形態素解析性能の検索システムに与える影響

多田智之　金岡秀信
オムロン株式会社　IT研究所
{tada, kana}@ari.ncl.omron.co.jp

1　はじめに

分かち書きのされていない日本語テキストを高速に検索する検索システムは、インデックスに登録する対象によって2種類に大別できる。

1)単語：形態素解析などによって分割された単語
2)文字列：文字n-gramなどの文字列

形態素解析の問題として、解析に失敗や曖昧性がある、辞書・文法規則の保守に労力を要する、などがあげられ、実装の容易な2)の方式が主流を占めてきた。

しかし、最近は形態素解析を組み込んだ1)の方式の検索システム製品も増えてきた[1][2][3]。自然言語入力、類似文検索、検索キーワードの自動追加、検索結果のクラスタリング、といった新しい機能では、検索要求として入力した文字列や検索対象の文章を意味のある文字列単位(=単語)に分割する必要があるためと思われる。また今後、単語処理に基づく更に高度な検索機能が出現すると予想される。

筆者は、単語ベースの検索技術をサポートすることを目的に、「**形態素解析の解析精度がある程度まで落ちても、テキスト情報検索システムの検索精度に大きな影響を与えない**」という仮説を立て、検証する。この仮説は、「形態素解析が単語を切り間違える場合でも、入力した検索文字列を同じ形態素解析で解析することにより同じ箇所で分割され、検索時に結局一致するのでは？」[4]という示唆が元になっている。

2　検証の環境

検索精度の評価にはBMIR-J1[5]を使用した。単なる文字列一致を正解とするのではなくBMIR-J1のような主題を考慮した正解で判定する必要と考えた。質問は60問全問を用い、適合性判定ではAランク、Bランクの両方を正解とした。精度の評価指標としてはPrecision / Recallを使った。

検索システムはWAIS-8b5.1[1]を用いた。このシステムの適合度計算法は極めて単純で、ある文書の適合度R(d)は以下の式で表わされる。

$$R(d) = \frac{\sum_{i=1}^{n} tf(i)}{S_{max}}$$

tf(i)：キーワードiの出現回数
n：検索入力に含まれるキーワード数
Smax：キーワード出現回数累積の最大値

すなわち、検索キーワードの出現回数のみで算出している。このように単純な適合度計算法を選択したのは、評価結果の考察を容易にするためである。

形態素解析システムは、オムロンが開発した高速な日本語形態素システムSuperMorpho-J[6]を使用した。形態素解析精度は使用する単語数を変化させることによりコントロールする。

3　形態素解析とn-gram

形態素解析の精度を変化させながら検索精度を評価する前に、比較実験としてn-gramをインデキシングする方式、複合語をインデキシングする方式を評価する。

文字種を考慮しない単純なn-gram方式としては、1-gramと2-gramの複合インデキシングが最も検索精度が高いという報告があるが[7]、今回使用したWAISの単純な適合度計算法では、2-gramのみのインデキシングの成績が良かったため、これを比較対照とする。1-gramと2-gramの複合インデキシングの成績が悪い理由は、1-gramの「の」のようなidfが小さい文字が主題と関係ない記事にも数多く含まれていて、Precisionを下げてしまうためである。

また、複合語をインデキシングする方法としては、形態素解析した単語の中で、助詞を挟まずに隣接する名詞相当の2単語を連結して1つの複合語とみなし、単語とともにインデキシングした。

[1] 京都大学の形態素解析 JUMAN で日本語化され一般配布されているバージョン。

図1にRecall 11点でのPrecision[8]の曲線を示す。また表1にPrecisionの平均値(non-interpolated)[8]とインデックスに含まれる語数を示す。Precisionの値はいずれも60問の平均値である。

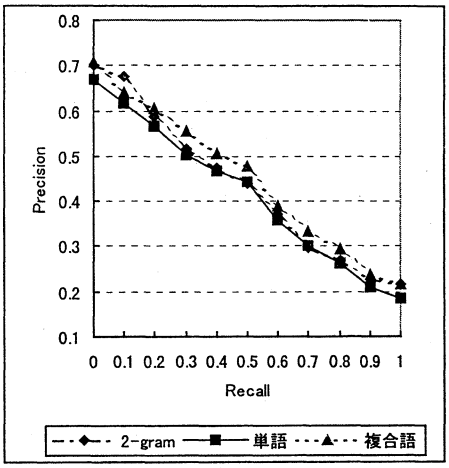

図1：単語、2-gram、複合語の検索精度

表1：単語、2-gram、複合語の精度と語数

	Precision 平均	異なり語数	総語数
単語	0.403	15,967	146,468
2-gram	0.422	75,815	442,897
複合語	0.438	38,327	190,098

この結果と論文[7]の結果と比較しながら考察する。論文[7]で用いられている形態素解析は8万語というかなり小規模なもので、この実験では比較の対象とはならない。それよりは論文[7]で提案されているタグ付きコーパスで学習する単語分割法 “Simple indexing based on segmentation proba-bilities” が比較的精度の高い形態素解析に、”Overlapping indexing”が複合語インデキシングに、相当すると見なす方が妥当である。そう考えると、この実験で得られた結果 精度が若干の差で、

