大規模名詞句フレーズのSMT訳への後処理型置き換えによるNMTの評価*

飯田 頌平† 龍 梓‡ 木村 龍一郎‡ 宇津呂 武仁‡ 三橋 朋晴§ 山本 幹雄‡
†東京電機大学 工学部 ‡筑波大学大学院 システム情報工学研究科 §日本特許情報機構

1 はじめに

ニューラルネット機械翻訳 (NMT) とは，Sutskever らによる Sequence-to-Sequence モデル [8] をベースとした手法であり，その長所は流暢な翻訳ができることである．しかし NMT では使える語彙の数に制約があり，専門用語の翻訳に弱いという欠点がある．この問題は新規の専門用語を多く含む特許文の翻訳において，特に深刻な問題となる．これまでの NMT システムにおける語彙数制限の研究では，未知語の出現箇所に位置情報を付加してアライメントを追跡し，その後単純な辞書検索により未知語を翻訳結果に置き換える手法 [6] などが知られている．しかし，このアプローチでは 専門用語の一部であっても単独の未知語として解釈していたため，特許文の翻訳には制限があった．

この点を改善したものとして，Long らによる大規模専門用語を含む特許文に対応した NMT システムに関する研究 [5] がある．この手法では，対訳特許文中の専門用語をトークンで置き換えて NMT モデルの訓練を行い，翻訳文中のトークンを統計的機械翻訳 (SMT) による翻訳結果で置き換える．しかし，この手法では元々の特許文に含まれていた数万語彙にもおよぶ名詞句フレーズを僅か数個のトークンに置き換えることで，本来は異なる名詞句フレーズを含んでいた別々の文章が，NMT モデル上同種のフレーズを含む文章として扱われてしまうことになる．そこで，本論文では，トークンを使用せずに NMT 訳中の専門用語を SMT の翻訳結果で直接置き換える NMT システムを提案する．

また，近年，SMT を使うアプローチとは別に，単語をサブワード単位または文字単位に分割することで未知語なしでコーパスを網羅した NMT モデルを訓練する研究が行われている．Sennrich らの手法 [7] では，語彙数を目的関数とした貪欲法を用いており，Wu らの手法 [11] および SentencePiece ツール [1] では，目的関数としてエントロピーを最尤推定して最適化している．本稿では，Wu らの手法 [11] がより優れた成果を挙げた [1] ことを受け，SentencePiece による語彙集合を利用した評価結果を示す．

2 NMT における名詞句フレーズの SMT 訳への置き換え方式

2.1 NMT モデルにおける語彙集合

本論文では，単語単位・形態素単位のものと，SentencePiece によるものの二種類の語彙集合を用いる．前者としては，英語文は単語単位を基本とし，さらに英文中の記号を空白で分割するために，Moses [3] の Tokenizer ツールを用いた．日本語は形態素解析ツール MeCab[2] (IPAdic) を使用して形態素単位の語彙集合を定義した．一方，後者としては，日本語文・英語文それぞれ SentencePiece 単位で処理した語彙集合を使用する．

2.2 言語知識に基づき選定した名詞句フレーズ

本論文では，日本語の入力文において，言語知識に基づく形態素解析ツールを用いて各形態素・単語に付与した品詞情報を利用し，名詞句を抽出するための品詞パターンを以下のように定義する．

$$(\text{名詞} \mid \text{接頭辞} \mid \text{動詞} \mid \text{形容詞}) + \text{名詞}$$

この品詞パターンを満たす形態素・単語列を名詞句フレーズとみなして抽出する．なお，形態素解析ツールには MeCab (IPAdic) を使用した．SentencePiece は品詞情報を持たないため，名詞句フレーズの抽出には利用しない．

*Evaluating NMT by Post-Replacement with Large-scale Noun Phrases Translated by SMT

†Shohei Iida, School of Engineering, Tokyo Denki University
‡Zi Long, Ryuichiro Kimura, Takehito Utsuro, Mikio Yamamoto, Graduate School of Systems and Information Engineering, University of Tsukuba
§Tomoharu Mistuhashi, Japan Patent Information Organization

