3A1-3

Wikipediaからの特定ドメインの雑談対話システムのための発話候補文集合の獲得

Acquisition of Domain-specific Utterance Candidate Sentences for a Chat Dialogue System from Wikipedia

杉本 俊 *1 Shun SUGIMOTO　植木 拓 *2 Taku UEKI　林 宏幸 *3 Hiroyuki HAYASHI　ニコルズ エリック *4 Eric NICHOLS　中野 幹生 *4 Mikio NAKANO

*1首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University

*2オーストラリア国立大学
College of Engineering and Computer Science, Australian National University

*3電気通信大学情報理工学部
Faculty of Information Science and Engineering, University of Electro-Communications

*4(株) ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan, Co., Ltd.

One approach to generating speech in a dialog system is to select sentences that are highly related to the user's utterance from among a set of utterance candidates prepared in advance. In this paper, we propose a method to automatically construct a domain-specific set of utterance candidate sentences for a chat dialogue system. Speciffcally, we generate a large amount of sentences from Wikipedia articles related to a specific domain, and exclude sentences that are inappropriate as system utterances using machine learning classifiers.

1. はじめに

対話システムは，ホテルの予約や道案内といった特定のタスクをこなすことを目的としたタスク指向型のもの，ユーザとの対話自体を目的とした非タスク指向型のものに分類することができる．後者に分類される雑談対話システムは，楽しみ，癒し，安心感など，心理的にプラスの影響をユーザに与えることから，近年活発に研究が行われている [東中 14]．

こういった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る [稲葉 12, 磯村 09, 水上 16]．しかし，現状では発話候補文は人手で作成されることが多く，莫大なコストがかかるという問題がある．本研究では，特定ドメインにおける雑談対話システムの発話候補文を自動で生成することを目的とし，豊富な知識を内包するWikipedia からの知識獲得を行う．例えば，料理に関する雑談を行うシステムを構築する場合，料理に関連するトピック (リンゴ，ラーメンなど) を用意し，それらと一致する Wikipedia 上の記事からテキスト情報を抽出する．その後，テキスト情報から大量の発話文の集合を構築し，機械学習を用いてシステムに利用可能な発話文のみ抽出を行う．

Wikipedia を利用するメリットはいくつかあげることができる．各記事には，ある一つのトピックについての情報が網羅され，体系化されている．書かれている情報には，問題もあるが一定の信頼性があることも示されている [日下 12]．Wikipedia 全体での総記事数は膨大であり，様々なドメインに関する十分な量の記事が存在する．このことから，ドメインに対する広範かつ専門的な知識を必要とする，特定ドメインの雑談対話システムに利用するコーパスとして，Wikipedia は非常に適していると言える．一方，問題も存在する．Wikipedia は誰でも簡単に編集を加えられるため，文章や見出しの構成などがページによって大きく異なる．また，各ページは一連の文章で構成されるため，そこから抜き出した一文は，前後の文脈を考慮しないと意味が理解できないこともある．そのため，抽出した文の集合の中から雑談対話システムに利用できる文とそうでない文を分類する必要がある．

本稿では，雑談対話システムに利用可能な文を自動的に判別する方法を提案する．利用可能でない雑談文の特徴として，日本語として不自然である，内容が専門的すぎる，文法的な欠陥がある，などがあげられる．しかし，こういった様々な感覚的要因をとらえて分類するのは，文の意味を深く理解する必要があるため，現状の技術では困難である．提案手法は，分類に利用する特徴を，それぞれの雑談文が持つ位置情報 (ページのどの範囲に記載されていたかという情報) に限定することで，雑談対話システムに利用可能な文を一定の精度で分類することができる．

本稿の構成は以下の通りである．2 節では，本研究と類似した課題に取り組んでいる関連研究について述べる．3 節では，提案手法について説明し，4 節において手法の評価を行う．5 節では考察を述べ，6 節ではまとめと今後の課題について示す．

2. 関連研究

[稲葉 14] では，Twitter をコーパスとした非タスク指向型の対話システムにおける発話文の自動生成を行なっている．正解発話としたツイートに多く出現する単語に対して高いスコアを設定し，それぞれのツイートにスコアリングを行なっている．しかし，本研究のような特定のドメインに対する豊富な知識が求められる雑談対話システムの候補文としては，Twitter での発言は情報の質・量ともに問題がある．また，この手法では，専門的な用語が多数を占める Wikipedia コーパスに対し

