Web データを利用したカタカナ異表記の自動獲得

増山 毅司
東京大学 大学院 総合文化研究科
tak@r.dl.itc.u-tokyo.ac.jp

中川 裕志
東京大学 情報基盤センター
nakagawa@dl.itc.u-tokyo.ac.jp

1 はじめに

英単語をカタカナ語に翻訳するとその表記に揺れが生じる. この揺れのことを以後カタカナ異表記と呼ぶ. 日本語では, 外来語のカタカナ異表記が多いため, テキスト情報を扱うときに問題になっている. 問題になっている分野としては, 情報検索, 情報要約, 機械翻訳, 質問応答が挙げられる.

例えば, 「spaghetti」の 6 種類のカタカナ異表記を Google で重複しないように検索した結果を表 1 に示す. 表 1 の第 1 列は, キーワードとして入力したカタカナ異表記を示し, 第 2 列は, Google 検索でヒットしたページ数を示している. 表 1 より, 6 種類の異表記全てが高頻度でヒットしていることがわかった. 次に, Google 検索の異表記チェック機能を利用して「スパゲッティ」を検索した. 異表記チェック機能は完全ではないため, 「スパゲッティ」以外にも「スパゲッティー」, 「スパゲティ」, 「スパゲティー」を含むページは得られたが, 「スパゲッテイ」や「スパゲテイ」を含むページは得られなかった. このように, 探しているページに含まれるキーワードが検索に与えるキーワードと異表記の関係にある場合, 探しているページが常に得られるとは限らないという問題がある.

このカタカナ異表記の問題は Web 検索時にのみ起こるわけではなく, 一見すると表記の統一が完全になされていると思われる新聞記事をコーパスとして利用する際にも起こることが報告されている (佐藤, 2004). 本研究では, カタカナ異表記の問題を自然言語処理の技術を使って解決する方法を提案し, その評価を行うことを目的とする.

表 1: 「spaghetti」のカタカナ異表記の Google 検索結果

異表記	ヒットしたページ数 (%)
スパゲッティ	187,000 (32.7%)
スパゲッティー	57,600 (10.1%)
スパゲッテイ	6,850 (1.2%)
スパゲティ	240,000 (41.9%)
スパゲティー	77,400 (13.5%)
スパゲテイ	3,800 (0.7%)
合計	572,650 (100%)

これまでにカタカナ異表記の生成や抽出に関する研究がいくつか報告されている. 獅子堀らは, 人手で「ヴェ」と「べ」の置換のようなカタカナ語の異表記変換ルールを作成して, その変換ルールを用いてカタカナ語の異表記を生成する方法を提案している (獅子堀他, 1993). この方法の問題としては, 人手でカタカナ語の異表記変換ルールを作成しているために, ルールの作成や更新に非常に手間とコストがかかることが挙げられる. カタカナ語には, 新語が多く, かつ, 異表記のバリエーションも多様なために, 人手で網羅的なルールの作成と更新を行うことは限界があると考える.

また, 増山らは, 表記の類似度を測る尺度である表記ペナルティを提案して, 文脈の類似度と組み合わせることでカタカナ語の異表記ペアを抽出している (Masuyama et al., 2004). 表記ペナルティとは, 編集距離を基本とした尺度で, 編集距離に「ア ↔ ァ」や「ヴ ↔ ブ」は「ア ↔ テ」よりも似ているなどというペナルティの差を導入したものである. このペナルティの調整は, 専門家が人手で行っている. 例えば, 人手で作成した表記ペナルティを使うと, 「ア」と「ァ」, 「ジ」と「ヂ」の置換, 及び, 「ー」の挿入や削除は表記ペナルティ値が 1, 「ハ」と「バ」, 「ウ」と「ヴ」の置換は表記ペナルティ値が 2, 「ア」と「ヤ」, 「ツ」と「ッ」の置換は表記ペナルティ値が 3 などという整数値を付与することができる. 増山らは, 表記ペナルティと文脈の類似度を用いることで, 再現率 97.4%, 適合率 89.1%, *F-measure* 値 93.1%の精度で異表記ペアを抽出している. この方法の問題としては, 表記ペナルティの調整が特定のコーパスに特化していること, 及び, その調整が人手で行われていることにより, カタカナ語の新語が増えた場合や異なるコーパスに適用した場合に, 表記の類似度を再調

整しなければならないことが挙げられる.

