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と単語抽出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示す．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(ファシリテーターは除く)

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ないな。
- 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)：広島の原爆ドーム、平和記念公園、厳島神社、もみじまんじゅう、牡蠣、広島の筆（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=True，他を正解値=False 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 maxTime=5400(90 分)，tWeight = 0.5，α=0.8 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，α

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
α		0.85
類似度閾値		0.8

表 5.1: パラメーターの設定

と類似度閾値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．

上記のパラメーターを基本とした上で提案手法は以下に述べる 2 種類を用いた．

① デフォルト

表 5.1 のパラメーターをそのまま使用する．

② 単語抽出なし

単語抽出を行わず，発言内容中の単語ベクトル全ての平均を取る．本手法を行う理由は単語抽出が精度上昇に貢献していることを確認するためである．単語抽出に関連しないパラメーターは全て表 5.1 のパラメーターをそのまま使用する．

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 MeCab で発言内容文からストップワードを取り除いたものに対してに 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．

③ LDA ベクトル

分散表現以外で単語の意味を考慮する比較手法として，分散表現の代わりに LDA で発言をベクトル化する手法を用いる．

3つ目の比較手法では MeCab で発言内容文からストップワードを取り除いたものを Bag-of-Words ベクトルに変換する．そして，Bag-of-Words ベクトルのトピック分布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．LDA はトピック数を 100 個とし，学習データには wikipedia ダンプデータを用いた．

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=True，通知を行わないと判定した時を予測値=False とする．次に，予測値=True かつ正解値=True であるものの個数を $TP(True\ Positive)$ または $Hits$(的中数)，予測値=False かつ正解値=True であるものの個数を $FP(False\ Negative)$ または $Misses$(見逃し数)，予測値=True かつ正解値=False 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=False かつ正解値=False 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．

$$Precision = \frac{Hits}{Hits + FalseAlarms} \tag{5.1}$$

$$Recall = \frac{Hits}{Hits + Misses} \tag{5.2}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

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

実験結果として，各手法の F 値を表 5.2 に，各手法の再現率と適合率及び 2 つの値の差を表 5.3 に示す．

また，各手法での話題変化判定における TP と FP の割合の和である P-SUM と

手法	F-measure
比較手法 1 (常に予測値=1)	0.404
比較手法 2 (TF-IDF ベクトル)	0.487
比較手法 3 (LDA ベクトル)	0.385
提案手法 1 (単語抽出あり)	0.515
提案手法 2 (単語抽出なし)	0.086

表 5.2: 実験結果 1

手法	平均評価指標		
	Precision	Recall	Difference
比較手法 1 (常に予測値=1)	0.257	1	0.743
比較手法 2 (TF-IDF ベクトル)	0.489	0.558	0.069
比較手法 3 (LDA ベクトル)	0.573	0.317	0.256
提案手法 1 (単語抽出あり)	0.515	0.552	0.037
提案手法 2 (単語抽出なし)	0.833	0.046	0.787

表 5.3: 実験結果 2

TN と FN の割合の和である N-SUM を表 5.4 に示す.

手法	平均割合	
	P-SUM	N-SUM
比較手法 1 (常に予測値=1)	1.0	0
比較手法 2 (TF-IDF ベクトル)	0.83896	0.16103
比較手法 3 (LDA ベクトル)	0.255	0.745
提案手法 1 (単語抽出あり)	0.58776	0.41223
提案手法 2 (単語抽出なし)	0.02737	0.97262

表 5.4: 実験結果 3

表 5.5 に提案手法 1 で正しく判定された発言の一部を示す.

title	body
地域	広島や長崎からそこまで離れていない沖縄は先ほどのプランに入れるのはどうでしょうか?
議論が進めやすいようにどんな感じの方か決めませんか？	どんな方か空想できれば、よりよい案が出せると思うので。英語が話せて、日本語が話せないので、「アメリカ人」。日程は 6 泊 7 日で金銭的にも余裕があるようなので、例えば、〇十代のご夫婦、カップルなど。 人数が何人なのか、性別も決めるといいかもしれません。

表 5.5: 正解発言データ

5.5 考察

実験結果から，以下の考察①，考察②，及び考察③が言える.

考察① 提案手法は比較手法よりも総合的に性能が高い

考察② 単語抽出を行うことで性能が上昇する

考察③ 同じスレッドに属する 2 発言を同じ話題だと規定することで性能がある程度まで上昇する

考察① 提案手法 1 は比較手法よりも総合的に性能が高い

表 5.2 が示すように，提案手法 1 は他のどの手法よりも高い F 値を出している．表 5.3 が示すように，適合率では提案手法 2 が最も高く，再現率では比較手法 1 が最も高い値を出している．しかし，提案手法 2 では再現率が，比較手法 1 では適合率が他の手法に比べてかなり低くなっており，結果として F 値の低下に繋がっている．また，表 5.3 が示すように，提案手法では再現率と適合率の間の差が 0.037 で最も小さくなっている．すなわち，適合率と再現率のバランスが最も良かったことが他の手法よりも高い F 値に繋がったと思われる．

そして，表 5.5 に示されるように旅行プランへの追加提案や新しい視点の提案等，議論において話題変化を起こす可能性の高い発言を検出できていることが確認できた．

考察② 単語抽出を行うことで性能が上昇する

表 5.2 が示すように，提案手法 1 は提案手法 2 よりも高い F 値を出している．一方で，表 5.3 が示すように提案手法 2 は非常に高い適合率を示すと同時に，非常に低い再現率を示している．提案手法 1 と提案手法 2 の間の違いは単語抽出を行うかどうかだけなので結果に影響を与えたのは単語抽出である．単語抽出が結果に

影響を与えた理由として，単語抽出によって平均を取る単語ベクトルの数が減ったことが考えられる．表 5.4 が示すように，提案手法 2 では TP と FP を合わせた話題が変化したと判定した数が他の手法に比べて非常に少なく，多くの発言が過去の発言のどれかと高い類似度を示していることが伺える．すなわち，平均を取る単語ベクトルの数が多くなることで単語に関係なく結果的に平均ベクトルが全て似たようなものとなり，発言の平均ベクトルの差が小さくなってしまったと想定できる．

考察③ 2 発言が同じスレッドである場合の補正値を高くすることで性能がある程度まで上昇する

α=0.85 が最も良い α の数値であったが，最良であった要因として最も大きいのは類似度閾値の値よりも大きかったことであると考える．すなわち，2 つの発言が返信関係によって同じスレッドに属している場合，同じ話題であると規定することによって現在の提案手法では性能がある程度まで上昇することが分かった．

しかし，話題変化を起こすとタグ付けで判定された発言の内，他の発言と同じスレッドに属しているものは 55%である．すなわち，話題を変える発言の半数以上は返信発言であり，α=0.85 の条件下では見逃されていることが分かる．

故に，2 つの発言が返信関係によって同じスレッドに属している場合，同じ話題であると判定することは必ずしも性能が上昇するわけではなく，あくまで「ある程度まで」であることも分かった．

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及び LDA を用いて発言をベクトル化する手法と比較した．

実験の結果，以下のことがわかった．

- 提案手法は再現率と適合率の間の差が最も小さく，比較手法よりも良い性能を示す
- 単語抽出を行うことで発言の平均ベクトル間の差が出やすくなり，性能が上昇する
- 同じスレッドに属する 2 発言を同じ話題だと規定することで性能が上昇するが，限界がある