Vol.2020-NL-246 No.22
Vol.2020-SLP-134 No.22
2020/12/3

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

出口 祥之[1,a)]　内山 将夫[2]　田村 晃裕[3]　二宮 崇[1]　隅田 英一郎[2]

概要：本論文では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する．従来法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サブワードトークン数の差がより小さくなるサブワード分割法を提案する．提案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．従来法と提案法を用いて翻訳性能を比較したところ，WAT ASPEC 英日・日英翻訳タスクと WMT14 英独・独英翻訳タスクにおいて，Transformer NMT モデルの性能が最大 0.81 BLEU ポイント改善した．

キーワード：ニューラル機械翻訳，サブワード，単語分割

Bilingual Subword Segmentation for Neural Machine Translation

Deguchi Hiroyuki[1,a)]　Utiyama Masao[2]　Tamura Akihiro[3]　Ninomiya Takashi[1]
Sumita Eiichiro[2]

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，バイトペア符号化（Byte Pair Encoding，以下 BPE）[13] やユニグラム言語モデル [8] 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，ユニグラム言語モデルは生文からサブワード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，語彙数の上限を制約と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．

本論文では対訳情報からサブワードを得る新たなサブワード分割法を提案する．提案法は，分かち書きされない言語を含む翻訳の性能を改善するため，ユニグラム言語モデル [8] に基づいたアルゴリズムとなっている．具体的には，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．提案法では原言語文と目的言語文のトークン数の差を最小化するため，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．

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法はデータ圧縮技術に基づきトークン数の最小化を行なうため，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” という語の翻訳の学習に寄与しない．一方で我々の提案法を用いると，日本語文と英語文のサブワードトーク

1 愛媛大学
2 情報通信研究機構
3 同志社大学
a) deguchi@ai.cs.ehime-u.ac.jp

ン数を近づけるため，これらの複合語は “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，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．

ただし，NMT の訓練時には対訳コーパスを使うことができるが，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．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予め学習しておく．そして，翻訳時には，学習した LSTM ベースのサブワード分割器により原言語文を分割する．

WAT Asian Scientific Paper Excerpt Corpus（以下，ASPEC）[10] 英日・日英翻訳タスクと 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 [8] 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 $P(\boldsymbol{x})$ を次式により表す．

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

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

ただし，$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ 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．

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

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は次式によって得られる．

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x} \in \mathcal{S}(X)} P(\boldsymbol{x}), \tag{4}$$

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 $P(\boldsymbol{x})$ 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．

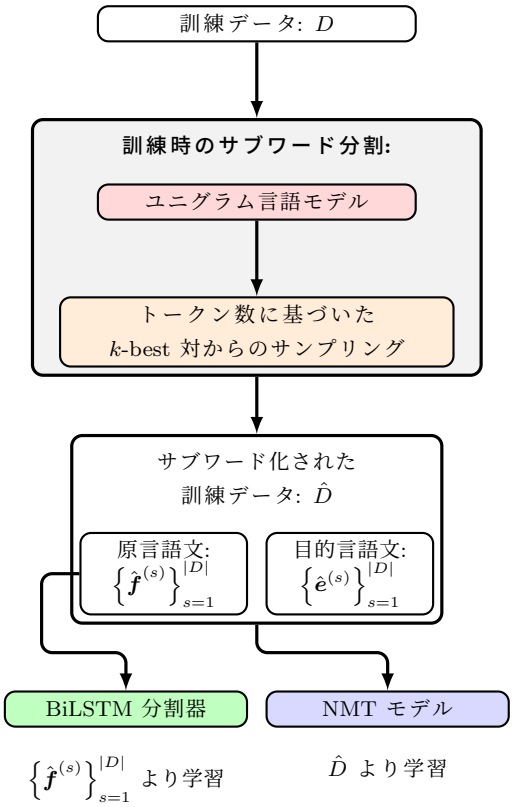

図 1 訓練時のサブワード分割

3. 提案法

本節では，対訳文からサブワードを獲得する提案手法を示す．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．ただし，NMT の訓練時には対訳データを利用できるが，翻訳時（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を獲得する．**図 1**，**図 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を分割する．

