翻訳精度に基づく固有名詞の翻訳手法の研究

高井公一 †,‡　服部元 ‡　米山暁夫 ‡　安田圭志 †　須藤克仁 †　中村哲 †
† 奈良先端科学技術大学院大学
‡ 株式会社 KDDI 総合研究所
takai.koichi.tc1@is.naist.jp, {ge-hattori, yoneyama}@kddi-research.jp
ke-yasuda@dsc.naist.jp, {sudoh, s-nakamura}@is.naist.jp

1　はじめに

近年における音声言語処理の技術の発展に伴い，機械翻訳を介した言語間音声コミュニケーションは旅行者にとって現実的なものとなりつつある．また，観光者が訪れるスポット，ランドマーク，飲食店などの固有名詞が多く存在し，そのカバレージが翻訳システムの精度に影響を及ぼすことが分かっている．この固有名詞の課題の対処として，クラスと言われる単語のグループを用いた手法や，サブワードモデルなど研究されてきたが，実用では，固有名詞対訳辞書の作成コストと，固有名詞クラスのアノテーションのコストの問題が課題である．

本論文では，まず，クラス言語モデルに必要なアノテーションの課題を解決し，固有名詞クラス定義を自動推定する手法を提案する．ここでは，日英対訳コーパスから自動翻訳指標を用いて，最適な固有名詞クラス定義を決定する．次に，機械翻訳の実験検証として，入力文の固有名詞クラスの推定結果により，固有名詞を正確に訳せるクラス言語モデルに基づく機械翻訳と，サブワードに基づく機械翻訳を使い分けることで，訳質が向上することを JParaCrawl の日英対訳コーパスを用いた実験で確認する．

2　関連研究

固有名詞に着目した機械翻訳のアプローチは大別して 2 つに分けられる．1 つ目は，サブワードモデルのアプローチで，学習コーパスの単語をサブワードもしくは文字単位に細かく分割し，未知語となる語彙を減らす手法 [1–4] である．2 つ目は，未知語を既知語に置き換え，既知語を含む文でニューラル機械翻訳（Neural Machine Translation：NMT）を行い，翻訳結果から置き換えた単語を戻す手法 [5] である．

サブワードを使用する手法は，未知語となる固有名詞が含まれていても文全体の訳は崩すことなく，流暢に翻訳できる．しかしながら，この手法では固有名詞自体の訳語が翻訳結果に出力されない問題や，同表記の固有名詞の訳語など多義性の問題から，翻訳の適切さとしての課題が残る．

次に，既知語への置き換え手法は，未知語を事前学習したクラスの中から選択し，指定したクラスの別の固有名詞の既知語に置き換えて処理する．既知語は翻訳モデルの学習コーパスの中から選択するため，学習量が十分なことから，文全体の訳も崩れず適切に翻訳できる．そして，未知語である固有名詞の翻訳に関しては，人手で対訳辞書を作成するため，正確に翻訳することができる．ここで，クラス言語モデルの利用は，音声認識の分野で，未知語の問題を解決するために用いられてきた [6–9]．この考えは，単語クラス付き対訳辞書を用いる手法などにより，コーパスベースの機械翻訳にも取り入れられている [10–13]．しかしながら，その多くは人手によりアノテーションされた学習データを用いて，教師あり学習をう方法で，学習コーパスを用意する必要がある点で課題が残る．

3　提案手法

本研究では，正確な固有名詞の翻訳が行える既知語への置き換え手法を用いる．まず，人手によるアノテーションを行わなくとも固有表現認識 (Named Entity Recognition:NER) モデルの学習が行える手法を提案する．従来研究と同様に，教師あり学習の手法を用いるが，固有名詞のクラス定義をするための学習データを自動構築する．また比較的簡単な原言語と目的言語の対訳辞書の作成は人手で行い，翻訳モデルに多数存在する固有名詞クラスから自動推定する仕組みを構築する．さらに学習した NER モデルにより，NMT モデルを切り替え，訳質を向上する手法を提案する．