[1] https://github.com/google/sentencepiece
[2] http://taku910.github.io/mecab/

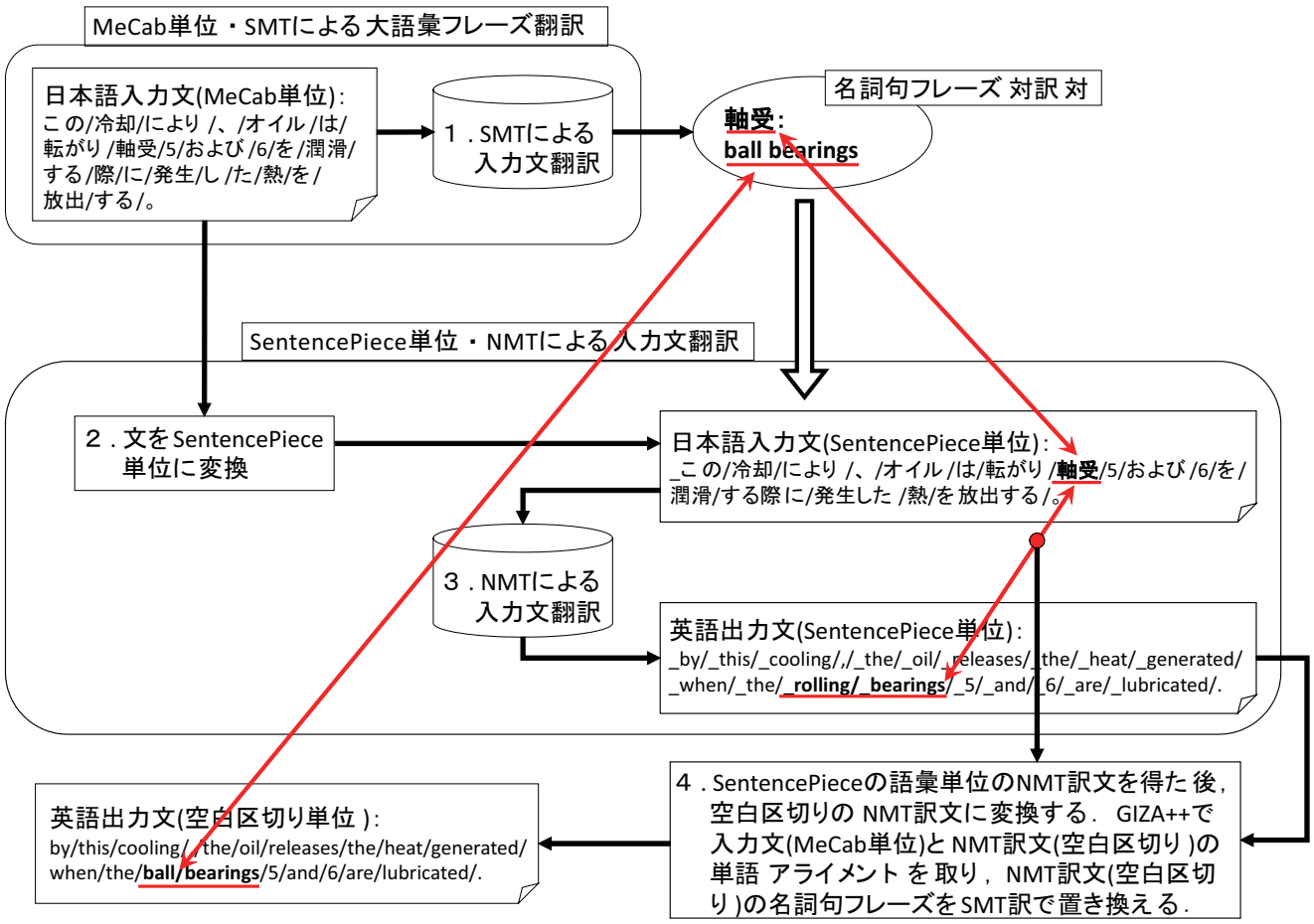

図 1: SentencePieceNMT による翻訳文中の名詞句フレーズを MecabSMT で置き換えるシステム

2.3 名詞句フレーズの SMT 訳への置き換え

図 1 に沿って，SentencePiece 単位の語彙に対応した NMT モデルを用いて，NMT 翻訳文中の名詞句フレーズを SMT 訳に置き換える流れを以下に示す．