本研究では, 表記ペナルティを自動的に調整するような方法を提案する. また, その表記ペナルティと文脈の類似度を組み合わせて異表記ペアの抽出を行う. 自動で作成した表記ペナルティを使うと, 人手で作成した表記ペナルティと同様な整数値を自動的に付与することができる. 専門家が人手で作成した表記ペナルティ (Masuyama et al., 2004) との性能比較を行ったところ, ほぼ同程度の性能であることがわかった. また, この表記ペナルティと文脈の類似度を組み合わせて, 毎日新聞などの新聞記事延べ 38 年分のコーパスからカタカナ語の異表記ペアの抽出を行ったところ, 再現率 98.8%, 適合率 86.5%, *F-measure* 値 92.2%の抽出精度が得られた.

本稿の構成は以下の通りである. 次節で, 表記ペナルティを統計的に導出する方法について述べる. 第 3 節で, 実験結果を示し, 人手で作成した表記ペナルティとの性能比較を行った結果について述べる. 第 4 節でまとめと今後の課題について述べる.

2 表記ペナルティの統計的な導出

本研究では, 表記ペナルティを統計的な方法で導出するために大量の異表記ペアを必要とする. また, この異表記ペアは, カタカナ語の新語にも対応したものである必要がある. 人手でこのような異表記ペアを作ることは網羅性の点で問題があるため, 本研究では, Web データを利用して異表記ペアの抽出を行う.

2.1 Web データからの異表記候補のカタカナ語を含むページの検索

本研究では, 4 つのサイト[1] から英単語とその日本語訳ペアの集合を入手した. 英単語とその日本語訳のペア数は, 全ての字種を含めると 62,497 ペアであったが, 本研究では, その中から, (vodka, ウォッカ) のような英単語とそのカタカナ表記の関係の 14,958 ペアのみを人手により選んだ.

本研究では, (vodka, ウォッカ) のような英単語とそのカタカナ表記のペアの集合に対して, 英単語を日本語ページ指定で検索する方法と英単語に「英和」というキーワードも加えて検索する方法の 2 つにより, 異表記候補のカタカナ語を含むページの検索を行った. Google 検索エンジンには, 辞書検索というものがあり, 「英和」と調べたい英単語を入力することでその英単語の日本語訳のページを検索することができる.

2.2 Web から抽出したページからの異表記候補ペアの収集

本研究では, 英単語を Google 検索することで得られた日本語訳の候補を含むページから, 英単語のカタカナ表記との編集距離が 1 のカタカナ語のみを異表記候補ペアとして抽出した. 例えば, 「vodka ウォッカ」の場合, (ウォッカ, ウォトカ), (ウォッカ, ウオッカ), (ウォッカ, ヴォッカ) が異表記候補ペアとして抽出された. 同様に, 編集距離が 2 と 3 の場合について調べたところ, 編集距離が 2 の場合では, (アイドル, ライバル), (ミリリットル, リットル), 編集距離が 3 の場合では, (アイスクリーム, ソフトクリーム), (アップロード, ダウンロード) などといった異表記ではないカタカナ語ペアが多数抽出された. そのため, 編集距離が 1 のカタカナ語ペアのみを異表記候補ペアとして抽出することにした.

2.3 文脈の類似度を用いた異表記ペアの抽出

第 2.2 節で抽出された異表記候補ペアは, 必ずしも異表記の関係ではなかった. そこで, 本研究では, 異表記候補のカタカナ語の文脈の類似度を用いて異表記ペアの抽出を行った. 文脈としては, 異表記候補のカタカナ語の前後数単語を扱った. 以後この前後数単語のことを文章コンテキストと呼ぶ.

