伏字を含むテキストの分ち書き処理

笠原 要　永田 昌明

NTT コミュニケーション科学基礎研究所
{kasahara.kaname,　nagata.masaaki}@lab.ntt.co.jp

1. はじめに

本稿では、一部の文字が記号等で置き換えられた伏字を含むテキストについて教師データを作成し、点推定で形態素に分ち書きし、「○」等の伏字に用いられる文字が伏字として使われているかを判定する方式を提案する。

消費者生成メディア(Consumer Generated Media, CGM)の記事には、犯罪に関わる違法情報や青少年に有害な内容等が含まれている場合があり社会問題となっている。CGMを提供する事業者は、記事の違法有害性を適宜判定して削除・通報等することが求められているが、多数の記事を目視確認する場合、人的稼動を要する問題がある。

そこで我々は、作業を効率化する支援方式を検討した[1]。有害/無害を判定し有害な表現をアノテートした記事のテキストデータを教師データとし、新たな記事に対して有害な記事の可能性があるか判定し、含まれる有害表現を抽出するものである。実験を行い、スパムフィルタ[2]を用いた場合、無害な記事を 99.5%の精度で58%取り除くことができることがわかった。しかし、有害な可能性がある形態素や文字を有害/無害ラベル付けする実験では、様々な特徴やラベル付け方法(CRF/SVM)を用いたが精度、再現率共に2割を超えなかった。この原因として属性(形態素，文字及び付随する情報)のスパースネスが考えられる。下記は掲示板サービス2ch(URL http://www.2ch.net)のテキストについて、個人情報の観点から一部置き換えた上で、MeCab で分かち書き処理したものである(以降の事例も同様)。「|」は形態素の境界を示す

アホ	だ	な	ｗｗｗｗｗｗｗｗ	もう	○	ね
電	○	高校	2	年	の	K	原	かなめ
以前	から	笠	○	に	執拗	な	電話

最初の「○ね」は「死ね」と推察される。同じような文字や形態素列の出現が複数あれば、これらは有害表現として抽出される可能性がある。一方「電○」や「笠○」のような、1つの単語を「○」で伏せた固有名詞についても、勿論形態素辞書にエントリーがないために、「○」の前後に分ち書き境界を設定してしまっている。固有名詞は「死ね」のような一般語に比べると異なりの出現数が多い一方、伏せる文字や伏せる場所は自由に設定できるので、このままでは出現頻度が少ない系列情報としなるため、期待された処理が行えない。

そこで本稿では、CGMコンテンツ中の伏字を含む形態素を正しく分かち書きする方法について検討を行う。

2. 伏字を含むテキストの分かち書き処理

2.1 伏字の特徴

本稿で扱う伏字とは、単語や形態素を構成する文字列中の一部を別の文字で置き換える操作及び、その文字とする。分ち書き処理を行う際に留意すべき伏字の特徴を例と共(2 ちゃんねるのテキストを一部変更して利用)に挙げる。

[伏字の数/場所]

形態素中の1文字で伏せる場合が比較的多いが、2文字以上を伏せる場合もある。また、離散的に伏せることも行われる。

(例) |謀|会社|(|○○○○－○業|)|
(例) |NTT|コ○ュニ○ーション|科学|基礎|研究|所|

そのため、少数の伏字に限定した分ち書き方式では、CGMのテキストには対応しづらい。

[伏字に用いられる文字の多義性]

伏字には記号文字(○、●、△、□、＝等)が多く使われるが、これらは本来、他の意味合いで使われている。、例えば「○」は、その形状から“丸”や“丸い”意味を表す記号や、アスキーアート(文字を要素として描かれた絵画)、項目を列挙する際の先頭文字等様々に使われる。多くの場合は1文字で1つの形態素を構成するが、先の「○ね」(死ね)のように、1文字でも伏字となる場合がある。そのため、分ち書きの処理を行う際には、伏字として用いられる文字が本来の用法で使われているのか、伏字として使われているのかを判定することが必要となる。

[伏字の復元]

一般名詞や用言の場合には、伏せた元の文字(「元字」と呼ぶ)をユニークに人間が推定できることが多いが、固有名詞や固有表現に関わる形態素の一部が伏字となっている場合には推定困難な場合がある。

