特定構文を用いた用語間の意味関係の抽出

○石川大介 †
藤原　譲 ⋆

Automatic extraction of semantic relationships among terms by SS-SANS

○ Daisuke ISHIKAWA †
Yuzuru FUJIWARA ⋆

SS-SANS method is to extract semantic relationships with automatic. This work is to apply SS-SANS extract in a associative relationships in the test collection which includes thesis data(300 thousands of japanese abstracts). The process uses templates between two terms for extraction and it extracted automaticaly 150 thousands of relationships. These relationships find out distinction between front term and end term. Front terms are called OBJ terms and end terms are called SBJ terms. Analyse of terms for OBJ terms or SBJ terms and their distributions. In conclusion, distinctions to use terms for OBJ terms or SBJ terms are explained.

1　はじめに

人間が考え、意思決定をしていく過程において、知識は欠くことができない。人間の思考を支える多くの機能はこの知識からなり、コンピュータに同様の処理を行わせるには、人と同等の知識が無ければならない。

学習の定義は様々だが、その中の一つに明示的な知識の獲得がある。では知識はどのように獲得されるかといえば、人は知識として情報が記述された本や論文などを手に取り、そこに書かれている文章を読み取ることにより自らの知識と結合させながら学習していく。本研究では、このようなより人間に近い形で学習ができないかを進めている。

人間にとって身近な情報源であり、特に生の知識情報が記述されているものとして、CD-ROM で配布されている論文 (NII-NACSIS コレクション) がある。この中の日本語要旨を形態素解析し、特定構文により用語関係を自動抽出し、得られた情報がどういう意味を表すのかについて検討した。

2　SS-SANS 法

関連関係、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][2]。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。

この方法は、目的とする文章からテンプレートを使って用語と構文を抽出する。ここでテンプレートとは品詞の並びを指す。実際の処理の流れは、テンプレートと同一の文章があった場合、用語の組合せかもしくは構文のどちらかが既知であった場合、もう一方を追加するという処理を反復させている。

3　実験手順

入力データには NII-NACSIS コレクション (1999 年度版)[3] の NTCIR1(語分割データ) を使用し、その中の各学会から集められた論文の日本語要旨 (約 33 万件) を用いた。

この入力データを、形態素解析ツール JUMAN[4] を用いて品詞情報を得る。また、語分割がなされている用語を複合語としている。これらの情報を要旨の文章に添付してデータ化をした。そしてこのデータについて SS-SANS 法の処理を行った。この時、初期条件として構文ファイルに「を 行う」という構文を 1 個入れ、用語ファイルを空にして処理を行った。この中では「複合名詞 助詞 動詞 複合名詞」というテンプレートを用いている。この実験で使用したシステムを図 1に示す。

これにより約 15 万種類の関連関係が得られた [5]。この関連関係の中で、最初の用語 (複合名詞) を目的用語 (OBJ)、もう一つの用語を主用語 (SBJ) と呼ぶことにする。今回の実験では、この目的用語と主用語の分布や割合に着目し、可視化について調べた。

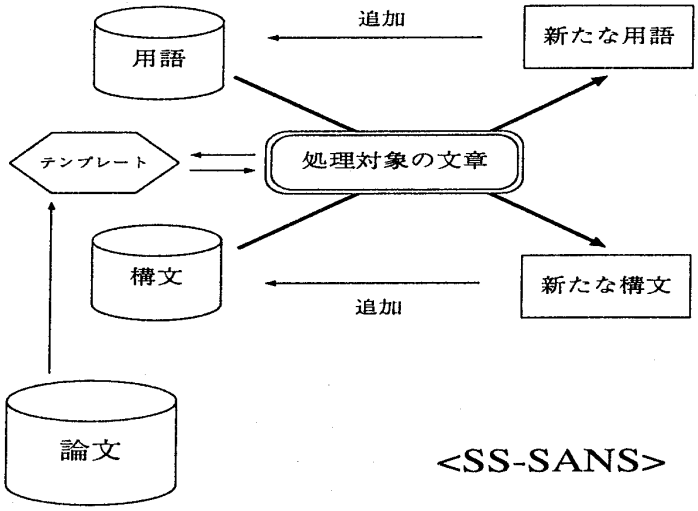

図 1: システムの概要

4 用語の分布

x 軸を目的用語として使われた回数、y 軸を主用語として使われた回数として、分布をプロットした。また、「学習」という語を含む用語の一部をプロットした。これらのグラフを図 2に示す。

