極大単語索引と複合語辞書を用いた高精度な全文検索
－知的検索ソフトウエア *MEISTER* における単語検索機能の改善と評価－

4Y－7

野口直彦　　菅野祐司　　稲葉光昭

{noguchi,kanno,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、実用化が進んでいる全文検索システムでは、通常は文字列の完全一致で検索を行うため、検索ノイズが増加する傾向がある。特に、単語境界が明確でない日本語の文書を対象とする場合、これは大きな問題となる。

我々は、単語ベースの索引（極大単語索引）方式を用いて、任意文字列に対して高速な全文検索が可能な、知的検索ソフトウエア *MEISTER* を開発した[5]。*MEISTER* では、文書から単語を切り出してその出現位置と共に索引に登録するので、n-gram 索引方式では解決不可能な上記の問題を部分的に解決可能である[1]。

本稿では、ノイズの除去をさらに高精度に行うための工夫（複合語辞書の構築／利用）について述べる。また、実験的に構築した複合語辞書を用いて、精度評価実験を行ったので、その結果についても報告する。

2 極大単語索引方式と単語検索機能

MEISTER では、辞書を用い、延長関係に関して極大な単語要素のみを文書から切り出してその出現位置と共に索引に登録する（極大単語索引）。検索時は、検索文字列の単語被覆を求め、その被覆を構成する各単語の全延長語に対応した索引情報（出現位置）の連接演算を行うことで、辞書単語でない検索文字列に対しても漏れのない検索を保証する[2][3]。

さらに、*MEISTER* では、極大単語索引方式の特徴を利用し、「検索文字列の延長語による切り出し位置を選択的に排除する」ことにより、検索ノイズの除去が可能である。（単語検索機能）この機能を、検索文字列が「スキー」である場合を例に、図1にて説明する。

文書からは、極大単語のみが切り出されて索引に登録される。文書中に出現する3つの「スキー」という文字列は、各々「アスキー」「スキーマ」「スキー」「スキー場」「アルペンスキー」という極大単語として切り出される。通常の文字列検索時は、「スキー」という検索文字列に対して単語被覆を求め（「スキー」1単語からなる被覆が求まる）、その全延長語の索引情報から文書中での「スキー」の全出現位置が求まる。一方、単語検索時は、検索文字列「スキー」を真に含む延長語で切り出された位置を排除する。その結果、「アスキー」「スキーマ」の出現位置を除去し、「スキー」の極大単語としての出現位置だけを検索結果とすることができる。（検索文字列が非単語の場合にも、同様に検索できる）

しかし、図1のように、「スキー場」「アルペンスキー」のような複合語が辞書に登録されている時には、上記単語検索機能ではこれらの出現位置も排除してしまう。この種の検索漏れを引き起こすことなく、ノイズだけを確実に除去するためには、これらが「スキー」の複合語であるという情報を保持した辞書が必要となる。

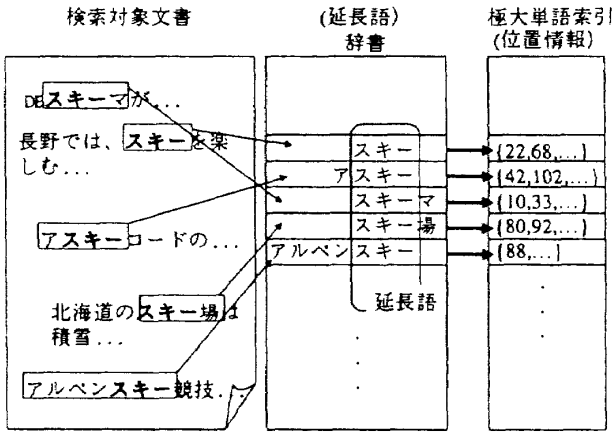

図1 極大単語索引を用いた全文検索

3 複合語辞書を用いた単語検索

ある語が複合語であるかどうか、また、どのような単位をその構成素とするべきか等々は、形態論／語構成論／統語論などに絡む、国語学上の問題であり、従来から様々な議論が行われている[4]。ここでは、そのような議論からは離れ、全文検索における高精度なノイズ除去という目的に絞り、次のような理想的状況を前提とする。

