2H4-3

文章構造情報を利用したイベント抽出法

An Event Extraction Method using Sentence Structure

櫻井 茂明[*1] 折原 良平[*1]
Shigeaki Sakurai Ryohei Orihara

[*1](株) 東芝 研究開発センター
Corporate Research & Development Center, Toshiba Corporation

This paper deals with a method that extracts important discussions from bulletin board sites, gives an order to them, and extracts characteristic expressions included in them. Especially, this paper proposes an extraction method of events in order to identify the important discussions, where the events represents the contents of the discussions and are composed of objects, actions, impressions, and so on. The method generates an attribute vector that characterizes an article included in discussions based on results of the parsing. The vector is composed of both attributes that expresses meaning of words and attributes that expresses sentence structure. The paper applies the method to articles collected from bulletin board sites written in English and verifies its effect.

1. はじめに

インターネット環境の普及に伴って、掲示板サイトにより、多数の人々が互いの意見を気軽に交換できる環境が整っている。このような議論の大半は、人々の関心を引く議論ではないものの、企業活動にすら影響を与える大きな問題へと発展する議論も少数ながら存在している。このような議論が起こった場合、関連する企業には多大なる被害が発生するため、大きな問題に発展する前にその問題の存在を認識し、適切な対策を実施することが求められている。

しかしながら、インターネット上には多数の掲示板サイトが存在しており、その中で日々多くの議論がなされているため、すべての議論を詳細に確認することはできない。従って、膨大な議論の中から問題として発展しそうな重要な議論を抽出することが必要である。

本問題に対して、我々のグループでは、イベントと呼ばれる議論の中心となる対象、行動、印象、によって議論を特徴付けることにより、分析者にとって注意する必要のある議論を抽出し、順位付けする方法を提案するとともに、当該議論を特徴付ける表現を抽出する方法を提案した [Sakurai and Orihara 06] [櫻井・折原 07]。

本論文では、より一層の分析性能の向上を目指すために、議論を特徴付けるイベントの抽出に、構文解析技術を導入する方法を検討し、英語掲示板サイトから得られるデータに適用して、その効果を検証する。

2. 掲示板サイトの分析法

本節では、掲示板サイトに内在する多数のスレッドの中から、注意の必要なスレッド (以下注意スレッド) を発見する現象レポートの生成法を説明する。

2.1 対象掲示板サイト

掲示板サイトと一口にいってもその形態には様々なものが存在する。しかしながら、典型的には、各掲示板サイトには、特定の URL 情報が付与された複数のスレッドが含まれており、その中に、時間情報、属性情報、テキスト情報からなる複数の記事が含まれている。本論文では、このような構成を持つ指定されたサイトを分析対象とする。

2.2 現象レポートの生成

現象レポートの生成は、4 種類の抽出処理を掲示板サイトに記述されているスレッドに順次適用することによって行われる。以下においては、各抽出処理を説明する。

記事抽出: 記事抽出は、掲示板サイトに記述されているスレッドからスレッドを構成する記事を抽出し、その記事の中から時間情報、属性情報、テキスト情報を抽出する。このような情報の抽出を行うために、記事抽出においては掲示板サイトごとに手動で作成されるラッパー関数を利用している。

イベント抽出: イベントとは注意スレッドかどうかを判断する上で必要となる、記事に記述されている主体、行動、感情などを代表するものであり、その存在が予め想定可能なものである。対象とするシステムの場合、会社名、機器名、不満といったものがイベントとして定義されている。イベント抽出では、分類モデルに基づいた方法とイベント関連単語に基づいた方法によってイベントを抽出しており、それらの結果を統合することにより、最終的な結果を出力している。

