情報検索用自己組織型シソーラスの評価

3F－1

奥村 晃、羽生田 博美、木下 哲男

新情報処理開発機構 新機能沖研究室

1. 自己組織型シソーラスを用いた情報検索

当研究室では異種知識を用いた協調問題解決のテストベッドとして協調型情報検索システムを開発中である[羽生田95]。ここで我々が目標としているのは「徐々に賢くなり、人間の意図や要求を推し量る」という新機能であり、それを実現するための機構の一つが「自己組織型シソーラスを用いた情報検索」である。

自己組織型シソーラスはシソーラス展開された単語の組合せの中から実際に検索結果に含まれる組み合わせに応じてシソーラス中の単語間にリンクを張り、次回以降の検索におけるシソーラス展開では、追加されたリンクで結合されたものを優先的に出力するものである。自己組織型シソーラスを用いた情報検索の動作概要は以下の通りである(図1)。

(1)　検索要求は自立語を複数含む日本語の名詞句である。(2)検索要求は検索コントローラによって形態素解析されて自立語が取り出され、シソーラスに渡される。(3) シソーラスは受け取った各自立語を同義語や上位/下位語等に展開した結果を検索コントローラに返す。(4)　検索コントローラは各展開結果をOR条件で結んだものをANDで結合して検索エージェントへの入力とする。(5) 検索エージェントは検索入力に従って記事を検索し結果を検索コントローラに返す。この時、検索入力中のどの単語の組み合わせで記事が見つかったかという情報も合わせて返す。(6) 検索コントローラは検索結果をユーザに返すと共に、発見された単語の組み合わせをシソーラスに知らせる。(7) シソーラスは受け取った単語の組み合わせに応じて単語間に新しいリンクを張る。これを共起関係リンクと呼ぶ。

このようにして追加された共起関係リンクを用いて、次回以降のシソーラス展開では複数の語義による展開が可能な場合に、共起関係リンクで結合された語義に限定して展開することにより検索の効率を向上させる。つまり、徐々にユーザの要求に近いものが検索されるようになる。

2. 自己組織型シソーラスの試作

自己組織型シソーラスは大きく分けて、初期シソーラス、シソーラス展開プログラム、自己組織型プログラムの三つの部分から成る。今回は評価用としてそれらの部分試作を行った。

初期シソーラスは国語辞典の語義文を解析して生成した。試作版では対象とする単語を名詞に限った。表記は標準的なもの一つに絞った。語義文は各名詞の各語義の最初の一文のみを用いた。関係としては同義語関係と上位語関係のみを扱うことにした。本来なら語義文を構文解析して上位語等を抽出するのであるが、今回は形態素解析のみを行い文末に出現する単語を上位語とした。ただし、語義文が一語だけから成る場合と同義語を示す記号が付随する場合は同義語として扱った。できあがったシソーラスは単語数が34654個、関係の数が43868個となった。

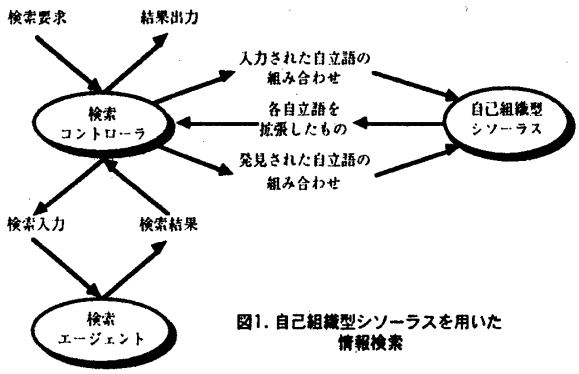

図1. 自己組織型シソーラスを用いた情報検索

An Evaluation of Self-organizable Thesaurus for Information Retrieval
Akira Okumura, Hiromi Haniuda, Tetsuo Kinoshita
Real World Computing Partnership, Novel Functions Oki Laboratory
c/o Multimedia Lab., Oki Electric Industry Co.,Ltd.
10-3, Shibaura 4-chome, Minato-ku, Tokyo 108, Japan

シソーラス展開プログラム、自己組織型プログラムはともにPrologで作成した。試作版ではシソーラスの規模が大きくないため、すべてを主記憶上のファクトとして読み込んで処理を行う。

3. 自己組織型シソーラスの評価

評価を行うためには評価尺度が必要である。自己組織型シソーラスは検索効率を向上させるために考案したものなので、評価尺度としては検索効率の向上の度合いを用いるのが最適であるが、これには次のような問題がある。

その一つは、本方式が実際の検索結果からの学習を拠り所にしているのに対し、そのような検索結果は一般に入手が困難であることである。この問題は実際によく使用される検索システムからデータを取ることができれば解決するのであるが、本プロジェクトで開発中の検索システムはまだ試作段階であり現状では多くのユーザに供用できる状態にない。

また、検索効率としては一般に再現率、適合率などが用いられるが、シソーラスを用いた検索の場合はその意味合いが曖昧である。なぜなら指定されたキーワード以外の単語を含む検索結果について、それが正解なのかどうか一概に決定できないからである。「パソコン」で検索して「ワークステーション」の記事が出力された場合、これが正解なのかどうかユーザによって、検索目的によって、またその時の気分によっても変わってくるであろう。

そこで今回は追加される共起関係リンクによって語義選択を行った結果、どの程度の絞り込みが可能となるかについて統計的に予測した結果を示す。

まず、試作した初期シソーラス中の単語で複数語義を持つものは7025個である。すべての単語を語義の数で分類した結果を表1に示す。この表を元に計算すると、語義数の単語数による加重平均が1.27であることが分かる。つまり、シソーラスによる語義選択によって検索の範囲が1.27分の1(=0.79)に狭まることになる。もしこの語義選択が正しく、ユーザの求める情報が狭められた検索範囲に含まれるのであれば、適合率は1.27倍に上がるものと考えることができる。

表1. 語義の数毎の単語数

語義の数	単語数	語義の数	単語数
1	27629	9	8
2	5577	10	2
3	1057	11	1
4	240	12	0
5	65	13	2
6	37	14	1
7	23	15	1
8	11		

4. おわりに

自己組織型シソーラスによる語義選択によってどの程度の効果が期待できるかを各単語の語義数により予測した。しかしながら、実際に効果的な語義選択方式を確立する上では様々な問題が残されている。

まず、複数の語義文から抽出された上位語が同じものである場合がある。この場合は他の単語間関係を解析しない限り複数語義として識別できない。他の単語間関係を解析するにはそれぞれ個別の解析規則を用意しなくてはならない。

また、上位語が「こと」、「もの」、「ひとつ」、「部分」などの形式的な名詞であった場合、シソーラス展開の結果としてこれらの単語を出力するのは不適切である。「こと」に係る動詞の関係する語を得るために動詞シソーラスを用意するなどの対策が必要である。

今後は、上述した問題を解決しつつシソーラスの拡充をはかり、実際の検索結果を用いてより現実的な評価実験を行う予定である。

参考文献

[奥村94] 奥村 晃、木下 哲男, 「情報検索用シソーラスの自己組織型について」, 信学技報, AI94-47 (1994-11).

[羽生田95] 羽生田 博美、木下 哲男, 「異種知識の協調利用に基づく情報検索システム」, 情報処理学会第50回全国大会, 3F-4 (1995).