4D-6

ブートストラッピング手法を用いた句表現の自動抽出

野村 雄司　　末永 高志　　高木 徹
株式会社 NTT データ 技術開発本部

1. はじめに

企業では，自社製品のより深い分析や業務効率化のために，消費者から直接寄せられる情報や，web 上の様々な情報を活用したいというニーズが高まっている．一般的な評判や感情の分析による製品の良し悪しの傾向把握だけでなく，製品の不具合や，消費者の苦情，要望，感謝といった具体的な情報を抽出し，改善活動や注意喚起につなげたいと考えている．

これを実現するためには，目的ごとに意味のある単位で情報を抽出する必要がある．「メールを着信した際，音が鳴らない」といった文があったとき，事象の発生条件となる「メールを着信した際」と，不具合事象となる「音が鳴らない」をそれぞれの句を対にした句表現として抽出することで，抽出した結果の背景を含めた理解が可能となる．

一般のキーワードに基づく分析では，キーワードとして「音」や「鳴らない」を個別に抽出することになる．しかし，これから重要な情報か判断をすることは難しく，句表現を抽出する方式の検討が必要となる．

これに対して，抽出元の情報は多種多様であり，同一企業内でも部門や業務によって記載方法が一定でないことが多い．また，web では特に個人ごとに記載方法が異なる．そのため，抽出元データごとに，抽出ルールを人手で作成して処理することは現実的ではない．

そこで，部門や業務といったドメインに依存せずに，少量の教師データを用いて，任意の句表現を抽出する手法を提案し，実データにより有効性を確認したので，報告する．

2. 関連研究

少量の教師データから語彙を獲得する従来手法として，ブートストラッピング手法[1,2]がある．この手法では，抽出対象の語（以下，インスタンスと呼ぶ）と抽出パターン（以下，パターンと呼ぶ）を交互に繰り返し自動抽出し，少数の正解インスタンスから大量のインスタンスを得る手法である．

現在，有効な方式として，Espresso[1]が知られている．これは，交互に繰り返し抽出していく過程で，多義性のあるインスタンスが抽出されると無関係なインスタンスに遷移する意味ドラフトと呼ばれる現象を抑制する機能を有する．具体的には，自己相互情報量を用いた信頼度スコア関数の導入により，実現している．

ブートストラッピングは，一般的には名詞で構成される語や，1～2 語に限定された評判や感情を表す語を抽出する用途で利用されるが，同様のロジックで句表現の抽出にも利用可能である．

なお，句表現の抽出手法として，坂地らの Cross-Bootstrapping による課題・効果表現対の抽出[3]がある．しかし，特許文書を対象としているため，抽出する表現は「名詞＋格助詞＋動詞」で構成されていることを前提としている．そのため，名詞句や「フリーズする」など助詞を伴わない表現の抽出はできない問題がある．

3. 句表現抽出方式

ドメインに依存せずに任意の情報を少ない作業コストで抽出するために，Espresso をベースとした句表現抽出を提案する．ここで，インスタンスとパターンの関係の定義した抽出テンプレートと信頼度スコア関数を検討が必要となる．

本検討では抽出テンプレートとして，”[head] 〈instance1〉 [mid] 〈instance2〉 [tail]”を用いた．[head]，[mid]，[tail]は，それぞれに位置するパターンを表す．また，[head]と[tail]に関しては最大で 2 句とした．

さらに，句表現の抽出では，名詞のみで構成される語や，語数が限定された語を抽出する場合と違い，パターンに一致するインスタンス候補が大量に抽出され，多くの誤りを含むと予想される．そこで，抽出されたインスタンス候補と，そのインスタンス候補を抽出したパターンが生成される元となった正解インスタンスを構成する品詞の一致度，および語尾の表層形，品詞の一致度を Espresso の信頼度スコアに反映させる方式を提案する．

Extracting Phrase Expressions via a Bootstrapping Method
Yuji Nomura, Takashi Suenaga and Toru Takaki
Research and Development Headquarters,
NTT Data Corporation

