様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２０ 日現在

機関番号：１３３０２
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１９５０

研究課題名（和文）オピニオンマイニングのための製品レビューからの暗黙的属性の抽出

研究課題名（英文）Extraction of Implicit Aspects for Opinion Mining on User Reviews

研究代表者
白井　清昭（Shirai, Kiyoaki）

北陸先端科学技術大学院大学・先端科学技術研究科・准教授

研究者番号：３０３０２９７０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：レビューにおいて，ユーザが製品の属性(例.モバイルフォンのbatteryやpriceなど)に対して意見を書くとき，明示的な単語を使わずに，暗に属性に対する意見を述べることがある．本研究ではこのような属性をレビュー文における「暗黙的属性」と呼び，これを自動的に抽出する．まず，大量のレビュー文から明示的属性の抽出を試み，明示的属性を含む文と含まない文を得る．次に，後者は何らかの暗黙的属性を含む文とみなし，類似度が高くかつ明示的属性を含む別の文を探索することで，その文の暗黙的属性をラベル付けする．最後に，暗黙的属性がラベル付けされた文のデータから暗黙的属性抽出モデルを深層学習により獲得する．

研究成果の学術的意義や社会的意義
従来の属性抽出に関する研究の多くは明示的属性を対象としていたのに対し，本研究では暗黙的属性を抽出の対象とする点に特徴がある．近年の自然言語処理技術は深層学習による手法が主流だが，最新の深層学習モデルを適用するために，暗黙的属性がラベル付けされたデータセットを自動的に構築する点に学術的意義がある．
ユーザレビューを分析し，製品やサービスに対する評判を明らかにするオピニオンマイニングは，ユーザにとって有益な情報をもたらす技術である．本研究の成果により，明示的属性だけではなく暗黙的属性を分析することが可能になり，オピニオンマイニングの精緻化が促進されるため，その社会的意義は大きい．

研究成果の概要（英文）：When users express their opinion toward an aspect of a product (e.g. “battery” or “price” of a mobile phone) in a review, they sometimes write opinions implicitly without using explicit words. In this study, such an implicitly expressed aspect is called “implicit aspect”. We propose a method to extract implicit aspects automatically. First, by extracting explicit aspects from a large amount of reviews, sentences containing or not containing explicit aspects are obtained. The latter is regarded as an implicit sentence, which may includes a certain implicit aspect. Next, the implicit aspect of the implicit sentence is determined by searching the similar sentence that includes the explicit aspect and propagating that explicit aspect as the implicit aspect of the sentence. Finally, a model to extract implicit aspects is trained by deep learning using the labeled sentences as training data.

研究分野：自然言語処理

キーワード：オピニオンマイニング　属性抽出　暗黙的属性

１．研究開始当初の背景

オピニオンマイニングは，ユーザによって書かれた大量のレビューを解析し，対象物(モバイルフォン，パソコン，デジカメなど)に対する世間の評判を明らかにする技術である．オピニオンマイニングで解決するべき主要な課題として，「属性抽出」と「極性判定」の 2 つが挙げられる．属性抽出とは，与えられたレビュー文に対し，意見が表明されている対象となる製品の属性(スマートフォンの「バッテリ」「操作性」「デザイン」など)を抽出(検出)するタスクである．一方，極性判定とは，レビュー文が属性に対して表明している意見が「肯定」「否定」「中立」のいずれであるかを判定するタスクである．ここでは前者の属性抽出に着目する．

一般に，評価対象となる属性は単語または名詞句で表わされる．例えば，“The battery of the phone lasts the whole day. ”というレビュー文において，意見が表明されている製品の属性は“battery”である．このように文中に明示されている属性を「明示的属性」と呼ぶ．一方，ユーザは属性を明示せずに暗黙に意見を表明することも多い．例えば，“Surprisingly, this phone lasts two days.”という文はバッテリに対する意見を表明しているが，“battery”という単語自体は文中に明示されていない．このように明示されていない属性を「暗黙的属性」と呼ぶ．

従来の属性抽出に関する研究の多くは，明示的属性を抽出することに焦点を当てており，暗黙的属性を抽出する研究はほとんど行われていない．一方，オピニオンマイニングでは，多くのレビュー文に対して属性を抽出し，それに対するユーザの意見の極性を判定することで，製品の評判を把握することが求められる．明示的属性を含むレビュー文のみをオピニオンマイニングの対象とすると，暗黙的に表明されている意見を見過ごすことになり，世間の評判を正確に把握できない可能性がある．したがって，オピニオンマイニングにおいて，暗黙的属性の抽出は重要な課題である．

暗黙的属性の抽出に関する先行研究は多くはないが，これらの研究に共通しているのは，属性と関連の深い単語，特に感情語(good, bad など極性を示唆する単語)のリストをあらかじめ作成し，そのような単語や感情語が文中に出現したとき，それに関連する属性を(文中になくても)抽出するという点である．このような手法は，基本的には単語を手がかりに暗黙的属性を抽出しており，比較的浅いレベルの自然言語解析に基づく手法である．一方，近年では，深層学習を自然言語処理に応用する研究が盛んに行われており，優れた成果を挙げている．ところが，深層学習の手法を基に暗黙的属性を抽出する研究はこれまで行われていなかった．

