バイリンガルサブワード分割のための EM アルゴリズム

松井 大樹 [1]　二宮 崇 [2]　田村 晃裕 [3]
[1] 愛媛大学　[2] 愛媛大学 大学院理工学研究科 [3] 同志社大学
matsui@ai.cs.ehime-u.ac.jp
ninomiya@cs.ehime-u.ac.jp
aktamura@mail.doshisha.ac.jp

概要

本論文ではニューラル機械翻訳のための EM アルゴリズムを用いたバイリンガルサブワード分割法を提案する．機械翻訳における一般的なサブワード分割では対訳関係を考慮せずに各言語ごとにサブワード分割を学習するため，機械翻訳タスクに適したサブワード分割になるとは限らない．本研究は対訳コーパスを用い，原言語文と目的言語文の対訳関係を考慮したサブワード分割のための EM アルゴリズムを提案する．提案手法は対訳情報を用いるため，より機械翻訳タスクに適したサブワードが得られると考えられる．従来法と提案手法を用いて翻訳性能を比較したところ，WAT ASPEC 英日・日英タスクにおいて，Transformer NMT モデルの性能が最大で 0.6 ポイント改善した．

1　はじめに

ニューラル機械翻訳（Neural Machine Translation, 以下 NMT） [1, 2, 3] では，予め指定した語彙に基づいて計算を行うため，翻訳時の原言語文に低頻度語や未知語が表れると翻訳性能が低下する．このような語彙の問題に対処するため，バイトペア符号化（Byte Pair Encoding, 以下 BPE）（Sennrich et al. 2016） [4] やユニグラム言語モデル（Kudo 2018） [5] などによるサブワード分割が現在広く用いられている．BPE によるサブワード分割は事前トークナイズを要すのに対し，SentencePiece（Kudo 2018） [6] によるユニグラム言語モデルは生文からサブワード列に直接分割するため，日本語や中国語といった分かち書きされていない言語においても形態素解析器を必要としない．

しかしながら，これらの分割法は対訳関係を考慮せず，各言語ごとにサブワード分割を学習するため，機械翻訳タスクに適したサブワード分割になるとは限らない．例として，日英翻訳において，“nonextended” と “延長されなかった” という対訳対があるとする．この場合，“nonextended” は “no next end ed” などよりも “non extend ed” のほうが優れた分割であり，“延長されなかった” は “延　長　され　な　かった” などよりも “延長　されな　かった” のほうが優れた分割である．これは NMT が各サブワードの対訳関係 “non” と “されな”，“extend” と “延長”，“ed” と “かった” を対応付けて学習できるためである．これらの問題を解決するために，対訳関係を考慮したバイリンガルサブワード分割 [7, 8] が提案されている．しかし，出口ら [7] のバイリンガルサブワード分割は原言語サブワード列と目的言語サブワード列のトークン長をそろえるものであり，トークン長が近いとはいえ各トークンがアライメント関係にあるとは必ずしも言えない．Hiraoka ら [8] のバイリンガルサブワード分割は NMT モデルとサブワード分割モデルが一体化しており，利用する場合には同時に NMT モデルの学習が必要となり，サブワード分割および機械翻訳モデルの学習に大きなコストを要する．

本論文では，対訳情報からサブワード列を得る新たなサブワード分割法を提案する．提案手法は，分かち書きされない言語を含む翻訳性能を改善するため，SentencePiece によるユニグラム言語モデル分割に基づいたサブワード列を得る．バイリンガルサブワード分割のための確率モデルを新たに定義し，原言語のサブワードと目的言語のサブワードが対となる確率を EM アルゴリズムを用いて求める．具体的に，提案手法は，ユニグラム言語モデルによって得られる原言語文と目的言語文それぞれの分割候補の組み合わせを求め，各対のサブワードのアライメント関係を取得し，ユニグラム言語モデルによる生起確率とアライメント確率を掛け合わせ，確率が最も大きくなるサブワード列対を選択する．提案手法を

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

用いることで，原言語文と目的言語文のトークンの対訳関係が整うことになり，言語間でトークンが１対１に対応付けされやすくなる．そのため，従来のサブワード分割法より NMT に適した分割が得られることが期待される．

本手法は原言語文と目的言語文の対訳アライメント関係を利用して分割するため，対訳コーパスが与えられる訓練時には問題は起きないが，原言語文しか与えられない翻訳時にはそのままではサブワード分割ができない．そこで提案手法では，EM アルゴリズムによって求められたアライメント確率の周辺化を行い，ユニグラム言語モデルによる生起確率と各原言語文サブワードの周辺確率を掛け合わせ，確率が最も大きくなるサブワード分割候補を選択する．

