応用

3.11 レビュー解析

―誤り分析におけるプロセスとプロダクト―

藤井 敦（東京工業大学）　乾 孝司（筑波大学）

レビュー解析とは

レビューとは物事の価値を論じた記述である．身近な出来事，新しい商品，国の政策について感じたこと，思ったこと，考えたことである．レビュー解析とは，雑多な文書群からレビューを集めて，誰が何に対してどんな価値を見出したのかを読み解く処理である．一つひとつはツブヤキやボヤキでも大量に集めて解析すれば，あるホテルに対する評判やその理由が分かるかもしれない．そこに泊まるかどうか迷っている旅行者やサービス向上を目指す当ホテルの支配人には有益な情報である．

技術の紹介

レビュー解析の単位はさまざまである．ここでは文を解析の単位とする．「私の友人は従業員の態度に腹を立てた.」と「私の友人は砂のお城に旗を立てた.」は何が違うのか？　レビュー解析は入力文から「意見保持者＝私の友人，対象＝従業員，属性＝態度，評価＝腹を立てた」といったレビューの構成要素を抽出して，評価を「肯定／否定」のような極性や「☆☆☆」のような数値で表現する．

レビュー以前に普通の文を解析する方法を考えよう．分かりやすい例として，隠れマルコフモデルは英語が苦手な生徒に似ている．英文法がよく分からずに，文を名詞や動詞といった品詞の並びとしか認識していない．教科書の英文をたくさん読んで，「品詞Ａと Ｂの連接」や「品詞Ｃと単語Ｄの対応」が起こる確率を暗記する．英文を読むときは，品詞を数珠繋ぎにして先頭の品詞から順番に対応する単語を出力したときに，自分が読んでいる英文と同じ単語列になる品詞列の候補から確率が最大の品詞列を選択する．

たとえば,「The（冠詞）price（名詞）of（前置詞）this（形容詞）car（名詞）is（動詞）reasonable（形容詞）.（記号）」のように各単語の品詞が推定される．ここで，レビューの構成要素である属性や評価極性を特殊な品詞として混ぜると，「The price（属性）of this car（対象）is reasonable（肯定）.」のように該当する情報を特定できるようになる．

Next NLP におけるタスク

誤り分析の対象は，極性分類を拡張してレビュー文の評価を肯定，否定，中立のいずれかに分類する処理である．使用したレビュー文は筑波大学文単位評価極性タグ付きコーパス（TSUKUBA コーパス）の一部であり，文数は肯定 1,379，否定 639，中立 682 である．全データは楽天データ公開から入手可能である☆1．サポートベクターマシンを用いて 10 分割交差検定を行った．単語の出現のみを分類の特徴量として用いた単純な手法を分析の対象とすることで，誤り分析自体の難しさと手法の複雑さによって生じる分析の難しさを分離した．

プロセスとプロダクト

誤り分析の成果物には Process（どうやるのか）と Product（何を得たのか）という 2 つの P がある．誤り分析の観点が目的や分析者によって異なる問題を解消するためには両方の P が重要である．「どうやるのか」は，誤りの原因を究明して手法を改善することを目的とし，誤りの事例を分析しながら原因を類型化した．1 つの誤りに複数の原因が該当する場合はすべてを列挙した．

☆1　http://rit.rakuten.co.jp/opendataj.html

大分類	中分類	小分類	具体例
狙いが外れた	正解の極性に対する支持が不十分	評価表現	肯定：「おいしい」 否定：「今ひとつ」
		表記ゆれ	肯定：「有難い」や「有り難い」は代表表記でない
		未知語	否定：「バサバサ」
		誤記	肯定：「気に入る」を「気に入れる」と誤記
		定型句	肯定：「気を利かす」 否定：「〜してほしい」
		特殊記号	肯定：「◎」 否定：「...」
		修辞疑問	否定：「〜があっても良いのではないでしょうか？」
		学習データ	疎問題やデータ偏向
		特徴語なし	中立に多い
	不正解を支持	類出語	肯定：「とても」 否定：「ただ」
想定していない	参照表現		「バス、トイレなしの予約でしたが、両方ついたお部屋」の「両方」が「バス、トイレ」を指す
	文間関係		全体的に肯定か否定に傾倒：極性の継続（「特筆すべきは」）や反転（「しかし」）
	領域知識		肯定：「3 回目の宿泊」はリピーターを示唆
	比較		否定：「料金の割にせまい」
	仮定		否定：「露天風呂があれば良かった」

表 -1　評価分類タスクに関する誤り事例の分類体系

「何を得たのか」については，誤りの原因をまず「本来の狙いが外れた」と「当該手法が想定していない事象」に分けた．前者を「正解の極性に対する支持が不十分だった」と「不正解を支持してしまった」に分けて，さらに個別の手法に関する事項に分けた．肯定と否定が相互に誤分類される場合は，正解に特徴的な単語の不足もしくは不正解に特徴的な単語の過剰が原因であるのに対して，中立が関与する場合はそもそも中立に特徴的な単語が少ないため誤りの傾向が異なった．「想定していない事象」は解決が見込まれる手法に細分した．結果的に，**表 -1** に示すような三階層の分類体系が作成された．「本来の狙いが外れた」には，評価表現や定型句の特定に起因する根本的な誤り，表記ゆれや誤記に起因するレビューに特有の誤り，「〜があっても良いのではないか？」といった修辞疑問による否定の強調を認識できない誤りがあった．「想定していない事象」には，参照表現や文間の関係といった談話に関する誤り，領域知識の欠如に起因する誤り，比較や仮定といった文の構造に起因する誤りがあった．

展望

誤り分析の結果（Product）を踏まえて，Process を共有するための分析作業マニュアルについて議論する．マニュアルの要素としてチュートリアル，リファレンス，トラブルシューティング，用語集を考える．チュートリアルは，教科書のように通読や演習を通して体系的な基礎知識を与える素材である．リファレンスは，誤り分析の最中に見つけた特定の事例をきっかけとして，さらに深く分析するための素材である．チュートリアルが最初から通読することを前提としているのに対して，リファレンスは索引のように特定の語句による逆引きを可能とする．トラブルシューティングは，先人が経験した誤り分析の「落とし穴」と脱出方法に関する事例を提供する．用語集は，マニュアルに出現する用語の解説である．こうした取り組みは学生の研究指導や若手研究者の育成といった教育目的の利用にも意義がある．

(2015 年 10 月 1 日受付)

藤井 敦（正会員）fujii@cs.titech.ac.jp
1998 年東京工業大学大学院博士課程修了．現在，同大学院情報理工学研究科准教授，博士（工学）．自然言語処理等の研究に従事．

乾 孝司（正会員）inui@cs.tsukuba.ac.jp
2004 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．日本学術振興会特別研究員等を経て，2009 年筑波大学大学院システム情報工学研究科助教．2015 年同准教授．現在に至る．博士（工学）．自然言語処理の研究に従事．