

Twitter からの同意・非同意表現抽出

大野 正樹

村上 明子

日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

ソーシャル・ネットワーキング・サービス (Social Networking Service, 以下 SNS と呼ぶ) 等の普及により, 人々がネットワーク上で議論する機会が増加した. ネットワーク上での議論が活発になると, 人手でその全てを把握することが難しくなるため, それらを自動的に分析する技術が重要となる. 本稿では, ネットワーク上での議論をオンラインディスカッションと呼ぶ. オンラインディスカッションの各発言は返答関係に基づく木構造で整理できるという特徴を持つ.

議題に対する参加者の態度を判別する際に, 「賛成」や「反対」などの同意・非同意表現が重要な手がかりとなる. これらの表現は発言者の意見を直接・間接的に示すため, オンラインディスカッションを分析する際に利用することができる.

同意・非同意表現を網羅的に用意し, 発言中にこれらの発言が含まれているか判定することで, 議題に対する参加者の態度を判別することは可能だが, ある表現を手手で網羅することは難しい. 提案手法はテキストデータからこれらの表現を自動獲得することで, これらの表現の獲得を支援することを目的とする.

本稿では返答関係を持つテキストから同意・非同意表現を獲得することを試みる. 返答構造を持つテキストから同意・非同意表現を自動獲得する場合には, 同意・非同意表現が出現する箇所を特定することと, そこからそれらの表現を獲得することが必要とされる. 提案手法は典型的な同意・非同意表現を手手で定め, それらと返答構造により同意・非同意表現が出現する箇所を特定する.

はじめに, 典型的な同意・非同意表現を手手で定める. 次に, ある発言に対する返答であり, なおかつこの表現を含む発言を用意したデータから獲得する. さらに, 各々の発言と兄弟関係にある発言を獲得する. 最後に形態素パターンと C-Value を用いてその発言に頻出する語を獲得する.

提案手法の有効性を評価するために, Twitter と

IBM 社内のディスカッションである InnovationJam¹ から得たデータをもとに評価実験を行った. Twitter から獲得した同意・非同意表現の正解率は 0.40 であり, InnovationJam から獲得した同意・非同意表現の正解率は 0.80 であった. 実験から同意表現と非同意表現ではその使われ方が異なることが分かった. 例えば, 実験に使用したデータでは, 発言者は同意する場合には明確に同意を表し, 同意しない場合には非同意表現は使わずに同意しない理由を述べるケースが多かった.

2 関連研究

提案手法は同意・非同意表現を網羅的に用意するために, 返答関係を持つテキストから同意・非同意表現を獲得する. 提案手法はテキストデータを用いて自動的に語彙を拡張するという側面と, 返答構造を持つテキストから何らかの情報を抽出するという側面を持っている. ここではその 2 つの側面のそれぞれに関連する手法を挙げる.

テキストデータを用いて自動的に語彙を拡張する研究として, Wang らの手法 [7] と He らの手法 [4], Talukdar らの手法 [6] を挙げる. Wang らは Web 上のテキストを使って語彙を拡張した [7]. また, He らは Web 上のテキストと検索エンジンに入力された検索語のログデータを使って語彙を拡張した [4]. Wang らの手法と He らの手法はマークアップ文書を対象にしているため, フリーテキストを対象にした提案手法とはその目的が異なる. Talukdar らはフリーテキストを対象にテキストに頻出するパターンを獲得し, 語彙を拡張した [6]. この手法は特定の表現ではなく単語を拡張することを目的としている.

返答構造を持つテキストから何らかの情報を抽出する研究として, Brody らの研究 [2] と羽鳥らの研究 [8] を挙げる. Brody らは Twitter に出現する Cooool の様な長音化した言葉をもとの言葉に戻し, 次に, その言葉の極性を推定した. 長音化した言葉は Twitter

¹<https://www.collaborationjam.com>

上でよく使用されているが正規の辞書には載っていないため、これらをもとの言葉を推定することは重要である。Brody らの手法は Twitter に出現した語にのみ着目しており、返答構造の情報を使用していない。提案手法は返答構造に着目し、Twitter から同意・非同意表現の獲得を試みる。羽鳥らはオンラインディスカッションの議論の内容を概観するために、オンラインディスカッションの発言から重要文を抽出した。この手法は返答構造に着目しているという意味で提案手法に似ているが、議論の要約を目的としている。

3 提案手法

本稿では返答関係を持つテキストから同意・非同意表現を獲得することを試みる。返答構造を持つテキストから同意・非同意表現を自動獲得する場合には、同意・非同意表現が出現する箇所を特定することと、そこからそれらの表現を獲得することが必要とされる。提案手法は典型的な同意・非同意表現を手で定め、それらと返答構造を用いて同意・非同意表現が出現する箇所を特定する。

