

第5章

評価実験

5.1 序言

本章では，COLLAGREE で行われた議論のデータを対象にした提案手法の評価実験について述べる．評価実験では COLLAGREE 上で行われた複数の議論データを用意し，提案手法と比較手法で実験を行う．結果として提案手法のほうが分散表現を用いていることで良い精度を出せるか確認する．

以下に本章の構成を述べる．まず，5.2 節では実験に用いたデータについて説明する．5.3 節では実験設定について述べ，5.4 節では評価実験の結果を示す．5.5 節では実験結果に対する考察を行い，最後に 5.6 節で本章のまとめを示す．

5.2 対象データ

5.2.1 議論データ

議論データはCOLLAGREEを用いた実験で収集されたデータを使用する。データの概要を以下に示す。

【実験概要】

実施年月 : 2017 年 3 月

グループ人数 : 2~3 名

議論時間 : 90 分前後

議論テーマ : 外国人観光客向けの日本旅行プランの決定

議論テーマ説明文 : みなさまに、外国人観光客向けの日本旅行プランを立てていただきます。想定される旅行者の条件は以下の通りです。

- 英語は話せるが、日本語は話せない
- 初めての日本旅行である
- 日程は6泊7日
- ホテルは自分たちで手配できる
- 旅行のために貯金したので、金銭的には余裕があり、国内をいろいろとまわることが可能である
- 来日、帰国の際の空港は、どこでもかまわない
- 2つのプランを比較したいと考えている（プランは2つ用意してください）

ファシリテータ：あり

5.2.2 評価データ

本研究では 5.2.1 節で説明した議論データに対し、次に述べる基準で伊藤孝行研究室の学生にアノテーションを行ってもらった。アノテーション担当者が基準を満たすと判断した発言に”1”のタグを、満たすと思われない発言に”0”のタグを付ける。

① それまで話題となっていた対象や事態とは異なる、新しい対象や事態への言及する発言

話されている内容が、以前と全く異なる対象や事態へと移行する位置でデータを区切る。

例 1:

(今までの話題:パック旅行はなぜ安いのかについて)

- A:ホテルが宿泊費の一部を出しているから安いのかな？
- B:おそらく。
- A:なるほど。
- B:沖縄行きも安いね。(今まで沖縄の話はされておらず、この後“沖縄行きのパック旅行”に話題が変わる(かもしれない))

例 2:

(今までの話題:外国人のツアー旅行の行き先について)

- A:他は寄らなくてもよいですか?(新しい行き先が出るように仕向けている)

② 既に言及された対象や事態の異なる側面への言及する発言

既に話題として取り上げられることについて、以前とは異なる側面から言及がなされる位置で区切る。

例 3:

(今までの話題:外国人のツアー旅行の行き先について)

- A:広島、長崎はどう？
- B:外国人観光客とか広島、長崎で見かけた覚えがない。
- A:ツアーに英語を話せるスタッフとか付けられるかな?(“ツアー旅行のスタッフ”に話題が変わる(かもしれない))

③ 議論のフェーズを移行させる(かもしれない) 発言

議論のフェーズを今までから移行させる(と思われる) 発言の位置で区切る。

例 4:

(今までの話題:外国人のツアー旅行の行き先について)

- A:八坂神社や清水寺など有名どころがたくさんありますし、魅力的だと思います

- B: 京都周辺ツアー清水寺、金閣寺、銀閣寺、伏見稲荷大社、嵐山、など日本
の建物や食べ物など広島長崎ツアー広島、長崎の戦争の地を見る事と、
それぞれの場所で食べ物建造物を見るツアー (地名を挙げる段階から、各地
点を結ぶツアープランへの作成段階に話題が変わる (かもしれない).)

例 5:

(今までの話題:外国人のツアー旅行のプランについて)

- A: 京都周辺ツアー清水寺、清水焼体験、抹茶・和菓子など体験、きもの体験、
金閣寺、嵐山、伏見稲荷大社その中で乗れそうなら屋形船などはどうでしょ
うか?
- B: 屋形船、風情があって良いと思います。
- (途中省略)
- C: まとめると、・京都周辺ツアー京都周辺（八坂神社、清水寺、金閣寺、
銀閣寺、伏見稲荷大社、嵐山、有馬温泉）、おいしい料理（豆腐など）、温
泉、6泊7日ツアー
・広島長崎ツアー広島（3日）：広島原爆ドーム、平和記念公園、厳島神
社、もみじまんじゅう、牡蠣、広島筆（メイクや書道なので使用する）、
お好み焼き、呉の戦艦、アナゴ（移動1日）長崎（3日）：ハウステンボス、
グラバー園、眼鏡橋、大浦天主堂、軍艦島、長崎ちゃんぽん、佐世保バー
ガー
この2プランで問題ないでしょうか？(初めて、2つのツアーの内容をまと
め、議論の収束に近づけた.)