２．研究の目的

本研究は，製品に関するレビュー文が与えられたとき，それから暗黙的属性を抽出する手法を確立することを目的とする．1.で述べたように，明示的属性ではなく暗黙的属性を抽出の対象としている点に独自性がある．

暗黙的属性を抽出する際，先行研究のような浅い自然言語処理に基づく手法ではなく，深層学習による深いレベルの自然言語理解手法に基づく手法を提案する．深層学習には大量のラベル付きデータが必要である．深層学習に適用できるほど大規模かつ正解の暗黙的属性がラベル付けされたレビュー文のデータセットは存在しない．本研究は，このようなデータセットを自動構築し，深層学習を暗黙的属性の抽出に適用する．

本研究では，暗黙的属性の抽出を分類問題として解く．まず，評価対象の製品が持つ代表的な属性を「暗黙的属性クラス」としてあらかじめ定義しておく．例えば，評価対象がスマートフォンのとき，price, design, battery などを暗黙的属性クラスとする．レビュー文が与えられたとき，その文によって暗に意見が表明されている属性が，暗黙的属性クラスのいずれであるかを分類する．なお，本研究における対象言語は英語とするが，提案手法は一部を除いて言語に依存しないため，他の言語に適用することも可能である．

３．研究の方法

提案手法の概要を図 1 に示す．本研究は[A]～[D]の 4 つのステップで構成される．

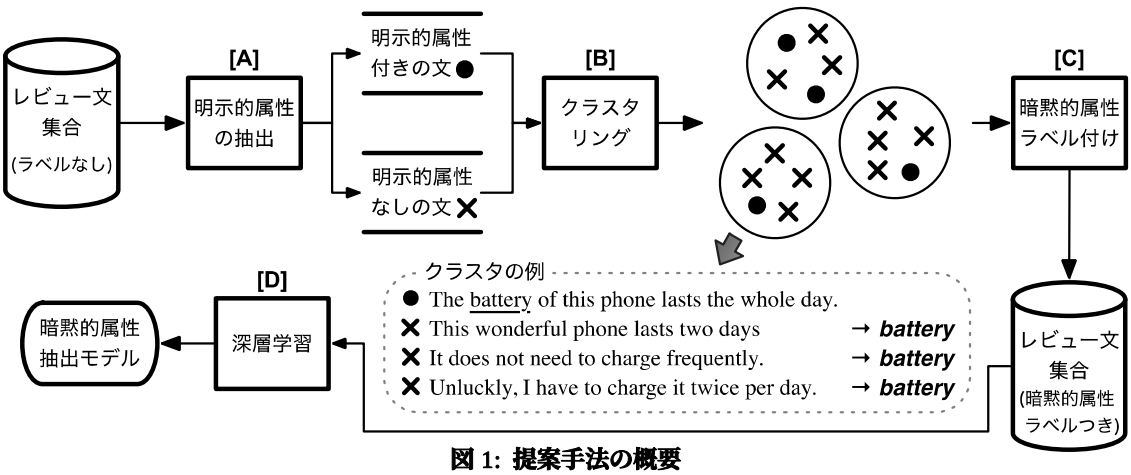

図 1: 提案手法の概要

ステップ[A]では，明示的属性を抽出する．具体的には，レビュー文から明示的属性を抽出するモデルを機械学習する．明示的属性がラベル付けされた既存のデータセットを使用し，これを訓練データとして，明示的属性を抽出するモデルを得る．学習アルゴリズムとして Conditional Random Fields (CRF)を用いる．属性抽出の標準的な手法に従い，CRF の学習素性として，対象単語の出現形や品詞，対象単語の前後 3 単語の出現形や品詞，対象単語の前の属性ラベルを使用する．次に，属性がラベル付けされていないレビュー文の集合を用意し，学習した CRF モデルを用いて，レビュー文から明示的属性を抽出する．これにより，レビュー文集合を，明示的属性が付与された文(●で示された文)と付与されていない文(×で示された文)に分ける．

ステップ[B]では，レビュー文のクラスタリングを行う．ここでの目的は，明示的か暗黙的かに関わらず，同じ属性に対して意見を表明している文をまとめたクラスタを作成することである．クラスタリングのために，個々のレビュー文をベクトルで表現する．文のベクトル表現を得る方法には様々なものがあるが，ここでは sparse composite document vector (SCDV) を採用する．クラスタリングアルゴリズムとしては，大量のレビュー文をクラスタリングの対象とするため，比較的計算量の小さい k-means を用いる．k-means 法ではクラスタ数をあらかじめ設定する必要があるが，1 つのクラスタが 20 文程度になるようにクラスタ数を決める．すなわち，Purity を高くする(異なる属性に言及した文が 1 つのクラスタにまとめられる誤りを少なくする)ため，比較的小さいクラスタを数多く生成する．

