情報検索用シソーラスの試み

1G-4

加納　英文　　宮原　末治　　小橋　史彦

（ＮＴＴ　ヒューマンインタフェース研究所）

１．まえがき

近年、データベースシステムに対し、検索精度の向上、検索の容易化、柔軟化への要求が高まってきている。著者らは、フルテキストに対し、キーワード、および、インデックス付けなどの事前加工を必要とせず、自然言語による検索が可能な文書情報検索システムを提案している。[1)2)]

本報告では、この文書情報検索システムに必要となるシソーラスの作成について報告する。

２．シソーラスの位置付け

従来の検索システムは、統制言語方式、フリータム方式、および、それらを組み合わせた方式が一般的であった。しかし、統制言語方式は、検索効率（特に再現率）は高いが、シソーラスのメンテナンス、ディスクリプタへの変換に高度な専門知識を必要とする。また、フリータム方式は、検索作業は容易であるが、検索式の作成が困難な上、検索効率（特に再現率）があまり期待できない。組合せ方式でも、専門的知識が必要であるといった欠点を持っている。

本文書情報検索システムは、蓄積したテキスト全文を検索対象とし、自然言語による質問文よりキーワードを抽出し、抽出したキーワードをシソーラスにより同義語・類義語に展開し、拡張したキーワードで検索を行なう。検索結果には展開した同義語・類義語の意味距離、文書構造情報による評価値が与えられ、優先順位付けされて出力される。これらの処理により、専門的知識がなくとも検索効率を高めることができる。

３．シソーラスの構成

本検索システムのためのシソーラスの役割を
①同義語展開による漏れのない検索
②関係種別による優先順位の決定
③類義語などを使った検索結果の絞り込み及び拡大
の３つと考え、そのために、本シソーラスでは同義語・類義語を以下のように分類することとした。図１にシソーラスの概念図を、表１にシソーラスの関係種別とその例を示す。

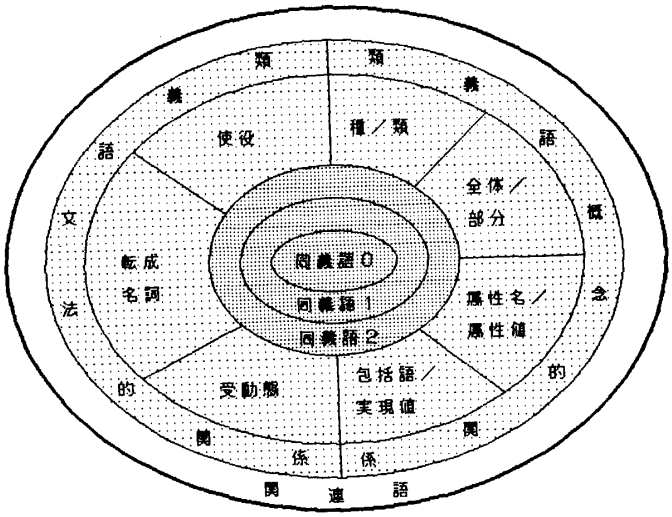

図１　シソーラスの概念図

表１　シソーラスの関係種別と例

関係	関係種別	例
同義語0	表記のゆれ	移り変わり／移り変り
	略語／完全語	ＩＭＦ／国際通貨基金
	カタカナ語／対訳語	コンピュータ／電子計算機
同義語1	言い替え語	お父様／おやじ
	カタカナ語／対訳語	イメージ／映像
同義語2	準同義語	誤字／あて字
	カタカナ語／対訳語	ヒッチハイク／無銭旅行する
類義語	種／類	衣服／背広、ジャンパー
	全体／部分	動物／頭、胴、首、手、足
	属性名／属性値	色／赤い、青い、白い
	包括語／実現値	挨拶／おはよう、こんにちは
	使役	悩む／悩ます
	転成名詞	喜ぶ／喜び
	受動態	教える／教わる
関連語	連想的関連	現品／見本、代用品、部品
	対語	教える／学ぶ

A Prototype Thesaurus for Infomation Retrieval.

