Bilingual Subword Segmentation for Neural Machine Translation

HIROYUKI DEGUCHI†‡, MASAO UTIYAMA††, AKIHIRO TAMURA†††, TAKASHI NINOMIYA† and EIICHIRO SUMITA†††

This paper proposes a new subword segmentation method for neural machine translation, called bilingual subword segmentation, which tokenizes sentences to minimize the difference between the number of subword units in a sentence and that in its translation. While existing methods tokenize a sentence without considering its translation, the proposed method tokenizes a sentence using subword units obtained from bilingual sentences and is thus suitable for machine translation. The method was evaluated on WAT Asian Scientific Paper Excerpt Corpus (ASPEC) English-to-Japanese, Japanese-to-English, English-to-Chinese, and Chinese-to-English translation tasks and WMT14 English-to-German and German-to-English translation tasks. The evaluation results reveal that the proposed method improves the performance of Transformer neural machine translation (up to +0.81 BLEU (%)).

Key Words: Neural Machine Translation, Subword, Word Segmentation

† 愛媛大学, Ehime University
‡ 現在, 奈良先端科学技术大学院大学, Nara Institute of Science and Technology (NAIST)
†† 国立研究開発法人情報通信研究機構, National Institute of Information and Communications Technology (NICT)
††† 同志社大学, Doshisha University
1 はじめに

ニューラル機械翻訳（Neural Machine Translation, 以下 NMT）では、予め指定した語彙に基づいて計算を行うため、翻訳時の入力文に低頻度語や未知語が現れると翻訳精度が低下する。このような語彙の問題に対処するため、バイナメペア符号化（Byte Pair Encoding, 以下 BPE）(Sennrich et al. 2016) やユーニグラム言語モデル (Kudo 2018) 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ライズを要すのに対し、ユー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ら書きられない言語においても形態素解析器を必要としない。BPE やユーニグラム言語モデルはどちらもデータ圧縮に基づいたアルゴリズムであり、語彙量の上限を制約したトークン数の最小化を行っている。

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 “設計法 (design method)” と “計測装置 (measurement instrument)” という複合語が訓練データに多数出現する場合を考える。従来のサブワード分割はデータ圧縮技術に基づきトークン数の最小化を行うため、これらの複合語が 1 つのサブワード単位に結合される。したがって、これらの訓練データは “計測法” という言葉の翻訳の学習に寄与しない。

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。提案法は、分から書きられない言語を含む翻訳の性能を改善するため、ユーニグラム言語モデル (Kudo 2018) による分割に基づいたサブワード列を得る。具体的に、提案法は、ユー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から、お互いのトークン数の差が小さくなるサブワード列を選択する方法である。提案法では、ユーニグラム言語モデルから得られる原言語文と目的言語文の最尤解を比較し、トークン数が多い言語側のトークン数に近づけるように、より細かい単位のサブワード分割を複数分割候補から選択する。提案法を用いることで、原言語文と目的言語文のトークン数の差が小さくなり、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。提案法では日本語文と英語文のサブワード分割を近づけるため、課題例として挙げた “設計法” と “計測装置” という複合語は “設計 (design)” と “法 (method)”, “計測 (measurement)” と “装置 (instrument)”, それぞれ 2 トークンに分解される。これにより、語彙量を減らす方法としては文字単位に分割するという方法も考えられるが、文字単位の分割を用いると文全体のトークン数が増える（系列長が長くなる）ため、系列長に依存した計算量が増加する。サブワード分割によって、語彙量の上限を制約して満たす中でトークン数を最小化することで、トレードオフの関係にある語彙量とトークン数（系列長）の問題に対処しているといえる。

なお、提案法はサブワード辞書自体を変えるものではないことに注意されたい。例えば、課題例の “設計法” の場合、“設計法”, “設計”, “法” のいずれもサブワード辞書内に含まれており、従来法では “設計法” が選択されるのに対して、提案法では “設計” と “法” が選択される。
NMTにおいて，“設計”と“法”、“計測”と“装置”というそれぞれのサブワードの訓練データが“計測法”という語の翻訳にも活用できるようになると考えられる。

そこで、本手法は原言語文と目的言語文の分割数を比較しながらそれぞれの文を分割するため、原言語文単体では分割ができない。NMTの訓練時には原言語文と目的言語文の分割数を比較するために対訳コーパスを用いることができるが、翻訳時には原言語文に対応する目的言語文が存在しないため、原言語文を分割することができない。そこで提案法では、対訳コーパスを用いてサブワード分割した訓練データの原言語文からLSTMベースのサブワード分割器を調べ学習し、翻訳時において訓練時の分割に近い候補を選択することで、訓練時と翻訳時の分割のギャップを小さくして翻訳性能の低下を防ぐ。具体的には、翻訳時に、学習したLSTMベースのサブワード分割器により原言語文のサブワード分割候補をランキングし、スコアが最大となる分割を選択する。