連絡先: 杉本 俊，株式会社 Nextremer，東京都板橋区成増 1-30-13 トーセイ三井生命ビル 10F，03-6915-6447，shun.sugimoto@nextremer.com

て有効性が低いと考えられる．

また，[太田 09] は，雑談対話システムのための Wikipedia からの発話候補文の抽出を行なっている．登場頻度の少ない単語から構成される文（珍しい文）は，対話を盛り上げるような内容であると仮定し，こういった文を抽出することを目的としている．それぞれの単語に対して，コーパス上における登場頻度を元にスコアリングを行い，スコアの高い文を抽出する．この手法では，ごく一部の特異な文しか抽出できず，特定ドメインの雑談対話システムとして利用するのは難しい．

本研究は雑談対話システムに利用する発話文の自動生成が目的であり，抽出対象がこれらの研究とは異なる．また，文の意味とは直接関連のない位置情報を分類に利用する．

3. 提案手法

本稿で提案する手法は，指定したトピックそれぞれに対して発話文の集合を構築する．具体的には，そのトピックの Wikipedia の記事から抽出したテキストに対してフィルタリングをして文の集合を取り出し，各々の文が発話文として自然になるように加工を加える．その後，それぞれの文が，Wikipedia の記事上のどこに現れていたかの情報（これを位置情報と呼ぶ）をもとに，発話候補文として相応しいかどうかの判定を行う．

3.1 Wikipedia からの文抽出

本項では，Wikipedia からの文抽出について説明する．まず，特定ドメインに属するトピック名で Wikipedia に検索をかける．トピック名と対応する記事が存在する場合には，その記事の 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．この時，抽出された各文に対して，以下の 6 つの情報を与える．

- topic name（トピック名）
- header name（ヘッダ名）
- header info（<h1>,<h2>,<h3>,<h4>）
- header id（上から何番目のヘッダに所属するか）
- paragraph id（ヘッダ中の何番目の段落に登場するか）
- sentence id（所属する段落の何文目に登場するか）

topic name 及び header name は，各文が何について書かれているかという情報であり，header info, header id, paragraph id, 及び sentence id は抽出された文がページ上のどこに存在したかという位置情報である．例えば，「冷凍庫に入れておくと長持ちする。」という文は，[バター, 保存法, <h2>, 05, 01, 00] という情報を持つことになる．

以上によって，トピックそれぞれにおける文の集合が構築される．抽出される文の数はトピックによって大きく異なり，100 以上の文が抽出できるトピックもあれば，1 文も抽出できないようなトピックも存在する．本稿では，10 以上の文が抽出できるトピックを対象として発話候補文の分類を行うことにする．

3.2 フィルタリング・口語体への変換

本項では，Wikipedia から抽出した生のテキストを発話文に変換するための下処理として，フィルタリング及び口語体への変換について説明する．フィルタリングとは，文に含まれる特定の表現を検出することで，対話システムに利用できない文を事前に排除することである．その後，残った文を口語体へ変換する．

現状の雑談対話システムでは，基本的にユーザとシステムの一問一答形式の対話を行う場合がほとんどであり，前の文脈を考慮して，気の利いた発言を行うことは困難である．よって，雑談対話システムが用意する発話文は，それぞれが単体で意味理解ができる文であることが要求される．一方，Wikipedia に存在する文は一連の長い文章の中に存在しており，前後の文脈を考慮して書かれた文が多数存在する．

表 1: フィルタリングの対象となる単語（例）

現在	現代	上述	先述	以下	再度
上記	同様	参照	次の	当時	本稿

表 2: フィルタリングの対象となる文頭の表現

品詞	接続詞	代名詞	助詞	
表現	“この〜”	“その〜”	“あの〜”	“どの〜”

