6C-3

危険性の視点からの評判分析

山本 智也 † 徳久 雅人 † 村田 真樹 †
鳥取大学 工学部 知能情報工学科 †

1 はじめに

事物に対する P/N という評価極性は議論が多い．ネガティブな事象には〈恐れ〉，〈怒り〉，〈悲しみ〉，〈嫌だ〉などの感情があり，危険性が強ければ〈恐れ〉に傾くと予想される．危険性の強弱が分かれば感情分析の役に立つと考えられる．そこで，本稿では危険性を表す表現の自動収集を目的とする．

本稿では *SO-score* による算出方法 [1] を使用する．危険だと思われる表現（例えば「危険」）と安全だと思われる表現（例えば「安全」）を *SO-score* に適応させる．また，本稿では話題を指定する．これは一般に「ホテル」の評判分析ではホテルの文書集合から情報を抽出することと同じ考え方である．

ここで問題点はコーパス全体からの *SO-score* の算出では偶然性が高く，かつ，単語の候補が多すぎることである．そこで，本稿では *SO-score* を多数算出し，統計をとる方法を提案する．

2 提案手法

まずブログ記事を n 分割し n 個の文書集合にする．次に各文書集合から指定する話題を含む文を抽出し，これらを話題文書集合とする．さらに話題文書集合から危険性および安全性のある文を抽出し，それぞれを危険文書集合，安全文書集合とする．話題文書集合から単語リストを作成し，その各単語について *SO-score* を算出する．*SO-score* は 1 単語につき n 通りが得られるので，単語ごとの *SO-score* の分布を求める．安定して出現し，かつ「危険」あるいは「安全」に傾く単語を選び出す．

2.1 ブログ記事の分割

分割することで文書集合に出現する単語にランダム性を持たせる．これは，時期ごとに分割すると「雪道」など時期に依存する単語が安定して出現しないためである．方法は，ブログ記事の日付ごとに通し番号を付け，その番号を分割数 n で割った余りによってグループ分けを行う．今回は 30 分割とする．ブログ記事は 2008 年 7 月 31 日から 2013 年 4 月 31 日までを使用する．

2.2 話題文書集合の作成

分割した各ブログ記事から指定する話題を含む文を抽出する．

2.3 危険文書集合と安全文書集合の作成

話題文書集合から危険性のありそうな文と安全性のありそうな文を抽出する．危険性のありそうな文は「危険」もしくは「危＋〈ひらがな一文字〉」のマッチング，安全性のありそうな文は「安全」のマッチングにより文をまずは粗く抽出する．

文章には多くの表現が存在する．その中でも，「〜だから危険」という意味をもつ文は「〜」の部分に危険性がありそうだが，「危険だから〜」という意味をもつ文はその部分においてむしろ危険を回避しているように思われる．そのため，「危険だから〜」という意味をもつ文を取り除くためのフィルタを作成した．

フィルタには SVM の 2 値分類を用いる．素性は「危険」あるいは「危＋〈ひらがな一文字〉」より前に出現する単語列には「L：」，後に出現する単語列には「R：」という文字列を付与したものである．単語列は unigram と bigram を使用しており，unigram については助詞を取り除いている．この素性を与えることで前と後に出現する単語の傾向をつかみ，「〜だから危険」と「危険だから〜」という文の分類ができると考えた．SVM のトレーニングデータは「自転車」の話題についての正例 423 文，負例 383 文を用意した．他の話題や安全性のある文についても同じトレーニングデータを使用する．

2.4 単語リストの作成

話題文書集合に出現する一般名詞と用言性名詞を抽出し，単語リストを作成する．各単語について，出現回数を，話題文書集合，危険文書集合，および，安全文書集合からカウントする．

2.5 *SO-score* の算出

Turney は評価極性の算出方法として，*SO-score* を用いる手法を提案した [1]．*SO-score* の計算式を以下に示す．

$$SO\text{-}score(t) = PMI(t, \text{“好評表現”}) - PMI(t, \text{“不評表現”})$$
$$PMI(a,b) = \log_2 \frac{p(a,b)}{p(a)p(b)}$$

この式は，語句 t の評価極性を算出する．$p(a,b)$ はコーパス内において語句 a と語句 b が同一文で共起する

Sentiment Analysis from the point of view of danger
†Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University

確率，$p(a)$ は a を含む文がコーパス内に出現する確率をそれぞれ表す．*SO-score* の値が正の場合は t が好評極性であり，負の場合は t が不評極性だと解釈する．

