7V-02

短歌・俳句翻訳のための異文化特性に基づく代替名詞抽出手法の検討

中辻佑弥[†1]　中島伸介[†1]　田中克己[†2, 3]　河合由起子[†1]
†1 京都産業大学　†2 福知山公立大学　†3 京都大学

1　はじめに

日本における短文の詩として，短歌・俳句がある．これらは日本文化の理解や知識に基づいており，短文となる詩を異文化の人々に共感してもらうことは難しい．本研究では，それら詩の独自の表現を異文化に合わせた表現に代替する詩翻訳手法を検討する．

美野らは，文脈を考慮したニューラル機械翻訳の精度向上を目的とし，目的言語側の前文の参照訳と機械翻訳結果の両方を文脈情報として用いる手法を提案している．提案手法を利用することで，BLUE スコアにおいて有効性を明らかにしている [1]．また，平田らは，Transformer ベースのモデルである GPT-2 を用いた俳句生成を目的に，散文である青空文庫を事前学習に用いることで，俳句の各条件を満たす俳句の割合は減少するものの，日本語としての破綻がなく意味の通る俳句を生成できる割合を増加させることができることを明らかにしている [2].

これらの研究では，日本語での俳句の自動生成する手法や，文脈の前後関係から英語を学習する手法が提案されている．本研究の目的は，異文化の人々に共感を得ることができる俳句の自動翻訳であり，また，短歌・俳句などの詩における名詞に着目して他文化言語を学習する手法が特異点である．

これまで我々は，詩の表現手法における俳句・短歌の名詞の重要性に着目し，日本特有の名詞を異文化の名詞に代替する手法を提案してきた [3]．その際，代替名詞抽出の検証を行い，ドイツ人・メキシコ人の 2 人に評価実験を行った．代替名詞抽出の結果，適合率は 0.83，再現率は 0.71 となり，有効性の確認ができた．一方，ユーザ評価検証では，ユーザの母国で代替した名詞に共感していなかった．また，短歌・俳句における古語の代替が出来ていなかった．そのため，古語の追加学習が課題だと考えられる．本稿では，Word2Vec に古語の追加学習を行うことで，更に共感を伴う代替名詞抽出を行う．また，名詞による感情や事象を表現する手法となる短歌，俳句を対象に，抽出された代替名詞を評価検証する．

A Study of Alternative Noun Extraction Methods Based on Cross-Cultural Characteristics for Tanka and Haiku Translation
†1 Nakatsuji Yuya　†1 Najajima Shinsuke　†2, 3 Tanaka Katsumi
†1 Yukiko Kawai
†1 Kyoto Sangyo University
†2 Fukuchiyama Public University
†3 Kyoto University

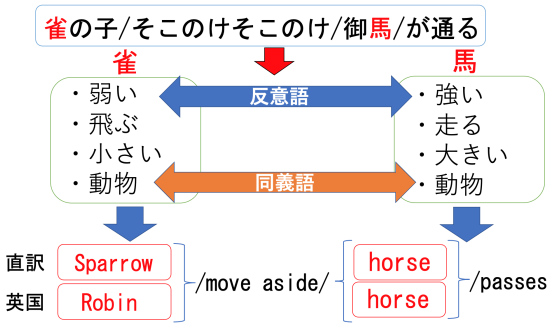

図 1: 詩（俳句）における翻訳の流れ

2　異文化特性の基づく代替名詞抽出システム

本研究では，短歌および俳句を形態素解析し，名詞を抽出する．抽出した名詞に対して Word2vec による和算演算によって異文化の人々が共感する代替名詞抽出を実現する．図 1 に詩翻訳システムの概要を示す．

まず，入力された短歌および俳句を形態素解析し，名詞を抽出する．短歌や俳句では，“美しい”や “素晴らしい”などの主観形容詞を排しつつ，名詞による感情や情緒を表現する．また，名詞の対比配置により感情等をより一層強調表現する．よって，本研究では名詞を対象とする．次に，抽出された名詞（例えば “雀”と “馬”）に対して，学習モデルやコーパスを用いて類義語を抽出する．抽出された類義語の同義語反意語を判定し，名詞と類義語（例えば雀に対する “弱い”，“飛ぶ”，“小さい”，“動物”）を用いて，和差演算より文化圏として国名を加えて，代替名詞を抽出する．最後に，原文を翻訳した後に原文の名詞を抽出した代替名詞に置換する．例えば，Google 翻訳では “Sparrow move aside horse passes”となるが，英国では “Robin move aside horse”となり，国ごとに典型的な名詞に代替され，共感が高くなることが期待できる．

2.1　Wikipedia と古語を含めた Word2Vec モデル生成

翻訳ための代替名詞抽出手法の流れを以下に示す．

1) 形態素解析より形態素に分解
2) 詩の音数形式により一般名詞を連結し名詞を抽出
3) 抽出された名詞の類義語集合を生成
4) 他の類義語集合と比較し，対比（同義語/反意語）する類義語を選定

表 1: 短歌 25 件，俳句 25 件における名詞総数

	形態素名詞総数	連結による名詞総数	辞書の名詞総数（古語無）	辞書の名詞総数（古語有）
短歌	158	96	44	60
俳句	103	69	30	35

表 2: 代替名詞抽出検証 (古語有無)

