3B-6

シソーラス統合支援法の検討

加納 英文　山階 正樹　小橋 史彦
(NTT ヒューマンインタフェース研究所)

1．はじめに

近年、商用データベースばかりでなく社内データベース等の利用が拡大してきている[1)]。これらの検索システムにおいて、品質のよい検索結果を得るためには、システム管理者あるいは利用者のニーズを反映してシソーラスを効率よく構築、更新できる技術が重要である。

これらの構築、更新において、既にあるシソーラスを利用して効率よくシソーラスを生成・拡張する方法が考えられる。例えば、科学技術用語のシソーラスに特定分野の専門用語シソーラスを統合して拡張を行う場合、類似した分野のシソーラスを統合して品質を高める場合、利用者が個々に所有しているシソーラスを一つにまとめる場合等があり、そこでシソーラスの統合技術が必要となる。

本報告では、シソーラスの階層構造に着目して2つのシソーラスを統合する場合に、機械処理でどこまで支援できるかどうかについて述べる。

2．シソーラスの統合

異種シソーラスを統合する際には、双方のシソーラスに登録されている用語間の概念関係を判別して統合を行なう必要がある。本手法では、それらの関係を双方のシソーラスで一致する用語が持つ上位の概念構造によって判別する。

統合作業のレベルを考えると、①自動処理が可能な場合、②機械による支援は可能であるが、人手による判断が必要な場合に分けることができる。以下に各々の処理内容を述べる。

2．1　自動処理可能な場合

(A) 例1に示すように、双方のシソーラス(a、b)で、全く独立なトリー構造(上位語に一致語がない場合)は、それぞれを新たなシソーラスで独立したトリーとすることができる。

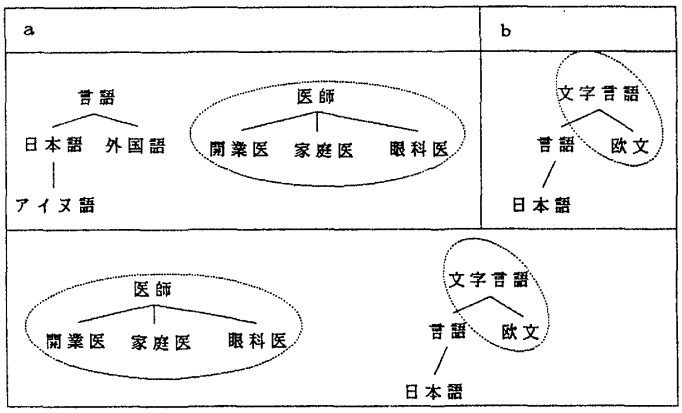
また例2に示すように、一致語があっても、その上位と下位の系列がすべて異なる場合には、多義語と判断して独立なトリー構造とみなすことができる。

(B) 同一の用語に対して双方のシソーラスで概念が完全に一致すると判別できる場合、その概念は統合して新シソーラスに組み込むことができる。

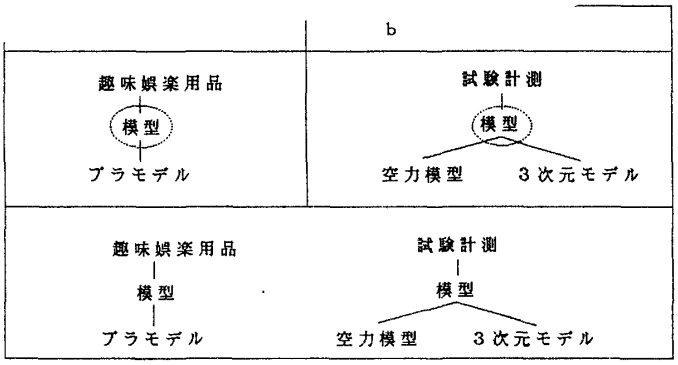
トリー構造の最上位の用語が一致する場合、その概念は完全一致すると仮定し、シソーラスの上位・下位関係では概念の包含関係が保たれることを考慮すると、トリーの上位系列が一致する用語間の概念は完全一致する。

