1L3-2in1

雑談システムにおけるトピックの関係性を考慮した発話選択手法

Topical Sentence Similarity for Response Selection in Casual Dialogue System

福田 拓也 *1　　若林 啓 *2
Takuya Fukuda　　Kei Wakabayashi

*1筑波大学 情報学群 知識情報・図書館学類
College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba

*2筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba

The response selection is one of the promising approaches to the casual dialogue system, which picks out an appropriate sentence from a massive candidate collection for system response to user's utterance. However, the existing methods have not explicitly addressed a phenomenon of topic drift, which is known as an indispensable feature to keep a conversation amusing in human dialogue. In this paper, we propose a method to measure a topical similarity between sentences to handle the topic drift in response selection based conversation. The preliminary experiment to inspect the correlation with human sense of ordering in topical similarity shows that our method works well when user's input utterance refers to a popular topic.

1. はじめに

近年，人間と対話を行うシステムの研究が活発になされており，私達の社会に広がりつつある．雑談システムとは人間との自然な会話を目指した対話システムであり，ELIZA や A.L.I.C.E が例として挙げられる．雑談システムの応答手法としては膨大な規則を用いたルールベース応答や対話ログを用いた用例ベース応答が存在する．これらには人的作業コストや多様なパターンを表現しきれないという問題が存在する．一方で，[稲葉 12] でも採られるような，予め用意した発話候補データからシステムの応答として発話候補文を選択する応答手法を発話選択と呼ぶ．発話候補データは，近年，Twitter のデータを用いた発話候補文の自動生成手法も提案されており，低コストで大量かつ質の高い発話候補文の獲得が可能となっている．

発話選択を行う上での重要な観点の 1 つとして，話題のつながりが考えられる．人間同士の会話では異なる話題が関わりをもって出現する場合が数多く存在し，図 1 中の会話では「センター試験」に関する話題から「健康」に関する話題へと遷移している．このとき話題のドリフト [Hobbs 90] が生じていると言える．話題のドリフトとは話題遷移を自然に働きかける，異なる話題同士の結合現象を表す．

このような高度な応答を雑談システムが意図的に行うためには話題の関係を捉えることが必要であると考えられる．本研究では関連する話題とその距離を求めることによる，話題の関係性を考慮した発話選択手法を提案する．これによって自然な話題転換をもたらす応答や対話の活性化が期待される．特に対話の活性化については [藤本 04] では人間同士の雑談では現在の話題と関連性はあるが当たり前過ぎない話題への遷移が対話の活性化へと繋がると指摘している．このような話題への遷移は，現在の話題に対して距離が離れた話題を選択することによって実現できると考える．

本研究は発話理解や話題転換のタイミングについてまで及ぶものではなく，あくまでも話題の距離についてのみ焦点を当てたものである．

連絡先: 筑波大学 情報学群 知識情報・図書館学類
〒 305-8550 茨城県つくば市春日 1-2
E-mail: s1211537@klis.tsukuba.ac.jp

No.	話者	発話
1	A	センター試験全然できなかった・・・
2	B	僕はセンター対策はしたから結構解けたよ
3	A	直前に風邪を引いたんだ．健康には気をつけたのに
4	B	体調には気をつけるべきだったね

図 1: 話題のドリフトの例

2. 関連研究

発話選択に関する研究では，[稲葉 12] はランキング学習を用いた発話選択を行う統計的応答手法を提案しており，[小新田 15] ではさらに発展させ，発話候補文の自動生成手法や話題転換手法も取り入れた対話エージェントの構築をしている．この研究では話題転換を示す語句が使用された，または対話が一定数経過した時，直前の発話からキーワードを取得し，その単語を含んだ発話文を選択している．しかしながら，取得する単語によっては現在の話題に対して不自然な話題転換をもたらす発話文を選択してしまう問題が生じるため，自然な話題のつながりを考慮する必要があると考えられる．