・全ての辞書単語は、単純語／複合語のどちらかである。
・各複合語は、二個以上の単純語の連鎖からなる。

この前提の下で、任意の辞書 D は、排他的な単純語集合 T と複合語集合 C の和集合として $(D = T \cup C, T \cap C = \phi)$ 定義される。今、「辞書 D が完備である」ということを次のように定義する。

【定義】C 中の各複合語を構成する単純語が全て T に含まれている時、D は完備であるという。

ある完備な辞書 D に関し、T 中の1つ以上の単純語からなる単純語連鎖に対して、それを構成素とする C 中の複合語を全て求めることができるような辞書を D の複合語辞書と呼び、CW_D で表す。以上のように構成された複合語辞書 CW_D と、D を利用して構成した極大単語索引 IDX を用いれば、次のようにして単語検索の精度を向上させることができる。

（1）検索文字列 str が D 中の単語である場合、
①CW_D を用いて str の複合語を全て求める。（$= CW(str)$）

Improving Retrieval Accuracy with Maximal-Word Index and Compound Word Dictionary,
Naohiko Noguchi,Yuji Kanno,Mitsuaki Inaba,
Multimedia Systems Research Laboratory,
Matsushita Electric Industrial,Co.,Ltd.
4-5-15 Higashi-Shinagawa, Shinagawaku,Tokyo Japan

②$CW(str)$中の各語とstrの出現位置をIDXから求め、その和集合を検索結果とする。
（2）検索文字列strがD中の単語でない場合、
①strの単語被覆を求め、その被覆を構成する各単語の全延長語に対応した出現位置をIDXから求め、それらの連接演算を行い、結果をNEとする。
②strの単純語分割を全て求め、各分割（単純語連鎖）に対しCW_0からその複合語を全て求める。（$= CW(str)$）
③$CW(str)$中の各語の出現位置をIDXから求め、それらとNEとの和集合を検索結果とする。

4 複合語辞書の構成手法

3で述べたような理想的な複合語辞書は、最終的には辞書単語を人手でチェックして構成する必要があるが、以下のような手法で、かなりの部分が自動化可能である。
（1）複合語候補／単純語候補への分類
原辞書の単語での分割を持たないものは、単純語候補として抽出し、一つでも分割を持つものは複合語候補とする。（自動分類ツール）
（2）複合語候補の単純語への分割
自動分割ツールを用いて、各複合語候補を単純語へと分割する。結果は人手でチェックする。
（3）新たな単純語の付加と単純語候補のチェック
（2）のチェックにより抽出された単純語を単純語候補として付加する。更に、単純語候補のうち長単位のものをチェック、必要ならば複合語候補として付加する。
（4）（2）～（4）を繰り返す。

今回は、（1）の自動分類ツール、（2）の自動分割ツールを作成して、以下の実験で用いる複合語辞書を比較的短時間（1週間人程度）で構築することができた。

5 実験および評価

複合語辞書を用いた単語検索の精度評価を行うために、まず、4で述べた構成手法に従って、ＥＤＲ基本語辞書（約２２万語）を原辞書とした複合語辞書を構築した。（複合語辞書Ａ）次に、ＥＤＲコーパス（約２１万文）を対象として、*MEISTER* を用いた文字列検索／単語検索を行い、その再現率／適合率を測定した。単語検索の際には、上記複合語辞書を利用する場合／しない場合を比較した。なお、検索結果の正誤判定は、ＥＤＲコーパスにおける単語分割を正解として判定した。また、ＥＤＲコーパスで認定されている単語で、元のＥＤＲ基本語辞書に含まれないものを追加した複合語辞書Ｂも構築し、比較実験を行った。

