ニューラル機械翻訳のためのバイリンガルサブワード分割の研究

Studies on Bilingual Subword Segmentation for Neural Machine Translation

愛媛大学大学院理工学研究科教授
二宮　崇

2001 年東京大学大学院理学系研究科情報科学専攻博士課程修了。博士（理学）。2017 年より愛媛大学大学院理工学研究科教授。自然言語処理の研究に従事。

奈良先端科学技術大学院大学
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術大学院大学博士後期課程に在学。自然言語処理の研究に従事。

国立研究開発法人情報通信研究機構（NICT）
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。

同志社大学理工学部情報システムデザイン学科准教授
田村　晃裕

2013 年東京工業大学大学院総合理工学研究科博士課程修了。博士（工学）。2020 年より同志社大学理工学部情報システムデザイン学科准教授。自然言語処理の研究に従事。

国立研究開発法人情報通信研究機構（NICT）フェロー
隅田　英一郎

1999 年京都大学大学院博士（工学）取得。日本アイ・ビー・エム、国際電気通信基礎技術研究所を経て、現在、情報通信研究機構。2016 年 NICT フェロー。機械翻訳の研究に従事。

1 はじめに

現在、ニューラルネットワークを用いた機械翻訳（ニューラル機械翻訳）が機械翻訳の主流となっている。注意機構に基づく LSTM（Long Short-Term Memory）[1] は初期のころから広く使用されてきたニューラル機械翻訳モデルである。このモデルは、原言語文（翻訳元言語の文）内の単語と目的言語文（翻訳先言語の文）内の単語間の関係を捉える言語間注意機構を用いることで高い精度を実現した。また、近年、トランスフォーマー（Transformer）モデル [2] が LSTM や畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を用いた手法と比べて高い精度を達成し、注目されている。トランスフォーマーモデルは、従来の言語間注意機構に加えて、同じ文中の単語間の関係を捉える自己注意機構を導入している。

これらのニューラル機械翻訳モデルは高い精度と流暢性を同時に実現するが、予め指定した語彙に基づいて学習および翻訳を行うため、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問題が知られている。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）[3]やユニグラム言語モデル[4]などによるサブワード分割が現在広く用いられている。サブワードは、一般に単語よりも短く、文字よりも長い単位であり、例えば、「pretraining」は「pre」と「train」と「ing」の３つのサブワードに分割されることが考えられる。サブワード分割は、形態素解析のように言語学的知見や規則に基づいた分割を行うのではなく、語彙量を指定した上で、コーパスから自動的にサブワード語彙を抽出し、分割することを特徴とする。語彙に「pretraining」という単語が登録されていなくても、「pre」「train」「ing」がサブワードとして語彙に登録されていれば、「pretraining」を完全な未知語として扱うのではなく、「pre」「train」「ing」をニューラル機械翻訳の入力/出力トークンとすることができ、未知語の問題を大きく緩和することができる。

BPEによるサブワード分割は事前トークナイズを必要とするのに対し、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、語彙量の上限を制約としたトークン数の最小化を行っている。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と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。

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法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」という語の翻訳の学習に寄与しない。

本稿は、国際会議「The 28th International Conference on Computational Linguistics（COLING'2020）」および国内ジャーナル「自然言語処理」において我々が提案したバイリンガルサブワード分割法[5][6]について解説する。提案法であるバイリンガルサブワード分割法は、ユニグラム言語モデル[4]に基づき、対訳コーパスを用いて行うサブワード分割手法である。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方法となっている。提案法はユニグラム言語モデルに基づくため、分かち書きされない言語にも適用可能である。

提案法では、ユ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もニューラル機械翻訳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トークン数を近づけるため、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、ニューラル機械翻訳において、「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。

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。ニューラル機械翻訳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予め学習し、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リランキングし、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）[7] 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、トランスフォーマーニューラル機械翻訳モデルの性能が最大 0.81 BLEU ポイント改善した。

2 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［4］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 $s=(s_1, s_2, \cdots, s_N)$ の生起確率 $P(s)$ を次式により表す。

$$P(s) = \prod_{i=1,\ldots,N} P(s_i)$$

$$\sum_{v \in V} P(v) = 1$$

ただし、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 $P(s_i)$ は EM アルゴリズムによって周辺尤度 L を最大化することにより推定される。

$$L = \sum_{x \in D} \log P(x) = \sum_{x \in D} \log \sum_{s \in S(x)} P(s)$$

ただし、D は対訳コーパスであり、x は D 中の原言語文または目的言語文であり、$S(x)$ は x のサブワード分割候補集合である。

x を入力文としたとき、x に対す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。

$$s^* = \mathrm{argmax}_{s \in S(x)} P(s)$$

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 $P(s)$ 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 バイリンガルサブワード分割法