(例)　|笠○|さん|は|見て|見ぬ|振り|だろ。|

上記例の「笠○」は人名である可能性が高いが、「笠原」だけではなく「笠井」や「笠置」のような2字の苗字のいずれも可能性がある。勿論、このテキストを含む記事

全体や関連する記事を読むことで、曖昧性を解消できる場合もあるが、違法有害の判定作業では、個別の記事を独立にデータとして受け取り作業する場合が多いので、文脈情報が使えないことを前提とする必要がある。さらに、下記のような抽象的な表現では、推定される元字が無数考えられる。

（例）|3|F|社員|の|○○|。|

（例）|○○|県|○○|市|○○|町|○|−|○○|

[複合操作]

情報の隠匿を主なる目的として、1つの形態素に伏字以外の文字操作も同時に行われるている場合がある

（例）|y○mada|長|と|t○naka|秘書|が|同居|中|らしい|

「y○mada」は、固有名詞「山田」をローマ字で表現し、アルファベットの一部を伏字したものと考えられる。さらに、「山田」の先頭文字をアルファベットで省略した上で、残りを伏字とした「Y○」のような表現も考えられる。

2.2 関連研究

伏字に対応した形態素解析処理の研究として、[3−4]の様な形態素辞書との類似照合の提案がある。辞書にある形態素について、文字の削除・置換・追加等の一定の編集距離内の操作で文字列を照合するものである。例えば文献[4]の実験では編集距離1での照合を行なっているが、複数の伏字に対応させる場合には処理速度が遅くなってしまう恐れがある。また、伏字の元字に曖昧性が高い場合には、どの文字が適切であるかを推定する必要があり、さらには、記事投稿者が抽象的な表現で連続する伏字を用いた場合には辞書ベースの形態素解析では対応できない点が問題である。

2.3 アプローチ

上記で説明した伏字の特徴を考慮し以下のような分ち書き処理のためのアプローチを設定した。

○ 伏字の数/場所を限定しない

実際の CGM コンテンツ中の形態素には複数の伏字が連続・離散的になされている場合があり、元の形態素と比べたときに編集距離が 1 よりも大きくなる場合が多くなる。辞書ベースの分ち書き処理では近接性の判定のための処理時間が多くなってしまう。そこで本稿では、編集距離に制約を付けずに実時間で処理可能な分ち書き処理を検討する。

○ 伏字に用いられる文字の多義性解消

分ち書き処理後に各種自然言語処理や応用を行う際には、伏字であるかどうかを判定できると処理精度の向上が期待されるため本稿では、分ち書きと同時に、伏字であるかどうか判定する方式を検討する。

○ 伏字を含む形態素に分類情報を付与

違法有害性に関わる言語表現の多くは特定の固有名詞や固有表現を前提としていない。例えば、

（例）|田中|は|横暴|で|ある|。|

（例）|笠○|は|横暴|で|ある|。|

（例）|○○|は|横暴|で|ある|。|

のような言語表現は、1つが有害であるならば、他も有害と判定される。そのため、分ち書きの利用を想定した分類情報の付与を行う。

○ 伏字以外の文字操作に依存しない分ち書き

情報を隠匿することを目的として、伏字だけではなく文字の省略やアルファベット化等の組合せが考えられるが、厳密な組合せ規則はないために、それを見出すことは容易ではない。一方、元字に関するアプローチで述べた通り、元の単語を推定できなくても前後の文字列の出現傾向から分ち書きすることが可能な場合もある。そこで本稿では具体的な文字操作を前提とせずに対応できる方法を検討する。

2.3 提案方法

伏字を含むテキストを分かち書きししたコーパスを教師データとして用意し、点推定による分ち書き方法[5]の適用を試みる。テキストを構成する文字列の各境界について、形態素境界であるか個別に推定する手法であり、特徴として、周囲の文字 n−gram、文字種 n−gram、単語情報等を利用して学習、推定するものである。この方法では、存在しない形態素が出現する恐れがある反面、注目する文字境界の周辺の文字に関する傾向から分ち書きをおこなうので、流行語や伏字のようなシステムとしては未知な形態素の分ち書きにも対応しうるメリットがある。

