

Twitter タイムラインからの話題の抽出とその評価

星 皓介[†] 山田 剛一[†] 絹川 博之[†]東京電機大学大学院 未来科学研究科[†]

1. はじめに

近年、ソーシャルメディアサービスの発展により人々の情報発信をする場が急速に増えてきている。特に、そのひとつである Twitter[1] が大きな成長を見せている。

Twitter のサービスの特徴の 1 つに、タイムラインと呼ばれる、自身の発言およびフォローしているユーザの発言が表示される場がある。興味のあるユーザをフォローすることで、ユーザ独自のタイムライン(ホームタイムライン)を作ることができる。

タイムラインは積極的に情報を取得することができ、有用であると考えられる。しかしながら、ユーザの興味の広がりとはフォローするユーザの増加に伴い、タイムラインには多様な情報が現れるようになる。また、現れる多くの話題は一般的になり、本来のユーザがもつ興味から離れたものとなっている。

そこで我々は、話題ごとにツイートをまとめ、その中で読むべき話題を推薦することを考えた。しかし、話題の抽出を行う際、ツイートあたりの情報量が少なく、タイムライン上の話題を抽出できない問題が現れた。140 文字の字数制限や思ったことをすぐに投稿する利用のされ方が原因となっている。この問題を解決するために、ツイートの連結を行い、情報量を増加させる必要があると考えた。

本論文では、ホームタイムラインを対象に同一話題のツイートを連結させる手法を提案する。

2. 話題抽出

2.1 話題とは

話題は話の内容を表すため、名詞で表現される。本研究では名詞の抽出をベースとする。特に名詞の中でも、より具体的な表現である語ほど話題としてふさわしいとする。また、形態素解析の際、未知語となる語が存在するが、これらは記号を除き固有名詞として扱う。

2.2 話題抽出時のツイート分割問題

タイムラインから話題抽出を行った結果、ツイートの内容が分割されると、話題が正しく抽出できないことがあった。例えば、話題の主題を表す言葉を含むツイートと話題の詳細を表す言葉を含むツイートに分かれる場合が挙げられる。これらは話題を抽出する際に問題となる。

ツイートが分かれて現れるのは、2 つのケースが存在する。ひとつは、ユーザのツイートしたい内容が 140 字を超えるとき。もうひとつは、ユーザが思いついたことをすぐに投稿し、あとから訂正、追記をするときである。Twitter の特性上、後者のケースが多く存在する。

3. ツイート連結手法

2 章を踏まえ、同一話題のツイートの連結手法を提案する。基本的に、2 つのツイートを比較し、連結可能か判定する手法である。これらの手法は、適用できる範囲が異なり、発言者が同一ユーザのツイート間の場合と、異なるユーザのツイート間の場合とで別の基準が必要となっている。前者は前論文[2]で述べた手法に新たに加えたものを説明する。後者は主な手法を述べる。

3.1 発言者が同じ場合

3.1.1 リツイートによる話題の差し込み

リツイートとは、あるユーザのツイートを、自身のフォロワーと共有する機能である。リツイートは、自身で内容を考える必要がなく容易に行えるため、どのタイミングで現れるのか予測が難しく、リツイートとの結合は単純には行えない。リツイートが結合しやすいツイートには次の 2 つが見られた。ひとつは、リツイート後に時間を置かず投稿されるツイートである。これは、リツイートに対して言及する場合に現れるもので、言及したいツイートのテキストをコピーし、通常のツイートとして投稿する Quote Tweet (QT) と呼ばれる利用のされ方が存在することからも分かる。もうひとつは、短い時間で出現する元のツイートの投稿者が同一であるリツイートである。これらが存在した場合、リツイートと連結させる。

3.1.2 共通の一般名詞

共通する名詞がツイートに含まれている場合、連結可能なことがある。ただし、固有名詞ではない場合、一般的に用いられる語が多く、頻出する語には注意が必要である。共通する名詞がどの程度存在するときに連結させるか、ツイートのテキストに含まれる単語の数から判断させる。

3.2 発言者が異なる場合

3.2.1 会話

日常の会話と違い、Twitter での会話はある 1 つの事柄について話すことが多く、話題が転換することはあまりない。そのため、会話の一連のやり取りをひとつにまとめることが可能である。

3.2.2 共通の一般名詞

3.1.2 と同様に共通する名詞がツイートに含まれていた場合、連結可能とする。ただし、発言するユーザが異なる場合、その意味や対象が異なることが多いため、共通する名詞を含む割合がより高いツイートを連結可能とする必要がある。

