有害性スコアリングによるWebテキストにおける隠語の発見

三谷 亮介　　小町 守　　松本 裕治　　隅田 飛鳥　　服部 元　　小野 智弘
奈良先端科学技術大学院大学　　KDDI研究所
{ryosuke-m, komachi, matsu}@.is.naist.jp　{as-sumida, gen ,ono}@kddilabs.jp

1　はじめに

近年インターネット上では，覚せい剤や大麻などの薬物の売買，未成年の売春などの違法行為や学校裏サイトでのいじめなどの様々な問題が蔓延しており，社会的な問題となっている．しかし，大量の投稿から，問題となる投稿だけを人手で管理するのは難しい．そのため，コンピュータによって，自動的に問題がある投稿を判別する技術が広く求められている．

コンピュータによって，有害な投稿を識別するためには様々な問題がある．その中でも，有害な語義を一般的な単語に隠して用いられる“隠語”は難しい問題の1つである．例えば，ある隠語「キノコ」では，通常用いられる「食用キノコ」としての意味と，有害な意味である「マジックマッシュルーム，幻覚キノコ」としての意味の両方を考慮する必要がある．そのため，単純なパターンマッチングだけでは，有害な情報のみを抽出することは難しい．さらに,「キノコ」のような一般的な名詞が隠語として使われている場合，無害な名詞としての用法が多く抽出され，有害な情報を人手で発見することが困難となる．

そこで，我々は以前，隠語を含む文脈の有害性識別タスクを提案した[3]．その研究では，人手で作成した隠語のリストに記載された隠語のみを有害性識別の対象としていた．しかし，変化の激しいインターネット上の有害情報を識別するためには，識別対象である隠語のリストを常に最新の状態へと更新し続ける必要がある．その一方で，隠語リストの更新には，有害な情報に関する専門的知識と多くの文章を読み込むための膨大な人的コストが必要であり，困難な作業である．

そのために，本研究では，隠語リストを拡張する作業をサポートすることを目的として「隠語の発見タスク」を提案する．また，隠語を含む文脈に対する有害性識別を応用した手法と，単語の頻度を用いた手法を提案する．さらに，ドメイン外文書を利用することで，低コストに性能を向上させることが可能なことを示す．

2　関連研究

Lauら[1]は，単語の新語義を発見するタスクに対して，文書に生起した単語を教師なしに同じ語義ごとにクラスタリングするWord Sense Inductionの手法を適用した．本研究とは，ある名詞が持つ隠語の有害な語義を新語義の1つとして考える点と，単語の用法に違いがあると期待できる2種類のコーパスを利用する点について類似する．しかし，Lauらは単語の語義をすべてを区別して扱っている点に対して，本研究では，ある単語が持つ語義を有害な語義と無害な語義の2つのみに分けることで，手法を簡略化し，隠語を獲得するという本タスクに特化させている点で異なる．

Mihalceaら[2]は，ある文脈中に生起した単語が生起した年代を当てるタスクWord Epoch Disambiguationを提案した．Word Epoch Disambiguationタスクでは，単語の用法の経時的な変化を捉えることで，dinnerやsurgeonなどの経時的に語義が変化した語の生起年代を推測する．本研究では，生起した年代や場所が異なる文書を“ドメインが異なるコーパス”として扱う点で共通する．しかし，Mihalceaらは，単語の用法から生起年代を推測する一方で，本研究では，ドメインや生起した年代から単語の用法を推測し，隠語の発見に応用する点で異なる．

3　隠語の有害性識別

3.1　隠語

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．本研究において取り扱う隠語を,「複数の語義を持ち，さらに，1つ以上の“有害”な語義を持つ名詞」と定義する．さらに,「ある隠語の語義が，インターネットサービス上において風紀を乱す行為，また，それを示唆する行為を指す」その時に，隠語が“有害性”を持っていると判断する．隠語とその語義の例を表1に挙げる．

