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 {tanji, ykaz}@nlp.nagaokaut.ac.jp
‡ 情報通信研究機構 {murata, stijn, torisawa}@nict.go.jp
§ 北陸先端科学技術大学 s0710017@jaist.ac.jp

1 はじめに

近年の Blog や Web 掲示板の発展に伴い、何らかのトラブルに遭遇したときやあるモノに関するトラブルについて調べるとき、Web を参照することが多くなった。その一方で、Web 上には膨大な量の情報が存在するため、必要な情報を得るためには時間や労力がかかってしまう。このとき、トラブルを表す記事を自動抽出することによって調査対象を絞り込み、コストを軽減することができる。

先行研究では、トラブルに関する名詞および Object-Trouble ペアを同定し、トラブルを表す表現を獲得している[1,3]。しかし、具体的な利用方法はそうした表現を用いてキーワード検索をすることにとどまっており、検索でヒットした文が実際にトラブルを表現しているかどうかはわからない。そこで、本研究では獲得した表現を素性として機械学習を行い、Web 文書集合の中からトラブルを表現している文のみを抽出することを狙う。

2 既存研究

トラブルに関する先行研究として、De Saeger et al.[1]や鳥澤[2]、Torisawa et al.[3]の研究がある。De Saeger et al.[1]は、「トラブル」等の下位概念語や物と動詞との係り受け関係等を用いて Object-Trouble ペアを同定している。また、鳥澤[2]は語の共起頻度や動詞項位置をもとに物に対する用途や準備に使用される表現を獲得している。これらを用いて、Torisawa et al.[3]はトラブルや利用方法等を視覚的にわかりやすく検索できるシステム「鳥式改」を作成している。鳥式改ではある物に対するトラブルが単語で表示されており、この単語をキーワードとして Web 検索することも可能としている。しかし、キーワードを含む文がトラブルを表しているとは限らず、現状では検索でヒットした文がトラブルを表しているかどうかはわからない。そこで、本研究では文レベルでトラブルを表す文を抽出することによって、トラブルについてより調べやすい環境を作ることを目指す。

類似した研究としては古瀬ら[4]の研究等、文抽出の研究が挙げられる。古瀬らは肯定、否定、中立的な評価や願望等の意見性を判別し、機械学習を用いて意見文の抽出を試みている。また、評価文書分類[5]にも類似している点がある。トラブルを表す文は否定極性をもつ評価文書を含むため、トラブルを表す文の抽出は否定極性の文書のみを得る場合に近い。ただし、我々は否定極性の評価とトラブルを区別して考えている。評価表現は「悪い」のような主観や、「故障した」等の物を対象とした経験が主である。トラブルはそれに加えて「借金する」といった人の行動に関する表現も広く扱うこととする。

3 手法

3.1 辞書の作成

本実験で使用したトラブルを表す表現の辞書について説明する。

トラブルを表す表現の品詞として、名詞、動詞および形容詞を用いた。De Saeger et al.[1]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し、名詞 20,429 種類、動詞 2,790 種類を辞書として用いた。同様の手法で獲得された形容詞に、Web で公開されている評価表現辞書 [1) 2)] のうち否定極性の形容詞を加え人手で選別したもの 954 種類を用いた。

次に、評価表現辞書 [2) 3)] にあった名詞と形容詞のペア等、トラブルを表すフレーズ 5,909 種類を用いた。さらに、「〜できない」「〜しにくい」等のトラブルを表しやすいパターン 14 種類や「びしょびしょ」「ズタズタ」等のトラブルを表しやすい擬音 110 種類を用いた。

3.2 分類器

本研究では、機械学習の手法として最大エントロピー法 [4)] およびサポートベクトルマシン法 [5)] を用いた。本実験では、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[6]。

また、辞書との単純マッチングを用いた方法として、3.1 で述べた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。トラブルと判定される例文を以下に示す。