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．

Vol.2020-NL-246 No.22
Vol.2020-SLP-134 No.22
2020/12/3

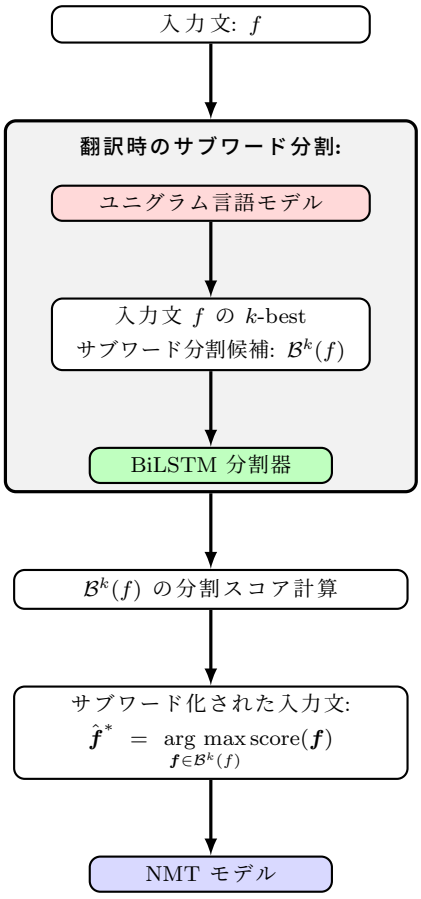

図 2 翻訳時のサブワード分割

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対を選択することで，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 $\mathcal{B}^k(f)$，$\mathcal{B}^k(e)$ の中から，対訳文対 (f, e) のサブワード列 $(\hat{\boldsymbol{f}}, \hat{\boldsymbol{e}})$ を得る．

$$(\hat{\boldsymbol{f}}, \hat{\boldsymbol{e}}) = \begin{cases} (\hat{\boldsymbol{u}}, \boldsymbol{e}^*) & \text{if } \text{len}(\boldsymbol{f}^*) < \text{len}(\boldsymbol{e}^*) \\ (\boldsymbol{f}^*, \hat{\boldsymbol{u}}) & \text{otherwise} \end{cases}, \tag{5}$$

ただし，len() はサブワードトークン数を返す関数であり，$\boldsymbol{f}^*$ と $\boldsymbol{e}^*$ 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 1-best サブワード列）である．そして，$\hat{\boldsymbol{u}}$ は，$\boldsymbol{v}^*$ を $\boldsymbol{f}^*$ と $\boldsymbol{e}^*$ のうち系列長の長い方としたとき，$\boldsymbol{v}^*$ とのトー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．

$$\hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}\in\mathcal{T}} P(\boldsymbol{u}), \tag{6}$$

$$\mathcal{T} = \arg\min_{\boldsymbol{u}\in\mathcal{B}^k} |\,\text{len}(\boldsymbol{u}) - \text{len}(\boldsymbol{v}^*)|, \tag{7}$$

$$\mathcal{B}^k = \begin{cases} \mathcal{B}^k(f) & \text{if } \text{len}(\boldsymbol{f}^*) < \text{len}(\boldsymbol{e}^*) \\ \mathcal{B}^k(e) & \text{otherwise} \end{cases}. \tag{8}$$

NMT モデルは，提案法により各対訳文をサブワード化した訓練データ $\hat{D} = \{(\hat{\boldsymbol{f}}^{(s)}, \hat{\boldsymbol{e}}^{(s)})\}_{s=1}^{|D|}$ 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 $\hat{D}$ の原言語文 $\{\hat{\boldsymbol{f}}^{(s)}\}_{s=1}^{|D|}$ 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，翻訳時の入力文 f のサブワード分割には BiLSTM 分割器を用いる．

BiLSTM 分割器は，n 個の文字からなる入力文字列 $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ の中からサブワードの境界点を識別する．BiLSTM 分割器は以下のような構造のニューラルネットワークである．

$$\mathbf{z} = \text{Embedding}(\mathbf{c}), \tag{9}$$

$$\mathbf{h} = \text{BiLSTM}(\mathbf{z}), \tag{10}$$

$$\mathbf{b} = \text{softmax}(\mathbf{h}W), \tag{11}$$

ただし，Embedding() は文字埋め込み層，$\mathbf{z}$ は文字列 $\mathbf{c}$ の d 次元埋め込み表現，BiLSTM() は BiLSTM 層，$\mathbf{h}$ は BiLSTM の隠れベクトル，softmax() は softmax 関数，$\mathbf{b}$ は BiLSTM 分割器の出力，$W \in \mathbb{R}^{d\times\{0,1\}}$ は隠れベクトル $\mathbf{h}$ 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 $\mathbf{b}_t = (b_{t,0}, b_{t,1})$ 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化された訓練データ $\hat{D}$ 中の全原言語文 $\hat{\boldsymbol{f}} \in \{\hat{\boldsymbol{f}}^{(s)}\}_{s=1}^{|D|}$ について，以下の目的関数 $\mathcal{L}_{segment}$ を最大化することにより学習される．

$$\mathcal{L}_{segment} = \sum_{t=1}^{n} \log b_{t,r_t}, \tag{12}$$

$$\text{where } r_t = \begin{cases} 0 & \text{if } c_t\text{はサブワードの開始点} \\ 1 & \text{otherwise} \end{cases}. \tag{13}$$

