2V-2

ソーシャルネットワークを情報源とした コミュニティ辞書自動生成の研究

宮本 和幸† 菱山 玲子†
早稲田大学大学院創造理工学研究科経営システム工学専攻†

1. まえがき

機械翻訳サービスは誤訳という問題を抱えており，辞書を強化することで改善が見込まれる．従来，この独自の辞書を作成する作業は主に手作業で行われており，多大なコストとなる．一方，インターネットの利用増加に伴い Web 上の情報が有益な情報と考えられ，解析対象の一つとされている．そこで本研究では，Twitter という情報源に着目し，特定のコミュニティに関する tweet からコミュニティ語彙を抽出し，コミュニティ対訳辞書の自動生成を目指す．ここでは，特定のコミュニティとして「早稲田」というコミュニティを対象とした．対訳の収集方法としては，集合知的な収集を試みた．なお，本研究において，コミュニティ語彙とは以下の通り定義する．

「公共空間で用いられるのと比較して，特定の地域や社会の中で明らかに頻繁に利用される，または，他と明らかに区別される地域や社会の中で特有の意味を持って共有される語彙（専門用語を含む）」

2. 関連研究

2.1 言語グリッドにおけるコミュニティ翻訳

言語グリッド[1]は，Web 上に分散した言語資源や計算資源を連携させることで，オープンスタンダードなプロトコルで質の高い言語サービスを提供するグリッドコンピューティング基盤である．この言 語グリッドを利用して開発された「スマート翻訳」[2]は，言語グリッド Toolbox のカスタマイズにより構築されており，コミュニティ辞書の連携を行うことができる．しかし，スマート翻訳の問題点は，機械翻訳の翻訳品質を確保するために不可欠となるコミュニティ辞書の構築機能が提供されていない点にある．例えば，このスマート翻訳を利用した実用サービスである「京大翻訳」[3]では，コミュニティ辞書として，京大に関連した用語を 15,000 語以上登録している．しかし，この辞書の作成作業は，対訳語の品質を維持するため，多言語で書かれた同一の内容の文書からに人手による収集を行っており，多大なコストがかかっている．この問題に対して，本研究の特徴は，対訳語としての品質低下を招くことなく，コミュニティ語彙の収集自動化を図り，現在手作業で行われている語彙コストを削減する点にある．

2.2 Web を情報源とするコミュニティ語彙抽出

石田ら[4]は，Web から専門辞書作成のための特徴語を抽出するシステムを提案している．この研究では，Web ページの重要度，単語量を加味することにより，任意のカテゴリに属する専門用語の抽出を行っている．しかし，この研究は専門用語となりうる用語の抽出のみを対象としており，それらの用語に対して訳語や説明文などの付加情報の獲得までは行っていない．専門辞書作成において，これらの情報は重要かつ高い精度が必要とされると考えられる．

この問題に対して，本研究の特徴は，対訳の獲得まで研究対象を広げ，また，Web よりも高いコミュニティ性の高い言語表現を含むと仮定される Twitter を情報源とする点にある．

3. 提案システム

図 1 に本研究の提案するシステムの概要図を示す．

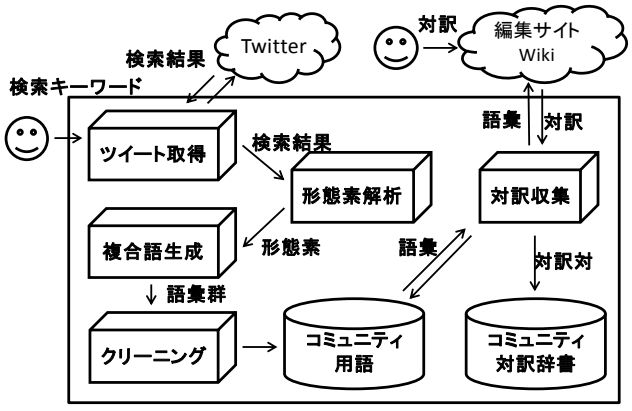

図 1. 提案システム概要図

ツイート取得モジュールでは，検索クエリを用いた検索，特定ユーザのツイート検索の二種類の方法でツイートを取得することができる．本研究では，対象コミュニティを「早稲田」としているため，以下の 3 つの検索結果を対象データとした．

- クエリ「早稲田」による検索結果
- クエリ「#早稲田」による検索結果
- 「waseda」をユーザ名に含むユーザの過去ツイート

形態素解析モジュールでは，取得したツイートに対して形態素解析を行い，名詞・固有名詞・未知語を抽出する．未知語とは，形態素解析の際に用いる辞書に登録されていない単語のことを指す．この未知語には，記号などのノイズが含まれる一方，辞書に登録されていない新しい単語が含まれる可能性が高い．形態素解析には，Java で記述された形態素解析器 Sen を用いる．

クリーニングモジュールでは，形態素解析モジュールで取得された単語に，「(」などの記号が含まれることがある．これは，これは抽出された用語の集合に対して，含まれているノイズ，一般的な用語の抽出を行い，削除する．一般的な単語の判定には，統計量である tf・idf 値を用いる．

対訳収集モジュールでは，コミュニティ語彙を，集合知的に対訳を獲得するために作成した Web サイトにアップロードする．アップロードした語彙に対して，対象となるコミュニティに属する人々によって入力された対訳を収集

Automatic Generation of the Community Dictionary Based on Social Network
† Major in Industrial and Management Systems Engineering, Graduate School of Creative Science and Engineering, Waseda University