WAT Asian Scientific Paper Excerpt Corpus（以下，ASPEC） [9] 英日・日英タスクにおいて，従来法と提案手法を用いた翻訳性能を比較したところ，Transformer NMT モデルの性能が改善した．

2 従来法

本節では提案手法の基礎となるユニグラム言語モデルに基づいたサブワード分割法（Kudo 2018）について説明する．ユニグラム言語モデルでは，各サブワードが独立に生起すると仮定し，サブワード列の生起確率 $P_{\mathrm{U}}(\boldsymbol{x})$ を次式により表す．

$$P_{\mathrm{U}}(\boldsymbol{x}) = \prod_{i=1}^{I} P(x_i) \tag{1}$$

$$\forall i \quad x_i \in V, \quad \sum_{x \in V} P(x) = 1$$

ただし，$\boldsymbol{x} = (x_1, x_2, ..., x_I)$ はサブワード列であり，V は語彙集合（サブワード辞書）である．各サブワードの生起確率 $P(x_i)$ は EM アルゴリズムによって周辺尤度 L_{lm} を最大化することにより推定される．

$$L_{lm} = \sum_{s=1}^{|D|} \log(P(X_s)) \tag{2}$$

$$= \sum_{s=1}^{|D|} \log\left(\sum_{\boldsymbol{x} \in S(X_s)} P_{\mathrm{U}}(\boldsymbol{x}) \right) \tag{3}$$

ただし，D は対訳コーパスであり，X_s は D の中の s 番目の原言語文または目的言語文であり，$S(X_s)$ は X_s の分割候補集合である．

生起確率が最大となるサブワード列（最尤解）は次式によって得られる．

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in S(X)} P_{\mathrm{U}}(\boldsymbol{x}) \tag{4}$$

ただし，X は原言語文または目的言語文である．また，k-best 分割候補も X に対するユニグラム言語モデルによって計算される確率 $P_{\mathrm{U}}(\boldsymbol{x})$ に基づいて得ることが出来る．ただし，サブワード列の生起確率は各サブワードの尤度の積の形で表されるため，系列長の短い（トークン数の少ない）サブワード列が高い確率を持つ傾向がある．

SentencePiece におけるユニグラム言語モデルを用いたサブワード分割は生文から直接学習できるため，日本語や中国語といった分かち書きされない言語においても単語分割器や形態素解析器を必要とせずに分割できるという特徴がある．

3 提案手法

本節では，提案手法となるバイリンガルサブワード分割のための EM アルゴリズムについて説明する．まず，サブワード文対の確率モデルの定義を与え，その後，EM アルゴリズムによるパラメータ更新式の導出，対訳コーパスのサブワード分割を行う手法，翻訳時のサブワード分割を行う手法について説明する．提案手法は NMT モデルや訓練法を修正する必要はなく，従来のサブワード分割法を置き換えるだけで適用可能である．

確率モデルは，ユニグラム言語モデルが出力するサブワード列の生起確率と原言語サブワードと目的言語サブワードのアライメント確率の積で与えられる．ただし，各原言語サブワード列と目的言語サブワード列及びそれらのアライメントは明示的に与えられておらず，隠れ状態となっている．そのため，潜在変数付き確率モデルの学習として有名な EM アルゴリズムを用いて，アライメント確率を学習する．次に，最も確率の高いサブワード列のアライメントを選択することで，訓練コーパスのサブワード分割を行う．

NMT の訓練時には対訳コーパスを利用できるが，翻訳時には対訳文が存在しない．そこで，原言語側サブワードのみを参照する周辺確率を求めることで，原言語文のサブワード分割を行う．

3.1 提案手法の確率モデル

原言語文 X と目的言語文 Y が与えられたとき，提案手法における確率モデルを次のように定義する．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

$$P(X,Y) = \sum_{\boldsymbol{x}\in S(X)} \sum_{\boldsymbol{y}\in S(Y)} P_{\mathrm{M}}(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

$$\approx \sum_{k=1}^{K} \sum_{l=1}^{L} P_{\mathrm{M}}(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)}) \tag{6}$$

ただし，X に対するサブワード分割候補 $S(X)$ のうち，サブワード生起確率 $P_{\mathrm{U}}(\boldsymbol{x})$ が高い top-K 個をそれぞれ $\boldsymbol{x}^{(1)},...,\boldsymbol{x}^{(k)},...,\boldsymbol{x}^{(K)}$，$Y$ に対するサブワード分割候補 $S(Y)$ のうち，サブワード生起確率 $P_{\mathrm{U}}(\boldsymbol{y})$ が高い top-L 個をそれぞれ $\boldsymbol{y}^{(1)},...,\boldsymbol{y}^{(l)},...,\boldsymbol{y}^{(L)}$ とする．また，P_{M} は原言語文のサブワード列 $\boldsymbol{x}$ と目的言語文のサブワード列 $\boldsymbol{y}$ に対する確率モデルであり，次式で定義する．

$$P_{\mathrm{M}}(\boldsymbol{x},\boldsymbol{y}) = P_{\mathrm{U}}(\boldsymbol{x})P_{\mathrm{U}}(\boldsymbol{y}) \prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})} \alpha_{uv} \tag{7}$$