5 目的用語と主用語の割合

それぞれの用語に対して、目的用語と主用語として使われた割合を求めた。実際には、全体のうち、目的用語として使われた値を求めた。以下の表は、左から 0.9 以上、0.45 ～ 0.55、0.1 以下となっている。O:S は目的用語 (O) と主用語 (S) の割合を表している。

使用数	(0.9 以上)	O:S
73	分散環境	67:6
32	陸上移動通信	30:2
32	ファジィ理論	29:3
22	最小 2 乗法	20:2
18	依存関係	17:1
18	ID 情報	18:0
16	降伏線理論	15:1
14	検索キー	13:1
11	非線形要素	11:0
10	グラフ理論	10:0

使用数	(0.45 ～ 0.55)	O:S
72	FEM 解析	33:39
52	顔画像	25:27
51	神経回路網	27:24
49	パターン認識	25:24
49	クラスター分析	24:25
32	視覚情報	17:15
30	文字認識	15:15
26	負荷分散	14:12
14	画像符号化	7: 7
10	帰納推論	5:5

使用数	(0.1 以下)	O:S
38	画質劣化	3:35
33	抑制作用	2:31
28	阻害効果	1:27
25	免疫染色	2:23
17	受信感度	1:16
16	伝送品質	1:15
13	物体認識	1:12
13	個人識別	0:13
12	画質評価	1:11
10	応答計算	0:10

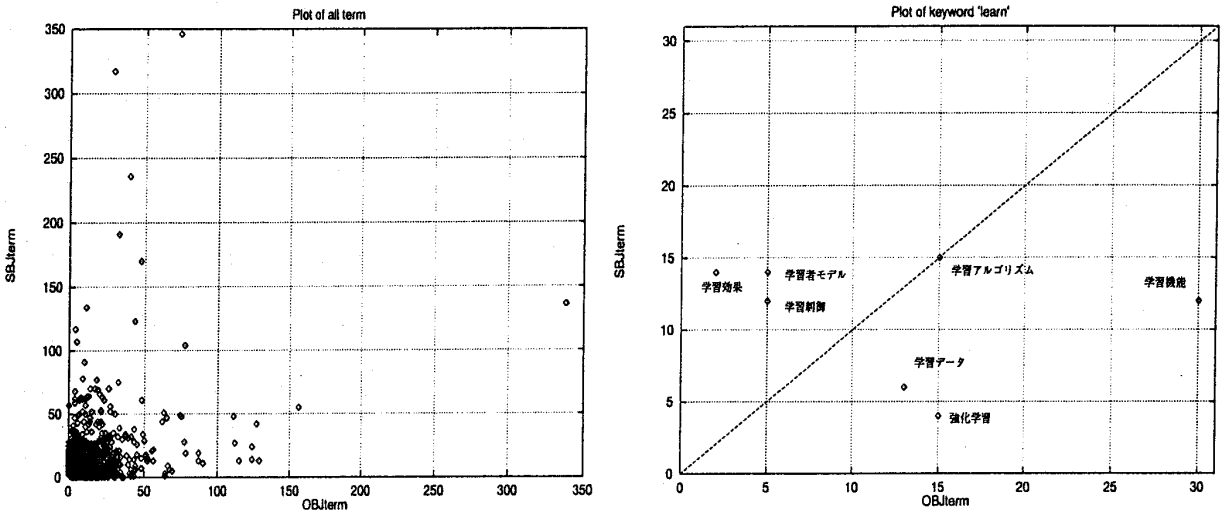

図 2: 用語の分布の様子

6 関連関係の可視化

図 3の左側は、「自然言語処理」という用語を軸に関連関係を展開をした様子である。前方の用語が「自然言語処理」の目的用語であり、後方の用語が主用語である。

また、今回の実験に用いたデータの一部を用いて可視化ツールを作成した。図 3の右側は、ブラウザ上で CGI により動作しているスナップショットである。

7 考察

用語の分布状態 (図 2を見ると、特に偏った所はなく均等な分布が現れた。また、「学習」を含む用語の分布は、ある程度の目的用語と主用語の違いが見受けられる。

目的用語と主用語の割合について、目的用語としての使われ方が全体の 0.9 以上の用語は、主に対象を表している。特に～対象、～変数という用語が多かった。これから判断すると、ほぼ確立している手法や道具、理論などを表す用語が多い。

次に、割合 0.45 ～ 0.55 の用語では、～条件、～方法、～特性、～データが多い。これらは、目的用語にも主用語にも成りうるという意味で、現在の研究分野で基礎と応用の中間に位置する研究段階の方法論や概念と言えそうである。