（古語無）	適合率	再現率	（古語有）	適合率	再現率
短歌（無）	0.814	0.443	短歌（有）	0.760	0.481
俳句（無）	0.846	0.431	俳句（有）	0.853	0.568
平均（無）	0.830	0.437	平均（有）	0.810	0.524

5) （4）の類義語と名詞と原文言語・翻訳言語国名により和差演算し代替名詞を抽出

本稿では詩翻訳として俳句と短歌を対象とするが，古語が多く，また助詞が排除されることが常用である．そのためコーパスや学習器に含まれず，名詞抽出および類義語抽出精度を低下させる．そこで，詩の音数形式（5-7-5，5-7-5-7-7）より名詞を連結し抽出する．例えば“天の原”は，“天（名詞-一般）”と“原（名詞-固有名詞-人名-姓）”が連結され抽出される．抽出された名詞より，閾値 Th_{wds} 以上を類義語集合とする．(4）では，詩に出現する複数の名詞の関係性を（3）の名詞の類義語集合より対比し検出する．本稿では，類義語集合から同義語または反意語を選定する．

2.2 和差演算による代替名詞抽出

生成したモデルを用いて，短歌と俳句の名詞の同義語および類義語を抽出し，原文言語国名および翻訳言語国名を用いて，和差演算より代替名詞を抽出する．名詞 w の代替名詞 w_{alt} は次の式で算出される．

$$w_{alt}(c_{alt}, w) = w + \sum_{i=0}^{n} w_i - \sum_{j=0}^{m} w_j + c_{alt} - c_{org} \quad (1)$$

ここで，c_{alt} は翻訳言語国名，c_{org} は原文言語国名である．また，w_i は名詞 w の類義語で n は w の同義語総数，m は w の反意語数である．例えば，詩に雀と馬が抽出された際の w=“雀”を英国文化として翻訳した際の代替名詞の算出式は，“雀” + {“動物”，“小さい”} - {“大きい”} + “英国” - “日本” となる．

3 代替名詞抽出結果の検証

提案手法により抽出した代替名詞を検証する．実験環境は，Python3，gensim（バージョン 3.8.3）を用いた．また，形態素解析器は MeCab，類義語抽出および演算は Word2vec，Wikipedia に古語[1]を追加し，50 次元のベクトルより，学習モデルを構築した．翻訳言語対象国は，“アメリカ”，“ドイツ”，“ブラジル”，“メキシコ”，“ニューカレドニア”とした．

[1]https://github.com/yoko-ot/Haiku

表 3: キリギリスの代替名詞 (代替名詞国：ドイツ)

同義・反意語	1	2	3	4
含まない	チェコ	オーストリア	フィンランド	ルーマニア
含む	アリゲーター	アザミ	モズ	モドキ

表 1 に抽出された名詞総数を示す．提案手法の音数形式連結と学習モデルより，適切に名詞が抽出された．ただし，“白妙” など辞書内に含まれない名詞は 4 割程あり，今後，古語を追加予定である．表 2 に，短歌・俳句より抽出された連結による名詞総数のうち，代替名詞の対象となる「花・地名・動物」を正解とした適合率，再現率を示す．古語を含むことで，適合率の平均は古語を含まない学習モデルより低くなったが，再現率は 0.524 と上昇した．F 値は，古語を含まない場合は平均は 0.573，含む場合は平均は 0.638 であった．以上の結果より，提案手法に古語を追加学習することで短歌・俳句より代替名詞の抽出精度の向上を確認できた．

表 3 は，抽出された名詞のうち，ドイツにおいて “キリギリス” を代替名詞にした例である．同義語と反意語を含まない場合は，国名が候補となっており，同義語と反意語を含む場合は，昆虫が抽出されており提案手法の有用性が確認できる．今後，ユーザ評価による検証を行う．

4 おわりに

本研究では，異文化の人々が共感を伴う翻訳を目的に，代替名詞抽出手法の検討を行った．提案手法の代替名詞抽出精度 F 値は 0.63 となり，古語の追加学習による F 値の向上が確認できた．今後，追加する古語を増やし，ユーザによる評価・検証を行う予定である．

謝辞

本研究の一部は，JSPS 科研費 19K12240，20H04293，22H03700 および京都産業大学先端科学技術研究所（M2001）の助成を受けたものである．ここに記して謝意を表す．

参考文献

[1] 美野秀弥，　伊藤均，後藤功雄，山田一郎，徳永健伸 “ニューラル機械翻訳での目的言語側の文脈の効果的な利用”，自然言語処理，Vol. 28, pp. 1162–1183 (2021).

[2] 平田航大，横山想一郎，山下倫央，川村秀憲ほか “Transformer による言語モデルを用いた俳句生成とその評価”，研究報告情報基礎とアクセス技術 (IFAT)，Vol. 2021, pp. 1–6 (2021).

[3] 中辻佑弥，中島伸介，田中克己，河合由起子 “他文化における共感を伴う詩翻訳手法の提案”，データベースシステム研究会 (2022).