Wikipedia のリンク情報に基づく話題遷移シナリオの自動生成

坂田 亘 [1,2] 吉越 卓見 [2] 田中 リベカ [2] 黒橋 禎夫 [2]
[1]LINE 株式会社 [2] 京都大学
wataru.sakata@linecorp.com, {takumiyoshiko, tanaka, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 はじめに

我々の日常会話において自分の興味ある話題に会話を誘導する振る舞いはよく見られる重要な対話行為である。本研究では**ターゲット指向対話**を導入し、そのような対話を実現するために話題遷移のシナリオを生成するシステムを構築する。ターゲット指向対話では対話相手の関心のある話題（**開始話題**）から雑談を開始し、全く別のある話題（**ターゲット話題**）に向けて能動的に話題の誘導を行う。こうした対話はセールストークを行うときなどに不可欠である。

提案システムでは情報源として Wikipedia を利用し各エントリを１つの話題とみなすことで、開始話題からターゲット話題への話題遷移経路の検出を行う。Wikipedia の文書内のハイパーリンクとそれを含む自然文には、エントリとエントリがどのように関係しているかの情報が含まれている。これをある話題から、リンク先エントリの話題への遷移として利用する。開始話題とターゲット話題には直接の繋がりがない場合が多いため、中継点となる話題を経て話題遷移を行うことが必要となる。Wikipedia のような webgraph 構造においては、多くのノードは他の全てのノードから小さい遷移数で到達できるという small-world 性が見られるため、数回の話題遷移で開始話題からターゲット話題へと遷移できる。

図 1 は複数の映画タイトルをターゲット話題として登録した際のシステムの概要図である。提案システムは開始話題が与えられると、そこから予め登録した複数のターゲット話題のうちいずれかに到達する最短の話題遷移経路を検出する。図 1 左には “吹田市” を開始話題として与えられた場合の話題遷移経路の候補が示されている。経路内の各リンクは Wikipedia 内のハイパーリンクと対応しており、話題間の関係を表す文が付随している。図 1 右は “吹田市” の話をしている対話相手に対して、システムが図 1 左の話題遷移経路を用いて映画を勧めようとしている対話の一部である。“日本沈没” へ話を近づけるためにシステムは『吹田市の万博会場跡地は万博記念公園 として整備され、 太陽の塔などがあるらしいですね。』と発話し、まず “太陽の塔” の話題へ話を転換しようとしている。このように適切に話題遷移経路が計算できれば、ユーザーの幅広い関心に対応しながらターゲット話題へ話を誘導する対話シナリオとして、これを用いることができる。

遷移における経路の選び方は複数存在するため、その中でより良い道筋を選択する必要がある。例えば図 1 では “降水” への話題の遷移は自然ではないと思われる。本研究では Wikipedia 上のリンク情報および文情報を利用して訓練データの自動構築を行い、話題遷移をスコア付けする方法を提案する。

実験では映画推薦のための話題遷移経路を見つけるシステムの構築を行い、獲得できた経路の分析を行う。さらに提案手法により適切な話題選択をすることで、比較対象の手法よりも自然な話題遷移が可能であることを確認する。

2 提案手法

2.1 最短話題遷移経路の探索

本項目では開始話題を入力としたときにターゲット話題のいずれかに到達する最短経路を求める方法について述べる。開始トピックが与えられてから実用的な時間で最短の経路を計算するためには幅優先探索を利用する。ターゲット話題それぞれを始点にとり、それぞれのリンクを逆向きに辿る幅優先探索を行うことで、各エントリから各ターゲット話題までの距離を計算できる。計算量は Wikipedia のエントリ数を D、ターゲット話題の数 T とすると $O(DT)$ となる。さらに各エントリに対して全ターゲット話題の中でエントリからの距離が最小となるようなものを計算し、その距離を保存しておく。これにより、あるエントリから最短でいずれかのターゲット話題の到達するために次に遷移するべき話題を定数

