4D-03

マイクロブログからの典型的使用場面付き辞書の構築

岡　利成[*]　　白井　清昭[†]

北陸先端科学技術大学院大学 先端科学技術研究科

1 はじめに

本論文は，単語の典型的使用場面の情報が付加された辞書をマイクロブログから自動的に構築する手法について述べる．本研究では，典型的使用場面を以下に述べる3種類と定義する．

時間 単語が特定の時間によく使われるとき，その時間．【朝】【昼】【夕方】【夜】【深夜】の5つと定義する．例えば,「おはよう」の典型的使用場面は【朝】である．

場所 単語が特定の場所でよく使われるとき，その場所．場所のカテゴリは47都道府県とする．例えば,「はいさい」の典型的使用場面は【沖縄県】である．

職業 単語が特定に職業の人によってよく使われるとき，その職業．職業カテゴリの詳細は後述する．例えば,「注射」の典型的使用場面は【医師】である．

典型的使用場面の情報が付与された辞書は，様々な自然言語処理応用システムで利用できる．例えば，雑談対話システムにおいて，ユーザが発話した時間や場所，ユーザの職業を推定し，それに応じて適切な応答を返すことができるようになる．

関連研究として，ウェブ検索エンジンの検索オプションを活用して特定の時間や場所に関連した文を検索し，時間や空間に依存した応答を生成する対話システムの研究 [1] や，緯度・経度の情報を人手で付与したコーパスを構築する研究 [3] などがある．本研究は，マイクロブログ (Twitter) を辞書構築のための情報源とする点，特定のアプリケーションを想定せずに汎用的な知識として辞書を構築する点，時間・場所に加えて職業も典型的使用場面の対象とする点に特色がある [4]．

2 提案手法

2.1 ツイートの収集

Twitter から，時間，場所，職業のメタデータが付与されたツイートを収集する．

2.1.1 場所のツイートの収集

既に述べたように，本研究における場所のカテゴリは都道府県である．Twitter API (Tweepy[1]) を用いて，「place:〈地名コード〉」を検索キーとして，地名コードがメタデータとして付与されているツイートを収集する．ここでの地名コードとは，都道府県を表わす位置情報である[2]．

2.1.2 職業のツイートの収集

ウェブサイト「13歳のハローワーク」[3]に掲載されている職業のうち，代表的と思われるものを選別し，職業カテゴリと定義する．最終的に【医師】【教師】【料理人】など44個の職業カテゴリを設定した．

Twitter ではユーザの職業はメタデータとして付与されていない．本研究では，ユーザの職業を半自動的に推定し，これを付与した．まず，それぞれの職業カテゴリについて，プロフィールに職業名を含むユーザを検索し，その中から真に職業カテゴリを職業とするユーザ1名を人手で選別する．次に，そのユーザがフォローしている人を辿り，その人のプロフィールに職業名が含まれていれば，そのユーザも職業ユーザとして特定する．上記の操作を職業ユーザが100人以上得られるまで繰り返す．

最後に，それぞれの職業ユーザについて，そのユーザが発信したツイートを Twitter API を用いて収集し，職業のメタデータが付与されたツイート集合を得る．

2.1.3 時間のツイートの収集

Twitter ではツイートの投稿時間がメタデータとして付与されているため，これを時間のメタデータとしてそのまま用いる．2.1.1 で収集した場所のツイートのデータを使用し，投稿時刻を【朝】(5:00-11:00),【昼】(11:00-16:00),【夕方】(16:00-19:00),【夜】(19:00-24:00),【深夜】(0:00-5:00) に分類した上で，時間のメタデータが付与されたツイート集合を得る．

2.2 候補単語の取得

ツイート集合から，辞書に登録するべき登録単語の候補を得る．MeCab[4] を用いてツイートを形態素解析し，品詞が名詞，動詞，形容詞，副詞の単語を候補単語として抽出する．また，本研究では，ハッシュタグも候補単語とする．Twitter を対象とした応用システムでは，典型的使用場面が付与されたハッシュタグも有用であると考えられるためである．ただし，ハッシュタグは単語分割せずに全体をそのまま候補単語とする．

Construction of Lexicon with Typical Situations of Words from Microblog

[*] Toshinari Oka , Japan Advanced Institute of Science and Technology, [†] Kiyoaki Shirai , Japan Advanced Institute of Science and Technology

[1] https://www.tweepy.org/

[2] https://help.twitter.com/ja/using-twitter/tweet-location

[3] https://www.13hw.com/jobidx/jobnameidx.html

[4] http://taku910.github.io/mecab/

2.3 典型的使用場面の特定

Kleinburg のバースト検出アルゴリズム [2] を基に，特定のカテゴリのみに頻出する単語を特定する．具体的には，時間，場所または職業のカテゴリを c とし，候補単語 w がカテゴリ c に頻出するスコアを式 (1) で求める．

$$\sigma(0, r_c, d_c) = -\ln\left[\binom{d_c}{r_c} p_0^{r_c}(1-p_0)^{d_c-r_c}\right] \quad (1)$$

