大規模コーパスからのカタカナ語の表記の揺れリストの自動構築

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

関根 聡
ランゲージ・クラフト研究所
sekine@languagecraft.com

1 はじめに

言語には, 同じ意味を表現する場合でもその表記に揺れが現れる. その中でも外来語からきているカタカナ語には表記の揺れが非常に多い. 表記の揺れとしては, (エンパイアステートビル, エンパイヤステートビル), (アニタ・ベーカ, アニタ・ベイカー), (ウインドオス, ウィンドウズ), (ゲーテ, ギョーテ) などが挙げられる. カタカナ語は文章中に 5%から 25%含まれているという研究報告 (黒田他, 1988) もあり, カタカナ語の表記の揺れは, 情報検索, 情報抽出, 質問応答などさまざまな自然言語処理分野で問題となっている.

一般に, コーパスが大規模になるほど多くの表記の揺れが存在することが考えられる. 例えば, 今回扱った新聞記事延べ 38 年分を例にとると, 「スパゲティ」には, 「スパゲッティ」, 「スパゲッティー」, 「スパゲッテイ」, 「スパゲティー」, 「スパゲテイ」などのように 6 種類以上の表記の揺れが存在する. このような表記の揺れを人手で探すのには手間とコストがかかるため, 本稿では, 自動的に表記の揺れを抽出できるような手法を提案することを目的とする.

これまでの研究としては, カタカナ語の変換ルールを用いて異表記を生成したり, 異表記を統一する研究が多かった (獅々堀他, 1993), (久保田他, 1993). しかし, これらの研究では, カタカナ語の意味について考慮していないため, バレーとバレエのようなカタカナ語については表記の揺れかどうかを判定できないという問題があった.

本稿では, 表記の類似性を測る尺度である編集距離を基本とした手法とコンテキストの類似性を測る尺度であるベクトル空間モデルの 2 つを組み合わせて, 大規模コーパスからカタカナ語の表記の揺れを自動的に抽出する手法を提案する. 延べ 38 年分の新聞記事を対象に実験を行った結果, 再現率 87.8%, 適合率 86.0%で表記の揺れを抽出することができた. この割合から推測すると, 同新聞記事から 148,618 個の表記の揺れを抽出できる.

本稿の構成は以下の通りである. 次節で, 大規模コーパスから自動的にカタカナ語の表記の揺れを抽出する手法について述べる. 第三節で, 実験結果を示し, 本手法の有効性について述べる. 最後に第四節で本稿のまとめをする.

2 大規模コーパスからのカタカナ語の表記の揺れの抽出方法

本手法は, 次の 3 ステップからなる.

1. コーパスからのカタカナ語の切り出し
2. 表記の類似度を用いた候補ペアの抽出
3. 意味的な類似度を用いた候補ペアの判定

本節では, これら 3 つのステップについて順に述べる.

2.1 コーパスからのカタカナ語の切り出し

本稿では, カタカナ, ・, ーの連続をパターンマッチングで切り出すことによってカタカナ語の抽出を行う. 例えば, 次のような 2 つのコンテキストがあった場合に, 本稿では, **太字**のカタカナ語のみを抽出する.

- 「奇跡の経済復興の父」といわれる故**ルートウィヒ・エアハルト**氏。
- もし**ソ**連や東欧諸国が統制志向を捨て、一九四八年に西独の**ルードウィッヒ・エアハルト**がとったような経済の自由化へと突き進めば、西**ドイツ**のように奇跡の復興を遂げるかもしれない。

2.2 表記の類似度を用いた候補ペアの抽出

本稿では, 切り出したカタカナ語の集合から表記の類似性を測る尺度である表記ペナルティを用いて候補ペアの抽出を行う. 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. 尚, ペナルティ値は, 訓練データによりチューニングしている. 例えば, 「ア」と「ァ」, 「ジ」と「ヂ」の入れ替え, 及び, 「ー」の挿入削除はペ

ネルティ1, 「ハ」と「バ」, 「ウ」と「ヴ」の入れ替えはペナルティ2, 「ア」と「ヤ」, 「ツ」と「ッ」の入れ替えはペナルティ3 等となっており, 最後に単語長による調整をしている.

