

3 G-4

シソーラスの自動構築及び維持システムの最適化

賴 静娟 * 張 曜冬 * 藤原 譲 **

(* 筑波大学工学研究科 ** 筑波大学電子情報工学系)

1はじめに

データベースの研究と開発において最も難しい問題の一つは意味処理、とくに同義性、多義性をもつ自然言語のアクセスと管理の方法である^[1]。この問題の解決のためにシソーラスが有効であるが、従来のシソーラス構築はほとんど手作業で、かつ単語間の意味的関係の判断を専門家の膨大な時間と知的労働に頼っているため、効率的には行かず、保守も困難である。

本論文では、幅広い専門知識標準用語集を利用してシソーラスを自動的に編集、保守するための方法に関する研究結果を報告する。本研究ではさらにこの方法の最適化について分析を行っている。本研究でいう最適化とは、シソーラス作成に必要な専門知識を最大限に抽出し、多義性の影響ができるだけ避けることを指す。最適化の結果、ほぼ完全な同義語集合を得た。

2 シソーラス自動作成システムの構成

われわれは標準用語集に基づいたシソーラス自動作成法の開発を行ってきたが、その基本的アプローチは英日対訳レコードをもつ標準用語集に含まれる専門家の知識を集約して、シソーラスの作成と維持を自動的に実現することである^{[2],[3]}。開発中のシソーラス自動作成システムは、現段階では同値関係抽出、造語規則利用、語義文解析という三つの機能によって構成されている。用語集を情報源として、既知の上下関係を参照して新たな同値関係を抽出する。同時に造語規則を利用して新しい上下関係を抽出し、シソーラスを補充、拡張する。また語義文解析によってより一般的な同値関係、上下関係を抽出できるようになり、その結果さらに大規模なシソーラスに対する更新と追加も容易になる^{[4],[5],[6]}。

3 同値関係と同値関係抽出のアルゴリズム

まず本研究で取り扱われている等価関係の定義を与える。等価関係とは複数の用語が同じ概念を表しているとみなされる場合の関係のことである。言語Aの概念から言語Bの概念への翻訳作業の結果はちょうど異なる言語の概念の間に、上述した意味での等価関係の橋をかけるようなものである。この等価関係は明らかに同値関係である。この等価関係には真の同義語と準同義語がある。

言語Aから言語Bへの対訳付用語集（これから辞書とも呼ぶ）をΩとする。われわれはこのような辞書をもとにし、同値関係Rの推移則（Transitivity Rule）即ちR(cell,電池)とR(電池,battery)ならば、R(cell,battery)を利用することによって、あるスタートワードの、言語Aに属する推移閉包S⁺と言語Bに属する推移閉包T⁺を作ることができる。この閉包が等価語集合（同義語集合ともいう）である。言語Aと言語Bによる対訳付用語集Ωの基で同値関係抽出のアルゴリズムを下に記述する。

言語A、言語Bの同義語集合S,Tを以下のように定義する。まずここでS' と T' が S と T のワーキングスペースとする。任意のスタートワード s ∈ A を選べ出すと、S⁰ = {s}, T⁰ = {} にする。

ここでワーキングスペース

$$T' = \{b_j | (s, b_j) \in \Omega\}$$

をこのように与える。さらに

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

によって

$$S' = \bigcup_{b_k \in T'} \{a_h | (a_h, b_k) \in \Omega\}$$

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

次の状態

$$S^i = S^{i+1} = S^{i+2} = \dots \dots$$

になると、アルゴリズムが終了する。S⁺ = Sⁱ ⊆ A が s の同義語集合となる。同じく s の同義語集合 T⁺ ⊆ B を得ることもできる。以下、このアルゴリズムを ALGORITHM (Ω, K) と記す。K はスタートワードの集合である。

4 実験の結果

この抽出実験は 64315 個のレコードをもつ英日対訳付用語集 Ω (英語、日本語) に基づいて行った。その一例として、化学分野の日本語用語をスタートワード集合とし、26 分野の用語集を辞書とした同義語抽出実験の結果を TABLE 1 に示す。

用語の本質的な性質の一つは多義性である。多義性を考慮する上で、如何に効果的に、かつ正確に同義性を抽出できるかがもっとも重要な課題である。これはいわゆる再現率 (RECALL RATE) と適合率 (RELEVANCE RATE) の問題である。一般に再現率と適合率とは相反する関係にあり、再現率が高くなると適合率が低くなる。当然シソーラスを設計する際に必要に応じて再現率と適合率の間のバランスを考慮しなければならない。上述した抽出アルゴリズムによって真の同義語と準同義語を抽出することができた。真の同義語、即ち同一の意味を有し表現の異なる用語はほぼ完全にシソーラスの中に含めることができた。例えば、簡略形とフルネーム、複合語の分解形と非分解形、用語の直接形と転置形、標準名と俗語などである。準同義語、即ち意味的重複の大きい用語もシソーラス中に取り込むことができた。実験の結果が示すように、本論文の方法によって高い再現率と高い適合率を得ることができ、かつ有効で実用性のあるシソーラスを構築できた。

