2H5-3

Twitter データに基づく関心事項提示システム

Interesting Information Extraction System from Twitter

高松誠志郎　　宮治裕
Seishiro Takamatsu　　Yutaka Miyaji

青山学院大学 社会情報学部
School of Social Informatics, Aoyama Gakuin University

This study proposes a method of constructing interesting information extraction system from Twitter. We focus on small talk ability that is an element of communication skills. Interest of the conversational partner is used as a effectual topic during the small talk. This system analyzes Twitter data and reveals interest of the partner. In the result of using our system, extracted elements proved that it is possible to use the interest as a topic.

1. 背景および目的

近年，コミュニケーション能力を重視する傾向が増大している．日本経団連による新卒採用に関するアンケート調査において，選考時にコミュニケーション能力を重視すると回答した割合は，ここ 5 年間継続的に 8 割を超えている．

一般的に，コミュニケーション能力は様々な要素を含有，意味しているが，その一要素として雑談力が挙げられる．円滑な人間関係を形成する上で，雑談を行う力は現代人に求められる力であると斎藤 [斎藤 10] は述べている．また斎藤は，雑談時に効果的な話題の例として，相手が関心のある話題を挙げている．しかしながら，相手が関心のある話題そのものを知ることは容易ではない．

その一方，近年 SNS が急速に普及し，そこには相手の興味や関心を示すデータが散在している．日本における SNS の普及率は約 6 割を超えており，代表的な SNS の一つである Twitter は，日本人の SNS 利用者のうち 3 割以上が利用している．

以上の背景から，SNS のデータを利用し，コミュニケーション相手の関心を事前に調査，円滑な雑談を遂行するための補助システムを構築することを本研究の目的とする．本システムでは，コミュニケーション相手の関心を提示し，併せて種々の関連情報を提供することで，会話準備のコスト削減を行う．

2. システム構成

本研究では，Python 用 Web アプリケーションフレームワークである Flask を用いて，コミュニケーション相手の関心語を提示するシステムを開発した．関心語の提示と同時に，ユーザの大まかな関心を把握するためのユーザ属性や，関心語の概要情報およびツイートデータの表示などを行う機能も搭載した．利用するユーザは，本システムに Web ブラウザからアクセスし，分析したい対象の Twitter スクリーンネームを入力することで，本システムを利用することができる．

2.1 システム概要

本システムの処理概要を示したシステム概要図を図 1 に示す．利用ユーザは分析対象ユーザの Twitter スクリーンネームを入力し，実行する．Twitter スクリーンネームを受け取ったシステムは，そのユーザのプロフィール情報およびツイートデータを Twitter から取得する．システムは，これら二つのデータを用いて，ユーザの特徴語を抽出する．このとき，抽出された特徴語を含むツイートを収集する．また，特徴語に対してユーザが抱く印象の分析も行う．これらの処理が終了し次第，各語の概要およびカテゴリ情報を Wikipedia から取得する．最後に，はてなキーワードを用いてユーザ属性を推定し，利用ユーザに対し結果を返却する．結果を受け取った利用ユーザは，特徴語を含むツイートのデータや，Wikipedia からの情報を適宜閲覧することができる．

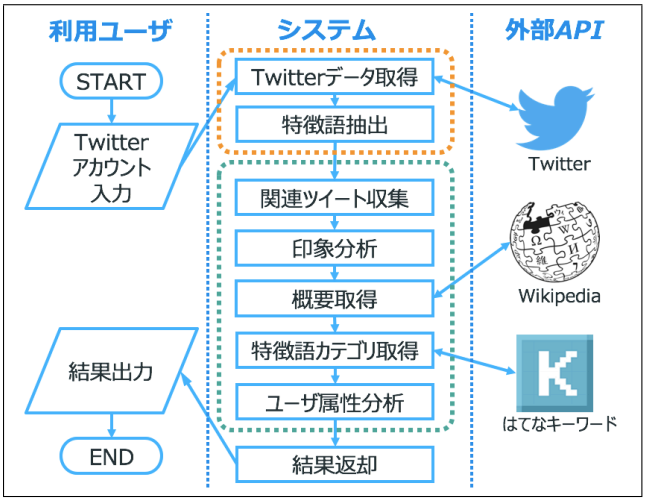

図 1: システム概要図

2.2 特徴語抽出

各語のスコアを算出し，スコア上位の語を，分析対象者の関心がある語として，特徴語とする．スコア算出にあたり，本研究では tf-idf 法を用いた．各特徴語のスコア $S_{u,i}$ は以下の式によって算出される [渡邊 12].

$$S_{u,i} = \frac{t_{u,i} + p_{u,i} \cdot C}{T_u} \cdot \log \frac{U}{U_i}$$

ここで，ユーザ u の語 i について，ツイート中の登場回数を $t_{u,i}$ で表し，プロフィール中の登場回数を $p_{u,i}$ で表す．C はツイート総数によって重みづけされたカウント回数である．また，

連絡先: 高松誠志郎，青山学院大学 社会情報学部，takataka4072715@gmail.com

T_u はユーザの分析対象となる全ての語の数を表す．ランダムに取得した Twitter ユーザ数を U，そのうち語 i を使用したユーザ数を U_i とする．この Twitter のランダムなユーザ集合は，2015 年 12 月に Streaming API を用いて収集した 2,000 人分のデータとした．

