# 用途別シソーラス構築法の検討

7F－4

加納　英文　　宮原　末治　　稲垣　博人　　福永　博信
ＮＴＴヒューマンインタフェース研究所

## 1. はじめに

近年、情報の電子化が進み、様々な情報がデータベース化されてきている。そこで、我々はエンドユーザーが情報を効率的に検索し、利用出来るように、自然言語による情報検索システムを提案している。[1)2)3)]

この自然言語による情報検索システムでは、異なった表現による揺れを吸収するため、同義語を中心としたシソーラスを用いることとし、約１６万語の単語を収集してきた。[4)]

検索対象の情報には、新聞情報の様に様々な言葉が使われ日々刻々と変化する情報と、マニュアル等の様に使用される言葉が固定的な情報とがある。前者の情報を検索する場合には、大規模なシソーラスを必要とするが、後者の場合は、その情報中に出現する言葉に関連し、小さく最適化したシソーラスの方がシステムとして有効であると考えられる。

本報告では、この用途に合わせ小さく最適化したシソーラスを、大規模なシソーラスの助けを借りて作成する方法を検討したので報告する。

## 2. シソーラス作成法

本シソーラスの作成法は、検索対象とするテキストから単語を抽出し、マスターシソーラス群から単語マッチングにより、同義語、類義語等の関係語を抽出するものである。この時点でマッチングしなかった単語に対しては人手により関係語を付与してシソーラスを作成し、運用に際しては、先に自動抽出したシソーラス（抽出シソーラス）と人手により作成したシソーラス（作成シソーラス）を統合する事により用途別シソーラスを作成するものである。

この作成方法によれば、一般的な単語については、機械的に関係語を付与することで、人手による作業は、用途別の単語に限ることができるため、用途に合わせ最適化したシソーラスを短期間で作成することができる。

概念図を図２－１に示し、以下にそれぞれの処理を述べる。

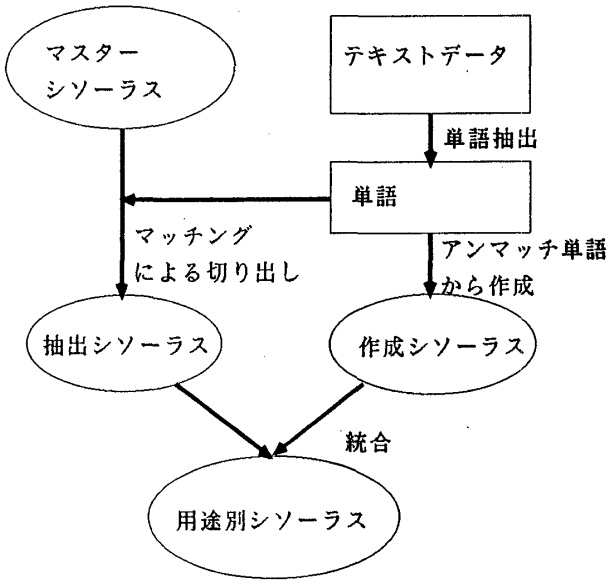



図２－１　用途別シソーラス作成概念図

①テキストからの単語抽出

テキストを形態素解析し、接続詞、副詞、非自立語等の不要語を除去する。

次に、動詞、形容詞、形容動詞の活用語尾を標準形に変換する。マスターシソーラスは、作成、あるいは、メンテナンスの容易さを考慮して、動詞は終止形、形容詞、形容動詞は連体形で持つこととする。

②マスターシソーラスからの抽出

シソーラスの見出し語と抽出単語のマッチングを取り、マッチした見出し語とその関係語を抽出する。抽出した関係語を関係種別毎に見出し語へと展開する。また、マッチしない抽出単語をアンマッチ単語として別ファイルに出力する。

③アンマッチ単語からシソーラスの作成

アンマッチ単語の中で類義語を必要とする単語を見出し語とし関係語を付与する。付与した関係語を関係語種別毎に見出し語に展開する。

---

A Study of Building a Thesaurus For Infomation Retrieval

Hidefumi Kanou　Sueharu Miyahara　Hirohito Inagaki　Hironobu Fukunaga
NTT Human Interface Laboratories