分類モデルに基づいたイベント抽出では、記事の中から抽出したテキスト情報に対して、形態素解析を実施することにより、語尾変化を吸収した語幹及びその品詞を決定する。また、当該テキストに属性を構成する語幹が含まれているかどうかを判定し、テキスト情報に対応する属性値ベクトルを生成する。ただし、分類モデル学習時において、tf-idf 値が指定したしきい値以上になる語幹を属性を構成する語幹とする。加えて、当該属性値ベクトルを SVM(Support Vector Machine)[Vapnik 95] によって学習したイベントごとの分類モデルに適用することにより、イベントを付与するかどうかを判定する。

一方、イベント関連単語に基づいたイベント抽出では、イベントに関連すると考えられる単語やフレーズを予めイベント関連単語として登録する。このイベント関連単語に対して形態素解析を実施し、形態素解析されたテキスト情報に形態素解析されたイベント関連単語が含まれているかどうかを判定することにより、イベントを付与するかどうかを判定する。

スレッド抽出: スレッド抽出は記事から抽出されたイベント抽出結果に基づいて、注意スレッドを抽出し、より注意する必要がある順に出力する。スレッド抽出は、はじめに、付与され

連絡先: 櫻井 茂明, (株) 東芝 研究開発センター, 〒 212-8582
神奈川県川崎市幸区小向東芝町 1, Tel:044-549-2406,
Fax:044-520-1268, E-mail:shigeaki.sakurai@toshiba.co.jp

たイベントのうち、会社イベントクラス (会社に関するイベントをまとめたもの) に含まれるイベントだけに注目し、当該スレッドで主に話題となっている会社を判定する。

また、不満イベントを付与された記事の件数を算出し、予め指定される最小不満イベント頻度以上の件数を含むスレッドを注意スレッドとして抽出する。最終的には、不満イベントの件数の多い順に、注意スレッドを順位付けして出力する。

現象抽出: 現象抽出は特定のスレッドに頻出する一方で、記事全体としては頻出しない表現を特徴的な表現 (以下現象) として抽出する。現象抽出は、予め参照記事の中から抽出品詞系列リストに含まれる品詞列と一致する表現を抽出し、その頻度を計算する。次に、入力された注意スレッドに対して、その中から抽出品詞系列リストに含まれる品詞列と一致する表現を抽出し、その頻度を計算する。この頻度を参照記事における当該表現の頻度と比較することにより、条件を満たす表現だけを抽出する。最終的には、特徴的な表現でないと予め分かっている表現を格納している不要語辞書を参照することにより、不要語辞書に格納されていない表現だけを現象として抽出する。

2.3 現象レポート

前節までに説明した抽出法に基づいて抽出したイベント、順位、現象をスレッドごとに出力することにより現象レポートを出力する。提案する現象レポートの生成法に基づいた現在のシステムの場合、不満イベントに加えて、会社イベントクラスに含まれるイベント、機器イベントクラス (機器に関するイベントをまとめたもの) に含まれるイベントといったイベントを扱っている。ただし、会社イベントクラスや機器イベントクラスに含まれるイベントに関しては、イベント関連単語を利用するだけで十分なイベントの再現率を得ることができるため、分類モデルを利用したイベントの抽出は現在のところ行っていない。図 1 は、現象レポートの一例を示しており、各イベントに対して最大 3 個の現象が抽出されている。

スレッド名 件数 機器名 件数 件数 現象名 件数

順位	スレッド	機器	不満総件数	現象1	現象2	現象3
1	Title A/40	Model A/20	30	blue screen/15	noise/12	CD-ROM/7
2	Title B/35	Model A/15	28	shutdown/12	heat/10	
3	Title C/30	Model B/15	25	performance/18	weight/11	
4	Title D/24	Model C/15	20	panel/16	brightness/12	stripe/10
5	Title E/20	Model B/15	15	memory/10	overheat/10	noise/9

図 1: 現象レポート

3. 構文解析の適用

構文解析は単語間の係り受け関係を解析することができ、形態素解析よりもテキストに関する多くの情報を得ることができる。このため、構文解析情報を利用することにより、現象レポートの分析性能の向上を図ることが期待できる。そこで、本節では、構文解析法を簡便に説明するとともに、構文解析情報を利用した分類モデルの学習法を検討する。

