Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

Web フィルタリング処理時における表記ゆれの動的解決

井手　厚[†1]　　東　藍[†1]　　松本　裕治[†1]

Web 上の文章には，意図的に誤変換された漢字などを含む，多くの表記ゆれ表現が存在しており，Web フィルタリングを行う上では表記ゆれに対応したシステムを構築することが有効である．本論文では表記ゆれ表現を同定する方法として，KAKASI による漢字かな変換機能と MeCab の分かち書き機能という２つの技術を利用した手法について提案する．意図的に誤変換された表記を抽出するために，KAKASI を用いた読み候補の作成を行い，その読み候補が妥当かどうかについての判断を MeCab を利用して行う．本手法の効果を実験によって確認した．

Dynamic Processing of Spelling Variations in Web Filtering.

Atsushi Ide[†1]　Ai Azuma[†1]　　Yuji Matsumoto[†1]

Web documents tend to include a number of spelling variations. Especially, in Japanese pages, some variations are intentionally used to hide improper words or expressions. This paper proposes to cope with this problem in two steps: expansion of possible pronunciation by KAKASI and morphological analysis by MeCab. Alter an exhaustive expansion of pronunciation of Kanji characters by KAKASI, and matching with the dictionary of improper expressions, Japanese morphological analyzer MeCab analyses the original sentence assuming the matched expressions existed in its system dictionary. We verify the effectiveness of our idea through experiments using sentences extracted from a real BBS.

[†1] 奈良先端科学技術大学院大学

■　1 はじめに

教育現場や子どものいる家庭においてインターネット環境を構築する際，子どもが不適切なコンテンツへアクセスしないための対策として，フィルタリング技術が開発されてきた．フィルタリング技術には大きく分けて，あらかじめ第三者機関[1]によって用意された URL のリストと照合して閲覧可能か否かを判定するようなレーティング方式，ブラック／ホワイトリスト方式と，有害となりうるキーワードをあらかじめ辞書として用意しておき，閲覧先のコンテンツをその辞書を用いて検査し，閲覧可否を動的に判断するというキーワード方式の２つが存在する．キーワード方式は，ベイズ分類器や SVM といった機械学習の手法と組み合わせることでかなり柔軟なフィルタリング性能を発揮することができ，スパムメールフィルタリングなどでも頻繁に利用されている．たとえば Sahami ら[1]はベイズ的アプローチで，Drucker ら[2]は SVM を用いてそれぞれキーワード式スパムフィルタリングを行い，その有効性を示している．市販の Web コンテンツフィルタの多くもこの方式を取り入れ，高い精度でのフィルタリングが達成されている．

しかしキーワード方式における欠点の１つに，表記ゆれという問題が存在する．たとえば誹謗中傷の意味合いをもつ「うざい」という単語は，このようにひらがな表記で表現されることもあれば，「ウザい」のようにカタカナまじりで表現されることもある．特に最近，ネットいじめなどの問題が大きくクローズアップされるようになり，ネット上の掲示板の書き込みの監視が強化されるようになると，「死ね」という単語を含む誹謗中傷を掲示板に書き込もうとし，フィルタリングツールあるいはサーバ側の監視によって書き込みを拒絶されたユーザが，今度は「氏ね」や「師ね」といったように意図的に誤変換を行って監視やフィルタリングをすり抜けようとする傾向が目立つようになってきた[2]．それによって，Web 上には多くの表記ゆれ表現が氾濫するようになってきている．先ほどの「うざい」「ウザい」という例のように，単にひらがなとカタカナの２種類の表記ゆれのみであれば比較的容易に対応することが可能であるが，「死ね」が「氏ね」「師ね」となることや，「うざい」が「宇座い」と表記されるような，漢字を含むような変換になった場合，あらかじめすべての変換を予測し，キーワード辞書に登録しておくことは困難である．こういった意図的な誤変換を含む表記ゆれ表現は，特に誹謗中傷などの領域で顕著にみられ，ネットいじめなどの被害を抑えるためには，これらの表現を同定し，フィルタリングの精度をさらに高めることが必要である．