一方、 0.1 以下の用語では、～検討、～結果、～報告、～研究など、結果に密接に関係している用語が多い。また状態に関する用語も見られ、主に主用語として使われる用語には、現在の研究課題となっている状況を示す用語も現れた。

8 まとめ

本研究では、論文から得られた 2 つの用語間を結ぶ関係を調べ、目的用語と主用語の関係を中心に調べた。その結果、目的用語と主用語に使われる用語の違いがあることが分かり、逆にある用語に対してどちらの要素が強いかによって、その大まかな属性が分かることが言える。

今回は 2 つの用語間のみに着目して用語の特徴を分析したが、今後は 3 用語以上で形成されているような用語間の関係（～と～を用いた～、～から～と～を導く～、など）を表す文章についても同様に検討する予定である。

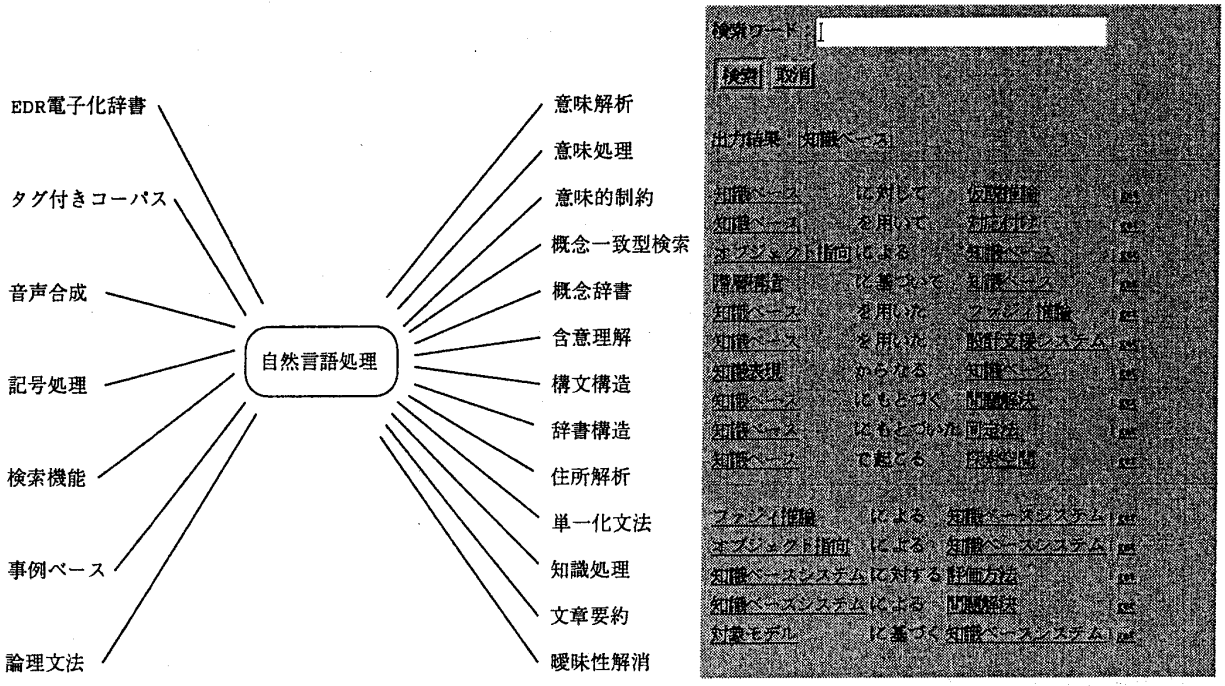

図 3: 関連関係の可視化

謝辞

本研究を行うにあたって、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] 畑口冬彦、藤原譲：SS-SANS を用いた関連関係を中心とした意味関係の抽出, 第 7 回研究報告会,pp91-94, 情報知識学会,1999

[3] NACSIS テストコレクション: http://research.nii.ac.jp/ntcir/index-ja.html

[4] 日本語形態素解析システム JUMAN： http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[5] 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 6 2 回全国大会,2001

石川大介 † 神奈川大学理学研究科情報科学専攻
藤原　譲 ⋆ 独立行政法人 工業所有権総合情報館
Daisuke ISHIKAWA † Department of Information Science, Kanagawa University
dais@goto.info.kanagawa-u.ac.jp
Yuzuru FUJIWARA ⋆ National Center for Industrial Property Information
fujiwara-yuzuru@ncipi.jpo.go.jp