小説会話文の翻訳へ向けた逆翻訳を用いた話者埋め込みの作成

長門亜優奈　松崎拓也
東京理科大学 理学部第一部 応用数学科
1421089@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

日本語の発話には，性別や性格などの発話者のキャラクター性が現れることが多い．一方，英語の発話は必ずしもそうではない．そのため，英語の小説を日本語に機械翻訳する際に，発話は発話者のキャラクター性に反する表現で翻訳されることがある．これを防ぐ方法として，発話の翻訳の際に，話者の特徴を反映した話者埋め込みを入力に加えることが考えられる．本論文では，その準備として，日本語の発話文における発話者の特徴の現れている部分をマスクし，その部分を当てることで話者埋め込みを訓練することを試みた．その結果，性別などのキャラクター性を話者埋め込みとして抽出できること，また，発話文における発話者の特徴の現れている部分を推測することに話者埋め込みの利用が効果的であることを示す．

1　はじめに

小説において，登場人物には性別や性格といったキャラクター性があり，特に日本語の発話にはそのような特徴がよく現れる．例えば，①「私は元気です」と②「俺は元気だぜ！」という２つの発話は同じ内容を持つが，①は女性的で丁寧，②は男性的で活発といった印象を受ける．一方で，英語では，①と②のどちらも “I’m fine.” と表現されるように，発話に登場人物の特徴が日本語ほど現れない．そのため，英語の小説を日本語に機械翻訳した際に，登場人物の発話として相応しくない文となる場合がある．この問題を解決することによって，小説の翻訳のためのより良い下訳を得ることができる．また，異なる言語の話者同士が翻訳を用いて自然な会話ができるようになるといった応用もありうる．

発話のキャラクター付けに関しては ChatBot へのキャラクター付けや，音声合成での話者埋め込みの利用が試みられてきた．ChatBot へのキャラクター付けについては，キャラクター応答の生成にポインタ生成機構を適用し，複数の異なるキャラクター応答を参照しながら少量のデータからキャラクター性をもった会話を実現する手法 [1] が試みられている．また，音声合成での話者埋め込みの利用については，目的話者の音声波形を用いて話者埋め込みを得る手法だけでなく，人間の知覚評価をフィードバックに用いて話者埋め込みを探索する人間参加型の手法 [2] が試みられている．ただし，これらの研究はいずれも，単一言語を対象としている点で，複数言語を対象とする翻訳に利用することを想定している本論文の手法とは異なる．

本研究の長期的目標は，英語の小説を日本語に翻訳する際に，英語の小説から発話者埋め込みを抽出することによって，翻訳された日本語の発話に登場人物のキャラクター性を反映させることである．本論文では，その第一段階として，逆翻訳 (Back Translation) を用いて，英語の小説から発話者埋め込みを生成する手法を提案する．キャラクター性を発話の翻訳に反映させるための話者埋め込みとしては，各話者の発話を日本語訳したときに終助詞 (ぞ・な・わ 等) や一人称の選択に現れる特徴を予測するための情報を，源言語 (英語) における人物描写から抽出すればよいと考えられる．そのような埋め込みモデルの学習データとしては小説の対訳が必要であるが，利用できる小説の対訳データは少ない．そこで， Sennrich ら [3] によって提案された，不十分な学習データに対応するための手法である逆翻訳を応用する．つまり，日本語の小説を英語に自動的に翻訳することで対訳データを作成し，そのデータを用いて英語の小説から発話者埋め込みを生成するモデルを訓練する．

青空文庫のデータを用いた実験の結果，話者の性別や著者の文体といった特徴を話者埋め込みとして抽出することに成功した．また，話者埋め込みによって発話のキャラクター性の現れる部分を推測する精度が向上した．

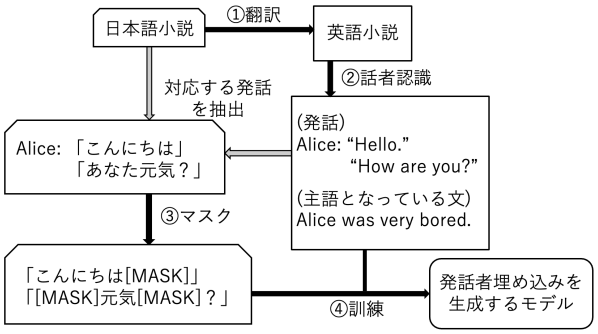

図 1　提案手法の流れ

2　手法

提案手法の処理の流れは図 1 の通りである．まず，与えられた日本語の小説のデータを英語のデータに翻訳した後，StanfordCoreNLP[1] を用いて，英語データ内の発話と発話者，および各発話者が主語となっている文を抽出する．次に，抽出された発話に対応する日本語の発話に対し，発話者の特徴が表れている部分を特殊トークン [MASK] で置き換え，置き換え前の語句を推測することで話者埋め込みを訓練する．以下，手法の詳細を述べる．

