トラブルを表す文のWeb からの抽出

丹治 広樹　村田 真樹　柿澤 康範　Stijn, De Saeger
鳥澤 健太郎　山本 和英

長岡技術科学大学
情報通信研究機構
北陸先端科学技術大学

発表者　斉藤章久

はじめに

- 何かしらのトラブルに遭遇したとき、Webを参照することが多くなった。
- しかし、膨大の情報量を持つWeb上では、欲しい情報を得るには時間や労力がかかってしまう。

↓

トラブルを表す記事を自動抽出することによってコスト軽減できるのではないのか？？

辞書の作成（1）

・トラブルを表す表現の品詞として、名詞・動詞・形容詞を用いた。

De Saeger et alの先行研究で獲得されたトラブルを表す名詞と、その名詞と係り受け関係にある動詞を人手で選別

↓

名詞は20429種類、動詞が2790種類

・同様に獲得した形容詞に、Webで公開されている評価表現辞書のうち否定極性の形容詞を加えて人手で選別

↓

形容詞は954種類

辞書の作成（2）

その他に

- 名詞と形容詞のペアのトラブルを表すフレーズ5909種類
- 「～できない」「～しにくい」当のパターンを14種類
- 「びしょびしょ」などの擬音のパターン110種類

分類器(1)

- 本研究では機械学習の手法として、最大エントロピー法およびサポートベクトルマシン法を用いた。

- サポートベクトルマシン法においてCは全ての実験で1、dは1と2を用いた。（Cはソフトマージンのためのパラメータ、dは多項式カーネルの次元数を表す）

分類器(2)

・辞書との単純マッチングを用いた方法

↓

辞書に記載されている表現を閾値以上の数だけ含む文はトラブルを表す文であると判定した。

・閾値については値を代えて実験し、経験的に最良の閾値2 を用いて評価を行った。

評価実験

- Web文書には、Yahoo!知恵袋および検索エンジン基盤TSUBAKIを使用した。
  （Yahoo!知恵袋は「question」だけを利用）

- 最大エントロピー法およびサポートベクトルマシン法で使用した素性を次のスライドに記した

使用した素性リスト(1)

素性番号	素性
S1	文章の長さ
S2	文章中の単語uni-gram
S3	文章中の単語bi-gram
S4	文章中の単語tri-gram
S5	文章に含まれる単語数
S6	文章の平均単語長
S7	文章中の各分の文末文字列
S8	最初の分の文末文字列
S9	最初の文の長さ
S10	最後の分の文末文字列

素性番号	素性
S11	最後の文の長さ
S12	名詞辞書とマッチした数
S13	マッチした名詞を含む単語uni-gram
S14	マッチした名詞を含む単語bi-gram
S15	マッチした名詞を含む単語tri-gram
S16	動詞辞書とマッチした数
S17	マッチした動詞を含む単語uni-gram
S18	マッチした動詞を含む単語bi-gram
S19	マッチした動詞を含む単語tri-gram
S20	形容詞辞書とマッチした数

使用した素性リスト（2）

素性番号	素性
S21	マッチした形容詞を含む単語uni-gram
S22	マッチした形容詞を含む単語bi-gran
S23	マッチした形容詞を含む単語tri-gran
S24	フレーズ辞書とマッチした数
S25	マッチしたフレーズ
S26	パターン辞書とマッチした数
S27	マッチしたパターン
S28	擬音辞書とマッチした数
S29	マッチした擬音
S30	全ての辞書でマッチした総数

基本実験

- 人手でトラブルか否かのタグを付与したYahoo!知恵袋1000文（うちトラブル文は281文）とTSUBAKIデータ1000文(うちトラブル文は128文)を用いて10分割交差検定を行った。

↓

これらの結果から最大エントロピー法、サポートベクトルマシン法のd=1および2において最良となる閾値を得た

オープンテスト

基本実験で用いたYahoo!知恵袋1000文とTSUBAKIデータ1000文を学習データとし、新たにタグを付けられたそれぞれの1000文をテストデータとして、オープニングテストを行った