[1] たとえばモバイルコンテンツ審査・運用監視機構（http://www.ema.or.jp/ema.html）など
[2] 本稿では，このような意図的な誤変換についても，広義の表記ゆれと定義することとする

Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

表記ゆれに関する問題は情報検索（Information retrieval）の領域などで取り上げられてきている．Jones ら[3]は検索に用られるクエリが，同じものを指しているにもかかわらず漢字，ひらがな，カタカナの３つの表記が用いられうることを指摘し，この違いを吸収するために，漢字かな変換プログラムである KAKASI[3]を利用して漢字をひらがなへと変換，そしてそれをさらにローマ字に変換し，プールしてあるクエリとの類似度を求めて同定を行うという方法を提案している．この方法は，クエリがあらかじめ単語として切り出されている場合には有効であるが，文章の中から表記ゆれの表現を見つけるというフィルタリングの課題にはうまく適用できない．Jones らの提案手法に従えば，たとえば「馬鹿な人」という句はまず形態素解析器によって「馬鹿な／人」と分割され，それぞれ「ばかな」「ひと」とひらがなに変換される．この適切な分割は，形態素解析器に「馬鹿」「人」という単語が含まれていることによって実現されている．一方，本稿で問題とされるような「場化な人」のように意図的に誤変換されたような表記ゆれ単語が含まれているときには，これらが辞書に含まれていないために，多くの場合において形態素解析器で適切に切り出すことができない．比較的正確に表記されることを前提としている検索クエリの場合とは異なり，あえて正確な表記を避けていることの多い Web 上の文章をフィルタリングするためには，誤表記が存在するという前提で分析を行う必要がある．

このような背景を踏まえ，本稿では漢字かな変換プログラムである KAKASI と，形態素解析エンジン MeCab[4]を利用し，この表記ゆれ問題を解決する方法について提案を行う．表記ゆれを含んだ文においては，すべての漢字が通常の読みをするとは限らないため，KAKASI を用いてあらゆる読みの候補を生成する．そして，各単語がひらがなで登録されているキーワード辞書と，生成された読みの候補を比較し，マッチしたキーワードを抽出する．このようにすれば多くの表記ゆれ表現を捕捉することが可能となる．しかし，漢字表現をすべてひらがな表現に変換することで，抽出された部分が全く無関係なものである可能性もある．そこで，マッチした部分を単語として仮定し，文として成立しうるかどうかの判断を形態素解析エンジン MeCab によって再解析を行うことで，誤った同定を回避させる．このように，すべての情報をひらがなとして処理することでマッチさせる幅を広げ，さらに文としての整合性を確認し，整合していると思われるもののみを抽出することで，表記ゆれの単語を含むより多くの単語を，キーワード辞書への登録作業を増やすことなしに，精度良く同定することが可能となる．実験によって，本手法によってどの程度の精度でキーワード辞書の単語が抽出されるかを検証した．

[3] http://kakasi.namazu.org/index.html.ja
[4] http://mecab.sourceforge.net/

■ 2 提案手法

■ 2.1 概要

まず，提案手法の処理の流れについて簡単に述べる．処理の流れは，

1. KAKASI のかな変換機能を用い，入力文に関するあらゆる読み候補を作成する
2. 読み候補中に，キーワード辞書とマッチする単語があれば抽出する
3. 抽出された単語を MeCab の辞書に一時的に登録し，オリジナルの入力文を分かち書きする

となる．以下，それぞれの処理に関する詳細を述べる．

■ 2.2 読み候補の作成（処理 1）

入力文の中に表記ゆれが含まれているという前提に立ち，KAKASI の漢字かな変換機能を用いた文の読み候補を生成する．たとえば入力文が「A くん肝血悪いです（A くんきもちわるいです）」であったときの読み候補を図１に示す．このように，すべての漢字の読みの組み合わせを読み候補とする．なお，この例においてはすべての漢字の単独の読みしか展開していないが，中には漢字の組み合わせ，すなわち単語になってはじめて発生する読みもある．したがって読み候補の中には単語としての読みも含んでいる．