例) セキュリティソフトを入れたら同種ソフトの<妨害>によりパソコンの電源が<切れ><なく>なって<困り>ました。

この例では、辞書に記載されている「妨害」「切れる」「ない」「困る」という 4 つの単語が含まれており、決定した閾値 2 以上であるためトラブルと判定された。

4 評価実験

4.1 基本実験

最大エントロピー法、サポートベクトルマシン法および辞書とのマッチングによる方法を用いて Web 文書集合からトラブルを表す文を抽出する実験を行った。Web 文書には、Yahoo!知恵袋 [6)] および検索エンジン基盤 TSUBAKI が提供している大量の Web 文書 [7)](以後、TSUBAKI データと呼ぶ)の原文を使用した。Yahoo!知恵袋は質問回答形式であり、文章の性質として question、

表1 使用した素性

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	文章中の各文の文末文字列
S8	最初の文の文末文字列
S9	最初の文の長さ
S10	最後の文の文末文字列
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
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-gram
S23	マッチした形容詞を含む単語 tri-gram
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

normal answer、best answer の 3 種類がある。このうち比較的トラブルの事例が含まれていた question を使用した。一方で、TSUBAKI データは Web 上のページをクロールしたものであり、実際の Web に近いデータである。

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表1に示す。このうち、文の長さは文字数が 1、2、5、10、20、30、50、70、100、150、200、250、300、350、400、450、500 文字以上の場合にこれらの数字を素性とした。平均単語長は出現した単語の文字数を平均し、0 からその値までの整数をすべて素性とした。文末文字列は文末から数えて 1〜10 文字を全て素性とした。辞書とマッチした数は、0 から辞書とマッチした数以下の整数をすべて素性とした。

表2 オープンテストの結果

	Yahoo!知恵袋			TSUBAKI データ		
	適合率	再現率	F 値	適合率	再現率	F 値
Baseline	0.263	1.000	0.416	0.126	1.000	0.224
DM	0.473	0.806	0.596	0.303	0.690	0.421
ME	0.592	0.840	0.695	0.338	0.619	0.437
SVM1	0.639	0.768	0.698	0.409	0.556	0.471
SVM2	0.633	0.715	0.671	0.386	0.540	0.450

基本実験として、人手でトラブルか否かのタグを付与した Yahoo!知恵袋 1,000 文(うちトラブルを表す文は 281 文)を用いて 10 分割交差検定を行った。同様に、TSUBAKI データ 1,000 文(うちトラブルを表す文は 128 文)を用いて 10 分割交差検定を行った。それらの結果から、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 F 値が最良となる閾値を得た。

本稿では、最大エントロピー法の確率値におけるトラブル分類の境界値と、サポートベクトルマシン法の分離平面との距離におけるトラブル分類の境界値を閾値と呼ぶこととする。

4.2 実験結果

4.1 の基本実験で用いた Yahoo!知恵袋 1,000 文を学習データ、新たにタグ付けした 1,000 文(うちトラブルを表す文は 263 文)をテストデータとしてオープンテストを行った。同様に、基本実験で用いた TSUBAKI データ 1,000 文を学習データ、新たにタグ付けした 1,000 文(うちトラブルを表す文は 126 文)をテストデータとしてオープンテストを行った。

最大エントロピー法(ME)、サポートベクトルマシン法の $d=1$ (SVM1)および 2 (SVM2)、辞書との単純マッチングによる方法(DM)について、トラブルを表す文を抽出した際の適合率、再現率および F 値を表2に示す。最大エントロピー法およびサポートベクトルマシン法の閾値には、基本実験において F 値が最良であった値を使用した。また、出力を全てトラブルと判定した場合をベースラインとした。

オープンテストの結果より、最大エントロピー法やサポートベクトルマシン法による機械学習の手法では、F 値がベースラインを大きく上回った。また、辞書との単純マッチングによる方法よりもよい結果となった。Yahoo!知恵袋、TSUBAKI データともにサポートベクトルマシン法の $d=1$ において最も高い F 値を得た。

5 考察

Yahoo!知恵袋および TSUBAKI データを対象として、サポートベクトルマシン法($d=1$)で実験を行ったときの再現率･適合率曲線を図 1 に示す。図 1 より、Yahoo!知

