ウィキペディアから抽出した人物エピソードの話し言葉への変換

柳 将吾　　佐藤 理史
名古屋大学大学院工学研究科
yanagi.shogo@i.mbox.nagoya-u.ac.jp

1 はじめに

対話エージェントに、より多くの話題や内容についての発話生成能力を持たせるための方法の一つに、外部情報源から取得したテキストを利用する方法がある。ニュースやウィキペディアなどの外部情報源のテキストは、一般に書き言葉で記述されているため、取得したテキストをそのまま用いると違和感が生じる。そのため、書き言葉で書かれたテキストを、話し言葉らしい形式に変換する必要がある。

我々は、このような変換の具体例として、ウィキペディアから抽出した人物エピソード (トリビア文) を、より話し言葉らしい形式の発話 (トリビア発話) に変換する問題に取り組んだ。理想的な変換を実現するためには多くの種類の変換が必要となるが、それらのうち、特に重要かつ必須と考える (1) 当該選手名補完、(2) 尊称挿入、(3) 文末変換、を実装した。本稿では、これらの内容について報告する。

2 システムの概要

図 1 に、想定するトリビア発話の利用シナリオを示す。このシナリオでは、特定のスポーツ競技に関する対話の終了時に、話題の中心となった選手のトリビアを、エージェントが発話することを想定している。この最後の発話を、ウィキペディアから抽出した文から自動生成することが、本研究の目標である。

作成したシステムの全体像を図 2 に示す。本システムは、Juman/KNP [1, 2]、および、MeCab (NEologd 辞書)[3, 4, 5] の 2 つのツールを利用する。システムの中核は、入力文をどのように編集すべきかを決定する 3 つのモジュール (当該選手名補完、尊称挿入、文末変換) である。これらのモジュールが生成した文字列置換コマンドを書き換えモジュールが実行し、出力文を生成する。以下では、中核となる 3 つのモジュールと、その性能評価について述べる。性能評価には、ウィキペディアから抽出したトリビア文 199 文を用いた。

A はエージェント (システム)、U はユーザーを表す

A　明日は、(ユーザー) さんが好きな卓球で、女子ダブルス三回戦、平野美宇選手の試合がありますね。
U　会場はどこ？
A　東京体育館で実施されるようです。
U　テレビで放送されるの？
A　XYZ で 15 時から生中継されるみたいですよ。
U　勝てそうかな？
A　対戦選手は、中国の**選手、++選手ですね、世界ランクは**位みたいです。
U　強そうだね、応援しないとね。
A　応援しましょう。
U　... (発話せず)
A　**平野選手は、2005 年 10 月号卓球王国の表紙を飾るなど、かつての福原愛さんのように、幼い頃から卓球がうまく、負けそうになると泣いて悔しがったので、マスコミから「第 2 の愛ちゃん」と呼ばれ一躍有名になったそうです。**

↑

2005 年 10 月号卓球王国の表紙を飾るなど、かつての福原愛のように、幼い頃から卓球がうまく、負けそうになると泣いて悔しがったので、マスコミから「第 2 の愛ちゃん」と呼ばれ一躍有名になった。
出典: ウィキペディア (平野美宇)

図 1: 想定するトリビア発話の利用例

3 当該選手名補完

ウィキペディアの文の大半は、ページタイトルを主題とする文であるが、日本語では、その主題は記述されないのが普通である。一方、図 1 に示したシナリオでは、「平野選手は」のように、誰のことを話しているのかを明示的に発話する必要がある。

当該選手名補完は、以下のように行う。

1. 「＜当該選手名＞+助詞『は・の・が・に』」が存在する場合、選手名の直後に「選手」を挿入する。(正確には、そのような編集コマンドを生成する。以下も同様である。)
2. 「彼女・彼 + 助詞『は・の・が』」が存在する場合、「彼女・彼」を当該選手名に書き換え、その直後に「選手」を挿入する。