表 1: 隠語の語義

隠語	無害な語義	有害な語義
スピード	速度	覚せい剤
草	植物	大麻
キノコ	茸類	マジックマッシュルーム

3.2 隠語コーパス

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 (1.3.12)[1] と形態素解析器 MeCab (0.991)[2] 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 254 種類に対して， 前後 20 形態素を 1 文脈として 8,097 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．

表 2 に，隠語コーパスのサンプルを示す．

3.3 有害性識別の手法

隠語の有害性をラベルとして付与した隠語コーパスから 100 文をテストデータ，残りをトレーニングデータとして利用する．トレーニングデータから単語 n-gram を素性として抽出し，分類器（SVM）をトレーニングする．テストでは，入力として有害性ラベルを外したテストデータを分類器に渡す．分類器は学習したモデルを用いて隠語の有害性識別を行い，入力された隠語に対して識別結果をラベルとして付与し，出力する．我々の過去の研究 [3] では，この手法により，高い性能 （F 値 91）で隠語の有害性識別が可能なことが示されている．

4 有害ドメイン文書を用いた隠語の発見

「隠語の有害性識別」タスクが隠語として使われる可能性がある単語に対する有害性の識別を行なっていたことに対して，本研究で取り組む「隠語の発見」タスクは，隠語以外も含む単語集合に対して隠語らしさを判定するタスクである．本研究では，既存の言語資源を用いて，与えられた隠語の候補から，有害な意味で使われる隠語を選び出すための手法を提案し，その検証を行う．

[1]UniDic http://www.tokuteicorpus.jp/dist/
[2]MeCab http://mecab.sourceforge.net/

4.1 隠語コーパスによる単語の有害性識別

本手法は，隠語コーパスから隠語の周辺文脈の情報のみを用いることで，有害性識別を隠語の発見タスクに応用したものである．覚せい剤を意味する隠語は複数存在することから，同じ意味を持つ隠語は同様の用法によって使われるのではないかという仮説に基づき，隠語の情報を使用せずとも文脈のみを用いることで，分類が可能ではないかと考えた．

隠語コーパスから隠語の周辺文脈の有害性を学習し，事例として与えられた隠語の候補とその周辺文脈に対して，有害性の有無を識別する．本研究では，1 つの隠語候補に対して有害ドメイン文書から T 個の事例を抽出し，テストデータとして用いる．次に，隠語コーパスから隠語の周辺文脈の有害性を学習する．そして，テストデータの各事例に対して，超平面からの距離を取得する．最後に，事例の結果の集合において，有害ラベルへの最大の超平面距離を取得し，その単語 w の隠語らしさのスコア S_w として用いる (式 1).

$$S_w = \max\{Distance(w_i), i = 1, 2, \cdots, T\} \quad (1)$$

4.2 ドメインの異なる文書間における名詞の生起頻度の活用

一方，隠語コーパスのアノテーションは作業コストが高いために，文章中における単語の生起数に着目した教師データを必要としない手法を提案する．Han ら [1] の研究を応用し，隠語の語義を新語義の 1 つとして考えることで，隠語の獲得タスクに適用する．ドメインによって単語が異なる語義で使用されるならば，それらの文書間において単語の頻度が異なるという仮説に基づき，有害な文脈を多く含むと思われる文書と，無害な文脈を多く含むと思われる文書から，それらの間における隠語候補の単語の頻度を比較する．本研究では，式 2 を用いて，単語 w の隠語らしさのスコアを算出し，ランキングを行う．

$$S_w = \frac{\text{有害な文書における単語 } w \text{ の出現頻度}}{\text{無害な文書における単語 } w \text{ の出現頻度}} \quad (2)$$

5 隠語の発見の実験

提案手法の有効性を検証するために，手法の比較実験を行う．

表 2: 隠語コーパスのサンプル

有害性	前文脈	隠語	後文脈
有	沖縄県警は 6 日，麻薬を含む乾燥	キノコ	(マジックマッシュルーム) を国際郵便で密輸入
有	まとめます！血溶き最強説 !?| ＢＬＯＧＴＯＰ｜	ネタ	の安全な隠し場所
無	逮捕された新聞記事で絵葉書作って	売り	ました。 1 枚 150 円で 20 万枚売れたので

