韓国語対訳データを利用した文字分割と音素分解による朝鮮語ニューラル機械翻訳

金 輝燦[1*]　　平澤 寅庄[2]　　小町 守[2]

[1]朝鮮大学校　[2]首都大学東京

kimhwichan1997@gmail.com, hirasawa-tosho@ed.tmu.ac.jp, komachi@tmu.ac.jp

1　はじめに

現在、多くの言語間でニューラルネットワークを用いた機械翻訳の研究が進められているが、朝鮮語[1]に関する研究は進められていない。その理由の一つとして、訓練に用いる対訳データがないことが挙げられる。ニューラル機械翻訳モデルの翻訳精度を向上させるためには、大規模な対訳データが必要であることが知られている。例えば、訓練用対訳データ内に 1 億以上の単語数がない場合、ニューラル機械翻訳はフレーズベース統計機械翻訳よりも精度が劣ることが報告されている [2] 。ニューラル機械翻訳モデルで高い翻訳精度を得るためには大規模でかつ質の高い対訳データが必要となり、朝鮮語から英語への翻訳は対訳データが乏しいため高い翻訳精度を実現することは困難である。

原言語・目的言語間の翻訳モデルを訓練する際に対訳データが不十分、または、対訳データが存在しない場合、原言語と目的言語の両言語において、十分な対訳データがある言語（ピボット言語）を仲介にし翻訳モデルを作成する手法がある [7]。また、朝鮮語は韓国語と非常によく似た言語であるため、Marujo ら [3] が提案しているように、ルールベースにより類似した言語間の変換を行う手法を用いて、韓国語の対訳データを朝鮮語に変換し、朝鮮語の疑似対訳データを作成することが考えられる。しかし、朝鮮語には前者のようなピボット言語となる言語がなく、韓国語から朝鮮語への変換は文脈を見て判断する必要があるため、後者のようなルールベースによる変換も困難である。

ピボット言語や疑似データの作成による対訳データの拡張のみでなく、対訳データに対して前処理を行うことで翻訳精度を向上させることができる。例えば、Costa-jussà ら [1] は独英間のニューラル機械翻訳タスクに取り組み、原言語を文字でトークン化し訓練することにより翻訳精度が向上することを示している。また、日本語・中国語間の翻訳においては、文で用いられる漢字を表意文字、ストロークまで分解することにより、より高い翻訳精度を得ることができる [8]。

そこで、本研究では韓国語と朝鮮語間の文法差異を吸収するために、韓国語の入力文を文字でトークン化、または音素に分解する手法を提案する。また、この手法を用いることにより、韓国語・英語の対訳データを利用して朝鮮語から英語への翻訳モデルを効果的に学習できることを示す。本研究の主な貢献は以下の通りである。

- 韓国語・英語対訳データの評価データセットを用いて、韓国語を人手により朝鮮語に変換し、朝鮮語・英語の翻訳精度を評価できる対訳データを作成した。
- 韓国語を文字でトークン化、または音素に分解し翻訳モデルを訓練することにより、そのモデルにおける韓国語から英語への翻訳精度を低下させることなく、朝鮮語から英語への翻訳精度が向上することを示した。

2　韓国語と朝鮮語

2.1　文法差異

表 1 に同じ意味の韓国語と朝鮮語の単語、または文節に現れている文法差異の例を示した。文法差異には分かち書き、頭音法則、合成語に関する違いがある。しかし、合成語に関する違いは、実験に用いた対訳データにおいて出現頻度が著しく低いため、分かち書きと頭音法則の差異を対象とする。

*本研究は首都大学東京の研修生として行った研究である。

[1]朝鮮語は主に朝鮮半島で扱われている言語であるが、大韓民国と朝鮮民主主義人民共和国でいくつかの文法的差異がある。本稿では大韓民国で用いられている朝鮮語を “韓国語”、朝鮮民主主義人民共和国で用いられている朝鮮語を国名をとって “朝鮮語” と呼ぶことにする。

