共起距離に基づく文特徴量を用いた派生談話認識に関する調査

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 †
k128567@ie.u-ryukyu.ac.jp

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 [1]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。

そこで我々は 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。これまでの研究成果として、Twitter からハッシュタグの有無に関わらず共起による談話のまとめを自動的に生成するシステムの提案 [2] とその基礎研究として、談話の定義のゆれについて考察 [3] を行ってきた。特に [3] では、談話定義における個人間の揺れの原因に談話の派生が大きく関わっていることが示唆された。すなわち、談話を自動抽出するためには談話の派生度合い（離れ度合い）を自動推定することが必要である。

本研究では談話の派生度合いを自動推定するために、[2] では考慮しなかった、システムに与える Seed Tweet Set に含まれる形態素を基準とした相対的な共起距離を特徴量として求め、この共起距離が離れるほど同じ談話であると感じにくくなることを検証した。

なお、共起とは 2 つの語がある範囲に同時に出現することを指している [1] 。

[1] http://togetter.com/

2　調査手法

先行事例 [3] で行ったアンケート結果と Seed Tweet Set からの共起の距離を比較する。

アンケートは 2012-01-05 14:34:40+09 から 2012-01-05 14:59:11+09 までに ie_list [2]内で行われた 115 件すべてを Tweet を母集団とした。すべての母集団に対して、2 件の Tweet を Seed Tweet Set と関連した話題であるか、各ツイートに対して「関連した Tweet である」「関連しない Tweet である」「どちらとも言えない」という評価をしてもらった。アンケートで提示した情報は、ユーザ名、Tweet 本文、投稿時間である。また、この母集団の平均文字数は平均 72.7 文字であった。Seed Tweet Set と関連した談話を中心的に発話していた 7 名にアンケートを依頼した。

これらの結果、Tweet を「全員関連した談話だとしたもの」「全員関連した談話ではないとしたもの」「どちらとも言えない点を含むものの概ね関連した談話だとしたもの」「どちらとも言えない点を含むものの概ね関連した談話でないとしたもの」に分類した。

分類した Tweet がそれぞれどのような Seed Tweet Set からの距離となるのか観測する。

Step1 : Seed Tweet Set の入力

談話の元となる Seed Tweet を入力する。本実験では 2Tweet を Seed Tweet Set として与えた。

Step2 : Seed Tweet Set の形態素解析とフィルタリング

入力された Seed Tweet Set を形態素解析し、名詞、動詞などの使用する形態素のみを取り出し、第 1 ステップ辞書に登録する。

Step3 : ステップ辞書の作成

[2] https://twitter.com/#!/nakarx/ie

母集団の中から、第 1 ステップ辞書に登録された形態素が出現している Tweet を検索し、発見した場合は、その Tweet の形態素を解析し、使用する形態素のみを取り出して新たな第 i ステップ辞書に登録してゆく。

Step4 : ステップ辞書の完成

Step3 で作成した第 i ステップ辞書を用いて Step3 を繰り返し実行することで、新たな i+1 ステップ辞書を繰り返し作成する。なお、新たなステップ辞書に登録する形態素は、それ以前のステップ辞書に登録されていない形態素に限る。ここで作成されたステップ辞書を用いて共起的距離の算出を行なう。ここで距離を D とすると、D は Seed Tweet Set : D=0、Seed Tweet Set から繋がった次のステップ辞書に登録された形態素は D=1、その次は D=2 となるような数である。

Step5 : 各 Tweet の距離の測定

Step4 で求めた距離を用いて式 1 のように各 Tweet の距離を算出する。

$$TweetDistance = \frac{\sum_{i=0}^{TweetLength} D(morpheme\_i)}{TweetLength} \quad (1)$$

- TweetLength : Tweet に含まれる各ステップ辞書に含まれていたすべての形態素の数。
- D(morpheme_i) : Tweet に含まれる全形態素 i の Seed Tweet Set からの距離。

3 調査結果

調査結果を図 1 に示す。Y 軸は Tweet Distance であり、X 軸は母集団の Tweet ID となっており、Tweet ID は時間が進むに連れ大きくなるので、X 軸が大きくなるほど時間が進んでいる。「全員関連した談話だとしたもの」では概ね Tweet Distance が 1 以下となった。また時間的経過も初期の段階に集中している。「どちらとも言えない点を含むものの概ね関連した談話だとしたもの」は若干 Tweet Distance の値が広がったが、概ね 2 以下の値となった。「どちらとも言えない点を含むものの概ね関連した談話でないとしたもの」は Tweet Distance の値自体は「どちらとも言えない点を含むものの概ね関連した談話だとしたもの」と変わらないが、時間的変化が大きく影響していることがわかった。「全員関連した談話ではないとしたもの」は Tweet Distance の値が非常に離散的であり、仮説とは異なる結果となった。

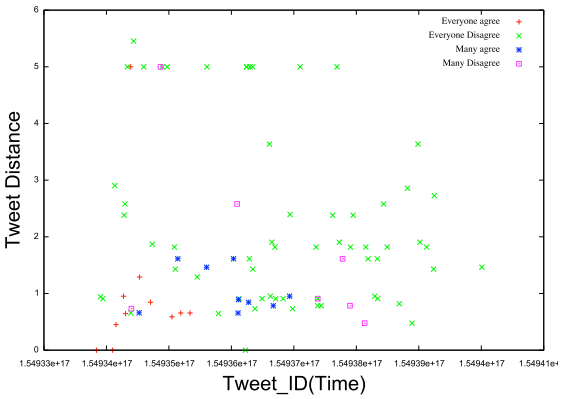

図 1: 共起の例

4 おわりに

本研究では談話の派生度合いを自動推定するため、共起距離が離れるほど同じ談話であると感じにくくなるという仮説を検証した。その結果、「全員関連した談話ではないとしたもの」以外では共起距離や時間である程度仮説に準じる結果となったが、「全員関連した談話ではないとしたもの」は非常に離散的な結果となっており、共起的距離から派生を検知するためには、これを他のアプローチで除去せねばならない可能性が示唆された。また「全員関連した談話ではないとしたもの」の中で共起的距離が小さかった Tweet の中に「Tweet 自体の形態素数が少なく、一般的な表現のみの Tweet」が存在しており、TF-IDF などの手法で離散をある程度抑えることが可能かもしれない。

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] 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

[3] 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