E-020

分割誤りに頑健な新語のカテゴリ分類

Category Classification of New Words Robust to Segmentation Errors

山田 達史† 松本 和幸† 吉田 稔† 北 研二†
Tatsushi Yamada Kazuyuki Matsumoto Minoru Yoshida Kenji Kita

1．はじめに

近年 Twitter などインターネット上の書き込み回数が飛躍的に増加している．tweet には，流行の言葉が多く含まれるため，既存のシソーラスで直接カテゴリ分類をすることは困難である．また，単語の文脈に基づく既存の単語意味解析の手法では形態素解析による分かち書き処理が前提となる．たとえば新語である“大阪都構想”という単語は形態素解析では“大阪／都／構想”と分割されるため解析することはできない．本研究では，シソーラスで分類されていない語を新語として定義し，分割誤りとなる新語でもカテゴリを付与できる手法を提案する．

2．関連研究

インターネット上の膨大なテキストデータを利用し，有用な情報を得るための研究は盛んにおこなわれている[1,2]．単語の共起関係を利用し特徴ベクトルを生成することを目的とした研究[3]や，ニュースなどのテキストデータから単語の共起ベクトルを生成し，それらをカテゴリ分類する研究[4]，単語の共起関係から単語を概念化して概念ベースを構築する研究[5]がおこなわれている．これらの研究では，分割誤りされる可能性のある新語へ対処については述べられていない．

3. 提案手法

提案手法では，以下の流れで新語のカテゴリを分類する．また，表 1 はテキストの例を出してカテゴリ分類までの流れを説明している．

Step1 形態素解析による分割誤りする新語を抽出する
Step2 新語周辺単語の出現頻度をカウント
Step3 周辺単語のベクトル生成
Step4 類似語を利用したカテゴリ分類

3.1 形態素解析による比較データ生成

図 1 では既存の辞書で分割誤りした単語を右側，比較データ作成のため形態素解析に成功するようにしたものを左側に記載している．既存の辞書で形態素解析すると，図 1 の右側のような分割誤りが起こる新語が存在する．本稿では，分割誤りした新語を対象に，類似語の検索によるカテゴリ判定の比較をし，問題点を探る．

3.2 周辺単語の抽出による類似語検索

分割誤りすると，新語を単語単位でカテゴリ分類することができないため，周辺単語を用いてカテゴリ分類することを考える．新語の分割誤りしたテキストデータに対し，新語の文字列以外の単語の出現頻度をカウントする．出現頻度が高く，重要度の高い単語を周辺単語とし，それらをキーとして類似語検索をおこなう．

† 徳島大学，Tokushima University

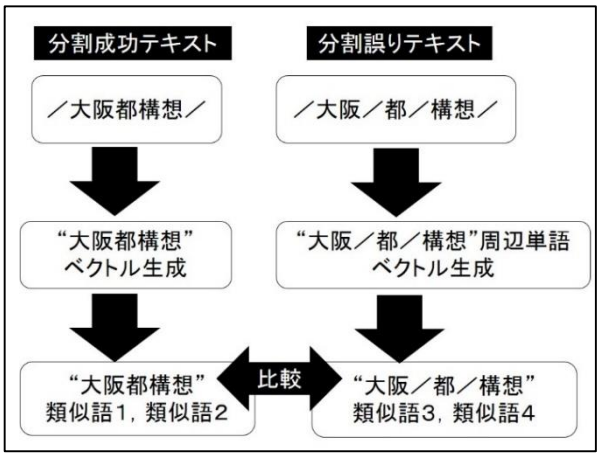

図 1 分割され方の違いによる類似語の比較

3.3 類似語を用いたカテゴリ分類

出現頻度の高かった周辺単語に対して，ベクトルを用いた類似語を抽出する．これらの類似語をシソーラスのカテゴリと紐付け，もっとも多いカテゴリを分割誤りした新語のカテゴリとする．分割誤りした新語のカテゴリと分割成功した新語のカテゴリを比較し，同じだった場合は正解とする．

表 1. テキストの例：“妖怪ウォッチのアニメみたよ”

	分割成功テキスト	分割誤りテキスト
形態素解析結果	妖怪ウォッチ／の／アニメ／み／た／よ	妖怪／ウォッチ／の／アニメ／み／た／よ
ベクトル化する単語	妖怪ウォッチ	アニメ
類似語	進撃の巨人，鋼の錬金術師	ドラえもん，タッチ
カテゴリ	アニメ	アニメ

4．実験に用いるデータとツール

本研究では，テキストコーパスとして Twitter[6]から収集した tweet コーパスを用いる．このコーパスは，2015 年 1 月～2015 年 4 月までの 4 ヵ月間において TwitterAPI により，シソーラスに登録されている単語をランダムでキーワードとして選択して検索クエリとすることで収集し，リツイートや短縮 URL などのノイズを除去した計 32,234,154 文からなる．形態素解析器には，MeCab ver.0.996[7]を用い，新語を含むテキストの形態素解析に適した辞書として，新語が定期的に自動更新されている mecab-ipadic-neologd[8]を用いる．類似語の検索のために，word2vec[9]を用いて，分かち書きテキストから単語ベクトルの生成をおこなう．実験においては，word2vec での学習時，特徴抽出のための窓幅を 10，ベクトルの次元数を 200 と設定した．

5. 予備実験

予備実験として分割誤りする単語に対しての類似語検索実験をおこなった．検索には，word2vec の検索用ツールである w2v-distance を用いる．表 2 に，実験で用いた，標準の形態素解析辞書(IPA 辞書)で分割誤りする新語データを記載する．

