

1 E-5

概念構造表現の為のシソーラス自動構築と精度向上の研究

頼 静娟* 北川 博之** 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

シソーラスは情報検索における役割がよく知られているが、一般に情報資源(情報ベースと称す)においては、シソーラスは情報に内在する概念構造を表現するのにも有効である。ここでいう情報ベースは大量かつ多様な情報の蓄積、管理、検索、演繹、数値計算機能のみならず、推論、学習、帰納と言った高度な機能も提供することができる情報資源であり従来のデータベースと区別している。類推などの機能に必要な情報ベースにおける情報構造の一つ——概念構造はシソーラスによって表現される。以下では、シソーラス自動構築法による情報の構造化の実現手法を示す。

2 自己組織型情報ベース

概念構造に基づく自己組織化によって情報ベースを構築しているが、その一環として本研究を行なっている。自己組織化とは既存情報が自分自身を利用して自動的に組織化を実現することを指す。シソーラス自動構築システムによって自己組織化機能を支援する。情報ベースにおいては、物理構造、概念構造、論理構造の三つを考える。物理構造とは文章が本来もつ内部的構造を指す。科学技術文献の場合には、タイトル、著者名、抄録、各章、参考文献により構成され、章が節からできている。近年、脚光を浴びているハイパーテキストシステムは主にこういった種類の構造を取り扱っている。概念構造とは専門分野の概念(用語)が孤立的に存在せず周りの概念(用語)と同義語関係、階層関係(上下関係と部分全体関係)などで関連づけられ体系化された構造を指す。詳細は次の節で示す。論理構造は事実知識に含まれる因果関係を主とした論理関係の集合である。因果関係には次の三つの因果律が考えられる。(1)原因結果(必然性あり):ある原因によって必ずある結果が出る場合。(2)要因結果(必然性あり):いくつかの原因から必ずある結果が導かれる場合。例えば化合物反応など。(3)理由結果(必然性なし):ある理由によっていくつかの結果が考えられる場合。論理構造は原因理由を結果で分類し、タキソノミーとして表現される。情報ベースにおける類推、学習、帰納などの機能を実現する上で概念構造と論理構造は重要である。

3 シソーラスによる概念構造の表現

我々が開発してきたシソーラス自動構築システムによっ

て同義語関係、階層関係、関連関係をもつシソーラスが構成される。概念構造の同義語関係と階層関係はシソーラスで表現されることができる。シソーラス自動構築システムによって原情報の用語を対象とする概念構造が得られる。シソーラスは情報ベースにおける情報アクセスと概念間の構造化という二つの役割を果たすことになっている。概念構造を利用して概念間の距離で類似関係が表現される。距離がパスで表され、距離が長ければ長いほど類似性が小さくなる。類似性は類推機能実現の前提となる。

4 シソーラス自動構築システム

我々はシソーラスを原情報および関係情報から自動的に構築するシソーラス自動構築システムを開発してきた。このシステムはより豊富な同義語関係、階層関係、関連関係をもつシソーラスを構築する為に四つの機能を提供する。次の二節に分けて述べる。

4.1 同義語関係抽出

この機能は多言語対訳関係をもつ原情報(同一言語によるものも可)を情報源として同義語集合の抽出を自動的に行なう。

《同義語関係自動抽出アルゴリズム》

言語Aと言語Bによる対訳関係をもつ原情報 Ω のもとでアルゴリズムを記述する。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 $s \in A$ を選べ出し、 $S^0 = \{s\}, T^0 = \{\}$ にする。

次に

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

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

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

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

により、 $T^i, S^i (i=1, 2, \dots)$ を順次求める。次の状態

Automatic Construction and Revision of Thesauri for Conceptual Structures.

Jingjuan Lai, Hiroyuki Kitagawa, and Yuzuru Fujiwara.

*Doctoral Degree Program in Engineering, Univ. of Tsukuba.