Web 上にある文章を発話文として選択することも広い意味での発話選択と捉えた際，[伊藤 15] ではカテゴリ情報を利用した意外性をもたらす発話選択を行っている．この研究では All About サイトにおいて，ユーザの発話と関連度が低い記事カテゴリに属し，かつユーザの発話と関連度が大きい記事を意外性が高い記事と判断し，その記事内の文章を発話として選択する．ここで，意外性をもたらす発話は話題を大きく遷移させる発話と解釈することができ，カテゴリ情報を用いて話題のつながりを間接的にではあるが考慮した手法と考えることができる．しかし，この手法で発話として選択することができるのは 1 記事内のデータのみであり，多様な応答を行うことは困難であり，また，選択した記事の内容を端的に表した文をどのように取得するかということを議論する必要がある．さらには話題数をカテゴリ数と見なしたとき，人間同士が行う会話の多様な

話題をカバーできるだけのカテゴリ数が必要となる．

対話システムにおける語の連想性・関連性に着目した研究については，[吉村 13] は概念連想や常識判断を行う自律応答システムを構築している．ここで提案される対話システムは概念のつながりに着目している点では本研究の考えと一致していると言える．しかし，ここでの概念は単語と対応しており，ある単語について最も関連度が高い単語を話題転換に用いているため結果的に同じような話題の会話になりやすく話題が行き詰まりやすい問題が生じると考えられる．また，関連度の低い単語を用いたとしても必ずしも明確に異なる話題へ遷移するとは限らない．それは単語としては関連度は低いがその背景にある話題が同じである場合が考えられるからである．

本研究では話題をある意味軸でまとめられた単語集合として扱う．これにより同じ話題を背景としてもつ単語の統制を行うことができ，またクエリの拡張という観点から見るとより多様な発話文を選択することが可能となる．話題の距離については今回，想起する話題の類似性を用いて判断する．これは同じ話題を想起するほど話題遷移の後も似た系統の話題を示すものとして考えたためである．

3. 提案手法

ここでは話題の距離の算出を始めとする発話文の話題認識，話題情報を用いた発話選択について具体的な方法を説明する．

3.1 話題の距離の算出

3.1.1 トピックモデルを用いた話題の生成

まず始めに，話題の関係性を捉えるに当たり，本研究ではトピックモデルの 1 つである Latent Dirichlet Allocation (LDA)[Blei 03] を用いた単語の話題推定を考える．トピックモデルとは様々なデータに隠された潜在的なトピックを推定する教師なし学習モデルであり，LDA は単語の背景にある話題を軸に確率的な枠組みを導入し，文書を複数の話題が混ぜ合わさったものと考えたモデルである．生成されるトピックは単語の確率分布で表され，確率値はそのトピックにおける単語の生成確率を表す．また，LDA ではサンプリングにより学習データ中の各単語トークンについてトピックを推論する．本研究では LDA により生成されるトピックを話題とする．

3.1.2 共起グラフの作成

次に，LDA によって生成された話題についてノードを話題，エッジの重みを共起度とする重み付き無向グラフを作成する．本研究では Augmented Expected Mutual Information (AEMI)[Chan 99] を用いて共起度を計算する．AEMI は確率を考慮した精細な共起度を測るための指標として用いられ，事象 a と事象 b が同時に成立する/しないとき，一方のみが成立するときの合計 4 つの場合を考慮している．本研究ではしきい値 (経験的に 0.0002 に設定) を定めることにより，重みが小さすぎるエッジは排除するものとする．

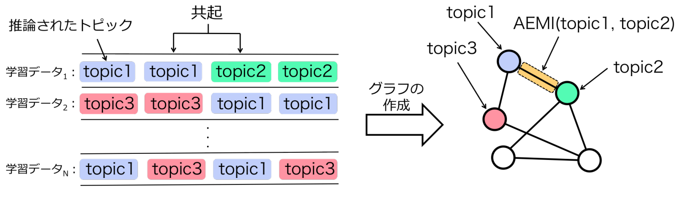

図 2: 話題の共起を用いてグラフを作成

3.1.3 距離の計算

作成した共起グラフにおいて，ある話題から距離 1 以内のノードを"想起される話題"と考える．そして，2 つの話題からそれぞれ想起される話題のうち，互いに共有している話題の割合が高い状態を密な結合関係と呼び，対して互いに共有している話題の割合が低い状態を疎な結合関係と呼ぶことにする．このとき想起する話題集合の類似性から，話題同士が密な結合関係にある場合は距離は近いと判断し，疎な結合関係にある場合は遠いと判断する．