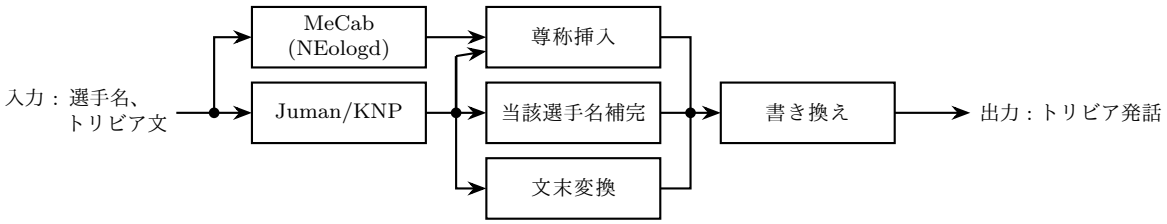

図 2: 書き言葉から話し言葉への変換システムの全体像

3. それ以外の場合、「＜当該選手名＞+選手+助詞『は・の』」を文頭に挿入する。助詞は、ほとんどの場合は「は」を選択するが、文頭の名詞が当該選手の属性を表すような名詞 (好物, 愛称, 趣味, 獲得賞金, 兄, 実家など) の場合は、「の」を選択する。

性能評価　トリビア文 199 文のうち、当該選手名、または、「彼女・彼」を含む文は 19 文 21 箇所あり、そのうち、15 文 18 箇所を正しく変換できた。残りの 180 文のうち、当該選手名を挿入すべき文は 170 文存在し、164 文に対して、許容できる助詞で当該選手名を補完することができた。

4　尊称挿入

本研究が想定する対話エージェントは、敬体 (です・ます体) を用いるエージェントである。このようなエージェントが尊称をつけずに他者に言及すると、強い違和感を覚える。ウィキペディアでは、ほとんどの場合、人名は尊称なしで現れるため、人名の直後に尊称を挿入する必要がある。

尊称挿入を実現するためには、人名検出と尊称選択が必要である。人名検出は、固有名検出の一部としてよく研究されているが、そこでは、尊称が人名検出の有力な手がかりのひとつとなっている [6]。尊称挿入のための人名検出では、尊称という手がかりなしに、人名を検出しなければならないため、一般の人名検出より難易度が高い。

さらに、誤検出の問題も存在する。たとえば、Juman/KNP は例文 (1a) の「古巣久光」を人名 (NE:PERSON) と判定し、MeCab (NEologd) は例文 (1b) の「雲梯」を人名 (固有名詞-人名) と判定する。

(1) a. 狩野舞子は、ウイングスパイカーとしてロンドンオリンピック代表に選抜されたが、その後、古巣久光製薬スプリングスにセッターとして復帰することが発表され注目を集めた。
b. 幼児教室の講師を務める母親が、脳の発達に雲梯が良いと本で読んだことから、生まれてすぐに雲梯に取り組んだ。

そこで、次のような方針を採用した。

1. 複数のツールを利用する。具体的には、Juman/KNP と MeCab (NEologd) を利用する。
2. それらのツールを全面的に信用せず、独自の人名判定を追加する。

具体的な方法は、以下のとおりである。

1. ツールによる候補収集
   KNP が NE:PERSON を付与した形態素列と、NEologd が固有名詞と判定した形態素 (列) を取り出し、それぞれの文字列を取得する。
2. Wiki 辞書による判定
   それぞれの候補に対し、後述する Wiki 辞書を用いて採用するかどうかを決定する。採用する場合は、挿入すべき尊称も決定する。
3. 候補の統合
   残った候補のうち、オーバーラップしているものを、ひとつにまとめる。

Wiki 辞書は、ウィキペディアの全タイトルに対し、それが人名かどうかの情報を付与した辞書である。この辞書は、各タイトルに対し、3 ビット (000_2 から 111_2) のスコアを保持する。それぞれのビットは、以下の事実に対応する。

b_0 記事の最初の文 (定義文) の最後の名詞句が職業を表す名詞句か否か
b_1 定義文の括弧内に年月日・月日が存在するか否か
b_2 記事の先頭の Infobox が人物を表す Infobox か否か

