

Title	複数の製品を対比する文の自動検出
Author(s)	花岡, 裕
Citation	
Issue Date	2014-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/12039
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士

概要

近年、ユーザレビューを対象とした評判分析は製品の購入を検討するユーザに有益な情報を提供する手段として注目されている。本研究ではユーザによるレビュー文のうち、特に複数の製品を対比した文に着目する。例えば「ipadはnexus7より重い」といった文のようにnexus7とipadを対比している文は、タブレット端末の購入を検討しているユーザにとって有益な情報となる。本論文は、複数の製品を対比した文(以下、製品対比文と呼ぶ)を検出することを目的とする。

本論文の提案手法の流れを以下に示す。まず、ユーザが評判を知りたい対象製品(P_t とする)についてレビュー文を取得する。次に、対象製品 P_t の競合製品リストを作成する。次に、製品対比文の検出を行う。最後に、製品対比文であると判定した文を出力する。

まず、対象製品のレビュー文を取得する。“ P_t レビュー”をクエリとして検索エンジンでウェブ検索を行う。得られたウェブページをHTMLタグ及び句点で分割したものをレビュー文の集合とする。

次に P_t と競合する製品のリストを作成する。競合製品とは、 P_t と同じ種類の製品で、購入を検討する際に比較の対象となりうる製品とする。ここではウェブサイト「価格.com」における P_t を含む製品カテゴリの製品リストをそのまま競合製品リストとする。

次に製品対比文の判定を行う。ここでは、与えられたレビュー文が製品対比文であるか否かを判定する2つの手法を提案した。1つはルールベースの手法、もう1つは機械学習に基づく手法である。

ルールベースの手法では二つのルールを設計した。一つは製品名が複数出現する文を製品対比文と判定するルールAである。ここで製品名は対象製品 P_t もしくは競合製品リスト中の製品とする。これは、製品名が2つ以上現れる文においては、それらの製品が対比されているという考えに基づく。ただし、“iPadもNexus7も十分高性能”のように、製品名が複数現れた場合でも、出現した製品を同列に扱って評価を行っているときなど、それらの製品を対比しているとはみなされない場合がある。そこで、レビュー文を節に分割し、2つ以上の節に異なる製品名が出現しなければならないとした。

もう一つのルールは、人手作成した対比表現辞書を基に対比文の判定を行うルールBである。対比表現辞書は、対比文に特徴的に現れる表現を対比表現とし、これを収集した辞書である。ルールBは、文中に辞書に登録された対比表現が出現したとき、それを製品対比文と判定する。ただし、文中に対比表現が出現するだけでは必ずしも製品対比文であるとは限らない。そこで、対比表現について、それが製品対比文に出現するための付加条件を設定した。

機械学習に基づいた手法では、対比文がタグ付けされた製品レビュー文の集合を訓練データとし、与えられた文が製品対比文か否かを判定する二値分類器を教師あり学習する。本論文で採用した学習アルゴリズムはSVMである。製品対比文の判定に有用であると考えられる以下の9つの学習素性を設計した。(1)文中の自立語。(2)対比表現辞書内の対比

表現の有無. (3) 評価表現. 評価を表す語が出現する文は製品対比文の可能性が高い. (4) ルール A. ルールベースの手法におけるルール A(複数の製品名を含む) を満たすか否か. (5) 製品名の数 (0,1,2 以上). 製品名を多く含む文ほど製品対比文の可能性が高い. (6) 用言文節の数 (0,1,2 以上). 用言文節とは, ここでは用言(動詞, 形容詞など)を主辞とする文節とする. 用言文節を多く含む長い文ほど製品対比文の可能性が高い. (7) 評価表現と製品名が共に出現するか否か. (8) 製品名の係り先用言と (9) 製品名の間接的係り先用言. いずれも製品名と係り受け関係にある用言に特徴的な単語が現れるため, 素性とした.

提案手法の評価実験を行った. 対象製品として 11 個の製品を選び, 上述の手法でレビュー文の集合を得た. これらの文に対し, 製品対比文か否かを人手で判定し, 評価用データとした. 評価指数として, 正解率, 精度, 再現率及び F 値の 4 つを用いた. なお, 本論文で提案するルールベースの手法および機械学習に基づく手法における学習素性は, $P_t=Nexus7$ のデータを精査し, 設計した. したがって, Nexus7 に対する評価はクロズドテストである.

まずルールベースの手法を評価した. ここではルール A, ルール B, 両者のいずれかの条件を満たすとき製品対比文と判定するルール A+B の 3 つを比較した. クロズドテストである $P_t=Nexus7$ では, F 値の値はルール A+B(0.401), ルール A(0.236), ルール B(0.176) の順に高かった. ただし, それ以外の製品を対象にしたオープンテストでは, これらのルールの F 値の順序が製品カテゴリによって変化し, 一概にどのルールが有効であるかを判定することは難しい. ルール A について, 製品カテゴリによって F 値が大きく変化する理由として, 製品名の記述の仕方が製品カテゴリによって異なり, そのため製品名の検出が正しく行えなかったことがあげられる. 一方ルール B については, 製品カテゴリによってクロズドテストとオープンテストで大きな差が無かったことから, 様々な製品カテゴリに適用可能なルールとして対比表現辞書は有効であるといえる.

次に機械学習による手法を評価した. F 値は異なる製品カテゴリのレビュー文を訓練データとした場合 (0.171) と同じ製品カテゴリのレビュー文を 5 分割交差して訓練データとした場合 (0.369) で大きく差が出たことから, 製品対比文の特徴は製品カテゴリによって異なることがわかった. また, 同じ製品カテゴリのレビュー文を訓練データとしたときの SVM の F 値が 0.369 に対し, ルール A+B の F 値が 0.180 であることから, ルールベースの手法よりも機械学習に基づく手法の方が優れている. これは, 機械学習によって製品カテゴリの違いによる対比文の特徴をある程度自動的に学習できるためと考えられる.

全体的に, 製品対比文判定の正解率が高いが, 精度, 再現率, F 値は低かった. これは, 製品対比文の全体に占める割合が約 3% と低く, 製品対比文の正例がデータ内に多く出現しないことが原因と考えられる.

次に, 個々の学習素性の有効性を調べるために, 素性集合から素性を 1 つ除いて SVM を学習し, 全素性を用いて学習した SVM と比較した. この実験では, 訓練データは同じ製品カテゴリの文とし, 5 分割交差検定を行った. 製品対比文の判定に最も有効であった素性は自立語であったが, この素性は他の素性に比べて数が多いことから, 製品対比文の判定に最も大きく貢献することは自然である. F 値を基準にすれば, 自立語の次に有効な

のは、対比表現、ルール A であり、それ以外の素性については大差が無かった。これらの素性が自立語を除く他の素性と比べて有効であることは、ルールベースの手法で採用されたルールがある程度妥当であることを示唆する。同様に、製品カテゴリ毎に素性の有効性も調べた。自立語が一番有効であることは全ての製品で共通しているが、それ以外の素性については製品カテゴリによってその有効性にばらつきがあり、製品対比文判定の F 値の向上に貢献しない製品カテゴリも存在した。このことから、製品対比文をより正確に判定するためには今回提案したものとは異なる素性が必要である。