WAT Asian Scientific Paper Excerpt Corpus（以下、ASPEC）（Nakazawa et al. 2016）英日・日英・英中・中英翻訳タスクとWMT14英独・独英翻訳タスクにおいて、従来法と提案法を用いた翻訳性能を比較したところ、Transformer NMTモデルの性能が最大0.81 BLEUポイント改善した。

2 従来法：ユニグラム言語モデルに基づいたサブワード分割

本節では提案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$P(\mathbf{x})$を次式により表す。

$$
P(\mathbf{x}) = \prod_{i=1}^{N} p(x_i), \quad (1)$$

$$\forall i \ x_i \in \mathcal{V}, \sum_{x \in \mathcal{V}} p(x) = 1, \quad (2)$$

ただし、$\mathbf{x} = (x_1, x_2, \ldots, x_N)$はサブワード列であり、$\mathcal{V}$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$\mathcal{L}_{lm}$を最大化することにより推定される。

$$
\mathcal{L}_{lm} = \sum_{s=1}^{D} \log(P(X^{(s)})) = \sum_{s=1}^{D} \log \left( \sum_{\mathbf{x} \in \mathcal{S}(X^{(s)})} P(\mathbf{x}) \right), \quad (3)
$$

ただし、$D$は対訳コーパスであり、$X^{(s)}$は$D$中の$s$番目の原言語文または目的言語文であり、$\mathcal{S}(X^{(s)})$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。

\[ x^* = \arg \max_{x \in S(X)} P(x), \quad (4) \]

ただし、\( X \) は入力文である。また、\( k\)-best 分割候補も入力文 \( X \) 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 \( P(x) \) に基づいて得ることができる。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。

このユニグラム言語モデルによる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長がある。

3 提案法

本節では、対訳文からサブワード列を得る提案手法を示す。我々の提案法では対訳文対でサブワードトークン数の差が最小になるような分割を行う。具体的には、原言語文と目的言語文それぞれのユミグラム言語モデルの最尤解のうち、トークン数の少ない（系列長の短い）側の文を、トークン数が多い側のトークン数に近づけるよう、より細かく分割された分割候補からサブワード列を選択する。ただし、NMT の訓練時には対訳コーパスを利用してできるが、翻訳時に（評価データ）には対訳文が存在しない。そこで、NMT の訓練時と翻訳時で異なる方法によりサブワード列を得る。図 1, 2 に NMT 訓練時のサブワード分割と翻訳時のサブワード分割をそれぞれ示す。NMT モデルの訓練時は、図 1 の通り、対訳データに基づくサブワード分割結果を用いて NMT モデルを学習する。一方で翻訳時には、図 2 の通り、対訳データのサブワード分割結果内の原言語文だけから予め学習しておいた LSTM ベースの単語分割器を用いて、翻訳対象の原言語文のサブワード分割候補をリランキングする。

提案法は NMT モデルや訓練法を修正する必要がなく、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。

3.1 訓練データのサブワード分割

訓練データ \( D \) におけるサブワード分割では、ユニグラム言語モデルによる分割候補からトークン数が近い候補を優先的に考慮することで、対訳文対 \( (f, e) \in D \) の分割を得る。具体的には、以下のようにして、原言語文と目的言語文それぞれの \( k\)-best の分割候補 \( B^k(f), B^k(e) \) の中から、対訳文対 \( (f, e) \) のサブワード列 \( (f, e) \) を得る。
図1 訓練時のサブワード分割

\[
(f, e) = \begin{cases} 
(\hat{u}, e^*) & \text{if } \text{len}(f^*) < \text{len}(e^*) \\
(f^*, \hat{u}) & \text{otherwise}
\end{cases},
\]

ただし, \text{len}() はサブワードトクン数を返す関数であり, \( f^* \) と \( e^* \) は, それぞれ, 原言語と目的言語文の最大の確率を持つサブワード列（ユニグラム言語モデルの尤度）である. そして, \( \hat{u} \) は, \( v^* \) を \( f^* \) と \( e^* \) のうち系列長の長い方としたとき, \( v^* \) とのトクン数の差が最小の候補の中から最大の確率を持つサブワード列であり, 以下の式で表される.
図2 翻訳時のサブワード分割

\[ \hat{u} = \arg\max_{u \in T} P(u), \quad (6) \]

\[ T = \arg\min_{u \in B^k} |\text{len}(\hat{u}) - \text{len}(\hat{v}^*)|, \quad (7) \]

\[ B^k = \begin{cases} B^k(f) & \text{if } \text{len}(\hat{f}^*) < \text{len}(\hat{e}^*) \\ B^k(e) & \text{otherwise} \end{cases} \quad (8) \]