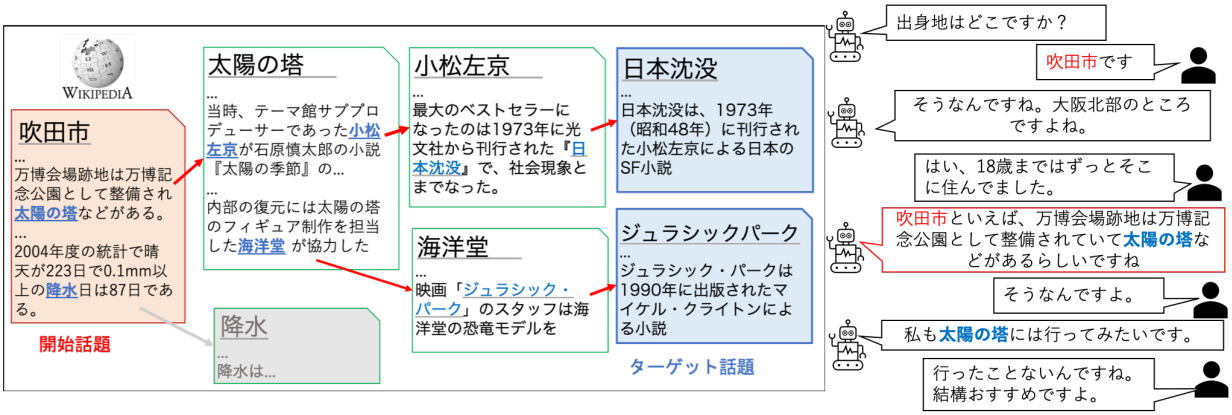

図 1　システム概要

時間で見つけられる。これを繰り返すことで開始話題からターゲット話題に到達するための経路を高速に求めることができる。

2.2　話題選択モデル

前節の方法では開始話題からターゲット話題までの複数の経路を得ることができるが、その中から良い経路を見つけることができれば便利である。ある話題と話題の関係を示す文が話題遷移に使う情報として自然であるかを識別できれば、そのような関係を示すリンクを優先的に辿ることで、良い話題遷移の経路を発見できると考えられる。本節では Wikipedia の文とリンク情報から教師データを自動構築し、それを用いることで話題の遷移先を適切に選択するモデルを構築する。

まず、実際の遷移の事例について観察する。Wikipedia の “手羽先唐揚げ” の記事には “バッファローウィング” および “白ごま” へのリンクとともに次のような記述が存在する。

(1)　手羽先唐揚げに類似するアメリカ料理として、バッファローの名物料理、**バッファローウィング**がある。

(2)　鶏の手羽先をから揚げにしてタレを塗り、塩・胡椒・**白ごま**などを振りかけて仕上げる。

これらの情報を話題遷移として利用することを考えると、バッファローウィングへの遷移と比べて、白ごまへの遷移は不自然に感じられる。

我々はこのようなリンク情報の観察を通して話題遷移の自然さに関する次のような性質に着目した。ある文が話題 A へのリンクを含んでおり、後に続く文で A が主語となる場合、その文の情報は A への話題遷移に活用しやすい。例えば、先ほどの “バッファローウィング” への遷移の例において、次の文で「バッファローウイングは〜」とバッファローウィングの詳しい説明が続くことは自然に感じられる。一方、“白ごま” の例において、次の文で白ごまが主語として現れる可能性は “バッファロー” の例と比べると低いと思われる。そこでリンク文字列とそれを含む文を入力として、次の文でそのリンク文字列が主語として現れるかどうか推論を行う判別器の訓練を行い、リンクのスコアの計算に用いることを考える。

疑似訓練データは Wikipedia 内の連続する 2 文を収集し、次のように自動的に生成できる。まず連続する 2 文のうち 1 文目に存在するリンクが示す語 l が 2 文目においては主語となっているような 2 文を収集する。主語の判定はリンク文字列が 2 文目において助詞 “は” の直前に存在しているか形態素解析器を用いることで判定を行う。リンクを示す語 l と 1 文目の文全体 s のペアを正例として用いる。また負例として、1 文目内から無作為に選択した名詞節 n と 1 文目の文全体 s のペアを利用する。こうして収集した疑似訓練データを利用し、正例ならば 1、負例に対しては 0 を出力するように判別器の訓練を行う。以上で述べた方法によって日本語 Wikipedia から訓練データを収集することで正例データ 73,197 件と負例データ 71,653 件が得られ、これらを全て用いて訓練を行った。

推論時にはリンク文字列とそれを含む文をモデルの入力として得られた出力の値をリンクのスコアとして利用する。複数のリンク候補から 1 つ選択する際には、このスコアの最も大きいリンクを選択する。候補のうち、最上位に同程度のスコアが並ぶ場合（最大スコアとの差が 0.01 未満）には、それらの中で IDF 値が高いリンク文字列に紐づくリンクを選

択する。モデルとしては日本語大規模コーパスで事前訓練済み[1)]の BERT[1] を利用する。

3 実験

3.1 Wikipedia データの前処理

Wikipedia 内に存在するリンク情報の中には話題遷移のために用いるには相応しくないものが多く見られる。特に “情報”、“平成” などのように多くのものと関連している一般的な話題への遷移は通常の対話ではあまり見られない。そこで次に示すような Wikipedia 内の一般的なエントリへのリンク情報は用いないよう除去を行った。

