The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

漢字分解したテキストによるニューラル機械翻訳

Neural Machine Translation with Kanji Decomposition

ビィシュウ グプタ Vishu Gupta　中村 亮裕 Akihiro Nakamura　福田 治輝 Haruki Fukuda　綱川 隆司 Takashi Tsunakawa　狩野 芳伸 Yoshinobu Kano
西田 昌史 Masafumi Nishida　西村 雅史 Masafumi Nishimura

静岡大学
Shizuoka University

This paper proposes a method for neural machine translation (NMT) with kanji decomposition of Japanese text. NMT models have restrictions of the vocabulary size, which can be solved by applying subword, character-level, or byte-based models. In Japanese text, the vocabulary size would not be minimized even in a character level because of kanji varieties. We report an experimental result of NMT model using Japanese text with kanji decomposition that is expected to satisfy both of decreasing vocabulary size and keeping kanji information.

1. はじめに

ニューラル機械翻訳モデルでは語彙は一定のサイズに限定される．この課題に対処する方法の一つとしてサブワード単位や文字単位で処理する方法が提案されている [Sennrich 16]．これらの方法の利点の一つとして未知語の問題がなくなることが挙げられる．文字単位であれば語彙サイズは原言語と目標言語のアルファベットの範囲内となり非常に小さくなる一方で，翻訳性能は維持される，または向上するとの報告もある [Lee 17]．さらに文字コードを用いてバイト単位にまで分割することで，言語に依存せず語彙サイズを 256 にまで小さくするモデルも提案されている [Costa-jussà 17]．

日本語では従来，形態素解析によりトークン化したものを機械翻訳の入出力としていた．これを文字単位とした場合，漢字のバリエーションが多いために語彙サイズは比較的大きくなる．語彙サイズを小さくするために漢字を読みに変換することはできるが，同音異義語の情報が失われてしまう．

本研究ではこれらの問題を抑えつつ，漢字よりも細かい単位への分割を適用するため，漢字をその構成要素に分解してからニューラル言語処理向けトークナイザ SentencePiece[*1] を適用したトークン列を生成し，ニューラル機械翻訳の入出力として扱う方法を提案し，その効果を検証した．

2. 提案手法

図 1 に提案手法の概要を示す．まず，コーパス中の日本語の各漢字を KRADFILE[*2] に基づく変換テーブルを参照して分解する．次に，SentencePiece により構成要素に分解した文字列を再度結合する．この日本語テキストを用いてニューラル機械翻訳モデルを学習させる．翻訳結果として得られたテキストは逆変換によりもとの漢字に戻して出力する．

連絡先: 綱川 隆司，静岡大学情報学部，〒 432-8011 静岡県浜松市中区城北 3 丁目 5-1，053-478-1487，tuna@inf.shizuoka.ac.jp

∗1 https://github.com/google/sentencepiece

∗2 http://www.edrdg.org/krad/kradinf.html

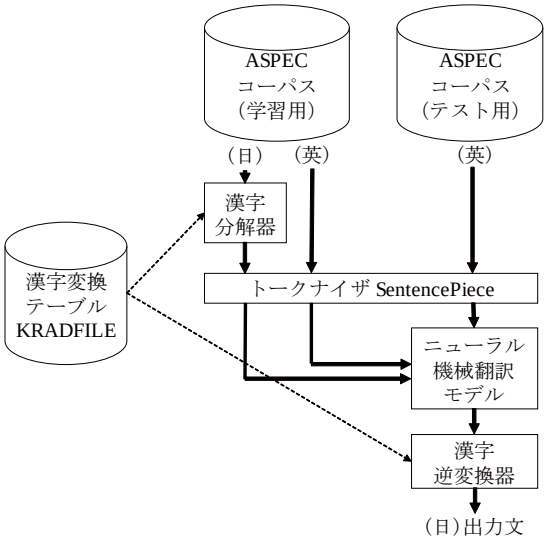

図 1: 提案手法概要

2.1 KRADFILE による漢字分解と逆変換

KRADFILE は，漢字の構成要素による検索を容易にするために提供されている，漢字とその構成要素の変換テーブルである．例えば，

- 哀：衣口𠂉
- 愛：心爪冖夂

のような漢字分解が示されている．構成要素はそれを表す代表の漢字である場合がある．例えば，“こざとへん” は “阡” で表される．

本稿では原則としてこのテーブルに従って漢字を置換する．ニューラル機械翻訳によって得られた文字列は，左方最長一致で合致するテーブルの項目に該当する漢字に逆変換する．しかし，この方法では組合せによる誤結合が発生するため，下記の処理を施している．

- 同じ分解結果となる異なる漢字を区別するため，分解結

果の末尾に判別用の文字を追加する.(例　跟：口艮足止 α，踉：口艮足止 β)

- 構成要素として用いられる漢字自体は，前後に判別用の文字を追加したパターンに変換する.(例　耳：#耳#)
- 左からの最長一致により誤結合が起こる単語に対応するため，一部の漢字の構成要素列の順序を入れ替える.(例　需：雨而　のままだと，需要→雨而女西→嬬西，となり逆変換に失敗するため，需：而雨　とする)