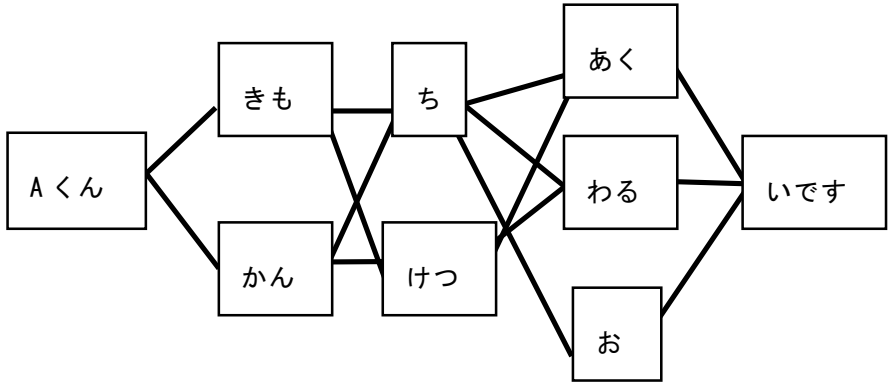

図１　「A くん肝血悪いです」の読み候補

Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

■ 2.3 辞書とマッチする単語の抽出（処理 2）

生成された読み候補集合の中から，キーワード辞書に含まれている単語をすべて抽出する．キーワード辞書の単語はすべてひらがなで登録しておく．図 1 の場合，辞書に「きもちわるい」という語の列が登録されていたとすると，図 1 の読みの中には「きもちわるい」と読める読みが存在するために，「肝血悪い」という語が抽出される．また，辞書に「でぶ」という単語が登録されていたとする．そして入力文が「街で武器屋を探す」となっていたとすると，読み候補集合の中には「まち『でぶ』　きやをさがす」のように「でぶ」を含むものが現れる．通常の読みからすれば，ここから「でぶ」という単語を抽出するのは不適切である可能性が高いが，ここの処理ではこのようなケースでも構わずに抽出する．したがってこの例では「で武」を抽出することになる．

■ 2.4 抽出単語の MeCab 辞書への登録と分かち書き処理（処理 3）

抽出された単語を，MeCab のユーザ辞書へ一時的に登録する．その後，オリジナルの入力文を MeCab で分かち書きする．たとえば「A くん肝血悪いです」という入力文があり，キーワード辞書に「きもちわるい」という単語が登録されていたとする．すると処理 1,2 の過程で「肝血悪い」という単語が抽出されるので，この単語を MeCab のユーザ辞書へ登録し，分かち書き処理を行う．すなわち「肝血悪い」という単語が存在すると仮定した状態で分かち書きを行うのである．処理の結果，最適解として，登録した単語が切り出されたものが出力されてきたら，その単語はキーワード辞書に登録していた単語であると同定する．逆に，先の例「街で武器屋を探す」の場合は「で武」という単語が抽出され MeCab 辞書へ登録されるが，この場合は MeCab が「街／で　／武器／屋／を／探す」と「街／で武／器／屋／を／探す」という分かち書き候補を比較し，よりもっともらしい前者を最適解として出力することになる．したがって，この例においては入力文中に「でぶ」という単語は含まれないと判断する．

■ 3 実験

■ 3.1 キーワード辞書について

評価実験で利用したキーワード辞書について述べる．辞書には，誹謗中傷語のうち，インターネット上の掲示板等で頻繁に目にし，かつ表記ゆれや誤変換が生じやすい 35 語を選択した．キーワード辞書に用いた語の一部を表 1 に示す．登録語はひらがなで表記され，活用語に関しては，活用形もすべて検索対象とするために活用形情報を付与した．

きもい 形容詞・アウオ段 形容詞
きしょい 形容詞・アウオ段 形容詞
いたい 形容詞・アウオ段 形容詞
きもちわるい 形容詞・アウオ段
むかつく 五段・カ行イ音便
くさい 形容詞・アウオ段
しぬ 五段・ナ行
うざい 形容詞・アウオ段
ぶす 名詞 一般
ぶさいく 名詞 一般
でぶ 名詞 一般
ばか 名詞 一般