ステップ 1 はじめに，MeCab 単位の語彙の日本語入力文を使用して，SMT による翻訳を行う．翻訳の過程で GIZA++により入力文と SMT 翻訳文の間の単語アライメントが得られるので，あらかじめ抽出した日本語名詞句フレーズとその SMT 訳からなる名詞句フレーズ対訳対を得る．

ステップ 2 形態素解析前の日本語入力文を SentencePiece の語彙単位に分割する．

ステップ 3 SentencePiece の語彙単位に分割された訓練コーパスを用いて訓練した NMT モデルにより入力文を翻訳する．

ステップ 4 SentencePiece の語彙単位の NMT 訳文を得た後，空白区切りの NMT 訳文に変換する．GIZA++によって入力日本語文 (MeCab 単位) と NMT 訳文 (空白区切り) の間で単語アライメントを取り，日本語名詞句フレーズとその NMT 訳からなる名詞句フレーズ対訳対を得る．そして NMT 訳文中の名詞句フレーズを，SMT による名詞句フレーズで置き換える．

単語がアライメントに含まれていない場合や，目的言語側で離れた位置に翻訳された場合など，単語アライメントによって翻訳結果が同定できない場合には，SMT による翻訳置き換えに失敗したものとして，NMT 訳をそのまま出力する．

3 評価

3.1 データセット

訓練・評価用データセットとして 110 万行からなる対訳特許文 (NTCIR-7 特許翻訳タスク [1] における日英対訳特許文．詳細は [5] 参照) のコーパスを使用した．そのうち検証用データとして 1,000 行を，SMT のチューニング用に 1,000 行をそれぞれ無作為に抽出して，残りのデータを NMT および SMT モデルの訓練用データとして用いた．NMT モデル訓練時の語彙

表 1: 自動評価結果

モデル	BLEU	名詞句再現率
MeCab NMT 名詞句置換なし	39.56	52.3% (1329/2539)
MeCab NMT 提案手法で名詞句置換	40.72	**57.6%** **(1463/2539)**
SentencePiece NMT 名詞句置換なし	39.84	51.3% (1302/2539)
SentencePiece NMT 提案手法で名詞句置換	**40.88**	55.8% (1417/2539)
MeCab NMT Long の手法 [5] で名詞句置換	40.36	**59.3%** **(1505/2539)**

集合としては，MeCab では日英それぞれ 4 万語とし，SentencePiece では日英ともに 1 万 8 千語の語彙とした．なお， 2.2 節の手法を適用した結果，検証用データから 2539 個の日本語名詞句フレーズが抽出された．この名詞句フレーズを対象として 2.3 節の置き換えを行う．

3.2 評価手順

SMT モデルの訓練のために，フレーズベース SMT モデル用のツールキットである Moses [3] を使用した．Moses [3] には英文の記号を分割するための Tokenizer や，単語アライメントを取得するツールの GIZA++も含まれている．NMT モデルの訓練では，エンコーダ・デコーダそれぞれに 3 層の LSTM 層を使用し，各 LSTM 層および埋め込み層の変数を 512 個とした．より詳細な設定は，Long らの設定 [5] に倣った．また本論文の検証では，日本語文を入力し英語文に翻訳する日英タスクを行った．

3.3 評価結果

BLEU スコアおよび 2.2 節で抽出した名詞句の再現率の評価結果を表 1 に示す．再現率は以下の手順で測定した．(1) GIZA++により日本語入力文と英語参照文の間で単語アライメントを取得し，名詞句フレーズ対訳対を得る，(2) 翻訳文に対して，名詞句フレーズの参照訳を照合して訳出の有無を検出する．