**Institute of Information Sciences and Electronics, Univ. of Tsukuba.

$T^i = T^{i+1} = T^{i+2} = \dots = T^+$
になると、

$S^i = S^{i+1} = S^{i+2} = \dots = S^+$
も成立し、アルゴリズムを終了する。

$$T^+ \subseteq B$$

がB中のsの同義語集合となる。同時にAにおけるsの同義語集合 $S^+ \subseteq A$ も得られる。

4.2 その他の機能

造語規則の利用 —— 用語の造られ方における種々の組合せパターンや規則性を造語規則という。この機能は複合語の構成パターンを利用することによって単語間の階層関係の抽出を実現する。複合語の構成パターンはさまざまであり、自動化に利用できるパターンはいくつかがある。

語義文解析 —— この機能は定義付き辞書を情報源として利用し定義文構造を分析することによって同義語関係と階層関係の抽出を実現する。JIS用語辞典や岩波情報科学辞典は定義付き辞書である。

既存シソーラスの利用 —— シソーラスの構築はいつもゼロから始まるとは限らない。実際既存シソーラスの利用も大変重要である。ISOのROOTシソーラスは同義語関係が少ないが、階層関係が豊富である。われわれのシソーラス自動構築システムは階層関係の抽出が不十分なのでROOTシソーラスなどのような専門家知識が豊富な資源の利用は有用である。

5 概念構造の精度向上

この節において主に同義語集合に対する精度向上の問題を検討する。用語は同義性と多義性をもつため、上記のアルゴリズムをそのまま使うと、抽出された同義語集合には意味が異なる用語（異義語）が含まれるケースの発生を避けられない。ここで、同義語集合の精度を向上させる方法を与える。異義語を含む同義語集合を混合集合という。混合集合における同義語の割合を示す目安として同義率で表す。混合集合Dの異義語の数はTとする。同義率 ψ_D を

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

で表す。同義率が高いほど同義語集合の精度がいい。

5.1 上位概念による分解

既知の階層関係を利用して、混合集合における用語間の同値関係を判別する方法である。つまり共通上位概念で同義語同士を結び付ける方法である。三つの同義語集合 $\{\alpha, \beta\}, \{\beta, \gamma, \delta\}, \{\delta, \epsilon\}$ の中で、 β と δ は多義性を持つ語と仮定する。同義語自動抽出アルゴリズム実行の結果、単純な推移的閉包では意味の異なる三つの同義語集合が同一の集合 $\{\alpha, \beta, \gamma, \delta, \epsilon\}$ にはいることになる。

B_1, B_2, B_3 はそれぞれの同義語集合の既知の共通上位概念とすると、この共通上位概念を利用して確実な同義語同士しか結び付かないように分解することができる。この方法によって、混合集合を分解することができる。

5.2 近接分野の限定

同義語集合のサイズをできるだけ大きくし、しかも異義語をできるだけ抑えるために、適当な近接分野を定める方法がある。近接分野を定める為の判断基準を示す。分野（領域）iの用語集合を R_i で表す。

「一方向性近接度 $\bar{\theta}_{ij}$ 」

$$\bar{\theta}_{ij} = \frac{|R_i \cap R_j|}{|R_i|}$$

一般に、 $\bar{\theta}_{ij} \neq \bar{\theta}_{ji}$ 。一方向性近接度だけでは、決して十分ではない。例えば、母集団の大きさの問題、異義語の問題などを配慮すべきである。従って、双方向性近接度 θ_{ij} を導入する方法や同義率を配慮する方法などで調整に使用する。

「双方向性近接度 θ_{ij}^P 」

$$\theta_{ij}^P = (\bar{\theta}_{ij} * \bar{\theta}_{ji})^{1/2}$$

「双方向性近接度 θ_{ij}^S 」

$$\theta_{ij}^S = (\bar{\theta}_{ij} + \bar{\theta}_{ji}) / 2$$

6 むすび

同義語関係および階層関係を利用する自動構築シソーラスによる情報ベースの概念構造化を実現する方法を示した。今後の課題としては概念構造の精度をさらに向上する概念関係以外たとえば論理関係、その他の関連に対応した関係などの構造を補充する必要がある。また、専門家の持つ既存知識の利用もまだ考える余地がある。

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 Databases. (44)FID Congress, Aug. 1988, Helsinki
- [2] Y. Fujiwara, J. He, G. Chang, N. Ohbo, H. Kitagawa and K. Yamaguchi: Self Organizing Information Systems for Material Design. Proceedings of - CAMSE '90, Aug. 1990, Tokyo