用語間の意味関係の抽出

—— SS-SANS：テンプレート構文を用いた関連関係抽出 ——

神奈川大学　理学部　情報科学科
畑口冬彦、藤原　譲

要旨

多種、多量、の情報が広域流通する情報化が、急速に進んでいる現在も情報処理機能は依然として数値計算と符号照合すなわち検索、演繹推論などが中心である。しかし情報の内容に関する高度な機能に対する必要性も強く認識されるようになってきた。

そこでと専門用語を概念の表現として捉え、その意味について記述、表現、理解、生成、処理の方式を明らかにして用語間の意味関係の抽出と、それに基づく用語による構造化知識の構築と意味理解、学習・思考機構解明の試みを報告する。

はじめに

インターネットのグローバルな普及と計算機の加速度的高性能化、低価格化で情報化がますます進んでいる。したがって情報の発信も利用も急速に多種、多様、広域化している。このことはこれまでのように数値計算とキーワード検索、演繹推論を主として行うのみでなく、豊富な情報･知識の有効な活用が求めらるようになってきた。

人間が持っている高度な思考機能をコンピュータに持たせるためには、情報を記憶させるだけではなく、意味を理解させる必要がある。そのためには情報を構造化して意味関係を記述することが必要となる。しかしながら、情報の構造化に関する体系的手法は整備されていない。本研究はそのために意味関係を抽出する方法について研究した。意味関係には、同値関係、階層関係、因果関係などの関係があるが本研究では因果関係を主たる対象とする。

意味記述には必要な意味関係を網羅的に認識、表現しなければならない。そのために用語の意味関係抽出システムの全体的な構成を図１に示す。ここで意味関係抽出と組織化のうちの SS-SANS (Semantically Specified Syntactic Analyses of Sentences) は構文解析による意味関係の抽出である。つまり思考機能の基礎となる、学習機能の一部にあたる。

対象文章としてこの実験で用いた情報は「NACSIS テストコレクション」学術情報センターのデータで、情報科学に関する多くの論文の題名と要旨に、構文解析を行ったものである。図１の網羅的な一次情報に対応する情報となる。このデータの単語の総数は 232,869 個である。

SS-SANS の対象としてこのデータを選んだ理由は、次の３点である。

1） 対象となる文章は日本語および英語なので、SS-SANS を用いる為には文章から単語を抽出し品詞情報も付加しなければならない。従って、構文解析された文章は処理が容易である。

