日本語全文検索における文字組ベースのランキングの評価

4Y－6

赤峯享　福島俊一
NECヒューマンメディア研究所

清古勇治
NEC情報サービス

1. はじめに

自然言語の検索要求文や既存のテキストを入力とし、それに関連するテキストを大規模なテキスト集合から高速／高精度に検索するシステムのニーズが高まっている。この検索方式は、(1)検索要求文からキーワードの名詞句を抽出する、(2)抽出した名詞句を元に検索条件のターム集合を作成する、(3)検索を実行し適合したテキストをスコア順に出力する、という手順で一般に行なわれている。単語の境界が曖昧な日本語においては、検索条件に用いるタームに長単位語を用いるか、短単位語を用いるか、文字組を用いるか等のバリエーションがあり、この選択が検索精度や検索速度に影響を与える[2, 3]。

本稿では、日本語テキストの全文検索において、長単位語、短単位語、文字組をベースとして検索条件を作成し、そのそれぞれに対して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[1]を用いて検索精度を評価した結果を報告する。

2. 評価のポイント

検索要求文から抽出された名詞句をもとにして、検索条件を作成する場合、タームとして、単語を用いるか、文字組を用いるか等によって検索条件にバリエーションが生じる。例えば、「全文検索」と「ランキング」という名詞句に対して、長単位語、短単位語、１文字組、２文字組でタームを作成した場合、ターム集合は以下のようになる。まず、これらのタームの単位のうち、どれが適切かの評価を行う。

- 長単位語：(全文検索, ランキング)
- 短単位語：(全文, 検索, ランキング)
- １文字組：(全, 文, 検, 索, ラ, ン, キ, ン, グ)
- ２文字組：(全文, 文検, 検索, ラン, ンキ, キン, ング)

仮に長単位語(全文検索)をタームにそのまま用いた場合は、適合率は高くなるが、「全文を検索する」を含むようなテキストはヒットせず、再現率が低くなってしまうことがある。一方、１文字組のみをタームに用いた場合は、余分なテキストを検索してしまい、適合率が低くなってしまう可能性がある。つまり、単一の種類のタームを用いるだけでは、必ずしも高精度の検索が可能になるとは限らない。そこで、以下のようなバリエーションも持たせることにした。

(1) 同種のタームだけでターム集合を作成するのではなく、異種のタームと組み合わせたターム集合を作成することも許す。例えば、２文字組と１文字組を組み合わせたターム集合を作成する。
(2) 該当テキストを絞り込むためのターム集合とランキングを行うためのターム集合を別のものにすることも許す。例えば、短単位語のターム集合で該当テキストの絞り込みを行い、ｎ文字組のターム集合でランキングを行う。ランキングでよく用いられるベクトル空間法では、本来、全テキストを対象にしてランキングするが、インターネットの検索エンジン等では、ブーリアン式で該当テキストを絞り込んだ上で、それらをランキングする２段構成が主流になっている。

3. 実験

3．１ 評価用テキスト集合

検索精度の評価は、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[1]を用いた。BMIR-J1 は、「菓子メーカー」や「円高対策のためのメーカーの海外進出」等の 60 件の検索要求文からなり、検索対象テキスト集合（600 件の新聞記事）に対して適合する正解テキストが記述されている。

3．２ 検索手順

今回、評価を行った検索手順を以下に示す。

(1) キーワード抽出：BMIR-J1 で用意されている検索要求文からキーワードを人手で抽出する。
(2) 検索条件のターム集合の作成：抽出したキーワードを元に検索用のターム集合とランキング用の重要度付きのターム集合を作成する。
(3) 検索およびランキング：検索用のターム集合を用いて該当テキストの絞り込みを行い、ランキング用のターム集合をもとに tf*idf 法でスコア計算をする。この部分の検索は、既存の全文検

Susumu Akamine* ,Toshikazu Fukushima* and Yuji Seiko**
*NEC Corporation, **NEC Information Service,Ltd
An evaluation of n-gram based ranking method for Japanese text retrieval

索エンジン[4]を利用した。また、tf および idf は、以下の式を用いた。

tf(X, t)=(1+log(テキストX中のタームtの出現数))／log(テキストXの文字数)

idf(t) =log(総テキスト数／タームtが出現したテキスト数)

