概念間の意味関係の自動抽出法とその応用例

○賴 静娟†
王 暁晶†
陳 漢雄†
藤原 譲†

The Method and its Application of Automatically Extracting Semantic Relationships among Concepts

Jingjuan Lai†
Xiaojing Wang†
Hanxiong Chen†
Yuzuru Fujiwara†

Improvement of thesauri compiled automatically is reported in this paper. Translational glossaries are used as information sources, and this is a unique feature of the present method. Reliability and practicality of the method are illustrated with examples. The algorithm of the method and refinement methods are explained and the results of experiments are shown.

1　はじめに

シソーラスが情報の効果的処理と情報検索の中で主要な役割を果たしていることは周知の通りである。長い間、シソーラスの構築は殆ど大量の時間のかかる専門家の作業を必要としていたため、シソーラス自動構築法の開発が要求されるようになった。シソーラス自動構築アプローチの中でいままで主に二つのタイプがある。一つは文献からシソーラスに必要な個別関係を編集するアプローチである。もう一つは既存シソーラスの統合である。我々は用語集をシソーラス構築の情報源とする全く新しいアプローチを提案した。本文の中で、概念間の等価関係における推移則に基づく同義語自動抽出アルゴリズムを記述した後、このアルゴリズムを英日対訳用語集に適用した実験結果の実例をいくつか示す。最後に、得られた同義語集合の最適化方法を示す。

2　同義語自動抽出アルゴリズム

用語集の対訳関係を等価関係と見なすことができる。概念 e と j の等価関係をペア (e,j) で表す。同義語自動抽出のメカニズムは次のように簡単に述べる。(e_1,j_1) と (j_1,e_2) が既知の等価関係を持つペアとする。等価関係の推移則によって、ペア (e_1,e_2) が新たな等価関係を持つペアとして得ることができる。従って、推移閉包 $T^+ = \{e_1,e_2,j_1\}$ が得られる。この T^+ は同義語集合と呼ぶ。

次に、言語 A と言語 B の対訳用語集 Ω に基づいた同義語自動抽出アルゴリズムを記述する。Ω はペア (a_i,b_j) によって構成され、但し、$a_i \in A$, $b_j \in B$。言語 A と言語 B にそれぞれ対応する同義語集合は S と T とする。便宜上、S' と T' はそれぞれ S と T に対

†筑波大学 電子・情報工学系
†Institute of Information Science and Elecrtonics, University of Tsukuba

応するワーキングスペースとする。まず、初めに、Aから概念“w”を取り出し、Sに入れる。従って、SとTの初期状態S^0とT^0はそれぞれ次の通り

$$S^0 = \{w\}$$
$$T^0 = \{\}$$

次に示した各ステップを踏んで、逐次にT^iと$S^i(i = 1, 2, ...)$を計算することができる。

$$T' = \bigcup_{a \in S^i} \{b | (a, b) \in \Omega\}$$

$$T^{i+1} = T^i \cup T'$$

$$S' = \bigcup_{b \in T^{i+1}} \{a | (a, b) \in \Omega\}$$

$$S^{i+1} = S^i \cup S'$$

もし、

$$S^i = S^{i+1}$$

或は

$$T^i = T^{i+1}$$

になるとき、ちょうど

$$T^{i+1} = T^{i+2}$$

或は

$$S^i = S^{i+1}$$

になれば、このプロセスが終了する。wに対応して、言語Bが同義語集合$T^{i+1} = T^+$，言語Aが同義語集合$S^{i+1} = S^+$が得られる。

3　実例及び議論

表1

用語集サイズ	64314
同義語集合の総数	95298
サイズ1の同義語集合の数	80404
サイズ2の同義語集合の数	11308
サイズ3以上の同義語集合の数	3586
同義語集合の最大サイズ	112

同義語自動抽出実験に使用した日英対訳用語集は26個分野からなる標準工業用語集である。この用語集に基づいた実験の結果は表1に示されている。

同義語集合の総数が用語集サイズより大きい理由は一つの概念に対応する英語と日本語同義語集合をそれぞれ計数したからである。実験結果の実例として、得られた同義語集合をいくつか示す。