2） 本文よりも題名や要旨の方が簡潔に記述しているので、関連関係を抽出するのに適している。

3） 専門的な用語の関連関係抽出はその分野の知識を必要とするので、情報科学の分野を選んだ。

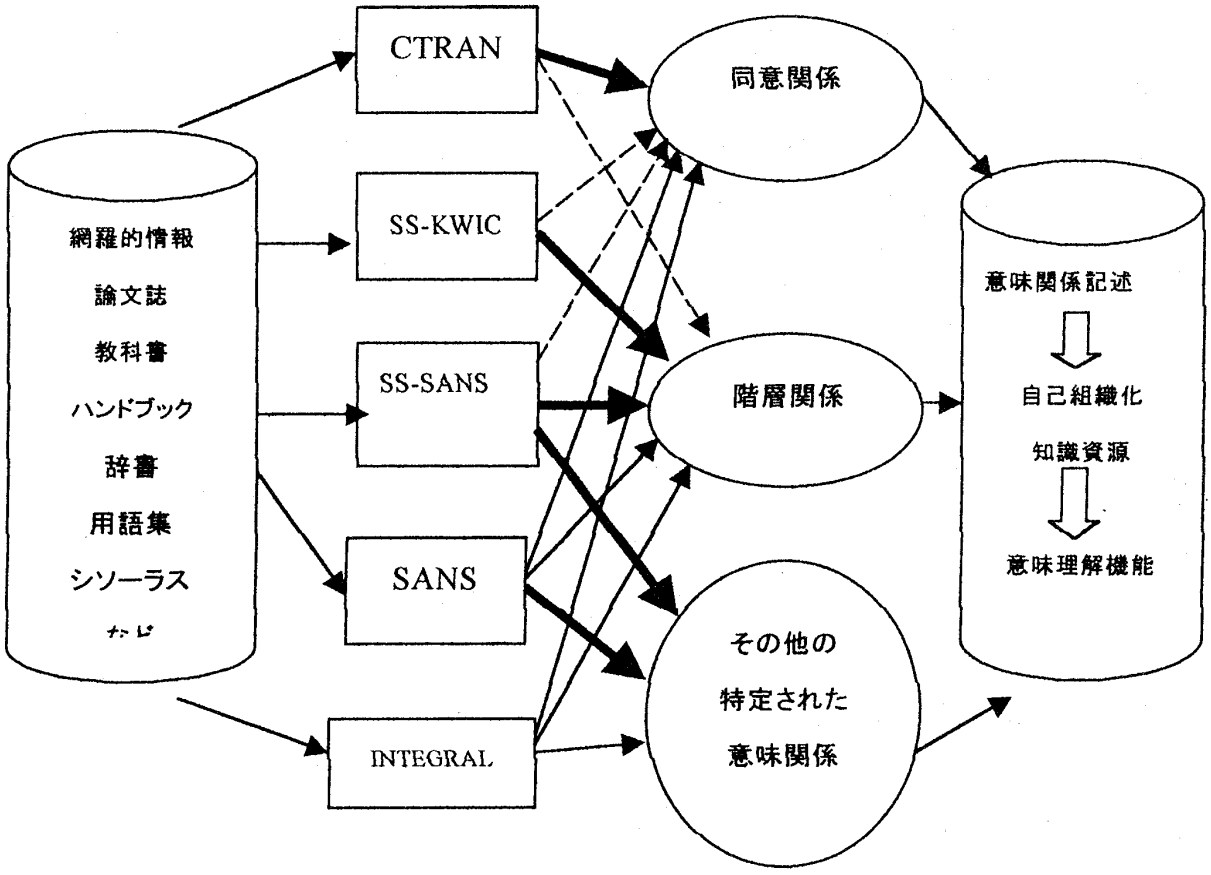

図１　意味関係抽出システムの構成

方法：最初に構文解析された文章から、SS-SANS に適した文章に変換する。不必要な品詞情報などを削除することにより、処理時間を短縮する。

処理の概要は図２に示すように１と２の行程は普通の文章を構文解析し、品詞情報を含んだ文章ファイルに変換する作業なので一度行えばよい。３～８の行程を繰り返し、用語間の関係を摘出し、新たな情報が抽出されなくなるまで行う。従って、この方法は意味関係抽出の構文パターンの学習を内臓した処理方式と言える。