オープンテストの結果（1）

	適合値	再現率	F値
Beseline	0.263	1.000	0.416
DM	0.473	0.806	0.596
ME	0.592	0.840	0.695
SVM1	0.639	0.768	0.698
SVM2	0.633	0.715	0.671

Yahoo!知恵袋でのオープンテスト結果である。

それぞれの項目は「DM」が辞書との単純マッチング法、「ME」が最大エントロピー法、「SVM1・2」はサポートベクトルマシン法のd=1および2のことである。

なお、「Beseline」は出力をすべてトラブルと判定した場合である。

オープンテストの結果（2）

	適合値	再現率	F値
Beseline	0.126	1.000	0.224
DM	0.303	0.690	0.421
ME	0.338	0.619	0.437
SVM1	0.409	0.556	0.471
SVM2	0.386	0.540	0.450

TSUBAKIデータでのオープンテスト結果である。

F値がBeselineのF値よりも大きく上回っていることがわかる。

また、どちらの場合でもSVM1が最も高いF値を得た

不正解な出力

①トラブルでない文をトラブルと判定したパターン

（例）【弊社は個人情報の紛失、破壊、改ざん、漏えいなどを防止するため、不正アクセス、コンピュータウィルス等に対する適切なセキュリティー対策を講じます。】

「紛失」「不正アクセス」などの表現が列挙されたためトラブルと判定されたと思われる

②トラブルの文をトラブルでないと判定したパターン

（例）【データが入ったe-amusement passを紛失した模様。】

「紛失した」という決定的な文があるが、辞書とマッチした回数が一回であり、素性としては影響力が薄かった可能性が高い。

α値からの重要性判断

- 最大エントロピーで求まるα値を正規化した値は、大きいほどシステムが判断する際に重要である素性であると示す。

- Yahoo!知恵袋、TSUBAKIデータにおける素性の正規化α値をこの後のスライドに記した。

  なお、重要度1はトラブルと判定する場合の正規化α値。重要度2はトラブルでないと判定する場合の正規化α値とする。

Yahoo!知恵袋における重要な素性

素性の単語	重要度1	素性の単語	重要度2
S2_が	0.664	S2_は	0.664
S30_1	0.662	S2_お	0.643
S2_しまい	0.633	S2_って	0.634
S30_3	0.627	S8_?	0.613
S3_のですが	0.617	S10_?	0.596
S20_1	0.616	S2_人	0.591
S3_わかりませ	0.615	S2_2	0.590
S4_わかりません	0.615	S2_聞き	0.589
S7_かりません。	0.614	S3_?	0.588
S16_1	0.607	S17_違う	0.585
…	…	…	…

「が」「のですが」などの逆接を表す素性がトラブルと判定することが多いようである。

逆に「？」といった疑問形の文末はトラブルと判定されにくい。

また、「わかりません」はトラブルで、「？」が文末の文はトラブルではないと判定されやすい。

TSUBAKIデータにおける重要な素性

素性の単語	重要度1	素性の単語	重要度2
S30_1	0.712	S2_・	0.635
S2_場合	0.674	S2_ます	0.598
S16_1	0.665	S2_や	0.593
S30_3	0.658	S2_ように	0.592
S3_のに	0.642	S2_.	0.589
S2_ほど	0.628	S1_150	0.582
S26_1	0.623	S2_たい	0.579
S2_あり	0.617	S2_を	0.571
S2_が	0.613	S2_か	0.568
S2_車	0.609	S2_ことは	0.566
…	…	…	…

「場合」「ほど」などがトラブルと判定することが多いようである。

「ます」「たい」といった疑問形の文末はトラブルと判定されにくい。

おわりに

結果

・比較的トラブルの事例が多く含まれるWeb文書ではF値0.698

・トラブルの事例がWeb全体と同程度の割合で含まれる文書ではF値0.471

を得ることできた。

今後の展望

- Yahoo!知恵袋のような記事には、質問文からトラブルを抽出することによって対応する回答文から解決策の文を見つけ出す
- トラブルを表す文の前後の分も考慮して、TSUBAKIデータのようなWeb文書からでも解決策の文を抽出。
- 辞書の拡張