ここで、式7、8より、分割候補の選択は、最尤解のトークン数の少ない側（系列長の短い側）
において行われる。\(^3\)

NMT モデルは、提案法により各対訳文をサブワード分割した訓練データ \(\mathcal{D} = \{ (f^{(s)}, e^{(s)}) \}_{s=1}^{\mathcal{D}} \) から学習される。

### 3.2 翻訳時のサブワード分割

翻訳時は入力文 \(f\) に対する対訳文 \(e\) が存在しないため、サブワード分割の入力に対訳文を用いることができない。そのため、予め 3.1 節で作成した訓練データ \(\mathcal{D} \) の原語文 \(\tilde{\mathcal{D}} \) から、文字ベースの双方向 LSTM（Bidirectional LSTM、以下 BiLSTM）を用いたサブワード分割器（以下、BiLSTM 分割器）を学習しておく。そして、翻訳時の分割は、ユニグラム言語モデルの \(k\)-best 分割候補を BiLSTM 分割器に入力し、各分割候補に対してスコア付けを行いランキングすることにより得られる。

BiLSTM 分割器は、\(n\) 個の文字からなる入力文字列 \(c = (c_1, c_2, \ldots, c_n)\) に対して、サブワードの開始文字が否かを表す境界タグを割り当て、サブワードの境界点を 2 値分類として識別する。BiLSTM 分割器は以下のよう構造のニューラルネットワークである。

\[
\begin{align*}
z &= \text{Embedding}(c), \\
h &= \text{BiLSTM}(z), \\
b &= \text{softmax}(hW),
\end{align*}
\]

ただし、Embedding() は文字埋め込み層、\( z \) は文字列 \( c \) の \(d\) 次元埋め込み表現、BiLSTM() は BiLSTM 層、\( h \) は BiLSTM の隠れベクトル、softmax() は softmax 関数、\( b \) は BiLSTM 分割器の出力、\( W \in \mathbb{R}^{d \times \{0,1\}} \) は隠れベクトル \( h \) の空間から境界タグ次元に写像するパラメータ行列である。ベクトル \( b_t = (b_{t,0}, b_{t,1}) \) は文字 \( c_t \) がサブワードの開始点か \( (b_{t,0}) \) 開始点でないか \( (b_{t,1}) \) の確率分布を表現している。BiLSTM 分割器は、3.1 節の方法でサブワード分割された訓練データ \(\tilde{\mathcal{D}}\) 中の全原語文 \(\tilde{f} \in \{ f^{(s)} \}_{s=1}^{\tilde{\mathcal{D}}} \) について、以下の目的関数 \(\mathcal{L}_{\text{segment}}\) を最大化することにより学習される。

\[
\mathcal{L}_{\text{segment}} = \sum_{t=1}^{n} \log b_{t,r_t},
\]

where \( r_t = \begin{cases} 0 & \text{if } c_t \text{ is subword's start point} \\ 1 & \text{otherwise} \end{cases} \).

\(^3\) ユニグラム言語モデルは各サブワードの尤度の積によって求められる生起確率に基づいて分割を行うため、\(k\)-best分割候補のうち、最尤解に近づくほどトークン数は少なくなる傾向がある。そのため、通常は最尤解よりもトークン数の少ない分割候補は得られない。したがって、提案法では最尤解のトークン数の少ない側（系列長の短い側）においてより細かく分割された分割候補を選択する。
翻訳時は次のようにして入力文 \( f \) をサブワード分割する。はじめに、ユニグラム言語モデルを用いて入力文 \( f \) の \( k \)-best サブワード分割候補 \( B^k(f) \) を得る。次に、各分割候補 \( f \in B^k(f) \) のスコア \( \text{score}(f) \) を、予め学習しておいた BiLSTM 分割器によって以下のように算出する。

\[
\text{score}(f) = \sum_{t=1}^{n} \log b_{t,r},
\]

(14)

最後に、最大のスコアを持つサブワード列を選択し、出力とする。

\[
f^* = \arg \max_{f \in B^k(f)} \text{score}(f).
\]

(15)

