1P-6

時系列テキストデータからの議論構造の可視化

平岡　美那子†　　大澤　幸生†

東京大学工学系研究科システム創成学専攻†

１．はじめに

我々人間は議論を通じて集団での意思決定を行う．過去に行われた議論の内容は，関係者が同じ知識を共有するために，議事録として保管されている場合が多い．議事録から Toulmin の議論モデルのような理論的な議論構造が抽出できれば，議論が不十分であった点を発見し再び議論する等，様々に活用することができる．しかし，自然言語で記述された個々の発言を自動的に厳密な議論モデルに変換することは難しく，また，現実の議論は，参加者それぞれの思考プロセスの違いにより論点のずれが生じ，そもそも議論モデルに則していない場合がある．そこで，本研究では，個々の発言に発言時間が付与されているような議事録について，出現する単語の特徴度を定義することで，ラフな議論構造を抽出し，話題の推移を可視化する手法を提案する．本手法を用いれば，議論の流れを議事録よりもすばやく把握することができ，発言者ごとの思考プロセスを読み取ることができる．以下の章で関連研究および本手法の詳細な説明と評価実験について述べ，考察を行う．

２．関連研究

議論支援システムや議事録の可視化に関連する研究はこれまでにも盛んに行われている．文献[1]では，議論の構造化・可視化を手助けするツールを用いて協調学習の効果を高める手法等について述べられている．また，堀田ら[2]はランダム・ツリーを利用して政策論議を構造化する手法を提案している．しかしどちらも，自動的に議論を構造化させるにはいたっていない．

松村ら[3]は，影響普及モデルを用いて，議論の発展のトリガを発見する手法を提案している．また，白松ら[4]は，長い議事録を話題遷移図として簡潔に提示できる手法を提案している．しかしこれらの手法は，発言者の違いを考慮していない点で，本手法とは異なっている．

Argumentation Visualization from time series text data
†Minako Hiraoka, †Yukio Ohsawa, Department of System Innovations, School of Engineering, The University of Tokyo

３．提案手法

本研究では，議事録に出現する語の特徴度を定義し，それをもとに，ラフな議論構造を決定する．ここで，所与の議事録データは，各発言 Si(i=1,…,n)について，発言時間 Ti と発言者の情報が分かっているものとする．

（１）単語文書行列の作成

発言の集合である議事録データを，単語文書行列 M に変換する．M の成分(j,i)は，単語 Wj の発言 Si における出現頻度を表す．

（２）語の特徴度の計算

議事録に出現する各単語 Wj について，議論の中での位置づけを定量化するために特徴度 Cj を計算する．本手法では，単語の出現頻度を時系列データとしてみたときの標準偏差の値を特徴度として用い，語の出現時間の平均を代表時間 Ai として用いる．計算式を以下に示す．

$$C_j{}^2 = \frac{1}{N_j}\sum_i (T_i \cdot M_{ji} - A_j)^2$$

$$A_j = \frac{1}{N_j}\sum_i T_i \cdot M_{ji}$$

（ただし $N_j = \sum_i M_{ji}$）

ここで，特徴度が高い語とは，特定の時間にだけ頻繁に出現し，トピック性が高いということを意味する．

（４）複数ピークのクラスタリング

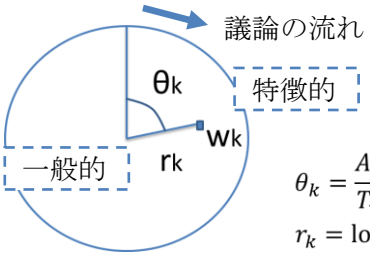
長い議論の中では，話題が循環し，出現頻度が複数回のピークを持つ語も存在すると考えられる．そのような語を，１つの特徴度・代表時間で表すのは適切ではない．そこで、全単語について混合ガウス分布を仮定した EM アルゴリズムによるクラスタリング[5]を適用し，ピークが複数検出される場合は，ピーク間の中間位置で出現頻度データを切り離し，異なる単語として処理を行うようにする．

（４）ラフな議論構造の可視化

最後に，ラフな議論構造マップを構築し可視化することを考える．議事録に含まれている全単語を表示することは現実的でないので，特徴度の高い順に K 個の単語を選択し表示することとする．はじめに，選択された単語

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

Wk(k=1,…,K)について，議論中の位置づけを読み取れるように，特徴度 Ck と代表時間 Tk を用いて図１のように配置する．



$$\theta_k = \frac{A_k}{T_n} \cdot 2\pi$$

$$r_k = \log(C_{max} - C_k)$$

図 1 単語のマップ上の位置づけ

次に，ラフな議論構造を，単語間の共起度をもとに構築する．ここで，各発言の中で最も特徴度の低い単語，すなわち一般的と考えられる単語をベース単語とおき，図２のような議論モデルを考える．この時，（１）同一発言内に含まれるベース単語と，それ例外の単語間（２）同一人物の発言のベース単語間，の２種類の間には，議論上の繋がりがあると考え，リンクを張ることとする．但し、ベース単語およびリンクは表示されている単語の間で計算する．