文章コンテキストの作成方法についてであるが, まず, ウォッカやウォトカのような異表記候補のカタカナ語を Google 検索して異表記候補のカタカナ語を含むページの抽出を行った. 次に, 検索の結果として得られたページに対して, 茶筌[2] を使って形態素解析を行い, 付与された品詞情報の中からストップワードを除いた名詞, 動詞, 形容詞, 副詞, 未知語のみを内容語として抽出した. ここで, ストップワードには, ひらがなのみの単語と 1 文字の単語を使用した. 最後に, 異表記候補のカタカナ語の前後各々50 語を取り出して文章コンテキストとした.

本研究では, 文脈の類似度としては, 式 (1) に示すコサイン類似度を用いた. ここで, 「場合」, 「知る」, 「聞く」などといった高頻度で出現する単語によってコサイン類似度が高くならないように, 単語 t の重みとしては, $\log(freq(t)+1)$ を用いた. $freq(t)$ は, 単語 t の文章コンテキスト中での出現頻度を示している.

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \quad (1)$$

[1] http://homepage2.nifty.com/katakanaEnglish/
http://wwwhoshi.cis.ibaraki.ac.jp/useful/useful15.html
http://www.ke.ics.saitama-u.ac.jp/jsgs/keywords.html
http://www.smalltown.ne.jp/~usata/pub/distfiles/skk-jisyo-extra-200307/SKK-JISYO.edict

[2] http://chasen.naist.jp/hiki/ChaSen/

本研究では, 0.00006 を閾値に設定し, それ以上のコサイン類似度であれば異表記ペアとして抽出を行った. (ウォッカ, ウォトカ), (ウォッカ, ウオッカ), (ウォッカ, ヴォッカ) のコサイン類似度は, それぞれ 0.00157, 0.00020, 0.00025 となり, 異表記ペアとして抽出された.

2.4 表記ペナルティの統計的な導出

第 2.3 節で抽出した異表記ペアの集合を使って, 表記ペナルティを統計的に導出する方法について述べる.

本研究では, 「異表記は, カタカナ語を構成する特定の文字または文字列の出現と関係して起こりやすい.」という特性を利用している. 例えば, (ウインブルドン, ウィンブルドン), (ウインドウズ, ウィンドウズ), (ウインク, ウィンク) の場合, 「イ」と「ィ」の置換は, 前後の「ウ」と「ン」の出現と関係して起こっていることがわかる. 本研究では, 以後, 挿入, 削除, 置換をオペレーションと呼び, オペレーションの前後の数文字を文字コンテキストと呼ぶ.

各ペアのオペレーションの対象になる文字を x とし, α, β, γ, δ を任意の 1 文字とすると, 本研究で扱う文字コンテキストは, 表 2 で示される. 例えば, (ウインブルドン, ウィンブルドン) の場合, 「イ」と「ィ」がオペレーションの対象になる文字で, 文字コンテキストの対象になる文字は, 「ウ」, 「ン」, 「ンブ」, 「ウとン」の 4 つである. この場合, ウの前には文字がないので, 直前の 2 文字は文字コンテキストとして扱わない. 同様にして, (ウイルス, ウィルス) の場合は, 「イ」と「ィ」がオペレーションの対象になり, 文字コンテキストの対象になる文字は, 「ウ」, 「ル」, 「ルス」, 「ウとル」の 4 つである. この 2 つの例しかない場合は, 文字コンテキストの数は 8 つになる.