Extraction of Topics on Twitter Timeline and Its Evaluation

[†] Kosuke HOSHI, Koichi YAMADA, Hiroshi KINUKAWA
Graduate School of Science and Technology for Future Life,
Tokyo Denki University

3.2.3 共通の固有名詞

ツイートを比較する際、共通する固有名詞が含まれていると連結可能なことが多い。ただし、固有名詞でも一般的によく使われる語は連結できるとは限らない。例えば、「日本」のような地域名がそれに当たる。地域名は非常に多く用いられ、また、何度も使われることがあるため、特別に扱う必要がある。3.2.2 共通の一般名詞と似た特徴であるが、固有名詞はより限定的に用いられるため、一致する固有名詞の数が少なくても、連結可能と判定させる。これは、Twitter の投稿文書が 140 文字以内であるため、一致する語に限界があるためだ。

例.

「エドウィンが倒産とはΣ(´-`)」

(2013-11-27 14:32:12)

「ネタかと思ったけどマジでエドウィンのサイトが503になった。」

(2013-11-27 14:37:03)

「エドウィンが倒産で、どこの会社だ？って思ったらEDWIN かよ！」

(2013-11-27 16:43:00)

4. ツイート連結実験

以上を踏まえ、5 名の被験ユーザのホームタイムラインを対象に、同一話題のツイートをまとめることができるか実験を行う。対象とするツイートは、話題をあらかじめ決めておき、タイムライン上にその話題が現れる範囲に含まれるツイートとする。そのため、ツイート数には被験ユーザと話題によりばらつきがある。実験回数は、被験ユーザごとに 3 回、合計 15 回となっている。

評価方法は、人手で作成した正解ツイート群とシステムが出力したツイート群とを比較する。その際、システムは同じ話題を含むツイートを別のグループに誤って分類することがあるため、出力されたどのツイート群と比較を行うか決める必要がある。この実験では単純に、正解に含まれるツイートを最も多く含むツイート群をシステムの比較対象とする。

4.1 結果

精度、再現率、F 値をそれぞれ求め、その平均を表 1 に示す。次に、発言者ごとの評価を行う。一人のユーザのツイートが話題ごとに連結できるかを評価するものである。対象ツイートと評価方法は上記のものと同様である。結果の平均を表 2 に示す。

表 1 連結実験の結果—平均

精度	再現率	F 値
0.864	0.464	0.661

表 2 発言者ごとの結果—平均

精度	再現率	F 値
0.902	0.788	0.824

4.2 考察

表 1 の結果より、再現率が特に低いことが分かる。これは、他のユーザのツイートとの連結に失敗することが原因となっており、表 2 では高い値を示していることから分かる。同一ユーザのツイート連結と異なるユーザのツイート連結では、ツイートの時間間隔を用いているかが大きく異なり、実験からも分かった。また、評価方法も原因となっている。今回用いた評価方法では、話題を含むツイートがちょうど 2 つに分類された場合、再現率が 0.5 と評価されてしまう。本来ひとつの話題として分類されるはずのツイートが、複数の話題として分類された場合でも評価可能な方法を用いる必要がある。

精度が低い原因は、異なる話題間でそのどちらにも属するツイートと結合可能なツイートが存在するとき、ひとつのツイートが複数の話題を結びつけてしまうからである。話題に属するすべてのツイートとの近さを測り、しきい値を用いて分類することも考えられるが、ひとつのツイートが話題を結びつける手がかりとなることもあるため、どちらが良いか一概には言えない。

また、短時間に複数の話題を含む多量の投稿をするユーザのツイートに、連結誤りが多く見られた。これは、連結の判定を時間に大きく依存しているのが原因となっており、話題の切り替えが早いユーザの特徴である。こういったユーザにはツイートの時間間隔を適応させるのが難しいため、ツイートの投稿量から判別し時間間隔を用いない方法が考えられる。同様の問題として、ニュースサイトの記事を連続でツイートすることで、起きるケースも見られた。これらのツイートの多くはツイートボタンから投稿されており、どの媒体から投稿されているかを調べることで、判別可能であると考えられる。

5. おわりに

本論文では、Twitter における同一話題のツイートの連結手法をいくつか提案した。実験の結果、同一ユーザのツイートの連結では、比較的高い値を示したが、異なるユーザのツイート間では値は低く、特に再現率は 0.464 と非常に低い値となった。異なるユーザのツイートに対して、有効な手法が必要であった。

参考文献

- [1] Twitter : <http://twitter.com/>
- [2] 星皓介, 山田剛一, 絹川博之, “Twitter における同一話題のツイートの連結と話題抽出”, 情報科学技術フォーラム講演論文集, pp.159-160, 2013.