本稿では, 訓練データを分析したところ, 表記ペナルティがある程度以上の場合には, カタカナ語の表記の揺れがほとんど見つからなかった. そのため、閾値を設定し、表記ペナルティがその閾値以下のカタカナ語のペアを表記の揺れの候補としている.

第 2.1 節の 2 つのコンテキストから切り出したカタカナ語に対して, (ルートウィヒ・エアハルト, ソ), (ルートウィヒ・エアハルト, ルードウィッヒ・エアハルト), (ルートウィヒ・エアハルト, ドイツ), (ソ, ルードウィッヒ・エアハルト), (ソ, ドイツ), (ルードウィッヒ・エアハルト, ドイツ) といった表記の揺れの組合せを考えた場合に, (ルートウィヒ・エアハルト, ルードウィッヒ・エアハルト) のみが表記ペナルティが 3 で候補ペアとして抽出される.

2.3 意味的な類似度を用いた候補ペアの判定

本稿では, 抽出した候補ペアに対して, コンテキストの類似性を測る尺度であるベクトル空間モデルを用いて表記の揺れかどうかの判定を行う.

まず, コンテキストの抽出方法についてであるが, 本稿では, 新聞記事延べ 38 年分から判定を行おうとしているカタカナ語が属する 10 記事を無作為に選んで JUMAN[1] により形態素解析を行い, その中からストップワードを除いた名詞, 動詞, 形容詞, 副詞, 未知語のみを抽出している. 尚, ストップワードには, ひらがな, 記号, 数字などを使用した.

次に, ベクトル空間モデルについてであるが, 本稿では, コサイン類似度を用いて表記の揺れかどうかの判定を行う. 判定を行おうとしているカタカナ語のペアのうち一方のカタカナ語が属するコンテキストのベクトルを $\mathbf{a}$ とし, 他方のカタカナ語が属するコンテキストのベクトルを $\mathbf{b}$ とすると, コサイン類似度は式 (1) で示される. 尚, 本稿では, 「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類似度が高くなることを防ぐために, 単語の重みとして $log(N+1)$ を用いた. ここで, N は, 単語の出現頻度を示している.

$$sim(\mathbf{a},\mathbf{b}) = cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \quad (1)$$

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

本稿では, 第 2.2 節で抽出される候補ペアに対して, コサイン類似度が 0.05 以上のカタカナ語のペアを表記の揺れと判定する. 第 2.2 節で候補ペアとして抽出された「ルートウィヒ・エアハルト」と「ルードウィッヒ・エアハルト」のそれぞれのコンテキスト, 及び, その類似度を示すと次のようになる. この場合, 類似度が 0.05 以上であるため, 本稿では, 「ルートウィヒ・エアハルト」と「ルードウィッヒ・エアハルト」を表記の揺れと判定する.

ルートウィヒ・エアハルト: 奇跡:0.7, 経済:1.9, 父:0.7, 復興:0.7, ···

ルードウィッヒ・エアハルト: 奇跡:1.1, 経済:2.4, 自由化:1.4, 復興:1.1, ···

コサイン類似度= 0.17

3 実験

3.1 使用データ・評価方法

本稿では, 実験に新聞記事延べ 38 年分 (4,678,040 記事) を使用した. 延べ 38 年分の新聞記事に出現するカタカナ語の異なり数は 1,102,108 語であった.

本稿では, 表記ペナルティが 1 から 12 の 2,590,240 個の表記の揺れ候補から無作為に 221 個を選んで評価したものを正解セットとした. ここで, 表記ペナルティが 10 から 12 では正解が 1 つもないため, 1 から 12 までででほぼ正解を網羅していると考えている.

次に, 評価方法についてであるが, 本稿では, 再現率, 適合率, $F-measure$ 値 (F 値) を用いて実験を行った.

$$\begin{aligned} \text{再現率 } (R) &= \text{抽出した正解が全正解に占める割合} \\ \text{適合率 } (P) &= \text{抽出結果中の正解の割合} \\ F \text{ 値} &= \frac{2RP}{R+P} \end{aligned}$$

3.2 実験結果

表記ペナルティが 1 から 12 の場合で実験した結果を表 1 に示す. 表 1 の第 1 列は, 表記ペナルティ値を示し, 第 2 列と第 3 列は, それぞれ表記ペナルティのみを用いた場合と表記ペナルティと意味的な類似度の両方を用いた場合の精度を示している. 尚, 表 1 中の数字は, それぞれの表記ペナルティ値における正解数/抽出数 (適合率) を示している.