StartWord: 耐炎性
JapaneseSet:
　　難燃性【化学】
　　耐炎性【化学】
　　耐燃性【化学】
　　耐火性【環境、安全工学】
　　耐炎性【環境、安全工学】
　　耐火度【環境、安全工学】

StartWord: （白熱）電球
JapaneseSet:
　　白熱電球【数学物理学】
　　（白熱）電球【数学物理学】
　　白熱ランプ【数学物理学】
　　白熱電球【建築】
　　光球【建築】

実験に使用した用語集は標準化されたものにもかかわらず、例のような同義語集合が数多く抽出され、これを真の同義語集合と呼ぶ。これは表現の多様性によるものと考えられる。

言葉のもう一つの特性－－多義性によって、下の例のようなノイズを持つ同義語集合も抽出されてしまった。

StartWord: ウインドウ
JapaneseSet:
　　窓【建築】
　　ウインドウ【建築】
　　可視放射【化学】
　　光【数学物理学】

窓【数学物理学】
光【化学】
可視放射【食品技術】
可視放射【測定試験法と機器】

もちろん、多義性以外に、同形異義語によってノイズを引き起こすことも有り得るのである。

同義語集合 D における同義率 ψ_D は次の式によって計算する。

$$\psi_D = \frac{\mid D \mid - T}{\mid D \mid}$$

但し、T は D の StartWord と異なる意味を持つ概念の総数である。ψ_D が大きいほど、D の信頼度も大きい。

4　最適化

ノイズを排除するために、同義語集合に対する最適化方法を二つ提案する。

4.1　上位概念による分解

この方法には、既知の階層関係を用いて同義語集合を分割する。即ち、共通の上位概念をもつ同義語同士しか結び付けないようにする方法である。例えば、本来、三つの真の同義語集合 $S_1 = \{a, b\}$，$S_2 = \{b, c, d\}$ と $S_3 = \{d, e\}$ の中で、b と d は多義語とする。上記したアルゴリズムを実行した結果、唯一の同義語集合 $S = \{a, b, c, d, e\}$ ができてしまう。このとき、もし S_1，S_2 と S_3 に対応する上位概念 B_1，B_2 と B_3 がそれぞれ知っていれば、B_1，B_2 と B_3 によって S を S_1，S_2 と S_3 に分解することができる。この方法で、新たな真の同義語同士を得ることができる。

4.2　精度の改良

この方法は前述の方法を違って、情報源を制御することによって得られる同義語集合の精度を調整するのである。即ち、ノイズをできるだけ排除するために、基礎となる分野の範囲を限定する。近接分野を限定する基準は次のように表される。

「一方向性近接度 $\vec{\theta}_{ij}$」

$$\vec{\theta}_{ij} = \frac{\mid R_i \cap R_j \mid}{\mid R_i \mid}$$

但し、R_i は分野 i に属する概念の集合である。一般に、$\vec{\theta}_{ij} \neq \vec{\theta}_{ji}$。一方向性近接度が母

集団の大きさの問題と異義語の問題を配慮していないため、双方向性近接度を導入する方法で調整に使用する。θ^g_{ij} で幾何平均を、θ^a_{ij} で算数平均を表す。

「双方向性近接度 θ^g_{ij}」

$$\theta^g_{ij} = (\vec{\theta}_{ij} * \vec{\theta}_{ji})^{1/2}$$

「双方向性近接度 θ^a_{ij}」

$$\theta^a_{ij} = (\vec{\theta}_{ij} + \vec{\theta}_{ji}) / 2$$

5　改良の結果

上述した改良案は、分野が近ければ、概念が類似した特定の意味で使われるという考えに基づいて提出した。近接度計算式を利用して分野間の距離を決めることができる。記号 C で表された数学物理分野を用いて上述した改良案を実証した。図 1 に示されたのは一方向性近接度の結果である。但し、横軸で分野を表し、縦軸で近接度を表し、CtoOTHER で C 分野からその他の分野への近接度を、OTHERtoC でその他の分野から C への近接度を表している。双方向性近接度の結果は図 2 に示されている。但し、$R + R$ で算術平均を、$R * R$ で幾何平均を表している。

