Vol.2016-NL-229 No.5
2016/12/21

雑談対話システムの話題遷移における自然性の自動評価

豊嶋 章宏[1,a)]　杉山 弘晃[2,b)]　吉野 幸一郎[1,c)]　中村 哲[1,d)]

概要：雑談対話システムにおいては，雑談における話題の幅を持たせるため，焦点語に着目した話題遷移の研究が行われている．しかしこうした話題遷移は，時として文脈に関連のない不自然な遷移を行ってしまい，対話の自然性を低下させる場合がある．そこで本稿では，雑談対話システムにおける話題遷移の自然性を自動評価する手法について検討する．具体的には，任意のユーザ発話に対してシステム発話が生成された際に，システム発話がどの程度自然に話題遷移できているかその自然性を推定するモデルの構築を行った．今回は，焦点語や対話行為，述語項，N-gram の 4 つの素性を用いて SVR による回帰モデルを学習して自然性を評価するモデルを構築し，人手で評価された値とモデルの評価値を比較した．

キーワード：対話システム，話題遷移，自然性評価

1. はじめに

近年，自然言語を介して人間と会話を行うことを目的とした雑談対話システムが注目を集めている．NTT のしゃべってコンシェル[*1]や Apple の Siri[*2]，Microsoft の Cortana[*3]などの対話システムにおいても，雑談は対話を円滑に進めるうえで重要な役割を持つと認識されており，雑談対話システムに関する研究は今後も増加していくことが予想される．

これまで雑談対話システムでは，対話における焦点語などを扱うことによって，焦点語を中心に話題遷移をする研究が行われてきた [1]．しかし，焦点語のみでは対話の前後の文脈を捉えきることが難しい．そのため，システム主導で話題遷移を行う場合に，不自然な話題遷移が行われてしまう問題があった．これに対し，話題遷移を行った前後でその遷移が自然かどうかを評価することができれば，こうした不自然な話題遷移を抑制できる可能性がある．

そこで，本研究では雑談対話システムにおける話題遷移の自然性を自動評価する手法について検討する．具体的には，あるユーザ発話に対してシステム発話が生成された際に，どの程度自然な対応であるかを評価する．このユーザ発話とシステム発話ペアの自然性を考慮することによって，明らかに不自然な話題遷移を抑制することができる．今回は，機械学習によって構築したモデルを用いて自然性の評価を行う．この自然性スコアによって対話履歴に対しての妥当性を計算することができるようになり，明らかに不自然な話題遷移を含む候補を抑制することができる．

[1] 奈良先端科学技術大学院大学　情報科学研究科
Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
[2] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 2-4 Hikaridai, Seika, Sagara, Kyoto 619-0237
[a)] toyoshima.akihiro.su4@is.naist.jp
[b)] sugiyama.hiroaki@lab.ntt.co.jp
[c)] koichiro@is.naist.jp
[d)] s-nakamura@is.naist.jp
[*1] https://www.nttdocomo.co.jp/service/shabette_concier/
[*2] http://www.apple.com/jp/ios/siri/
[*3] https://www.microsoft.com/en-us/windows/cortana

2. 対話の破綻と自然性

対話の自然性に関する研究として対話破綻検出に関する研究がある [2]，[3]．対話破綻は人と対話システムが対話を進めていく際に，直前の話題や発話と上手く繋がらない発話をシステムが行うことで生じる．対話破綻検出では，こうしたシステム発話によって生じた破綻を検出する．システム発話が対話履歴に対して対話破綻を引き起こすか事前に検出できれば，そうした対話破綻を引き起こす可能性のあるシステム発話を抑制することが可能となる．本研究で取り扱う問題は対話破綻検出と類似しているが，これまでの対話破綻検出の研究では破綻有無のみを対象としているのに対して，本研究ではその自然性を数値として定量評価しているという点で異なる．これにより，ユーザ発話に対してシステム発話候補が生成された際に，候補から最も自然である発話を数値に基づいて選択することが可能となる．