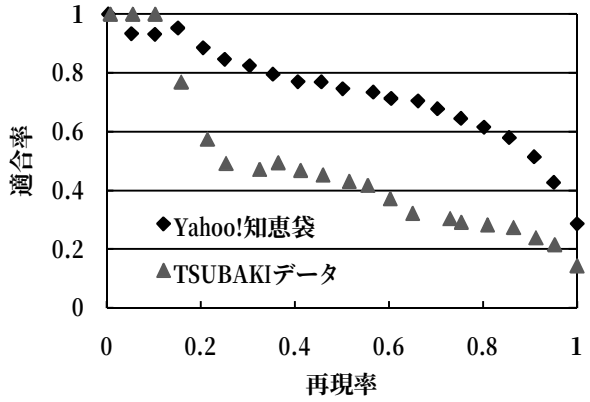

図1 トラブルの分類における再現率･適合率曲線

恵袋を対象としたときは再現率が0.4のとき適合率が0.8程度あることがわかる。また、適合率が 0.5 のとき再現率は 0.9 程度あることがわかる。よって、少量であっても正確に抽出したいときや、多少の間違いが含まれていても網羅的に抽出したいとき等、状況による使い分けが可能であると考えられる。

同様に、TSUBAKI データを対象としたときは再現率 0.2 で適合率が 0.9 程度あることがわかる。TSUBAKI データは文体が多岐に渡るため、Web 上にある膨大なデータに対して効力を発揮する。Yahoo!知恵袋の結果と比較すると再現率が悪いが、Web 全体を対象とできるため、精度よく大量のデータが必要なときに有効だと考えられる。

システムの出力が不正解であった事例について検証する。トラブルでない文をトラブルと判定した事例として、トラブルを表す表現を多用した文があった。例として、「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」という文がある。「紛失」「不正アクセス」等の表現が列挙されたためトラブルと判定されたと考えられる。また、「口の中で消えてしまうような口どけの生地とクリーム」といった文も見られた。「消えてしまう」という表現を誤ってトラブルと受け取ってしまった可能性が高い。これらに対して、「紛失-防止する」等の係り受けを見る必要があると考える。

逆に、トラブルを表す文をトラブルでないと判定した事例には、表現1つのみでトラブルと判断すべき ものがあった。例として、「データが入った e-amusement pass を紛失した模様。」という文があった。「紛失した」という部分が決定的な手がかりであるが、辞書とマッチした回数は 1 回であり、素性としての影響力が薄かった可能性がある。過去形は経験を表し、トラブルを表す動詞の過去形やサ変名詞＋「した」はトラブルを表す傾向が強いと考えられる。よって、過去形に重みを置くような素性を加える等の対策により改善されると考える。また、辞書とマッチしなかった文も見られたため、辞書を拡張することも必要である。

表3 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
・・・	・・・	・・・	・・・

表4 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
S2_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
・・・	・・・	・・・	・・・

最大エントロピー法で求まるα 値を正規化した値(以後、正規化α 値と呼ぶ)を求めた。この値が大きいほどシステムが判定する際に重要な素性であることを示している[7,8]。Yahoo!知恵袋における素性の正規化α 値を表3に示す。同様に、TSUBAKI データにおける素性の正規化α 値を表4に示す。ここで、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、TSUBAKI データともに、辞書とマッチした総数を表す素性がトラブルと判定する上で大きな影響をもつようである。また、「が」「のですが」「のに」等、逆接を表す表現を含む文では「最近PCを買ったのですが、DVD ドライブがついていませんでした」といったようにトラブルを表す場合が多いと推測できる。これらは重要かつ文書に依存しない素性であるといえる。

TSUBAKI データにおいて、「場合」「ほど」といった表現はトラブル、「ます」「たい」といった文末表現はトラブルでないと判定されやすいことがわかる。しかしこれらはYahoo!知恵袋における重要な素性の上位には出てきて

いない。Yahoo!知恵袋においては「わかりません」はトラブル、「？」といった疑問形の文末はトラブルでないと判定されやすいことがわかる。Yahoo!知恵袋では「検索してもわかりませんでした。どなたか教えてください」というような書き込みがあるため、「わかりません」がトラブルと判定する上で重要となる。また、トラブルと関係のない文は「おでんにオススメの具材は何ですか？」といった知恵や知識を問う質問が多いため、「？」はトラブルでないことを表しやすい。これは、質問回答型の掲示板特有の傾向であると考えられる。このように、出現しやすい表現や文末表現は対象とする文書に依存する素性であるといえる。