図 1

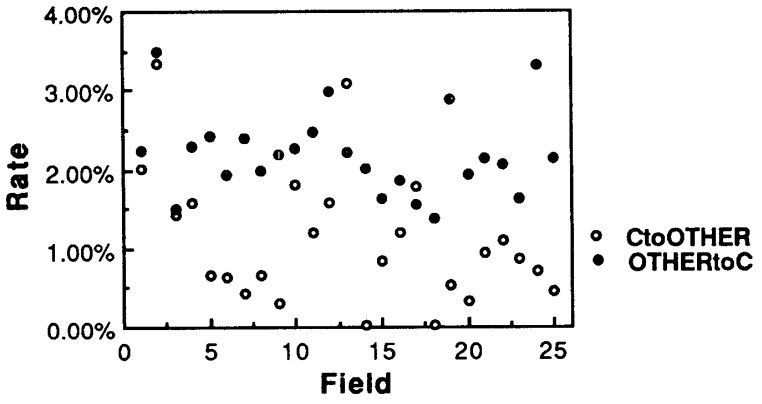
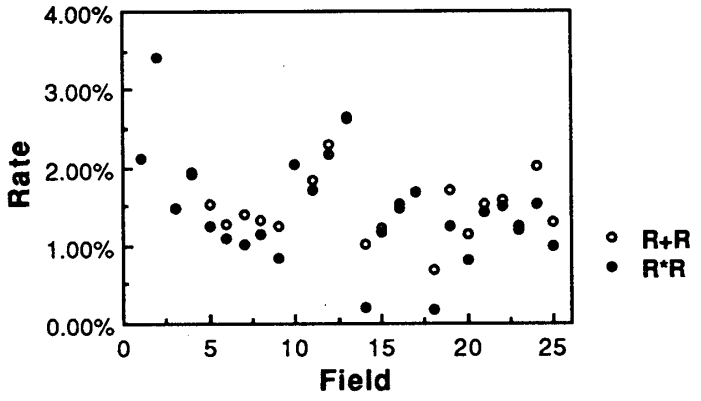


図 2



6　むすび

対訳用語集に基づく同義語自動抽出の改良アルゴリズムについて説明してきた。この方法で有用なシソーラスが得られることは実験の結果によって示された。同時に、この方法の有用性と実用性も示された。本文で報告した内容はわれわれのシステムの一部分にしか過ぎない。構築されたシソーラスは検索システムにだけではなく、類推、帰納推論および仮説推論など高度な機能のための情報構造化に利用が可能となった。

参考文献

[1] Y.Fujiwara, W.G.Lee, Y.Ishikawa, T.Yamagishi, A.Nishioka , K.Hatada, N.Ohbo and S.Fujiwara: A Dynamic Thesaurus for Intelligent Access to Research Database. Proc. of (44)FID Congres, Aug. 1988, Hels inki

[2] Y.Fujiwara, N.Ohbo, T.Itoh, M.Morita, K.Sawai, T.Kawasaki and S.Fujiwara: Multilingual Thesaurus for Internationally Distributed Information Systems. Information, Communication and Technology Transfer, 1987, pp.47-54

[3] A.Ghose and A.S.Dhawle: Problems of Thesaurus Construction. Journal of American Society for Information Science, July 1977 , pp.211-217

[4] Y.Fujiwara, J.He, G.Chang, N.Ohbo, H.Kitagawa and K.Yamaguchi: Self Organizing Information Systems for Material Design. Proc. of CAMSE'90, Aug 1990, Tokyo

[5] U.Guntzer et al.: Automatic Thesaurus Construction by Machine Learing from Retrieval Sessions. Information Processing and Managment, 25(3), 1989, pp.265-273

[6] D.Soergel: Automatic and Semi-Automatic Methods as an Aid in the Construction of Indexing Language and Thesauri. Int. Classif. 1(1), 1974, pp.34-39

[7] J.Lai, H.Kitagawa and Y.Fujiwara: Structuralization of Information by the Automatically constructed Thesaurus. Information Media, 7(4), July 1992, pp.25-32