表 1　辞書に用いた語の一部

■ 3.2 入力文について

次に，評価実験で利用した文章について述べる．実験に利用した文章は，ネット上のいくつかの掲示板上に実際に用いられていた書き込みで，辞書に登録した語を含んでいる 420 文である．利用した文の一部を表 2，表 3 に示す．これらの文のうち，辞書に登録されている単語であると同定されるべき部分は 373 カ所読みの可能性としては含まれてはいるがその単語とは関係のない部分（「その場から逃げ出した→その『場か』ら逃げ出した」など）は 120 カ所あった．

まじあいつら師ねばいいのに　　【師ね→死ね】
馬路肝意よなぁ　【肝意→きもい】
まじ肝いから WWW ちょーうざい.死ね.　【肝い→きもい】【うざい】【死ね】
コイツマジ機知害　【機知害→きちがい】
部細工なおんながでかい顔して歩いてやがる。　　【部細工→ぶさいく】
○○○（平仮名名字）たつきばか 【ばか】

表 2　辞書にある語と同定されるべき部分を含んだ文の例

Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

今日バイト先でぶっ飛ばされた・・・・　（でぶ）
市場の失敗に伴なう財の提供　　（う財）
お着物の柄の出方や一部細工の仕様変更がある場合がございます。　（部細工）
ちょうど日本のなまはげみたいなのかな。　　（はげ）
激しくその姿を変化させているのだと思います。　（くそ）

表３　辞書にある語を含んでいるが，同定されるべきではない部分を含む文の例

■　3.3 KAKASI および MeCab について

KAKASI のバージョンは 2.3.4，MeCab のバージョンは 0.98 である．MeCab で利用する辞書は IPA 辞書である．どちらの辞書情報についても，手を入れることなく配布時の状態のまま利用した．

■　3.4 結果

第２節で述べた手続きを自動で行うスクリプトを作成し，性能を測定した．この手続きを経ると，たとえば「馬路肝意よなぁ（まじきもいよなぁ）」という文は，「きもい」という読みが存在するので「肝意」という単語が MeCab のユーザ辞書へ登録され，

馬路　　名詞,固有名詞,人名,姓,*,*,馬路,ウマジ,ウマジ
肝意　　形容詞,自立,*,*,形容詞・アウオ段,基本形,きもい,,
よ　　助詞,終助詞,*,*,*,*,よ,ヨ,ヨ
なぁ　　助詞,終助詞,*,*,*,*,なぁ,ナァ,ナー

のように分解される．この例の場合，未知の語である「肝意」が MeCab によってうまく切り出されている（すなわち，MeCab の分かち書き機能の出力が「馬路／肝意／よ／なぁ」となる）ため，「きもい」のことであると推測することが可能である．本手法の精度を検討するため，入力文すべての結果について，Precision 値，Recall 値を算出した．Precision 値は 0.93,Recall 値は 0.97 であった[5]．また，True Positive(TP;辞書に登録されている単語であると同定されるべきものがきちんと同定される)，False Negative（FN;同定されるべきものが同定されない），False Positive（FP;同定されるべきでないものが同定される），True Negative（TN;同定されるべきでないものが同定されない）の４つの値を算出し，その結果を表４に示す．True Positive Rate は 0.97，True Negative Rate は 0.775 であった．

[5] Recall に関しては，全データではなく，選択された文だけを対象としているため，参考値である

	同定されるべき	*同定されるべきではない*
登録語として同定	362	27
同定せず	11	93

表４　実験結果

■　4 考察およびエラー分析

実験の結果より，True Positive Rate が 0.97 と高いことから，漢字をひらがなに変換し，読み候補を生成することでほとんどのキーワード語を同定することが示唆された．一方 True Negative Rate は 0.775 であり，必ずしも低い値ではないのだが，同定すべきではない部分から２割強の割合で同定してしまっている．以下，改善の可能性を探るために，False Negative および False Positive となった例について，その特徴を検討する．

■　4.1 False Negative の特徴

FN となった文章の例を表５に示す．FN となった部分で最も多かったのは，前後のカタカナあるいはひらがなと一体化して名詞であると解釈されたものである．たとえば，
「バカカタヤマのせいで全部オジャンさ」
という文の場合，キーワード辞書にある「ばか」が含まれているため，同定されるべきなのだが，MeCab の解析では