さらに、Wiki 辞書は、スポーツ選手か否かの情報も保持する。これは、b_0 と b_2 の決定で用いた手がかりに基づく。

表 1: 尊称挿入の結果

	OK	未挿入	誤挿入	R	P
KNP	45	30	7	60	87
NEologd	66	10	15	87	81
両者の統合	72	8	11	90	87
提案法	73	7	8	**91**	**90**

Wiki 辞書を用いた人名判定は、以下のように行う。

1. 確からしい候補 (KNP の NE:PERSON と NEologd の固有名詞-人名) でも、明らかに人名でない (スコアが 2 以下) 場合は採用しない。
2. それ以外の候補 (NEologd の人名以外の固有名詞) は、明らかに人名 (スコアが 7) の場合のみ採用する。

尊称としては、人物 (人名) がスポーツ選手である場合は「選手」、それ以外の場合は「さん」を用いる。

性能評価　表 1 に、尊称挿入の実験結果を示す[1]。この表の NEologd は、固有名詞-人名のみの結果であり、「両者の統合」は、KNP と NEologd での結果を統合したものである。2 つのツールの結果を統合すると、オーバーラップを一つにまとめる過程で、誤検出が減少する場合がある[2]。この表より、Wiki 辞書によりさらなる性能向上がもたらされていることがわかる。

5　文末変換

話し言葉は、書き言葉とは異なる文末形式をとるのが普通である。ここでは、話し言葉らしい文末形式にするため、4 つの変換を実装した。文末述語部の同定と解析には述語部解析システム Panzer [7] を用い、生成には HaoriBricks3 [8] を用いた。以下では、常体・敬体変換を除く 3 つの変換について説明する。

5.1　体言止めに対する述語補完

「読み上げの際に体言止めを用いると唐突で高圧的な印象を与える」[9] と指摘されているように、体言止めの形式は、話し言葉にはそぐわない。

そこで、文末が体言の場合、以下のように述語を補完する。

1. 体言がサ変名詞以外の場合は、「だ」を補完する。
2. 体言がサ変名詞の場合は、直前の文節を調べ、連体修飾の形式ならば「だ」を、それ以外の場合は「する」を補完する。

述語補完の具体例を以下に示す。

(2) a. 趣味は足のネイルアートだ　(普通名詞)
　　b. 趣味はバナナ・グッズの収集だ
　　(連体修飾の形式+サ変名詞)
　　c. 2003 年、近畿大学に進学する　(サ変名詞)

性能評価　199 文中、体言止め補完が必要である文は 57 文存在し、そのうち 54 文が正しい述語を補完できた。残りの 3 文の誤りは、次の理由による。

1. 体言止めでは、ボイス (能動態・受動態) の情報が欠落する。欠落しているボイスが受動態の場合は、正しく復元できない。(「観光大使に選出 (された) 」)
2. 体言止めでは、しばしば直前の格要素が複合名詞化される (「1 次予選を突破 (した)」→「1 次予選突破」) が、省略された格助詞は復元できない。

5.2　テンスの変更

書き言葉では、過去の事実がテンス無標 (現在形) で書かれる場合がある。

(3) リオデジャネイロオリンピックでは、準々決勝戦ブレイディ・エリソンとの試合で敗退し 8 位に終わる。

しかしながら、話し言葉では、過去の事実はテンス有標 (タ形) で伝えるのが普通である。

当然のことながら、トリビア文には、現在のことを記述した文もある。

(4) また、お菓子作りも趣味にしている。

ここでは、述語が状態を表すのか、それとも動作を表すのかに基づいて、必要に応じてテンスを変更する。具体的には、

- テンスが無標の場合で、かつ、述語が動作を表す述語の場合、テンスを有標に変更する。

述語は、以下の場合に、状態を表すと判定し、それ以外の場合は、動作を表すと判定する。

1. 述語が形容詞または判定詞の場合
2. 述語が動詞で、かつ、状態動詞 (いる、ある、由来する、など) のリストに含まれる場合

[1] 尊称選択の正否は問わない。
[2] 一方がファーストネーム (「チャールズ」) を検出し、他方が人名全体 (「チャールズ・クレイグ」) を検出した場合など。