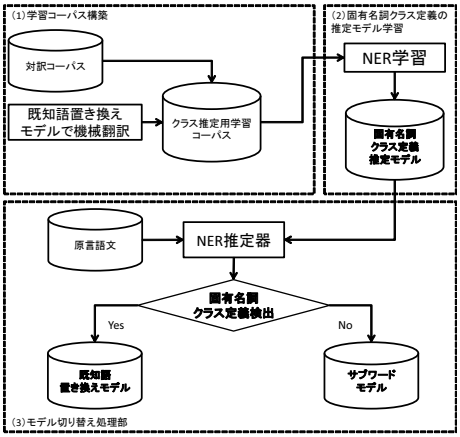

図 1　提案手法の処理概要

図 1 に示す通り，提案手法は，データ構築処理部，固有名詞クラス定義の推定モデル学習，およびモデル切り替え処理部からなる．データ構築処理部は，日英対訳コーパスから，自動翻訳指標を活用して，最適な固有名詞クラス定義を決定する．作成した推定モデルのデータから，固有名詞クラス定義の推定モデルを学習する．そしてモデル切り替え処理部は，実際に翻訳処理する際の仕組みを想定し，SentencePiece モデルと既知語への置き換えモデルの切り替えを NER モデルで分岐する．学習データは，モデル切り替え処理部にある翻訳モデルの学習や適応は行わず，固有名詞クラス定義推定モデルの学習にのみ活用する．これらの 3 つの処理の詳細を説明する．

3.1　データ構築

図 2 に，データ構築の処理手順を示す．日英対訳コーパスをベースに，日本語と英語の両言語に固有名詞が含まれる文を抽出する．次に，複数存在する固有名詞クラスの中から最も翻訳性能が高くなるクラスを，翻訳の自動評価により決定して，学習データを作成する．固有名詞クラス定義の自動付与の手順を以下に記す．

1. 固有名詞クラス定義のうち一つを選択し，既知語置き換えモデルに固有名詞を登録する．
2. 選定した 1000 文の原言語文から，上記の 1 の辞書で，既知語置き換えモデルで翻訳する．
3. 1 で登録した固有名詞を辞書から取り除く．
4. 上記の 1 から 3 をすべての固有名詞クラスに対して行う．

そして，上記の手順で得られた固有名詞クラスごとの翻訳結果に対して自動評価を行う．目的言語文を

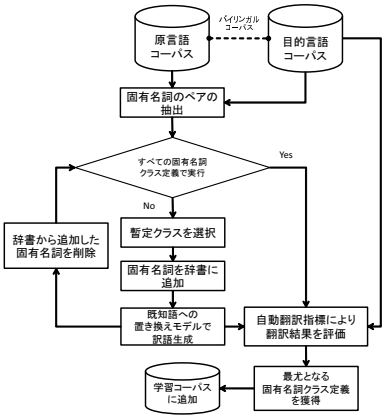

図 2　データ構築処理

参照訳として，次式により最適な固有名詞クラス ($\hat{c}$)[1] を得る．

$$\hat{c} = S_{BLEU}(T_{REF}, T^{c}_{MT}) \tag{1}$$

C，T_{REF}，T^{c}_{MT}，はそれぞれ，固有名詞クラスの集合，対訳コーパス中の目的言語文，固有名詞クラス定義 C としたときの翻訳結果である．S_{BLEU} は，T_{REF}，T^{c}_{MT} の自動評価指標 BLEU スコア [14] で次式により計算する．

$$BLEU = BP_{BLEU} \times exp(\frac{1}{N}\sum_{n=1}^{N} log(Pn)) \tag{2}$$

ここで，$N = 4$，BP_{BLEU} は翻訳文が参照文と比較して短い場合に用いるペナルティ係数である．BP_{BLEU} は翻訳文の単語数を c，正解文の単語数を r とし，以下の式で計算される．

$$BP_{BLEU} = \begin{cases} 1 & (c > r) \\ e^{1-\frac{r}{c}} & (c \leq r) \end{cases} \tag{3}$$

このようにして対訳コーパスの原言語文に対して，$\hat{c}$ を付与することができる．

3.2　固有名詞のクラス推定モデル

前節の学習データから，固有名詞クラスを付与する NER モデルを学習する．NER モデルは指定した単語に対して，単語自身の表層標記情報や周囲の単語列の情報を用いて，固有名詞クラスを推定する手法が研究されてきた．そして，固有名詞のクラス定義は翻訳エンジンごとに定義されており，詳細なものでは形態素情報が用いられている．本研究の NER モデルでは，入力文に含まれる各単語をそれぞれ $N+1$ カテゴリ（翻訳システムによって定義され