以下に、特に複合語辞書を利用する効果が大きいと思われるカタカナ語についての実験結果を示す。表１は、カタカナ３文字からなる単語についての検索結果であり、数値欄は、**正解数／検索箇所数**を表す。複合語辞書Ａを利用した時、利用しない場合に比べて再現率、適合率共に大幅に向上することが分かる。また、複合語辞書Ｂを用いた場合は、再現率、適合率共にほぼ１となった。図２は、表１の各々のケースを再現率－適合率グラフにプロットしたものである。これから、次のことが言える。

① *MEISTER* の単語検索機能を用いることで、適合率は向上するが、再現率はやや下落する（検索漏れが生じる）
② しかし、複合語辞書を用いることで、再現率を落とすことなく、適合率の向上が図れる
③ 辞書単語を増やしていくと、ナイーブな単語検索では再現率が更に下落（適合率は上がる）するが、これは、複合語辞書を新たに構築することで防ぐことができる

検索文字列	文字列検索	複合語辞書Ａ		複合語辞書Ｂ	
		非利用	利用	非利用	利用
スキー	136/327	96/166	136/207	80/ 80	136/136
スパイ	85/113	76/ 80	85/ 89	37/ 41	85/ 89
テスト	360/390	336/338	360/362	232/232	360/360
ブルー	45/114	22/ 86	45/109	11/ 11	45/ 45
ライブ	19/212	9/ 15	19/ 25	5/ 5	19/ 19
キング	14/208	8/ 32	12/ 36	3/ 3	14/ 14
チーム	378/382	347/349	375/377	291/291	375/375
ロック	143/501	119/177	142/200	88/ 88	143/143
再現率	1.000	0.778	0.998	0.522	0.996
適合率	0.454	0.837	0.866	0.997	0.997

表１ カタカナ３文字語（3583語）の検索結果

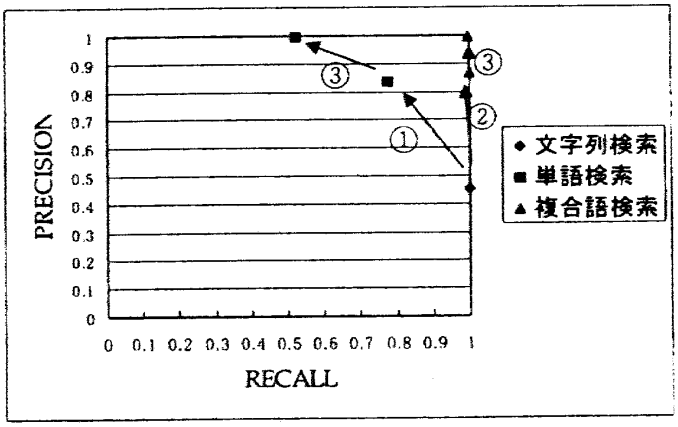

図２ カタカナ３文字語の検索における再現率／適合率

6 おわりに

本稿では、n-gram 索引等、従来の全文検索手法では困難だった検索ノイズの除去を高精度に行うしくみについて述べた。５で示した実験結果は、形態素解析等を行う手法と比べても遜色のない精度である。また、形態素解析を行う場合、検索精度はその解析精度に依存するので、精度向上には解析手法／辞書等の複雑なチューンアップが必要である。一方、本手法では検索精度は複合語辞書の精度のみに依存し、単語を増強するだけで改善していくことが可能である。今後は、複合語辞書構築の更なる自動化と、複合語検索の高速化を図っていく。

参考文献

[1]稲葉光昭 他：極大単語索引方式を用いた知的検索ソフトウエア *MEISTER*－ランキングライブラリの機能と特長－，第５５回情処全大，3N-3(1997)．
[2]菅野祐司 他：極大単語索引方式を用いた知的検索ソフトウエア *MEISTER*－辞書・索引ライブラリの機能と特長－，第５５回情処全大，3N-2(1997)．
[3]倉知一晃 他：日本語文書に対する新しい索引検索方式 －索引作成と検索の原理－，第５０回情処全大，4F-2(1995)．
[4]斎藤倫明他（編）：「語構成」，ひつじ書房(1997)．
[5]野口直彦 他：極大単語索引方式を用いた知的検索ソフトウエア *MEISTER*－概要－，第５５回情処全大，3N-1(1997)．