3N－5

極大単語索引を用いた知的検索ソフトウェア *MEISTER* －英語文書検索への拡張－

福重貴雄　野口直彦　菅野祐司　佐藤光弘　野本昌子　稲葉光昭
{fuku,noguchi,kanno,msato,nomoto,inaba}@trl.mei.co.jp
松下電器産業（株）マルチメディアシステム研究所

1　はじめに

近年、インターネットの発達によって、英語文書検索に対する要求がますます高まってきている。本稿では、筆者らが開発中の知的検索ソフトウェア *MEISTER*[1] の、英語文書検索への拡張について述べる。

本拡張は、知的検索ソフトウェア *MEISTER* において英語文書を検索できるようにするものである。筆者らは、応用システムとして WWW 上の文書を検索する「ホームページ知的検索システム」の英語版を開発し、現在 (財) 地方自治情報センター様の「Explore Japan」(http://search.nippon-net.or.jp/english/search.html) ほかで、運用中である。

2　英語文書検索システムの全体構成

本拡張による英語文書検索システムは、基本的に、日本語文書検索システムの辞書および辞書ライブラリを英語処理用のもので置き換えたものであり、索引ライブラリ、関連キーワードライブラリ、ランキングライブラリは、日本語文書検索用のものと共通である。表 1 に、本システムの適用例における辞書構成を示す。うち、未知語辞書は、索引作成時に生成され、システム辞書 3 の複合語辞書は、4で述べる複合語抽出モジュールにより、オフラインで生成されたものである。

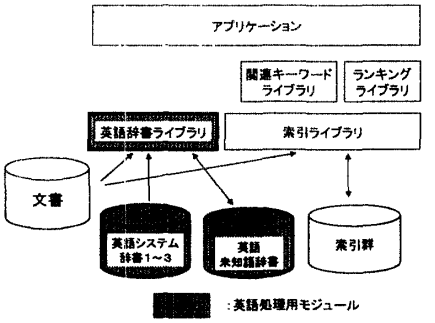

図 1　英語文書検索システムの全体構成

表 1　辞書構成例

辞書	内容	エントリ数	サイズ
システム辞書 1	英語基本語	13 万	1.8 MB
システム辞書 2	日本語地名、人名	4.5 万	0.5 MB
システム辞書 3	複合語	11 万	1.7 MB
未知語辞書	文書中の未知語 (文書サイズ: 22 MB)	3.6 万	1.0 MB

Maximal-Extension Indexing method for Smart TExt Retrieval *MEISTER*: Extension for English Language Text Retrieval.
Yoshio Fukushige, Naohiko Noguchi, Yuji Kanno, Mitsuhiro Sato, Masako Nomoto, Mitsuaki Inaba
Multimedia Systems Research Laboratory,
Matsushita Electric Industrial Co., Ltd.
4-5-15 Higashi-Shinagawa Shinagawa-ku Tokyo 140 Japan

3　英語文書検索システムの基本機能

本拡張による英語文書検索システムでは、以下の機能を実現している。

活用一致検索機能 大文字小文字は区別せずに、検索語の活用形に一致する語を含む文書を検索する。不規則活用にも対応している。

完全一致検索機能 大文字小文字は区別せずに、検索語と完全に一致する語を含む文書を検索する。

未知語処理機能 索引作成時に、システム辞書に登録していない語を未知語として抽出し、未知語辞書に登録し、ユーザによる検索を許す。

複合語検索機能 "White House" のように、空白を挟んだ語列 (複合語) による検索機能。

各種論理演算機能 (日本語検索と共通) 検索条件として複数の語を指定し、AND 検索や OR 検索など、各種論理演算を使った検索を行なう。

検索結果ランキング機能 (日本語検索と共通) 検索結果に対して、ランキングを行なう。

関連キーワード提示機能 (日本語検索と共通) ランキング結果の上位の文書から、特徴的なキーワードを抽出し、関連キーワードとしてユーザに提示する。

4　複合語処理について

(1) 複合語処理の必要性

情報検索における検索語としての複合語の重要性は、日本語の情報検索においても、すでに広く認識されている ([2] など)。

しかし、英語文書検索においては、日本語の場合以上に、複合語処理が重要になる。日本語で通常単語として扱われるような語も、英語では空白を挟んだ複合語となる場合が多く、複合語として表現して初めて意味が十分限定される場合が多いからである。たとえば、「食卓」を表す英語の表現は "dining table" であるが、"table" だけでは、「表」という意味と区別できない。

また、システムが関連キーワードとして固有名詞を提示する場合は、複合語として提示することがより重要になる。たとえば、祇園祭における「宵山祭」を表す関連キーワードとして、"yoiyama" を提示しても、祇園祭に関する知識の少ない利用者、とくに非日本語話者にとっては意味が明瞭でない。この場合、"yoiyama festival" という複合語の形式で提示することが望ましい。

表 2 に、筆者らが開発したホームページ知的検索システムで「"sports" と "event" を含むページ」を検索した場合に、関連キーワードとして表示されるもの上位 10 語 (句) までを、複合語なしの場合と、ありの場合、それぞれについて示す。