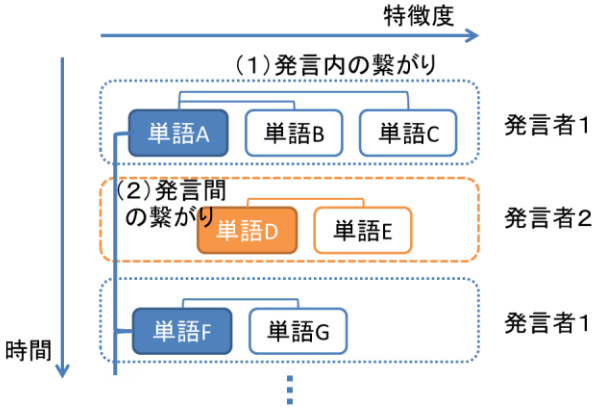


図 2 ラフな議論構造

４．実装と評価

上述の手法に基づき，議論可視化ツールを実装した．日本語処理には cabocha を使い，混合ガウス分布によるクラスタリングと図のプロットには R を用いた．2004 年 11 月 10 日に行われた自民両代表の党首討論の議事録[6]を用いて可視化を行った．この議事録には発言時間の情報がなかったため，議事録における改行を発言の区切りとし，発言時間を最初から順番に 1，2…とした．要素として抽出する単語は名詞のみとした．抽出された要素（名詞）のうち，数詞や代名詞と，全体を通して出現頻度が 3 以下であるものを重要語ではないと判断して除外した．全発言数は 102，ノイズ除去後の総語数は 207 であった．本手法で特徴語 50 語を抽出し可視化させた結果を表１および図 2 に示す．図 2 から，議論の中で出てきた話題に関連する語のクラスターが形成され，話題が推移している事がわかる．また，２人の発言者の使用単語やリンクの結びつきの違いから，２人が同じ論点について議論を戦わせている部分と，議論がすれ違っている部分が分かりやすく表示されている．

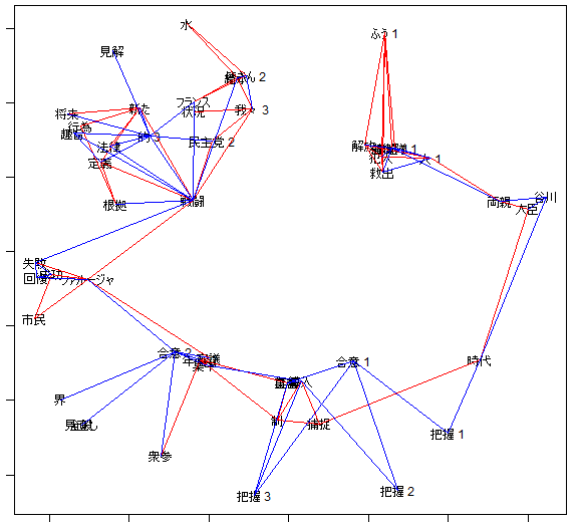


図 3 党首討論データの可視化結果
（赤：自民党，青：民主党）

５．おわりに

発言がなされた時間情報と発言者情報を利用して，ラフな議論構造を可視化する手法を提案した．今後は，この図をベースにして，過去の議事録からよりよい議論を発展させていくための手法についても研究していきたい．

参考文献

[1] Paul A. Kirschner, Simon J. Buckingham Shum and Chad S. Carr (Eds.): Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer-Verlag: London, 2003

[2] 堀田昌英, 榎戸輝揚, 岩橋伸卓: 多元的議論構造の可視化手法:社会技術としての政策論議支援, 社会技術研究論文集 1, 67-76, 2003

[3] 松村真宏, 加藤優, 大澤幸生, 石塚満: 議論構造の可視化による論点の発見と理解, 知識と情報：日本知能情報ファジィ学会誌, 554-564, 2003

[4] 白松俊, 駒谷和範, 尾形哲也, 高橋徹, 奥乃博: 話題遷移図の可視化と話題遷移に応じた関連情報提示による議事録閲覧支援, 言語処理学会第 15 回年次大会, D2-1, 2009

[5] Jeff A. Bilmes: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, International Computer Science Institute 4.510, 126, 1998

[6]国会議事録検索システム, http://kokkai.ndl.go.jp/, (2014-1-13 アクセス)

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.