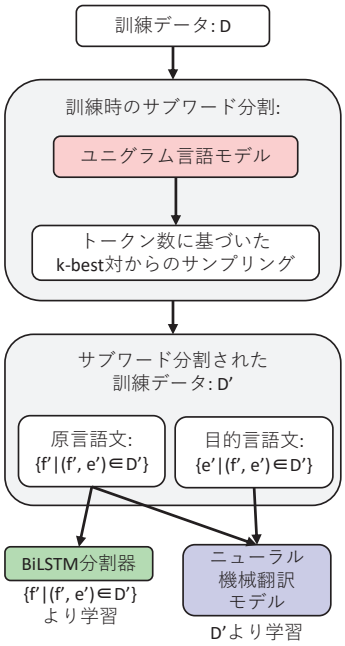

図 1　訓練時のサブワード分割

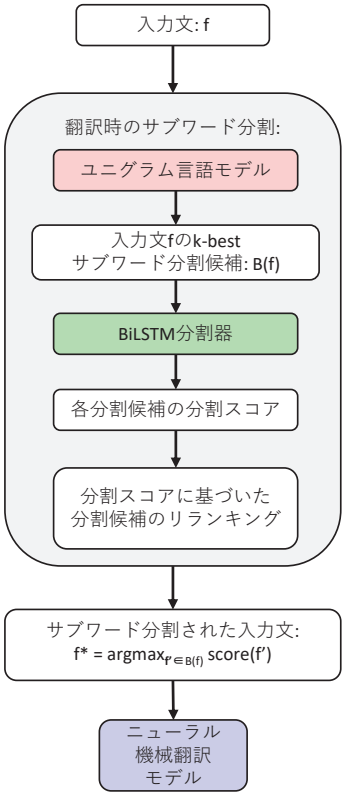

図 2　翻訳時のサブワード分割

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ニ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、ニューラル機械翻訳の訓練時には対訳コーパスを利用できるが、翻訳時（評価データ）には対訳文が存在しない。そこで、ニューラル機械翻訳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 にニューラル機械翻訳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。ニューラル機械翻訳モデルの訓練時は、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ニューラル機械翻訳モデルを学習する。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。

提案法はニューラル機械翻訳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

対訳コーパスの訓練データに対す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、対訳文対の分割を得る。

具体的には、訓練データ D 中の各原言語文 f と目的言語文 e の対に対し、次の手続きを行うことによって、サブワード分割を得る。

1. ユニグラム言語モデルにより、原言語文 f と目的言語文 e それぞれの k-best の分割候補 B(f)、B(e) を得る。
2. B(f) の最大確率のサブワード列を f*、B(e) の最大確率のサブワード列を e* とする。
3. f* と e* のトークン数（サブワード数）を比較する。

（ア）f* と e* のトークン数が同じ場合は、(f*, e*) を出力する。

（イ）f* の方が長く（トークン数が多く）、e* の方が短い（トークン数が少ない）場合は、B（e）の中で f* とトークン数が最も近い候補の中で最大確率のサブワード列 e** を探し出し、(f*, e**) を出力する。

（ウ）f* の方が短く、e* の方が長い場合は、（イ）における f* と e* を入れ替えて同じ処理を行う。B(f) の中で e* とトークン数が最も近い候補の中で最大確率のサブワード列 f** を探し出し、(f**, e*) を出力する。

トークン数が多いということはより細かいサブワードに分割されている、ということである。3. のステップにおいて、原言語文 f と目的言語文 e に対する最大確率のサブワード列 f* と e* のうち、より長い方のサブワード列を出力としてまず確定する。つまり、より細かく分割されているサブワード列を出力として確定する。短い方の最大確率のサブワード列は採用せず、分割候補の中からより細かく分割された（より長い）サブワード列を探索し、長い方のサブワード列と同じ長さのサブワード列探し出し、その中で最大確率のサブワード列を出力する。

上記の提案法によって訓練データ D の各対訳文をサブワード分割した訓練データを D'とする。ニューラル機械翻訳モデルは訓練データ 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 D'の原言語文だけを集めたデータから、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。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リ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 $c = (c_1, c_2, \cdots, c_n)$ に対して、サブワードの開始文字か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な構造のニューラルネットワークである。

Z = Embedding(c),

H = BiLSTM(Z),

B = softmax(HW)

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 の中間表現、softmax は softmax 関数、B は BiLSTM 分割器の出力、$W \in R^{d \times \{0,1\}}$ は中間表現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 $B_i = (b_{i,0}, b_{i,1})$ は文字 c_i がサブワードの開始点か ($b_{i,0}$)、開始点でないか ($b_{i,1}$) 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 D' 中の原言語文 f' ((f', e') ∈ D') について、以下の目的関数 Lsegment を最大化することにより学習される。

$$L_{segment} = \sum_{i=1,\cdots,n} \log B_{i,r(i)}$$

r(i) = 0 if c_i はサブワードの開始点

r(i) = 1 otherwise

翻訳時は次のようにして入力文 f のサブワード分割を行う。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(f) を得る。次に、各分割候補 f' ∈ B(f) のスコア 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。

$$score(f') = \sum_{i=1,\cdots,n} \log B_{i,r(i)}$$

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。

$f^* = \mathrm{argmax}_{f' \in B(f)}\ score(f')$

以上により得られたサブワード列 f* をニューラル機械翻訳モデルに入力し、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 [4]）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 [8]」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、ニューラル機械翻訳システムとして Transformer base モデル [2] を用いた。