ステップ[C]では，暗黙的属性のラベル付けを行う．まず，得られたクラスタのラベルを決定する．ここでのクラスタのラベルとは，クラスタ内のレビュー文が意見を表明している属性とする．クラスタ内に明示的属性が付与された文が含まれるとき，その明示的属性をクラスタラベルとする．異なる明示的属性が付与された文が混在するとき，多数決によってクラスタラベルを決定する．なお，明示的属性が付与された文が含まれないクラスタについては，後続の処理を行わない．図 1 の「クラスタの例」では，battery を明示的属性として持つ文が 1 つ存在するため，クラスタラベルを battery とする．次に，同じクラスタに含まれ，かつ明示的属性が付与されていない文に対し，クラスタラベルを暗黙的属性としてラベル付けし，またラベル付けされた文を暗黙的属性が付与された文として収集する．図 1 の例では 3 つの文が battery を暗黙的属性とする文として収集される．また，より多くの文を収集するために，暗黙的属性クラスの同義語リストを用意し，これをクラスタラベルとする文も収集する．例えば，battery については，charger，power などの同義語リストを用意し，これをクラスタラベルとするクラスタからも battery を暗黙的属性とするレビュー文を収集する．

ステップ[D]では，深層学習によって暗黙的属性抽出モデルを学習する．分類モデルとして BERT(Bidirectional Encoder Representations from Transformers)を採用する．前のステップで構築した暗黙的属性がラベル付けされたコーパスを用いて，事前学習済みの BERT モデルのファインチューニングを行い，暗黙的属性を抽出するモデルを得る．

４．研究成果

モバイルフォンとパソコンの 2 つのジャンルに関するレビューを対象に暗黙的属性を抽出するシステムを構築した．モバイルフォンについては battery, case, design, size, screen, price の 6 つを，パソコンについては screen, software, windows, interface, price の 5 つを暗黙的属性クラスとして定義した．

まず，明示的属性を抽出するモデルを CRF により学習した．学習データとして SemEval-2014 のラップトップ・ドメインの属性ラベル付きデータを用いた．CRF モデルによる明示的属性抽出の精度，再現率，F 値はそれぞれ 0.77，0.64，0.70 であった．

次に，Amazon Review Data に含まれる 10,000 件のレビュー文に対してクラスタリングを行った．その後，クラスタラベルの決定，クラスタ内のレビュー文に対するクラスタラベルの付与の手続きを経て，暗黙的属性がラベル付けされたレビュー文を収集した．モバイルフォンとパソコンのそれぞれについて，およそ 1600 件程度の暗黙的属性ラベル付きレビュー文を得た．この一部をサンプリングし，付与された暗黙的属性が正しいかを人手によって評価した．その結果，暗黙的属性付与の正解率は，モバイルフォンについては，battery: 0.82，case: 0.74，design: 0.58，size: 0.14，screen: 0.76，price: 0.78，パソコンについては，screen: 0.70，software: 0.64，windows: 0.72，interface: 0.62，price: 0.56 といった結果が得られた．モバイルフォンにおける size など一部の頻度が少ない暗黙的属性については正解率が低かったものの，全体的にはある程度高い正解率が得られた．

上記の手続きで構築した暗黙的属性がラベル付けされたコーパスを用いて，事前学習済み BERT モデルをファインチューニングし，暗黙的属性を判定するモデルを学習した．提案手法の評価のため，(1)明示的属性のラベル付きデータのみを用いるモデル，(2)暗黙的属性のラベル付きデータのみを用いるモデル，(3)両者を併用したモデルを比較した．モバイルフォンについては，6 つの暗黙的属性の分類の F 値のマクロ平均は，モデル(1)では 0.058 と非常に低かったのに対し，モデル(2)では 0.52，モデル(3)では 0.60 となった．本研究で構築した暗黙的属性のラベル付きデータを用いることで分類の F 値が大きく向上することを確認した．パソコンについても，暗黙的属性のラベル付きデータを BERT のファインチューニングに用いることで，5 つの暗黙的属性の分類の F 値のマクロ平均が大幅に向上することを確認した．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Aye Aye Mar, Kiyoaki Shirai
２．発表標題 Automatic Construction of Annotated Corpus with Implicit Aspect
３．学会等名 The Thirteenth Edition of the Language Resources and Evaluation Conference, pp. 6985-6991（国際学会）
４．発表年 2022年

１．発表者名 中村 駆, 白井 清昭
２．発表標題 評判情報分析のための製品属性タグ付きコーパスの半自動構築
３．学会等名 情報処理学会第84回全国大会，第2分冊，pp.35-36
４．発表年 2022年

１．発表者名 曽田颯人，白井清昭
２．発表標題 商品レビューの複数の観点からの有用性の評価
３．学会等名 言語処理学会第27回年次大会，pp.518-522
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関