3L2-R-12-5

話題の結晶化による論点発見支援

The arguing point discovery support with visualization of the relationship between topics

藤塚 理[*1]
Osamu Fujitsuka

大澤幸生[*1]
Yukio Ohsawa

[*1] 東京大学工学系研究科
School of Engineering, the University of Tokyo

In this paper, we tried to enable people to find the chance from dialogue logs. In our method we divide dialogue logs to appropriate segments based on vocabularies co-occurrence and visualize the relationship between the segments. To find the important relationship we focused on mutual information which can indicate chance points. Our goal is not only to provide deeper understandings on the topics but also to find elusive concepts included in the topics. We analyzed the dialogue log by our system and succeeded in finding the elusive concepts.

1. はじめに

今日，会社などの組織だけではなく学校の授業や研究室であっても，新しいアイディア創出や一つのテーマに対する様々な視点を学ぶために，人との対話や議論は欠かせないものとなっている．しかし対話から新たな発見を見出すことは必ずしも容易なことではなく、対話ログの分析研究はいくつか行われきた．

前野や瀬々は，共起度に基づいて単語クラスタリングを用いて対話ログを視覚化し，視覚化したグラフから話題を抽出している[前野 07][瀬々 07]．さらにデータ結晶化[前野 07]を用いてクラスタ間の関係を掴んでいる．斉藤は時系列で区切り話題を出現クラスタとして獲得し，その推移を視覚的に表示することで対話ログの概要把握の支援を行った[斉藤 10]．これらの手法においては，単語のクラスタリングをする際には話題数や話題内容を事前に分かっていることが必要であり、複雑な話題が混在する対話を対象にした支援としては不十分である．

松村による構造化マップ[松村 03]では，議事録から話題の単位を同定し，同定した話題単位間の関連から話題遷移をフローチャート形式で構造化し，読者が話題の展開を把握できるよう支援している．しかし議論全体の流れを読者が把握することを目的としていること，時系列で離れた話題間の関係性が視覚化されないことなどの点で新たな発見を促すには不十分である．

そこで本稿では，対話ログから話題の単位を同定し話題の遷移を視覚化するとともに，話題間の関係性を相互情報量に基づいて提示することにより、新たな論点の発見の支援を目指す．

2. 提案システム

本章ではシステムの概要と提案手法について説明する．

2.1 システム概要

システム概要について示す．以下の流れに沿って行う．全体の流れを図 1 に示す．

(1) 形態素解析

まず本文を奈良先端科学技術大学院大学で開発された MeCab を用いて形態素解析をする．形態素解析とは，言語で意味をもつ最小単位に分割することである．ここでは名詞のみを取り出した．

連絡先: 藤塚理，東京大学大学院工学系研究科，〒113-8656
東京都文京区本郷 7-3-1，8254197611@mail.ecc.u-tokyo.ac.jp

(2) 共起度測定

本文を任意の大きさに区切り隣り合ったセグメント間の共起度を計算する．共起度計算には cosine 値を用いた．共起度計算の前に、tf-idf 法を基にした計算式(i)

$$\mathrm{Score}(x,w) = \mathrm{Freq}(x,w) \times \{\log_2(allfreq/freq(x)) + 1\} \quad \text{(i)}$$

但し、Freq(x,w) : 単語 x の話題 w 内での頻度
allfreq : 全文中での全単語頻度
freq(x) : 全文中での単語 x の頻度

により求めた値を成分とする各セグメントの特徴ベクトルを算出する．上位数単語（指定可能）の特徴ベクトルとする．

(3) 話題の同定

共起度の高い隣り合ったセグメント間を結合していく．指定の数のセグメント数になるまで結合を繰り返すことで，話題の切れ目を同定する．時系列順に、話題 1、話題 2、話題 3、・・・と名前をつけていく．

(4) 話題の遷移の視覚化

同定した話題間の共起度(cosine 値)を基にして，まずは話題の遷移を視覚化する．具体的には、共起度が閾値より高ければ対話間に関係があるとみなして時系列を考慮して矢印を引く．対話は時系列に沿って進行しており，時間的に離れた話題間で話題の遷移はしないと考えるのが妥当である．よって，時間軸で比較的近い話題間の遷移のみに注目して視覚化をする．

(5) 話題間の関係性の視覚化

時間軸では離れており話題の遷移は考えづらいものの意味のあるつながりを，相互情報量(*後述)を用いて視覚化する．相互情報量は単語間に定義される量である．話題 x に含まれる単語群$\{X_i\}$と話題 y の単語群$\{Y_j\}$の相互情報量は、単語X_iと単語Y_jの相互情報量を mutual(i,j)として，

$$\mathrm{MI} = \sum_i \sum_j mutual(i,j)$$

として計算する．MI 値が高い話題同士を(4)で描いたグラフに、リンクとして結びつける．

(6) 話題ごとの特徴単語の提示

話題ごとの特徴語を tf-idf 法を基にした計算式(i)により特定し，提示する．特徴語を提示することで，話題内容や話題間の関係性の理解促進を促す．

2.2 相互情報量

(1) 相互情報量

相互情報量は事象 A と事象 B, その生起確率 p(A)と p(B)を考えた時に, 以下の式に表される量である.

$$\mathrm{mutual}(A, B) = \log p(A,B) - \log p(A) - \log p(B)$$

相互情報量はある語が共起相手の事象の情報をどの程度持っているかを示す指標である. しかし相互情報量がどのような意味を持つかは様々な解釈が可能である.