3.1 返答構造に着目した探索範囲の絞り込み

効率良く同意・非同意表現を獲得するために、それらが多く出現する箇所を特定する。探索範囲を絞り込むために、本稿では同意・非同意表現の出現に関して下記の仮定をおく。

1. 典型的な同意・非同意表現を含む発言は何らかの議論に関わる発言である可能性が高い。
2. その発言と兄弟関係にある発言も議論に関わる発言である可能性が高い。従って、その発言は同意・非同意表現を含む可能性が高い。

この仮定に従って探索範囲を絞り込むことで、探索の効率を上げることができる。探索範囲を絞り込む手順を下記に示す。

1. 「賛成」や「反対」などの典型的な同意・非同意表現を手で定める。
2. ある発言に対する返答であり、なおかつそれらの表現を含む発言 $T = \{T_1, T_2, \dots, T_n\}$ を用意したデータから獲得する。

3. 各々の発言と兄弟関係にある発言 $Tr_T = \{Tr_{T_1}, Tr_{T_2}, \dots, Tr_{T_n}\}$ を獲得する。

図 1 に返答構造の例を示す。ここでは T_1 に対する返答と T_5 に対する返答がある。 T_6 が典型的な同意・非同意表現を含んでいるため、 T_5 と T_5 に対する返答が何らかの議論に関わる発言である可能性が高いとみなし、探索の対象とする。

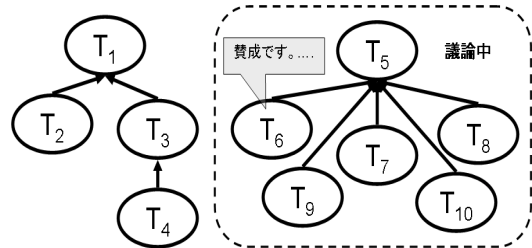


図 1: 発言の返答構造の例

3.2 表現の獲得

探索の対象のテキストから同意・非同意表現を獲得する。仮定 (2) から探索の対象のテキストに頻出する表現を同意・非同意表現とみなす。

始めに同意・非同意表現として、「人称代名詞 M^* 動詞 M^* 」という形態素パターンを仮定し、このパターンに当てはまる文字列群 W を同意・非同意表現の候補として抽出した。 M^* はワイルドカードであり、任意の形態素が該当する。

次に文字列群 W の要素 w を C-Value[3] によってスコア付けする。C-Value を求める式を下記に示す。

$$C\text{-value}(w) = \log(|w|) \cdot f(w) - \frac{1}{|S_w|} \sum_{v \in S_w} f(v)$$

$|w|$ は w の長さを、 $f(w)$ は w の出現回数を表している。また、 S_w は w の部分文字列全体を表している。

最後に、C-Value の高い文字列を同意・非同意表現とみなす。より多くの同意・非同意表現を獲得するならば、ここで得られた表現を典型的な同意・非同意表現に加えて、提案手法を対象のテキストに適用する。

4 評価実験

4.1 実験概要

提案手法の有効性を確認するために、評価実験を行った。実験は提案手法によって獲得した n 個の表現が

同意・非同意表現であるか、その正解率を測るものである。ある表現が同意・非同意表現であるかの判断は人手で行った。

実験では返答構造を持つテキストとして、Twitter と InnovationJam のデータを用いた。InnovationJam とは IBM 社内のオンラインディスカッションであり、発言には発言者の名前などの社内プロフィールがつく。実験では 2008 年に投稿された 28,896 件の発言を使用する。Twitter のデータとして、2012 年 1 月 1 日から 1 月 11 日までに投稿された 18,692,185 件の発言を使用する。このデータは Twitter Stream API²によって集めた。

InnovationJam はその大部分が英語で構成されている。Twitter を知識源として同意・非同意表現を獲得した場合と比較が行いやすいように、典型的な同意・非同意表現として {agree, amazing, good, disagree, bad, awesome} を選び、英語の同意・非同意表現を獲得することにした。

実験に使用するデータから Twitter 特有の情報であるハッシュタグやユーザ ID などを取り除いた。Twitter 上には多くのスパムが存在することが指摘されているが、それらの影響を考察するために、収集したデータからスパムを取り除くことをしなかった。