以上により得られたサブワード列 \( f^* \) をNMTモデルに入力し、翻訳を行う。

4 実験

4.1 実験設定

提案法と従来法（ユニグラム言語モデル (Kudo 2018)）の翻訳性能を比較した。また、従来法として、ユニグラム言語モデルによって得られる複数のサブワード分割候補について周辺尤度を最大化する“サブワード正則化 (Kudo and Richardson 2018)” とも性能を比較した。複数サブワード分割候補を得るためのユニグラム言語モデルにはSentencepiece⁴を用いた。全実験において、NMTシステムとしてTransformer base (Vaswani et al. 2017) モデルを用いた。

4.1.1 データセット

翻訳性能はWAT ASPEC 日英・英日（以下 ASPEC 日-英）翻訳タスク⁵ (Nakazawa et al. 2016) を用いて評価した。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それぞれ独立に行い、サブワードの語彙量は、原言語側と目的言語側でそれぞれ16,000になるように設定した。ミニバッチの大きさは約10,000トークンになるよう設定した。NMTの訓練には訓練データの上位150万文対を使用し、データの前処理はWATベースラインシステム⁶に従った。開発データと評価データのデータ数はそれぞれ1,790、1,812文対であった。

4.1.2 ハイパーパラメータ

全NMTモデルにおいて、パラメータ最適化にはAdam (Kingma and Ba 2014) を用い、\( \beta_1 = 0.9, \beta_2 = 0.98 \) とした。モデルのパラメータ更新は10万回行った。学習率は4,000回更新時で5e-4となるように線形に増加させ、以降は更新回数の逆平方根に比例して減衰させた (Vaswani et al.

⁴ https://github.com/google/sentencepiece
⁵ http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
⁶ http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/baseline/dataPreparationJE.html
2017). ドロップアウトの確率は 0.1 に設定した。NMT モデルの損失関数にはラベル平滑化交差エントロピー (Szegedy et al. 2016) を用い、平滑化 $\epsilon$ は 0.1 に設定した。パラメータの更新 1,000 回毎にモデルを保存し、性能評価時には、訓練終了時点で前 5 つ分のモデルパラメータを平均化したモデルを用いた。翻訳文の生成にはビーム探索を用い、ビーム幅は 4、文長正則化パラメータは 0.6 (Wu et al. 2016) とした。

提案法のハイパラメータに関して、ユニグラム言語モデルから得るサブワード分割候補数 $k$ は開発データで調整し、5 に設定した。BiLSTM 分割器の埋め込み次元は $d = 256$ とし、BiLSTM 層は 2 層スタックした。文字埋め込み層、BiLSTM 層、出力層のパラメータは全て $[-0.1, 0.1]$ の一様分布で初期化した。BiLSTM 分割器のパラメータ最適化には Adam を用い、$eta_1 = 0.9, \beta_2 = 0.98$ とした。モデルのパラメータ更新は 10 エポック分行った。学習率は 5e-4、ドロップアウトの確率は 0.1、ミニバッチの大きさは約 256 文にそれぞれ設定した。

サブワード正則化を用いたモデルでは、提案法と条件を揃えるため、ユニグラム言語モデルの最大スコアのサブワード列を翻訳する 1-best デコードを使用した。

### 4.2 実験結果

表 1 に実験結果を示す。表中の“ユニグラム言語モデル”、“サブワード正則化”、“BiSW” はそれぞれ、ユニグラム言語モデル、サブワード正則化、提案法を用いた NMT モデルを示している。翻訳性能は BLEU (Papineni et al. 2002) で評価し、評価方法は WAT Automatic Evaluation Procedures に基づく。また、ブートストラップ再サンプリングによる有意差検定 (Koehn 2004) を実施し、有意水準は 5% とした ($p \leq 0.05$)。表 1 中の“†”は“BiSW”が“ユニグラム言語モデル”に対して、“‡”は“BiSW”が“サブワード正則化”に対して、有意に高いことを示す。

表 1 から分かるとおり、提案法“BiSW”は日英、英日翻訳の両言語方向において“ユニグラム言語モデル”および“サブワード正則化”より性能が改善されている。“BiSW”を用いることで“ユニグラム言語モデル”に対し、日英、英日翻訳においてそれぞれ 0.81, 0.10 BLEU ポイント、“サブワード正則化”に対し、それぞれ 0.53, 0.19 BLEU ポイントの性能改善が確認された。また、両言語方向において、提案法はベースラインの“ユニグラム言語モデル”，及び“サ

|                | 日英 | 英日 |
|----------------|-----|------|
| ユニグラム言語モデル | 28.58 | 43.19 |
| サブワード正則化 | 28.86 | 43.10 |
| BiSW（提案法） | †‡29.39 | †‡43.29 |