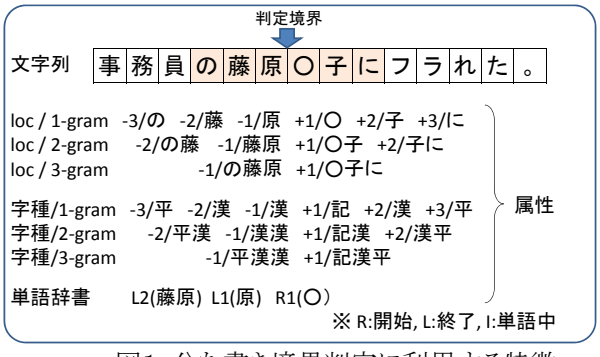

図1 分ち書き境界判定に利用する特徴

次に、伏字に用いられる文字（例えば「○」）を含む形態素について、伏字が含まれるか判定する。同時、あるいは続いて、伏字を含む形態素については違法有害処理を効率化させるための分類を行う

。

(1) 記号(伏字を含まない)
(2) 伏字を含む形態素
一般語、組織名、人名、地名、日付・時間、人工物、数字、その他

(2)の一般語以外のサブ分類については、IREX 固有表現を一部集約して用いる。

これらを分ち書きされた「○」記号を含む形態素全てにラベル付けしたものを教師データとし、CRF を用いてラベル推定を行った。

推定方法としては、伏字である/無いを分類した後に、伏字が含まれる場合にサブ分類を推定する2段階の方法と、これらを一括で分類する一段階の分類方法を実験的に比較し、最適化する。

3. 実験

3.1 実験方法

教師データとして、掲示板サービス2ちゃんねるに投稿された告発に関わる記事から「○」記号を含む 4,951 文を抽出し、MeCab[5]で分ち書きを行い、その結果を目視で修正した。評価データとして、違法有害表現抽出に関する[1]の実験で用いた 10-20 代女性を主要ターゲットとして提供されているモバイル系 CGM サイトのブログ記事及び記事コメント 72 万件から「○」を含む 439 文を抽出し、同様の分ち書き作業を行った。これらの結果について表1に示す。

2つのコーパスデータそれぞれについて、「○」を含む形態素に対して行った分類作業の結果を表2に示す。教師データと評価データを比べると、「○」が伏字として用いられている割合は教師データのほうが多いため、どちらもCGMコンテンツである点は共通しているが、利用者層はテストデータの方が限定のため、「○」用いられ方の傾向が多少異なっている可能性がある。

表 1. 実験に用いたコーパス一覧

		Train(2ch)	Test(mobile)
文数		4,951	439
形態素数	全体	202,768	31,825
	「○」含む	6,900	684
「○」文字数		8,342	873

表2. 「○」を含む形態素の分類作業結果

分類		Train(2ch)	Test(mobile)
記号		1,446	499
伏字	一般語	510	51
	人名	2,921	45
	組織名	1,341	24
	数字	124	13
	場所	448	7
	日付・時間	26	6
	人工物	84	39

点推定による分ち書き方法としては、Kytea[5]に準拠して LIBLINEAR[6]を用い、正則化最小二乗法で学習・推定した。

3.2 結果1:分かち書き

表3に教師データを4分割した交差検定の結果を示す。通常の辞書ベースの形態素解析(MeCab に Naist 辞書を利用)の場合、伏字を含む形態素の多くは、1文字の記号として「○」を処理してしまうため精度が1割程度にとどまるが、点推定を用いた時には、伏字を含む形態素が 78%の精度で獲得できていることが分かる。

一方、テストデータで分ち書きの精度を評価した結果(表4)、全体及び伏字記号を含む形態素の精度は多少落ちてしまっている。処理するテキストと同じドメインのコーパスを作成することが好ましいと考えられる。

表3 分ち書き処理の評価結果(CV=4)

分ち書き方法	精度		
	文字境界	形態素	伏字を含む形態素
点推定	0.970	0.926	0.780
MeCab	0.897	0.827	0.093

表4 分ち書き処理の評価結果(テストデータ)

