論文の内容の要旨

大規模コーパスからのカタカナ語の異表記リストの自動構築

増山 毅司

英語や日本語のような言語には, 同じ意味を表現する場合でもその表記に揺れが現れるという特徴がある. 特に日本語の場合は, 外来語と呼ばれる外国から借用しているカタカナ語に表記の揺れが非常に多く, 情報検索, 情報要約, 機械翻訳, 質問応答などさまざまな自然言語処理分野で問題になっている.

表 1 に外来語の例を示す. 表 1 の第 1 列は英単語を示し, 第 2 列は外来語の異表記の例を示している. このような異表記を大規模なコーパスから探してこようとすると非常に手間とコストがかかるという問題がある. 本論文では, このような問題を解決するために, 大規模なコーパスから自動的にカタカナ語の異表記リストを構築することを目的とする.

表 1: 外来語の異表記の例

英単語	外来語の異表記
Cameron Diaz	キャメロン・ディアス, キャメロン・ディアズ, キャメロンディアス
detail	ディテール, ディティール, ディテイル, ディテェール
idea	アィデア, アイデア, アイディア, アイデイア
Mozart	モーツァルト, モーツアルト
OK	オーケー, オッケー, オッケイ
spaghetti	スパゲッティ, スパゲティ, スパゲティー
vision	ビジョン, ヴィジョン

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多く報告されている. それらの研究は, 大きく 2 つに分けられる. 1 つは, 人手でカタカナ異表記の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異表記を生成・抽出する方法である. もう 1 つは, カタカナ異表記を自動的に抽出方法である.

前者の問題点としては, 人手で変換ルールを作成しているために, 変換ルールの維持・管理に非常に手間とコストがかかることが挙げられる. 外来語には, 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, 人手でルールを作成するには限界があると考える.

後者の問題点としては, 表記の類似度を測る重みの調整が人手で行われているために, 外来語の新語が増えた場合に重みを調整し直さなければならないことが挙げられる.

本論文では, この重み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する. 表記ペナルティを使うと, 例えば, 「ア」と「ァ」, 「ズ」と「ス」の置き換え, 及び, 「ー」の挿入と削除は表記ペナルティが 1, 「ィ」と「ー」, 「ヴ」と「ブ」の置き換えは表記ペナルティが 2, 「ト」と「ツ」, 「ヴ」と「ウ」の置き換えは表記ペナルティが 3 などという値を自動的に決定することができる.

本論文では, 「vodka ウォッカ」のような英単語と外来語のリストを用意し, Google と呼ばれる検索エンジンを用いて, 日本語ページ指定検索や「英和」というキーワードを加えた検索により「vodka」を含むページの抽出を行う. 次に得られたページから「ウォッカ」との編集距離が 1 のカタカナ語ペアを異表記候補ペアとして抽出する. この場合, (ウォッカ, ウォトカ), (ウォッカ, ウオッカ), (ウォッカ, ヴォッカ) が編集距離 1 で異表記候補ペアとして抽出される. そして, 異表記候補ペアの各々をもう一度 Google 検索してカタカナ語が属する文章コンテキストの抽出を行い, コサイン類似度がある閾値以上の場合は異表記ペアとして抽出する.

本論文では, 「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と共起して起こる」という特性を利用して表記ペナルティの計算を行う. まず, 得られた異表記ペアに対して, 挿入, 削除, 置換が起こった文字の前後数文字を文字コンテキスト ($context$) とし, 各文字コンテキストに対してオペレーション ($x \leftrightarrow y$) が起こる確率を式 (1) により計算する.

$$\begin{aligned} P_{x\leftrightarrow y}(context_i) &= P(x \leftrightarrow y | context_i) \\ &\approx \frac{f(context_i, x \leftrightarrow y) + 1}{f(context_i) + 2} \qquad (i = 1, 2, \cdots, 5) \end{aligned} \tag{1}$$

次に得られた確率から, オペレーション ($x \leftrightarrow y$) が起こる確率を最大化するような文字コンテキストを式 (2) により求める.

$$\hat{context} = \underset{context_i}{\operatorname{argmax}} P_{x\leftrightarrow y}(context_i) \qquad (i = 1, 2, \cdots, 5) \tag{2}$$

最後に, 式 (2) から求めた文字コンテキスト ($\hat{context}$) を使って, オペレーション ($x \leftrightarrow y$) に対する表記ペナルティ ($SP_{x\leftrightarrow y}$) を式 (3) により求める. なお, 本論文では, 式 (3) により得られた値の整数部分を表記ペナルティとしている.

$$SP_{x\leftrightarrow y} = \frac{1}{P_{x\leftrightarrow y}(\hat{context})} \tag{3}$$

式 (3) により, オペレーションが特定の文字または文字列と共起して起こる場合は, 表記ペナルティの小さい値が得られ, そうでない場合は表記ペナルティの大きい値を得ることができる.