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 $\mathcal{B}^k(f)$ を得る．次に，各分割候補 $\boldsymbol{f} \in \mathcal{B}^k(f)$ のスコア score($\boldsymbol{f}$) 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．

$$\text{score}(\boldsymbol{f}) = \sum_{t=1}^{n} \log b_{t,r_t}. \tag{14}$$

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．

$$\hat{\boldsymbol{f}}^* = \arg\max_{\boldsymbol{f}\in\mathcal{B}^k(f)} \text{score}(\boldsymbol{f}). \tag{15}$$

以上により獲得したサブワード列 $\hat{\boldsymbol{f}}^*$ を NMT モデルに入力し，翻訳を行う．

Vol.2020-NL-246 No.22
Vol.2020-SLP-134 No.22
2020/12/3

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 [8]）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 [9]” 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 Sentencepiece[*1] を用いた．全実験において，NMT システムとして Transformer base [15] モデルを用いた．

4.1.1 データセット

翻訳性能は WAT ASPEC 英日・日英翻訳タスク[*2][10] を用いて評価した．サブワードの語彙数は，原言語と目的言語でそれぞれ独立して 16,000 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[*3]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[6] 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 [15]．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 [14] を用い，平滑化 ϵ は 0.1 に設定した．モデルはパラメータの更新 1,000 回毎に保存し，モデル性能の評価時には，訓練終了時点から前 5 つのモデルの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[16] とした．

提案法のハイパー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，$\beta_1 = 0.9$, $\beta_2 = 0.98$ 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との比較では，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 デコードを用いた．

[*1] https://github.com/google/sentencepiece
[*2] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[*3] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/baseline/dataPreparationJE.html

表 1 ASPEC における翻訳性能の比較（BLEU(%)）

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

	適合率	再現率	F 値
Ja	97.05	97.44	97.24
En	98.82	99.22	99.02

表 3 オラクル分割の翻訳性能（BLEU(%)）

	ASPEC	
	日英	英日
BiSW	29.39	43.29
オラクル分割	29.49	43.49

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 [11] で評価し，評価方法は WAT Automatic Evaluation Procedures[*4]に従った．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 [7] 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て，BiLSTM 分割器の性能を評価した．結果を**表 2** に示す．表 2 のとおり，

[*4] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html#automatic_evaluation_systems.html

表 4 BiLSTM 分割器を用いないときとの翻訳性能比較（BLEU(%)）

	ASPEC	
	日英	英日
BiSW	29.39	43.29
BiSW w/o BiLSTM	28.80	43.00

BiLSTM 分割器の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 中の “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分かるが，その差は小さい．オラクル分割との性能差は，日英，英日翻訳においてそれぞれ 0.10，0.20 BLEU ポイントに留まっており，これは表 2 に示したとおり，BiLSTM 分割器の分割性能が高く正解分割に近い分割を得られているためであると考えられる．

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化するため 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 1-best サブワード列（$\boldsymbol{f}^*$）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 1-best サブワード列（$\boldsymbol{f}^*$）で固定する．つまり，常に $\boldsymbol{v}^* = \boldsymbol{f}^*$ で固定する．これにより，訓練時の複数分割候補からの探索は常に目的言語側のみで行われ，原言語側の 1-best サブワード列のトークン数に最も近い候補が選択されるようになる．

表 4 に実験結果を示す．表中の “BiSW w/o BiLSTM” が BiLSTM を用いない場合の手法を示している．表 4 より，BiLSTM 分割器を用いないと翻訳性能が低下することが確認された．具体的には，“BiSW w/o BiLSTM” は “BiLSTM” と比較して，日英，英日翻訳においてそれぞれ 0.59，0.29 BLEU ポイント性能が低下することが分かる．この実験結果より，目的言語側のみの探索では翻訳に適した分割を獲得するのには十分でなく，原言語，目的言語それぞれの分割候補から双方向に探索する必要があると考えられる．したがって，翻訳時の入力文を分割するための BiLSTM 分割器は必要であると考えられる．

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実際の例を示す．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．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，原言語文と目的言語文間でサブワードが対応付けられやすくなり，NMT モデルの学習を支援できるようになると考えられる．

