

平成 21 年 12 月 3 日
株式会社 KDDI 研究所
独立行政法人情報通信研究機構

「くだけた表現」の自動判読技術の開発に成功 ～インターネットや携帯電話特有の伏せ字やギャル文字も解読可能に～

株式会社 KDDI 研究所（本社：埼玉県ふじみ野市、代表取締役所長：秋葉 重幸）は、独立行政法人情報通信研究機構（本部：東京都小金井市、理事長：宮原 秀夫）からの委託研究である「インターネット上の違法・有害情報検出技術の研究開発」の一部として、従来の言語解析技術では正しく解析することが困難だった、ホームページやブログ上で使われる口語やギャル文字などの「くだけた表現」を、正規な表現に自動修正する技術を開発しました。

本技術の利用により、Web 上の文書を高精度に解析し、違法・有害な情報のフィルタリング精度を向上させたり、掲示板の口コミ情報から商品の評判を高精度に分析したりすることができるようになります。

【背景】

近年、インターネットの普及によりブログや Web 掲示板などを通して、ユーザが情報を取得したり、発信したりする機会が増えています。Web 上の文書には誹謗・中傷や犯罪予告などの違法・有害な書き込みも含まれており、これらの表現を言語解析によって自動的に検出するフィルタリング技術に注目が集まっています。また、Web 上の書き込みを解析することでユーザの意見や動向を把握し、マーケティングに利用する技術なども期待されています。しかし、Web 上の文書には口語やギャル文字、伏せ字などが多数含まれている上に、日々新しい言葉も現れており、従来の言語解析技術では正しく解析することが困難でした。そのため、言語解析を利用したフィルタリングや評判解析など応用技術の精度も十分ではなく、実用レベルのものが実現できないという課題がありました。

【今回の成果】

こうした問題を解決するために、Web 上のくだけた表現を言語解析に適した、正規な表現へと自動的に修正する技術を開発しました。この技術では解析不能なくだけた表現を検出し、その修正候補となる表現を新聞文書などの正規な表現を多く含む文書から自動的に検索して取得します。取得した複数の修正候補の中から適切な表現を選ぶために、同じような文脈で頻繁に使われている表現かどうか、元のくだけた表現から大きく変化し過ぎていないかどうか、修正後の文章が日本語として自然かどうか、といった指標を計算することで、文脈に最適な表現に修正します。本技術を使えば、例えば「わナ=Uは」（「わたしは」を意味する）のようなギャル文字や「オ●マ大統領」（「オバマ大統領」を意味する）のような伏せ字もそれぞれ正しい表記に修正し、高精度に言語解析を行うことが可能となります。本技術を商用のブログ記事に適用した結果、従来の形態素解析器では解析できなかったくだけた表現を最大 38% 減少させることを確認しました。

【今後の展望】

今後は本技術を違法・有害情報フィルタリング技術などに応用していく予定です。

以上

概要: ブログなどに見られるくだけた表現を言語解析に適した表現に自動変換する

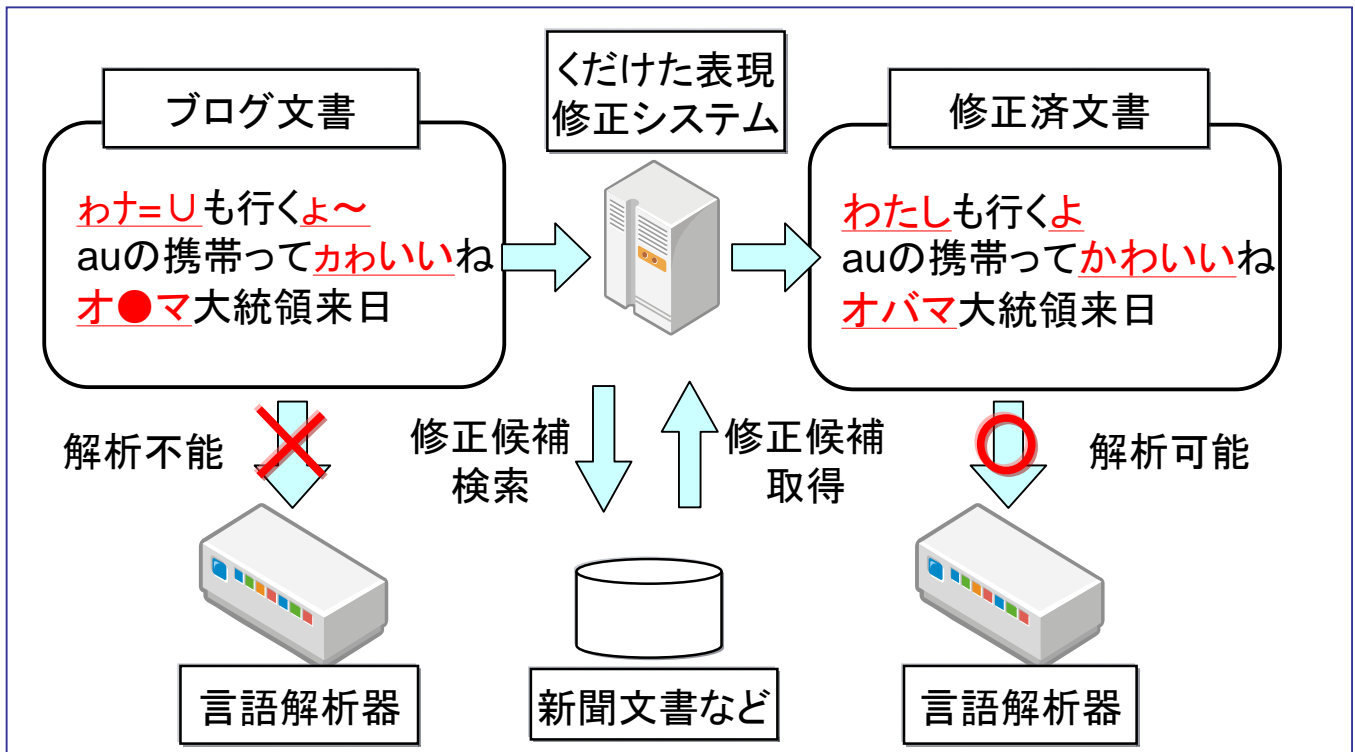
特長: 修正候補を新聞文書や他のWeb文書から自動的に検索

⇒ **ブログテキストを自動的に正規化し、高精度に解析することが可能**

性能: ブログ上の解析不能な語を平均 **30%** 減少させることに成功。

カテゴリによっては **38%** 減少。(恋愛・結婚など若い女性が興味のあるカテゴリ)

応用先: 情報フィルタリング、評判解析など



くだけた表現修正技術の詳細

要素技術1: 修正候補の自動取得

ブログ文: できるかどうか**かわ**分かりません

検索文生成: どう***分**かり

検索結果 : どう**か**は**分**かり

(=修正候補) : どう**か**分**か**り

: どう**し****た****ら****い****い****か**分**か**り

※編集距離とは、二つの文字列がどの程度異なっているかを表す指標。一方の文字列を他方の文字列に変換するために必要な挿入、削除、置換の最小回数。例:「フォーラム」から「ファーム」への編集は「オ」を「ア」に置換し、「ラ」を削除する方法が編集回数最小となるため、編集距離は2である。

要素技術2: 修正候補の選択方法

多数の修正候補から3つの指標を用いて最適なものを選択

1. 検索結果における出現頻度

2. 置換文字列間の編集距離※

3. 統計的言語モデルを用いて修正後の表現の自然さを推定