2.1　青空文庫の日英機械翻訳

利用できる日本語と英語の小説対訳データは多くは存在しない．そのため，日本語の小説を英語に翻訳することにより，対訳データを作成した．具体的には，青空文庫[2] の小説 13,772 編を，JParaCrawl version 3 [4] を用いて訓練された transformer モデル (large モデル)[3] を用いて英語に翻訳した．

2.2　話者認識

作成した英語の小説データから，発話とその話者，および各話者が主語となっている文を抽出した．具体的には， StanfordCoreNLP を利用して，発話者認識処理の結果から話者と発話内容の組を抽出し，係り受け解析の結果からある動詞の主語 (nsubj) が抽出した話者となっている文を抽出する．次に，日本語の小説から，英語の小説で抽出された発話に対応する文を抽出することにより，発話の日英対訳データを作成する．

2.3　発話のマスク

抽出された日本語の発話に対し，キャラクター性が現れやすい，以下の部分をマスクする．

代名詞　一人称や二人称のような代名詞をマスクする．具体的には， MeCab によって「代名詞・一般」とタグ付けされた単語を [MASK] に置き換え，その単語を正解として保存する．

例：私は元気です． → [MASK] は元気です．
（正解）私

終助詞　文末で話者の態度を示す，「ぞ」「な」といった終助詞をマスクする．具体的には，MeCab によって「終助詞」とタグ付けされた単語を [MASK] に置き換え，その単語を正解として保存する．

例：大丈夫ですよ． → 大丈夫です [MASK].
（正解）よ

敬語（動詞）　敬語の動詞と敬語表現が存在する動詞をマスクする．具体的には，青空文庫の全文から MeCab によって「動詞」とタグ付けされた単語のうち，敬語であるものと敬語表現が存在するものを抽出し， [MASK] に置き換える．その単語を正解，その単語の対応する敬語表現（もしくは敬語でない表現）を不正解として保存する．

例：先生は仰った．→先生は [MASK] た．
（正解）仰っ，（不正解）言っ

その他の文末表現　発話の文末のうち， MeCab によって「終助詞」とタグ付けされないものについて，「です」,「ます」,「だ」を [MASK] に置き換え，その単語を正解として保存する．

例：元気です． → 元気 [MASK].
（正解）です

文末が MASK でないものは，文末に [MASK] を挿入し，空文字列を正解として保存する．

例：私は元気！→私は元気 [MASK] ！
（正解）“”

2.4　話者埋め込みモデルおよび訓練方法

話者の埋め込みとして，ある話者のキャラクター性を知る手がかりとなるであろう文章を英語 BERT に入力したときの [CLS] ベクトルを用いる．より具体的には，2.2 節で述べた特定の話者の発話およびその話者を主語とする節を含む文を，小説中での出現順に並べたものを $t_1, t_2, \ldots, t_n$ とするとき，これらを [SEP] トークンを挟んで連結した t_1 [SEP] t_2 [SEP] . . . [SEP] t_n のうち，BERT への最大

1）https://stanfordnlp.github.io/CoreNLP/
2）https://www.aozora.gr.jp
3）https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl

入力長である 512 トークンを先頭から切り出したもの（これを $\mathbf{t}$ とする）を英語 BERT（bert-base-uncased）への入力とする．以下では上記の操作で得られた [CLS] ベクトルを e-BERT$_{\mathrm{CLS}}(\mathbf{t})$ と表記する．

上記の，英語 BERT を用いた話者埋め込みモデルの訓練は以下のように行う．まず，話者 s の日本語の発話に対し，2.3 節で述べたマスク処理を行なったものを日本語 BERT に入力する．日本語 BERT としては，東北大乾研究室によって訓練されたもの（bert-base-japanese-v3）を用いる．次に，入力の各トークンに対する日本語 BERT の出力ベクトルそれぞれに対し，話者 s の埋め込み e-BERT$_{\mathrm{CLS}}(\mathbf{t})$ を加え，語彙中の各トークンに対するスコアへと変換するモジュール（いわゆる言語モデルヘッド）に入力し，マスクされた語の推測を行う．空文字列が正解である [MASK] については，日本語 BERT の語彙のうち未使用のもの（[unused0]）を正解とする．損失関数としては交差エントロピーを用い，マスクされたトークンに対してのみ損失を計算する．

訓練の際は，上で述べた処理モジュールのうち日本語 BERT，英語 BERT，言語モデルヘッドの全てをパラメータチューニングの対象とする方法と，このうちいくつかのパラメータを固定する方法が考えられる．これらについては実験で比較する．