表 1 ASPEC 日-英における翻訳性能の比較 (BLEU(％))

† [http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html#automatic_evaluation_systems.html](http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html#automatic_evaluation_systems.html)
ブリード正則化”より有意に性能が高く、提案法の有効性が確認できる。

5 考察

5.1 BiLSTM 分割器の性能とオラクル分割

本節では翻訳時に用いる BiLSTM 分割器のリーキングによる分割性能を考察する。分割性能の評価には ASPEC の評価データを用い、参照訳を使用して 3.1 節の手法により分割した原言語文の分割結果を正解の分割とみなした。結果を表 2 に示す。表 2 のとおり、BiLSTM 分割器のリーキングによる分割性能は非常に高く、正解分割に近いサブワード分割を得ていることが分かる。

さらに、正解分割の原言語文を用いた翻訳性能を評価した。この正解分割に対する翻訳性能は提案法の性能の上限値を示しているといえる。結果を表 3 に示す。表 3 の “オラクル分割”が正解分割に対する翻訳性能を示している。表 3 より、“オラクル分割”は BiSW より高い翻訳性能となっていることが確認できる。“BiSW”の翻訳性能は “オラクル分割”と比較して、日英、英日翻訳においてそれぞれ 0.10、0.20 BLEU ポイント低く、これは BiLSTM 分割器の予測の誤差によりオラクル分割と異なる分割が得られたため性能が低下したと考えられる。

5.2 BiLSTM 分割器の必要性

提案法では翻訳時の入力文をサブワード分割するため BiLSTM 分割器を必要とする。本節では翻訳時の BiLSTM 分割器の必要性を確認するため、BiLSTM 分割器を用いたときと用いないときの翻訳性能を比較した。BiLSTM 分割器を用いない場合は、翻訳時の原言語文の分割結果としては、ユニグラム言語モデルの最尤解 (f* ) を用いる。また、訓練時と翻訳時とのギャップを失くすため、訓練時には、原言語文をユニグラム言語モデルの最尤解 (f* ) で固定する。つまり

| 適合率 (%) | 再現率 (%) | F 値 (%) |
|-----------|-----------|----------|
| 日        | 97.05     | 97.44    | 97.24    |
| 英        | 98.82     | 99.22    | 99.02    |

表 2 ASPEC 日-英の評価データにおける BiLSTM 分割器のサブワード分割性能

|               | ASPEC 日-英 |
|---------------|-------------|
|               | 日英 英日   |
| BiSW          | 29.39 43.29 |
| オラクル分割   | 29.49 43.49 |

表 3 オラクル分割の翻訳性能 (BLEU(%))

641
5.3 提案法によるサブワード分割の例

本節では従来法“ユニグラム言語モデル”と提案法で得られるサブワードの違いを実例で確認する。表5に、APSEC 日英の訓練データに対して従来法と提案法をそれぞれ適用した際の例を示す。表5より、従来法では複数の意味から成るサブワードが1トークンに結合されているのに対し、提案法ではそれらが分解されていることが分かる。また、表中に訳文中の対応箇所を示す。表より、“設計法”が訳文中の“design method”と対応付して“設計”と“法”に分割されていることが確認できる。これは、従来法が生起確率のみに基づいて分割されるのに対し、提案法では対訳相手の分割情報を参照しているためであるといえる。これにより、原言語

|                | ASPEC 日-英 | 日英 | 英日 |
|----------------|-------------|-----|-----|
| BiSW           | ↑29.39      | ↓43.29 |
| BiSW w/o BiLSTM| 28.80       | 43.00 |

表4 BiLSTM分割器を用いないときの翻訳性能比較 (BLEU(％))

| ユニグラム言語モデル | BiSW | 訳文中の対応箇所   |
|---------------------|------|----------------|
| helper              | help er | ヘルパーエル |
| basically           | basic ally | 基本的には |
| focused             | focus ed | に着目して |
| popularization      | popular ization | 普及 |
| 第三者              | 第三者 | the third person |
| 骨密度              | 骨密度 | bone density |
| 設計法              | 設計法 | design method |

表5 ASPEC 日英翻訳の訓練データにおけるサブワードの例
文と目的言語文間に付与される頻度を測定した結果、NMT モデルの学習を支援できるようになると期待される。ただし、“popularization”の語彙学において対応箇所は“普及”という1トークンであるのに対し、提案法による分割では“popular”と“ization”に分割されている。これらは単語単位ではなく、文単位でトークン数を近づけるため、“普及”以外のトークンの分割を参照し、語彙学の対応箇所とトークン数の対応がない分割を行ったと考えられる。