本研究では, 異表記ペアが与えられたときに, 各文字コンテキスト (CLC) に対してオペレーション ($x \leftrightarrow y$) が起こる確率を式 (2) により計算する. ここで, $x \leftrightarrow y$ の x と y は異なる文字を示している.

$$\mathrm{P}(x \leftrightarrow y|CLC) \approx \frac{f(CLC, x \leftrightarrow y) + 1}{f(CLC) + 2} \tag{2}$$

また, 式 (2) の $f(CLC)$ と $f(CLC, x \leftrightarrow y)$ を次式により計算する.

$$\begin{aligned} f(CLC) &= CLC \text{ が出現するペア数} \\ f(CLC, x \leftrightarrow y) &= CLC \text{ が出現し, かつ, } x \leftrightarrow y \text{ が起こるペア数} \end{aligned}$$

本研究では, データスパースネスの問題に対応するために, 式 (2) のようにラプラス法により分子に 1, 分母に 2 を加えてスムージングを行っている.

次に, カタカナ語を構成する特定の文字または文字列の出現と関係して起こるという異表記の特性を利用し, 式 (2) により得られた確率から, $x \leftrightarrow y$ が起こる確率を最大化するような文字コンテキスト ($C\hat{L}C$) を式 (3) により求める.

$$C\hat{L}C = \underset{CLC}{\arg\max} \ \mathrm{P}_{x \leftrightarrow y}(CLC) \tag{3}$$

最後に, 式 (3) から求めた $C\hat{L}C$ を使って, $x \leftrightarrow y$ に対する表記ペナルティ ($SP_{x \leftrightarrow y}$) を式 (4) により求める. ここで, $\lfloor a \rfloor$ は, a を越えない最大の整数を示している.

$$SP_{x \leftrightarrow y} = \left\lfloor \frac{1}{\mathrm{P}_{x \leftrightarrow y}(C\hat{L}C)} \right\rfloor \tag{4}$$

式 (4) により, オペレーションが特定の文字または文字列の出現と関係して起こる場合は, 表記ペナルティの小さい整数値が得られ, そうでない場合は表記ペナルティの大きい整数値を得ることができる. 式 (4) により得られたオペレーションとそのオペレーションに対する SP 値の一部を表 3 に示す.

表 3 の第 1 列と第 3 列は, オペレーションを示し, 第 2 列と第 4 列は, オペレーションに対する SP 値を示している. ここで, 1 文字のみのオペレーションは挿入または削除を示し, $x \leftrightarrow y$ は文字 x と文字 y の置換を示している.

表 2: 本研究で扱う文字コンテキスト (x: オペレーションの対象になる文字, α, β, γ, δ: 任意の 1 文字)

文字コンテキストのパターン	文字コンテキストの説明
$\alpha\beta x$	x の直前に出現する 2 文字
βx	x の直前に出現する 1 文字
$x\gamma$	x の直後に出現する 1 文字
$x\gamma\delta$	x の直後に出現する 2 文字
$\beta x\gamma$	x の直前と直後に出現する 1 文字

表 3: オペレーションとそのオペレーションに対する SP 値の一部

オペレーション	SP 値	オペレーション	SP 値
・	1	グ ↔ ク	2
ー	1	ヴ ↔ ブ	2
ウ ↔ ー	1	ト ↔ ツ	3
オ ↔ ォ	1	ヴ ↔ ウ	3

3 人手で作成した表記ペナルティとの性能比較

本研究では, 第 2.4 節で自動で作成した *SP* と人手で作成した *SP* (Masuyama et al., 2004) との性能比較を行った.

人手で作成した *SP* とは, 各オペレーションに対する *SP* 値を人が決めているものである. 本研究では, 専門家が作成したものを用いた. 人手で作成した *SP* を使うと,「ア」と「ァ」の置換や「ー」の挿入・削除は, *SP* 値が 1,「ウ」と「ヴ」の置換は *SP* 値が 2,「ア」と「ヤ」の置換は *SP* 値が 3 などという整数値が与えられる.

