4D-4

音韻類似による書き言葉修正システム

渡邊　剛† 天沼　博† 松澤　和光†
†神奈川大学大学院　工学研究科　電気電子情報工学専攻

1. はじめに

コンピュータに言葉を処理させる際の問題点の一つとして、単語や文法が正しくない文章の処理が困難なことがある。特に最近は、SNS や Twitter などのウェブ上での私的なコミュニケーションが流行しており、そこでは通常は音声で語られるような「話し言葉」が電子テキスト化されてコンピュータ処理の対象となっている。こうした話し言葉は公式サイトの説明文やニュース記事などにおける「書き言葉」とは異なり、従来の言語処理技術が基盤としてきた文法的な制約が満たされていないばかりか、そもそも単語の形そのものがさまざまに変化し崩れたり置き換わったりしていて、言語処理上の大きな問題となっている。

例えば挨拶語「おはよう」が Web 上では「おはよー・おはよっ・おはっ・おっはー」等に変化したりする。このように様々な年代層の人間がさまざまに変形した話し言葉を日々生み出し利用しているのが、現代のネット上での言語状態と言えるだろう。これらの崩された話し言葉が含まれる文を例えば形態素解析しようとした場合、未知語と判定されたり、解析に失敗して別の単語に分解されたりしてしまう。そうかと言って、話し言葉を元の書き言葉に変換する対応データを用意したり、形態素解析ツールの利用辞書に新しい単語として組み込んだりするのは非常に煩雑であり、なによりも頻繁に更新され変化していく話し言葉の現状に追いつけない恐れがある。

そこで本研究では、変形された単語としての話し言葉を、対応する元の書き言葉に自動的に修正するシステムを提案する。この修正には、単語と単語の間の音韻的な距離を計算する手法を用いる。また、「話し言葉」化される単語は特定の単語に集中していると考え、これを話し言葉 DB として用意して修正に用いる

Written language amendment system by phoneme resemblance
†Tsuyoshi Watanabe, †Hiroshi Amanuma,
†Kazumitsu Matsuzawa,
†Graduate School of Engineering,Kanagawa University

2. 提案手法

2.1 話し言葉から書き言葉への修正法

話し言葉から書き言葉への修正では、「話し言葉」の抽出、話し言葉 DB から修正候補の選択、音韻距離計算から成る。

話し言葉 DB は、Twitter で収集したつぶやきを元に話し言葉とそれに対応するデータで構成したもので、修正対象文に対して本 DB に登録されている話し言葉で修正を行う。

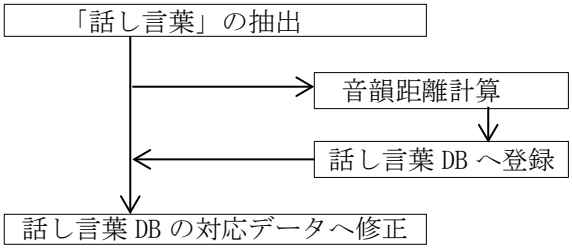

図 1. システム概要

2.1.1 修正対象の「話し言葉」の抽出方法

まず、修正対象文から話し言葉を抽出するために、次の処理を行う。

1) 形態素解析ツール Chasen により形態素に分割。
2) 話し言葉特有の誤った分割に対処するため、別途作成した品詞ルールに従い、形態素を再結合。
3) 上記で再結合した形態素と、一般の単語辞書をマッチングし、マッチすれば処理対象から除外。

上記の処理結果として残ったものを「話し言葉」として処理を進める。

2.1.2 修正候補の選択

前項で「話し言葉」として処理対象に残った単語と話し言葉 DB に登録されている話し言葉についてマッチングを行い、マッチした話し言葉を対応するデータに置き換えることで修正を行う。また、話し言葉 DB にマッチする単語がなかった場合は、次項の処理を行う。

2.1.3　音韻的距離の計算

音韻を利用した距離計算により、処理対象の「話し言葉」と、話し言葉 DB との距離計算を行う。この結果として、距離が近い DB 内の話し言葉の対応データとして新たに登録する。その後、再び修正候補の選択を行う。