複合語 > n-gram > 単語

となったことと整合が取れる。

また、この実験では複合語インデキシングを行なった場合でも、総語数は単語インデキシングの1.3倍に収まり、n-gramの総語数よりも少ない。

4 形態素解析の精度変化

形態素解析の解析精度が変化した場合、検索精度にどのような影響が出るかを調べるために、形態素解析の単語を減らして意図的に形態素解析精度を落とす。単語を減らす手続きは、単語の重要性、出現頻度を考慮せず全くランダムな方法で行なった。また、形態素解析精度の評価は、BMIR-J1の質問60問分の文章と、検索対象の記事600件の中から正解記事30件を選んで対象とした。表1に単語数と検索精度の関係を示す。

表2：単語数と形態素解析精度

単語数	質問での精度	記事での精度
17万	1.00	0.976
15万	0.831	0.888
13万	0.760	0.812
11万	0.645	0.722
9万	0.550	0.642

形態素解析の精度は以下の計算式でおこなった。

$$\text{精度} = \frac{\text{形態素解析が出力した正解単語数}}{\text{目視検査で決定した正解単語数}}$$

正解データを作成するために目視で検査したところ、質問60問には242単語が、記事30件には8549単語が含まれていた。全くランダムな方法で単語数を減らしたため、解析精度の低下は著しい。

この実験では、形態素解析に失敗した場合の処理が重要となる。検索漏れを少なくするために、未登録語を含む同じ文字種の連続する範囲を1単語とするのではなく、短めに分割されるようにした。すなわち、たまたま部分文字列に一致した単語を優先とし、何の単語も一致しなかった部分は文字種に応じて以下のようにしている。

漢字、ひらがな：1文字を1単語とする

例)「北陸」「地方」が辞書になく「陸地」がある場合、「北陸地方」は「北」「陸地」「方」

カタカナ：一連のカタカナを1単語とする

例)「ファクシミリ」が辞書になく「ミリ」がある場合、「ファクシミリ」は「ファクシ」「ミリ」

5 検索精度への影響

上記の形態素解析を検索システムに適用し、検索精度を評価する。検索精度はPrecisionの平均値

(non-interpolated)の 60 問平均で比較する。その結果を図 2 に示す。

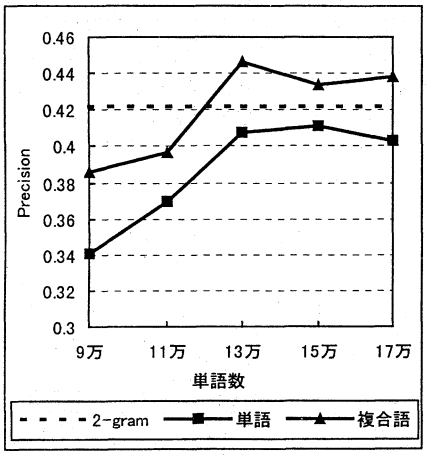

図 2：形態素解析の精度と検索精度

この図で分かるように、単語を 13 万語にまで減らしても検索精度はあまり変化せず、更に単語数を減らすと徐々に検索精度が低下する。これは筆者の立てた仮説と一致する振る舞いをしていると言える。

形態素解析の精度が低くなるのに、途中までは若干検索精度が上がる傾向があるのは、単語が短く分断されて結果的に正解文書に一致するキーワードの数が増えてしまうためである。更に形態素解析精度が下がって分断される単語が増えると、今度は誤って一致するキーワードの数が上回ると考えられる。

6 おわりに

単語や n-gram をインデキシングした全文検索システムで、主題に着目した評価テストコレクションをによって Partial Match 方式の再評価を行ない、形態素解析法式は n-gram と比較して遜色が無いことを確認した。また、形態素解析の精度を意図的に落とし、ある程度までは形態素解析精度が落ちても検索精度に影響が無いことを確かめた。

この結果を更に確かめていくためには、もっと統計的に信頼できるデータを集める必要がある。最近一般公開された評価テストコレクション BMIR-J2 [9]では、5080 件という統計的に十分な検索対象記事が用意されている。また、結果が辞書に含まれる単語にも左右されやすいと思われるため、統計的に評価できるよう いろいろなケースで実験するべきである。

今回は検索の適合度計算の方法として最も単純な方法を用いたが、Vector Space Model や Probabilistic Model など高度な適合度計算法を適用してみる必要がある。

ただし BMIR-J1/J2 の要求する正解は、質問文章中に現れる表層的なキーワードだけでは正確に発見しきれない。ユーザが本当に欲する情報を得るための検索システムはどのようなものかを示唆している。統計的な手法の改善を進めていくだけでなく、言語処理をベースとしたもっと高度な検索手法を開発していく必要がある。

謝辞

株式会社 日本経済新聞社の協力によって、社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を利用。

また、上記の評価テストコレクション BMIR-J1 の構築に携わった方々に感謝いたします。

参考文献

[1] オムロン アルファテック：Verity Search'97 http://www.omron-at.co.jp
[2] Justsystem：Search20
[3] コマツソフト：VextSearch
[4] 徳永：情報検索と自然言語処理、言語処理学会第 2 回チュートリアル、pp.60-77、1996.
[5] 福島ら：日本語情報検索システム評価用テストコレクション BMIR-J1、自然言語処理シンポジウム「大規模資源と自然言語処理」、1996.
[6] 多田ら：高速日本語形態素解析ソフト「SuperMorpho-J」、言語処理学会第 4 回年次会、1998.
[7] Ogawa et al.：Overlapping statistical word indexing: A new indexing method for Japanese text、SIGIR97 pp.226-234、1997
[8] Harman et al.：Overview of the Third Text REtrieval Conference (TREC3)、pp.A5-A13、1995.
[9] 木谷ら：日本語情報検索システム評価用テストコレクション BMIR-J2、情処研究報 DBS 114-3 pp.15-22、1998