

第1章

情報処理学会 第80回全国大会

投稿論文

情報処理学会 第80回全国大会 (The 80th National Convention of IPSJ) に投稿した論文を示す.

分散表現を用いた話題変化判定

芳野 魁^{1,a)}

受付日 2015年3月4日, 採録日 2015年8月1日

概要: 大規模意見集約システム COLLAGREE ではファシリテーターと呼ばれる人物が議論のマネジメントを行っているが、長時間に渡って大人数での議論の動向をマネジメントし続けるのは困難である。ファシリテーターが画面を見なければならない時間を減らし、負担を軽減する工夫があることが望ましい。ファシリテーターが画面を見るべきタイミングは議論の話題が変化したときであると考えられる。すなわち、ファシリテーターの代わりに自動的に議論中の話題の変化を観測することが必要である。本研究ではファシリテーターに話題の変化を伝えるために、分散表現を用いて発言の繋がり度合いを数値化し、繋がり度合いが小さいもの、即ち話題の変化を起こしうる発言を検出することを目指す。

キーワード: 自然言語処理, 議論支援, 話題遷移, COLLAGREE, 分散表現

A Topic Change Judgment Method based on Distributed Representation

KAI YOSHINO^{1,a)}

Received: March 4, 2015, Accepted: August 1, 2015

Abstract:

Keywords: NLP, discussion support, topic shift, COLLAGREE, word embedding

1. はじめに

近年、Web 上での大規模な議論活動が活発になっているが、現在一般的に使われている “2ちゃんねる” や “Twitter” といったシステムでは整理や収束を行うことが困難である。困難である原因として議論の管理を行う者がいないことを挙げることができ、「炎上」と呼ばれる議論の無秩序の状態が頻繁に観測されている。つまり、議論を整理・収束させるには議論のマネジメントを行う人物が必要である。大規模な意見集約を目的とした大規模意見集約システム COLLAGREE²が開発された。COLLAGREE では、掲示板のような議論プラットフォームをベースにしており、自由に意見を投稿することができる。COLLAGREE では議

論を秩序的に進行し、収束させるためにファシリテーターと呼ばれる人物が議論のマネジメントを行っている。しかし、ファシリテーターは人間であり、長時間に渡って大人数での議論の動向をマネジメントし続けるのは大きな負担がかかり困難である。

COLLAGREE で大規模かつ長時間の議論を収束させるためには、ファシリテーターが必要な時には画面を見るようにして、他の時は見なくても済むようにすることで画面に向き合う時間を減らす工夫があることが望ましい。ファシリテーターが画面を見るべきタイミングは議論の話題が変化する、または話題の変化を起こしうる発言が投稿されたときである。以前の議論の内容から外れた発言が投稿された時、ファシリテーターが適切な発言をすることで、脱線や炎上を避けて議論を発展させながら収束させることができる。すなわち、ファシリテーターの代わりに自動的に議論中の話題の変化を判定することが求められている。

現在、COLLAGREE 上で使用されている議論支援機能で

¹ 情報処理学会
IPSJ, Chiyoda, Tokyo 101-0062, Japan

^{†1} 現在、名古屋工業大学
Presently with Nagoya Institute of Technology

^{a)} joho.taro@ipsj.or.jp

はファシリテーターの作業量の減少には繋がりにくい。

近年、単語を実数ベクトルで表現する技術である分散表現は自然言語処理の分野において多くの研究で使われており、機械翻訳を始めとする単語の意味が重要となる分野で精度の向上が確認されている。分散表現を用いることで、従来の手法より単語の意味を考慮した処理が可能になり、内積計算を用いることで単語間の類似度を計算できる。

本論文では、新たに投稿された発言を調査し、分散表現を用いてファシリテーターの代わりに自動的に話題の変化を判定する手法を提案する。具体的には新しく投稿された発言を過去に投稿された発言と1つ1つ比較して類似度を計算する。類似度の計算は三段階で行われる。全ての過去の発言との総合類似度を計算し類似した発言があるかどうかで話題の変化の判定を行う。COLLAGREE上で実際に行われた議論データを対象に評価実験を行う。提案手法がファシリテーターの代わりの自動的な話題の変化判定において有用であることを示す。