文書頻度の大きいエントリ Wikipedia の各エントリに対して文書頻度を計算し、文書頻度の大きいものから上位 1,000 件のエントリへのリンクを除去対象とした。この対象となったエントリの例として “情報”、“スポーツ” などがある。

被リンク数の多いエントリ 多くの Wikipedia 文書からリンクされているエントリもまた過度に一般的なエントリが多い。被リンク数が 10,000 以上のエントリへのリンクを除外の対象とした。この対象となったエントリの例として “アメリカ合衆国”、“英語” などがある。

年月日や年号についてのエントリ 「アメリカンフットボールは 1921 年には...」などの文のように、年月日または年号などのリンクが Wikipedia 内には多く見られたが後続の文の主語として用いたとしても話題遷移の自然さが低いと思われる。これらはパターンマッチングによって除去を行った。

また、丸括弧内のリンク、箇条書きとなっている文内のリンクは利用しないこととした。最終的にシステムが用いる Wikipedia のエントリは約 100 万件、リンクは約 1,700 万件となった。

3.2 話題遷移経路の分析

ユーザの関心ある話題を開始話題として収集を行いそこからターゲット話題までの経路を計算するとどのような経路が得られるか調査を行った。映画推薦のための話題遷移経路を生成するシステムを想定し、ターゲット話題として複数の映画タイトルのエントリを利用した。具体的には Wikipedia のエントリ「映画の一覧」[2)]から収集を行い、結果として 409 件の映画リストを取得した。

表 1 取得した経路長・経路数・1 エントリあたりのリンク数（リンク数/E）の平均値、中央値、最小値、最大値

	平均値	中央値	最小値	最大値
経路長	2.58	3.0	1	5
経路数	14.3	7.5	1	113
リンク数/E	3.2	2.7	1	31

開始話題は次のようにクラウドソーシングで収集した。まず、会話の導入となるような簡単な質問を作業者に投げかけユーザ発話を収集を行なった。「あなたの趣味は何ですか」「好きなアーティストは誰ですか」などユーザ発話の収集のための対話導入用の発話を 10 個用意し、各質問につき 100 件のユーザー発話を収集することにより 500 件のユーザ発話を収集した。次にそれぞれのユーザ発話をクエリとして Wikipedia のエントリを対象に検索を行い、検索結果の各エントリを開始話題の候補とした。情報検索には Elasticsearch[3)] を利用した。収集した開始話題候補のうち、ユーザクエリに対して適当な開始話題でないと想定されるものを除去し、結果として 432 件の開始話題が得られた。

収集した開始トピックをもとに得られた経路の統計情報について表 1 に示す。話題経路の長さは平均で 3 リンク未満であり、多くの開始話題に対して数回の話題遷移でターゲット話題に到達できている。

3.3 遷移先話題の選択モデルの評価

次に話題選択モデルの評価を行った。評価のためにベースラインとして RANDOM、IDF を採用した。RANDOM は遷移候補の中から無作為に 1 つ選択する。IDF は各遷移候補に対しリンク文字列の IDF を計算し、IDF が最も大きい遷移を選択する。

評価にはクラウドソーシングを用い、RANDOM、IDF、提案手法のうち 2 つずつを対象に、どちらの手法による話題選択が自然か比較した。最初に、IDF と RANDOM を比較した。まず、Wikipedia からリンクを複数含む文を収集し、各文について、IDF と RANDOM で 1 つずつリンクを選択した。選択したリンクが同一だった場合はその文を除外し、異なるリンクを選択した文のみをクラウドソーシングに用いた。図 2 のように、リンクを含む文をもとに簡単なルールで生成した疑似対話文と各手法が選択した 2 つのリンクを提示し、どちらのリンクへ話題遷

1） http://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese
2） https://ja.wikipedia.org/wiki/映画の一覧
3） https://www.elastic.co/jp/elasticsearch/

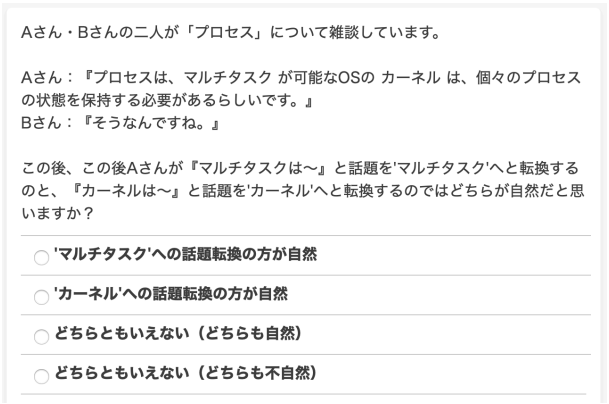

図 2　クラウドソーシング画面

表 2　話題選択モデルの評価結果

IDF	RANDOM	引き分け	合計
137	43	220	400
提案手法	RANDOM	引き分け	合計
123	50	227	400
提案手法	IDF	引き分け	合計
122	88	210	400