また、ファシリテーターによる議論をコントロールするような発言も含む。

例 6:

- F:もし現在の旅先候補でよろしければ、具体的なプランづくりに移行したい
と思います。よろしいでしょうか？
- F:残り 20 分を切りました。皆様、いかがでしょうか？

以上の基準に沿ってタグを付けてもらい，“1”のタグが過半数より多く付けられた
発言を正解値=1，他を正解値=0 とした。

5.3 実験設定

5.3.1 パラメーター

本実験ではパラメーターは次の通りに設定した。前処理にて用いる okapiBM25
のパラメーターは $k1=2$, $b=0.75$ とし、LexRank のパラメーターは $n=50$, $threshold=0.7$ とした。また、重み付けを用いて文章から抽出する単語の数は 5 個とし
た。分散表現として用いる fastText は次元数を 100 次元とし、学習データには
wikipedia ダンプデータを用いた。総合類似度の計算に用いるパラメーターは $max-$
 $Time=5400(90 \text{ 分})$, $tWeight = 0.5$ とし、総合類似度の閾値は 0.8 とした。表 5.1
に実験の設定をまとめる。

okapiBM25 のパラメーターは一般的に妥当とされる [?] ものをを用いた。LexRank
のパラメーターの場合、 n は 50 以上でも結果に差がなかったことから 50 とし、

okapiBM25	k1	2
	b	0.75
LexRank	n	50
	threshold	0.7
抽出単語数		5
fastText	次元	100
	学習データ	wikipedia ダンプデータ
maxTime		5400
tWeight		0.5
類似度閾値		0.8

表 5.1: パラメーターの設定

threshold は結果が最も良かったものを用いた．抽出単語数も同様に最も結果が良かったものを用いた．

5.3.2 比較手法

① 常時通知

1 つ目の比較手法は，最も単純でわかりやすい発言の内容に関係なく常に通知する手法を用いる．

② TF-IDF ベクトル

単語の意味は考慮せず出現頻度に基づく比較手法として，分散表現の代わりに TF-IDF で発言をベクトル化する手法を用いる．

2 つ目の比較手法では okapiBM25 の代わりに TF-IDF を用いて連想配列を作成する．そして，連想配列から単語の重みを要素として持つベクトルに変換する．発

言内容の類似度計算は提案手法と同じで Cosine 類似度を用い、以降の総合類似度も提案手法と同じである。

5.3.3 評価指標

本実験では評価指標として適合率 (Precision), 再現率 (Recall), F 値 (F-measure) の 3 種類の指標を用いる。

適合率, 再現率, F 値はそれぞれ次のようにして求める。まず, 発言の通知を行うと判定した時を予測値=1, 通知を行わないと判定した時を予測値=0 とする。次に, 予測値=1 かつ正解値=1 であるものの個数を $TP(True Positive)$ または $hits$ (的中数), 予測値=0 かつ正解値=1 であるものの個数を $FN(False Negative)$ または $misses$ (見逃し数), 予測値=1 かつ正解値=0 であるものの個数を $FP(False Positive)$ または $falseAlarms$ (誤警報数) として数える。また, 予測値=0 かつ正解値=0 であるものの個数を $TN(True Negative)$ として数える。そして, 式 5.1, 式 5.2 及び式 5.3 に従って適合率, 再現率, F 値を計算する。

$$Precision = \frac{hits}{hits + falseAlarms} \quad (5.1)$$

$$Recall = \frac{hits}{hits + misses} \quad (5.2)$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (5.3)$$

3 つの値はどれも値が高いほど判定精度が高いことを示す。

5.4 実験結果

実験結果を表 5.2 に示す.

手法	平均評価指標		
	Precision	Recall	F-measure
比較手法 1	0.256579335	1	0.404074977
比較手法 2	0.75	0.105429708	0.180947229
提案手法	0.517148273	0.558192262	0.509950195

表 5.2: 実験結果

また, 各手法の TP,TN,FP,FN の割合の平均, 及び TP と FP の平均割合の和である P-SUM と TN と FN の平均割合の和である N-SUM を表 5.3 に示す.