(2) チャンス発見学における相互情報量

ここではチャンス発見学における解釈を説明する. チャンス発見学においては, 二つの事象に共通要因があり, それが珍しい事象である場合に相互情報量が高くなると解釈する[大澤 06] . 単純化して事象 A は原因 X と原因 Y から, 事象 B は原因 Y と原因 Z から生起しており, X, Y, Z は互いに独立であると考えると, この時,

$$p(A) = p(X)\,p(Y),\ p(B) = p(Y)\,p(Z),\ p(A, B) = p(X)p(Y)\,p(Z)$$

となり,

$$\mathrm{mutual}(A, B) = -\log(Y)$$

と変形できる. つまり共通要因となっている原因 Y は生起確率が低く珍しい事象の時に相互情報量は高くなり, 相互情報量は稀である共通要因が存在を示唆している. 相互情報量の高い話題同士を提示することにより, 気づきにくい共通要因の発見を促すことが出来る.

3. 提案システムの適用例

本章では提案システムを対話ログへ適用した例を示す.

3.1 適用した対話ログ事例

対話ログとして, 次世代の ICT を使った教育についての二人による対話を使用した. ICT と教育という大枠はあったものの, 自由に議論をしている. 全 23680 字、324 発言の対話であった.

3.2 対話ログへのシステム適用

対話テキストを適用し作成した GUI のイメージを図 2 に示す. 右にシステムにより作成したグラフを, 左に各話題の内容が理解しやすいように特徴単語を表示する.

3.3 話題の推移の検討

図2に話題の流れが示されている. グラフは大きな二つの流れがグラフ下部で合流する形となっているが, 実際の議論においても「日本とアメリカの教育に関する話」と「情報革命に関する話」があり, それら二つをふまえて「日本の未来の教育についての話」がなされている. 実際の会話の流れと一致しており, このグラフを参照することによって議論の流れや議論の内容理解が素早く進むことが期待される.

また時間的に離れており話題間のつながりは薄いと考えられ実際に cosine 値も低くなっているが, 相互情報量により結ばれている話題間にリンクも相互情報量によって抽出されている. 「過去のマイクロコンピュータによる情報革命」に関する話題と「未来のコンピュータで可能なこと」に関する話題へのリンクや, 「日本の国際競争力」の話題と「将来の IT 教育と電波」の話にリンクが張られている. 出現単語の類似性という観点では類似度は低いながら, 2 つの話題間の共通の事象という観点で考えることが可能であると思われる話題である. しかしこの関係性については被験者を多くし, 関係性から新たな発見を得られたかを検討していく必要がある.

4. おわりに

本稿では対話ログの話題の遷移を視覚化するとともに、話題間の重要な関係性を相互情報量を基にして提示した. 対話ログの話題遷移だけからでは提示されない関係性を提示したが, 提示した関係性から新たな論点の発見がなされたかについては多くの被験者を通してのさらなる実験が必要である. また話題間の隠れた共通因子の存在をリンクによって示したが, その隠れた因子についての情報は提示されていない.

これを踏まえ, 今後の課題・展望として,

- この手法の有用性のさらなる検討
- 内部情報同士の関連性を示すだけでなく, 外部情報を用いて隠れた話題を提供する.

などのことが必要であろう.

参考文献

[前野 07] 前野義晴: コミュニケーションから探る組織の見えない黒幕, 人工知能学会論文誌 22 巻 4 号 C, 2007.

[松村 03] 松村真宏: 議論構造の可視化による論点の発見と理解, 日本知能情報ファジィ学会誌 Vol. 15, No. 5, pp.554-564, 2003.

[大澤 06] 大澤幸生: チャンス発見のデータ分析, 東京電機大学出版局 , 2006.

[斉藤 10] 斉藤正孝, 片上大輔, 新田克己: データ結晶化を用いた対話ログの時系列解析, 情報処理学会研究報告 Vol.2010-CE-103 No.2, , 2010.

[瀬々 07] 瀬々佳奈: KeyGraph とデータ結晶化を用いた交渉ログからのシナリオ抽出支援システム, 電子情報通信学会 , 2007.

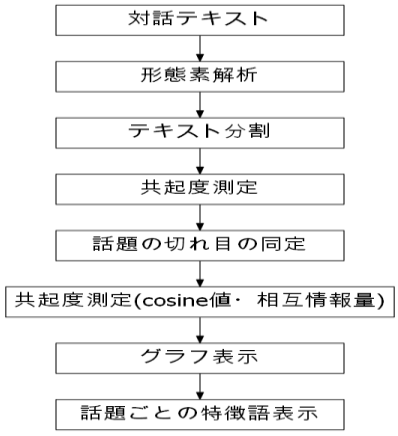

図 1 システムの構成図

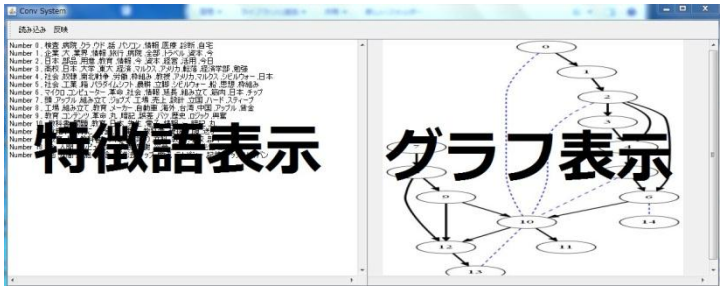

図 2 システム提示例