この話題間の密または疎な結合関係を求める方法として，話題間の関連度を考慮し，重み付きの Jaccard 距離 [Flavio 10] を用いる．このとき話題間の直接エッジの重みは Jaccard 係数のスコアを上昇させるため重要な要素と考え，これを考慮するために話題間に仮想ノードを設ける．算出対象は，各話題について想起される話題のみとする．

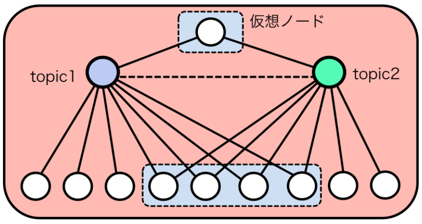

図 3: Jaccard 係数

3.2 発話文の話題認識

ユーザの発話文から話題を認識する方法を図 4 を用いて説明する．提案手法では発話文中のキーワードの話題がその発話文全体の話題を表すものと考える．始めにユーザの発話について名詞単語を抽出し，単語の IDF の値に基づきキーワードを求める．IDF とは逆文書頻度とも呼ばれ，一般に単語の重みとして利用されるものである．本研究では IDF の算出は LDA を用いた話題生成の際に用いるデータセットと同じものを使用し，最も IDF が高い単語をキーワードとして選択する (手順 1)．

次に求めたキーワードから話題を決定する．LDA によって生成された話題は単語の確率分布で表されており，生成される全ての話題は学習データに出現した全単語についての生成確率を持っている．このため発話文中の各単語について，単語の生成確率がしきい値 (経験的に 0.0005 に設定) を上回る話題候補から生成したものとして考える (手順 2)．この段階でキーワードについての話題候補が 1 つに求まる場合，その話題をキーワードの話題とする．そうでない場合，1 発話文中に含まれる話題の種類は多くないと考え，同じ話題を候補としてもつ単語がある場合にそれらの単語について，共通する話題以外の話題候補は削除する (手順 3)．

手順 3 を終えた段階でもキーワードの話題候補が複数ある場合，話題の共起度を用いて他の単語の話題候補と最も関連度の高いキーワードの話題候補を求める (手順 4)．

3.3 話題情報を用いた発話選択

話題情報を用いて発話候補データから発話文を選択する方法を図 5 を用いて説明する．ここで行うことは話題を端的に表す発話文を選択することである．始めに目的の話題において，生成確率が上位 30 件の単語集合を作成する (手順 1)．その後，その生成確率の比率に従ってランダムに 5 つの単語を選択する．これは 1 つの話題情報からより多様な発話候補文を選択するためである．この選択された単語について，IDF を素性値

