5C-3

Twitterからの談話自動抽出

堀川敦弘 †1　　當間愛晃 †2

†琉球大学 工学部情報工学科

1 はじめに

1.1 研究背景

新しいコミュニケーションサービスであるTwitter[1]は、情報の収集や発信の場として利用されており、談話や議論を行う場としても利用されている。また、さまざまな研究対象としても注目を浴びてきた。

通常Twitterで談話や議論をまとめるために用いられる機能の一つにハッシュタグがある。これは情報の発信者が自らのTweetにどの様な内容であるのかという情報を付与するものであるが、Twitter上ではハッシュタグが付与されていない談話も数多く行われており、現状では、これらをまとめるために、Togetter [2]などのサービスのように人手でまとめる方法しか存在せず、とても効率的とはいいがたい。そこで本研究では、Twitterからハッシュタグの有無に関わらず談話のまとめを自動的に生成するシステムを提案する。

また、本研究の提案手法を用いれば、談話をTwitterからリアルタイムに抽出することも可能である。Togetterなどのサービスにはリアルタイム性が皆無であり、まとめられている談話はすでに終了していることが多い。提案手法を用いればまとめられている談話を読み返しながら談話に参加することができ、新しい形の談話支援ツールとなることが可能である。

1.2 先行研究

与儀ら [2] はTweet群から議論の要約を自動生成する基礎研究として、Tweetを機械学習で種類ごとに分類するという研究を行った。与儀らの研究ではハッシュタグなどでまとめられている議論から要約を自動生成することが目的であり、談話抽出を目的とすることと、ハッシュタグの有無に関わらない点で本研究と異なっている。

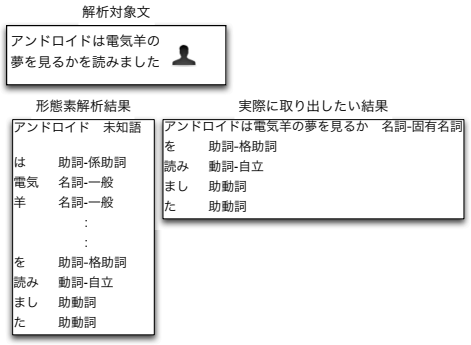

図 1: 形態素解析で問題が生じる例

1.3 研究内容

談話に関連したTweetであるかの判別には共起を用いる。なお、本研究では、ある情報発信者のTweetを閲覧した人間が、そのTweetに影響されたTweetをしたとき、談話がおこったとする、本研究ではSeed Tweet setを与え、このSeed Tweet setに関連した談話を抽出する事が本研究の目的である。

小野ら [1] は共起を用いてWebページにメタデータを付与する際、共起情報の収集をコーパスを用いることによって実現しているが、本研究では言葉の経時変化やその場で作られた造語などに対応するためTwitterやWikipedia[3]などから動的に共起情報を収集する手法を提案する。また、本研究ではTwitterやWikipediaなどから動的に共起情報を収集するため文を単語に分解することが必要である。しかし、通常の形態素解析では、解析に用いる辞書に登録されていない単語は「未知語」として検出されるか、「既知の単語の組み合わせ」として出力されるので、Twitterなどスラングが多用される問題を正しく処理することができない (図1)。そこで本研究ではGoogle Suggest API [4]を用いて形態素解析で検出できない語を解決することを提案する。

更なる問題点としてTwitterから談話を抽出するにあたって、「スラング等が大量に使用されている」「1Tweetの文字数が最大140文字と少ない」などの条件が談話に関連したTweetであるか判別することを難しくしてい

An automatic extraction of a specified discourse on Twitter
†Dept. of Information Engineering, Univ. of the Ryukyus
†1 Atuhiro HORIKAWA
†2 Naruaki TOMA

