極大部分文字列を用いたWebテキストの語義曖昧性解消

三谷亮介　小町守　松本裕治
{ryosuke-m,komachi,matsu}@is.naist.jp
奈良先端科学技術大学院大学

隅田飛鳥
as-sumida@kddilabs.jp
株式会社KDDI研究所

1. はじめに

近年，Web上に存在するテキストデータは増え続けている．また，Webテキストを処理するための頑健な技術の開発は様々な方面から求められている．評判情報抽出のように，意味解析や文書分類といった基盤技術が求められるアプリケーションがあるが，Web文書に対してこれらの基盤技術はうまく適用できない場合がある．たとえば，Webテキストに多く含まれる固有名詞を中心とした映画のタイトルや署名といったような複数の単語からなる表現(スターウォーズ等)は，文の意味を捉えるにあたりよい手がかりになるが，単語による分割を行うと複数の素性として扱われてしまう場合がある(スター,ウォーズ)．文書分類や語義曖昧性解消など，文脈からトピックを把握する必要があるタスクでは，関連性のある文字や単語をまとめあげる必要がある．また，既存のアプリケーションは単語分割が正しく行われる前提で構築されているが，新聞記事で学習された単語分割器は必ずしもWebテキストに対して頑健であるとは限らないという問題点もある．

新聞記事などの一般分野の記事とは異なり，Webテキストには，常に新しい情報が流入するために，Web発信の流行語等が含まれる可能性が高い．服部らの研究[1]によって，特定の分野におけるWebテキストにおいては，特に未知語の出現率が高いことがわかっている．教師あり形態素解析を行う際，単語の分割境界を決めるために様々な単語が登録されている辞書を用いるが，辞書に登録されていない語(未知語)が多いと解析の精度が低下してしまうという問題がある．新語は常に生み出され続けるため，辞書を最新に更新し続けるには多大なコストがかかる．

そこで，本研究では，Webテキストの意味解析に向けた頑健な基盤技術の開発の第一歩として，語義曖昧性解消タスクの1つ，隠語の有害性判定タスクに取り組み，意味解析の素性としての極大部分文字列の有効性を検証する．極大部分文字列の取得には辞書が不要であるため，頑健な解析を行うことができ，語義曖昧性解消タスクにおいてbag-of-wordsのベースラインと比較して性能が向上することを示す．

2. 関連研究

岡野原らは文書分類タスクにおいて，極大部分文字列集合を素性とすることで，bag-of-wordsを素性とした場合よりも高い精度を得られることを示した[2]．岡野原らとの研究とは極大部分文字列集合を素性とする点で共通しているが，局所的な文脈を見ることもできる語義曖昧性解消と，大局的な情報しか参照できない文書分類とでタスクが異なっている．

村本ら[3]はWebテキストにおける固有名詞の意味カテゴリの曖昧性を解消する研究を行った．彼らの研究と語義曖昧性解消というタスクは類似しているが，本研究では，隠語の曖昧性解消を行うため，品詞が固有名詞だけではない点，また，テキストの分割粒度として極大部分文字列を用いる点で異なる．

坪井ら[4]は単語n-gramと系列パターンマイニングによって得た部分文字列を用いてe-mailとWeb文書の著者推定を行う研究を行った．単語と違う分割粒度を用いる点で本研究と類似するが，本研究では長さが極大の部分文字列を用いる点，また，隠語の有害性判定タスクに取り組む点で異なる．

3. 極大部分文字列の抽出

文書中の全部分文字列の生起において，ある部分文字列q_1の生起がq_2と同じ位置であるような部分文字列はいくつかの集合に分割される．各集合において，長さが極大となる，

かつ，出現回数が 2 回以上の部分文字列を極大部分文字列と呼ぶ．文書中の全部分文字列をすべて列挙すると計算量は文書長 d の 2 乗に比例する．岡野原らの研究[2]では，拡張接尾辞配列を用いて，効率的に全部分文字列を列挙し，極大部分文字列を得る方法が提案された．

次の 6 つのテキスト集合から，大麻のイベントであるマリファナマーチを含む極大部分文字列を抽出した結果を表 1 に示す．

1. ...ンキーがマリファナマーチをして...
2. ...やマリファナマーチなんかでは LSD...
3. ？今年もマリファナマーチやります...
4. ...年前に大阪のマリファナマーチに...
5. ...も来年のマリファナマーチで「私...
6. ...？今年もマリファナマーチやりま...