本論文の構成を以下に示す。2章ではCOLLAGREEの概要と議論支援及び本研究で支援することを目標とするファシリテーターについて説明し、関連研究についても説明するとともに本研究との違いを述べる。更に本研究の重要な要素である重み付けと分散表現についても詳しく述べる。次に、??章では話題変化判定システムの全体モデル説明を行い、??章では分散表現を用いた発言内容の類似度計算について説明する。そして、5章では話題の変化判定の評価実験について説明する。最後に??章で本論文のまとめと考察を示す。

2. 関連研究

議論支援を行う研究では小谷ら?は好意的発言影響度を取り入れた議論支援システムを開発した。システムは議論中の発言の意図や内容に加えて、発言に対するリアクション(同意、非同意、意見)などから議論進行をモニタリングしている。モニタリングの結果を基にして議論の活性化や深化に対して参加者が果たしている役割を”好意的発言影響度”として定量化して表示する。小谷らは一般参加者や学習者の議論活性化及び収束に向けた支援を目的としているが、本研究はファシリテーターに対する支援を目的としている点で異なる。

話題遷移を扱う研究では別所?は単語の共起頻度行列を特異値分解で次元を削減して作成した単語の概念ベクトルを用いて、連結された新聞記事を元の形になるように分割をする実験を行っている。テキストセグメンテーションの1手法であるTextTilingを用いてブロック間の類似度を計算する際にブロック中の単語の中で自立語にのみ概念ベクトルを付与し、左右のブロックの和ベクトル(または重心ベクトル)の余弦測度を求め、類似度(または結束度)としている。テキストセグメンテーションは基本的にある地点

Aで話題に沿って分割をするか考える際に、地点Aより先の情報を使うことができる。すなわち、ある程度の文章が現れてから話題が変わったかどうかの判定をしており、リアルタイムでの動作を想定していない。本研究はリアルタイムでの動作を想定し、ある発言Bが話題の変化を起こすかどうか判定する際にBより先の情報を使うことなく判定している点で異なる。中村ら?は連結された新聞記事をLDAを用いてトピック変化点を検出して分割する実験を行った。中村らは次の手順でトピック変化点を検出した。

- (1) 現在地点より前の L_0 個の形態素を処理対象範囲 $B_0(L_0 \leq L : L = \text{文章長の下限})$ とし、 B_0 から複数のトピック変化点候補箇所を抽出する。
- (2) B_0 を各トピック変化点候補箇所 で2ブロックに分割し、2ブロック間の類似度を計算する。そして類似度が最も小さくなるトピック変化点候補箇所 D_1 とその類似度 S_1 を求める。文書ブロック間の類似度はLDAによって求められるブロックのトピック混合比ベクトルを用いて算出する。
- (3) B_0 を D_1 で分割し、現在地点に近い側のブロックを B_1 とする。
以降は以下4.と5.の処理を繰り返す ($n \geq 2$)。
- (4) ブロック B_{n-1} に対し2.と同様の手順でトピック変化点候補箇所 D_n とその類似度 S_n を求める。
- (5) $S_n \geq S_{n-1}$ の場合、 D_{n-1} をトピック変化点として処理を終了する。 $S_n < S_{n-1}$ の場合、 B_{n-1} を D_n で分割し現在地点に近い側のブロックを B_n として、処理を継続する。

なお、上記(1)におけるトピック変化点候補は1文中でトピックが変化することは考えにくいことと計算コストを考慮し、文境界(句点出現位置)をトピック変化点候補としている。

トピックモデルはトピックに基づいて学習を行っていて、基本的に学習の際にトピックの数を指定する必要があり、指定に伴いトピック数が制限されてしまう。また、教師なし学習であるので必ずしも話題を捉えた学習がされるとは限らない。すなわち、トピックの数によっては変化に対応できない話題が存在することがあり得る。本研究はトピックの指定が必要なLDAによるトピックモデルとは違い、トピックに関係なく単語の共起頻度を学習するfastTextを用いている点で異なる。更に、本研究は中村らの研究とは話題の変化点を判定する際に複数候補から類似度が最小である候補点を選ぶのではなく、候補点を逐一調べ閾値を下回るかどうかで変化点を判定している点や対象とするデータが新聞記事ではなく議論である点でも異なる。