表 2　分割誤りする新語

分割誤りする新語	mecab-ipadic-neologd	カテゴリ
／大阪／都／構想	／大阪都／構想／	政治
／妖怪／ウォッチ／	／妖怪ウォッチ／	アニメ
／こち／亀／	／こち亀／	アニメ
／進撃／の／巨人	／進撃の巨人／	アニメ
／鋼／の／錬金術／師	／鋼の錬金術師／	アニメ
／きゃ／り／ー／ぱみゅぱみゅ／	／きゃりーぱみゅぱみゅ／	アーティスト
／もも／クロ／	／ももクロ／	アーティスト

表 3 に，表 2 で示した「分割誤りする新語」を検索クエリとして入力した場合，表 4 に，mecab-ipadic-neologd により分かち書きしたコーパスを用いて検索した場合のそれぞれの類似語の例と，正解率を示す．検索結果の上位 10 件のうち表 2 のカテゴリと一致した場合を正解とする．出力された正解の数を A とし，正解率を$P_{Category}$とすると，以下の式により正解率を表すことができる．

$$P_{Category} = \frac{A}{10} \times 100\ [\%]$$

「分割誤りする新語」に対しては，分割位置にスペースを挿入し，“妖怪 ウォッチ”のように検索することで，双方の単語ベクトルが合成されることにより，好ましい結果が得られる場合と，そうでない場合とがあった．

表 3　IPA 辞書で分割誤りされたまま検索した結果

分割誤りする新語（2 語の組合せ）	類似語の例	$P_{Category}$[%]
大阪＋都	京都	0
大阪＋構想	橋下	50
都＋構想	維新	60
進撃＋の	巨人	10
進撃＋巨人	エレン	40
妖怪＋ウォッチ	ウォッチメダル	80
こち＋亀	該当なし	0
きゃ＋り	び	0
もも＋クロ	杏果	30

表 4　mecab-ipadic-neologd により分割された単語の組合せを用いた結果

mecab-ipadic-neologd による組合せ	類似語の例	$P_{Category}$[%]
大阪都	住民投票	100
構想	二重行政	90
大阪都＋構想	住民投票	90
妖怪ウォッチ	妖怪ウォッチ 2	70
こち亀	バクマン	100
進撃の巨人	アルミンクリスタ	80
鋼の錬金術師	該当なし	0
きゃりーぱみゅぱみゅ	該当なし	0
ももクロ	ももいろクローバー Z	70

表 3，表 4 より，分割に誤りがない場合は比較的正確なカテゴリ判定がおこなえることがわかる．また，「こち亀」や「きゃりーぱみゅぱみゅ」といった，ひらがなで構成されている語や先頭または末尾にひらがなを含んだ語の場合，前後の助詞などによってコーパスにおける分割位置が一定でなくなる可能性があるため，単語単体で分割された単語の組み合わせを利用してもカテゴリ判定をすることが困難であった．また，mecab-ipadic-neologd を用いることにより，最新の単語に対しても対応できることが分かった．しかし，登録漏れが無いわけではないため，登録されていない単語の場合にでも検索ができる仕組みが必要であることもわかった．

6．おわりに

提案手法では，対象となる新語をコーパスから検索し，周辺単語を取得し，それらの類似語を用いてカテゴリ分類する．予備実験の結果，辞書に未登録の単語でも，分割された単語の単語ベクトルを組み合わせることで，カテゴリ分類できる可能性があった．細かく分割され過ぎる場合や，ひらがなで構成される単語の場合において，検索が困難であることが分かった．また，mecab-ipa-neologd を用いることで，最新の単語に対応できるが，最長一致での分割の副作用があるため，必ずしも最適な結果が得られるとは限らない．

今後は，周辺単語に基づく類似語検索の実験をおこない評価する予定である．また，予備実験において問題点として明らかとなった，単語単位での分割位置と，文中の分割位置とで差異がある場合への対処方法を検討したい．また，分割位置に依存せずに周辺単語を抽出しカテゴリ分類する手法を実装し，評価実験をおこなう予定である．

謝辞

本研究の一部は，科学研究費補助金（基盤研究（C)15K00425，若手研究（B)15K16077）の補助を受けて行った．

参考文献

[1]. 渡邊恵太，加藤昇平，“ユーザ興味を反映した情報推薦のための潜在的ディリクレ配分法を用いた協調フィルタリング”，信学技報．NLC, 言語理解とコミュニケーション，2014-01-30，113，429，pp. 15-20.
[2]. 近藤光正，中辻真，田中明通，“Wikipedia に基づく Web 閲覧履歴からの潜在的興味キーワード抽出”，信学論(D), Vol. 96, No. 5, pp. 1199-1211, 2013.
[3]. 福元伸也，渕田孝康，“単語の共起関係を利用した概念的特徴ベクトルの生成”，DEIM Forum 2015 B4-4.
[4]. 尾脇 拓朗，福元伸也，“単語の意味を考慮した共起ベクトルによるテキスト分類”，DEIM Forum 2014 C6-2.
[5]. 別所克人，内山俊郎，内山匡，“全単語間共起を考慮した概念ベース生成手法”，信学技報，IE2014-41, PRMU2010-29, MI2010-29(2010-5).
[6]. Twitter: https://twitter.com/
[7]. 日本語形態素解析器 MeCab ver.0.996: http://taku910.github.io/mecab/
[8]. mecab-ipa-neologd: https://github.com/neologd/mecab-ipadic-neologd
[9]. Word2Vec: https://code.google.com/p/word2vec/