表 1 より, 表記ペナルティと意味的な類似度の両方を用いた場合は, 再現率 (87.8%) と適合率 (86.0%) でバランス良く高い精度が得られていることがわかる. また, 表記ペナルティのみを用いた場合を見ると, 表記ペナルティ値が 4 以上の表記の揺れの抽出においては, 適合率の低下を招いていることから, 表記ペナルティ値が 1 から 3 の範囲で表記の揺れを抽出することは全体的な精度 (F 値) を向上させるために有効であることがわかる.

表 1: 表記ペナルティ値を 1 から 12 とした場合の実験結果

表記ペナルティ値	表記ペナルティのみ 正解数/抽出数 (%)	表記ペナルティと意味的な類似度 正解数/抽出数 (%)
1	30/33 (90.9%)	29/29 (100%)
2	9/17 (52.9%)	9/13 (69.2%)
3	5/13 (38.5%)	5/8 (62.5%)
4	2/14 (14.3%)	-
5	0/30 (0.0%)	-
6	1/14 (7.1%)	-
7	1/20 (5.0%)	-
8	0/14 (0.0%)	-
9	1/12 (8.3%)	-
10	0/16 (0.0%)	-
11	0/17 (0.0%)	-
12	0/21 (0.0%)	-
再現率	49/49 (100%)	43/49 (87.8%)
適合率	49/221 (22.2%)	43/50 (86.0%)
F 値	36.3%	86.9%

次に, システムが表記ペナルティ値を 1 から 3 にした場合の実験結果を表 2 に示す. 表 1 と異なる点は, 表記ペナルティのみの場合も 4 から 12 の正解を取りこぼすために再現率が下がるという点である.

表 1 と表 2 を比べると, 表記ペナルティのみの F 値が 42.2%も向上していることがわかる. しかし, 表記ペナルティだけでなく意味的な類似度も用いた方がさらに 8.4%も F 値が高いことがわかる. そのため, 意味的な類似度も考慮に入れた方がより正確に表記の揺れを抽出できると考える.

本稿では, 第 1 節で述べた「スパゲティ」の 6 種類の表記の揺れがどれくらい多く抽出できるかについても調べた. その結果を表 3 に示す. 表 3 の第 1 列は, 表記ペナルティ値を示し, 第 2 列は, それぞれの表記ペナルティ値における抽出数 (再現率) を示している. 表 3 より, 6 種類の表記の揺れの全ての組合せを抽出できていることがわかる.

表 2: 表記ペナルティを 1 から 3 とした場合の実験結果

表記ペナルティ値	表記ペナルティのみ 正解数/抽出数 (%)	表記ペナルティと意味的な類似度 正解数/抽出数 (%)
1	30/33 (90.9%)	29/29 (100%)
2	9/17 (52.9%)	9/13 (69.2%)
3	5/13 (38.5%)	5/8 (62.5%)
再現率	44/49 (89.8%)	43/49 (87.8%)
適合率	44/63 (69.8%)	43/50 (86.0%)
F 値	78.5%	86.9%

表 3: 「スパゲティ」の表記の揺れの抽出結果

表記ペナルティ値	表記ペナルティと意味的な類似度 抽出数 (再現率)
1	7 (100%)
2	6 (100%)
3	2 (100%)
合計	15 (100%)

表 2 より, 表記ペナルティと意味的な類似度の両方を用いた場合に, 再現率が 87.8%で, 適合率が 86.0%の精度で表記の揺れを抽出できることがわかった. 次に, 新聞記事延べ 38 年分におけるカタカナ語の表記の揺れの抽出可能数を表 4 に示す. 表 4 の第 1 列は, 表記ペナルティ値を示し, 第 2 列は, 抽出可能な表記の揺れの数を正解数/抽出数 (適合率) で示している. 例えば, 表記ペナルティ値が 2 の場合, 表記ペナルティと意味的な類似度の両方を用いると 51,645 個の表記の揺れを抽出でき, 表 2 に示した実験結果から, 適合率は 69.2% であるので, そのうちの 35,754 個の表記の揺れが正解であると推測する. このような形でペナルティ値が 1 から 3 までの正解数を推測すると, 表 4 のように新聞記事延べ 38 年分から 148,618 個の表記の揺れを抽出できると推測する.

