特定ドメイン雑談対話システムのための Wikipediaを用いた発話文の生成

Generating System Utterance Candidates Using Wikipedia for a Chat-Oriented Dialogue System in a Specific Domain

杉本 俊[1*†] 植木 拓[2†] 林 宏幸[3†] ニコルズ エリック[4] 中野 幹生[4]
Shun Sugimoto[1] Taku Ueki[2] Hiroyuki Hayashi[3] Eric Nichols[4] Mikio Nakano[4]

[1] 首都大学東京 [2] オーストラリア国立大学
[1] Tokyo Metropolitan University [2] Australian National University
[3] 電気通信大学 [4] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[3] University of Electro-Communications [4] Honda Research Institute Japan, Co., Ltd.

Abstract: One approach to generating system utterances in a chat-oriented dialog system is selecting a sentence that is related to the user's utterance from a set of utterance candidates prepared in advance. This paper proposes a method for automatically constructing a set of utterance candidates for a chat-oriented dialogue system in a specific domain. It generates a large amount of sentences from Wikipedia articles related to the domain of the system, and then excludes sentences inappropriate as system utterances using a classifier that exploits position information in the Wikipedia articles.

1 はじめに

近年，雑談対話システムに関連する様々な研究が行なわれているが [東中 14]，どのような話題でも扱うオープンドメインのシステムは，ユーザの発話内容が非常に多岐にわたるために，まだ適切な応答が困難である．そこで，我々は，特定の話題を扱う特定ドメインの雑談対話システムの研究を行っている．

雑談対話システムにおいて，発話文の候補を大量に用意しておき，その中からユーザの質問に関連が高いものをシステムに出力するという手法がよくとられているが [稲葉 12]，発話候補文集合を人手で作成するのは労力がかかる．発話候補文集合を自動構築する方法として，Twitter[1]を用いる方法が提案されているが [稲葉 14]，本研究では，より信頼できる情報源として，Wikipediaを用いる方法を提案する．具体的には，Wikipedia 日本語版[2]の記事にあるテキスト情報から，語彙情報を用いたフィルタリングと口語体への変換を行った後，記事中の位置情報を用いて，システム発話候補として不適切な文を除外する方法を提案する．

*連絡先： 首都大学東京大学院システムデザイン研究科
〒 191-0065 東京都日野市旭が丘 6 丁目 6
E-mail: sugimoto-shun@ed.tmu.ac.jp
†本研究は (株)Nextremer において行った．
[1]https://twitter.com/
[2]https://ja.wikipedia.org/

2 提案手法

提案手法では，まず当該ドメインに関連する Wikipedia の記事から，HTML の<p>タグに属する文のみをテキストとして抽出し，句点 (。) 区切りで文に分割する．次に，文脈がないと理解できないような文を除外する．具体的には，「先述」，「参照」，「この」，「その」などの，文脈依存の文であることを示す語を含む文を除外する．

その後，残った文を，人手で構築した規則を用いて口語体に変換する．この規則を用いることで，「乳酸菌は通常、腸内細菌として棲息しているが、ヨーグルトの乳酸菌は、腸内定着することはできない。」という文は，「乳酸菌は通常、腸内細菌として棲息していますけど、ヨーグルトの乳酸菌は、腸内定着することはできないんですよ。」[3]のように変換される．

また，文中にトピック（記事名）が存在しない場合には，トピックの補完を行う．文中にトピックが存在しない文に対して，その文の文頭に「“[トピック] って，”」を付け加える．例えば，“餃子” の記事に属する文から口語体に変換された文「ポーランドやスロバキアでは

[3]この論文では，クリエイティブ・コモンズ・表示・継承ライセンス 3.0（http://creativecommons.org/licenses/by-sa/3.0/）のもとで公表されたウィキペディア (https://ja.wikipedia.org/) の「ヨーグルト」及び「餃子」の項目を二次利用した．（参照日 2017 年 07 月 30 日）