表6にASPEC日本翻訳の評価データ（評価データの日本語）に対して従来法と提案法をそれぞれ適用した実例の例を示す。表6より、評価データにおいても、対訳文（参照訳）を参照することなく、BiLSTM分割器により、言語間で1対1のサブワードの対応付けを取りやすい単位に分解されていることが分かる。ただし、提案法において“透水性”と分割された例は、語彙学の対応箇所が“preamability”であるのに対し、従来法の“透水性”よりもトークン数の差が大きくなっている。これは、文単位でトークン数を近づけた訓練データからBiLSTM分割器を学習させているため、もしくは、BiLSTM分割器が誤って予測しているためであると考えられる。

5.4 ハイパーパラメータ$k$に対する敏感さ
提案法では分割候補の候補数$k$がハイパーパラメータとなっている。本節では、ハイパーパラメータ$k$の値によって提案法の翻訳性能がどの程度変化するかを考察する。図3に、ASPEC日本翻訳の開発データにおける$k$を変化させたときの翻訳性能を示す。$k$の値を${[2, 10], 15, 20, 50, 100}$の中で変化させたときの翻訳性能を評価した。

図3より、一部例外はあるものの$k$が50を超えるまでは概ね翻訳性能が改善していることが確認できた。

5.5 尤度に基づいた対訳文対のサブワード分割手法
Kudo(2018)の文献では“the unigram language model is reformulated as an entropy encoder that minimizes the total code length for the text. According to Shannon’s coding theorem, the optimal code length for a symbol $s$ is $-\log p_s$, where $p_s$ is the occurrence probability of $s$.”と述べられている。このことから、サブワード分割した文のトークン数とそのサブワード列の尤度の間には関係性があると考えられる。そこで本節では、提案手法において、文のトークン数
の代わりに尤度を用いた場合の性能を考察する。つまり、原言語文と目的言語文のトークン数の差を小さくする代わりに、尤度の差が小さくなるように分割を行った場合の性能を評価する。具体的には、式 5 から式 8 において、\( \text{len()} \) を \( -\log P() \) で置き換えて計算することで、原言語文と目的言語文をサブワード分割した際の尤度の差が小さくなるような分割候補を選択する。

表 7 に結果を示す。表中の“BiSW（トークンベース）”と“BiSW（尤度ベース）”はそれぞれトークン数に基づいた提案法と尤度に基づいた提案法を示している。表 7 より、尤度に基づいた手法は日英翻訳においてはユニグラム言語モデルよりも翻訳性能が改善されているが、両言語方向においてトークン数に基づいた提案法よりも性能が低い。これは、尤度とトークン数の間の関係はあるものの完全に一致していないためであると考えられ、関係の度合いはユニグラム言語モデルの性能に依存していると考えられる。

さらに、ASPEC 日-英の訓練データおよび評価データにおいて、対訳文対間のサブワードトークン数の差の平均を調査した。結果を表 8 に示す。表 8 より、トークン数ベースと尤度ベースのどちらの手法も訓練データ、評価データの両方においてユニグラム言語モデルよりもトークン数の差は縮まっている。また、尤度ベースよりもトークン数ベースの提案法のほうがよりトークン数の差が小さくなっていることも確認できる。この結果より、トークン数ベースの提案法

### 図 3 ハイパラメータ \( k \) に対する敏感性（開発データにおける翻訳性能）

|        | ASPEC 日-英 |        |        |
|--------|-------------|--------|--------|
|        | 日英 | 英日 |
| ユニグラム言語モデル | 28.58 | 43.19 |
| BiSW（トークンベース） | 29.39 | 43.29 |
| BiSW（尤度ベース） | 29.28 | 43.09 |

表 7 尤度ベースの手法との比較 (BLEU(%))
5.6 分から書きされた言語対及び分から書きされない言語対に対する提案法の有効性

本節では、英独翻訳のような分から書きされている言語対、及び、日中翻訳のような分から書きされない言語対の翻訳に対する提案法の有効性を検証する。

翻訳性能の評価には、それぞれ WMT14 英独・独英（以下 WMT14 英-独）翻訳タスクと ASPEC 日中・中日（以下 ASPEC 日-中）翻訳タスクを用いた。

本実験における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辞書を共有して行った。サブワードの語彙量は、WMT14 英-独で 37,000、ASPEC 日-中で 16,000 に設定し、NMT モデル内の原言語側と目的言語側の埋め込み層を共有した。ミニバッチの大きさは WMT14 英-独で約 25,000 トークン、ASPEC 日-中で約 6,000 トークンになるよう設定した。WMT14 英-独の訓練データにおいて、各文をサブワード分割した後、250 トークンを超える文と原言語/目的言語文のトークン数の比が 1.5 を超えるものを除去した。ハイパラメータ \( k \) は開発データで調整し、2 に設定した。