バカカタヤマ　　名詞,一般,*,*,*,*,*
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
せい　　名詞,非自立,一般,*,*,*,せい,セイ,セイ
で　　助詞,格助詞,一般,*,*,*,で,デ,デ
全部　　名詞,副詞可能,*,*,*,*,全部,ゼンブ,ゼンブ
オジャン　　名詞,一般,*,*,*,*,*
さ　　助詞,終助詞,*,*,*,*,さ,サ,サ

Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

という結果になる．同様に「○○○[6]たつきばか」というひらがなのみで構成された文は，

○○○たつきばか　　名詞,一般,*,*,*,*,*

となり，辞書に登録されている「ばか」という語が検出されない．MeCab に限らず，一般的な形態素解析器においては，辞書にない語は字種でまとめて未知語と推定するという処理を行っているが，多くの文章においてはカタカナ名詞＋カタカナ名詞という連接あるいはひらがな名詞＋ひらがな名詞という連接はあまり出現しないために，1 つの名詞として処理されてしまっているものと思われる．また，個人名＋「氏ね」という場合も同定できなかった．これは「氏」が人名の接尾辞と解釈されてしまうことによって生じている．この問題は本提案の枠組みにおいては解決が難しいため，別のアプローチを考える必要があるだろう．

さらに，KAKASI の辞書に登録されていな漢字を使用していた場合も同定できなかった．これについては，KAKASI で利用している辞書にさらに読み情報を追加することで対処することが可能である．実際，未登録語で FN となったものについては，KAKASI の辞書に登録を登録を行うことでキーワード語が同定できたことを確認した．

バカカタヤマのせいで全部オジャンさ　【バカ→ばか】 猿並の屎女　【屎→くそ（KAKASI辞書未登録語）】 ○○○○（漢字個人名）氏ね。　【氏ね→しね】 ○○○[6]たつきばか　【ばか】 生きる価値ないキモブスババア　【ブス】）

表５ FN となった文の例

■　4.2 False Positive の特徴

続いて，FP となった文章の例を表６に示す．「毎日のおそうざいは、身近に手に入る材料で、誰からも好まれる料理を。」という文は「うざい」という文字列を含んではいるが「おそうざい」の一部なので同定されるべきではないのであるが，同定されてしまう．MeCab の出力は以下のようになる．

毎日　名詞,副詞可能,*,*,*,*,毎日,マイニチ,マイニチ
の　助詞,連体化,*,*,*,*,の,ノ,ノ
おそ　形容詞,自立,*,*,形容詞・アウオ段,ガル接続,おそい,オソ,オソ
うざい　形容詞,自立,*,*,形容詞・アウオ段,基本形,うざい,,
は　助詞,係助詞,*,*,*,*,は,ハ,ワ
、　記号,読点,*,*,*,*,、,,、,,
身近　名詞,形容動詞語幹,*,*,*,*,身近,ミヂカ,ミジカ
に　助詞,副詞化,*,*,*,*,に,ニ,ニ
手　名詞,一般,*,*,*,*,手,テ,テ
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
入る　動詞,自立,*,*,五段・ラ行,基本形,入る,ハイル,ハイル
材料　名詞,一般,*,*,*,*,材料,ザイリョウ,ザイリョー
で　助詞,格助詞,一般,*,*,*,で,デ,デ
（以下省略）

これは，MeCab で利用している IPA 辞書に「そうざい」という単語が入っていないこと，そして「うざい」という単語を切り出しても前後で形態素の分割がうまくいってしまう（本例では「おそ」が「おそい」の活用として切り出せる）ことによって生じたものと推測される．今回の評価実験で FP となった部分の多くはこのケースに当てはまる．なお，表６の４つの例についてそれぞれ「そうざい」「伴なう」「あん肝」「ねたきり」という単語を MeCab の辞書に追加し，再び処理を行ったところ，キーワード辞書の単語が検出されなくなったことからも，MeCab の辞書のさらなる拡充が FP の割合を下げることにつながることが示唆される．