5.1 隠語候補の選定

対象とする単語は，隠語コーパスにおいてラベル付けされた有害と無害の両方の文脈に 1 回以上出現した名詞 2,020 語である．また，それらに含まれる隠語の数は 57 語である．

5.2 有害ドメイン文書コーパス

有害ドメイン文書コーパスは，薬物とアダルトの 2 つの分野の Web ページから独自に構築した．一般的な新聞記事などの文書と比べて，薬物の乱用などのドメインに依存した有害な表現が多く使われていると考えられる．記事数は 31,333 件である．

5.3 ベースラインと提案手法

ランダムな単語選択

本研究で対象とする隠語 57 語を含む名詞 2,020 語を対象として，ランダムに並べ替えを行った場合をベースラインとする．

有害性識別による隠語の発見

本研究では，有害性識別のための分類器には LIBSVM [3]Ver3.00，素性は bag-of-words のみを使用した．そして，対象となる 2,020 単語に関して，事例数 T=100 として，1 単語 100 件の文脈を有害ドメイン文書から抽出し，有害性識別のテスト事例として使用する．また，単語 w の隠語らしさのスコア S_w を式 1 にて計算する．

ドメインの異なる文書間における名詞の生起頻度の活用

本研究では，2 種類の無害/有害文書集合を使用して実験を行う．まず，隠語コーパスの有害性ラベルに基づいた手法である．隠語コーパスに付与されている有害性のラベルに基づき文脈単位でコーパスを 2 分割し，有害と無害のそれぞれの文脈の集合を別ドメインの文書として扱う．

次に，有害ドメイン外データである青空文庫[4]コーパスを無害な文書として使用する手法である．本研究では，青空文庫が著作権の有効期限（著者の没後 50 年間）が切れている作品が多数集められている点に着目し，無害な語義としての用法が期待できると考えた．2012 年 10 月にコーパスを作成した際の作品数は 11,836 件である．この青空文庫コーパスと有害ドメイン文書コーパスをそれぞれ，無害/有害文書として扱う．そして，それぞれの文書中に生起する単語の頻度を計算し，式 2 によって，それぞれの隠語候補のスコアを計算した．

5.4 評価尺度

評価には，各手法が算出した隠語候補に対する隠語らしさのスコアのランキングから，上位 K 語に含まれる隠語のカバー率を用いて評価を行う．

5.5 実験結果

各手法において，隠語らしさのスコアによるランキング結果から上位 100-2000 個までの隠語のカバー率の推移を図 1 に示す．

6 実験結果の考察

表 1 の結果から，提案手法のいずれもがランダムに単語を並べ替えるベースラインよりも高い性能で隠語を発見できることがわかった．また，各手法におけるランキング結果を表 3 に示す．

有害性識別を用いた手法では，ランキング上位 500 件以降の時，高いカバー率で隠語を発見することができた．特に，アップやヤクなどの隠語コーパス内で有害/無害文脈間の頻度にあまり違いが見られなかった語を高くスコア付けすることができた．

隠語らしさのランキングにおける上位 500 件においては，隠語コーパス内の単語の頻度に基づく手法が効

[3]LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4]青空文庫 http://www.aozora.gr.jp/