④抽出シソーラスと作成シソーラスの統合

抽出シソーラスと作成シソーラス間の重複単語を関係語と共に出力し、人手により重複単語が多義語か否かを判断し、同一語の場合は、統合する。

## 3. 用途別シソーラス作成法の評価

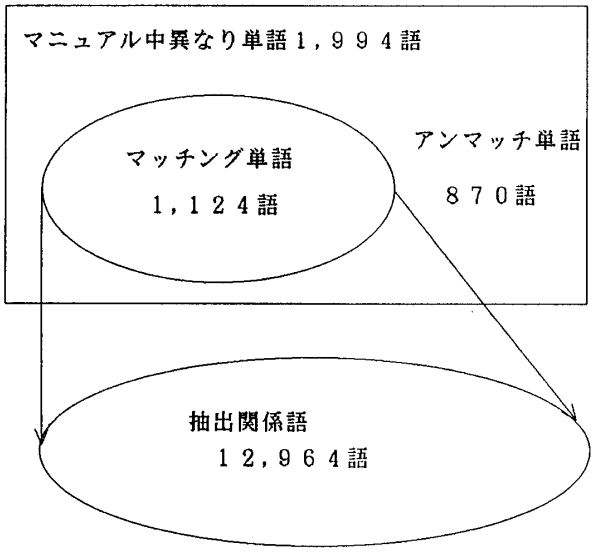
本シソーラス構築法の評価を、ワークステーションのマニュアルを対象に、現在作成中（約６０％）の一般語シソーラスから関係語を抽出する事により行った。

### 3. 1 マニュアルからの関係語の抽出

マニュアル中から抽出した総異なり単語１９９４語

マッチング単語　１１２４語（約５６．３７％）
アンマッチ単語　　８７０語（約４３．７３％）

マッチング単語から関係語を抽出したところ
関係語　１２，９６４語が抽出できた。




### 3. 2 抽出関係語の評価

マニュアル中の５文書を抽出し、文書中の異なり単語１９語について、抽出した１４５語の関係語について調査した。

適正語　　７２語（約７７％）
漏れ　　２１語（約２３％）

この漏れた関係語のうち２語は、分野に依存した関係であり、残りの１９語は、一般的な関係であった。現在進めている、一般語シソーラスの完成により自動的に抽出できる関係語は、９５％以上の精度になる見通しである。

また、抽出関係語の残りの７３語（約５０％）は、マニュアル中の単語とは多義となる語である。これらの関係語は、メモリ容量を多くするが、検索結果には影響を及ぼさないものである。

### 3. 3 アンマッチ単語の評価

アンマッチ単語を抽出調査したところ、一般語、専門語、それぞれ約半分ずつであった。このうち、一般語については、先に述べた一般語シソーラスの完成により、その約９割はカバー可能である。また、専門語については、コンピュータ分野のマスターシソーラスを作成中であり、それによるマッチングを予定している。

## 4. まとめ

用語が限定される情報検索システムのため小さく最適化したシソーラスを、大規模なシソーラスから機械的に切り出し、その不足分を人手により作成する用途別のシソーラス構築方法を検討した。ワークステーション用マニュアルから、現在作成中の一般語シソーラスを使った抽出実験を行ったところ、マニュアルからの抽出関係語について、約８割の精度で自動抽出できた。さらに、一般語シソーラスの完成により精度を９５％以上に向上できる見通しを得た。

また、アンマッチ単語についても、一般語シソーラスの完成、および、コンピュータ分野のマスタシソーラスの作成により、大幅に語数を減少できる見通しである。

最後に、本研究の機会を与えて戴いた当研究所・小橋主幹研究員、実験に協力していただいた、ＮＴＴ技術移転（株）小見係長、酒井主任、土倉社員に感謝いたします。

## 参考文献

１）稲垣ほか：係り受け関係を用いた類似検索システム：第３９回情報処理学会全国大会
２）宮原ほか：文書情報蓄積検索システムの検討：第３９回情報処理学会全国大会
３）福永ほか：語の類犠性と結合関係を考慮したテキスト検索：通信学会全国大会
４）加納ほか：情報検索用シソーラスの試み：第３９回情報処理学会全国大会