3. 応答ペアの自然性評価

ユーザ発話とシステム発話がそれぞれ与えられた際に，

ユーザ発話に対してシステム発話がどの程度自然であるかを定量評価するモデルの構築を行う．これによって，不自然な話題遷移が行われていないかを検出することができる．本研究では，推定モデルを構築する際にユーザ発話とシステム発話のペアに対して自然さが 7 段階で付与されているデータを使用する．データに付与されている自然さの値は離散値だが 7 段階の度合いであるため，これを予測する手法として SVR (Support Vector Regression) を用いる．従来研究においても，対話の相互的な品質評価を行う際に SVR による推定モデルが最も効果的であったことが報告されている [4]．ユーザ発話とシステム発話間より抽出可能な特徴を素性として作成し，素性情報からユーザ発話に対するシステム発話の自然さを評価する回帰モデルを構築する．

具体的には，焦点語，対話行為，述語項，N-gram をユーザ発話とシステム発話から抽出し，これらの共起情報を SVR の学習における素性として使用する．焦点語は，発話文の中心となる単語である．焦点語は，従来研究の雑談対話システム [1] の対話遷移を考慮する際にも使用されており，自然さを推定する際にも有用であることが考えられる．対話行為は，対話における発話文の役割を表しており，”質問”や”自己開示”などの対話行為がある．対話行為は，ユーザ発話とシステム発話の発話文ペアから対話行為の遷移ペアとして抽出される．対話中に出現する対話行為の遷移ペアには，ある程度決まったパターンがあると考えられ，自然性の推定に有効に働くことが考えられる．述語項は，発話文を構成する述語と対応する項のペアを表している．述語項は発話における内容語で構成され，これが極端に変わるような話題遷移は自然でない可能性がある．N-gram は，発話文を単語区切りにした際の単語連接である．本研究のタスクと類似したタスクとして，会話の空白部分を選択するような会話文完成問題があるが，このタスクにおいて N-gram 素性は会話文の流れの自然さを推定する際に有用であることが示されている [5]．本研究で扱う自然性推定のタスクでも，N-gram の素性が有効であることが考えられる．

4. SVR による自然性評価器の構築

本節では，3 節で述べた自然性評価タスクに対して推定モデルを構築し，その評価を行った．4.1 項では，実験データから素性を抽出する手法について述べる．4.2 項では，モデルを構築および評価する際に使用する実験データについて述べる．4.3 項では，評価手法について述べる．

4.1 自然性評価器に用いる素性

発話文および応答文から素性を作成するための各特徴量の抽出方法について述べる．はじめに，焦点語，対話行為，述語項を抽出する．これらの素性の抽出には，リッチインデクサ[*4]を用いて発話文を解析した結果を用いる．リッチインデクサは，テキストデータの集合から人名や組織，評判情報などの情報を抽出する．本研究ではリッチインデクサを用いて焦点語，対話行為，述語項を発話文から抽出する．例えば，”8 月に京都へ行きました”という発話文をリッチインデクサを用いて解析すると，焦点語として”京都”，対話行為として”自己開示_事実”，述語項として”行く”と”8 月”という述語項のペアがそれぞれ抽出できる．この対話行為としては，目黒ら [6] が定義した対話行為を用いており，人手で対話行為がアノテーションされた対話データをもとに SVM で学習したモデルを用いて付与されている．発話文と応答文それぞれに対してこれらの特徴量を抽出し，その共起を素性とする．**図 1** に，発話文と応答文から焦点語，対話行為，述語項を抽出して素性を作成する例を示す．User はユーザ発話を示し，System はユーザ発話に応じたシステム発話を示している．ユーザ発話から，焦点語として”京都”，対話行為として”自己開示_事実”，述語項として”行く・8 月”がそれぞれ抽出されている．システム発話から，焦点語として”新幹線”，対話行為として”質問_事実”，述語項として”行かれる・新幹線”がそれぞれ抽出されている．そのため，焦点語として”京都-新幹線”，対話行為として”自己開示_事実-質問_事実”，述語項として”行く・8 月-行かれる・新幹線”といったそれぞれの特徴量の共起情報が素性として抽出される．