本稿では，好評表現を「安全」，不評表現を「危険」または「危＋〈ひらがな一文字〉」とし，安全/危険文書集合ならびに話題文書集合から確率を求める．

2.6 統計

1 単語につき n 個の *SO-score* が得られるので，その単語のヒストグラムを作成する．ヒストグラムは縦軸に度数，横軸に *SO-score* の階級とする．階級の幅は 2 とする．なお，−1 以上 1 未満の階級がある．

2.7 評価極性の決定

度数の総和が 30 となり，かつ −1 以上 1 未満の階級の度数が最大でないヒストグラムを有する単語を選び出す．選んだ単語のうち，最大度数が負の階級の場合は危険性，正の階級の場合は安全性のある単語とする．

3 実験

3.1 単語の抽出

「自転車」,「自動車」,「病院」という各話題について危険性のある単語の抽出を行った．ブログ記事全体において各話題文書集合の総数はそれぞれ 1,318,256 文，828,448 文，2,192,923 文である．各話題について評価極性の決まった単語を全て表 1，表 2，表 3 に示す．

表 1 :「自転車」の危険および安全な単語

危険	片手	メール	お婆さん	結構	不注意	当然
	道	携帯	目	傘	スピード	ブレーキ
	一	状態	練習	運転	灯火	非常
安全	子ども	日	ヘルメット	ルール	交通	則
	利用者	高齢者	対策	整備	確認	認証
	教室	推進	組立	教育	基準	指導
	環境	マーク	協会			

表 2 :「自動車」の危険および安全な単語

危険	致死	過失致死	タクシー	バイク	違反	気
	自転車	運転	人			
安全	システム	衝突	開発	技術	協会	乗用車
	制度	リコール	生産	製品	性能	対策
	問題	確保	センター	バス	仮称	一貫
	コンテナ	原因	関係	世界	向上	法律
	海陸	環境	構造	基準	品質	運送

表 3 :「病院」の危険および安全な単語

危険	状態	生命	医者	後	母	命	救急
	診察						
安全	看護	対策	医療	地域	管理	機関	安心
	委員会						

3.2 評価

下記の精度と再現率で評価する．結果を表 4 に示す．

精度： 自動で抽出した危険性のある単語のうち，人の判断で危険性があるとされた割合．5 名に，危険/安全の両単語を示し，判定してもらう．

再現率： 一般に危険性のありそうとされる単語が抽出できた割合．5 名に，危険性のありそうな単語を 10 個ずつ連想してもらう．その単語のうち，いくつの単語が自動で抽出できているのかを評価する．

表 4 ：精度と再現率

話題	精度	再現率	抽出数
「自転車」	0.59	0.22	18
「自動車」	0.67	0.08	9
「病院」	0.28	0.06	8

3.3 フィルタとヒストグラムの性能評価

フィルタとヒストグラムの性能評価を行った．フィルタを使用しない場合，ヒストグラムを使用しない場合，両方を使用しない場合の 3 通りにおける精度と再現率を求めた．対象とする話題は「自転車」とし，1 名で行う．精度を求める際，ヒストグラムを使用しない場合は膨大な量の単語を抽出するため，ランダムに 20 件のサンプリングをする．再現率については，3.2 節で連想してもらった単語を用いる．その結果を表 5 に示す．

表 5 ：フィルタとヒストグラムを使用しない場合の評価

手法	精度	再現率	抽出数
フィルタ無し	0.50	0.22	24
ヒストグラム無し	0.20±0.18	0.52	3,228
両方無し	0.10±0.13	0.50	3,266

（± 値は，危険率 5% の誤差）

3.4 考察

3.2 節の再現率が低い原因を分析した．抽出失敗の例として「自転車」の場合には「パンク」や「事故」などがあった．これらの単語と共起する単語を調べたところ，「自転車」と「事故」の共起の場合には「歩行者」や「死亡」などが多く出現している．

「事故」や「死亡」など危険性の最たる表現を話題に応じて不評表現に定めることで，危険性を表す単語の抽出性能は向上すると予想される．なお,「怖い」,「恐ろしい」,「安心」などの感情表現語はいずれの話題にも適応できる可能性がある．これらの表現を用いて実験を行うと，得られる単語の種類が変わり，3.1 節と同数程度の単語が抽出できた．

4 おわりに

ヒストグラムを用いることで単語の候補を絞り込むことができた．ヒストグラムを用いない場合は，候補が多いため必要な単語を含みやすく，再現率が上がるが，精度が下がった．以上により，危険性を表す単語を安定して自動収集する一つの方法を示すことができた．

参考文献

[1] Turney, P. D:“Thumbs Up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews.” In Proc. of ACL2002, pp.417-424, 2002.