ただし，$A(\boldsymbol{x},\boldsymbol{y})$ は原言語のサブワード列 $\boldsymbol{x}$ と目的言語のサブワード列 $\boldsymbol{y}$ の間のアライメントを返す関数であり，アライメントは対応するサブワード対の集合とする．また，α_{uv} は原言語側サブワード u と目的言語側サブワード v が対応する確率である．

3.2 アライメント確率 α_{uv} の算出

ユニグラム言語モデル P_{U} とアライメントを返す関数 A は所与のものとして，EM アルゴリズムを用いてアライメント確率 α_{uv} を求める．

$$P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x},\boldsymbol{y}) = P_{\mathrm{U}}(\boldsymbol{x})P_{\mathrm{U}}(\boldsymbol{y}) \prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})} \alpha_{uv}^{\mathrm{old}} \tag{8}$$

$$P_{\mathrm{M}}^{\mathrm{new}}(\boldsymbol{x},\boldsymbol{y}) = P_{\mathrm{U}}(\boldsymbol{x})P_{\mathrm{U}}(\boldsymbol{y}) \prod_{u,v\in A(\boldsymbol{x},\boldsymbol{y})} \alpha_{uv}^{\mathrm{new}} \tag{9}$$

$$Q = \sum_n \sum_k \sum_l \left(\frac{P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)}) \log P_{\mathrm{M}}^{\mathrm{new}}(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)})}{\sum_{k'}\sum_{l'} P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')})} \right) \tag{10}$$

$\alpha_{uv}^{\mathrm{new}}$ に関して Q 関数を最大化することにより，$\alpha_{uv}^{\mathrm{new}}$ の更新式を得る．

$$\alpha_{uv}^{\mathrm{new}} = \frac{\sum_n \sum_k \sum_l E_{nkluv}}{\sum_{u''}\sum_{v''}\sum_n \sum_k \sum_l E_{nklu''v''}} \tag{11}$$

$$E_{nkluv} = \frac{P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)})}{\sum_{k'}\sum_{l'} P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')})} C_{nkluv} \tag{12}$$

ただし，C_{nkluv} は，n 番目の文対における原言語のサブワード列 $\boldsymbol{x}_n^{(k)}$ と目的言語のサブワード列 $\boldsymbol{y}_n^{(l)}$ に対し，サブワード u と v がアライメント関係となっている回数である．

3.3 訓練データのサブワード分割

訓練データ D の各文対 X，Y に対して，次式に従ってサブワード列 $\boldsymbol{x}^{(\hat{k})}$，$\boldsymbol{y}^{(\hat{l})}$ を求め，サブワード文対として採用する．

$$\hat{k},\hat{l} = \arg\max_{k,l} P_{\mathrm{M}}(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)}) \tag{13}$$

3.4 翻訳時のサブワード分割

翻訳時におけるサブワード分割では，アライメント確率を目的言語側サブワードで周辺化することによって原言語側サブワードの確率を求める．テストデータの各文 X に対して，次式に従ってサブワード列 $\boldsymbol{x}^{(\hat{k})}$ を求め，サブワード文として採用する．

$$\alpha'_u = \sum_{v\in V_{\mathrm{target}}} \alpha_{uv} \tag{14}$$

$$\hat{k} = \arg\max_k P_{\mathrm{M}'}(\boldsymbol{x}^{(k)}) \tag{15}$$

$$P_{\mathrm{M}'}(\boldsymbol{x}) = P_{\mathrm{U}}(\boldsymbol{x}) \prod_{u\in\boldsymbol{x}} \alpha'_u \tag{16}$$

ただし，V_{target} は目的言語側のサブワード集合である．

4 実験

4.1 実験設定

提案手法と従来法（ユニグラム言語モデル）の翻訳性能を比較した．ユニグラム言語モデルによるサブワード列候補集合を得るために，SentencePiece[1] を利用した．原言語側サブワードと目的言語側サブワードのアライメントを得るために，fast_align[2] [10] を利用した．NMT には Fairseq [11] を使用し，Transformer base（Vaswani et al. 2017） [12] モデルを利用した．翻訳性能を評価するために，sacreBLEU を利用した．sacreBLEU [13] の日本語のトークナイズには ja-mecab [14] を，英語のトークナイズには 13a を利用した．