Hidefumi Kanou,Sueharu Miyahara,Fumihiko Obashi

NTT Human Interface Laboratories

①同義語０：語彙の示す概念が等しく、かつ文体的価値も同一の語彙
あらゆる文脈で言い替え可能で、周囲の語彙とも文体的に調和する語彙
②同義語１：語彙の示す概念は同一であるが、文体的価値の異なる語彙
文脈によっては言い替えたときに、慣習、ニュアンスなどの違いにより周囲の語彙と不調和を起こす語彙
③同義語２：語彙の示す概念が類似してはいるが、厳密には相違する語彙
言い替えは不可能
④類義語　：語彙と語彙の間に概念的、あるいは文法的に関係がある語彙
⑤関連語　：同義語・類義語の関係はないが、ある視点に基くと何らかの関連がある語彙

同義語を０～２と３つのレベルに分けたのは、同義語を収集する中で、同義語には、いかなる場合でも同義な関係、指し示す概念だけが同義な関係、また、非常に類似した関係があり、検索結果に優先順位を与える場合に有効となると考えたためである。

また、カタカナ語／対訳語について、当初は、全て同義語０であると考えていたが、作業をしていく中で、同義語０、同義語１、同義語２があることが解り、関係種別を分けることとした。

ところで、語と語の関係においては、分野に依存した同義・類義関係と依存しない関係が存在する。そこで、シソーラスは、分野に依存しない一般語のシソーラスと分野に依存する専門語のシソーラスに分けて作成する。検索システムに適用するときは、一般語のシソーラスと検索対象とする分野の専門語シソーラスを組み合わせて使用する。

４．シソーラスの作成手順と作成規模

シソーラスの作成手順は、次の通りである。
①表記のゆれ・略語／完全語など関係種別ごとに辞書などを参考に言葉を集めファイル化する。
②収集した語彙の関係を見直す。
③収集した語彙の多義／同義を判定する。
④関係種別毎のファイルを結合する。

収集した見出し語の総計は約１６万４千語であり、表２に関係種別による語彙の内訳を示す。各関係ごとの比率は、

同義語０　：　約３０．０％　（４９，２００語）
同義語１　：　約３２．９％　（５４，０００語）
同義語２　：　約　１．３％　（　２，１００語）
類義語　　：　約１０．２％　（１６，７００語）
関連語　　：　約２５．６％　（４２，０００語）

であり、全体の６０％以上が同義語関係である。

表２　関係種別による語彙の内訳

関係	関係種別	語数
同義語０	表記のゆれ	３５，０００
	略語／完全語	４，２００
	カタカナ語／対訳語	１０，０００
同義語１	言い替え語	４０，０００
	カタカナ語／対訳語	１４，０００
同義語２	準同義語	２，０００
	カタカナ語／対訳語	１００
類義語	種／類	１６，０００
	全体／部分	３００
	属性名／属性値	１００
	包括語／実現値	１００
	使役	１００
	転成名詞	５０
	受動態	５０
関連語	連想的関連	３０，０００
	対語	１２，０００
合計		１６４，０００

５．まとめ

今回は、自然言語による、柔軟な検索を可能とするために、従来あまり集められていなかった同義語を中心に語彙を収集した。今後は、今回小量であった類義語についても収集する予定である。

さらに、検索システムによるシソーラスの評価、一般語・専門語シソーラスの結びつけ、および、シソーラス更新支援ツールの整備について検討していく予定である。

最後に、本シソーラスの作成に協力していただいた当研究所の壁谷主幹研究員、山階主任研究員、ＮＴＴ技術移転株式会社の小見係長、酒井主任に感謝致します。

参考文献

１）宮原他：文書蓄積検索システムの検討、情報処理学会第３９回後期全国大会
２）稲垣他：係り受け関係を用いた類似文書検索システム、情報処理学会第３９回後期全国大会