文法差異	韓国語	朝鮮語	日本語	割合
分かち書き	많은 것	많은것	多くのもの	86.9
頭音法則	농구	롱구	バスケット	19.6
	이행	리행	履行	
		이행	移行	
合成語	바닷가	바다가	浜辺	0.3

表 1: 韓国語と朝鮮語の文法差異の例と、韓国語の評価データ内で文法差異が現れている文の割合 (%) 。

分かち書き　韓国語と朝鮮語では形式名詞や固有名詞を含む単語や数量表現等の分かち書きが異なる。韓国語と朝鮮語は主に助詞が現れたら分かち書きをするが、朝鮮語では助詞の次に来る単語が形式名詞であったら続け書きをする。表 1 の例では、「多くのもの」を意味する韓国語は「많은 것」と書かれ、「은（の）」が助詞であるため分かち書きされるが、「것（もの）」が形式名詞であるため朝鮮語では続け書きして「많은것」と書かれる。このように、分かち書きを韓国語文法から朝鮮語文法へ変換するには、前後の単語の品詞を見て判断する必要がある。

頭音法則　韓国語では、語頭で子音である「ㄹ」が母音である「ㅏ, ㅑ, ㅛ, ㅜ, ㅠ , ㅣ, ㅖ」と結合すると、子音が脱落して「ㅇ」 になり、それ以外の母音と結合すると「ㄴ」となるが、朝鮮語では子音は脱落しない。表 1 の「バスケット」を意味する単語は韓国語では頭音法則が適用され「농구」となるが、朝鮮語では「롱구」となる。また、韓国語には頭音法則により多義語になってしまう単語がある。表 1 の「履行、移行」を意味する単語は、韓国語ではどちらも「이행」となるが、朝鮮語では「 리행、이행」となる。このように韓国語には多義語が存在するため、文脈を考慮しなければ頭音法則の差異を吸収することは困難である。

2.2　朝鮮語の評価データ作成

News Korean-English parallel corpus[2]の評価データセットを朝鮮語母語話者一名が人手により朝鮮語の文法に変換し、朝鮮語の評価データセットを作成した。表 1 に評価データ内で文法差異が現れている文の割合を示した。この表からも分かち書き、および頭音法則が韓国語・朝鮮語間の主な文法差異であることが分かる。

[2]https://github.com/jungyeul/korean-parallel-corpora

3　韓国語対訳データを利用した朝鮮語機械翻訳モデル

本研究では韓国語や朝鮮語の入力文を文字レベルでトークン化、または音素に分解することで、韓国語と朝鮮語間の文法差異の影響を低減させ、韓国語対訳データを使用して朝鮮語の機械翻訳モデルを効果的に訓練する手法を提案する。また、以降の文では分かち書きのスペースをわかりやすく␣と表記することにする。

文字レベル　文字レベルでのトークン化では、単語を文字ごとに分かち書きを行う。例えば、表 1 に挙げた「多くのもの」を意味する単語は、韓国語では「많은␣것」、朝鮮語では「많은것」と書かれるが、文字レベルでトークン化すると「많␣은␣것」となり、両言語での違いがなくなる。よって、文字レベルでトークン化することにより分かち書きの差異を無視することが可能となる。

単語レベル（音素 BPE）　単語に含まれる文字を音素（母音と子音）に分解する。これにより、頭音法則の差異を低減させる。例えば、「バスケット」を意味する単語は韓国語では「농구」、朝鮮語では「롱구」と書かれ、文字単位で見ると 1/2 しか共通していないが、音素に分解すると韓国語は「ㄴㅗㅇㄱㅜ」、朝鮮語は「ㄹㅗㅇㄱㅜ」となり、トークンの 4/5 が共通している。このように、音素に分解することで頭音法則の差異を低減することが可能となる。

また、分かち書きは入力文と同じように単語、または文節で分かち書きをする。「롱구는␣운동」という文を音素に分解する場合、「ㄹㅗㅇㄱㅜㄴㅡㄴ␣ㅇㅜㄴㄷㅛㅇ」というようにする。この音素に分解した文に対して BPE (Byte Pair Encoding [5]) を適用することにより、単語、または文節の分かち書きを考慮しつつ、音素単位に分割することが可能となる。

