SS-SANS 法を用いた意味関係の自動抽出

2X-06

石川大介 池村匡哉 近藤雄裕 杉田勝彦　森本貴之 藤原譲

神奈川大学大学院 理学研究科 情報科学

1　はじめに

本研究は、現在の計算機では処理が困難である高度な機能として、類推、帰納推論、仮説推論、発想、連想などや、更にこれらを複合した問題解決、設計、意志決定、評価を最終目標としている。そのためには、まず大量の情報を資源として、情報解析をし、それに見合う情報構造に記述し、その情報を用いて自己組織化するという第一段階がある。ここで、情報を抽出して意味関係を構造化する手法の一つとして、概念間の関係を自動抽出する SS-SANS 法がある。本研究では、この SS-SANS 法を用いて大量の情報を処理し、その結果について考察した。

2　SS-SANS 法

用語間には同値関係や階層関係などがあるが、特に因果関係などの関連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一定の構文を利用して、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用語と構文を得る。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

入力データとして、NACSIS テストコレクション (1999 年度版)[3] の NTCIR1 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。今回の実験では、「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 SS-SANS 法を行った。ここで複合名詞とは、NTCIR1 の語分割データにより、複数の名詞から成り立っている名詞列のことを示す。

この入力データを、形態素解析ツール JUMAN[4] を用いて前処理を施し、文字と品詞のみで形成されたファイルを読み込んで SS-SANS 法の処理をした。この時、初期条件として構文ファイルに「を 行う」という構文を 1 個入れ、用語ファイルを空にして処理を行った。

4　実験結果

抽出された構文数と用語数 (関連関係) の変化を図 1に示す。横軸は処理回数、縦軸は (種類の) 個数である。

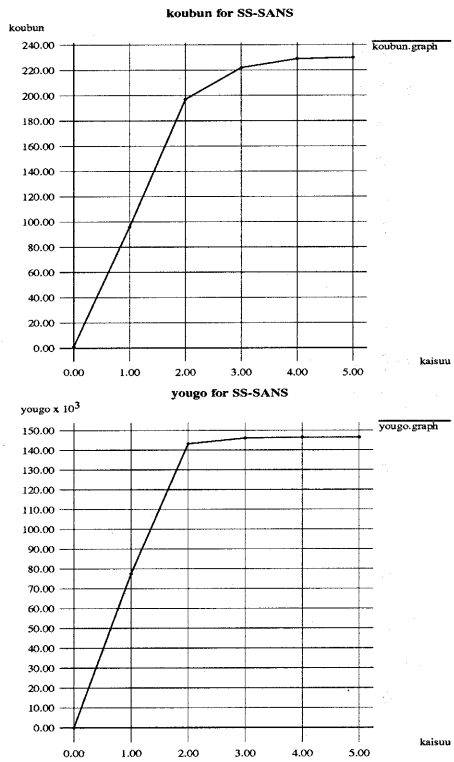

図 1: 構文数と用語数の変化

4.1　抽出された情報

抽出した用語間の関連関係は 146457 個、構文数は 230 個であり、この処理に要した反復数は 5 回であっ

Automatic extraction of semantic relationships by SS-SANS

Daisuke ISHIKAWA, Masaya IKEMURA, Takahiro KONDO, Katsuhiko SUGITA, Takayuki MORIMOTO, Yuzuru FUJIWARA

Department of Information Science, Kanagawa University

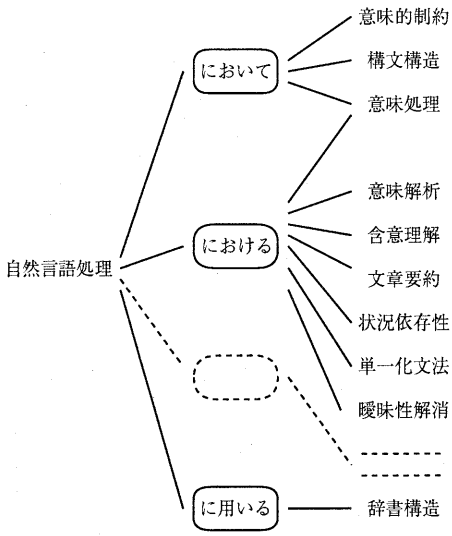

図 2: 「自然言語処理」の関連関係 (一部)

た。また用語を個々に数えると、重複していない全用語数は 182120 個であった。

関連関係で、2 個以上の構文が当てはまった用語の組み合わせは、1279 個あった。特に多くの構文に当てはまった組み合わせを以下に示す。

構文数	用語	用語
7	有限要素法	数値解析
6	説明変数	重回帰分析
5	有限要素法	磁界解析
4	有限要素法	弾塑性解析
4	有限要素法	数値計算

同様に、得られた構文の中で、特に多くの用語間を結んだ構文を以下に示す。

用語数	構文
25866	に よる
21871	に おける
12422	を 用いた
8240	と して
7807	を 用いて

抽出した関連関係の中で、「自然言語処理」という用語に着目して得られた用語間の関係を図示したものを図 2示す。

4.2 抽出されなかったデータ

得られた文章は 146686 種類あり、逆に抽出用テンプレートと同一の全文章は入力データ中に 162238 種類あった。これにより、今回の抽出結果は全体のほぼ 9 割を網羅できたことが分かった。さらにその内訳を調べると、抽出されなかった関連関係は 14365 種類と約 1 割であるのに対し、構文数は 2856 種類あり、得られた構文の 10 倍の構文が未抽出となった。未抽出の構文は、「が 好んだ」「を 食べる」などの論文中では余り使わないであろう表現の構文や、「から 割り出した」「を 振り分ける」といったかなり特殊な表現の構文で占められ、その他は構文解析エラーや品詞は正しいが日本語として間違っている構文であった。

5 考察

図 1から一つの構文から 1 回目の処理で約半分は抽出でき、2 回目で用語と構文の両方の抽出数が収束していることが伺える。

未抽出データ、特に構文については次のように考えられる。抽出された文章が 9 割に達していたことから考えて、全文章は存在する全構文数の約 1 割の構文で、全体の約 9 割の用語間を結んでいることになる。つまり、学術的な文章に使われる構文というものは、主として一部の構文を使った表現のみから形成されているからであろう。

6 まとめ

今回は 2 個の用語間のみに着目して抽出を行ったが、更に 3 個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検討して対応させていく予定である。その他、ノイズとなっている用語データの除去や、構文データに活用形の情報いれる、サ変動詞にも対応させる、などの細かい改良も挙げられる。

そして、当面の目標は C-TRAN 法や SS-KWEIC 法で既に構造化されているデータと共有できるような構造を成して、互いに違う用語関係を相互作用させていくことにより、意味理解や情報構造化を進めることである。

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