表 1 にあるような単語が含まれるような文は，前後の文脈を想定しており，これらの単語が指し示す対象がその一文には存在しない場合がほとんどである．そのため，文単体では意味理解ができない．本手法ではこういった単語が含まれる文を除外する．また，文頭の表現が表 2 に該当する場合も除外する．これらを文頭に含む文も，前後の文脈ありきの文であることや，編集者の表現方法 (もしくはミス) によってそもそも文として成立していない場合が多い．

文字数によるフィルタリングも行う．文字数が短い文は，当たり前すぎたり（e.g.「箸を使って食べる。」），文として成立していない場合が多い．逆に文字数が長い文は，一文に情報を詰め込みすぎており，ユーザが理解するのに苦労したり，不快感を覚える可能性がある．こういった文は雑談対話システムに利用する文としては不適切なので，極端に文字数が長い文や短い文は除外する．なお，Wikipedia から一文として抽出した段階で，8 文字以下または 80 文字以上の文を対象とする．

次に，口語体への変換を行う．Wikipedia は文語的な表現で書かれており，ここから抽出したテキストをそのまま対話システムに出力すると，非常に硬い表現となり違和感を覚える．そのため，これを口語的な表現へと変換する必要がある．本手法では，変換ルールとして以下のルールを作成し，口語体への変換を行う．ユーザが違和感を覚えない発話文にするために，変換をするべきポイントは，主に，読点 (、) の直前および句点 (。) の直前の 2 点である．例えば，読点前が動詞の場合には，「[動詞の活用形] + “て、”」となるように変換し，逆接の接続助詞 (“が”) である場合には，「“〜ですけど”」などのように変換することで表現が柔らかくなる．また，文末には句点が存在するので，その直前をですます調にする必要もある．

このルールを用いることで，「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。」と変換される．

[口語体への変換ルール]

1. **読点 (、) の前**

 1.1 **読点前が動詞**

  1.1.1 活用形が五段の「ガ行, ナ行, バ行, マ行」
  → 「連用タ接続＋“で”」

  1.1.2 活用形が [1.1.3] 以外 → 「連用タ接続＋“て”」

  1.1.3 「“おり”」(‘おる’ の連用形) → 「“いて”」

1.1.4 「“し”」(サ変の連用形) → 「“して”」

1.2 **読点前が形容詞**

1.2.1 活用形が「連用形」 → 「形容詞＋ “て”」

1.3 **読点前が接続助詞「“が”」**

1.3.1 「動詞＋ “が”」
→「動詞（連用形）＋ “ますけど”」

1.3.2 「動詞＋ “たが”」
→「動詞＋ “たんですけど”」

1.3.3 「形容詞＋ “が”」 → 「形容詞＋ “ですけど”」

1.3.4 「“であるが”, “だが”」 → 「“ですけど”」

1.3.5 「“だったが”」 → 「“でしたが”」

1.3.6 「“たが”」 → 「“たんですけど”」

1.3.7 「“ないが”」 → 「“ないんですけど”」

1.4 **その他の処理**

1.4.1 「“であり”」→ 「“で”」

2. **文末（句点 (。) の前）**

2.1 「名詞＋ “という”」 → 「“んですよ”」追加
2.2 「名詞, 形容詞, 副詞」 → 「“ですよね”」追加
2.3 「動詞, 動詞＋助詞, 動詞＋助動詞, “形容詞＋助動詞”」
→ 「“んですよ”」追加
2.4 「助動詞＋ “という”」
→「 助動詞＋ “そうですよ”」
2.5 「“である”」 → 「“なんですよ”」
2.6 「“であった”, “だった”」 → 「“だったんですよ”」

3. **その他の処理（文中）**

3.1 「“であった”」 → 「“だった”」

以上が変換のルールとなる．また，文中にトピック名が存在しない場合には，トピック名の補完を行う．文中にトピック名が存在しない文に対して，その文の文頭に「“[トピック名] って，”」を付け加える．例えば，“餃子” というトピックに属する文「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」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．

3.3 位置情報に基づく発話文の分類

ここでは，発話文の分類手法について説明する．分類の際に用いる特徴量として，文抽出時に得られる [header info, header id, paragraph id, sentence id] を利用する．分類に用いる教師データは以下の 3 つの基準のもとでアノテーションを行う．