文字レベル（音素 BPE）　入力文に対して、文字レベルでのトークン化を行ったあとに、音素に分解する。文字レベルでのトークン化で分かち書き、音素に分解することで頭音法則の差異を吸収できるため、両方行うことで、２つの差異を同時に吸収することが可能となる。例えば、「롱구는␣운동」という文を文字レベルでトークン化し、音素に分解する場合、「ㄹㅗㅇ␣ㄱㅜ␣ㄴㅡㄴ␣ㅇㅜㄴ␣ㄷㅛㅇ」というようす

hyperparameter	value
Embedding size	512
Hidden layer size	1,024
Enc./Dec. depth	1
Enc./Dec. recurrence transition depth	2
Tie decoder embeddings	yes
Layer normalization	yes
Hidden/Embedding dropout	0.5
Source/Target Word dropout	0.3
Label smoothing	0.2
Optimizer	adam
Learning rate	0.0005
Batch size (tokens)	1,000
Early stoping patience	10
Validation interval	8,000

表 2: ハイパーパラメータ

る。この文に対して BPE を適用することで、文字レベルの分かち書きを考慮しつつ、音素単位に分割することが可能となる。

4 朝鮮語から英語への翻訳実験

4.1 実験設定

本研究で用いたモデルは双方向 LSTM で、実装には Nematus[3]を使用した。また、Sennrich ら [6] を参考としハイパーパラメータを調整した（表 2)。

本研究では News Korean-English parallel corpus を使用しモデルの訓練を行った。モデルの評価には朝鮮語文法に変換した News Korean-English parallel corpus を使用した（2.2 節)。この対訳データは大文字と小文字が truecasing されていないため、すべての入力文対に対して Moses script[4] を用いて tokenization, truecasing を行った。また、訓練データに対しては単語数が 200 以上の文は削除している。表 3 に訓練、開発、テストデータの統計情報を示した。評価の際には翻訳モデルの出力文を Moses script を用いて detruecasing, detokenization を行い、sacreBLEU [4] を用いて BLEU を評価した。本研究では入力言語である韓国語、朝鮮語の単語レベルのデータに加えて、以下の 4 つの前処理を施したデータを用いた。表 4 に各前処理後のデータ情報を示した。

[3]https://github.com/EdinburghNLP/nematus
[4]https://github.com/moses-smt/mosesdecoder

	文数	英語	韓国語	朝鮮語
訓練	93,975	2,297,744	1,567,469	-
開発	1,000	25,804	18,126	15,613
テスト	2,000	53,904	36,641	32,345

表 3: News Korean-English parallel corpus およびその評価データの朝鮮語翻訳の統計情報。訓練、評価データの文数と各言語での単語数を整理した表である。

	タイプ	トークン	平均文長
単語	213,552	1,567,469	16.67
単語（文字 BPE）	32,083	2,057,155	21.89
文字	15,372	4,231,099	45.02
単語（音素 BPE）	29,442	2,091,575	22.25
文字（音素 BPE）	1,736	4,316,529	45.93
単語	53,222	2,297,744	24.45
単語（文字 BPE）	16,024	2,494,763	26.54

表 4: 各前処理後のデータ情報（上：韓国語、下：英語)。訓練データでのタイプ、トークン、平均文長を整理した表である。

単語（文字 BPE） Sennrich ら [6] に従って、単語分割済みの韓国語・英語側それぞれに文字単位の BPE を適用した。オペレーション数は 30k 、最低出現頻度を 10 と設定した。また、以下の韓国語の前処理に対して、英語側はすべて単語（文字 BPE）を使用した。

文字 文字レベルでのトークン化を行う。データ内に含まれる、英語と漢字は文字レベルに区切らずそのまま単語として扱う。また、トークンタイプを頻度上位 1,700 のものに限定した。

単語（音素 BPE） 音素に分解し、BPE を行った。オペレーション数は 30k 、最低出現頻度を 10 と設定した。また、音素への分解には hgtk[5] を用いた。

