1P-4

Twitterにおける身内的表現を用いた談話同定に関する検討

堀川敦弘 †　當間愛晃 ‡　赤嶺有平 ‡　山田孝治 ‡　遠藤聡志 ‡
琉球大学大学院 理工学研究科 情報工学専攻 †　琉球大学 工学部 情報工学科 ‡

1　はじめに

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や Mention 関係がある。しかしながら、Twitter 上ではハッシュタグや Mention 関係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 *など人手でまとめる手法しか存在せず、効率的とはいいがたい。

そこで本研究では Seed Tweet Set として幾つかの Tweet をシステムに与えると、システムがユーザのタイムラインから Mention 関係やハッシュタグの有無にかかわらず談話を自動的に抽出するシステムの構築を目指している。

今日までの研究成果として、単純共起法による談話自動同定システムの提案 [1] などを行った。これは、あるタイムラインから、特定の談話を自動同定するため、談話に属した複数の Seed Tweet Set をシステムに入力し、入力ツイートにおける単語の共起関係を用いてタイムラインから談話を抽出するという手法である。これらの結果から単純共起だけでは談話同定することのできない Tweet が存在することが示された。

そこで我々は、単純共起のみで談話同定できない Tweet の中には、Seed Tweet Set から共起的つながりがなくとも、ユーザ間で使用される身内的な共起関係が存在すると仮定しその検証および抽出方法の検討を行った。

2　提案手法

ある期間中にユーザ同士が Tweet しあったものの中で、他ユーザ間ではほとんど用いられていない表現は、その期間中において身内的な表現の可能性が高いと考えられる。そこで我々は、二者間での Reply のやりとりとタイムライン上の全てのユーザの Reply 関係を用いて、idf を適用した二者間共起辞書の作成し、これを用いることで身内的表現の抽出を提案する。

2.1 節で作成方法を、2.2 節で利用方法について述べる。

2.1　idf を適用した二者間共起辞書の作成

Step1：収集期間とユーザ ID の入力

システムに入力として二者間共起の収集期間、2 件のユーザ ID を与える。システムは収集期間の範囲内で、与えられたユーザ ID 同士で Reply している Tweet を収集する。

Step2：形態素解析とノイズ除去

Step1 で収集した Tweet を 1Tweet ごとに MeCab を用いて形態素解析し、名詞と動詞のみ取り出す。さらに各形態素を基本形に変換する。なお、@から始まるユーザ名も除去する。

Step3：二者間共起辞書の構築

各ツイートに出現した形態素の組み合わせを共起辞書に登録する。すでに登録されている共起については出現回数をインクリメントする。

Step4：idf 辞書の構築

事前に取得した多人数の Reply のやりとりから、形態素の共起を作成し、その出現回数を計測し idf 辞書に格納する。なお、本論文の実験では 497,168 件の Tweet を使用し idf 辞書を構築した。

Step5：idf を適用した二者間共起辞書の作成

Step3、Step4 で作成した辞書を用いて各共起の Point を求め二者間共起辞書に登録する。Point は式 (1) により求める。

$$Point = Co\_occur\_dic(i) \times log\frac{idf\_dict(i)}{Pair\_dict(n_i)} \quad (1)$$

Co_occur_dic(i):共起辞書における共起 i の出現回数

idf_dict(i):idf 辞書に登録された共起 i の総出現回数

Pair_dict(n_i):ユーザ n の二者間共起辞書に登録された共起 i の出現回数

A study to extract same discourse identification with relative expressions on Twitter.

†Atuhiro HORIKAWA, Graduate School of Engineering and Science, University of The Ryukyus

‡NaruakiTOMA, Yuhei AKAMINE, Koji YAMADA, Satoshi ENDO, Dept. of Information Engineering, Univ. of the Ryukyus