例3において、最上位で一致する「機械部品」の

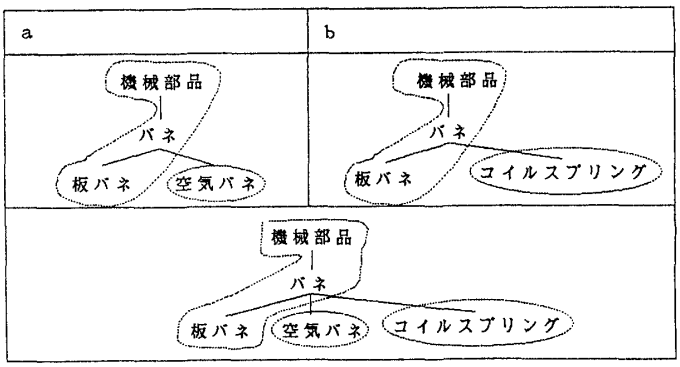
例1　独立なトリー構造を持つ場合



例2　一致語が多義語の場合



例3　概念一致が判別できる場合



A Stady of Thesauri Integration using Computer Aid

Hidefumi KANOU, Masaki YAMASINA, Fumihiko OBASHI

NTT Human Interface Laboratories

概念は完全一致し、「バネ」、「板バネ」も概念の上位系列が一致するため、自動統合が可能である。

２．２　機械支援可能な場合

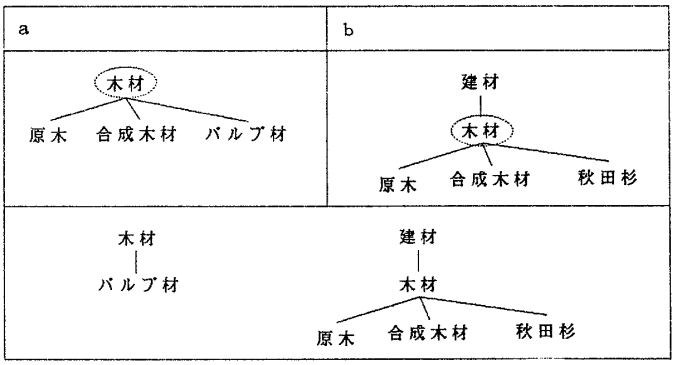
①一致語の一方が最上位で他方はそれ以外の場合
②一致語の直上位の概念が包含関係にある場合
③一致語が概念的重なり関係にある場合

には、双方のシソーラスで、同一の用語が存在しても、それらが示している概念の広がりが異なる場合があり、その関係を自動判別することは困難である。このため、これらの場合では、統合位置の候補を提示することはできるものの、最終的な判断は人手によらざるを得ない。

例４（①の場合）では、ａにおける「木材」がより広い概念を持っている。そのため、この場合には、概念の包含関係を人手で判断して統合する必要がある。この例では、ａにおける「木材」直下の「原木」「合成木材」は「建材」直下の「木材」の下に位置づけられるが、「パルプ材」については、「木材」というより広い概念の下に位置づけることになる。

なお、直上位の概念が人手の判断で完全一致と判別された場合はその直下での統合については、２．１（Ｂ）の場合と同様の処理となる。

例４　包含関係にある場合



３．概念の対応関係の分布

実験に用いるＮＥＥＤＳ－ＩＲシソーラス[2]、ＮＫＭＥＤＩＡシソーラス[3]の用語数、それらで一致する用語数は以下のとおりである。

ＮＥＥＤＳ－ＩＲの用語数：：１６，３３９語
ＮＫＭＥＤＩＡ　の用語数：：１０，２９０語
一致する用語　　：：　４，１７１語

自動処理で判別可能な概念数は、１８．８％程度であり、判別困難な概念は８１．２％程度である。

表１　一致語の概念関係

	自動判別可能	自動判別困難
語数	７８３	３，３８８

４．シソーラス統合時の処理種別による用語の分布

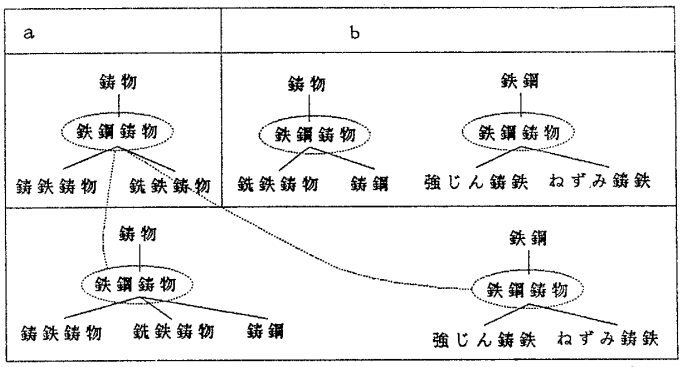
表２にＮＥＥＤＳ－ＩＲシソーラスの用語を、新たな統合シソーラスに対応づける際の処理種別による用語数の分布を示す。約４９．９％が自動処理で位置づけ可能であり、約５０．１％は機械支援はできるものの最終的には人手による判定が必要である。この中には、いま着目している語を完全一致と判断できたときには、それより下位の用語について自動的に処理でき、統合作業を効率的にできる場合もある。

表２　処理種別による用語の分布

	自動処理可能	機械支援可能
語数	８，８６１	８，８８４

なお、表２において総和がＮＥＥＤＳ－ＩＲシソーラスの用語数より大きくなっているのは、以下に例を示すように、両シソーラス間で一致する用語が１：１ではなく１：Ｎに対応しＮ個に対して統合の可能性が生じるためである。

例５　対応が１：２になる場合



５．むすび

本報告では、シソーラス統合を目的に、異なったシソーラスにおける用語の概念的な対応関係を上位構造に着目して判断する方法を検討した。その結果、商用の記事検索用シソーラスを用いた実験では、約半数の用語については自動処理で統合でき、残りの半数についても結合位置を機械支援できる見通しを得た。

今後は、本方法の高能率化をはかるため、複合語の処理法を検討するとともに、シソーラスの構築、維持管理に必要な各種支援機能の検討を行う。

［参考文献］

１）データベース白書１９８７：データベース振興センター
２）ＮＥＥＤＳ－ＩＲシソーラス：日本経済新聞社
３）ＮＫＭＥＤＩＡシソーラス：日刊工業新聞社