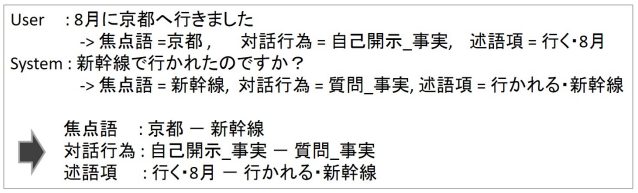

図 1 素性の作成例

次に，共起 N-gram を用いた素性について述べる．単語 N-gram を作成する際には，発話文および応答文に対して形態素解析を行い単語区切りの形式にする．形態素解析を行う際に使用する形態素解析器として，Jtag[*5]を用いる．この際に，動詞や形容詞など用言の活用形はすべて終止形に戻している．発話文中の単語連接から単語 N-gram を作成する．発話文と応答文からそれぞれ単語 N-gram を抽出し，組み合わせて共起ペアを作成することで共起 N-gram 素性を作成する．本研究では，単語 N-gram として 1-gram，2-gram，3-gram を用いて，発話文と応答文からこれらを組み合わせることで作成可能な 9 通りの共起パターンを共起 N-gram 素性として使用する．しかし，これらの手順より作成した共起 N-gram 素性は非常にスパースになることから，頻度が 10 より大きい共起 N-gram のみを素性として採用した．

[*4] http://www.ntt.co.jp/svlab/activity/category_2/product2_07.html
[*5] http://www.ntt.co.jp/svlab/activity/category_2/product2_06.html

このようにして 4 つの特徴量からそれぞれ素性を作成した．本研究では，これらの 4 つの素性とその組み合わせから構成される 6 つの素性を用いる (**表 1**)．以後，焦点語，対話行為，述語項より作成された素性を組み合わせたものを 3feat.，4 つの素性を組み合わせたものを 4feat. と記載する．SVR の実装としては LIBLINEAR[*6]を使用した．

表 1 使用する素性の一覧

素性名	用いる素性
焦点語	焦点語
対話行為	対話行為
述語項	述語項
N-gram	N-gram
3feat.	焦点語，対話行為，述語項
4feat.	焦点語，対話行為，述語項，N-gram

4.2 実験データ

本研究では，筆者らが作成した発話・応答文のペアに対して 7 段階自然さが付与されているデータを用いる．実験データは，500 文の発話文に対してそれぞれ 100 文の応答文と自然さの値が付与されている．発話文は雑談対話コーパス [1] および Twitter よりそれぞれ 250 文ずつ抽出して作成しており，1 文で話題や意味が分かる文を採用している．応答文は，それぞれの入力文に対して人手により 70 文，システムによる自動生成で 30 文作成している．人手で作成する際には，10 名のアノテータによって 7 文ずつ作成している．この際，対象となる発話文を文節単位でマスキングしており，ユーザは発話文に対して自然である応答文の作成が困難となるような条件が設定してある．このため，文法上は正しいが話題遷移が自然にできていないような応答文が付与されている．マスキングの条件として，マスキングしていない文が 42 文，30%の文節をマスキングしている文が 14 文，60%の文節をマスキングしている文が 14 文となるようにアノテータに対して発話文提示を行い，応答文の付与を行ってもらっている．マスキングされた発話文の一例を表 2 に示す．

表 2 マスキングされた発話文の一例

マスキングの割合 (%)	発話文
0	今日はいい天気ですね
30	今日はいい**
60	**いい**

このようにして作成した計 50,000 ペアの発話・応答文のペアに対して 6 人の被験者によって自然さの値を付与する．被験者は，発話文に対して応答文が自然な文であるかどうかを目視で評価する．自然さの値は 1〜7 の値を付与するものとし，値が大きいほど自然な文 (違和感がない文)，値が小さいほど不自然な文（違和感がある文) として判断する．**表 3** に，”1 皿ずつでてくるようなコースだとなかなかお腹いっぱいにならなかったりしますよね．”という発話文に対する応答文と応答文に対応した自然さの一例を示す．SVR を用いて自然性評価器を構築する際には 6 人の被験者による評価値を平均した値を用いる．