ただし，$p_0 = \dfrac{R}{D}$，$R = \sum_{c\in C} r_c$，$D = \sum_{c\in C} d_c$

r_c はカテゴリが c で単語 w を含むツイートの数，d_c はカテゴリが c であるツイートの数，C はカテゴリの集合である．p_0 はデータセット全体における候補単語 w の平均出現確率を表わす．カテゴリ c における単語 w の出現確率が p_0 と比較して大きな差があるとき，$\sigma(0, r_c, d_c)$ は大きい値を取る．

カテゴリ c のそれぞれについて，式 (2) と式 (3) の条件を満たす単語を選択し，典型的使用場面の辞書を構築する．基本的には，$\sigma(0, r_c, d_c)$ が閾値 K よりも大きい単語を選択する．ただし，$\sigma(0, r_c, d_c)$ は，単語 w がカテゴリ c に極端に出現しない場合にも大きくなるので，平均以上にカテゴリ c によく出現する単語を抽出するために式 (3) の条件を設定する．

$$\sigma(0, r_c, d_c) > K \quad (2) \qquad \frac{r_c}{d_c} > \frac{R}{D}(= p_0) \quad (3)$$

予備実験では，時間の辞書を構築する際，同一ユーザが同じ単語を何回も発信しているとき，その単語に誤った時間カテゴリが割り当てられることが多かった．例えば,【深夜】のカテゴリとして「頑ばる」が誤って獲得されたが，この単語は一人のユーザのみによって同じ時間帯に繰り返し発信されていた．このような誤りを回避するため，時間の辞書については，単語と時間カテゴリの相関関係を，その単語を含むツイート数ではなく，その単語を含むツイートを発信したユーザ数で評価する．すなわち，r_c をカテゴリが c で単語 w を含むツイートを発信したユーザの数，d_c をカテゴリが c であるツイートを発信したユーザの数とし，同様に式 (1)，(2)，(3) によって単語の典型的使用場面のカテゴリを特定する．

3 評価実験

2019 年 1 月から 12 月にかけて，時間，場所，職業のメタデータが付与されたツイートを収集した．表 1 は，収集したツイートの総数，最大もしくは最小のツイート数を収集したカテゴリとそのツイート数，カテゴリ当たりの平均ツイート数を示している．

構築した辞書の概要を表 2 に示す．式 (2) における閾値 K は，時間，場所，職業のそれぞれについて，どのカテゴリについても最低 50 個の単語を取得するように設定した．

表 1: 収集したツイートの概要

	総数	最大	最小	平均
時間	20.3M	6.4M【夜】	1.7M【深夜】	4.06M
場所	8.4M	1.7M【東京都】	30K【鳥取県】	0.17M
職業	8.6M	0.5M【声優】	23K【消防士】	0.19M

表 2: 自動構築した典型的使用場面付き辞書の概要

	単語数	カテゴリ数	単語/カテゴリ
時間	1,149	5	230.40
場所	16,554	47	364.51
職業	7,440	44	170.66

次に，構築した辞書を評価する．それぞれのカテゴリについて，式 (1) のスコアが高い上位 20 件の単語について，そのカテゴリが適切であるかどうかを人手で判定し，カテゴリ毎に正解率 (カテゴリが適切と判定した単語の割合) を求めた．判定は著者 2 名で行った．カテゴリ毎の正解率のマイクロ平均，最高値・最低値とそのカテゴリ，ならびに二者の判定の一致度を示す κ 係数を表 3 に示す．

表 3: 自動構築した辞書の評価

	正解率			κ
	平均	最高	最低	
時間	0.59	0.88【朝】	0.28【夕方】	0.83
場所	0.79	1.00【福岡県】	0.38【宮崎県】	0.77
職業	0.52	0.83【プログラマ】	0.15【主婦】	0.53

場所の辞書については，正解率が高いことから，特定の都道府県に関連した単語をうまく獲得できたことがわかった．一方，時間・職業の辞書については，カテゴリによっては正解率が高いものの，全体的には 6 割以下の正解率であり，改善の余地がある．時間の辞書については，正解と判定した単語の多くは，そのカテゴリの時間帯に放送されるテレビ・ラジオ番組のハッシュタグであり,「おはよう」などの一般的な単語は少なかった．職業の辞書については，ユーザの職業の自動推定の誤りが不適切な単語が獲得された原因のひとつと考えられる．

今後は，誤りの原因について精査し，より質の高い辞書を構築する手法を探究していきたい．

参考文献

[1] 服部峻. Web 知識を用いた時空間依存な対話システムの試作. 信学技報 人工知能と知識処理, 110(105), AI2010-3, pp. 13–18, 2010.

[2] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397, 2003.

[3] 松田耕史, 佐々木彬, 岡崎直観, 乾健太郎. 場所参照表現タグ付きコーパスの構築と評価. 情報処理学会研究報告, Vol. 2015-NL-220, No. 12, pp. 1–10, 2015.

[4] 岡利成. マイクロブログからの典型的使用場面付き辞書の構築. 修士論文, 北陸先端科学技術大学院大学, 3 2020.