4Q－1

全文検索における例文検索

鈴木 克志, 藤井 洋一, 望月 泰行, 丸山 冬樹

三菱電機(株) パーソナル情報機器開発研究所

1　はじめに

最近、全文検索システムの開発が活発化している。しかしながら文書検索の観点からは工夫の余地が多く、特に、大量文書に対する運用においては、１回の検索結果の量が多いため絞り込みの手間が大変である。そこで、ユーザの絞り込み支援を目的として、例文検索機能の研究を行なっている。これは、全文検索結果に絞り込みをかけ、検索要求文と同じ係り受け構造（格構造）を有する文を検索する機能である。この機能により文書の中から検索要求文と類似した文を検索することが可能になり、従来の and 条件による絞り込み検索と比較して、絞り込み効果が高いものと期待される。

2　例文検索の概要

従来の全文検索では、「○○電機が何かを開発したこと」に関する内容の文書ファイルを検索しようとした場合、「○○電機 & 開発」で検索すると、２つの文字列がともに存在する文書を検索する。このとき、２つの文字列の文書中の関連は考慮されないので、検索の適合率が悪くなる。しかし、「○○電機が開発した」という文字列そのもので検索しても、「○○電機が昨年開発した、、、」という表現が検索されず、検索の再現率が悪い。すなわち、離れた文字列要素間の関連を考慮した検索が行なえず、絞り込み効果に限界があった。

3　実現方式

図１に実現方式を示す。

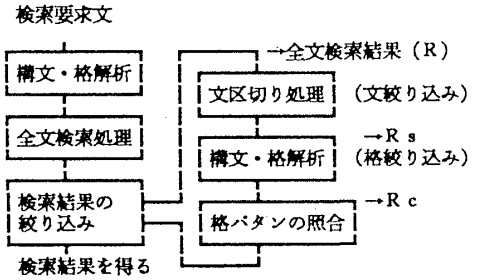

図１．例文検索

(1) 検索要求文の構文・格解析

構文・格解析処理は、二つの自立語単語の組み合わせから成る検索要求文に対して、構文解析を行ない、格関係を求める。例えば、「○○電機が開発する」という要求文からは、「動作主（開発、○○電機）」を抽出する。

(2) 全文検索処理

検索要求文の二単語に対して「単語１ AND 単語２」の条件式で全文検索を行ない、二単語が共に存在する文書を抽出する。

(3) 文区切り処理（文絞り込み）

文書ファイル中を検索し、二つの単語が同一文中に存在する部分を照合候補文として求める。例えば、「一方、○○電機が昨年に開発し、今年発売した炊飯器は、売れ行きが好調で先日増産が決定された。」このとき、シソーラスを適用し、二単語のいずれか一方の同義語も照合候補文として許容している。

(4) 格解析処理（格絞り込み）

照合候補文を構文・格解析し、二単語に対する格関係を計算する。格関係は、図２（次頁）に示すようなパタンを用いる。検索（照合）漏れを少なくするため、直接的な係り受け関係にあるもの以外に、拡張格構造として埋め込み関係にあるものも定義する。

(5) 照合処理

Search for Example Sentences in Full Text Search System

Katsushi SUZUKI, Youichi FUJII, Yasuyuki MOCHIZUKI, Fuyuki MARUYAMA

Mitsubishi Electric Corp.

検索要求文から求めた格関係と、照合候補文から求めた格関係とを照合し、成功すれば、照合候補文を検索結果と認定する。

(C1) 基本格構造（NがVする等）
(C2) 受動態など格構造が変形したもの（NによるV等）
(C3) Nの並列化、複合名詞化（～とNがVする、＊N＊がVする等）
(C4) Vの複合用言化（Nが～Vする等）
(C5) Vの埋め込み（NがVを～する、NがVに～する、等）
...　　(注) Vはサ変を含む

図2．格構造の分類の概要

4　実験と評価

既存の全文検索システムをベースに、新聞記事1年分（約65000文書、朝日新聞社提供）を対象として実験を行なった。格解析・照合ソフトは試作中であり、その部分をシミュレートする形の机上実験により行なった。図3に実験結果を示す。

	例文1	例文2	例文3	
R	35	198	234	←全文検索件数
Rs	14	31	182	←文絞り込みの結果
Rc	11	7	164	←格絞り込みの結果
Rh	12	14	180	←人間が関連文書と判断
Rs∩Rh	12	8	171	←文絞り込みが適合
～Rs∩Rh	0	6	9	←文絞り込みの漏れ
Rs∩～Rh	2	23	11	←文絞り込みのゴミ
Rc∩Rh	11	5	160	←格絞り込みが適合
～Rc∩Rh	1	9	20	←格絞り込みの漏れ
Rc∩～Rh	0	2	4	←格絞り込みのゴミ

例文1：「三菱電機が開発する」　例文2：「政府が予測する」
例文3：「合弁を設立する」or「合弁で設立する」
(注) 数字は、検索された文書の件数である。

図3．実験結果

「（名詞）Nが（サ変述語）Vする」という例文で検索する状況を評価した。「N and V」の全文検索件数Rを求め、Rの中で、NとVが同一文に存在する検索件数Rsをプログラムで求めた。さらに、Rsの中からNとVが格関係にある検索件数Rcを机上で求めた。

また、これとは別にRの中から「NがVする」という文に内容が関連する文書を人間が判断して得られた件数をRhとした。Rhは作業者の主観に依存する。検索状況としては、「ある電機メーカが何らかの開発を行なった」など3種類の状況を仮定し、関連する記事をRの中から抽出した。

その結果、3例文の合計は、Rc＝182、Rh＝206、Rc∩Rh＝176であり、適合率（Rc∩Rh／Rc）＝89．8％、再現率（Rc∩Rh／Rh）＝85．4％となった。

考察すべき点は多いが重要なものを列挙する：

(1) Rcの件数は、格解析プログラムの能力に大きく依存する。例文1の場合、プログラムが拡張格構造C5を扱わないと検索漏れが大きくなることがわかった（Rcのうち66％がC5に属す）。

(2) シソーラスの扱いが重要である。例文2の検索漏れの原因のほとんどは、「米政府が～の見通しを明らかにした」のように、同義語の定義次第（例：予測と見通し）で照合可能であった。

(3) 文絞り込みは、格絞り込みと比較して検索漏れが少なくゴミが多くなる（Rs⊃Rc）。

(4) 記事の対象分野による絞り込み効果の揺れは当然ある。例文1は、対象記事の書き方に典型性が強く、効果が大きい。例文3の検索漏れの多くは、1記事の量が平均して多く、「合弁を進めるため、共同出資で設立し…」のように記事内の表現の多様性が多かった。

(5) 二単語の共起の強さ（頻度）も影響する。

(6) 関連研究には、参考文献【1】がある。【1】では、形態素解析結果を利用して係り受け関係を抽出し絞り込みを行なっている。しかし、格解析を行なわない場合、照合不適合による検索漏れが多くなるという問題点がある。特に上記で指摘した埋め込み関係にある拡張格構造の扱いが、絞り込みの際に大きく影響する。

5　おわりに

格照合による絞り込み効果が確認され、全文検索を補完する機能としての有効性、及び自然言語処理による情報検索の高機能化の方向性を示すことが出来た。さらに広範詳細な実験評価と実利用可能な機能の切り出しが今後の課題である。

参考文献

【1】管野、他：フルテキストデータベースの技術動向、電子情報通信学会研究報告 DE90-34(1990).