3.1 構文解析

一口に構文解析といっても多くの手法が提案されている。本論文では、論文 [平川・天野 89] [中里 他 86] に記載されている構文解析を実施することにより、入力されたテキストの一文に対応する構文木を生成する。本構文解析では、はじめに、テキストの一文に対して、単語及びその属性を記述してある辞書を参照することにより、与えられたテキストの一文を文節単位にまとめて、単語及び当該単語に関連する属性が付随している列に分解する。次に、体言及び体言に準ずる語の他の語に対する関係を示している格助詞などの表層的な特徴を利用することにより、名詞とその名詞に最も近い動詞の間に係り受け関係を設定し、単純な構文木を生成する。次に、係り受け可能な名詞、動詞の組を、当該の単純な構文木の中から取り出して、当該の組み合わせに対して、必要ならば複数の係り受け候補を生成する。このとき、当該係り受け候補に対しては、意味係り受け関係を記述してある辞書を参照することにより、設定可能な意味係り受け関係及びその確信度に基づいてスコアを計算して付与する。最終的には、文節間の意味的な制約を考慮した上で、係り受け候補群の中から最適な係り受け候補を決定し、最もスコアが高くなる構文木を生成する。また、各文節を展開して、各ノードが単語からなる構文木を生成する。

例として、「先日コンピュータを買いましたが、その処理速度に満足していません。」を構文解析して得られる構文木を図 2 に示す。図の構文木においては、各ノードに見出し語 (文節を代表する単語) が表示されており、見出し語間がリンクによって関連付けられている。また、見出し語に含まれる単語の属性の一部が見出し語の右側にラベルとして付与されており、見出し語間の意味係り受け関係がリンク上に付与されている。

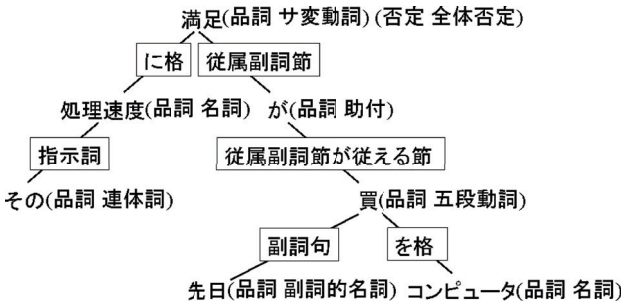

図 2: 構文木の例

3.2 属性値ベクトルの生成

従来の分類モデルに基づいたイベント抽出では、テキスト情報を形態素解析して得られる語幹のうち、その評価値である tf-idf 値が指定したしきい値以上となる語幹を選択し、属性値ベクトルを構成していた。このような属性値ベクトルでは、「処理速度に満足しているが、デザインに満足していない」、「デザインに満足しているが、処理速度に満足していない」といった、構成される単語は同一だが、意味の異なる文章を区別することができない。このような文章を区別するには、単語間の係り受け関係を解析する必要がある。このため、構文解析情報を利用することにより、より文章の意味を反映した属性値ベクトルを構成することが期待できる。

一方、構文解析結果は木構造として与えられるため、属性値ベクトルを構成するには何らかの工夫が必要となる。論文 [鹿島 他 06] に提案されている方法では、構文木を一般化したラベル付き順序木に対して、ラベル付き順序木向けのカーネル関数を設計している。本カーネル関数の場合、ラベル付き順序木に含まれる部分構造を基準として、属性値ベクトルを構成することができるため、構造情報をうまく取り込んだ特徴付けを行うことができる。しかしながら、ラベル付き順序木に含まれる部分構造を抽出する際には、ラベル間の多数の組み合わせを考慮する必要があるため、属性値ベクトルの計算量が大きくな