ピエルクと呼ばれるんですよ」は，この処理によって，「餃子って，ポーランドやスロバキアではピエルクと呼ばれるんですよ」のように変換される．日本語の文章では，既に登場した主題を省略することが多いため，この処理を行うことで，雑談対話システムが出力してもユーザが理解できる文となる．

以上の処理によっても，まだシステム発話として不適切な文が残る．たとえば，一文で意味内容が理解できなかったり，トピックと内容が対応していなかったりする．そこで，その文が Wikipedia の記事のどこかから取り出されたかの位置情報（表 1）を用いて，不適切な文を除外する．これらの位置情報をどのように組み合わせて用いるかは，データから学習する．

3 評価

提案手法のうち，最後のステップである，記事中の位置情報を用いた不適切文除外の評価を行った．我々は料理に関する雑談対話システムを構築しているため，料理分野に関するトピックを無作為に 60 個選択し，Wikipedia からの文抽出，フィルタリング，及び口語体への変換を行なった．なお，品詞の識別には MeCab[4] を用いた．この処理によって得られた 2,427 文に対して，大学生 3 人でアノテーションを行なった．アノテーションの基準は以下の 3 つである．

(1) 日本語として不自然である
(2) 一文で意味・内容を理解できない
(3) トピックと内容が対応していない

これら 3 つの基準のどれかに当てはまる場合，システム発話として不適切な文であると判断した．以降，不適切な文を利用不可能文，それ以外を利用可能文と呼ぶ．

多数決の結果，利用可能文が 1,476 文，利用不可能文が 951 文得られた．3 名の評価者による判定の一致率（Fleiss の κ）は，0.703 となり，多少のばらつきが確認された．このアノテーションデータからロジスティック回帰に基づく判別器を学習した．判別器の評価のため，データを抽出元の記事（トピック）で 5 分割（各 12 トピック）し，交差検定を行なった．データ全体における利用可能文と利用不可能文の割合には偏りがあるため，各検定におけるトレーニングデータでは，利用可能文と利用不可能文の割合が同じになるようにランダムに利用可能文を排除した．

交差検定の結果，すべての特徴量を用いた場合の精度が 65.2%であり，すべての文を利用可能文と判断した場合の精度 60.8%を上回った．また，用いる特徴量の組み合わせをすべて試したところ，(f3) を除く 3 つの特徴量を利用した場合の精度が 65.5%で最も高かった．

[4] http://taku910.github.io/mecab/

表 1: 利用する特徴量

特徴量	説明
f1	ヘッダの種類 (<h1>, <h2>, <h3>, <h4>)
f2	所属するヘッダの番号
f3	所属する段落の番号
f4	所属する段落における文の番号

4 おわりに

本稿では，Wikiepedia から，特定ドメインの雑談対話システムの発話候補文集合を作成する方法を提案した．提案手法における，システム発話として不適切な文の自動検出の精度は十分とは言えないものの，記事中の位置情報という単純な情報が，複雑な要因によって決定される，発話候補文の利用可能性と一定の関連性を持つことを示した．判別器によって得られる尤度を利用し，発話候補文をランキングすることで，利用可能文抽出の適合率を高めることも可能である．

今後は，記事中の位置情報以外の特徴量を用いることで利用可能性判定の精度を向上させる．また，利用可能文と利用不可能文のラベル付けを行った三人の一致率が十分高い値でないことが判別器の学習に悪影響を及ぼしていると考えられることから，ラベル付けの基準をより明確にして行く．

参考文献

[東中 14] 東中 ： 雑談対話システムに向けた取り組み, 人工知能学会研究会資料 SIG-SLUD-70, pp. 65–70. 2014.

[稲葉 12] 稲葉, 平井, 鳥海, 石井 ： 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, 電子情報通信学会論文誌, J95-D(6), pp. 1390–1400. 2012.

[稲葉 14] 稲葉, 神園, 高橋 ： Twitter を用いた非タスク指向型対話システムのための発話候補文獲得, 人工知能学会論文誌, 29(1), pp. 21–31. 2014.