2.2 SentencePiece の適用

提案方法では漢字分解した文字列をそのまま入力するのではなく，SentencePiece によって得られたトークン列として入力する．SentencePiece はサブワード分割アルゴリズムの一種を採用しており，バイトペアエンコーディング (BPE) により文字列から直接分割を学習する．これにより語彙サイズと分割単位のバランスをとることができる．

本研究では英語テキスト，および日本語テキストの漢字分解後の文字列に対して SentencePiece を適用し，得られた分割をトークンとしてニューラル機械翻訳の入力とする．出力結果のトークン列を単純に結合し，漢字への逆変換を施す．

下記は，日本語文 “現在，筋ジストロフィー患者の移動介助において文書マニュアルを使用している。” に対して単に SentencePiece を適用した場合と，漢字への分割を施してから SentencePiece を適用し，漢字への逆変換を施した場合の例を示す．トークンの区切り目はアンダースコア (_) で示している．

SentencePiece 現在_，_筋ジストロフィー_患者の_移動_介助_において_文書_マニュアル_を使用し_ている_。

漢字分割＋ SentencePiece ＋逆変換（区切り削除前） 現_|_丿一土_，_筋ジストロフィー_|_口口心_者_の移_|_一日力里丿_介助_において_文_書_マニュアル_をノ一化_口_用_している_。

SentencePiece による分割では，助詞が結合するケースはあるものの概ね従来の形態素解析器に近い結果となっているのに対し，漢字分割したケースではより細かく分割され，また漢字の構成部分の途中に区切りが来るケースが散見された．特に，“|” や “口” は単独のトークンとして扱われる傾向にあった．評価実験では翻訳結果の出力の際に区切りを削除し結合してから逆変換を施すため，逆変換に失敗するケースはほぼみられない．

3. 実験

科学技術論文抄録の対訳からなる ASPEC コーパス [Nakazawa 16] の日英対訳文対を用いて英日方向の翻訳実験を行った．学習用データ 100 万文対を訓練に用いてニューラル機械翻訳モデルを学習し，テスト用データ 1812 文対を用いて評価した．ニューラル機械翻訳モデルは OpenNMT[Klein 17] を用いて構築し，語彙サイズを 40000，トークン列長の最大数を 150 とした．漢字分解の効果を検証するため，日本語テキストに対して直接 SentencePiece による分割を施した場合，および SentencePiece を適用せずに漢字分解後の文字単位で分割した場合との比較を行った．評価時には分割方法による差異を解消するため，参照文および出力文に対して MeCab[*3] による形態素解析を施してから BLEU スコアを算出した．

*3　http://taku910.github.io/mecab/

表 1: 実験結果

適用手法	BLEU	BP
SentencePiece	0.2480	0.914
漢字分割+SentencePiece	0.2408	0.911
漢字分割	0.1260	0.670

表 1 に各実験設定における BLEU スコア，およびスコア算出の際に用いる BP (Brevity Penalty) を示した．BP は出力文のトークン数が参照文のトークン数より短い場合にかかる係数であり，小さいほど出力文長が短いことを表す．漢字分解を適用した場合のスコアは適用しない場合をやや下回っており，漢字分解の効果は確認できなかった．また，漢字分解を適用しても出力文長にはほぼ変化がみられなかった．SentencePiece 適用後の語彙サイズは漢字分解適用後もあまり変化していないと考えられる．一方，SentencePiece を適用しない場合は低い精度にとどまった．これは，漢字分解後の文字単位の入力トークン列が非常に長くなるために十分な学習ができず，出力文も短くなったことによるものと考えられる．

4. おわりに

本稿ではニューラル機械翻訳による英日翻訳において，語彙サイズの制約を軽減するために日本語の漢字分解を適用した文字列を入出力に用いる方法を提案した．実験結果からは SentencePiece によるトークン化を適用した場合，漢字分解によりトークンが比較的細かく分割されるという結果が得られるものの，漢字分解による精度向上はみられなかった．今後の課題として，漢字の部首と読みを用いるなどの漢字分解方法の検討や，文字ベースのニューラル機械翻訳 [Lee 17] の適用が挙げられる．

参考文献

[Costa-jussà 17] Costa-jussà, M. R., Escolano, C., and Fonollosa, J. A. R.: Byte-based Neural Machine Translation, in *Proc. of the 1st Workshop on Subword and Character Level Models in NLP*, pp. 154–158 (2017)

[Klein 17] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation, *ArXiv:1701.02810* (2017)

[Lee 17] Lee, J., Cho, K., and Hofmann, T.: Fully Character-Level Neural Machine Translation without Explicit Segmentation, *Trans. of the Association for Computational Linguistics*, Vol. 5, pp. 365–378 (2017)

[Nakazawa 16] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, in *Proc. of the 9th International Conference on Language Resources and Evaluation*, pp. 2204–2208 (2016)

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725 (2016)