本研究では, 毎日新聞などの新聞記事延べ 38 年分のコーパス (4,678,040 記事) の中から *SP* 値が 1 から 12 の異表記候補ペアを 682 個選んで評価したものを正解セットとした. なお, コーパス中のカタカナ語の異なり数は, 1,102,108 であった. ここで, *SP* 値が 10 から 12 では正解が 1 つもなかったため, 1 から 12 まででほぼ正解を網羅していると考える.

表 4 は, 各 *SP* 値において, 抽出することができた正解ペアの割合を示している. 例えば, *SP* 値が 2 の場合, 人手で作成した *SP* を用いると 207 個の異表記候補ペアを抽出できて, そのうち 162 個のペアが正解であったため, 正解ペアの割合は 78.3% になった. 同様にして, 自動で作成した *SP* の場合は, 148 個の異表記候補ペアを抽出できて, そのうち 133 個のペアが正解であったため, 正解ペアの割合は 89.9%になった. 本研究では, 自動で作成した *SP* 値と人手で作成した *SP* 値の相関係数を調べたところ, 0.76 であった. そのため, 両者には強い相関があることがわかった.

表 4: カタカナ異表記ペアの割合

SP 値	人手で作成した *SP*	自動で作成した *SP*
1	216/221 (97.7%)	262/286 (91.6%)
2	162/207 (78.3%)	133/148 (89.9%)
3	70/99 (70.7%)	51/90 (56.7%)
4	2/14 (14.3%)	2/26 (7.7%)
5	0/29 (0.0%)	0/16 (0.0%)
6	0/13 (0.0%)	2/34 (5.9%)
7	1/20 (5.0%)	1/39 (2.6%)
8	0/13 (0.0%)	1/15 (6.7%)
9	1/12 (8.3%)	0/8 (0.0%)
10	0/16 (0.0%)	0/5 (0.0%)
11	0/17 (0.0%)	0/12 (0.0%)
12	0/21 (0.0%)	0/3 (0.0%)

次に, 人手で作成した *SP* と自動で作成した *SP* の異表記ペアの抽出精度に有意差があるかを調べた. 表 5 は, *SP* が 1 から 3 で, コサイン類似度を 0.05 以上にした場合の異表記ペアの抽出精度を示している. なお, 評価は, 次に示す再現率, 適合率, *F-measure* 値を用いて行った.

$$\text{再現率 } (R) = \text{抽出した正解ペアが全正解ペアに占める割合}$$

$$\text{適合率 } (P) = \text{抽出したペア中の正解ペアの割合}$$

$$F-measure \text{ 値} = \frac{2RP}{R+P}$$

表 5 の精度について, 対応のある t 検定を行ったところ, 棄却域 5%で平均値に有意差がないことがわかった. そのため, ほぼ同程度の性能が得られていることがわかった.

表 5: 自動で作成した *SP* と人手で作成した *SP* の性能

評価方法	人手で作成した *SP* 文脈の類似度	自動で作成した *SP* 文脈の類似度
再現率	417/420 (99.3%)	415/420 (98.8%)
適合率	417/480 (86.9%)	415/480 (86.5%)
F-measure 値	92.7%	92.2%

4 まとめと今後の課題

本研究では, 表記の類似度を自動的に調整するような表記ペナルティという尺度を提案した. 人手で作成した表記ペナルティとの性能比較を行ったところ, ほぼ同程度の性能が得られていることがわかった.

今後の課題としては, 英語以外の言語の単語とそのカタカナ表記のペアの集合も加えて表記ペナルティを作成することが挙げられる. 例えば, ドイツ語の単語とそのカタカナ表記のペアとしては, (Arbeit, アルバイト) が挙げられる.

参考文献

T. Masuyama, S. Sekine, and H. Nakagawa. Automatic Construction of Japanese KATAKANA Variant List from Large Corpus. *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 2:1214–1219. 2004.

佐藤 理史. 2004. 異表記同語認定のための辞書編纂. *NL*, 161(14):97–104. 2004.

獅々堀 正幹, 青江 順一. カタカナ異表記の生成および統一手法. *NL*, 94(5):33–40. 1993.