表 9 に WMT14 英独・独英翻訳、ASPEC 日中・中日翻訳の実験結果を示す。表 9 より、両言語方向において、提案法 “BiSW” を用いることで従来法 “ユニグラム言語モデル” と比べて翻訳性能が改善されることが確認された。具体的には、英独、独英、日中、中日翻訳において、 “BiSW” は “ユニグラム言語モデル” と比べてそれぞれ 0.32、0.02、0.11、0.12 BLEU ポイント

|                  | ASPEC 日-英 |
|------------------|------------|
|                  | train test|
| ユニグラム言語モデル | 7.83 6.07 |
| BiSW（トークン数ベース） | 6.74 4.98 |
| BiSW（尤度ベース） | 7.10 5.38 |

表 8 対訳文対間のサブワードトークン数の差の平均

|                  | WMT14 英独 | ASPEC 日中 |
|------------------|------------|
|                  | 英独 独英 日中 中日|
| ユニグラム言語モデル | 26.45 30.62 35.21 47.59 |
| BiSW（提案法） | 26.77 30.64 35.33 47.71 |

表 9 WMT14 英独及び ASPEC 日-中翻訳における提案法の有効性 (BLEU(%)）

\(^8\) https://www.statmt.org/wmt14/translation-task.html
性能が改善された。実験結果より、分から書きされた言語対及び分から書きされない言語対に対しても提案法の有効性が確認された。

6 関連研究

BPE (Sennrich et al. 2016) とユニグラム言語モデル (Kudo 2018) はサブワード分割法として広く用いられている。BPE は辞書式圧縮に基づいたサブワード分割アルゴリズムであり、指定した語彙量を上限として、出現回数順に隣接するサブワードを再帰的に結合する。BPE は簡単なアルゴリズムで実装が容易なため多くの NMT システムで採用されているが、決定的アルゴリズムであるため複数の分割候補を得ることができない。

ユニグラム言語モデルは尤度に基づいたサブワード分割アルゴリズムである。各サブワードの生起確率は EM アルゴリズムによって推定される。ユニグラム言語モデルは BPE と比べてアルゴリズムが複雑であるが、尤度に基づいた複数のサブワード分割候補を得られ、かつ、事前トークナースを必要とせず生産から直接学習できるという特徴がある。本研究の実験ではユニグラム言語モデルのリファレンス実装である SentencePiece (Kudo and Richardson 2018) を用いた。

サブワード正則化 (Kudo 2018) は複数のサブワード分割候補を用いた NMT の訓練法であり、サンプリングされた分割候補の周辺尤度を最大化する。サブワード正則化を NMT に組み込むには、訓練時にパラメータを更新することに動的にサブワード分割をサンプリングする必要があり、NMT の訓練処理を修正する必要がある。

BPE-dropout (Provilkov et al. 2020) はサブワード正則化を用いるように BPE を拡張した手法である。BPE-dropout では、隣接サブワードの結合を確率的に棄却することで複数のサブワード分割候補が得られる。ただし、P(x|X) のような尤度に基づいた k-best 候補を得ることはできない。

単語やサブワードへの分割を行わずに文字単位で翻訳を行う NMT モデルも提案されている。Cherry ら (Cherry et al. 2018) は単語単位やサブワード単位の NMT よりも文字単位の NMT の翻訳性能が高くなると報告している。ただし、Cherry らは文字単位の NMT の問題点として計算量の多さとモデリングの難しさがあることを述べている。我々の手法は NMT モデルの出力の粒度について文字単位の NMT の長所と短所（翻訳性能とモデリング・計算量）のバランスをとったものとも考えられる。

Ataman ら (Ataman et al. 2017; Ataman and Federico 2018b) や Huck ら (Huck et al. 2017) は言語学に基づくサブワード分割を提案している。Ataman ら (Ataman et al. 2017; Ataman and Federico 2018b) は教師なし形態学に基づく“Linguistically Motivated Vocabulary Reduction (LMVR)” を用いることで BPE より翻訳性能が向上することを示した。Huck ら (Huck et al.
2017) はサブワード分割においてステミングや複合語分割などによる言語学的な知識を用いた分割を組み合わせることで、翻訳性能が改善することを示した。また、Atamanら (Ataman and Federico 2018a) は単語を n-gram 文字で分解することで形態学的にリッチな言語を含む翻訳が改善することを示している。