手法	平均割合					
	TP	TN	FP	FN	P-SUM	N-SUM
比較手法 1	0.251993	0	0.748006	0	1.0	0
比較手法 2	0.0239234	0.730462	0.0079744	0.2280701	0.0414673	0.9585326
提案手法	0.1419457	0.5980861	0.1499202	0.1100478	0.2918660	0.7081339

表 5.3: 実験結果 2

5.5 考察

実験結果から次のことが言える.

考察 1 提案手法は適合率と再現率の良いバランスで判定をすることができる

考察 2 提案手法の精度は重み付けに依存する

考察① 提案手法は適合率と再現率の良いバランスで判定をすることができる

表 5.2 が示すように、提案手法は比較手法よりも高い F 値を出している。原因を究明するために、他 2 つの比較手法の問題点を考える。比較手法 1 は常に話題が変化したと判定するため見逃しが無く再現率が高いが、対価として何も除外しないので適合率は低くなってしまう。極端さが F 値の低下に繋がったと言える。

一方、比較手法 2 では TF-IDF による発言ベクトルを用いて発言内容の類似度を計算しているが、TF-IDF では文字の出現頻度のみを使用していることと全ての単語を発言ベクトルに含んでいることから新しく投稿された発言の内容文と過去の発言の内容文の両方に同じ単語が含まれている程、類似度が大きくなる。すなわち、過去の発言に全く登場していない新しい単語を多く含んだ内容文を持つ発言でない限り、過去の発言と同じ話題であると判定されやすい。結果として、比較手法 2 では過去の発言との差がよほど顕著な発言でない限り、話題が変化したと判定しなくなるため適合率は高くなるが、対価として見逃しが増え再現率が低くなる。事実、表 5.3 が示すように比較手法 2 では比較手法 1 とは逆に発言の殆どが話題の変化を起こさないものとして判定されている。図 5.1 に比較手法 2 で発言の繋がりを図示したものの拡大図を示す。

図 5.1 で示されるように殆どの発言がかなりの数の発言と同じ話題であると判定されていることが確認できる。

以上の 2 つの比較手法の問題点に対する考察を踏まえて提案手法が比較手法と異なる点を考える。異なる点として、上位の単語のみを類似度計算に用いている点と分散表現を用いている点が挙げられる。比較手法 2 と違って、上位の単語の



図 5.1: 比較手法 2 による繋がりグラフ-拡大図

みが用いられることで同じ単語が含まれているだけでは類似度が上がるとは限らなくなり，結果として話題が変化したと判定する回数が増え比較手法2よりも高い再現率を示したと考えられる．図5.2に提案手法で発言の繋がりを図示したものを示す．