4.2 実験結果

n を変化させたときの、Twitter と InnovationJam のデータから同意・非同意表現を獲得した際の正解率を表 1 に示す。どちらも $n = 10$ としたとき正解率が最も高く、InnovationJam から獲得した同意・非同意表現の正解率は 0.80 であり、Twitter から獲得した同意・非同意表現の正解率は 0.40 であった。また、どちらも n の値を大きくするにつれて正解率が下がった。InnovationJam を知識源とした場合の方が、Twitter に比べて効率的に同意・非同意表現を獲得できる。しかし、Twitter を知識源とした場合の正解率は 0.40 であることから、同意・非同意表現を獲得する知識源として Twitter が有用であることが示せた。

実験の際には、Twitter 上のデータからスパムを取り除かずに実験データを作成したが、結果を見るとスパムの影響は見られなかった。典型的な同意・非同意表現と返答構造を用いた探索範囲の絞り込みが効果的であったと考えられる。

獲得した同意・非同意表現の例を図 2 に示す。InnovationJam で使用される同意・非同意表現と、Twitter

表 1: 獲得した同意・非同意表現の正解率

	Twitter	InnovationJam
$n = 10$	0.40	0.80
$n = 15$	0.33	0.67
$n = 20$	0.25	0.70

表 2: 獲得した同意・非同意表現の例

Twitter	InnovationJam
I agree	I like the idea
I 'm sure	you are right
I don't think	I also think

で使用される同意・非同意表現ではその表現の性格が異なった。これはオンラインディスカッションの形態の影響によるものと考えられる。InnovationJam は社内のディスカッションであるため参加者はフォーマルな表現を使うが、Twitter は友達間の議論であるためくだけた表現が使用される。この結果は、オンラインディスカッションを分析する際には、その形態に併せて言語資源を用意しなければいけないことを示唆している。

4.3 考察

4.3.1 オンラインディスカッションの特徴

実験データでは、同意する場合には明確に同意を表すが、同意しない場合には非同意表現を使わずに同意しない理由を述べるケースが多かった。そのため、非同意表現は同意表現に比べて出現頻度が少なかった。この理由として、オンライン上の議論では相手の顔が見えないため、ユーザが相手に対するネガティブな内容の発言を避けていることが考えられる。多くのユーザーに発言批判されたユーザがその直後に Twitter を退会したケースが今回の実験データにあった。

Twitter ではユーザは非公式 RT を用いて意見を述べることもある。非公式 RT とは、「Happy new year,too! :) RT @XXXX happy new year :)」のように、あるユーザーの発言を含んだ発言を投稿することである。実験に使用したデータを調べたところ、非公式 RT のうちおよそ 23%がなんらかのコメントを含んで発言していた。今回の実験では、Twitter では非公式 RT を発言に対する返答とみなさなかったが、これも返答で

²<https://dev.twitter.com/docs/streaming-api>

Twitter には “you are” が “u r” と表記するなど、
 くださった表現が多かった。この表現は非常に多様であり、
 これを上手くまとめることができなかったことが Twitter を知識源とした場合の正解率が低い原因である
 とが考えることができる。また Brody らが指摘するように
 Cooooool の様な長音化した言葉も多く出現する [2]。Twitter から効率的に同意・非同意表現を獲得する
 場合、これらのくださった表現をもとの言葉に戻す必要がある

実験に使用したテストデータを眺めると、言語によって同意・非同意表現の出現頻度が異なることが分かった。英語の発言に比べて、日本語の発言には同意・非同意表現の出現回数が少ない。これらは言語の語彙の大きさによって起きるものではなく、言語を使用する人々の習慣あるいは Twitter の使い方によるものであると言える。

本稿では、「賛成」や「反対」などの同意・非同意表現を返答構造を持つテキストから自動獲得する方法を述べた。提案手法は人手で定めた典型的な同意・非同意表現と返答構造により同意・非同意表現が出現する箇所を特定した。

参考文献

- Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1576, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

- [2] Samuel Brody and Nicholas Diakopoulos. Coooooooooooooooooolllllllllllll!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 562–570, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.
- [3] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, Vol. 3, No. 2, pp. 115–130, 2000.
- [4] Y. He and D. Xin. Seisa: set expansion by iterative similarity aggregation. In *Proceedings of the 20th international conference on World wide web*, pp. 427–436. ACM, 2011.
- [5] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600. ACM, 2010.
- [6] P.P. Talukdar, T. Brants, M. Liberman, and F. Pereira. A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pp. 141–148. Association for Computational Linguistics, 2006.
- [7] R.C. Wang and W.W. Cohen. Language-independent set expansion of named entities using the web. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pp. 342–350. IEEE, 2007.
- [8] 羽鳥潤, 村上明子. スレッド構造と語彙的連鎖を用いたオンラインディスカッションからの重要文・トピックの抽出. In 言語処理学会第16回年次大会発表論文集 (NLP2010). March 2010.