ドイツやフランスのオーケストラは素晴らしいですね

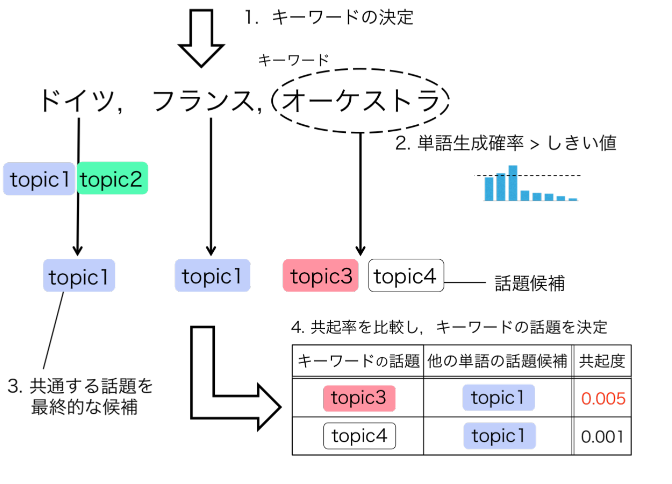

図 4: 発話文の話題認識の流れ

とする話題の特徴ベクトルを求める (手順 2)．対して Bow 表現に変換した各発話候補データについて，同様に IDF を素性値とする発話候補文の特徴ベクトルを作成する (手順 3)．以上の 2 つの特徴ベクトルについてコサイン類似度を算出し，類似度が高い発話候補文を目的の話題を表す発話文とする．

また，本手法では発話選択について，複数の話題情報を組み合わせて用いることが可能である．IDF をその単語が与える話題の情報量として見なす．このとき IDF がより高い単語であれば，ある特定の話題について多くの情報をもたらし，IDF が低い単語であれば話題についての情報量は少ないと判断することができる．また，話題ごとに IDF を足し合わせることにより，発話文がどの話題についてどの程度の情報量を含んでいるかを導くことができると考えられる．

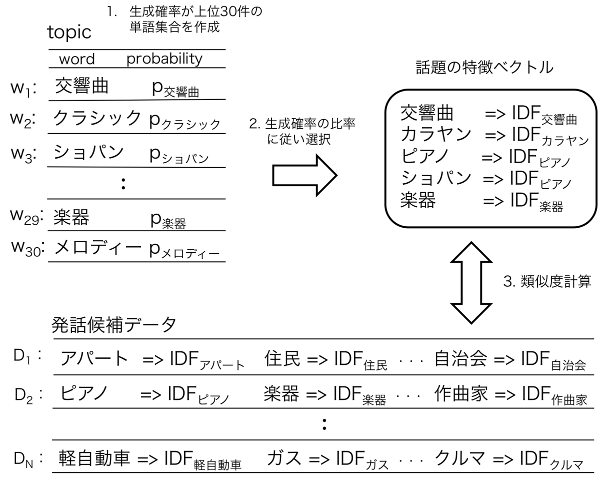

図 5: 発話選択の流れ

4. 評価実験

実験では被験者の感覚と似た形で話題の距離に従った発話文の選択が行えたかを提案手法と被験者が並べた発話文の順序の相関を求め，評価する．

4.1 実験用データ

4.1.1 LDA の学習用データと前処理

本研究では LDA の学習用データとして Yahoo!知恵袋における，研究機関向け公開用データ [*1] を使用した．

LDA による学習を行う際の前処理について説明する．始めに Yahoo!知恵袋の質問とベストアンサーのセットを 1 文書とする．その後，各文書について形態素解析により名詞情報を抽出し，BoW 表現に変換する．形態素解析の辞書は Web 上の言語資源を活用した mecab-ipadic-neologd[*2] を使用する．あらかじめ本データ特有のカテゴリである「Yahoo! JAPAN」に属する文書は学習データとして含めないものとした．次に，BoW 表現の文書集合から文書長が上位の文書をカテゴリごとに収集する．以上によって収集された 238,327 件の文書を学習データとして用いて，4,994 トピックを生成した．

4.1.2 発話候補データと前処理

本研究で用いる発話候補データは LDA の学習用データと同様に Yahoo!知恵袋を用いる．これは Yahoo!知恵袋では口語調になっているデータが多いため，発話文としても利用することができると考えたためである．「。」や「？」など文の区切りを示す記号で抜き出し，20 文字以上 100 文字以下の名詞を少なくとも 1 つ含む文書を発話候補として収集した．このとき，「ところで」や「しかも」などの接続詞を含む文書は削除した．以上によって収集した 216,100 件の文書を発話候補データとして用いた．