1)　固有名詞クラスの選択により訳語が変化する原言語文に限る

るクラス数 N と，それ以外のクラス）に分類する形で学習する．NER モデルに関しては，従来研究の手法をいくつか用いて検証する．

3.3 モデル切り替え処理

実際の翻訳実行には，固有名詞を正確に翻訳するために，既知語への置き換えモデルと SentencePiece モデルを切り替えて，ハイブリッド処理を行う．図 1 の (3) にある，原言語文を NER 推定し，推定結果により処理を分岐する．推定結果が固有名詞クラスの場合は，既知語への置き換えモデルを用いる．固有名詞と推定した固有名詞クラスを辞書に登録し，同クラスの代表単語と置き換えて文全体を翻訳する．そして，翻訳結果の代表単語を登録した固有名詞に戻して訳語を得る．一方，推定結果が固有名詞クラス以外の場合は，SentencePiece モデルで文全体を翻訳し，訳語を得る．

4 実験

本実験では日英機械翻訳を対象に，提案手法の有効性を確認する．また手法の有効性の確認のため，ハイブリッド手法としない個々のモデルで比較評価する．全評価データを SentencePiece モデルに入力した場合をベースラインとする．そして既知語への置き換え手法の比較として，評価データの固有名詞の存在を既知として，固有名詞クラスを人手で付与したもので検証した．

4.1 実験条件

既知語への置き換えモデルは，国立研究開発法人情報通信研究機構のみんなの自動翻訳@ TexTra[2] を使用した．Transformer モデルのニューラル機械翻訳で固有名詞のクラスを選択し，辞書登録機能がある．そして SentencePiece モデルは，JParaCrawl の日英対訳コーパスの学習済みモデル[3] を使用した．

NER モデルの学習手法は，3 つの手法を検証した．1 つ目は，BiLSTM-CRF [15, 16] で，推定対処の単語を，前方向と後方向からそれぞれ単語列を順列処理する手法である．2 つ目は，GRN[4] [17] で，分散表現化 [18] した単語を入力することで分類する．分散表現の取り組みは，テキスト分類 [19, 20] にも応用された手法である．3 つ目は，BERT-NER [21] で，マルチレイヤー Tramsformer [22] である．近年の研究で，BERT モデルをベースした研究が活発で，NER タスクの研究もなされている [23]．本研究では，事前学習されたモデルをベースに，構築したコーパスを用いて FineTuning して適応する．そして BiLSTM-CRF と GRN では，前述の 1000 文の学習コーパスを用いて学習を行った．BERT-NER は，事前学習された 12 層, 隠れ層次元数 768 のモデル[5] を用い，前述の 1000 文の学習コーパスで FineTuning して学習させた．表 2 に，本研究の実験で使用した NER モデルのハイパーパラメータを示す．学習コーパスはオープンデータの JParaCrawl の日英対訳コーパスを使用した．1000 万文の JParaCrawl の日英対訳コーパスから，日本語と英語の両言語に固有名詞が含まれる文を抽出し，さらにその中から，自動算出した訳語の精度 [24] の高い 1000 文を選定して，学習コーパスとした．固有名詞クラス定義の種類は，評価データの旅行ドメインで利用する 14 クラスを用い，最終的に得られた固有名詞クラス定義と学習データ数を表 2 に示す．データセットの中で出現割合が最も多いクラスは日本人の姓で，最も少ないクラスはキャラクター名であった．

表 1 ハイパーパラメータ

パラメータ	BiLSTM-CRF	GRN	BERT-NER
mini batch size	32	10	32
epoch	15	200	4
optimizer	Adam	SGD	Adam

表 2 固有名詞クラス定義と学習データ数

定義名	データ数	定義名	データ数
アトラクション	21	宿泊施設	26
施設	21	国名	26
ランドマーク	34	食べ物	34
組織	26	ショップ	27
姓 (和名)	36	名 (和名)	44
姓 (洋名)	24	名 (洋名)	35
キャラクター	20	お土産	22

4.2 評価方法

NER モデルの性能を 3 つの指標で評価する．1 つ目が既知語置換率，2 つ目が固有名詞正解率，そして 3 つ目が BLEU による自動評価である．NER モ

2) https://mt-auto-minhon-mlt.ucri.jgn-x.jp/
3) http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/release/1.0/pretrained_models/ja-en/base.tar.gz
4) Gated Relation Network to Enhance Convolutional Neural Network for Named Entity Recognition
5) https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

表 3　固有名詞クラス推定と BLEU 評価

手法	既知語置換率	固有名詞正解率	BLEU
人手付与	100.00	100.00	39.26
SentencePiece（ベースライン）	0.00	89.86	35.73
BiLSTM-CRF	24.29	89.86	37.40
GRN	45.71	91.30	35.22
BERT-NER	**94.29**	**98.55**	**37.55**