3.3 再現率が低下した原因

第 3.2 節の表 2 を見ると, 表記ペナルティ値が 1 のときに, 意味的な類似度を用いた表記の揺れの判定により, 1

表 4: 新聞記事延べ 38 年分におけるカタカナ語の表記の揺れの抽出可能数

表記ペナルティ値	正解数/抽出数 (適合率)
1	91,966/91,966 (100%)
2	35,754/51,645 (69.2%)
3	20,898/33,436 (62.5%)
合計	148,618/177,047 (83.9%)

個の正解を抽出できていないことがわかる．本稿では, 訓練データに対して, 正解が抽出できない原因について調べた．訓練データにおいて抽出できなかった例を次に示す.

- **アロックサンワ**, **アロック・サンワ** (福井県の建材商社)

  アロックサンワ: 富山国体に出場する福井県選手団名簿の記事に出現し, 選手名 (会社名) のように選手名とペアで使われている.

  アロック・サンワ: 会社説明の記事で会社名として使われている.

「アロックサンワ」と「アロック・サンワ」の表記の揺れは, それぞれのカタカナ語が出現するコンテキストが異なるために抽出できなかったと考える．今回は, コンテキストが異なることによる影響を避けるために, 10 記事をコンテキストとして使用したが, この例の場合は, それぞれ 1 記事のみにしか出現していなかったため, コンテキストの違いによる影響が大きかったと考える.

3.4 適合率が低下した原因

第 3.2 節の表 2 を見ると, 表記ペナルティ値が 2 と 3 のときに, 意味的な類似度を用いた表記の揺れの判定により, それぞれ 4 個と 3 個の不正解を抽出していることがわかる．訓練データにおいてシステムが間違って抽出した不正解の例を次に示す.

- **カート**, **カード**

  カート: ショッピングカート, レーシングカート, スポーツカートなどの略で使われている.

  カード: クレジットカードやキャッシュカードの略, 及び, 試合の組合せ (好カードなど) の意味で使われいる.

不正解を間違って抽出する原因としては, コンテキストが似ていることが挙げられる．カートとカードの例では, 利用, 最多, 記録, 客, 目指す, チーム, 優勝, 高い, 活躍, 出場などといった単語が共通して多く出現するために意味的な類似度が高くなったと考える.

4 まとめと今後の課題

本稿では, 表記ペナルティと意味的な類似度を用いて大規模コーパスから自動的にカタカナ語の表記の揺れリストを構築する手法を提案した．新聞記事延べ 38 年分を対象にして実験を行った結果, 再現率 87.8%, 適合率 86.0%で表記の揺れを抽出することができた.

今後の課題としては, カタカナ語以外の表記の揺れについても検討していくつもりである．また, 今回は, 新聞記事延べ 38 年分を対象に実験を行ったが, Web データにはより多くの表記の揺れが存在すると考える．例えば, 「スパゲティ」の表記の揺れを例にとると, 第 1 節で述べた表記の揺れ以外に「スッパゲティ」, 「スッパゲティー」, 「スパゲーティ」, 「スパゲーティー」, 「スパゲテー」, 「スパッゲティ」, 「スパッゲティー」などといった 7 種類以上の表記の揺れが存在する．そのため, 今後は, 大規模な Web データからも表記の揺れを抽出できるように検討していくつもりである.

参考文献

黒田 淳子, 松永 義文. 1988. 日本語文におけるカタカナ英語の研究. **自然言語処理研究会**, 68-3, 1988.

獅々堀 正幹, 青江 順一. 1993. カタカナ異表記の生成および統一手法. **自然言語処理研究会**, 94-5, 1993.

久保田 淳市, 庄田 幸恵, 河合 眞宏, 玉川 博文, 杉村 領一. 1993. カタカナ表記の統一方式. **自然言語処理研究会**, 97-16, 1993.