本論文では, 長年自然言語処理の研究に従事している専門家が手動で作成した表記ペナルティと精度比較したところ, ほぼ同程度の精度で重みが調整できていることがわかった.

次に, 本論文では, この表記ペナルティという表記の類似性を測る尺度を用いて大規模コーパスから自動的にカタカナ語の異表記リストを構築する方法を提案する. 本論文で提案する異表記リストの構築方法の流れを図1に示す.

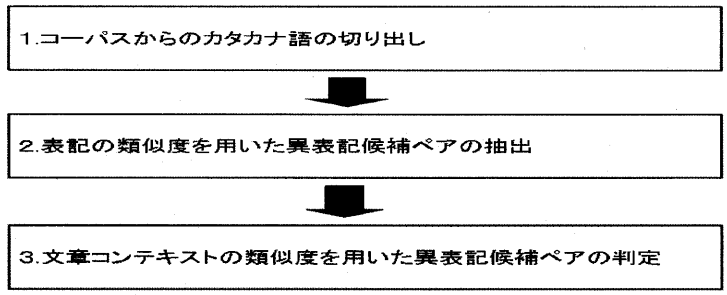

図 1: 異表記リストの構築方法の流れ

図1の1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, 本論文では, **太字** のカタカナ語のみを抽出する.

- 吉本興業は「てんねんでんねん**ミネラルウォーター**」を発売したが, 予想をはるかに上回る**ヒット**商品となった. このため「笑いは健康のもと」にちなみ「健康」を**テーマ**とした商品づくりを推進することにした.
- 快適な生活環境を実現するために欠かせないものの中に “安全できれいな水” がある. しかもより高い水質, おいしい水や健康によい影響を与える機能を持つ水なども求められている. 水は “ただ” ではなくなり**ミネラルウオータ**は飲用だけでなく料理に使う人も多くなっている.

図1の2では, 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異表記候補ペアの抽出を行う. 例えば, 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, (テーマ, ヒット), (テーマ, ミネラルウォーター), (テーマ, ミネラルウオータ), (ヒット, ミネラルウォーター), (ヒット, ミネラルウオータ), (ミネラルウォーター, ミネラルウオータ) のカタカナ語の組み合わせを考えた場合に, (ミネラルウォーター, ミネラルウオータ) のみが表記ペナルティが2となり, 閾値3よりも小さいため異表記候補ペアとして抽出される.

ここで, 切り出されたカタカナ語の数が多いほど, 比較に要する計算量の問題が無視できなくなる. そこで, 本論文では, 最初の1文字は同じカタカナ文字であるという条件や3文字未満のカタカナ語は無視するという条件などにより異表記候補ペアの絞り込みを行っている.

図1の3では, 抽出した候補ペアに対して, カタカナ語が属する文章コンテキストの類似性を測る尺度である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. 例えば, 異表記候補ペアとし

て抽出された「ミネラルウォーター」と「ミネラルウオータ」の文章コンテキスト, 及び, コサイン類似度を示すと次のようになる. この場合, コサイン類似度が 0.19 で閾値の 0.05 よりも大きいため, 本論文では, 「ミネラルウォーター」と「ミネラルウオータ」を異表記ペアであると判定する.

ミネラルウォーター: 影響:1.1, 健康:1.4, 水:1.6, 料理:0.7, $\cdots$

ミネラルウオータ: 影響:0.7, 健康:0.7, 水:3.4, 料理:1.4, $\cdots$

$$\begin{aligned} \text{コサイン類似度} &= \frac{1.1_{\text{影響}} \times 0.7_{\text{影響}} + 1.4_{\text{健康}} \times 0.7_{\text{健康}} + \cdots}{\sqrt{1.1^2_{\text{影響}} + 1.4^2_{\text{健康}} + 1.6^2_{\text{水}} + \cdots} \times \sqrt{0.7^2_{\text{影響}} + 0.7^2_{\text{健康}} + 3.4^2_{\text{水}} + \cdots}} \\ &= 0.19 \end{aligned}$$

本論文では, 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 91.5%, 適合率 91.7%, $F-measure$ 値 91.6%で異表記リストを構築することができた. また, 市販の文書作成ソフトウエアの「表記揺れチェック機能」との性能比較, 及び, 検索エンジンとの性能比較を行ったところ, 高い精度で異表記リストを構築できていることが確認できた. さらに, 異表記リストをテキスト分類問題に適用し, 特に適合率の向上に異表記リストが有効であることがわかった.