表 1.極大部分文字列の抽出

頻度	出現した部分文字列
2回	のマリファナマーチ
6回	マリファナマーチ
2回	？今年もマリファナマーチやります、…

ある文章において，映画のタイトルや署名といったような複数の単語からなる表現の出現は，文章の意味を捉える際に良い指標となる．例として挙げた，マリファナマーチもその 1 つである．しかし，従来のような単語による分割を行うと表現特有の意味を失ってしまう場合がある．このような表現を，極大部分文字列集合として抽出することにより，複数の単語からなる表現を 1 つの素性として用いることができる．

4. Web テキストの隠語の有害性判定

本研究では語義曖昧性解消タスクの 1 つとして「隠語の有害性判定」タスクに取り組む．隠語とは特定の社会集団の中でのみ通用する意味を持つ語である．隠語の種類を表 3 に示す．隠語が曖昧であるとは，隠語が複数の意味を持つ事を示している（H→いやらしい，ヘロイン）．隠語が明確であるとは，隠語が 1 つの意味しか持たないことを示す(NAIST→奈良先端大)．隠語の有害性は，隠語の出現文脈において，その隠語が指す意味が，他人を蔑んだり，中傷したりする表現や違法行為そのもの，またはそれを助長するものと判断した場合，有と判断する．

表 2.隠語の種類

	有害	有害/無害	無害
曖昧	H	キノコ、スピード	MBA、NLP
明確	托い		NAIST

隠語の有害性判定タスクとは，隠語を含む文脈に対して，その文脈が有害/無害のどちらの意味で使用されているのかを当てるタスクである．たとえば，以下の

1. おいしいキノコご飯を食べに行こう．
2. 幻覚キノコの売買に携わる．
3. 合法キノコを食べたらスピード違反で捕まった．

という文を考える．この場合，1.の文脈においては「キノコ」は一般的な「食用キノコ」の意味で用いられているが，2.の文脈における「キノコ」はマジックマッシュルームなどの薬物を示す有害な意味で用いられている．隠語の有害性判定タスクは，隠語が文脈において有害な意味で用いられているかどうかを判別するタスクである．一方，3.のように複数の曖昧性のある隠語が含まれる場合もあるため，文書分類タスクとは異なる．

この技術を利用することで，Web ページ管理者が違法な書き込みを削除することのサポートを行うことができると考えている．通常，単語のパターンマッチングだけでは「キノコ」のような日常でよく使われる一般的な名詞の隠語の場合，無害な文脈を多数抽出してしまい，有害な書き込みの検出が困難となる．有害な隠語の曖昧性を解消することができれば有害表現だけを効果的に検出することが可能となり，管理者の負担を軽減することができる．

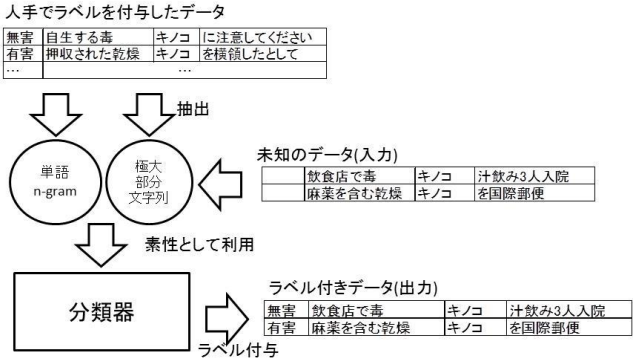

図 1.隠語の有害性判定方法

隠語の有害性判定の方法を図 1 に示す．ある文脈における隠語の有害性を人手によってラベルが付与されたデータから，単語 n-gram と極大部分文字列を抽出し，素性として分類器を学習させる．テスト時には，入力として，ラベルが付与されていない隠語とその周辺文脈を分類器に渡す．分類器は学習したモデルを用いて隠語の有害性判定を行い，入力された隠語に対して，判定結果をラベルとして付与し，出力する．

隠語の周辺文脈の抽出や，分類器のトレーニングには，テキストを何らかの粒度で分割する必要がある．本タスクでは，Web テキストを対象とするため，一般分野の新聞記事などを対象に解析した場合と比較して，単語分割の精度が期待できない．そこで，極大部分文字列を素性として用いることで，有害性判定性能の向上を図る．

5. 極大部分文字列を用いた隠語の有害性判定実験

分類器の訓練における素性として，隠語とその周辺文脈から単語 n-gram と極大部分文字列を抽出し，隠語の有害性判定性能を比較した．

5.1 使用データ

本研究で使用する隠語コーパスの作成にあたって，独自に Web テキストの収集と有害性のアノテーションを行った．収集した Web テキストを対象に，形態素解析用辞書 UniDic[1](1.3.12)と形態素解析器 MeCab[2](0.991)を用いて形態素解析を行った．次に，テキスト中に含まれる隠語に対して，前後 20 形態素を 1 文脈として 8,100 文脈を抽出，各文脈における隠語が持つ有害性の有無を人手でアノテーションした．実験では，隠語コーパスからランダムに 100 文を 10 回抽出し，テストデータとした．

5.2 ベースライン

ベースラインとして，単語 1-gram による bag-of-words を用いた．単語はコーパス作成時と同様に MeCab と UniDic を用いて分割を行った．極大部分文字列の抽出には esaxx[3](0.0.3)を用いた．教師なし解析である極大部分文字列集合の抽出では，すべてのデータから極大部分文字列を抽出した．また，隠語を含む文脈の分類には SVM[light][4] (V6.02)を使って，線形カーネルによる分類を行った．分類器のトレーニングには，以下の 7 通りの素性を用いた．

1. 単語 1-gram.
2. 単語 1～2-gram.
3. 単語 1～3-gram.
4. 極大部分文字列のみ.
5. 単語 1-gram＋極大部分文字列.
6. 単語 1～2-gram＋極大部分文字列.
7. 単語 1～3-gram＋極大部分文字列.