3 実験

実験では，青空文庫から作成した対訳データのうち話者埋め込み訓練時にエポック数の決定にのみ使用した話者の発話 (話者数 6,449, 発話総数 72,762) に対して訓練した話者埋め込みモデルを適用し，その結果を観察した．また，話者埋め込みを用いた場合とそうでない場合での [MASK] の穴埋めの正解率を評価した．さらに，ProjectGutenberg [4] の小説 22,001 編に対して訓練された話者埋め込みモデルを適用し，その結果を観察した．話者埋め込みの観察は表 2 の 3 つの設定のうち，すべてのパラメーターをチューニングした場合のモデルを用いた．

3.1 青空文庫作品の話者埋め込み

英語に機械翻訳した青空文庫の小説に対して，訓練した話者埋め込みモデルを適用し，得られた話者埋め込みに対して主成分分析 (PCA) を行った．表 1 は第 1 主成分得点が上位と下位の話者 10 人ずつとその作品・著者をまとめたものである．また，図 2

4） https://www.gutenberg.org

表 1 第 1 主成分の上位下位

下位		上位	
話者	作品	話者	作品
Kasuke	「入れ札」菊池寛	Kuroe	「キャラコさん」久生十蘭
Koshu	「大菩薩峠」中里介山	her	「キャラコさん」久生十蘭
Iori	「銭形平次捕物控」野村胡堂	Yoshie	「キャラコさん」久生十蘭
his	「寺坂吉右衛門の逃亡」直木三十五	Noriko	「杉子」宮本百合子
Yasuke	「影を踏まれた女」岡本綺堂	Haruko	「野ざらし」豊島与志雄
Isuke	「早耳三次捕物聞書」林不忘	Suzue	「ある夫婦の歴史」岸田国士
Iori	「宮本武蔵」吉川英治	her	「今朝の雪」宮本百合子
Yoriharu	「あさひの鎧」国枝史郎	Charako	「キャラコさん」久生十蘭
his	「余裕のことなど」伊丹万作	his	「杉垣」宮本百合子
Yamada	「私本太平記」吉川英治	Madam	「影のない犯人」坂口安吾

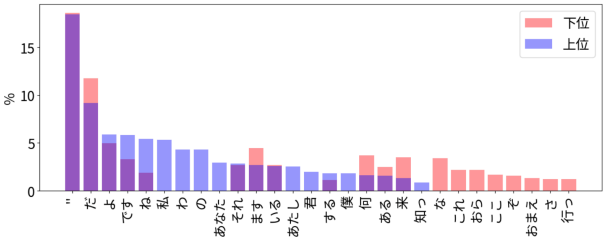

図 2 第 1 主成分の上位下位話者の発話の [MASK]

は第 1 主成分得点の上位 30 人の話者と下位 30 人の話者の発話において，マスクされた部分の単語の割合を示す．

表 1 の話者をみると，上位の話者は女性，下位の話者は男性が多いことが分かる．ここで, 図 2 において上位と下位いずれかのみに登場する単語に注目すると，上位の話者の発話には「わ」「の」「あなた」「あたし」，下位の話者の発話には「ぞ」,「おまえ」,「おら」が登場している．「あたし」「あなた」という単語は女性，「おら」,「おまえ」という単語は男性の一人称と二人称として一般的に用いられる単語である．また，「わ」「の」「ぞ」はいずれも終助詞であるが，「わ」「の」は女性，「ぞ」は男性に多く用いられる．これらのことから，第 1 主成分には話者の性別という特徴が表われていると考えられる．

また，表 1 の作品をみると，主成分得点の上位・下位それぞれで著者が同じ作品の話者が多いことが分かる．他の主成分得点についても上位や下位に著者が同じ作品の話者が多いものがあった．このことから，話者埋め込みは何らかの著者の文体の特徴も抽出していると考えられる．

3.2 穴埋めの正解率の評価

表 2 に話者埋め込みを用いた場合とそうでない場合での [MASK] の穴埋めの正解率を示す．評価データのうち，[MASK] の正解が空文字列のものは 39,141 件，[MASK] の正解が空文字列以外のものは 163,472 件であった．表 2 で事前訓練済み BERT の

表 2　[MASK] の穴埋めの正解率の評価

手法		[MASK] の正解が空文字列	[MASK] の正解が空文字列以外
事前訓練済み BERT のみ		—	29.47
話者埋め込みを利用	すべてのパラメータをチューニング	94.99	67.67
	日本語 BERT 本体のみ Freeze	94.25	54.44
	日本語 BERT 本体も言語モデルヘッドも Freeze	76.56	43.06