(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) トピックについて書かれていない

これら 3 つの基準のどれかに当てはまる場合，利用不可能文であると判断する．以降，対話システムに利用できる文を利用可能文，利用できない文を利用不可能文と定義する．上述したフィルタリングによって，基準「(1) 日本語として不自然である」に当てはまる文の多くを排除できる．よって，「(2) 一文で意味内容が理解できない」及び「(3) トピックについて書かれていない」に当てはまる文を分類によって排除する必要がある．Wikipedia の傾向として，段落や記事の後半に登場する文は前後の文脈に影響を受けやすく，またトピックに関する直接的な情報ではなくなる場合が多い．よって，基準 (2),(3) に影響を与える要因は，抽出された文の位置情報であると考えることができる．例えば「うどん」トピックの場合，記事の上部では，うどんそのものに関する基本的な情報を述べている一方で，記事の下部では「ほうとう」や「うどん用小麦」に関する情報に推移する．トピックが「うどん」である以上は，うどんに関する直接的な情報を取得するべきだと考える．

本提案手法では，それぞれの対話文が記事上で存在した位置が，トピックと対話文の距離 (関連性) に関係していると仮説を立てる．具体的には，文の抽出時に得られる header info, header id, paragraph id, 及び sentence id の 4 つで構成される位置情報を用いて，それぞれの発話文が雑談対話システムに利用できるかどうかを分類する．

4. 評価

本節では，提案手法の精度評価を行う．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，品詞の識別には形態素解析ソフトウェア「MeCab」[*1] を用いた．その後，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．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は，フライスの一致係数 κ を用いて $\kappa = 0.64$ 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をトピックで 5 分割（各 12 トピック）し，ロジスティック回帰による 5 交差検定を行なう．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一となるようにランダムに利用可能文を排除した．

利用する特徴量を表 3 に示す．(f1) はヘッダーの種類を示し，(f2) は文の所属するヘッダが上から何番目に登場するかを示す．(f3) は文の所属する段落がヘッダ中において何番目に登場するかを示し，(f4) は，所属する段落の何番目に文が登場するかを示している．

表 4 は，4 つの特徴量 (f1),(f2),(f3), 及び (f4) を利用する全ての組み合わせ (15 通り) における，分類の精度を示すものである．これを見ると，(f3) を除く 3 つの特徴量を利用した精度が 0.655 となり最も高くなった．また，(f3) のみを用いた分類精度が極端に低い値 0.500 になり，(f3) は分類に利用する特徴量として不適切であることがわかる．

図 1 は最も分類の精度が高かった (f1),(f2), 及び (f4) を用いた場合の ROC 曲線である．これは分類の閾値を 0.01 間隔で [0, 1.0] の範囲を変化させたときの，利用可能文抽出の真陽性率 (True Positive Rate) 及び偽陽性率 (False Positive Rate) を示す．図 1 の線上にある点は，閾値 0.41 において精度が最大値である 0.676 をとった点を示す．

5. 考察

本手法では，Wikiepedia の位置情報を用いて雑談対話システムに利用する発話文の分類を行なった．分類精度は実用的とは言えないものの，位置情報という単調な情報が，複雑な要因

*1 http://taku910.github.io/mecab/

表 3:

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

表 4: 特徴量の組み合わせによる分類精度の変化

特徴量	精度	特徴量	精度
f1	0.640	f2+f4	0.653
f2	0.645	f3+f4	0.602
f3	0.500	f1+f2+f3	0.640
f4	0.604	f1+f2+f4	0.655
f1+f2	0.640	f1+f3+f4	0.631
f1+f3	0.640	f2+f3+f4	0.648
f1+f4	0.629	f1+f2+f3+f4	0.652
f2+f3	0.637		