移するのが自然かを作業者に尋ねた。作業者は「どちらともいえない（どちらも自然)」、「どちらともいえない（どちらも不自然)」を選択することもできる。このとき、提示する順序によってバイアスが生じないようにはじめの 2 つの選択肢の順序は無作為になるようにした。1 件につき 3 人の作業者が評価し、一方のリンクのみが 2 票以上獲得した場合に、そのリンクを選択した手法に 1 加点した。それ以外の場合は引き分けとしてカウントした。これを 400 件評価し、提案手法と RANDOM、提案手法と IDF についても同様に比較した。

結果を表 2 に示す。提案手法は IDF、RANDOM に対して遷移先としてより自然なリンクを選択できていることがわかる。次に事例を表 3 に示す。１つ目は提案手法の選んだリンクが IDF の選んだリンクよりも自然である事例である。文内でより重要な役割を示す語を選択できている。2 つ目は IDF の方が自然であるとされた事例である。“輸血” と “血液製剤” は文内での重要度は同様であると思われるが “血液製剤” の方が話題としての面白さがあることから、話題遷移先としてより自然であると判断されたと考えられる。話題としての面白さを考慮したリンクの選択については今後の課題とする。

表 3　提案手法と IDF の選択事例：下線は提案手法の選択したリンク、下波線は IDF の選択したリンクを指す。

遷移元	文	正解手法
ワイリーエックス	ワイリーエックスは、プロテクティブ・オプティクス社が 1986 年に設立したブランドでアメリカ軍 、 連邦捜査局 、 アメリカ中央情報局 、アメリカ合衆国の警察など向けに目の保護を目的とした サングラス・ゴーグルを製造しているメーカーらしいです。	提案手法
献血	献血とは 輸血 や 血液製剤 製造のために無償で 血液を提供することであるらしいです。	IDF

4　関連研究

対話システムの研究において Wikipedia などの外部知識を活用し知識を持った対話を行うシステムへの研究は盛んに行われている [2, 3, 4]。しかし、これらのシステムではユーザーの発話に対して受動的に応答するのみで、システムが能動的に話題を選択するような機構は準備されていない。システムが話題の遷移を行う対話システム構築への取り組みも存在し [5, 6]、例えば Zhou らは ConceptNet [7] 上の話題の経路をもとに、推薦対話のデータセットを構築した。これらの研究においては遷移元と遷移先がどのような関係を持つかといった情報の利用は行われていない。我々の研究は Wikipedia に存在する多様な話題の関係性の知識を用いて話題遷移のシナリオの生成を行なっており、これと異なる。また、Wikipedia の small-world 性を用いて 2 つの遠い概念の関係性を明らかにする取り組み [8, 9, 10] も存在するが、この知識を対話遷移に利用する研究は我々の知る限り存在しない。

5　まとめと今後の課題

本研究では Wikipedia のリンク情報に基づき話題遷移シナリオを生成する新たなタスクに取り組んだ。実験では多くの開始話題から数回の遷移でターゲット話題へ到達できることを確認した。また、提案手法はベースラインよりも自然な話題を選択できることを示した。獲得したシナリオを基に流暢な会話を行うシステムを作ることが今後の課題である。

謝辞

この研究は国立情報学研究所 (NII) CRIS と LINE 株式会社とが推進する NII CRIS 共同研究の助成を受けています。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL2019*, pp. 4171–4186, Minneapolis, Minnesota.

[2] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *ICLR2019*, Ernest N. Morial Convention Center, New Orleans.

[3] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech2019*, pp. 1891–1895, Graz, Austria.

[4] Jaehun Jung, Bokyung Son, and Sungwon Lyu. AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue. In *EMNLP2020*, pp. 3484–3497, Online. Association for Computational Linguistics.

[5] 平岡拓也, Graham Neubig, Sakriani Sakti, 戸田智基, 中村哲. 説得対話システムにおける話題誘導に基づく対話制御. 情報処理学会 第 94 回音声言語情報処理研究会, 東京, 2012.

[6] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. In *COLING2020*, pp. 4128–4139, Online.

[7] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI2017*, pp. 4444–4451, San Francisco, California, USA.

[8] Robert West, Pineau Joelle, and Precup Doina. Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI2009*, pp. 1598–1603, California, USA.

[9] Robert West and Jure Leskovec. Human wayfinding in information networks. In *WWW2012*, pp. 619–628, New York, USA.

[10] Robert West and Jure Leskovec. Automatic versus human navigation in information networks. In *ICWSM2012*, pp. 143–144, Dublin, Ireland.