るといった問題があった。そこで、本論文では、構文解析結果の情報をより簡便に取り込んだ属性値ベクトルの構成法を検討する。

見出し語に付随する属性は見出し語の意味を規定する上で重要な属性 (以下意味属性) である。通常見出し語には多数の意味属性が付与されているが、その意味属性の重要度は同じというわけではない。文章の意味を大きく左右するものからちょっとしたニュアンスの違いを表現するものなど様々である。これに対して、すべての意味属性を利用した場合、属性値ベクトルが高次元になるため、学習器による過学習が発生する危険性が増大する。そこで、見出し語に付随する意味属性の中から重要と思われる意味属性を選択し、当該意味属性によって見出し語を特徴付けることにする。また、見出し語は、構文木において特定の位置に位置付けられており、構造的な情報も保持している。構文木に含まれるあらゆる部分構造を考慮するのは、計算負荷が高いとしても、全く構造情報を利用しないのでは、もとのテキストの意味を大きく変えてしまう危険性がある。そこで、見出し語の構文木における木の深さ、下位に連結する枝の数、見出し語に連結する枝に付与されているラベルなど、比較的容易に決定できる情報 (以下構造属性) によって、見出し語を特徴付けることにより、構文木の構造情報を部分的に反映した属性値ベクトルを構成することとする。これにより、従来の語幹の有り、無しに基づいた属性値ベクトルよりも元のテキストの意味を反映した属性値ベクトルを構成することが期待できる。

すなわち、構文解析結果を反映した属性値ベクトルとして、図 1 に示すような属性値ベクトルを構成することができる。図の例の場合、各見出し語が、構造属性「深さ」と意味属性「否定」及び「態」によって特徴付けられており、各見出し語の属性を連結することにより属性値ベクトルが構成されている。ただし、構文木の最上位のノードの深さを 0 とする。また、深さ「-1」は対応する見出し語がテキスト内に含まれていないことを表すとする。

表 1: 構文解析結果を反映した属性値ベクトル

satisfy			computer			···
深さ	否定	態	深さ	否定	態	···
0	全体否定	受動	-1	無し	無し	···

以上のように属性値ベクトルを構成することにより、構文木に対応する属性値ベクトルを構成することができるものの、通常、テキストは複数の文章から構成されている。構文解析においては、ひとつの文章に対してひとつの構文木を構成しているため、ひとつのテキストに対応する属性値ベクトルを構成するには、複数の構文木からひとつの属性値ベクトルを構成する必要がある。一方、重要な文章はより上位に記述される可能性が高いと考えられるため、同一の見出し語が複数回出現した場合には、対応する意味属性及び構造属性としては、最初に発見された見出し語のものを優先することにする。また、同一見出し語における矛盾した内容の記述に対応する見出し語の抽出を避けるため、同一の見出し語が発見された段階で、当該の見出し語の下位に付随する見出し語を削除することにする。このような削除を実施することにより、「A 社の製品に満足しているが、B 社の製品に満足していない」といった文章が与えられている場合に、2 回目に出現する「満足」に対応する「A 社」、「製品」(構文木においては、B 社 ··· の部分が上位に登場する) に関する情報を削除する効果を期待することができる。

次に、見出し語の深さに注目してみると、構文木の上位に位置する見出し語は下位に位置する見出し語よりも重要度が高いと考えられる。従って、見出し語選択時において、上位に位置する見出し語を重要視することにより、より妥当な見出し語に高い評価値を与えることが期待できる。そこで、式 (1) によって深さによって補正した tf-idf 値を計算することにする。本式においては、前半の大括弧の部分が見出し語の深さの逆数の平均値を表しており、後半の大括弧の部分が tf-idf 値を表している。

$$ev(w_i) = \{\sum_{l=1}^{l=t_{ij}} \sum_{j=1}^{j=D} \frac{1}{(depth(i,j,l)+1)\cdot t_{ij}\cdot d_i}\} \cdot\{\frac{1}{D}\cdot \log_2(\frac{D}{d_i})\cdot \sum_j \frac{t_{ij}}{n_j}\} \quad (1)$$