*http://togetter.com/

2.2 idfを適用した二者間共起辞書の利用

Step1：タイムラインの入力

同定したい談話を含んだタイムラインの収集期間を入力する。

Step2：形態素解析とノイズ除去

収集したタイムラインに対して、二者間共起辞書の作成のStep2と同様の形態素解析処理を行う。

Step3：idfを適用した二者間共起辞書の利用

Tweet毎にStep2の結果得られた形態素の組み合わせが二者間共起辞書に存在するか調べる。存在していた場合、その全ての共起関係のPointを足しあわせ、全ての共起関係の個数で割ることで得られる数値をそのTweetの2ユーザ間での身内度とする。本稿ではこの身内度の高いTweetをそのままSeed Tweet Setと同一談話とみなしてランク付けし出力している。

3 実験

提案手法を、我々 [2] が以前実施した談話同定アンケートの結果 (D1) と、D1と同様の手法で新たに実施したアンケートの結果 (D2) の二件のデータセットで検証した。

二者間共起の収集期間はD1では5ヶ月、D2では6ヶ月とした。

二者間関係にはそれぞれのデータセットでSeed Tweet Setを発話したユーザ2名を採用した。

3.1 データセットについて

実験にD1とD2のデータセットを使用する。これらのデータセットは、Tweet100件程度のタイムラインと2件のSeed Tweet Setを回答者に提示し、各TweetがSeed Tweet Setと同じ談話に「属する」「おそらく属する」「属さない」のうちどれにあたるかを選択してもらったアンケートである。詳細を表1に示す。

アンケート結果で、「(A) 回答者の全員がSeed Tweet Setと同じ談話に属している、またはおそらく属している」と回答したTweetの件数を表1における正解数とした。成功件数とは、単体共起法で (A) のTweetが閾値以上にランク付けされ成功した件数を示している。同様に失敗件数とは単純共起法で、(A) のTweetが閾値を超えず失敗した件数である。

表 1: データセットの詳細情報

	Tweet	回答者	閾値件数 (正解数)	成功件数	失敗件数
D1	115 件	7 人	20 件	10 件	10 件
D2	100 件	8 人	27 件	10 件	17 件

3.2 実験結果と考察

実験結果を表2に示す。表2においてリランクとは表1の失敗件数に含まれた (A) のTweetが提案手法で表2の閾値件数以上にリランクされ成功した件数を示している。

表 2: 提案手法の出力結果

	リランク	成功数 (All)	成功数 (Some)
D1	3 件	11	16
D2	4 件	9	20

実験結果より、単純共起法では下位ランクだったTweetを上位ランクにリランクすることができた。

表2において成功数 (All) とは提案手法において (A) のTweetが閾値件数を超え成功した件数である。この結果より、データセットD1とD2においては談話同定の評価値の一つとして提案手法を用いることで、単純共起法の適合率を上昇させることができると考えられる。

単純共起法および提案手法を用いても評価値が空となってしまい上位にリランクできなかったTweetがD1で1件、D2で6件存在した。これらについて観察した結果、Tweetを形態素解析した品詞の中に名詞や動詞が存在しないなどの理由から有効な共起が構築できなかったことに原因があると考えられる。これらのツイートはツイート自体が短い場合や、造語、洒落などを含む場合があり、共起という手法が単純には通用しない可能性が高く、別の手法を考案する必要がある。

4 今後の対応

表2の成功数 (some) は、提案手法においていくつかのユーザがSeed Tweet Setと同じ談話に属している、またはおそらく属していると回答したTweetが閾値件数を超え成功した件数である。これは全員にとっての正解ではなく、特定のユーザを対象とした正解を示すことができれば適合率が上昇する可能性を示していると考えられる。単純共起法で閾値以上のランクであったにもかかわらず提案手法で閾値以下に下落したTweetはD1で2件、D2で5件存在したが、すべて入力した2者間関係とは異なるユーザ関係のツイートであった。また、リランクできなかったTweetのうちD1で5件、D2で9件が入力した二者間関係とはことなる関係によるTweetであった。これらの事から、入力する二者間関係の変更や、複数の二者間関係の組み合わせなどを行い、検証する必要がある。

参考文献

[1] 堀川敦弘：Twitterからの談話自動抽出, 情報処理学会 第74回全国大会, 5C-3, pp.31-32, 2012

[2] 堀川敦弘：twitterからの談話自動同定法の一検討, 第22回インテリジェント・システム・シンポジウム (FAN2012), 2012