2.2　音韻的距離の計算法

音韻近似計算を行うには、「話し言葉」と話し言葉 DB の単語をそれぞれモーラへと変換してから距離を算出する。

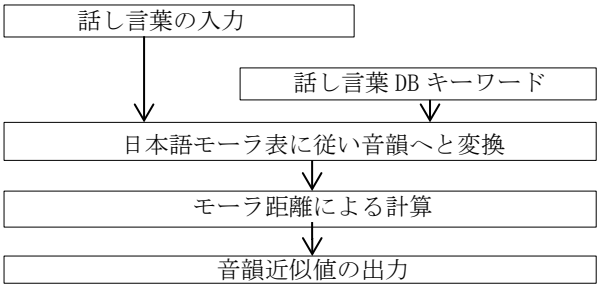

図 2.　音韻近似計算システムの概要

2.2.1　モーラへの変換

単語をモーラへと変換するため、日本語モーラ表[1]を参考に英字による発音表現への変換表を作成した。作成したモーラ変換表には、音韻距離計算を行う時に処理を容易にするため 50 音、濁音、長音、撥音、拗音とそれぞれに対応した英字表現が下記のルールに従い表記されている。

1) 発音は 1 モーラにつき 2 つの英字で表す。
2) 母音も処理の都合上「Aa,Ai,Au,Ae,Ao」と表記。
3) 「キャ」などの拗音は「ka」等、前のアルファベットを小文字にして区別。
4) 長音「ー」は「Xx」と表記。
5) 「ヨウ」など、母音オ行＋ウの場合、発音は「ヨォ」なので、「Yo,Xo」と表記、また、この場合の Xo は Xx と同等として処理をする。

表 1.　モーラ変換表(一部抜粋)

音	発音
ア、イ、ウ、エ、オ	Aa,Ai,Au,Ae,Ao
カ、サ、タ、ナ	Ka,Sa,Ta,Na
キャ、シェ、ニョ	ka,se,no
オウ、ノウ、ヨウ、	Ao,Xo,No,Xo,Yo,Xo

2.2.2　モーラ距離の考え方

モーラに変換した語はモーラ毎に完全一致以外は次のようなルールに従い距離を計算する。

1) 母音のみ一致
2) 子音のみ一致
3) 子音が近似

上記の条件の場合はそれぞれ設定した距離が加算される。

次に下記の場合は異なる音だが発音が似ている音であるといえるので、距離を減算する。

1) 「K:k」等の直音：拗音の距離
2) 「B:V」「H:F」など外来音との距離
3) その他発音が近いものに関して

4. 実験

4.1　音韻距離に関する評価実験

話し言葉 DB として試みに 100 語規模を作成し、別に収集した Twitter 頻出単語(1000 語)に対して音韻距離の計算を行った。この結果、例えば表２に示すように、音韻距離 1.0 前後をしきい値とすれば正しい書き言葉に修正できる見通しを得た。

4.2　書き言葉修正システムに関する評価実験

実際につぶやかれているつぶやき 10000 文を本システムに入力し、上記 4.1 で試作した話し言葉 DB を用いて、書き言葉への修正がどの程度行えているかを確認している。現在までのサンプリング調査の結果では、使われている話し言葉のうち 30%程度を正しい書き言葉に修正できる見通しを得ている。

表２．実験項目１の結果(一部抜粋)

対象の単語	DB キーワード	音韻類似値
オハヨウ	オハヨウ	0
オハヨー	オハヨウ	0
オハヨッ	オハヨウ	0.25
オハガキ	オハヨウ	3.75
オヤスミ	オハヨウ	46.875

5. おわりに

本研究では音韻類似を用いた話し言葉の修正について検討した。崩された話し言葉は、たとえ初めて聞く者にでも元の書き言葉の変形であることが容易にわからなければ意味がない。よって、話し言葉と音韻的に近い書き言葉を修正候補とすることが妥当と考えたからである。逆に、書き言葉から話し言葉への変形がどのようなメカニズムから行われるかを解明する生成側からのアプローチも考えられ、今後の課題である。

【参考資料】

[1]日本語モーラ表：日本語-Wikipedia-
URL:http://ja.wikipedia.org/wiki/日本語