ただし、w_i を i 番目の見出し語、D をテキストの総数、d_i を i 番目の見出し語をもつテキストの数、n_j を j 番目のテキストに含まれる見出し語の数、t_{ij} を j 番目のテキストに含まれる i 番目の見出し語の数、$depth(i,j,l)$ を j 番目のテキストに含まれる i 番目の見出し語が l 回目に出現した場合における構文木の深さとする。

4. 数値実験

4.1 実験方法

3 つの英語掲示板サイトから収集した 10,002 件の記事データを実験データとして利用する。各記事データに対しては、当該記事に不満を含む内容が記載されているかどうかがイベントとしてラベル付けされており、記事データから不満イベントを抽出するための分類モデルを学習する。具体的には、実験データに対して、5 cross-validation に基づいた実験を実施する。この 5 回の実験によって得られる再現率と適合率の平均値を比較することにより、イベント抽出性能に対する効果を検証する。

実験では、構文解析の効果を検証するために、形態素解析に基づいて生成された属性値ベクトルと、構文解析に基づいて生成された属性値ベクトルを用いた比較実験を行う。このとき、tf-idf 値に対応するしきい値としては、数種類のしきい値を適用した場合に、良好な結果を示した 0.0001 を利用する。本しきい値の場合、3,678 個の語幹によって形態素解析に対応する属性値ベクトルが構成されている。そこで、構文解析に基づいた属性値ベクトルにおいても、3,678 個の見出し語に基づいて属性値ベクトルを構成することにする。また、構文解析に基づいた方法の場合、見出し語の選択基準として tf-idf 値を利用した方法と見出し語の深さ情報を加味した方法 (tf-idf 改)、構造属性として見出し語の深さ情報を利用した方法と利用しない方法、意味属性として否定 (全否定、部分否定)、態 (受動、能動)、相 (進行形、完了形、完了進行形、近未来) を利用した方法と利用しない方法、を比較することにする。

4.2 実験結果

実験結果を表 2 及び表 3 に示す。表においては、左側に条件が記載されており、条件に対応する再現率及び適合率がその右側に順に記載されている。

4.3 考察

いずれの条件においても構文解析に基づいたイベント抽出性能は、形態素解析に基づいたイベント抽出性能よりも劣っており、期待した効果を得ることはできなかった。tf-idf+深さ未

表 2: 構文解析の効果 1

		再現率	適合率
形態素解析		0.662	0.599
tf-idf 改	深さ利用	0.480	0.384
	深さ未利用	0.503	0.457
tf-idf	深さ未利用	0.505	0.461

表 3: 構文解析の効果 2

		再現率	適合率
tf-idf 改	「否定」	0.508	0.456
+深さ未利用	「態」	0.502	0.453
	「相」	0.500	0.447
	「否定」+「態」	0.493	0.456
	「否定」+「相」	0.486	0.460
	「態」+「相」	0.484	0.439
tf-idf	「否定」	0.503	0.450
+深さ未利用	「態」	0.506	0.447
	「相」	0.506	0.455
	「否定」+「態」	0.500	0.452
	「否定」+「相」	0.490	0.447
	「態」+「相」	0.497	0.441

利用の場合、構文解析における属性値ベクトルと形態素解析における属性値ベクトルとの違いは、基本的には見出し語か単語かの違いである。このため、この違いにより、著しくイベント抽出性能が劣化しているということは、構文木を構成する見出し語が劣化の原因と考えられる。記事データの場合、引用記号が頻繁に文章内に挿入されており、この引用記号によって文書が正しく一文に切り出されていない可能性がある。このため、妥当な見出し語が取り出されなかったことが原因のひとつと考えられる。また、形態素解析に基づいた方法では、語幹の同一の単語に関しては、品詞の違いに関わらず同一の単語とみなしているのに対して、見出し語の場合には、このような見出し語の集約が行われていない。この集約が行われていないことも性能劣化の原因と考えられる。

