同値関係自動抽出法に関する検討

森本 貴之† 後藤 智範† 藤原 譲‡
神奈川大学 理学部† 独立行政法人 工業所有権総合情報館‡

１．はじめに

膨大な情報や知識を適切に利用するためには、情報の内容に関する、より高度な処理機能が要求される。このような要求に対して、我々は以前より専門用語を最小単位とし、意味関係に基づいた概念構造の生成とその利用に関する研究を行ってきている。そしてこれまでに意味関係を自動的に抽出、統合、調整するシステム[1]および概念構造を利用するためのアプリケーションの開発[2]を行ってきている。しかしながら、現在までに完成しているプロトタイプでは自動抽出された意味関係が精度等の面でまだ十分ではないため、高度な知的処理を行うことができない。本研究では、意味関係のひとつである同値関係の自動抽出法である C-TRAN 法の意味関係抽出精度の改善に関する検討を行う。

２．C-TRAN 法の問題点

C-TRAN 法は日英対訳辞書等から得られる対訳を同値関係とし、それらを推移的に繋ぐことで同義語集合を得る手法である[3]。図 1 に例を示す。日本語の「コンピュータ」の対訳として「computer」、計算機の対訳として「computer」、「workstation」が存在したとする。このとき、推移則よりこれらすべての用語において同義語集合が得られる。以上のように、C-TRAN 法は用語の表層情報のみを用いることで同義語集合を抽出することが可能である。

しかしながら、すべての場合においてこの推移則が成立するわけではない。そして適切でない同義語集合が抽出された場合、我々の研究目的である知識・情報の構造化に及ぼす影響は無視できない。特に、推論や仮説生成といった知的処理は概念構造のナビゲーションによって実現可能であり[1]、誤った概念構造のナビゲーションは誤った知的処理に繋がる。そこで、C-TRAN 法に実際にどのような問題が存在するか調査を行い、それらの問題をまとめることにより抽出精度改善のための指針を導く。

同義語集合

コンピュータ	=	computer
計算機	=	computer
	=	workstation

図 1．C-TRAN 法のアルゴリズム

３．実験

問題点を見つけ出すため、まず実際に同義語集合の抽出を行う。用いた同値関係（対訳）データは以下のとおりである。ただし、データは日本語と英語をそれぞれ 1 用語の対（1 組と呼ぶ）にしたものである（対訳が複数存在したものはすべて 1 用語対になるように展開する）。

- 情報処理用語辞典（オーム社）の日英対訳索引：対訳は 12933 組
- コンピュータ用語辞典（日外アソシエーツ）：対訳は 34887 組

これらのデータを統合し、重複を取り除いた 38683 組の同値関係を C-TRAN 法の入力データとする。

以下に抽出された同義語集合の結果を示す。

- 対訳が 1 用語のみの同義語：17445 組
- 3 用語以上（日本語、英語の両用語を含む）から構成される同義語集合：6311 集合

3 用語以上から構成される同義語集合の分布を図 2 に示す。図 2 のグラフにおいて、縦軸と横軸はそれぞれ同義語集合の大きさ（構成用語数）とその個数（集合の数）を示す。

４．考察

3 章の実験で得られた結果を人手で解析した結

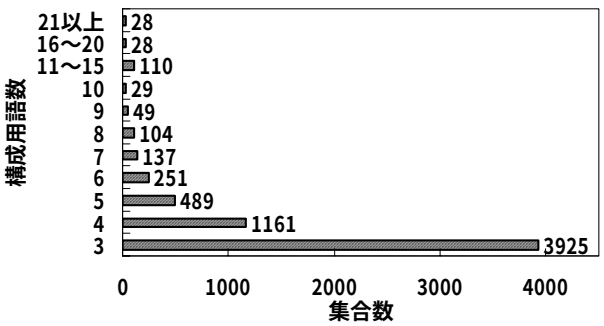

図 2．抽出された同義語集合

Improvements on Automatic Extraction of Equivalent Relationships
†Takayuki Morimoto, †Tomonori Gotoh, ‡Yuzuru Fujiwara
†Faculty of Science, Kanagawa University
‡National Center for Industrial Property Information

