3U-07

情報検索のためのカタカナ語表記揺れ処理
- “Spruce”によるカタカナ語拡張 -

奥村 薫
Microsoft Corporation

1. 始めに

全文検索においてカタカナ語の表記揺れは、再現率を下げる主要原因の一つである。揺らぎ語辞書による技法では、新語・専門用語などに対応しにくく、一方、アルゴリズムによる正規化では、思わぬ単語が同一視されてしまうことがありうる。

本稿では、クエリ時に類似カタカナ語を自動生成して検索するカタカナ語拡張技法を提起し、その性能を評価する。これは副作用をほとんど引き起こさず、かつ揺らぎ語の大部分をカバーする。なお、このカタカナ語表記揺れ処理は、IR 用形態素解析プログラム“Spruce”の一部として、エンカルタ総合大百科 2000、Windows2000 IIS などに搭載されている。

2. 既存技法

(1) 揺らぎ語辞書

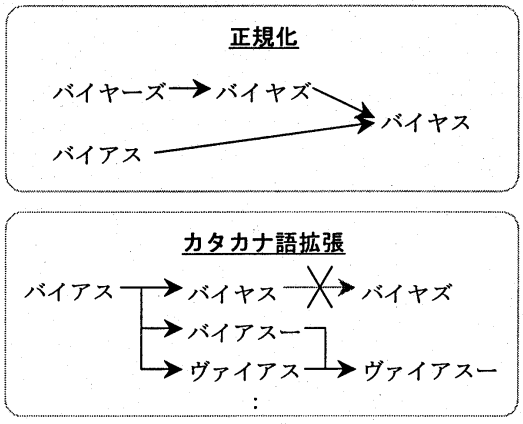
辞書に登録されたもののみを扱うので、原理上過剰な同一視は起こらない。一方、特に揺らぎやすい新語や専門用語などに対応しにくい。また、メンテナンスの手間がかかる。

(2) アルゴリズムによる正規化

インデクス時・クエリ時双方で、ルールに従ってカタカナ語部分列を変換し、代表元に帰着させるもの。メンテナンスが容易で、再現率も高いが、異なった意味の語を同一視して、適合率の低下を招くことがある。特に、ルールの複数適応・繰り返し適応で副作用がおきやすい。

3. クエリ時カタカナ語拡張の概要

正規化がカタカナ語全体の空間を商群に分割するとすれば、クエリ時拡張では語全体を距離空間として扱うのにも似ている。インデクスとしてはオリジナル語をそのまま登録する。クエリのカタカナ語



にコスト付きのルールを適応して複数の類似語を生成し、それぞれに対してクエリを行う。

実際には存在しないカタカナ語も多数生成されるが、それらはインデクスにはヒットしない。

(1) 変換テーブル

変換前	変換後	コスト
シャ	シア	3
シア	シャ	3

なお、コストはあえて非対称に設定することを許している。長音記号などを抜くのはよいが、挿入する際の変換コストを高めに設定しないと、生成語数が急速に増加する為である。

(2) 変換抑止テーブル

シャ/シアはギリシャ/ギリシア等、もっとも頻繁な揺らぎのひとつである一方、シアトル/シャトルでは副作用となる。そこで、より長い部分列に対して、この場合には変換を行わないことを指定できるようにした。これらを併用することで、きめ細かいコントロールが出来るようになった。

シャト	シアト	NoExpand
シアト	シャト	NoExpand

Katakana Word Variant Handling for IR – Katakana Expansion in “Spruce”.
Kaoru Okumura, Natural Language Group, Microsoft Corporation
One Microsoft Way, Redmond WA 98052-6399, U.S.A. e-mail: kaoruo@microsoft.com

（3）判定基準

ルール群を累積コストが閾値に達するまで再帰的に適応する。閾値はカタカナ語長の関数である。

4. 質的評価

（1）テスト・コーパス

さまざまな Web サイトに関する短い説明集から抽出したカタカナ語 19,000 語をインデクス対象及びクエリとして使用し、3 つの技法を比較する。

（2）使用ルール

ほぼ同一のルールを「正規化」と「拡張」に使用した。ただし正規化では、文字列ペアを同一視するのに対し、「拡張」ではペアにコストを与えている。よって、「拡張」で得られるヒット対は、正規化で得られるものの部分集合である。

（3）再現率の計算法

「正規化」技法で得られたカタカナ語の対に対して、人間が類義語(適)であるか、違う意味の言葉かを判定し、適を全体集合とした。よって、この数値は「正規化」に対する相対的な再現率である。

「揺らぎ辞書」技法に対しては、両方のカタカナ語がともに IME 辞書に入っていた場合にヒットすると仮定した。IME 辞書に登録されているものは、語として認知されており、揺らぎ辞書を作成する場合にも、この程度がカバーされるであろうという推論による。

（4）再現率・適合率

総カタカナ語中

	適合率	相対再現率
揺らぎ辞書	100.0%	40.4%
正規化	53.3%	100.0%
クエリ拡張	98.3%	85.3%

正規化の適合率は、特に短い単語で低くなっている。そこで、平均語長が 3 及び 4 文字以上のカタカナ語に限定した集計結果もここに提示しよう。

3 文字以上のカタカナ語中

	適合率	相対再現率
正規化	65.3%	100.0%
クエリ拡張	98.5%	86.4%

4 文字以上のカタカナ語中

	適合率	相対再現率
正規化	87.4%	100.0%
クエリ拡張	98.5%	95.6%

ユーザが求める語がヒット文献中に全く含まれていない確率が 34.7%, 12.6%となる正規化技法と比較すると、揺らぎ語の 15.6%あるいは 4.4%を見逃す方がかなり望ましいと言えよう。(大量の文献に対する検索では、一般にユーザは見逃しを気づかないことが多い)

5. 量的評価

生成カタカナ語数

前節のコーパス 19,000 語に対する平均生成語数は **6.5** 語であった。

語長と生成語数のグラフは次のとおり。

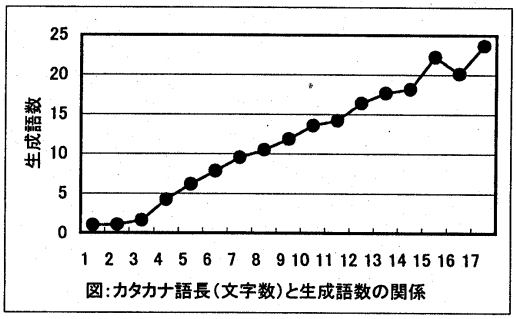


図:カタカナ語長(文字数)と生成語数の関係

速度

形態素解析プログラム Spruce のカタカナ語拡張有りと無しのモードを比較する。

測定環境 300MHz Pentium, 64MB Windows 2000

拡張無し：7619 文字/秒

拡張有り：7485 文字/秒

よって、当処理によるインデクス時間の増加は、1.8%程度であった。

6. 今後の課題

現在は、各変換コスト及び閾値をあらかじめ設定しているが、カタカナ語の頻度や揺らぎの発生頻度から、最適なコストや閾値関数を自動的にチューニングする仕組みも考えていきたい。

参考文献

〔1〕 Patrick Halstead, 奥村薫："ロバストな日本語形態素解析－辞書依存性の低いハイブリッドアルゴリズムの提案－", 情報処理学会第５４回全国大会 P.2-55,56 (1998)