表 3 応答文と自然さの一例

応答文	自然さ
いつもコース料理とは別に何品か頼んでいます	7
時間に余裕がないと食べに行けないです	4
脚って中々細くならないよね	1

4.3 評価手法

構築したモデルの評価手法として 10 分割交差検定を用いる．使用する実験データの数が 50,000 ペアであるため，実験データを 5,000 ペア毎の 10 のデータセットに分割し，1 つのデータセットをテストデータ，それ以外の 9 つのデータセットを学習データとして使用して評価を行う．これを分割した 10 のデータセットすべてに対して行い，平均を計算することでモデルの評価を行う．また本研究では，評価尺度として以下のピアソンの相関係数を用いる．

$$r = \frac{\sum_{i}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i}^{N}(y_i - \overline{y})^2}} \quad (1)$$

式 (1) において，x は自然性の推定値，y はアノテーションされた自然性の値，N はデータの総数をそれぞれ示している．ピアソンの相関係数は，絶対値が 1 に近いほどデータ間の相関が高いことを示す．この相関係数を求めることにより，構築した自然性評価器の推定の精度を評価する．

4.4 評価と考察

表 1 に示した 6 パターンの素性を利用して 6 つのモデルを構築し，ピアソンの相関係数を用いてモデルによる予測値とアノテーションされた値の相関を評価した結果を**表 4** に示す．

表 4 評価結果

モデル名	相関値
焦点語	-0.007
対話行為	0.123
述語項	-0.089
N-gram	**0.312**
3feat.	0.263
4feat.	**0.382**

[*6] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

表 4 が示すように，各素性を単体で使用して学習した SVR のモデルの中では N-gram を素性として使用したモデルが最も高い相関を示した．すべてのモデルを比較した場合では，4feat. が最も高い相関を示した．一方で，焦点語や述語項のみを素性として使用したモデルの相関は低かった．この要因として，述語項や焦点語として抽出される名詞や動詞の組み合わせが多岐にわたるため，これらの素性がスパースになるためだと考えられる．日本語語彙体系 [7] や EDR 概念辞書 [8] に基づいて，これらの単語を上位概念の単語にマッピングすることや，Word2Vec[9] を用いて低次元の分散表現にマッピングすることで素性のスパースネスが解消され，これらの素性が有効に働くのではないかと考えられる．次に，相関度と発話文の関連性を検証するために発話文ごとの相関値の平均を計算した．**表 5** に相関値が高い発話文と相関値の平均，**表 6** に相関値が低い発話文とその相関値の平均をそれぞれ示す．

表 5 相関値が高い発話文と相関値例

発話文	相関値
スポーツはされますか?	0.753
お休みの日は出かけたりとかされますか?	0.723
この夏はどこか行かれましたか?	0.719
もし宝くじが当たったら何をしますか?	0.709

表 6 相関値が低い発話文と相関値例

発話文	相関値
さっさとひだまりの支払い番号おくれよ	-0.008
でも，春は私の場合，花粉でつらい季節ではありますね.	-0.012
朧って漢字だとかっこいいよな	0.21
奈良の遺跡とか本気で発掘したかった	0.048

表 5 に示す発話文は相関値の平均が 0.7 を超えたテストデータの値とモデルによる評価値の間に強い相関がある例である．これらの発話文は，Yes/No で回答する質問や，What や When を質問するようなクローズドクエスチョンであり，システム回答の候補も限られるためモデルによる予測精度が高いのではないかと考えられる．一方で，表 6 に示す発話文は相関値がほぼ 0 であり，テストデータの値とモデルによる評価値が無相関であることを示している．これらの発話はそれ自体が頻度の少ない例であることに加えて，応答のバリエーションも多くなるような例であるため，予測が上手くいかないのではないかと考えられる．

最後に，使用する素性と評価結果の関係について事例分析を行う．**図 2** に素性を組み合わせることで改善された事例，**図 3** に素性を組み合わせても改善されなかった事例をそれぞれ示す．