[1] http://twitter.com/

[2] http://togetter.com/

[3] http://ja.wikipedia.org/

[4] http://google.com/complete/search?output=toolbar&q="クエリ"&hl=ja

る。本研究ではこれらの問題の解決方法を模索する。なお、本研究では談話の始まりと終わりの検出は行わない。

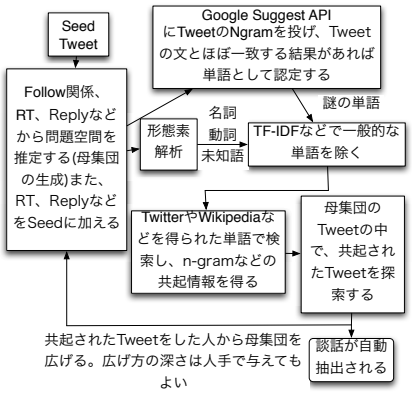

図 2: 提案手法の概要

2 提案手法

提案手法の概要として図 2 を示す。本手法では入力された Seed Tweets set と Followers から関係者発言一覧を取得し、これを母集団とする。この母集団から単語共起を利用して関連する談話発言抽出を行う。

Step1： Seed Tweet Set の入力

談話を自動抽出する際、談話に含まれている 1 つ以上の Tweet を Seed tweets set として人手で与える。このとき、与える Seed tweets set はより多く ReTweet や Reply を受けたものが理想的であると考えられる。なお、ReTweet とは他のユーザの Tweet を再投稿することを指し、Reply とは特定の Tweet への返信の事を指す。

Step2： 母集団の生成

与えられた Seed tweets set の Follow 関係や ReTweet された範囲、Reply などから談話を抽出する母集団を生成する。この母集団の範囲は、本研究の談話の定義による。

Step3： Seed tweets set の拡張

Seed tweets set の ReTweet や Reply なども Seed tweets set に含める。さらに、ハッシュタグ等が付与されているとき、同様のハッシュタグが付加されている Tweet を Seed tweets set に加える。

Step4： 単語の抽出

Seed tweets set となった Tweet を形態素解析と Google Suggest API の両方にかけて、「名詞」、「動詞」、「未知語」などの抽出を行なう。Google Suggest API には Seed tweets set の n-gram を処理させ、Suggest 結果に Seed tweets set とほぼ一致する場所があれば、それを新しい単語として認定する。

Step5： 一般的な語の排除

Step3 の結果から TF-IDF などにより一般的な単語を排除する。

Step6： 共起情報の作成

Seed tweets set から得られた単語から共起情報を作成する。ここで、共起情報を自動収集する場合は、Wikipedia や Twitter、Blog などに Seed tweets set から得られた単語で検索をかけることで共起情報を収集する。

Step7： 談話の抽出

母集団の中に共起された単語があるか調べる。共起の割合で閾値を定め、閾値以上の Tweet は談話に参加しているとする。

Step8： 繰り返し処理

談話に参加している Tweet が判明したので、それを Seed tweets set に加え、Step2 から繰り返し談話の抽出を行なう。なお、繰り返す回数は本研究では人手で与える。

上記に本研究の提案手法を示した。なお、これらの手法は今後の実験結果を考慮した上で改善する予定である。

3 現状と今後の対応

母集団を生成するなどの工程で逐次 Twitter にアクセスし続けると、Twitter の API 制限で一定時間 Twitter にアクセスできなくなる。この問題を回避するため、Twitter に実験用の問題空間を仮定して、その問題空間内の全ての Tweet、ユーザ情報、Follow 関係などをローカルに取得しており、今後このデータを実験に用いる。

結果の検証は、本手法で作成したシステムの結果と、母集団から人手で作成した結果をくらべ、その適合率と再現率を示す予定である。また、共起情報の取得先などを変更して複数の実験を行い、それぞれの手法ごとの結果を示す予定である。

参考文献

[1] 小野 裕作, “共起情報を用いた Web ページを特徴付けるメタデータ生成方式の検討と検索への応用”, 第 19 回インテリジェントシステムシンポジウム FAN2009 論文集, pp.462-465, 2009

[2] 与儀 涼子,“Twitter 上で行われる談話要約のための、文脈を表現する指標構築のための検討”, 第 9 回情報科学技術フォーラム (FIT2010), E-024, 2010