文字（音素 BPE） 文字レベルでトークン化したデータを音素に分解し、BPE を行った。オペレーション数は 1k とした。

4.2 結果

表 5 は用いたモデルと評価データでの BLEU の値を整理した表である。また、モデルの選択は韓国語、

[5]https://github.com/bluedisk/hangul-toolkit

	韓国語		朝鮮語		分かち書き		頭音法則	
モデル	開発	テスト	開発	テスト	韓国語	朝鮮語	韓国語	朝鮮語
Sennrich and Zhang [6]	-	10.37	-	-	-	-	-	-
単語	7.01	7.41	5.22	5.30	7.65	5.33	8.28	5.48
単語（文字 BPE）	9.50	9.61	9.0	9.32	9.80	9.48	10.27	9.45
文字	**10.20**	9.75	**10.05**	9.70	10.04	9.99	10.33	10.20
単語（音素 BPE）	9.28	**9.88**	9.04	9.01	**10.10**	9.16	**10.68**	9.10
文字（音素 BPE）	10.00	9.76	9.97	**9.75**	10.04	**10.03**	10.43	**10.37**

表 5: 韓国語・朝鮮語から英語辺の翻訳タスクにおける各モデルの評価

朝鮮語の各開発データを使いモデルを選択した。文字モデルが韓国語・朝鮮語のいずれの場合も開発データで最も値が高くなっていて、この時テストデータでは単語（文字 BPE）と比べて韓国語は +0.14、朝鮮語は +0.38 ほど改善がみられる。

5 考察

ここでは各前処理により文法差異を吸収できていることを示すためにテストデータから、分かち書きに差異が現れている文、頭音法則に差異が現れている文、の 2 つのサブセットを抽出しモデルをテストした。各サブセットで文字（音素 BPE）モデルの方が BLEU の値の差が縮まっていることを示す。

分かち書き　表 5 に分かち書きのサブセットでテストをした結果を示した。文字（音素 BPE）モデルが朝鮮語のテストで最も値が高い。また、韓国語・朝鮮語間の BLEU の差も 0.01 となり、分かち書きの文法差異をうまく吸収していることがわかる。

頭音法則　表 5 に頭音法則のサブセットでテストをした結果を示した。頭音法則のサブセットでも、文字（音素 BPE）モデルが朝鮮語のテストで最も値が高く、韓国語・朝鮮語間の BLEU の差も 0.06 となり、頭音法則の差異を吸収できていることがわかる。

6 おわりに

本研究では朝鮮語のリソース問題を解決するために、韓国語や朝鮮語の入力文を文字レベルでトークン化、または音素に分解するという手法を提案した。対訳データの作成やルールベースによる疑似データの作成などはコストが高いが、この手法はシンプルで、かつ、韓国語と朝鮮語間の文法差異を吸収し、朝鮮語から英語への翻訳精度が向上することを示した。

しかし、韓国語と朝鮮語の間に存在する差異は文法差異のみではない。単語の出現分布の差異は大きな違いの一つであるが、本研究では、韓国語を人手により朝鮮語文法に変換したため、考慮されていない。そこで、今後は朝鮮のニュース記事の英語訳データを用いて、単語における差異も考慮した対訳データを作成していきたい。

参考文献

[1] Marta R. Costa-jussà and José A. R. Fonollosa. Character-based neural machine translation. In *ACL*, pp. 357–361, 2016.

[2] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *WMT*, pp. 28–39, 2017.

[3] Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *EAMT*, pp. 129–136, 2011.

[4] Matt Post. A call for clarity in reporting BLEU scores. In *WMT*, pp. 186–191, 2018.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, pp. 1715–1725, 2016.

[6] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *ACL*, pp. 211–221, 2019.

[7] Haifeng Wang, Hua Wu, and Zhanyi Liu. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *ACL*, pp. 874–881, 2006.

[8] Longtu Zhang and Mamoru Komachi. Neural machine translation of logographic language using sub-character level information. In *WMT*, pp. 17–25, 2018.