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

表 1　実験データ

データ種別	文書数（文数）		
	学習データ	評価データ	全データ
アップデート情報	4 (339)	10 (886)	47 (3,882)
投稿記事	6 (186)	44 (1331)	3,473 (109,760)

表 2 データセットごとのタグ付与数

データ種別	タグ種別	学習データ	評価データ
アップデート情報	条件	24	89
	症状	29	103
投稿記事	条件	14	72
	症状	20	94

4. 評価実験

本評価では，製品をスマートフォンとし，その故障や不具合情報を分析，検索することを想定する．不具合が発生する「条件」と「症状」の抽出精度により有効性を確認した．実験データには，消費者の声を基に企業内に蓄積された文書を想定した通信キャリアの公式ホームページ上の製品アップデート情報[*1] と，スマートフォンに関するユーザ投稿サイト[*2] の，ドメインの異なる 2 種類のデータを用いた．不具合の条件および症状の表現部分に人手で正解タグを付与したデータのうち，初期に与える正解データと評価用データを表１のように無作為に分割した．また，各データセットに付与したタグ数は表２の通りである．

抽出テンプレートに本手法と同一のものを用い，Espresso の信頼度スコアだけを利用したものをベースラインとして，本手法との抽出精度の比較を行った．抽出精度の結果は，表３，表４の通りである．

5. 考察

アップデート情報，投稿記事ともに，ベースラインと比較して F 値が向上している．「タッチパネルが動作しなくなる」といった表現に加えて，「電源 ON 時」のような名詞句や「通信できない」のような助詞を伴わない表現も抽出できていることを確認した．

また，投稿記事では，アップデート情報と比べても適合率，再現率ともに低い精度となっている．これは，実験データ中には抽出対象となる表現，およびその周りの表現が一度しか出現しないものが多く，信頼度評価で正解インスタンスと不正解インスタンスに差が生じなかったためである．また，ベースラインと比較しても精度の向上が小さかった要因として，形態素解析誤りが生じ，品詞の構成の一致度による信頼度評価が正しく機能しなかったことが挙げられる．

*1　http://www.nttdocomo.co.jp/support/utilization/product_update/
*2　http://sp.oshiete.goo.ne.jp/

表 3　アップデート情報における抽出精度（%）

抽出方法	適合率	再現率	F 値
ベースライン	66	73	69
提案手法	92	85	89

表 4　投稿記事における抽出精度（%）

抽出方法	適合率	再現率	F 値
ベースライン	16	42	24
提案手法	22	39	28

6. まとめと今後の課題

Espresso をベースとするブートストラッピング手法を拡張することで，文書ドメインや抽出対象とする表現に依存しない少量の教師データを用いた方式を提案した．製品アップデート情報のように，比較的定型表現で記述された文書に対しては，適切に抽出することができた．これはコールセンタなど企業に寄せられた声をある程度決まった形式で業務担当者が記述した文書などには適用可能と考えられる．

一方で，web 掲示板のように消費者が投稿した記事では，ベースラインと比べてもわずかな向上しか見られなかった． これに対して，今後は，文書中に一度しか存在しない表現に対応するために，インスタンスからパターンを生成する際に，不要な表現を検出することの検証が必要である．

なお，本手法では対となる句表現を抽出することを目的としたが，抽出対象の表現が文ごとに分かれて記述されている場合にも抽出できるよう，句表現を構成する要素を対とするインスタンスとして抽出する方法の検討も課題の一つである．

参考文献

[1] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, In Proc. of COLING-ACL, pp.113-120(2006)
[2] Thelen, M. and Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern context, In Proc. of EMNLP, pp.214-221(2002)
[3] 坂地泰紀，野中尋史，酒井浩之，増山繁: 特許文書からのブートストラップ手法を用いた課題・効果表現対の抽出，情報処理学会研究報告，Vol.2009-NL-192，No.14，pp.1-8(2009)

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.