7 おわりに

本論文では、対訳文からサブワード列を得る、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。WAT ASPEC 英日・日英・英中・中英翻訳タスクと WMT14 英独・独英翻訳タスクにおいて、提案法を用いることで Transformer NMT モデルの性能が最大 0.81 BLEU ポイント改善した。実験と考察により、対訳文とのサブワードトークン数の差を小さくすることで翻訳性能が改善されることを示した。今後は他の言語対での提案法の有効性も確認していきたい。

謝 辞

本論文は国際会議 The 28th International Conference on Computational Linguistics (COLING'2020) に採択された論文に基づいて日本語で書き直し、説明を追加したものである。

本研究成果は、国立研究開発法人情報通信研究機構の委託研究により得られたものである。また、本研究の一部は JSPS 科研費 20K19864 の助成を受けたものである。ここに謝意を表する。

参考文献

Ataman, D. and Federico, M. (2018a). “Compositional Representation of Morphologically-Rich Input for Neural Machine Translation.” In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 305–311, Melbourne, Australia. Association for Computational Linguistics.

Ataman, D. and Federico, M. (2018b). “An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation.” In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers), pp. 97–110, Boston, MA. Association for Machine Translation in the Americas.

Ataman, D., Negri, M., Turchi, M., and Federico, M. (2017). “Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English.” arXiv preprint arXiv:1707.09879.
Cherry, C., Foster, G., Bapna, A., Firat, O., and Macherey, W. (2018). “Revisiting Character-Based Neural Machine Translation with Capacity and Compression.” In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

Huck, M., Riess, S., and Fraser, A. (2017). “Target-side Word Segmentation Strategies for Neural Machine Translation.” In Proceedings of the 2nd Conference on Machine Translation, pp. 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). “Adam: A Method for Stochastic Optimization.” CoRR, abs/1412.6980.

Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 388–395, Barcelona, Spain. Association for Computational Linguistics.

Kudo, T. (2018). “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. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). “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. Association for Computational Linguistics.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). “ASPEC: Asian Scientific Paper Excerpt Corpus.” In Proceedings of LREC 2016, pp. 2201–2206.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “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. Association for Computational Linguistics.

Provilkov, I., Emelianenko, D., and Voita, E. (2020). “BPE-Dropout: Simple and Effective Subword Regularization.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1882–1892, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). “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. Association
for Computational Linguistics.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). “Rethinking the Inception Architecture for Computer Vision.” In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). “Attention is All you Need.” In Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” arXiv preprint arXiv:1609.08144.

略歴
出口 祥之：2019年愛媛大学工学部情報工学科卒業. 2021年同大学院理工学研究科博士前期課程修了. 2021年より奈良先端科学技术大学院大学博士後期課程に在学.
内山 将夫：1992年筑波大学卒業. 1997年同大学院工学研究科修了. 博士 (工学). 現在. 国立研究開発法人情報通信研究機構上席研究員. 主な研究分野は機械翻訳. 言語処理学会会員.
田村 晃裕：2005年東京工業大学工学部情報工学科卒業. 2007年同大学院総合理工学研究科修士課程修了. 2013年同大学院総合理工学研究科博士課程修了. 日本電気株式会社. 国立研究開発法人情報通信研究機構にて研究員として務めた後. 2017年より愛媛大学大学院理工学研究科助教. 2020年より同高校大学理工学部准教授となり. 現在に至る. 博士 (工学). 言語処理学会. 情報処理学会. 人工知能学会. ACL 各会員.
二宮 晃：1996年東京大学理学部情報科学研究科卒業. 1998年同大学院理学系研究科修士課程修了. 2001年同大学院理学系研究科博士課程修了. 同年より科学技術振興事業団研究員. 2006年より東京大学情報基盤センター講師. 2010年より愛媛大学大学院理工学研究科准教授. 2017年同教授. 博士 (理学). 言語処理学会. アジア太平洋機械翻訳協会. 情報処理学会. 人工知能学会. 電子情報通信学会. 日本データベース学会. ACL 各会員.
隅田英一郎：1982年電気通信大学大学院修士課程修了. 1999年京都大学大学院博士（工学）取得. 1982年〜1991年（株）日本アイ・ビー・エム東京基礎研究所研究員. 1992年〜2009年国際電気通信基礎技術研究所研究員. 主幹研究員. 室長. 2007年〜現在 国立研究開発法人情報通信研究機構 (NICT) 先進
的音声翻訳研究開発推進センター（ASTREC）副センター長。2016年NICTフェロー。機械翻訳の研究に従事。

(2020年10月31日 受付)
(2021年2月8日 再受付)
(2021年3月12日 採録)