ベースラインとして，MeCab 単位の訓練文を 3.2 節の設定で訓練したモデルを用意した．そしてそのモデルに対して提案手法で名詞句の置き換えを行った結果，BLEU スコアが 1.16 ポイント，名詞句再現率が 5.3%向上した．一方，SentencePiece 単位の訓練文を 3.2 節の設定で訓練したモデルも作成したところ，MeCab 単位のモデルに比べると，BLEU スコアは 0.28 ポイント向上したが，名詞句再現率は 1.0%下がった．SentencePiece 単位のモデルに対し提案手法で名詞句の置き換えを行うと，BLEU スコアが 1.04 ポイント，名詞句再現率が 4.5%向上した．

また，Long の手法 [5] との定量的な比較も行った．MeCab 単位のモデルに対して名詞句の置き換えを行ったところ，BLEU スコアが 0.80 ポイント，名詞句再現率が 7.0%向上した．提案手法は Long のモデル [5] よりもより高い BLEU スコアとなるものの，名詞句の再現率では Long のモデル [5] の方がより優れた結果となった．

無作為に選定した少量の検証文に対して人手評価を行ったところ，MeCab 単位の NMT モデルよりも SentencePiece 単位の NMT モデルの方がより正確な意味合いの翻訳結果が得られた．具体的には，品詞や節の並びがより自然な傾向があった．しかし入力文と翻訳結果の間の単語アライメントを取得しようとすると，SentencePiece 単位で行うよりも MeCab 単位を用いた方が適切な名詞句フレーズ翻訳対が得られた．また，提案手法による置き換えを行うと，BLEU や名詞句再現率は改善するものの，品詞や節の並びが不自然になる例も散見された．

また，図 2 において，提案手法と Long の手法 [5] それぞれを用いフレーズの翻訳誤りを改善した例を示す．比較対象は，SentencePieceNMT(名詞句置換なし)，SentencePieceNMT の名詞句フレーズを提案手法により SMT 訳で置き換えたもの，および，MeCabNMT の名詞句フレーズを Long 手法 [5] により SMT 訳で置き換えたものの三種類である．SentencePieceNMT(名詞句置換なし) では，日本語名詞句「軸受」は “rolling bearings” と誤訳されるが，SMT 訳に置き換えることにより，両手法ともに参照訳 “ball bearings” を得ることができた．Long の手法 [5] では，BLEU の観点で見れば提案手法より精度が落ちており，トークンを用いることによって，文中のトークン箇所以外の翻訳に悪影響が出ていることが示された．一方で，Long の手法では，入力文と NMT 訳との間で単語アライメントを取り直すプロセスが無いため，名詞句の置き換えは本手法よりも正確に行えると考えられる．

4 関連研究

NMT と SMT のハイブリッドモデルに関連して，He らによる手法 [2] や Wang らの手法 [9, 10] が提案されている．He らの手法 [2] では，NMT モデル・SMT モデル・n-gram 言語モデル等のスコアを計算し，対数尤度モデルにより統合する．一方，Wang らの手法 [9, 10] では，NMT の確率値の枠組みの中に SMT によるスコアを統合している．これらの論文の評価実験での中英翻訳タスクの場合，中国語の語彙集合は中国語形態素