図 2 および図 3 は発話文と応答文が与えられたときの評価結果を示している．発話文と応答文の右に記述している”自己開示_事実”や”質問_事実”は文の対話行為をそれぞれ示している．また表中のスコアはアノテータにより付与されたスコアの平均値を示している．図 2 は N-gram のみを素性としたモデルの推定結果は，アノテータが付与した値と離れた値であるが，素性を組み合わせるとアノテータが付与した値に近い値に改善された例を示している．”自己開示_事実”は，自身の経験を相手に開示する役割，”質問_事実”は相手に質問する役割を対話行為として持つ．”自己開示_事実”から”質問_事実”へ対話行為が遷移することは，実際の会話でも起こりやすい．そのため，対話行為の素性を組み合わせることで，N-gram 素性ではうまく評価できない例に対応できたことが考えられる．一方で，図 3 に示す事例は素性を組み合わせても本来の評価値に近づけることができていない．この要因として，本研究で使用した 4 つの素性は入力文の意味が近い場合に自然であると評価してしまうため，入力文の意味が近いが自然ではない文をうまく評価できないためであることが考えられる．

発話文 : 8月に京都へ行きました。(自己開示_事実)
応答文 : 貴船のあたりは静かで涼しいですが、川床は行かれましたか？(質問_事実)

スコア	焦点語	対話行為	述語項	N-gram	3feat.	4feat.
6.2	3.72	4.07	0	0.47	4.25	4.21

図 2 改善された事例

発話文 : 8月に京都へ行きました。(自己開示_事実)
応答文 : 京都はもう行きましたか？(質問_事実)

スコア	焦点語	対話行為	述語項	N-gram	3feat.	4feat.
1	3.72	4.07	3.70	5.08	4.26	6.31

図 3 改善されなかった事例

5. おわりに

今回，ユーザ発話に対してシステムの応答発話がどの程度自然であるか推定するためのモデルの構築に取り組んだ．モデルの構築には SVR を使用し，異なる特徴量を用いた複数のモデルに対してピアソンの相関係数を用いて評価を行った．結果として，発話間の焦点語，対話行為，述語項，N-gram の共起情報をそれぞれ素性として SVR で学習したモデルが最も相関が高くなった．一方で，焦点語や述語項といった素性はスパースネスの問題があるため有効に働かなかった．そのため，シソーラスや Word2Vec 等を用いて素性を圧縮してスパースネスを解消することでモデルの推定精度が向上することが考えられる．

参考文献

[1] Higashinaka, R. Imamura, K. Meguro, T. Miyazaki, C. Kobayashi, N. Sugiyama, H. Hirano, T. Makino, T. and Matsuo, Y. : Towards an open-domain conversational system fully based on natural language processing, COLING, pp. 928-939 (2014).
[2] Funakoshi, K. Higashinaka, R. Inaba, M. Kobayashi, Y. Sugawara, S. Takanashi, K. Otsuka, H. Koiso, and H.

Bono, M. : On Dialogue Breakdown: Annotation and Detection - dialogue breakdown detection challenge, Second Workshop on Chatbots and Conversational Agent Technologies (2016).
[3] Martinovsky, B., Traum, D.: The error is the cue: Breakdwon in human-machine interaction. In: Proc. Error Handling in Spoken Dialogue Systems. pp. 1116 (2003).
[4] Schmitt, A. Schatz, B. and Minker, W. : Modeling and Predicting Quality in Spoken Human-Computer Interaction, SIGDIAL, pp.173-184, (2011).
[5] 堂坂浩二, 坂本裕磨, 高瀬淳, : 隣接発話らしさを利用した英語会話文完成問題の回答手法, 第 30 回人工知能学会全国大会, (2016).
[6] 目黒豊美, 東中竜一郎, 杉山弘晃, 南泰浩 : 意味属性パターンを用いたマイクロブログ中の発言に対する自動対話行為付与, 情報処理学会研究報告書, Vol.2013-SLP-98, No.1, pp.1-6 , (2013).
[7] Ikehara, S. Miyazaki, M. Shirai, S. Yokoo, A. Nakaiwa, H. Ogura, K. Oyama, Y. and Hayashi, Y. : GoiTaikei-A Japanese Lexicon, Iwanami Shoten (1997).
[8] NICT. : EDR Electronic Dictionary, NICT (1999).
[9] Mikolov, T. Sutskever, I. Chen, K. Corrado, G. Jeffrey, D. : Distributed Representations of Words and Phrases and their Compositionality, NIPS, (2013).