5結果

表１は構文ファイルが「[NN]″を″″行う″[NN]」という情報から、文章より抽出された関連

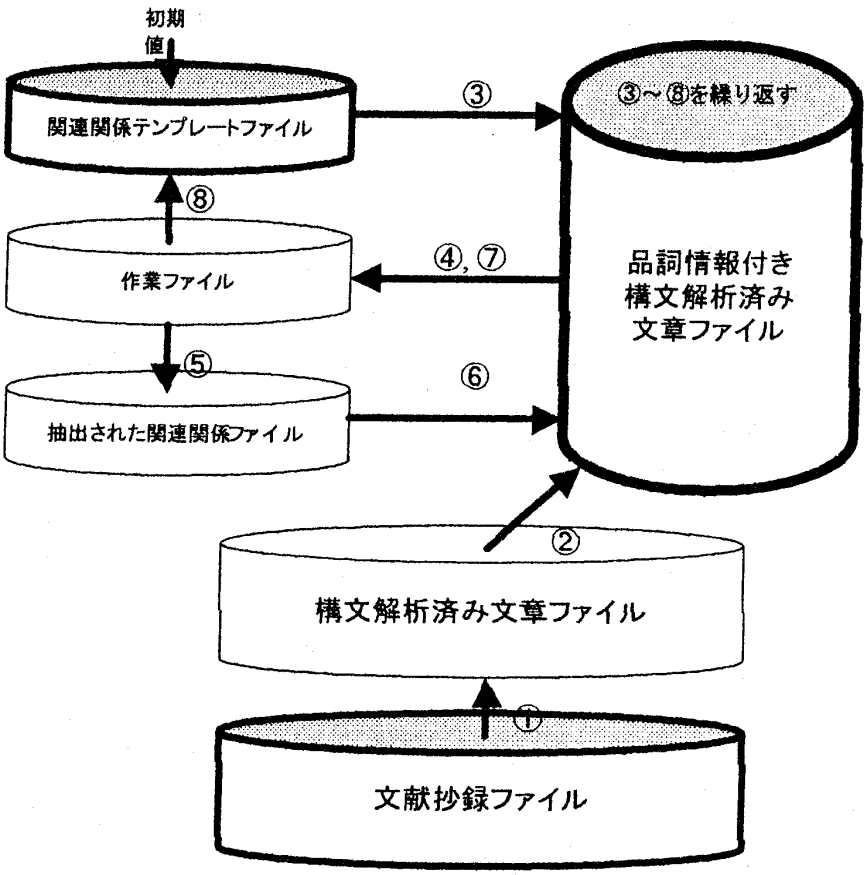

図２　システム構成、

行程の数	関連関係の数	構文の数
0	0	1
1	24	2
2	51	7
3	162	21
4	889	41
5	2026	58
6	2174	67
7	2233	73
8	2251	73

表１.抽出された関連関係と構文の数の例

関係と構文の数である。その結果、構文の数が７３個から増えていない。つまり関連関係から新しい構文が抽出されなかったのである。すると同様の構文数(７３個)からでは関連関係も２２５１個から増えないことが解る。つまりこの場合は８回以上の学習をしても成果が望めない。しかしながら、この例では十分な構文と関連関係を抽出したといえる。逆に「[NN]″と″[NN]″を″″用いた″[NN]」の例だと構文数が１個で関連関係が６個しか抽出されない。

むすび

SS-SANS による因果関係の抽出実験は、自己組織的学習を行うことにより、多くの関連関係や構文が抽出できることを示した。このことから逆に SS-SANS が学習機能を持つことにより多様な意味関係抽出機能を持つことが明らかとなった。

また今回の結果は、他の自然科学と比べると因果関係が著しく少ないことが示された。これは対象の情報が論文の全文ではなく抄録のみであったことに一因があるが、また情報科学がまだ学問として若く、知識の体系化が理工学の他分野に比し未整備であり、学問として十分に成熟していないこと、とくに理工学の特徴である実験科学的手法が活かされていないことに対応しているようである。

しかし SS-SANS により関連関係を抽出することで、C-TRAN や SS-KWIC で抽出される同値関係と階層関係をより正確なものにすることができる。一方 SS-SANS も同値関係と階層関係から、表現の多様化、抽象化などのより関連関係の抽出がより網羅的になる。異なる手法の相互間の学習が意味関係の有効な構造化を可能にすることが示された。とくに学習を円滑に進めるには、双対性、相対性、内部構造、動的構造などにも対応した構造化が必須であることも示された。

謝辞:
本研究を進めるにあたり、貴重なデータを提供いただいた学術情報センターの TMREC に感謝致します。

参考文献:

1)学術情報センター「タグ付きコパースにおける形態素・単語分割とタグ付けの基準」, 1998 (http://www.rd.nacsis.ac.jp/~ntcadm/workshop/sampleatr-d-t.html)

2) H. Sano, Y. Fujiwara, ″Syntactic and semantic structure analysis of article titles″, J. Inf. Sci. Prierciples of Practice 19 p119-124 1993

3) Yuzuru Fujiwara, ″The Model for Self Structured Semantic Relationship of Information for Its Advanced Utilization″, IFID, 19(2), p8-10, 1994

4) Yuzuru Fujiwara and Ye Liu, "The Homogenized Bipartite Model for Self Organization of Knowledge and Information", IFID 2(1)p13-17, 1998