デルは対象単語のクラス推定を，15 クラス（固有名詞クラス 14 と，固有名詞でないクラス）から推定し，固有名詞クラスと推定した場合は既知語への置き換えモデルを用いて翻訳する．既知語置換率は，固有名詞クラスを推定できた率である．次に，辞書登録した場合の固有名詞の翻訳の正確さの指標として，固有名詞正解率を算出する．最後に，固有名詞の翻訳を含めた文全体の翻訳の評価として，自動評価指標の BLEU 評価を行う．NER 推定結果が自動翻訳全体に及ぼす影響を．人手で作成した目的言語文の訳語を用いて評価比較する．

評価の対訳データは，独自に収集した，岐阜タクシーでの翻訳社会実証データの 2379 文の中から，両言語に固有名詞が含まれる 261 文を選定し，そこから評価データとして 69 文をランダムに抽出した．固有名詞辞書のデータ作成は，翻訳者が人手で作成し，複数のクラス定義を付与し，最終的に 285 件をエントリーした．

4.3　実験結果

表 3 は，固有名詞に人手でアノテーション付与する手法，ベースラインの SentencePiece モデル，および提案手法の NER モデルによるハイブリッド手法の比較評価結果である．2 列目に既知語置換率，3 列目に固有名詞正解率，4 列目に BLEU スコアを記述した．既知語への置換手法は，全ての条件において，基盤となる NMT と対訳辞書のエントリーは同じで，クラスの付与方法のみが異なる．表 3 から，人手付与の手法は，アノテーションのコストは必要になるが，固有名詞正解率が 100%と全て正確に翻訳ができて，BLEU スコアも最も高い数値であった．次に，固有名詞のアノテーション情報のない SentencePiece モデルは，固有名詞正解率が 89.86%で，BLEU スコアが 35.73%と高精度であった．この場合は自動で固有名詞判定は行わないが，翻訳結果から事後分析し，固有名詞正解率を算出した．固有名詞正解率が高い理由は，評価データには「日本」や「岐阜」など比較的収集しやすい用語があり，評価データの一部が，学習データの JParaCrawl に含まれていたと考えられる．付録の表 4 に評価結果の一部を示す．一方で，人手付与と比較すると，固有名詞正解率と BLEU スコアが低く，評価データの固有名詞が 10%近くが訳せていない．提案手法では，BiLSTM-CRF と GRN は，既知語置換率は低かったが，固有名詞正解率がベースラインより軽微改善できた．最も評価結果が良かったものは BERT-NER で，既知語置換率が 94.29%と非常に高い検出率となった．そして固有名詞正解率が 98.55%と高精度で，SentencePiece モデルと比べて大きく改善し，アノテーションコストが必要な人手付与に近い精度が得られた．さらに BLEU スコアについても，人手付与には劣るが，SentencePiece モデルよりも高い訳質が得られた．自動構築した NER モデルによるハイブリッド手法は，アノテーションコストなしに，人手付与の手法に近い性能と言える．

5　まとめ

本論文では，クラスベースの翻訳システムへの利用を目的とし，固有名詞クラスを自動的に付与する方法を提案した．本手法は，対訳コーパスと翻訳システムとを用いて，学習データを構築し，次に構築したデータから NER モデルによる固有名詞クラス推定モデルを学習した．評価実験では，このモデルから自動的にテストセットの固有名詞を推定し，推定した固有名詞クラスをニューラル機械翻訳の辞書に登録し，得られた翻訳結果を自動評価した．実験結果によると，提案手法による固有名詞クラス推定の既知語置換率が 94.29%で，自動評価評価も人手のアノテーションに迫るスコアで，高精度な結果が得られた．これらの結果から，提案手法は，人手による学習データの作成なしに，クラスベース翻訳システムの固有名詞辞書拡張を可能にしたと言える．今後の検討課題として，コーパスサイズを増やして固有名詞クラス推定の精度を改善する予定である．

6　謝辞

本研究は，総務省「グローバルコミュニケーション計画の推進 -多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環で収集したデータにより実施したものです．

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 1715–1725, 2016.

[2] Z. Chen Q.V. Le M. Norouzi W.Macherey M. Krikun Y. Cao Q. Gao K. Macherey J.Klingner A. Shah M. Johnson X. Liu L. Kaiser S.Gouws Y. Kato T. Kudo H. Kazawa K. Stevens G.Kurian N. Patil W. Wang C. Young J. Smith J. Riesa A. Rudnick O. Vinyals G. Corrado M. Hughes Y. Wu, M. Schuster and J.Dean. Google' s neural machine translation system:bridging the gap between human and machine-translation, 2016.[Online; accessed 28-December-2017]. http://www.pluto.ai.kyutech.ac.jp/NLP/.