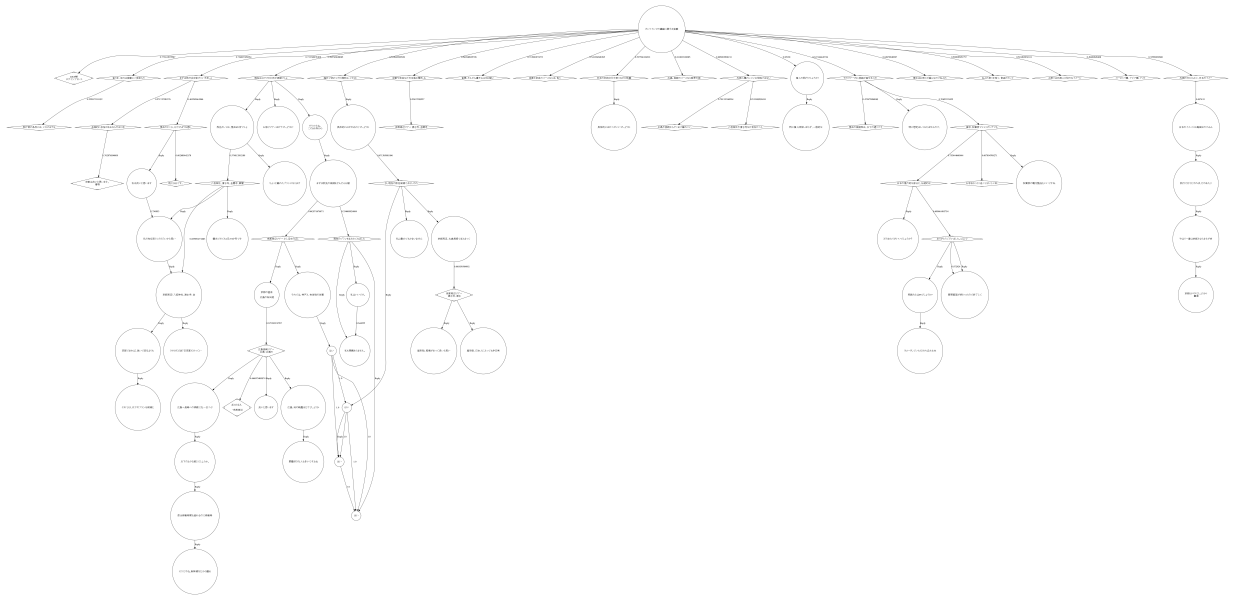


図 5.2: 提案手法による繋がりグラフ

図5.1とは違い，1つの発言に対して同じ話題であると判断された発言の数は少なく，比較手法2に比べ類似度が低くなっていることが伺える．また，分散表現を用いることで字面の異なる単語でも類似性を示すことができるため，再現率と共に適合率を上昇させることに繋がったと考えられ，故に適合率と再現率の間に大きな差を生むことなく判定をすることができたと考えられる．

考察② 提案手法の精度は重み付けに依存する

表5.4, table:MissesRemark にそれぞれ *falseAlarms*, *misses* となった発言の例を示す．まず，表5.4の上段にあるような，議論の最初の方において意見を促

title	body	抽出単語
まずは旅先候補を	まずは旅先候補をどんどんお願いします。	どんどん, お願い, 書き込み, 皆様
NULL	八坂神社、清水寺、金閣寺、銀閣寺、伏見稲荷大社、嵐山、有馬温泉 今あがっているのは、このぐらいでしょうか？	あがっ, 今, 伏見稲荷大社, 嵐山, 有馬温泉

表 5.4: falseAlarms 発言データ

title	body	抽出単語
NULL	やはり東京は人気ですね、アキバ行きたいって外国のかたが多いイメージです	外国, 行き, 多い, イメージ, アキバ
NULL	1日目大阪 (USJ) 2日目京都 (大原, 嵐山, 金閣寺, 二年坂) 3日目京都 (大原, 嵐山, 金閣寺, 二年坂) 4日目金沢 5日目金沢 6日目箱根 7日目東京 と言う感じでしょうか	金沢, 金閣寺, 二年坂, 大原, 嵐山

表 5.5: misses 発言データ

すようなファシリテーターが行うような発言が falseAlarms として誤判定されることが多かった。原因として過去に投稿された発言が少ないため他の発言との差が顕著になってしまったと考えられる。

一方、表 5.5 の上段にあるような他の発言に対する雑談っぽさのある他の視点を提供しようとするような返信発言が misses として誤判定されることが多かった。原因として本研究における手法では返信関係にある 2 発言間の類似度は返信関係にない 2 発言間に比べて上がりやすくなっていることが考えられる。

また、まとめを行う発言に弱いことが示された。原因としてまずはそもそも、表 5.4, 5.5 の両方にまとめを行う発言があるように人間でも時と場合によって判断が分かれることが挙げられる。既に投稿されたまとめ発言との差やタイミング等、提案手法で考慮しきれていない要素が人の判断に影響していると考えられる。他の原因として、含まれている単語数が多いことが挙げられる。以前の発言に含まれている単語が集まっていることから、表 5.4 の下段で示されているように偏り無く抽出されることがある。逆に表 5.5 の下段で示されている発言のように偏った内容の抽出がされることがある。上記のような不確定要素が結果に影響を与えたと考える。

5.6 結言

本章では本研究で提案する話題変化の判定手法が有用であることを実験により確認した。伊藤孝行研究室の学生に、COLLAGREE で行われた議論データに対して基準を満たすと思われる発言にタグを付けてもらって評価データとした。評価実験では発言の内容に関係なく常に通知する手法と分散表現の代わりに TF-IDF を用いて発言をベクトル化する手法から比較した。

実験の結果、提案手法は適合率と再現率の偏りのない良いバランスを保つ結果

を示すことが分かり，比較手法よりも高い F 値を出すことがわかった．