2.3 印象分析

抽出された特徴語に対し，分析対象ユーザが持つ印象を，ポジティブかネガティブかの二値で判定する．この判定には，小林ら [小林 05] および東山ら [東山 08] によって整備された日本語評価極性辞書を用いる．形態素解析の結果として出力される語の原形を調べ，その語が日本語評価極性辞書中に存在する場合，当該ツイート内の全ての名詞に対し，ポジティブまたはネガティブのカウントを加える．この値から割合を算出し，分析対象ユーザが持つ特徴語に対する印象値とした．

2.4 ユーザ属性推定

本システムでは，株式会社はてなが提供するサービスのはてなキーワードを利用し，ユーザ属性の推定を行う．はじめに，特徴語をカンマ区切りにしたデータを，はてなキーワード自動リンク API に対し送信する．この API は，送信された語がはてなキーワード内に存在する場合，その語のカテゴリ情報を返却する．この返却データを利用し，ユーザ属性を算出する．

カテゴリ毎の値は，特徴語に付随しているカテゴリに対し，その語のスコア値を足し合わせることで算出する．返却された全てのカテゴリに対し上記の処理を行い，各カテゴリの値を全体の値によって割ることで，そのカテゴリが占めるユーザ属性の割合を求める．

3. 実験

本システムが実際に分析対象者の関心語を抽出できているかを明らかにするため，検証実験を行った．実験対象者は Twitter 利用者 12 名とした．被験者は自らの Twitter スクリーンネームを本システムに入力し，提示された各分析結果に対してアンケートに回答する．表示される内容の例を図 2 に示す．

本システムが提示した特徴語に対する評価では，関心があると回答した割合は 65 %であり，話の種になると回答した割合は 62 %であった．特徴語に対して表示される印象値に関しては，55 %が正しいと評価された．

本システムが提示したユーザ属性は，1 名の被験者を除いて肯定的な評価を得た．また，本システムが提示したユーザ属性と，被験者が自分自身に当てはまると考えたユーザ属性の合致率は 52 %であった．

本システムが備える概要表示やツイート表示といった機能に対する評価も，1 名を除いて役立つと回答された．

以上の実験結果から，本システムが提示した特徴語は，半数以上の割合で話題の種として利用できることが明らかとなった．したがって，特徴語は話題として十分に機能し得ると考える．

特徴語の印象は，半数の語で正しいと回答された．残りの半数の語が正確な印象値を算出できなかった理由として，印象値の算出方法に原因がある．本システムでは，印象語が一つのみの場合，印象値が極端な値を示すため，ユーザ意識との隔たりが発生したと考える．

ユーザ属性の推定に関しても，半数以上の合致率が示され，大まかな情報を表示するという目的は達成された．ただし，ユーザ間における合致率には大きな差異が見られる．これは，はてなキーワードのカテゴリの偏りによるものであると考える．

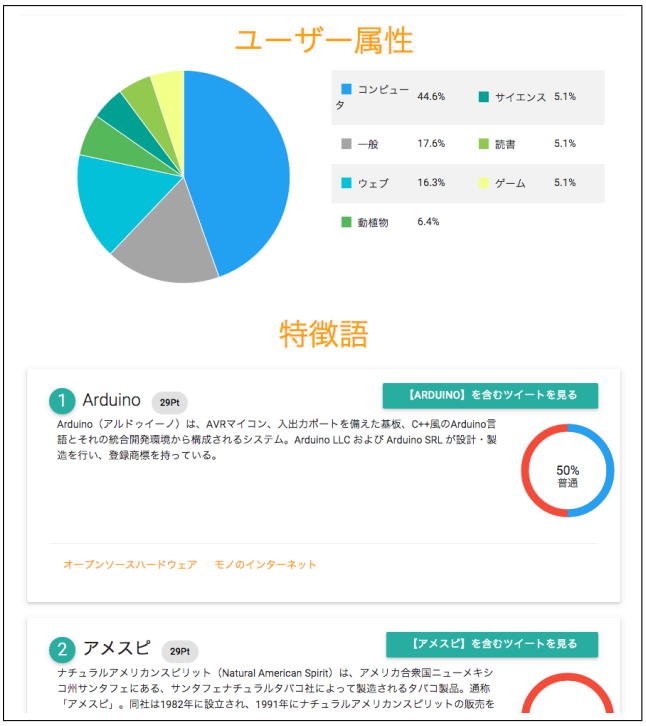

図 2: 結果例

4. まとめ

本研究では，SNS のデータを利用し，コミュニケーション相手の関心を事前に調査し，円滑な雑談を遂行するための補助を目的とし，システムの構築を行った．検証実験の結果，本システムが関心事項としてユーザに提示する特徴語は，会話を行う上で十分に有用な情報であることが示された．また本システムの機能の有効性も示され，人力で相手の関心を見抜くコストを削減し，より多くの情報を収集することが可能となった．

本システムの課題として，特徴語抽出の手法やユーザ属性の推定手法が挙げられる．特徴語の抽出に時間的価値基準を設け，はてなキーワード以外のユーザ属性の推定手法を用いる必要がある．

参考文献

[斎藤 10] 齋藤孝. 雑談力が上がる話し方—30 秒でうちとける会話のルール. ダイヤモンド社, 2010.

[渡邊 12] 恵太渡邊, 昇平加藤. Twitter における語の関連性に着目したユーザ興味語抽出手法の提案 (人工知能学会全国大会 (第 26 回) 文化, 科学技術と未来) – (データマイニング). 人工知能学会全国大会論文集, Vol. 26, pp. 1–4, 2012.

[小林 05] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[東山 08] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会年次大会発表論文集, pp. 584–587, March 2008.