使用している素性について、素性選択[9]を試した。素性選択では、素性をひとつずつ抜いた状態でテストしてF値が向上した場合、最も上がり幅が大きかった素性を抜いて次の選択を行う。しかし、本研究では選択におけるF値の変動がすべて2%未満であり、かつ使用するデータによってF値が不安定な動きをしたため、抜くべき素性の選択が行えなかった。この要因として、訓練データおよびテストデータが少なかったことや、いくつかの素性が似通っているためにひとつが抜けても相互に補完してしまったことが考えられる。この問題について、訓練データおよびテストデータの拡張や素性の再検討が必要であると考える。

6 今後の展望

今後は、辞書の拡張や素性の検討等によってトラブルを表す文の抽出精度のさらなる向上を目指す。

Yahoo!知恵袋のような記事においては、文章が質問と回答に分離しているという特性がある。この特性を利用して、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が記された文を見つけ出すことを検討している。さらに次のステップではTSUBAKIデータからでも解決策を見つけられるように、トラブルを表す文の前後の文も考慮して解決策が記された文の抽出を行う予定である。

7 おわりに

本研究では、トラブルを表す表現を素性として機械学習を行い、Web文書集合の中からトラブルを表現している文のみを抽出することを目標に実験を行った。2種類のWeb文書を対象に機械学習を行った結果、比較的トラブルの事例が多く含まれるWeb文書ではサポートベクトルマシン法で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同様にF値0.471を得た。今後は、より良い素性の検討等による精度向上や、抽出したトラブルを表す文の周辺文脈から解決策を抽出することを目指す。

参考文献

[1] Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama. Looking for trouble. In Proc. of The 22nd International Conference on Computational Linguistics (Coling2008), 2008.

[2] 鳥澤健太郎. 対象の用途と準備を表す表現の自動獲得. 自然言語処理, Vol.13, No.2, pp.125-144, 2006.

[3] Kentaro Torisawa, Stijn De Saeger, Yasunori Kakizawa, Jun'ichi Kazama, Masaki Murata, Daisuke Noguchi, and Asuka Sumida. TORISHIKI-KAI, an Autogenerated Web Search Directory. In ISUC2008, pp.179-186, 2008.

[4] 古瀬蔵, 廣嶋伸章, 山田節夫, 片岡良治. ブログ記事からの意見文検索. 情報処理学会 自然言語処理研究会, NL176-18, pp.121-128, 2006.

[5] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol.13, No.3, pp.201-242, 2006.

[6] 工藤拓, 松本裕治. Support Vector Machine による日本語係り受け解析. 情報処理学会 自然言語処理研究会, NL138-11, pp.25-32, 2000.

[7] 村田真樹, 内元清貴, 馬青, 井佐原均. 機械学習手法を用いた名詞句の指示性の推定. 自然言語処理, Vol.7, No.1, pp.31-50, 2000.

[8] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 鳥澤健太郎. ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析. 言語処理学会 第15回年次大会 発表論文集, 2009.

[9] 村田真樹, 金丸敏幸, 白土保, 井佐原均. 入力文の格助詞ごとに学習データを分割した機械学習による受身文の能動文への変換における格助詞の変換. システム制御情報学会論文誌, Vol.21, No.6, pp.165-175, 2008.

使用した言語資源およびツール

1) 高村大也. 単語感情極性対応表, http://www.lr.pi.titech.ac.jp/~takamura/index_j.html

2) 小林のぞみ, 乾健太郎, 松本裕治. 評価表現辞書, http://www.syncha.org/evaluative_expressions.html

3) 鍛治伸裕. 評価表現辞書, http://www.tkl.iis.u-tokyo.ac.jp/~kaji/

4) Masao Uchiyama. Maximum Entropy Modeling Package. http://www2.nict.go.jp/x/x161/members/mutiyama/software.html , 2006.

5) Taku Kudo. TinySVM : Support Vector Machines. http://chasen.org/~taku/software/TinySVM/ , 2002.

6) ヤフー株式会社. 「Yahoo!知恵袋-研究機関提供用データ データ仕様書 国立情報学研究所(NII)提供版ver1.0」. 国立情報学研究所, 2007.

7) 新里圭司, 黒橋禎夫. 検索エンジン基盤 TSUBAKI. 2007.