毎日のおそうざいは、身近に手に入る材料で、誰からも好まれる料理を。　（うざい） 市場の失敗に伴なう財の提供　（う財） あん肝いただきます　（肝い） 所沢市ねたきり老人等介護者手当支給要綱　（市ね）

表６ FP となった文の例

[6] ○○○の部分にはひらがなの名字が入っている

Vol.2010-NL-196 No.7
Vol.2010-SLP-81 No.7
2010/5/27

■ 5 まとめと今後の課題

本稿では，KAKASI を用いて漢字を読みに変換し，読み候補を生成，その読み候補からキーワード辞書にマッチするものを抽出し，再度 MeCab を通して文の整合性を検証することで，表記ゆれが生じやすいような単語を含む文から，キーワード辞書の登録語を精度よく同定することが可能であることを示した．

より精度を高めるためには，False Negative および False Positive の割合を下げる必要があるが，それには MeCab，あるいは KAKASI の辞書を拡充することで可能であることが示唆された．特に MeCab の辞書の拡充は多少手間のかかる作業となるが，FP となるようなパターンは限られているため，表記ゆれのようにあらゆる漢字の組み合わせなどを考慮してキーワード辞書を拡充させるよりもはるかに手間は少ないものと思われる．また，FP となってしまったとしても，その文章において，該当単語以外のものが適切に抽出されてさえいれば，その後の確率的フィルタリングの処理などによって誤りはほとんど目立たないものとなるだろう．これは今後の検討課題である．

本稿で扱った表記ゆれ事例は，「師ね（しね）」「場化（ばか）」「ウザい」のように，ひらがなに変換すれば完全に一致する単語のみを扱った．一方，実際の掲示板やチャットなどの書き込みには，「ウザぃ」「場あ化」のように小書き文を利用した表現を用いたり，文字間に音を伸ばすためのかなを挿入したりするものも多く見られる．特に最近の子どもの間では小書き文字を多用した表現を用いることが流行となっており，こういった表記ゆれの問題に対しても対処をする必要がある．本稿で提案した手法をそのまま適用しただけではこの問題を解決することはできないが，ひらがな（あるいはローマ字）の読み候補を生成した際，たとえば Navarro[4]が紹介するような近似文字列照合の手法が比較的容易に適用でき，近似しているものを含めた単語を抽出することができる．この手法を取り入れ，さらにどの程度広範囲に表記ゆれ表現を抽出できるかについても今後検討していきたい．

さらに，既存の未知語処理手法と本手法を組み合わせることでさらなる精度上昇を期待することができる．たとえば東ら[5]は形態素解析と同時に未知語候補を追加することにより，未知語処理を行う手法を提案している．このような手法で未知語と思われる部分を絞った上で，本稿の提案手法を適用することにより，精度がさらに高まることが期待される．

また，本提案の実用性を高めるためには，処理の時間も考慮に入れなくてはならない．特に KAKASI による読み候補の作成については，漢字の組み合わせによっては組み合わせ爆発をおこす．したがって，実装時には，読みを展開していく過程で適宜不必要な読みを削除していくようにしたが，それでも長い文の入力に対しては最大で 10 秒ほど処理に時間がかかってしまうこともあった．処理時間の削減にも今後取り組んでいかなくてはならない．

参考文献

1) Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization - Papers from the AAAI Workshop, Technical Report WS-98-05,(1998)55-62

2) Drucker, H.D., Wu, D., Vapnik, V.: Support Vector Machines for spam categorization. IEEE Transactions on Neural Networks, Vol. 10(5).(1999)1048-1054

3) Jones, R., Bartz, K., Subasic, P., Rey, B.: Automatically generating related queries in Japanese. Lang Resources & Evaluation, (2006)40:219-232

4) G. Navarro. A guided tour to approximate string matching. ACM. Computing Surveys, Vol. 33, No. 1, March 2001, pp. 31-88.

5) 東藍, 浅原正幸, 松本裕治. 条件付確率場による日本語未知語処理. 情報処理学会研究報告 2006-NL-173, pp. 67-74, 2006.