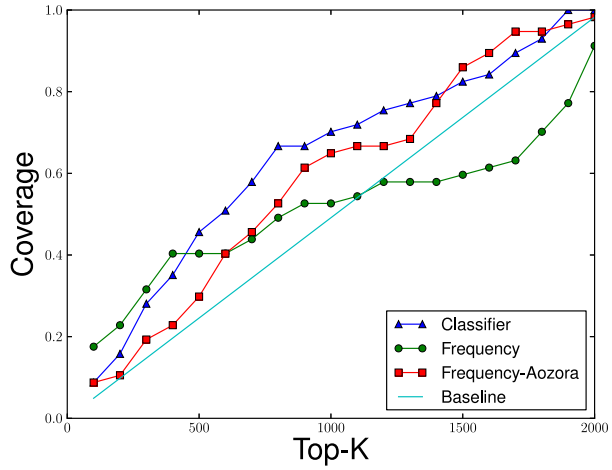

図 1: 上位 K 語の隠語候補における隠語発見のカバー率

率よく隠語を発見している．例えば，「サポート」といった隠語は有害文脈中で無害な文脈における頻度よりも高い頻度で登場しており，これらの隠語を有害と判定できるトレーニングデータが少なかった有害性識別による手法よりも，高くスコアリングすることができた．しかし，隠語コーパスの頻度に基づく手法では，無害な文脈において登場する頻度が多い単語に対して，高いスコアを与えることが難しく，上位 1,000 件以降は隠語の獲得効率が低下した．その一方で，ドメイン外のデータである青空文庫を用いた手法は，隠語コーパスのデータを使った手法よりも効果的に隠語を発見することが可能なことがわかった．実際に，「罰」などの隠語の有害文書内での異常な生起を検知することができた．この手法は有害性識別を用いた手法と比べて，性能では劣るものの，アノテーションなどの作業コスト無しで実行可能な手法でありながら，アノテーションデータから頻度を取得した手法よりもカバー率における AUC において 14 ポイント高い性能で隠語が発見できた．

最後に，本研究において提案した手法では発見することが困難な隠語を表 4 に示す．これらの隠語は，隠語コーパス中に有害な語義として出現する回数が少なく，さらに，ドメインが異なる文書と比較しても頻度に差が見られなかった語である．これらの語に関しては，シソーラスを用いて類義語を取得し，同じグループ内での隠語らしさを比較する手法や，異なるドメインのデータを適用するなど，新しいアプローチで抽出を行う必要があると考えている．

表 3: 各種法により獲得した隠語のランキング結果

隠語	有害性識別	頻度（隠語コーパス）	頻度（青空文庫）
ヤク	126	1336	132
アップ	943	1938	968
サポート	1155	22	23
罰	1622	1660	293

表 4: 有害性スコアリングでは発見が困難な隠語

隠語	有害性識別	頻度（隠語コーパス）	頻度（青空文庫）
コーラ	1647	1685	1681
援助	1794	1831	1821

7 おわりに

本研究では，人手で更新し続けるには高コストな作業である隠語のリストの更新をサポートするという目的で，文書中の単語から隠語を発見する「隠語の発見タスク」を提案し，その解決に取り組んだ．本研究で提案した手法は，そのいずれもがベースラインとして設定したランダムに単語を選ぶ手法よりも高い性能で隠語の発見が可能なことを示した．また，有害性識別を用いる手法は頻度に基づく手法よりも，カバー率における AUC において 24 ポイント高い性能で隠語の発見が可能である．一方で，ドメイン外のデータである青空文庫を使って，有害文書との頻度の差を比較することで，アノテーションデータの頻度を用いる手法よりも高い性能で隠語の発見を行うことが可能であることを示した．

今後の課題として，今回提案した手法では発見することが困難な隠語に対して，シソーラスなどの高度な意味情報を用いるなどしてアプローチする必要があると考えている．また，隠語を新語義の 1 つとして考え，新語義発見タスクで Lau ら [1] が用いた hierarchical Dirichlet process などの教師なし手法を適応することでも，文脈中で生起した単語の語義に合った隠語らしさのスコアをランキングすることが可能であると考えている．

参考文献

[1] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word sense induction for novel sense detection. In *Proc. of EACL*, pp. 591–601, 2012.

[2] Rada Mihalcea and Vivi Nastase. Word epoch disambiguation: Finding how words change over time. In *Proc. of ACL*, pp. 259–263, 2012.

[3] 三谷亮介, 小町守, 松本裕治, 隅田飛鳥. 極大部分文字列を用いた web テキストの語義曖昧性解消. 言語処理学会第 18 回年次大会, pp. 1292–1295, 2012.