次に、見出し語の選択基準、構造属性、意味属性の影響を見てみることにする。見出し語の深さを構造情報として利用した場合と、深さ情報を利用せずに当該の見出し語が記事中に存在すれば 1、存在しなければ 0 とした場合とを比較してみると、深さ情報を利用した方がイベント抽出性能が劣化している。見出し語の位置的情報を加味した分類モデルを学習した方が、イベント抽出性能が向上すると期待していたが、そのような効果を観測することはできなかった。同一の見出し語も深さによって分割することにより、過学習が起きてしまったことがその原因と考えられる。

また、見出し語の選択基準に関しては、他の実験データに関しては、深さ情報を加味した tf-idf の方が若干良いイベント抽出性能を記録したものの、本論文に記載している実験データでは、従来の tf-idf の方が若干良いイベント抽出性能を記録している。このため、どちらの選択基準が良いかは一概にはいえない。より構文木の上位にある見出し語が重要と考えて tf-idf 値を補正してみたが、必ずしも上位にある見出し語が分類モデルに寄与しているとはいえないようである。

一方、今回の実験においては、多数付与される意味属性のうち、意味的に重要と思われる 3 つの意味属性を属性値ベクトルを構成する意味属性として採用している。このうち、「否定」に関する意味属性に関しては、若干イベント抽出性能の向上を観測できたものの、「態」や「相」に関しては、その効果を観測することはできなかった。特定の意味属性を持つ見出し語の数は、相対的には少ないため、意味属性の効果があまりでなかったものと考えられる。

上記のように、現状では、構文解析情報に基づいた方法は必ずしも良好な結果を得られていない。しかしながら、構文解析により得られる情報は、文の意味的な違いを評価する上では有用であると考えられるため、構文解析情報の利用を今後再検討していく予定である。

5. まとめと今後の課題

本論文では、構文解析情報を簡易に組み込んで属性値ベクトルを構成する方法を提案し、その効果を英語掲示板サイトから収集した記事データに対して適用し、その効果を検証した。構文解析の結果与えられる意味属性のひとつである「否定」に関しては、若干の性能向上の可能性が確認されたものの、全体的には、従来の形態素解析に基づいた手法よりもイベント抽出性能が劣化しており、構文解析情報の利用方法を再検討する必要がある。

今後の課題としては、構文解析情報の利用方法の見直しに加えて、他の手法に基づいたイベント抽出法の性能改善を検討する予定である。例えば、ブースティングやバギングなどのアンサンブル学習の効果があるとの別の実験結果も得られており、具体的な導入方法を検討していく予定である。一方、システム利用部門からは、掲示板サイトを限定せずに評判情報を分析したいとのニーズも寄せられているため、この方面での分析方法も検討していく予定である。

参考文献

[Hsu et al. 86] Hsu, C. -W., Chang, C. -C., and Lin, C. -J.: A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~ cjlin/libsvm/.

[平川・天野 89] 平川 秀樹, 天野 真家: 構文/意味優先規則による日本語解析, 人工知能学会 第 3 回全国大会, 8-4, 363-366 (1989).

[鹿島 他 06] 鹿島 久嗣, 坂本 比呂志, 小柳 光生: 木構造データに対するカーネル関数の設計と解析, 人工知能学会論文誌, 21, 1, 113-121 (2006).

[中里 他 86] 中里 茂美, 堤 義直, 平川 秀樹, 天野 真家: 日英機械翻訳システムにおける日本語解析について (2), 情報処理学会 第 32 回全国大会, 3S-8, 1611-1612 (1986).

[Sakurai and Orihara 06] Sakurai, S. and Orihara, R.: Discovery of Important Threads from Bulletin Board Sites, Int. J. of Information Technology and Intelligent Computing, 1, 1, 217-228 (2006).

[櫻井・折原 07] 櫻井 茂明, 折原 良平: 掲示板サイト分析における重要議論抽出と特徴表現抽出, 日本知能情報ファジィ学会誌, 19, 1, 13-21 (2007).

[Vapnik 95] Vapnik, V. N. : The Nature of Statistical Learning Theory, Springer, (1995).