1Q-08

テキストに含まれる未知語の推定手法の提案

越智 雅人　　黒田 久泰

愛媛大学大学院理工学研究科

1. はじめに

インターネットが普及し，ニュースやブログなどからユーザは多くの情報を得ている．また，逆に Twitter や SNS などでは，ユーザが日夜書き込みを行い，膨大なテキストが増え続けている．それらのなかには，ユーザにとって分からない単語（略語などの未知語）を含むことも珍しくない．このような未知語が存在すると，ユーザは Web で未知語を検索して調べないといけないため煩わしい．また，商品などの推薦を行う場合，レビュー文を用いることもあるが，そのとき未知語を含んだレビュー文が存在することも多々ある．本研究では，そのような未知語がどのようなものであるのかを推定する手法を提案する．

2. 関連研究

牛久保らの研究[1]では，Twitter 上の未知語の意味推定方式について述べている．この研究では，コーパスに Yahoo!blog 検索 API を用いて，未知語を含むブログを使用している．そして，取得した文章に対して形態素解析を行い，名詞，形容詞，動詞だけを抽出し，20 のカテゴリに分類する．そのとき，未知語と同じカテゴリになった単語を同類語とみなし，その同類語からもコーパスを取得することで，コーパスの量を補っている．また，単純な名詞の複合語は 1 つの名詞として扱っている．最後に，コーパスから取得した形容詞，動詞の最頻出語を未知語を表す語として出力する．この研究の結果として，「ピングドラム」という未知語は動詞として「見る」が出力されている．「ピングドラム」はアニメであるため正しいと言える．しかし，形容詞は取得できない．他の結果として，動詞は間違っており，形容詞は取得できないものがほとんどであった．考察によると，形容詞は未知語，同類語を含む文の中にあまり含まれてい

Proposal of a method for estimating unknown words in the text

Masato OCHI, Hisayasu KURODA

Graduate School of Science and Engineering, Ehime University

なかったと考えられると述べている．動詞については，ブログは個人の日記のようなもののため，未知語や同類語に関する動詞より，感想などを伝えるための動詞が多く出現したのではないかと述べている．

この手法では，最後は単純に動詞，形容詞の最頻出に頼っている．しかし，単純に最頻出を使用した場合，未知語を説明していない動詞，形容詞が選ばれる場合があるため問題である．そのため，単純な形態素解析ではなく，係り受け解析などを用いる必要があると考えられる．また，カテゴリ分類にこの研究ではシンプソン係数を用いている．しかし，単純にシンプソン係数を用いた場合，調べる 2 つの単語の検索結果の差が極端に大きいと，誤った結果を示す場合があるため，改善の余地がある．

3. 提案手法

本研究では，未知語を名詞に限定し，ユーザが入力した未知語に関係する名詞，形容詞，動詞をアソシエーション分析を用いて提示する手法を提案する．提案手法の流れを図 1 に示す．

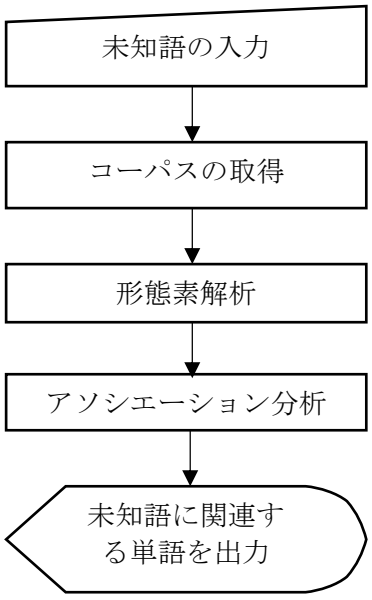

図 1　提案手法のフローチャート

3.1 コーパスの取得

TwitterAPI を用いて，ユーザが入力した未知語をキーワードとして含むツイートをランダムに 600 件取得する．

3.2 形態素解析

得られたコーパスに対して形態素解析を行い，各ツイートから名詞，形容詞，動詞だけを抽出する．形態素解析器には MeCab を用いた．その際，牛久保らの研究と同様，連続して出現する名詞は 1 つの要素として抽出する．また，多くのツイートには URL を含み，形態素解析器はそれを名詞として認識してしまうため，例外として URL は除外した．動詞などは活用形があり，アソシエーション分析を行う上で，同じ動詞を別の要素として分析してしまうのは問題があるため原形を用いた．

3.3 アソシエーション分析

アソシエーション分析とは，買い物をする際の商品の組み合わせから，関連性などを抽出する場合によく用いられる手法である．「商品 A を買う人は商品 B もよく買う」のようなルールがあった場合，それを「{A}=>{B}」のように表す．本研究ではこれを応用し，商品を単語に置き換えて考える．評価指標として支持度，確信度，リフトなどがある．支持度は，ある文章の中に，単語 A と単語 B が同時に出現する確率のことである．確信度とは，ある文章の中に単語 A が存在し，そのうえで単語 B が出現する確率（条件付き確率）のことである．アソシエーション分析には R 言語を用いた．

アソシエーション分析によって得られた結果から，未知語を含むルールのみを抽出し，ユーザに提示する．

4. 実験

実験として，表 1 に示す 2 個の未知語を想定し，本手法を用いた．この表には，各未知語から得られたルールの総数と，未知語をキーワードとして含むルール数も併せて載せておく．また表 2 と表 3 に各未知語に関する関連語を示す．ただし，ルールの左辺，もしくは右辺に複数の単語があるルールより，どちらも 1 つの単語からなるルールを優先している．

表 1　未知語の種類とルール数

未知語	ルール総数	キーワードを含んだルール数
イクメン	173	104
終活	5159	2849

表 2　未知語「イクメン」に対する結果

関連語	品詞	支持度	確信度
サイズ	名詞	0.065	0.983
がんばる	動詞	0.068	0.984
小さい	形容詞	0.067	0.968
仕事	名詞	0.081	0.986
働く	動詞	0.045	0.976
子育て	名詞	0.090	0.880
やる	動詞	0.044	0.833
てる	動詞	0.106	0.800

表 3　未知語「終活」に対する結果

関連語	品詞	支持度	確信度
終末	名詞	0.044	1.000
ひとりさま	名詞	0.044	1.000
過ごす	動詞	0.046	1.000
コラム相続 遺言書終活 川越市	名詞	0.043	1.000
遺言書	名詞	0.051	1.000
着	名詞	0.121	0.892
旅立ち	名詞	0.121	0.892
人生ラスト	名詞	0.121	0.892
吟味	名詞	0.121	0.892
テーマ	名詞	0.123	0.893

5. 考察

「イクメン」では「仕事」，「働く」などイクメンを表す名詞が取得できている．また，「終活」でも「終末」，「遺言書」，「人生ラスト」など終活を表す名詞が多く取得できている．しかし，「サイズ」や「吟味」などよく意味の分からない語も含まれている．

動詞については，イクメンでは「働く」などの意味のある単語がある一方，「やる」，「てる」など意味のない単語も含まれている．これは「出現する」の「する」のように名詞＋動詞の形のものが多くコーパスに多く含まれていたと考えられる．終活では，意味のある「過ごす」という動詞が 1 つだけ取得できた．

形容詞は，両者ともほとんど取得することができなかった．これは，Twitter 自体に形容詞を含むツイートが少ない傾向にあるのか，もしくは取得したツイートが少ないことが原因として挙げられる．

参考文献

[1]牛久保 祐樹，藤田 茂：“Twitter 上の未知語の意味推定方式”，平成 23 年度情報処理学会関西支部支部大会論文集，2011.