3．3 検索条件のターム集合

実験を行ったターム集合を以下に示す。

(A) 検索用：長単位語、ランキング用：長単位語
(B) 検索用：短単位語、ランキング用：短単位語
(C) 検索用：１文字組、ランキング用：１文字組
(D) 検索用：２文字組、ランキング用：２文字組
(E) 検索用：n文字組、ランキング用：n文字組
(F) 検索用：短単位語、ランキング用：n文字組

なお、長単位語は名詞が連続する部分とし、短単位語は、広辞苑のエントリーを参考にして長単位語を分割した。また、n文字組は、短単位語のタームをさらに漢字1,2文字組、カタカナ1,2,3文字組に分割して作成したものである。

ランキング用のタームの重要度は、(A)(B)(C)(D)の場合は一律に1.0を与えた。また、(E)(F)の場合は、漢字1文字組の重要度は漢字2文字組の1/2に、カタカナ2文字組の重要度はカタカナ3文字組の重要度の1/2に、カタカナ1文字の重要度はカタカナ3文字組の重要度の1/10にした。さらに(E)(F)では、n文字組に分割することで分割前の各短単位語の重要度がばらつくことを防ぐために、短単位語のレベルで重要度が等しくなるように、n文字組のタームの重要度を正規化した。例えば、「全文検索」に(F)の条件を適用すると、検索用のターム集合は（全文,検索）になり、ランキング用のターム集合は（全文,全,文,検索,検,索）で、重要度は2文字組が1.0、１文字組が0.5になる。

4．評価結果

評価は、0.1刻みの再現率に対する適合率の平均で行った。表1に評価結果を示す。(A)の長単位語を用いた検索の精度が極端に悪く、(B)の短単位語を用いた検索の精度が最も良かった。(B)と比較すると、(C)の１文字組は、明らかに精度が落ちているが、(D)の2文字組とはあまり差がなく、(E)のn文字組とはほとんど差がなかった。また、該当文書を短単位語で絞り込んだ(F)とn文字組で絞り込んだ(E)の間には、全く差が現れなかった。

表1:　評価結果

再現率	適合率					
	(A)	(B)	(C)	(D)	(E)	(F)
0.0	0.564	0.803	0.743	0.807	0.772	0.772
0.1	0.514	0.789	0.662	0.762	0.772	0.772
0.2	0.489	0.680	0.546	0.648	0.665	0.665
0.3	0.369	0.560	0.474	0.533	0.561	0.561
0.4	0.333	0.519	0.442	0.500	0.523	0.523
0.5	0.307	0.487	0.424	0.460	0.502	0.502
0.6	0.274	0.441	0.396	0.430	0.442	0.442
0.7	0.254	0.386	0.327	0.370	0.388	0.388
0.8	0.234	0.347	0.282	0.334	0.358	0.358
0.9	0.228	0.306	0.240	0.296	0.299	0.299
1.0	0.210	0.266	0.212	0.256	0.253	0.253

5．おわりに

ランキングのスコア計算として、一般的に用いられている tf*idf 法を用いて、文字組ベースと単語ベースで検索精度を比較した。漢字２文字組・カタカナ３文字組を用いれば、単語ベースと同等の検索精度になることがわかった。つまり、漢字２文字組・カタカナ３文字組以下の tf, idf 値を予め求めておけば、高精度な検索が高速に行えることになる。今回の評価は、小規模なテストコレクションで実施しため、評価結果に偏りが生じた可能性があり、さらなる吟味が必要である。今後、大規模なテストコレクションでも評価する予定である。

[1]株式会社日本経済新聞の協力によって、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。

[2]H. Fujii and W. B. Croft, A Comparison of Indexing Techniques for Japanese Text Retrieval, Proc. of SIGIR'93, 1993.

[3]小川ほか、短単位キーワードに基づくテキストデータベースシステム、情処 DBS 研資料、DBS-90-6、1992年。

[4] 福島ほか:全文検索システム RetrievalExpress の開発と評価、言語処理学会第３回年次大会 発表論文集、A4-3、1997年。