表 2: 伝聞モダリティ付与の結果

	○	△	×	計
伝聞モダリティ存在文	—	—	—	4
伝聞モダリティ付与文	126	15	20	161
伝聞モダリティ非付与文	29	4	1	34
計	155	19	21	199

性能評価　199 文中、テンスが無標の文が 127 文存在した。このうち、正しく有標に変更した文が 45 文、正しく無標のままとした文が 76 文、不適切に有標に変更した文が 6 文あった。エラーは、状態動詞リストの不備による。

5.3　伝聞モダリティ付与

トリビア文には、よく知られた事実を伝える文と、あまり知られていない人物の情報 (たとえば、趣味) やエピソードを伝える文がある。前者は、断定のモダリティで問題ないが、後者は、情報源がウィキペディアであることも考慮すると、伝聞のモダリティで発話することが望ましい。つまり、どちらのモダリティを採用すべきかは、文の事実性の度合いに依存する。

伝聞モダリティ付与では、原則として、文末に伝聞の助動詞「そうだ」を付与する。ただし、「文の事実性が高い」と判定した場合は、付与しない。

事実性の判定には、ヒューリスティックを用いる。今回のトリビア文の対象はスポーツ選手なので、オリンピックや世界選手権の入賞などを伝える文を事実性が高い文をみなし、たとえば、大会名と「〜メダル」という単語が同時に含まれる場合は事実性が高いと判断する。

性能評価　伝聞モダリティ付与の評価結果を表 2 に示す。伝聞モダリティの付与の適切さは、グレーな部分も多いため、許容の段階を○、△、×の 3 段階で判定した。199 文中、トリビア文にすでに伝聞モダリティを表す助動詞が存在した文が 4 文あり、残りの 195 文が付与対象となる。これらのうち、161 文に「そうだ」を付与し、34 文には付与しなかった。明らかに誤りと考えられる文は合計で 21 文であり、大部分の文に対しては、妥当な変換が得られた。なお、現在の事実性判定ヒューリスティックはスポーツ分野に過度に特化しているため、汎用性に乏しい。

6　変換の具体例

以下に、トリビア発話への自動変換例を示す。

(5) a. 趣味は足のネイルアート
　　b. 三宅宏実選手の趣味は足のネイルアートだそうです。
　　(当該選手名補完「の」, +だ, +伝聞, +敬体)

(6) a. リオデジャネイロオリンピックでは、準々決勝戦ブレイディ・エリソンとの試合で敗退し 8 位に終わる。
　　b. 古川高晴選手は、リオデジャネイロオリンピックでは、準々決勝戦ブレイディ・エリソンさんとの試合で敗退し 8 位に終わりました。
　　(当該選手名補完「は」, 「さん」挿入, +タ, +敬体)

(7) a. ピアノも得意で、相当の腕前である。
　　b. 入江陵介選手は、ピアノも得意で、相当の腕前だそうです。
　　(当該選手名補完「は」, +伝聞, +敬体)

謝辞　本研究は、トヨタ自動車株式会社との共同研究として実施した。

参考文献

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[3] https://taku910.github.io/mecab/
[4] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会発表論文集, pp. 875–878, 2017.
[5] https://github.com/neologd/mecab-ipadic-neologd
[6] 竹元義美, 福島俊一, 山田洋志. 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出. 情報処理学会論文誌, Vol. 42, No. 6, pp. 1580–1591, 2001.
[7] 佐野正裕, 佐藤理史, 宮田玲. 文末述語における機能表現検出と文間接続関係推定への応用, 言語処理学会第 26 回年次大会発表論文集, 2020.
[8] 佐藤理史. HaoriBricks: ブロック玩具に学ぶ日本語文章生成ライブラリ. 言語処理学会第 23 回年次大会発表論文集, pp. 20–23, 2017.
[9] 林由紀子, 松原茂樹. 自然な読み上げ音声出力のための書き言葉から話し言葉へのテキスト変換. 情報処理学会研究報告. Vol. NL-179, pp. 49–54, 2007.