4.2 実験方法

実験では評価データを作成し，それらをユーザの発話と見なす．この評価データは発話候補データの中からそれぞれ異なる話題を含んだ文書になるようにあらかじめ人手で選別した発話文とする．まず始めに，発話文の話題認識手法 (3.2) に基づき評価データの話題を求める．そして求めた話題に対して距離が近い話題，遠い話題，中間の話題を選択する．その後，話題情報を用いた発話選択手法 (3.3) に従って単語の IDF を素性値とする話題の特徴ベクトルを作成し，発話文を選択する．以上の手順により 1 つの評価データに対して 3 つの発話文を得ることができる．このときの発話文の順序を，提案手法による話題の距離に従って並べた発話文の順序とする．

それらの発話文をランダムに並び替え，クラウドソーシングを用いて被験者により判断される話題間の距離に従い，それらの発話文を再度並べ替えてもらう．そして提案手法による話題の距離に従って並べた発話文の順序と被験者が並べた発話文の順序の相関を求め，提案手法が正しく話題の距離に従って発話文を選択できたかを評価する．この時，評価データに対して意味の通った応答ではなく，あくまで話題の距離に基づき判断してもらう．実験では 30 件の評価データを作成し，1 つの評価データにつき 33 件，計 990 件の実験データを用いて提案手法を評価した．また，実験はクラウドソーシングサービスの 1 つである Lancers[*3] を利用した．

4.3 評価手法

提案手法を用いてシステムが並べた発話文と被験者が並べた発話文の順序に相関性があるかを調べるためにスピアマンの順位相関係数を用いる．順位相関係数 ρ は順位が完全に一致しているとき+1 をとり，完全に逆のとき-1 をとる．順位相関係数は 2 つの順位の相関を求めるものであり，システム側の発

*1 http://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html
*2 https://github.com/neologd/mecab-ipadic-neologd
*3 http://www.lancers.jp

話文の順番と被験者が並べた発話文の順序について平均順位相関係数 $\bar{\rho}$ を算出する．

4.4 結果と考察

実験結果を図 6 に示す．本実験により算出した各評価データの平均順位相関係数の最大値は 0.985，最小値は-0.318 であり，評価データ全体の平均は 0.496 であった．評価データのうち，70 %のデータが正の相関 ($0.4 < |\bar{\rho}| \leqq 0.7$) を示し，30 %が特に高い正の相関 ($0.7 < |\bar{\rho}| < 1.0$) を示した．評価データのうち，70 %のデータが正の相関 ($0.4 < |\bar{\rho}| \leqq 0.7$) を示し，30 %が特に高い正の相関 ($0.7 < |\bar{\rho}| < 1.0$) を示した．しかしながら 23 %のデータについてはほどんど相関を示さず ($0 < |\bar{\rho}| \leqq 0.2$)，また，負の相関の示すデータ ($0.2 < |\bar{\rho}| \leqq 0.4$) も存在した．このことから提案手法により相関を示す話題もあれば，相関を示さない話題もあることが分かる．相関を示した評価データに対しては人間の感覚と近い形で話題の距離に従った発話文の選択が行えたと考える．

また，正の相関を示した評価データを図 7 に示す．正の相関を示した評価データの話題とほどんど相関なしと判断された評価データの話題を比較してみると，話題の距離の大きさに違いが見られた．正の相関を示した評価データに対してほどんど相関なしと判断された評価データでは，評価データの話題とそれから想起される話題の距離はどれも遠いと判断され，さらにその差は微小なものであった．このことから評価データの話題は想起されるほとんどの話題に対して疎な結合であったと考えられる．そのような話題に対して密な結合をもつ話題を生成する必要があるため，その話題についての学習データをより多様化することや，生成する話題数を増やすことがほどんど相関なしと判断されたデータを減らす対策として考えられる．

