1G-7

全文探索と多様な表現

福永博信, 齋藤珠喜
NTTヒューマンインタフェース研究所

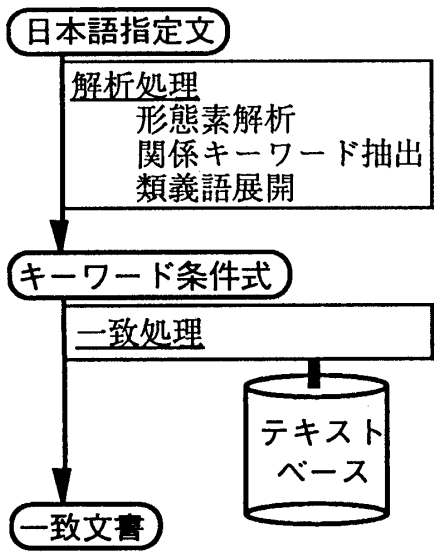
1．はじめに

筆者らは全文検索方式による計算機可読文（テキストベース）の検索システムを構築している。テキストベースの検索手法としてはキーワード法が一般的であるが、次のような欠点を持つ。1)キーワード付与作業が必要。2)検索結果は指定したキーワードに大きく依存。3)キーワードが抽出されていない情報は検索不能。これらの問題を解決するために、本システムでは全文検索方式を採用した。

2．システムの特徴

システム設計の方針を次に示す。1)検索対象は前処理を施さない生の文書ファイルとする。2)検索指定は日本語文で行い、特に制限は設けない。3)検索は実時間処理とする。4)再現性適合性の確保と高速化を目指す。

以上の方針より、検索指定文と検索対象文書間の実時間のキーワードマッチングによる検索方式を採用した。即ち、検索指定文からキーワード条件式を作成し、それを満足する文書を検索するという方式である。次にその概要を示す。



(1)解析処理
日本語検索指定文を形態素解析して、キーワード条件式を得る。

(2)一致処理
テキストベース中の各文書が(1)で得たキーワード条件式を満たすかどうかの一致判定処理を行い、一致文書を検索結果として出力する。

図1　処理概要

3．多様な表現

自然言語では一般的に同一の内容を表す表現は多数存在する。従って文の一致処理においては、それらの多様な表現ををどう処理するかがシステムの性能を左右する。多用な表現について整理すると次のようになる。

語選択	類義語	連結,結合,...
	上位語・下位語	バグを取る／デバッグする
	複合語	大小比較／大小の比較
	対義語	教える／教わる
語用法	サ変名詞	大小を比較する
	[動詞・名詞用法]	大小の比較
	名詞概念の拡張	ファイル(内容)を表示する
態	能動表現	AがBを使う
	受動表現	BがAに使われる
	使役表現	AにBを使わせる
可能表現	可能動詞	書き込める
	助動詞	書ける
	補助用言	実行できる
	複合語	実行可能だ
	可能の用言	実行することができる
否定表現	接頭語	不成立
	複合語	実行不可
	助動詞	実行せず　／　実行しない
	形容詞	更新することはない
	排他事象	真ではない　→　偽である
叙述と連体修飾		破壊的だ　←→　破壊的な

表1　多様な表現

(1)語の選択
表現に用いられている語は異なるが、同一の内容を表す表現は多数存在する。最も単純な例は、同義語・類義語を用いた表現の例である。上位・下位語の関係にある語を用いたものもある。また、日本語の場合、特に複合語を用いる例も多く見られる。

(2)語の用法
同じ語が用いられた場合でも、その用法が異なる表現がある。例えば、サ変名詞には動詞的に用いる用法と名詞的に用いる用法とがある。また、文脈によって名詞の表す概念が変化するような例も見られる。

(3)態
動作の表現では、能動態／受動態あるいは使役による表現が可能である。

(4)可能表現
可能を表すには、表1に示したような表現が可能

Japanese Full Text Search and Various Expression
Hironobu FUKUNAGA, Tamaki SAITO
NTT Human Interface Labs.