[3] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybridword-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 1054–1063, 2016.

[4] M.R. Costa-Juss'a and J.A.R. Fonollosa. Characterbased neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 357–361, 2016.

[5] Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. Towards zero unknown word in neural machine translation. In *Proceedings of the 25th Annual Meeting of the International Joint Conferences on Artificial Intelligence (IJCAI)*, p. 2852–2858, 2016.

[6] Issam Bazzi and James R. Glass. Modeling out-of-vocabulary words for robust speech recognition. In *ICSLP*, 2000.

[7] Hiroaki K. Genichiro K. Yoshihiko O. Yoshinori S. Yamamoto, H. Out-of-vocabulary word recognition with a hierarchical language model using multiple markov model. In *IEICE D-II 87(2)*, pp. 2104–2111, 2004.

[8] Shuntaro I. Yoshinori S. Yamamoto, H. Multi-class composite n-gram based on connection. In *ICASSP*, p. 533–536, 1999.

[9] Welly NAPTALI, Masatoshi TSUCHIYA, and Seiichi NAKAGAWA. Class based n-gram language model for new words using out-of-vocabulary similarity. In *IEICE*, p. 2308–2317, 2012.

[10] Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. Translation estimation for technical terms using corpus collected from the web. In *newblock In: Proceedings of the Pacific Association for Computational Linguistics*, pp. 325–331, 2005.

[11] Keiji Yasuda, Panikos Heracleous, Akio Ishikawa, Masayuki Hashimoto, Kazunori Matsumoto, and Fumiaki Sugaya. Building a location dependent dictionary for speech translation systems. In *CICLING*, 2017.

[12] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 1064–1074, 2016.

[13] Sebastien Jean, Kyunghyun Cho, Yoshua Bengio, and Roland Memisevic. Onusing very large target vocabulary for neuralmachine translation. In *28th NIPS*, p. 1–10, 2014.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[15] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *NAACL-HLT*, p. 260–270, 2016.

[16] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, p. 2227–2237, 2018.

[17] G. Dingy J. Louz Y. Zhangx H. Cheny, Z. Linz and B. Karlssonz. Grn: Gated relation network to enhance convolutional neural network for named entity recognition. 2019.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26. Curran Associates*, 2013.

[19] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, p. 1746–1751, 2014.

[20] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. Convolutional neural networks for modeling sentences. In *COLING*, p. 655–665, 2014.

[21] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. Tuning multilingual transformers for named entity recognition on slavic languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 89–93, 2019.

[22] Noam S. Niki P. Jakob U. Llion J. Aidan N G. Łukasz K. Ashish, V. and P. Illia. bidirectional transformers for language understanding. In *In Advances in neural information processing systems*, p. 5998–6008, 2017.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[24] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. Jparacrawl: A large scale web-based english-japanese parallel corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation(LREC 2020)*, p. 3603–3609, 2020.

A 付録

表 4 例：評価文と翻訳結果

原言語文：**岐阜**でおもしろい場所はどこですか？	
人手付与	Where is the interesting place in **Gifu**?
SentencePiece	Where are the interesting places in **Gifu**?
BiLSTM-CRF	Where is interesting place in **Gifu**?
GRN	Where is interesting place in **Gifu**?
BERT-NER	Where is interesting place in **Gifu**?
原言語文：あと二三分で**一夜城**に着きます	
人手付与	We'll get to **Ichiya Castle** in a few minute
SentencePiece	It takes a couple of minutes to arrive at the castle overnight
BiLSTM-CRF	It takes a couple of minutes to arrive at the castle overnight
GRN	We'll arrive in **Ichiya Castle** in a couple of minutes
BERT-NER	We'll get to **Ichiya Castle** in a couple of minutes
原言語文：川のこちらが**犬山市**です。	
人手付与	This is the **Inuyama City**
SentencePiece	This river is **Inuyama City**
BiLSTM-CRF	This river is **Inuyama City**
GRN	This is the **Inuyama City**
BERT-NER	This is **Inuyama City**
原言語文：何回か行きました。**朴葉味噌**か肉の料理がおいしかった。	
人手付与	I went several times. The **Hobo miso** or meat dish was delicious
SentencePiece	I went several times to eat Pakha miso or meat
BiLSTM-CRF	I went several times to eat Pakha miso or meat
GRN	I went several times to eat Pakha miso or meat
BERT-NER	I went several times. The food in **Hobo miso** or meat was delicious