彦 石川大介 池村匡哉 藤原譲: 構造化された知識を基にした情報検索システム、第9回研究報告会、pp.75-80、情報知識学会、2001
- [3] 近藤雄裕、藤原譲: 意味関係抽出による概念の構造化、第9回研究報告会、pp.71-74、情報知識学会、2001
- [4] 石川大介、池村匡哉、近藤雄裕、杉田勝彦、森本貴之、藤原譲: SS-SANS 法を用いた意味関係の自動抽出、情報処理学会第62回全国大会、2001
- [5] 石川大介、藤原譲: 特定構文を用いた用語間の意味関係の抽出、第9回研究報告会、pp.67-70、情報知識学会、2001
- [6] NII-NACSIS テストコレクション:
<http://research.nii.ac.jp/ntcir/index-ja.html>
- [7] 日本語形態素解析システム JUMAN:
<http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html>