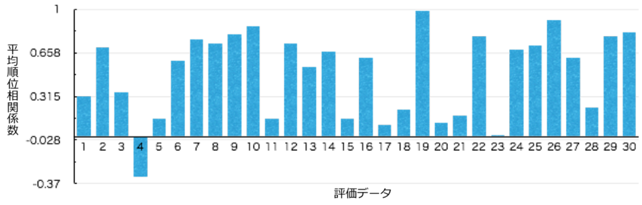

図 6: 評価データ毎の平均順位相関係数

評価データ（話題: 食材）	野菜はよく食べているほうだと思うのですが，肉，魚はあまりたべません
発話A（話題: 調味料）	ごぼうとにんじんをごま油で炒めて，だしの素，酒，みりん，砂糖，醤油で味つけています
発話B（話題: 虫）	あるコンビニでサラダを購入したらその中に羽の付いた虫が混入していました
発話C（話題: 栄養）	卵のタンパク質はもっとも良質で，ビタミンや鉄分も豊富にふくんでいます
システムの選択順序	A C B
被験者の選択順序の内訳	A C B：24，C A B：6，C B A：1，B A C：1，A B C：1

図 7: 正の相関を示した評価データ

5. おわりに

本研究では話題を抽象的な枠組みで捉え，そのつながりに着目し，話題の関係性を考慮した発話選択手法を提案した．その中で話題の距離を算出し，それに基づいた発話選択を目指した．

実験では被験者の感覚と似た形で話題の距離に従った発話文の選択が行えるかを判断するため，提案手法と被験者が並べた発話文の順序の相関を調べた．結果としては多くの評価データについて正の相関を示すことができた．しかしながら，提案手法は話題の関係性のみ考慮したものであり，入力文に対する意味を理解した応答ではない．このため今後の課題としては既存の対話エージェントに対して本手法の機構をどのように組み込むかが挙げられる．

組み込む対話エージェントについては，人間同士で行われる会話中に出現する話題は膨大であり，LDA によってその全てを網羅した話題集合を生成することは難しい．現状では特定ドメインの対話エージェントに焦点を絞り，提案手法を実装することが現実的と考える．また，LDA については，話題の階層構造を組み込んだモデルを用いることでより話題の粒度を考慮した，より多様な話題を生成できるものと考える．今後は上記の事柄を含め，本手法によって期待される自然な話題転換や対話の活性化についての評価も行っていきたいと考える．

謝辞

本研究の一部は，JSPS 科研費（課題番号 25280110,25540159）および筑波大学図書館情報メディア系プロジェクト研究（Research Projects of Facalty of Library, Information and Media Science）の助成によって行われた．

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003)

[Chan 99] Chan, P.: A non-invasive learning approach to building web user profiles, *KDD-99 Workshop on Web Usage Anal- ysis and User Profiling* (1999)

[Flavio 10] Flavio, C., Ravi, K., Sandeep, P., and Sergei, V.: Finding the Jaccard Median, in *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pp. 293–311 (2010)

[Hobbs 90] Hobbs, J.: Topic Drift, in B, D. ed., *Conversational Organization and Its Development*, pp. 3–22, Ablex Publishing (1990)

[伊藤 15] 伊藤 直之, 西川 侑吾, 大野 和久：語の意外度に基づき話題展開する非タスク指向型対話システム, 人工知能学会全国大会論文集, Vol. 29, pp. 1–4 (2015)

[稲葉 12] 稲葉 通将, 平井 尚樹, 鳥海 不二夫, 石井 健一郎：非タスク指向型対話エージェントのための統計的応答手法 (自然言語処理), 電子情報通信学会論文誌. D, 情報・システム, Vol. 95, No. 6, pp. 1390–1400 (2012)

[吉村 13] 吉村 枝里子, 芋野 美紗子, 土屋 誠司, 渡部 広一：知的会話処理における連想応答手法, 人工知能学会論文誌, Vol. 28, No. 2, pp. 100–111 (2013)

[小新田 15] 小新田 亮人, 稲葉 通将, 高橋 健一：Twitter から獲得した発話を用いる対話エージェントの構築, 人工知能学会全国大会論文集, Vol. 29, pp. 1–4 (2015)

[藤本 04] 藤本 英輝, 高梨 克也, 河野 恭之, 木戸出 正継：概念的関連性に基づく雑談の話題転換点分析, 第 18 回人工知能学会全国大会 (2004)