5.3 評価尺度

有害性判定の結果を正解率，適合率，再現率，F 値を用いて評価を行った．全 1,000 件のテストデータにおける有害，無害ラベルの分布には偏りが見られる．有害ラベルがついた隠語とその文脈の事例は 694 件であり，無害ラベルがついた事例は 306 件である．

[1] http://www.tokuteicorpus.jp/dist/

[2] http://mecab.sourceforge.net/

[3] http://code.google.com/p/esaxx/

[4] http://svmlight.joachims.org/

表 3.隠語の有害性判定結果

	正解率	適合率	再現率	F値
単語1-gram(ベースライン)	93.7	92.30	85.9	89.0
単語1〜2-gram	94.6	92.54	88.1	90.2
単語1〜3-gram	94.6	94.15	88.1	91.0
極大部分文字列	94.3	93.86	86.9	90.2
極大部分文字列+単語1-gram	94.4	92.83	88.1	90.4
極大部分文字列+単語1〜2-gram	94.4	93.50	88.4	90.9
極大部分文字列+単語1〜3-gram	94.4	94.14	86.8	90.3

6. 考察

10 回の実験結果の平均値を表 3 に示す．実験の結果，単語 1-gram のベースラインと比べて，極大部分文字列を素性として用いたことで，すべての評価尺度において性能が向上した．また，単語 1-gram を用いた手法と，単語 1〜3gram までを用いた手法を比べると，より長い単語列を用いたほうが全ての評価尺度において性能が向上した．そのため，本タスクでは UniDic における短単位より長い粒度の単位が効果的であることがわかった．極大部分文字列を用いた提案手法は単語 1-gram と単語 1〜2-gram を素性に用いた手法と同程度の F 値を達成しているが，提案手法は単語分割されたデータを必要としないという点で優れる．また，提案手法極大部分文字列の出現頻度をパラメータとして持つが，このパラメータを調整することによってさらに性能をチューニングすることが可能である．

さらに，極大部分文字列は単語と併用して素性に加えることができる．このタスクにおいては単語 1〜2gram と極大部分文字列の 2 つの素性を用いることで再現率が最も高くなった．隠語の有害性判定タスクでは，有害な表現を漏れなく検知するという目的から再現率を高くすることが求められるため，単語 n-gram と極大部分文字列を組み合わせた手法が有害性判定タスクに対して有効であることがわかった．

7. エラー分析

分類を間違ってしまった事例，全 396 件の内，270 件が有害を無害として判定したものだった．その中には，人間が実際に見れば判断できる文脈も多数含まれた．例として，隠語「葉っぱ」(有害:大麻,無害:植物の茎や枝につくもの)を含む文脈 1 を挙げる．また，人間が見てもよくわからない例として「weed」(有害:大麻,無害:雑草)を含む文脈 2 を挙げる．

1. ...間違って葉っぱが一枚でもポケットに入ってたりしたら，逮捕され，裁判...
2. ...諦めちゃダメなんだ！Weloveweed!この記事にコメント...

エラー分析を踏まえ，次の 3 つの手法を用いることで性能改善を考えている．

1 つ目は，素性ごとに重みを設定する方法である．現在は，全素性の重みを等しく設定している．有害な文脈に頻出する素性などに対して重み付けすることで判定性能を改善できると考えている．

2 つ目は，素性抽出の際，局所的な文脈を考慮することである．現在は，文脈中に含まれる各要素の位置を考慮していないが，対象となる単語の周辺を考慮したり，出現位置を素性にすることで，統語的な手がかりを用いることができる．

3 つ目は，有害性に曖昧性のない隠語リストを用いて，教師なしに有害性のある文脈を自動的に獲得する手法である．隠語リストにない名詞は一般の辞書から取得できるため，大規模にデータを獲得することができる．

8. おわりに

本研究では，極大部分文字列を素性として，Web テキストに対する隠語の有害性判定タスクに取り組んだ．極大部分文字列を用いると，複数の単語からなる表現を 1 つの素性として用いることが可能となる．極大部分文字列と単語 1〜2-gram を素性に用いた結果，再現率は 88.44 となった．

参考文献

[1]服部峻，亀田弘之　Web テキストにおける未知語の頻度調査 IEICE Technical Report 2010-05 pp.7-12
[2]Daisuke Okanohara,Junichi Tsujii　The Categorization with all Substring Features. SDM 2009-04 pp.838-846
[3]村本英明，鍛冶伸裕，吉永直樹，喜連川優 Wikipedia と Web テキストを利用した固有名の意味カテゴリの曖昧性解消　言語処理学会第 17 回年次大会 2011-03 pp.774-777
[4]Yuta Tsuboi,Yuji Matsumoto Authorship Identification for Heterogeneous Documents 自然言語処理研究会報告 2002-03 pp.17-24