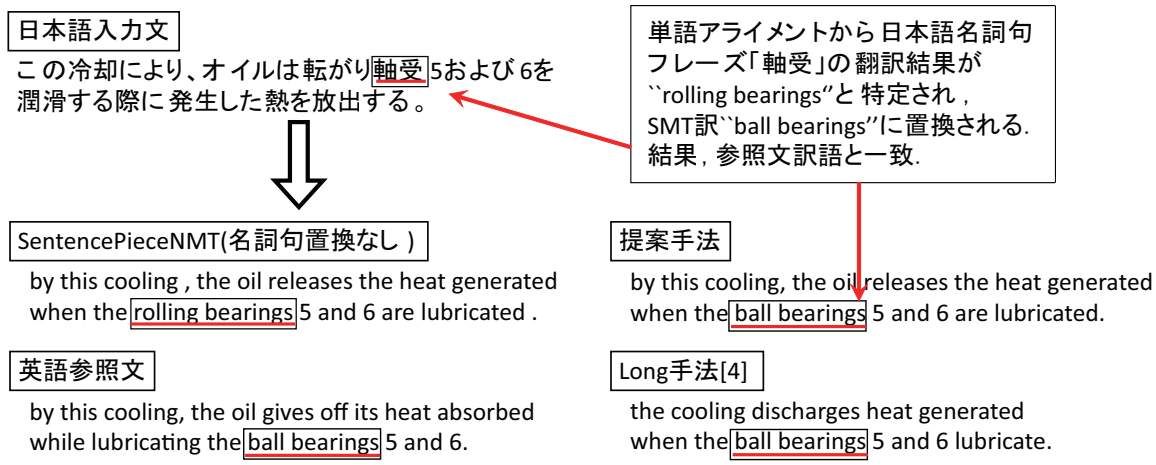

図 2: 提案手法による改善例

の単位となり，NMT 訓練時には，頻度上位 3 万語以外の語彙が未知語となる．そのため，NMT モデルにおける未知語に対しては，本来の NMT モデル以外の部分で対処している．これに対して，本論文で採用した SentencePiece を適用する場合には，NMT モデルにおいて未知語となる語彙がないため，本来の NMT モデルの枠組みの中だけで翻訳過程が完結する．

その他，[5] の発展手法として，[4] においては，エントロピーを用いてフレーズを選定する手法を提案している．この手法は，MeCab のような外部辞書知識を使用せずに名詞句フレーズを抽出できる点が長所である．この手法を取り入れつつ，対訳コーパスの語彙集合を SentencePiece によって決定することによって，本論文の手法においても，完全に外部辞書知識を不要にできる可能性がある．

5 おわりに

本論文では，NMT モデルと SMT モデルを組み合わせ，さらに MeCab 単位と SentencePiece 単位の語彙集合を使い分けることにより，大語彙の専門用語に対応して特許文を翻訳するシステムを提案した．そして日英翻訳において，提案手法を用いない NMT システムよりも優れた自動評価結果を得られることを示した．今後の課題として，入力文と NMT 翻訳文の間の単語アライメントをとる際に，前後の単語を含む誤りのアライメントとなる場合がよくあるので，英語側の文法・前置詞情報等の情報を併用することによって，単語アライメントの性能を改善することが挙げられる．

参考文献

[1] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, pp. 97–106, 2008.

[2] W. He, Z. He, H. Wu, and H. Wang. Improved neural machine translation with SMT features. In *Proc.30th AAAI*, pp. 151–157, 2016.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[4] Z. Long, R. Kimura, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. Neural machine translation model with a large vocabulary selected by branching entropy. In *Proc. MT Summit XVI*, pp. 227–240, 2017.

[5] Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. In *Proc. 3rd WAT*, pp. 47–57, 2016.

[6] M. Luong, I. Sutskever, O. Vinyals, Q. V. Le, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proc. 53rd ACL*, pp. 11–19, 2015.

[7] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pp. 1715–1725, 2016.

[8] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural machine translation. In *Proc. 27th NIPS*, pp. 3104–3112, 2014.

[9] X. Wang, Z. Lu, Z. Tu, H. Li, D. Xiong, and M. Zhang. Neural machine translation advisedby statistical machine translation. In *Proc.31st AAAI*, pp. 3330–3336, 2017.

[10] X. Wang, Z. Tu, D. Xiong, and M. Zhang. Translating phrases in neural machine translation. In *Proc. EMNLP*, pp. 1421–1432, 2017.

[11] Y. Wu, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. `http://arxiv.org/abs/1609.08144`, 2016. [Online; accessed 28-December-2017].