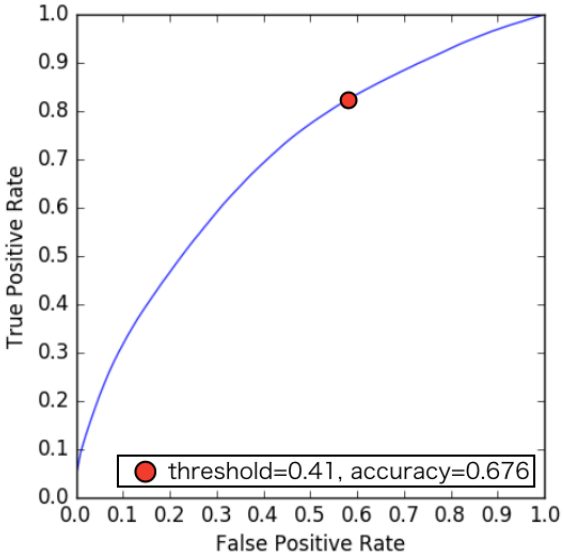

図 1: 特徴量「f1+f2+f4」を利用した場合の ROC 曲線

によって決定される雑談文の利用可能性と一定の関連性を持つことを示した．学習によって得られた位置情報による尤度を利用し，発話候補文に対してランキングすることで，より確実に利用可能文を抽出することも可能である．

6. おわりに

本稿では，特定ドメインの雑談対話システムのための発話候補文集合の獲得を行うために，Wikipedia からの発話文の抽出を行い，その発話文が持つ記事上での位置情報を特徴量として発話文の分類を行なった．実用に十分な精度とは言えないが，全文抽出と比較するとわずかではあるが良い結果を得ることができた．

今回の評価実験は 10 文以上の文を抽出できるトピックに限定したものであり，数文しか抽出できないようなトピックにおいては結果が異なる可能性がある．抽出文の数が少ないトピックはヘッダ数も少ないため，位置情報が重要であるとは考えにくい．よって，このようなトピックに対しては異なる分類手法が必要となる．

今後の課題として，次のことがあげられる．分類に利用する特徴量を位置情報のみに限定した場合，分類の精度が実用レベルに達しない．よって，位置情報とは異なる何らかの特徴量を追加することで，分類精度を向上させる必要がある．アノテーションデータでの三人の評価者の一致率が $\kappa = 0.64$ という低い値であることから，原理的に難しい問題であると考えられるため，より明確なアノテーションの基準を設けることで，問題設定をより明確にする必要もある．

また，Wikipedia から抽出される文には専門的な文が多いため，一文に対する情報量が多くなりがちである．そのため，ユーザは発話文を理解するのに時間がかかり不快感を覚えることも考えられる．本研究では，このような発話文を「マニアックな文」と定義し，今後の研究においてはマニアックな文を分類によって排除する手法も検討する．

本稿では独自の口語体変換ルールを用いたが，これが完全なものであるとは言い難い．将来的には [鍛治 04] で示されているような「言い換え」の技術を利用することも検討したい．また，料理ドメイン以外の他ドメインに対しても適用可能であるかどうかの検討はされていない．よって，本手法の他ドメインへの拡張性も検証する必要がある．

参考文献

[東中 14] 東中 竜一郎 ： 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 通将, 平井 尚樹, 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, vol. J95–D(6), pp. 1390–1400. 2012.

[磯村 09] 磯村 直樹, 鳥海, 石井 ： HMM による非タスク指向型対話システムの評価, 電子情報通信学会論文誌, vol. J92–D(4), pp. 542–551. 2009.

[水上 16] 水上 雅博, Lasguido Nio, 木付 英士, 野村 敏男, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 戸田 智基, 中村 哲 ： 快適度推定に基づく用例ベース対話システム (BibTex) 人工知能学会論文誌, 31–1. 2016 年 1 月.

[日下 12] 日下 九八 ： ウィキペディア ： その信頼性と社会的役割, 情報管理, vol. 55, no. 1, pp. 2–12. 2012.

[稲葉 14] 稲葉 通将, 神園 彩香, 高橋 健一 ： 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, vol. 29, no. 1, pp. 21–31. 2014.

[太田 09] 太田 知宏, 鳥海 不二夫, 石井 健太郎 ： 発話生成を目的とした Wikipedia からの文抽出, 人工知能学会全国大会講演論文集, 2009.

[鍛治 04] 鍛治 伸裕, 岡本 雅史, 黒橋 禎夫 ： WWW を用いた書き言葉特有語彙から話し言葉語彙への言い換え, 自然言語処理, vol. 11, no. 5, pp. 19–39. 2004.