分ち書き方法	精度		
	文字境界	形態素	伏字を含む形態素
点推定	0.955	0.878	0.679
MeCab	0.937	0.866	0.278

3.3 結果2: 伏字記号を含む形態素の分類

本稿では、分ち書き処理と分類情報付与を独立と仮定し、個々の処理の評価をまず行うこととした。ここでは、人が正しく分ち書きしたコーパスデータを入力して、「○」文字を含む形態素が正しく分類できるか評価を行った結果について表5に示す。

1-pass は、形態素に伏字を含んでいるかの分類と同時に伏字を含む形態素の分類を行う方法であり、2-pass は、これらを段階的に行った結果である。いずれの分類方法でも2つの分類それぞれについて、9 割前後の高い精度が得られている。

次に、2-pass の分類において、伏字の形態素に関す

る分類の精度について、表6に示す。交差検定では、件数の多い分類については概ね5割以上の精度を与えているが、件数の少ない日付や人工物は、精度が4割以下にとどまっている。これらについては、教師データ作成の時に、分類のバランスを考慮して作成することで改善されると期待される。テストデータについては、一般語を除いて、3 割以下の分類精度となっている。原因の分析が必要である。

表 5「〇」を含む形態素の分類結果

分類方法	精度	
	交差検定(CV=4)	テストデータ
1-pass	0.927	0.889
2-pass	0.946	0.934

表5 伏字を含む形態素の分類結果

分類	交差検定(CV=4)		テストデータ	
	正解件数	精度	正解件数	精度
一般語	510	0.526	51	0.520
人名	2921	0.736	45	0.270
組織名	1341	0.495	24	0.172
数字	124	0.776	13	0.250
場所	448	0.698	7	0.200
日付	26	0.250	6	―
人工物	84	0.383	39	―

4. おわりに

本稿では、伏字を含むテキストを分かち書きし、伏字として良く用いられる記号を含む形態素が伏字であるかどうかを判定し、さらに、違法有害表現の抽出の特徴として利用できるような分類が可能であるか検討した。

分ち書きしたテキストコーパスを教師データとして用いて分ち書き境界の点推定[5]を試みた所、伏字記号を含む形態素についても 7 割弱適切に行えることを示した。比較対象の MeCab については、辞書に教師データに現れた未知語を追加しなかったため、精度は相対的に低かったため、今後はこれを追加した上で、再度比較評価を行う予定である。

また、伏字記号を含む形態素の分類情報を付与して教師データとして用いた分類実験を行い、9割程度の精度で伏字が含まれているか、あるいは単なる記号として用いられているか判定できることがわかった。今後はこれらを一体に評価を行う予定である。一方、伏字を含む固有表現を構成する単語については、訓練データでの交差検定では 5 割程度、テストデータではその半分以下となってしまっている。分類に失敗した事例の分析とともに、分類カテゴリーや分類方法について改善を行う予定である。また、文献[8]では、分ち書き処理とともに品詞付与に点推定を用いている。ラベル付与という観点では、伏字を含む形態素のラベル付けに応用可能なアプローチであり、方式比較を行いたい。

参考文献

[1] 笠原要, 藤野昭典, 永田昌明: テキストに基づく違法有害記事の削除作業支援方式, 言語処理学会年次大会, 2011

[2] G. Robinson. A Statistical Approach to the Spam Problem. Linux Journal, No. 107, 2003. [3] Stoyan Mihov and Klaus U. Schulz. Fast approximate search in large dictionaries. Computational Linguistics, 30(4), pp.451-477, 2004.

[4] 藤 邦子, 今村 賢治, 松尾 義博, 菊井 玄一郎, 誤字脱字や伏字を許容する近似辞書照合技術, 言語処理学会第 17 回年次大会, pp.1143-1146 (2011).

[5] Graham Neubig, 中田 陽介, 森 信介. 点推定と能動学習を用いた自動単語分割器の分野適応言語処理学会第 16 回年次大会(NLP2010), 2010

[6] T. Kudo, Y. Matsumoto, “Chunking with support vector machines”, Proc. of the 2nd Meeting North American Chapter of the Association for Computational Linguistics, 2001.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A
library for large linear classification. Journal of Machine Learning Research, 9:1871-1874, 2008.

[8] 中田陽介, Graham Neubig, 森 信介, 河原 達也. 点予測による形態素解析, 情報処理学会 第198回自然言語処理研究会(NL-198), 2010.