し，DB へ格納する．この際に使用する Web サイトは，いつでもアクセス可能な状態である．

4. 評価実験

本研究の提案するシステムの有効性を検証するために以下のように実験目的を設定する．

1. Twitter から収集した用語の「コミュニティ性」の検証
2. 集合知的な訳語の収集に関する検証

上記に挙げた 1，2 の実験目的のために，それぞれ以下のように実験を行った．

4.1 コミュニティ判定

実験目的 1 を検証するための手順を以下に示す．

Step1 キーワード「早稲田」で検索エンジンを用いて検索を行い，検索結果上位 50 件に対して被験者が「早稲田」と関連性のある単語を手動で抽出(語群 Web)

Step2 ツイート取得モジュールより取得した，早稲田に関する約 2,500 の tweet からコミュニティ用語を抽出(語群 Twt)

Step3 語群 Web，Twt から，それぞれ 100 個の単語をランダムに抽出

Step4 語群 Web，Twt からランダムに抽出された単語に対して，被験者が，コミュニティ性の高さを示すコミュニティ値を評価．この際，知らない単語が出現した（判定不能な）場合は 0 を記入してもらう（0 の個数を CountZero とする）

Step5 語群 Web，Twt に対して与えられたコミュニティ値から，それぞれ CommunityScore(以下,CS)を求める

$$\mathrm{CommunityScore} = \frac{\sum_{i=1}^{100} \mathrm{Ci}}{\mathrm{N} - \mathrm{CountZero}}$$

4.2 Web サイトによる集合知的対訳収集

実験目的 2 を検証するための手順を以下に示す．

Step1 被験者が常にアクセス可能な Web サイトに，誰もが編集可能な Web ページ（Wiki）を作成

Step2 語群 T に属する単語（149 単語）を Web ページに掲載し，被験者に対応する訳語を記入し，編集してもらう

Step3 被験者 11 名に，約 1 カ月の期間で実施

5. 実験結果と考察

コミュニティ性判定の実験の結果を表 1，図 2 に示す．表 1 は CS の平均値，図 2 は被験者ごとの CS をそれぞれ W，T について表している．

表 1 が示す通り，CS(Twt) の平均値は CS(Web)よりも，約 0.8 高く，CS(Web)と比べると約 30%向上していた．このことより，Twitter という情報から Web よりもコミュニティ性の高い語彙の抽出に成功したと考えることができ，Twitter が有用な情報源となりうるといえる．また，被験者 10 名の全員に対して，CS(Twt)が CS(Web)よりも高い値を示していることから同様のことが考えられる．

表 1．CommunityScore の平均値

CS(Web)	CS(Twt)
2.751	3.546

集合知的対訳収集の実験結果を表 2 に示す．表 2 が示す通り，149 の語彙に対して 121 の対訳が付けられた．これは全体の約 80%にあたり，これらのほぼすべての対訳は精度的に問題のない質で獲得することができた．

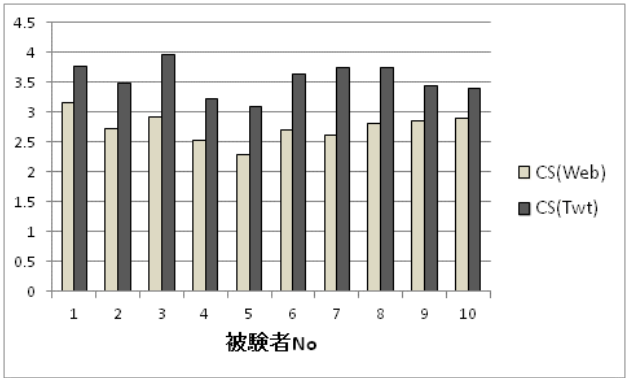

図 2．W,S に対する被験者ごとの CS

また，これらの 121 の対訳の中に，付加情報を含んだ対訳が 10 個含まれていた．ここでいう付加情報とは，対象となる語彙からのみでは表現できない，対訳付与者の知識が含まれている情報のことである．そして，このような付加情報を持った語彙が原文に含まれる文章を翻訳する際に，翻訳結果の精度は大幅に向上すると考えられる．

さらに，Web ページから対訳を抽出できる語彙の数は 8 個にとどまった．ここでいう Web ページとは，同一の内容を多言語で記述している Web ページのことであり，その多くは「English 版はこちら」などのリンクによって導かれる．この語彙が少ないということは，Web ページから対訳を抽出することができない多くの語彙に対して，集合知的に対訳を獲得することができたということであり，本研究の提案するシステムが語彙獲得，対訳獲得に有効であると考えることができる．

表 2．対訳の収集結果

対訳が付けられた用語	121
付加情報がある対訳	10
Webページから対訳を抽出できる用語	8

6. まとめと今後の課題

本研究では，コミュニティ辞書の作成におけるコスト削減のために自動化を目指し，情報源としてコミュニティ性の高い語彙が含まれているという仮定の元，Twitter からのコミュニティ語彙の獲得を試みた．評価実験により，本研究の提案システムが有効である可能性を示した．

今後の課題としては，利用可能な規模のコミュニティ辞書が実際の翻訳に与える影響の検証や，ユーザのプロフィールの推定を行うことにより，より有益な情報を効率的に獲得することが考えられる．

7. 参考文献

[1] T. Ishida: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp. 96-100, keynote address, 2006.

[2] 稲葉利江子他，言語グリッドを用いたスマート翻訳，AAMT ジャーナル，No.49， pp.23-29， 2011.

[3] 京大翻訳：*http://langrid2.nict.go.jp/smarttrans.html*

[4] 石田百人，ページの重要度，単語量を加味した専門用語抽出システム，早稲田大学卒業論文，2011