表 5 ASPEC 日英の訓練データにおけるサブワードの例

ユニグラム言語モデル	BiSW
helper	help er
basically	basic ally
focused	focus ed
popularization	popular ization
第三　者	第　三　者
骨密度	骨　密度
設計法	設計　法

表 6 ASPEC 日英の評価データにおけるサブワードの例

ユニグラム言語モデル	BiSW
密度分布	密度　分布
分散型	分散　型
透　水性	透　水　性

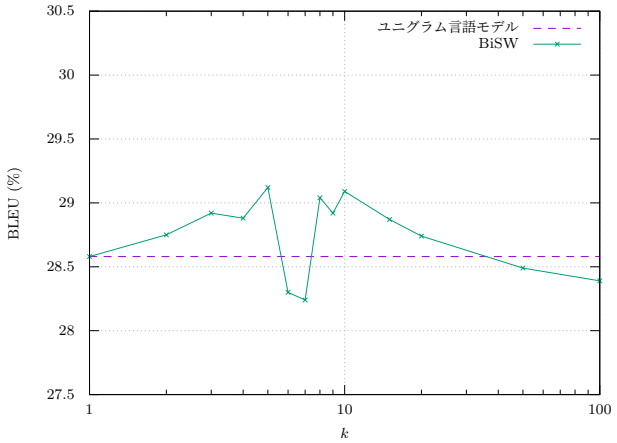

図 3 ハイパーパラメータ k に対する敏感さ（開発データにおける翻訳性能）

表 6 に APSEC 日英の評価データ（評価データの日本語文）に対して従来法と提案法をそれぞれ適用した実際の例を示す．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．

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からの探索幅 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 日英の開発データにおける k を変化させ

表 7 尤度ベースの手法との比較（BLEU(%)）

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

表 8 対訳文対間のサブワードトークン数の差の平均

	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

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

工藤 [8] 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，サブワード化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，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化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（トークン数ベース）" 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に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により，原言語文と目的言語文のトークン数の差を小さくする分割を行うという目的が達成できていることが確認できる．

表 9 WMT14 英独・独英翻訳における提案法の有効性（BLEU(%)）

	WMT14	
	英独	独英
ユニグラム言語モデル	26.45	30.62
BiSW	**26.77**	**30.64**

5.6 分かち書きされた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ち書きされている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，WMT14 英独・独英翻訳タスク[*5]を用いた．

本実験におけるサブワードの語彙数は，原言語と目的言語で辞書を共有して 37,000 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約 25,000 トークンになるよう設定した．訓練データの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 を超えるものを訓練データから除去した．ハイパーパラメータ 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 BLEU ポイント性能が改善された．実験結果より，分かち書きされた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 [13] とユニグラム言語モデル [8] 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，指定した語彙数を上限として，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

[*5] https://www.statmt.org/wmt14/translation-task.html

ナイズを必要とせず生文から直接学習できるという特長がある．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 [9] を用いた．

サブワード正則化 [8] 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ご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 [12] はサブワード正則化を用いられるように 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，$P(\boldsymbol{x}|X)$ 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 [4] 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も述べている．我々の手法は NMT モデルの入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 [3]，[2] や Huck ら [5] は言語学に基づくサブワード分割を提案している．Ataman ら [3]，[2] は教師なし形態学習に基づく “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 [5] 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 [1] 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

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

[2] Ataman, D. and Federico, M.: An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Boston, MA, Association for Machine Translation in the Americas, pp. 97–110 (online), available from ⟨https://www.aclweb.org/anthology/W18-1810⟩ (2018).

[3] Ataman, D., Negri, M., Turchi, M. and Federico, M.: Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English (2017).

[4] Cherry, C., Foster, G., Bapna, A., Firat, O. and Macherey, W.: Revisiting Character-Based Neural Machine Translation with Capacity and Compression, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 4295–4305 (2018).

[5] Huck, M., Riess, S. and Fraser, A.: Target-side Word Segmentation Strategies for Neural Machine Translation, *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 56–67 (2017).

[6] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (2014).

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

[8] Kudo, T.: Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, Association for Computational Linguistics, pp. 66–75 (2018).

[9] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, Association for Computational Linguistics, pp. 66–71 (2018).

[10] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proc. of LREC 2016*, pp. 2204–2208 (2016).

[11] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311–318 (2002).

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

[13] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (2016).
[14] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826 (2016).
[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pp. 5998–6008 (2017).
[16] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al.: Google's neural machine translation system: Bridging the gap between human and machine translation, *arXiv preprint arXiv:1609.08144* (2016).