である。即ち可能の意味を含む語を用いる方法、助動詞で可能の意味を添える方法、可能を表す語(辞)と結合した語を用いる方法などがある。

(5)否定表現

否定を表す表現も、表1に示したように、否定の意味を持つ語を用いる方法、助動詞で打ち消す方法、否定を表す語(辞)と結合した語を用いる方法などがある。また、排他的な事象で否定を表すような例もある。

(6)叙述表現と修飾表現

用言と相言には、叙述的に表現する方法と修飾的に表現する方法とがある。

このように色々な表現法による多様な表現が存在する。それらの表現を日本語文の階層に対応づけると次のようになる。

	単語	文節	単位文
類義語	連結する 結合する つなぐ		
上位語 ・下位語	デバッグ する		バグを取る
複合語	大小比較		大小の比較
サ変名詞 の用法			大小を比較する 大小の比較
能動態 受動態 使役表現			AがBを使う BがAに使われる Aにを使わせる
可能表現	実行可能	実行できる	実行することができる
否定表現	未実行だ	実行せず	実行をしない

表2　表現の階層

階層をまたがる多様な表現の同一視は困難であり、現在のシステムでは対象外としている。

4.処理概要

2章で述べたように、本システムの処理は解析処理と一致処理からなる。

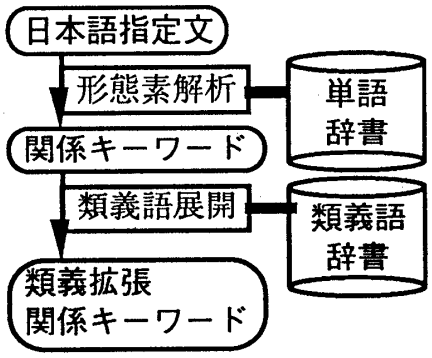


図2　解析処理部

解析処理部は、日本語の検索指定文からキーワード条件式を作る。この処理は日本語文を形態素解析して関係キーワードを作り出す部分と、関係キーワードを類義語に展開して類義拡張関係キーワード(=キーワード条件式)へ拡張する2つの部分からなる。

関係キーワードとは文から抜き出したキーワード同士を格情報(結合価)の関係で結び合わせた、構造を持つキーワードである。例えば、「文を検索する」という文からは関係キーワード　(検索　(obj　文))　が抽出される。これはキーワード「文」と「検索」が抽出され、さらに「文」は「検索」の対象であることを示している。このステップにおいて、3章で述べた多様な表現のうち、複合語，サ変名詞の動詞名詞用法，態については同一視できるように標準の形に変換する。即ち、複合語は単純語の列に、サ変名詞の名詞用法は動詞用法に、受動態は能動態に変換する。

類義拡張関係キーワードとは、類義語辞書を参照して関係キーワードの語の部分を類義語に拡張したものである。前述の例に於て、類義語辞書に「文」，「検索」の類義語として、「テキスト」,「探す」がある場合　([検索　探す]　(obj　[文　テキスト]))　なる類義拡張関係キーワードが得られる。本システムではこれをキーワード条件式として用いている。この拡張によって、類義語による多様な表現は同一視できるようになる。

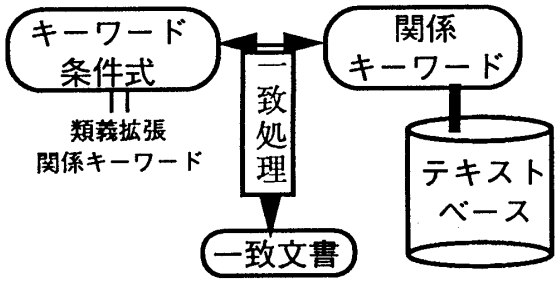


図3　一致処理部

一致処理部では、テキストベース中の各文書に対して、処理部で作り出したキーワード条件式を満たしているものを一致文書としてを抜き出す。この時テキストベースの各文書から関係キーワードを抽出し、それとキーワード条件式との間で一致の検定を行う。

5.おわりに

本稿では全文検索方式によるテキストベース検索システムについて紹介した。現在、計算機マニュアルの検索を対象として研究を行っており、LISPマシンELIS上にシステムを構築している。今後は階層をまたがる多様な表現についても同一視できるようにシステムを拡張していく。

文献

福永ほか：「語の類義性と結合関係を考慮したテキスト検索」平成1年信学春季全国大会論文集D-304