5 最適化

TABLE 1 に示したような高い適合率が得られるとはいっても、多義性に由来するノイズが混在しないわけでもない。ノイズとは同義語集合のなかにスタートワードと違う概念を表す用語のことである。このような集合をノイズセットと呼ぶ。このノイズ問題にどう対処するかが問題であろう。要求されるシソーラス精度の基準によって最適化の方法も異なってくる。

最適化の方法 1 (單一分野限定による最適化法) θ を 26 個

TABLE 1

分野・分野番号	テスト分野：化 学	
	シノニム・セット	ノイズ・セット
化学	D	473
環境工学、安全工学	G	506
材料	T	501
金属工学	U	486
化学技術	V	527
木材、紙及び繊維	W	542
エネルギー技術	J	495
食品技術及びタバコ	I	496
医学	F	497
共通分野	A	541
測定試験法及び機器	B	598
数学物理学	C	569
天文地理	E	539
農林水産業	H	482
電気技術	K	537
通信	L	523
制御、コンピュータ	M	534
機械工学	N	473
軍事技術	O	473
生産工学	P	621
運輸及び交通	Q	488
建設	R	555
鉱山工学	S	473
消費財(者)サービス	X	526
管理科学	Y	487
社会科学	Z	497

分野中の一つとする。 $K(\theta)$ を分野 θ に関する K とする。ALGORITHM ($\theta, K(\theta)$) によって同義語集合を作成すれば、最も精度の高い同義語集合を得ることができる。この方法は厳格な判断基準が要求される場合に最適である。

最適化の方法 2 (関連分野限定による最適化法) ある分野 θ に対して、 θ の関連分野を $\Sigma(\theta)$ で表すとする。ALGORITHM ($\Sigma(\theta), K(\theta)$) によって同義語集合を抽出する。この方法はある程度広義の同義関係が要求される場合に最適である。その理由は、関連分野の間で使用する用語が一致するものが多いからである。

最適化の方法 3 (同義語集合サイズの制限による最適化法) 同義語集合サイズをある定数で決めてしまうのも一つの案であるが、実験の結果、以下の二つの方法が有力と思われる。その一、同義語集合において先に抽出された用語ほど $K(\theta)$ に属する元のスタートワードに近い意味を有するという事実を利用する。抽出処理の過程で、同義語集合の中に分野 θ の用語が繰り返し出現する。そのときに何回目で抽出を中止するかを事前に決めておくことによってノイズを減らすことができる。その二、実験の結果より、同義語集合のサイズが一方的に増大するのもノイズの入った同義語集合の特徴の一つと分かったので、このような同義語集合に対して先頭の一部分だけを採用することによってノイズをかなり減らすことができる。

最適化の方法 4 (二次処理による最適化法) まず分野 θ の $K(\theta)$ に対して ALGORITHM ($\Omega, K(\theta)$) を実行する。得られた結果のノイズをチェックする。ノイズの入っている同義語セットの数がゼロであるような分野 β を取り出す。このような分野の集合を $\Sigma(\beta)$ で表す。二回目は ALGORITHM ($\Sigma(\beta), K(\theta)$) を実行すれば最適な結果を得ることができる。この方法によって、再現率を少々犠牲にするが、高い適合率を得ることができる。さらに抽出された同義語集合において異語源から各自の使い方で偶然の一一致によって発生するノイズも避けられる。

6 むすび

今回の実験結果には、標準用語集を情報源に用いて計算機による自動構成法は実用化の階段にまで達することが示され、シソーラス自動作成法における大きな進歩と言える。将来の目標は、A Iのような高度な情報システムに十分適用できるシソーラスを構築することである。

参考文献

- [1] F.W.ランカスター著、松村多美子、鈴木祐滋訳：情報システムのためのシソーラスの構築と利用。（社）情報科学技術協会、1989年
- [2] Y.Fujiiwara, W.G.Lee, Y.Ishikawa, T.Yamagishi, A.Nishioka, K.Hatada, N.Ohbo, and S.Fujiiwara: A Dynamic Thesaurus for Intelligent Access to Research Databases. (44)FID Congres, Aug. 1988, Holeinki
- [3] Y.Fujiiwara, N.Ohbo, T.Itoh, M.Morita, K.Sawai, T.Kawasaki and S.Fujiiwara: MULTILINGUAL THESAURI FOR INTERNATIONALLY DISTRIBUTED INFORMATION SYSTEMS. Information, Communication, and Technology Transfer. 1987, pp.47-54
- [4] A.Ghose, A.S.Dhawle : Problems of Thesaurus Construction. Journal of American Society for Information Science. July 1977, pp.211-217
- [5] 藤原謙、李元揆、張曉冬、北川博之、大保信夫：科学技術用語集に基づくシソーラスの自動作成。情報学シンポジウム、pp.63-69
- [6] Yuzuru Fujiiwara, Jihong He, Gyoto Chang, Nobuo Ohbo, Hiroyuki Kitagawa and Kazunori Yamaguchi : Self Organizing Information Systems for Material Design. Proceedings of - CAMSE '90, Aug. 1990, Tokyo