4.2 データセットとハイパーパラメータ

データセットには WAT ASPEC 英日・日英翻訳タスク[3] を用いた．NMT の訓練には訓練データのうち，100 万文対 (train-1.txt) を利用した．開発データ

1） https://github.com/google/SentencePiece
2） https://github.com/clab/fast_align
3） http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

	英日	日英
ユニグラム言語モデル（従来手法）	27.4	26.7
バイリンガルサブワード分割（提案手法）	**27.8**	**27.3**

表 1　ASPEC 英-日における翻訳性能の比較（BLEU（%））

従来法	
原言語サブワード	_A k at uki _disease
出力結果	ア キ スキー 病
提案手法	
原言語サブワード	_A ka tu ki _disease
出力結果	アカ ツ キ 病
正解データ	
	アカツキ病

表 2　提案手法の良い例

従来法	
原言語サブワード	_a _expressway
出力結果	高速道路
提案手法	
原言語サブワード	_a _express way
出力結果	急 行 路
正解データ	
	高速道路

表 3　提案手法の悪い例

とテストデータのデータ数はそれぞれ 1,790，1,812 文対であった．ユニグラム言語モデルの学習は，原言語側と目的言語側で独立して行い，辞書サイズはどちらも 16,000 に設定した．候補数は原言語側と目的言語側それぞれユニグラム言語モデルによるサブワード生起確率が高い上位 10 通り（top-k=top-l=10）とした．すべての Transformer base モデルにおいて，パラメータの最適化には adam（Kingma and Ba 2014）[15]，学習率は 1e-4，バッチサイズは 128 とし，その他のパラメータは Fairseq のデフォルトのままとした．学習は 30 エポックで終了させ，各エポックのモデルのうち，開発データ上で最も性能のよかったものを利用してテストデータの翻訳を行った．実験はランダムシードを変えて 2 度行い，その平均を実験結果とした．

4.3　実験結果

実験結果の BLEU [16] スコアを表 1 に示す．表 1 から分かる通り，バイリンガルサブワード分割は英日，日英翻訳の両言語方向において，ユニグラム言語モデルより性能が改善されている．バイリンガルサブワード分割を用いることでユニグラム言語モデルに対し，英日・日英翻訳においてそれぞれ 0.4, 0.6BLEU ポイントの性能改善が確認された．

4.4　考察

提案手法によるサブワード分割について考察する．バイリンガルサブワード分割を用いることで良く翻訳できた例を表 2 に示す．提案手法は従来手法よりも原言語サブワードが出力結果に対応するように分割されていることから，正確に翻訳できたと考えられる．バイリンガルサブワード分割を用いることで悪くなった例を表 3 に示す．提案手法は従来手法よりも細かく分割されており，“express” と “急行”，“way” と “路” を関連付けるよう学習していることから，“express way” を “急 行 路” と不正確に翻訳したと考えられる．

5　まとめ

本論文では，確率モデルと EM アルゴリズムを用いたニューラル機械翻訳のための新たなサブワード分割法を提案した．アライメント確率を導入したバイリンガルサブワード分割のための確率モデルの定義を与え，そのアライメント確率を求める EM アルゴリズムを導出した．訓練コーパスに対するサブワード分割は，提案する確率モデルを用いて，最も確率の高いサブワード列対を選択することで実現した．翻訳時には，目的言語側のサブワード列が得られないため直接提案モデルを適用することはできないが，アライメント確率を周辺化することで，原言語側だけで定義される確率モデルを与えた．実験の結果，WAT ASPEC 英日・日英翻訳タスクにおいて，Transformer NMT モデルの性能が改善し，提案手法の有効性を確認した．

今後の課題として，条件付き確率を導入することでアライメントの方向性を考慮した確率モデルに拡張することや，英日以外の言語対での実験等が挙げられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）による助成および JSPS 科研費 JP21K12031 による助成を受けたものです.

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[3] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[4] 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 (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[5] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[7] 出口祥之, 内山将夫, 田村晃裕, 二宮崇, 隅田英一郎. ニューラル機械翻訳のためのバイリンガルなサブワード分割. 自然言語処理, Vol. 28, No. 2, pp. 632–650, 2021.

[8] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Joint optimization of tokenization and downstream model. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 244–255, Online, August 2021. Association for Computational Linguistics.

[9] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[10] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[11] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Sufeng Duan and Hai Zhao. Attention is all you need for Chinese word segmentation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3862–3872, Online, November 2020. Association for Computational Linguistics.

[13] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[14] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[16] 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**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).