果、適切でない同義語集合は以下の 4 つに分けられる。以降では(1)～(4)に関して具体例を挙げながら解説する。

(1) 日本語での読み方が同じ場合
(2) 略語が同じ場合
(3) general term の場合
(4) 多義性によって意味がずれていく場合

まず(1)であるが、C-TRAN 法では用語の表層的な情報しか用いないため、表記が同じ場合は同じ用語として取り扱われる。したがって異なる英語表記の用語であっても、日本語表記（カタカナ）が同じ場合は同義語として取り扱われる。実験では「ホール（Hall, hole）」、「ラン（run, LAN）」が見つかっている。

(2)も(1)と同様に同一表記の問題で、省略表現（略語）が同じであれば同義語集合として処理されてしまう。例（CR, MT）を以下に示す。

CR：カード読取り装置（card reader）、制御レジスタ（control register）、改行復帰（carriage return）

MT：機械翻訳（machine translation）、磁気テープ（magnetic tape）

次に(3)の general term の場合であるが、一般的な用語であるが故、対訳の数が多くまたそれらの用語のもつ意味が微妙に異なることがある。実験で見つかった例としては、以下に示す「手法」を含む同義語集合（日本語と英語の用語をそれぞれ分けて記す）などが挙げられる。

日本語：手法, アプローチ,導入, 方法, 技術, 技法, 設置, インストレーション, エンジニアリング, 方法論, 設定

英 語 : approach, method, technique, installation engineering, technology, methodology, establishment, setting, settling time

この例では、「手法」には「approach」「method」「technique」の対訳がある。さらにそれらの用語が微妙に意味の異なる別の対訳を持つため、同義語集合として不適切となる。

最後に(4)であるが、これは(3)に近いもので、より専門的な用語に関しても複数の用語の同値関係を辿ることで少しずつ意味の齟齬が生じてしまう場合である。図 3、4 に異なる 2 タイプの例を示す。まず、図 3 は「せん孔段」が含まれる同義語集合に現れた用語の一部を本来の意味的にまとめたものである。これらの用語がすべて同義語集合に含まれる原因は「せん孔段」の対訳として「card row」、「punch row」以外に、上位概念といえる「row」が存在するためである。そのため、上位概念の「row」から辿ることのできる「行」、ついには「line」や「回線」まで繋がる。図 4 は図 3 と異なり、上位概念は存在しないが少しずつ意味がずれていく例である。

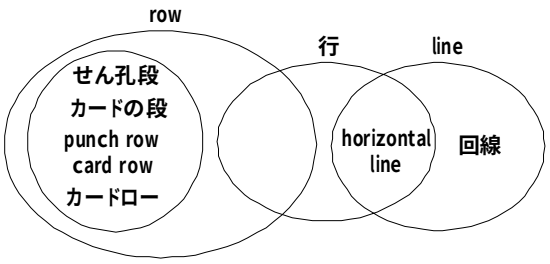

図 3．多義性の問題 (1)

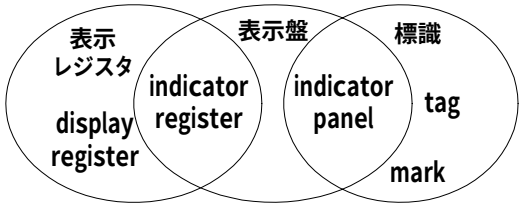

図 4．多義性の問題 (2)

ここまで述べてきた(1)～(4)の問題に対しての現在検討中である解決策について簡単に述べる。まず(1)と(2)に関しては、それぞれ元になっている用語（英語）の情報（用いられている単語の違い）を利用することで大部分に対処することが可能であると考えられる。次に(3)と(4)に関してはグラフ理論の導入による対処法の検討を行っている。

5．終わりに

情報や知識の持つ意味内容に対する高度な処理の実現に向けて、知識の構造化に関する研究を行っている。知識の構造化においては表現の多様性を取り扱うことは重要な問題である。そして、本研究はそれらを取り扱うための同値関係自動抽出における精度改善に関するものである。今後は改善手法の確立および具体的な実装を行う予定である。

参考文献

[1] 近藤 雄裕他, 意味関係抽出による概念の構造化, 情報処理学会第 62 回(平成 13 年前期)全国大会講演論文集(3), pp199-200, 2001.
[2] 森本貴之他, 構造化された知識を基にした情報検索システム, 情報知識学会第 9 回(2001 年度)研究報告会演 論文集, pp75-80, 2001.
[3] J. Lai, et al. *An information-base system based on the self-organization of concepts represented by terms*, Int. Journal of Terminology, vol. 3(2) pp313-334, 1996.