みを用いた (話者埋め込みを用いない) 場合と話者埋め込みを用いた場合を比較すると，話者埋め込みを用いた場合の方が正解率が高いことが分かる．次に，話者埋め込みを用いる場合のなかでは，[MASK] の正解が空文字列・空文字列以外のいずれもすべてのパラメーターをチューニングした場合・日本語 BERT 本体のみ Freeze した場合・日本語 BERT 本体も言語モデルヘッドも Freeze した場合の順で正解率が高かった．このうちどの設定で訓練した話者埋め込みモデル (英語 BERT) が発話の翻訳に最も有効かを検証するのは今後の課題である．一方で，事前訓練済み BERT のみを用いた場合と BERT 本体も言語モデルヘッドも Freeze した場合を比較すると，後者の方が正解率が約 1.5 倍高く，発話者ベクトルが発話のキャラクター性が現れる部分を推測しやすくしていることが分かる．

3.3　ProjectGutenberg 作品の話者埋め込み

話者埋め込みモデルは，2 節で述べたように逆翻訳によって作成したデータで訓練した．この項では，このモデルをもともと英語で書かれた作品を収集した ProjectGutenberg 中の小説に適用した結果について述べる．まず，3.1 項と同様に話者埋め込みに対する PCA を行なった．その結果，青空文庫と同様に第 1 主成分得点に関しては，下位には男性の話者が多く，上位には女性の話者が多いという特徴が表れた．また，いくつかの主成分得点の上位下位には著者が同じ作品の話者が表れた．

キャラクター性の近い人物の話者埋め込みは近い値をとるはずである．そのため，話者埋め込みの cos 類似度を測ることにより，特定の人物に近い人物が分かると考えられる．表 3 は Lewis Carroll の “Alice’s Adventures in Wonderland” の登場人物 Alice に近い話者，すなわち話者埋め込みの cos 類似度が上位のものを並べたものである．表 3 の話者をみると，女性が多いことが分かる．特に，“Alice’s Adventures Under Ground” のような異なる作品に登場する同一人物である Alice が多く，キャラクター性の近い人物の話者埋め込みは近い値をとっていることが分かる．また，著者が同じ作品の話者がいくつかあり，話者埋め込みは何らかの著者の文体の特徴も抽出していることがこのことからも分かる．

表 3　Alice に近い話者

cos 類似度	話者	作品
1.000	Alice	“Alice’s Adventures in Wonderland” (Carroll)
0.972	Pat	“Alice’s Adventures in Wonderland” (Carroll)
0.947	Alice	“Alice’s Adventures Under Ground” (Carroll)
0.899	ALICE	“Alice’s Adventures in Wonderland” (Carroll)
0.890	Tess Kenway	“The Corner House Girls Growing Up” (Hill)
0.885	DOROTHY DAINTY	“Dorothy Dainty at Glenmore” (Brooks)
0.885	Alice	“Alice’s Adventures in Wonderland” (Carroll)
0.885	Mary Hooper	“Betty Wales on the campus” (Dunton)
0.882	Alice	“The escape of Alice: A Christmas fantasy” (Starrett)
0.881	Amy	“Reels and Spindles: A Story of Mill Life” (Raymond)
0.881	her	“The Westminster Alice” (Saki)
0.881	Arabella	“Dorothy Dainty at Glenmore” (Brooks)
0.880	Elsa	“The Christmas Makers’ Club” (Sawyer)
0.877	Elsa Danforth	“The Christmas Makers’ Club” (Sawyer)

4　おわりに

本論文では，英語小説の会話文を日本語へ翻訳する際にキャラクター性を正しく反映するのための前段階として，逆翻訳を用いた話者埋め込みの作成の手法を提案した．この結果，作成した話者埋め込みには性別や著者などの特徴が抽出されており，話者埋め込みを用いることにより，発話におけるキャラクター性の現れる部分を推測する精度が向上した．今後は，発話におけるキャラクター性の現れる部分の推測の精度をさらに向上させ，本研究の長期目的である英語の小説を翻訳する際に，英語の小説から発話者埋め込みを抽出することによって，翻訳された日本語の発話に登場人物のキャラクター性を反映させるための研究を進める予定である．

参考文献

[1] 奥井颯平, 中辻真. ポインタ生成機構を用いたキャラクター応答生成の検証. 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 1I4GS201–1I4GS201, 2020.

[2] 宇田川健太, 齋藤佑樹, 猿渡洋ほか. 人間の知覚評価フィードバックによる音声合成の話者適応. 聴覚研究会資料= Proceedings of the auditory research meeting, pp. 297–302. 日本音響学会, 2021.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6704–6710, Marseille, France, June 2022. European Language Resources Association.