表 2　sports, event に対する関連キーワード抽出結果

複合語なし	複合語あり
arena	arena
athelete	athelete
festival	games
games	kokutai
hiroshima	national sport
kokutai	national sports festival
soccer	soccer
stadium	sports event
welfare	stadium
venue	world cup

(2) 複合語抽出モジュール

複合語検索や複合語関連キーワードの抽出には、複合語を登録した辞書が必要になるが、一般に世の中のすべての複合語をあらかじめ辞書に登録しておくことは不可能なので、対象文書に合った複合語辞書を用意することが重要になる。筆者らは、対象文書からオフラインで自動的に複合語辞書を作る複合語抽出モジュールを作成し、現在評価中である。同モジュールは、以下のように機能する。

1. 対象文書に対して形態素解析を行なう。
2. 先頭に 1 つ以下の形容詞のついた名詞連続を複合語候補として抽出する。
3. 複合語候補に対して、複合語フィルタを適用する。
4. フィルタを通過した候補から、複合語辞書を作る。

フィルタ適用前の候補には、品詞の推定間違いによるもの、名詞句の切れ目を誤ったものなど、不適切なものが多く含まれている。予備調査の結果、表 3 に示すように、とくに 3 語以上からなる複合語候補の正解率が低かったので、今回は 3 語以上からなる複合語の正解率を上げるために以下の 3 種類のフィルタを用意した。

二分木フィルタ 複合語は二分木構造を持ち、複合語全体が認定されるには、各部分構造が単語、または認定されている複合語でなければならない、とする。

共起フィルタ 二分木フィルタの条件に加え、姉妹となる部分木の主要部 (右端の語と仮定) の共起関係が 2 語対として観測されていなければならないとする。

固有名詞対応共起フィルタ 共起フィルタにおいて、左側の語 (非主要部) が固有名詞であれば、共起関係が観測されていなくてもよいとする。固有名詞かどうかは、綴りが日本語 (ローマ字) の音韻パタンになっているかで判定。

表 3　複合語候補の正解率

	候補数	正解率 (推定)	正解数 (推定)
2 語のもの	93,403	0.83	78,000
3 語以上のもの	42,894	0.65	28,000

(評価実験に使った文書データからのサンプル調査)

(3) 評価実験

フィルタの効果を見るため、日本国内の英語で書かれたホームページの中から約 22 MB (約 280 万語) のテキストデータを対象文書として選び、うち 3 語以上からなる複合語候補約 5 千候補について、人手で正誤判定の後、各フィルタ適用後の正解率・再現率を調べた。再現率については、フィルタ適用前の正解集合を全体とした。結果は、表 4 の通りである。

表 4　サンプルに対する各フィルタ適用時結果

	候補数	正解数	正解率	再現率
適用前	5,362	3,497	0.65	1
二分木	3,077	2,164	0.70	0.62
共起	983	819	0.83	0.23
固有名詞対応共起	1,491	1,196	0.80	0.34

二分木フィルタ使用時は、再現率は高いが、正解率があまり改善されなかった。一方、共起フィルタでは、正解率は高くなるが、再現率も低くなった。これに対して、固有名詞対応の共起フィルタでは、再現率をそれほど下げることなく、ある程度の正解率を確保できた。これは、国内の英語ホームページにおいては、"Ajiro fishery port" や "Matsushita Electric Industrial" のように、日本語の地名、人名を冠した表現が複合語として多く現れることによると思われる。

(4) 頻度に基づいた抽出との比較

従来、統計情報を使った複合語抽出手法が多く提案されている[3]が、今回対象としたデータに対して、頻度によって複合語候補を絞り込んだ場合の結果を図 2 に示す。

同図は、頻度 n 以上の候補集合について、正解率・再現率を表したものである。最小頻度を 3 以上にすると、正解率を固有名詞対応の共起フィルタ使用時と同程度まで上げられるが、再現率は同フィルタ使用時の 1/3 に下がってしまう。これは、今回対象としたような規模のデータに対する処理方法としては、統計情報を用いる方法はうまく機能しないことを示している。

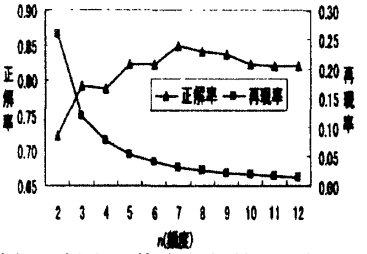

図 2　頻度 n 以上の複合語候補の正解率と再現率

5　今後の課題

今後、複合語抽出モジュールの改良や、関連語・派生語による検索機能の提供を行なう予定である。とくに、複合語抽出に関しては、今回対象にしなかった 2 語の複合語の正解率改善を目指す。また、大規模データに対する本手法の効果についても調査する。

謝辞

当研究に関して (財) 地方自治情報センター様のデータを利用させていただきました。ここに感謝いたします。

参考文献

[1] 野口直彦 他: 極大単語索引を用いた知的検索ソフトウェア *MEISTER* – 大規模文書検索への応用 –, 第 55 回情処全大 3N-1 (1997).
[2] 神林隆 他: インターネット情報探索に適したキーワード抽出, 情処研報, 97-NL-118-13 (1997).
[3] 中渡瀬秀一 他: 統計的手法によるテキストからの重要語抽出メカニズム, 情処研報, 95-FI-39-6 (1995).