翻訳性能は WAT ASPEC 日英・英日（以下 ASPEC 日 - 英）翻訳タスク[2] [7] 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。ニューラル機械翻訳モデル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。

全ニューラル機械翻訳モデルにおいて、パラメータ最適化には Adam [9] 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[2]。ドロップアウトの確率は 0.1 に設定した。ニューラル機械翻訳モデルの損失関数にはラベル平滑化交差エントロピー [10] を用い、平滑化 ε 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、訓練終了時点から前 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 [11] 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では、提案法と条件

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

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 ASPEC 日 - 英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 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、提案法を用いたニューラル機械翻訳モデルを示している。翻訳性能は BLEU [12] 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 [13] を実施し、有意水準は 5% とした（$p \leqq 0.05$）。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「サブワード正則化」より有意に性能が高く、提案法の有効性が確認できる。

4.3 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 2 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 2 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

表 2 ASPEC 日英翻訳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	訳文中の対応箇所
helper	help er	ヘル パー
basically	basic ally	基本的に は
focused	focus ed	に着目し た
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 3 ASPEC 日英翻訳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	訳文中の対応箇所
密度分布	密度 分布	density distribution
分散型	分散 型	dispersion type
透 水性	透 水 性	permeability

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」と対応付いて「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、ニューラル機械翻訳モデルの学習を支援できるようになると考えられる。ただし、「popularization」の訳文中の対応箇所は「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、訳文中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 3 に APSEC 日英翻訳の評価データ（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 3 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、訳文中の対応箇所が「permeability」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。

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

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。

4.4 分かち書きされた言語対及び分かち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ち書きされている言語対、及び、日中翻訳のような分かち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[5]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、ニューラル機械翻訳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。ハイパーパラメータkは開発データで調整し、2に設定した。

表4　WMT14英-独及びASPEC日-中翻訳における提案法の有効性（BLEU（%））

	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4にWMT14英独・独英翻訳、ASPEC日中・中日翻訳の実験結果を示す。表4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及び分かち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。

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

5 関連研究

BPE [3] とユニグラム言語モデル [4] 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 [8] を用いた。

サブワード正則化 [4] は複数のサブワード分割候補を用いたニューラル機械翻訳の訓練法であり、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ニューラル機械翻訳に組み込むには、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、ニューラル機械翻訳の訓練処理を修正する必要がある。

BPE-Dropout [14] 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(s|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ニューラル機械翻訳モデルも提案されている。Cherryら [15] は単語単位やサブワード単位のニューラル機械翻訳よりも文字単位のニューラル機械翻訳の翻訳

性能が高くなると報告している。ただし、Cherry らは文字単位のニューラル機械翻訳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。我々の手法はニューラル機械翻訳モデルの入出力の粒度について文字単位のニューラル機械翻訳の長所と短所（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考えられる。

Ataman ら[16]、[17] や Huck ら[18] は言語学に基づくサブワード分割を提案している。Ataman ら[16]、[17] 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[18] 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[19] 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。

6 おわりに

本稿は、我々が提案するバイリンガルサブワード分割法[5] [6] について紹介した。提案法は、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サブワード分割法である。WAT ASPEC 英日・日英・英中・中英翻訳タスクと WMT14 英独・独英翻訳タスクの実験を行い、提案法を用いることでトランスフォーマーニューラル機械翻訳モデルの性能が最大 0.81 BLEU ポイント改善されることが示された。これらの実験結果より、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たと考えている。今後は他の言語対でも提案法の有効性を確認していきたい。

謝辞

本稿は、国際会議「The 28th International Conference on Computational Linguistics（COLING'2020）」に採択された論文[5]および国内ジャーナル「自然言語処理」に採択された論文[6] に基づいて、それらの論文を再構成し、解説したものである。

これらの研究成果は、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] T. Luong, H. Pham and C. D. Manning.（2015）Effective Approaches to Attention-based Neural Machine Translation. In Proc. of EMNLP 2015, pp. 1412–1421.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin.（2017）. Attention is All you Need. Advances in Neural Information Processing Systems 30, pp. 5998–6008.

[3] R. Sennrich, B. Haddow and A. Birch.（2016）. Neural Machine Translation of Rare Words with Subword Units. In Proc. of ACL 2016, pp. 1715–1725.

[4] T. Kudo.（2018）. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Proc. of ACL 2018, pp. 66–75.

[5] H. Deguchi, M. Utiyama, A. Tamura, T. Ninomiya, E. Sumita.（2020）. Bilingual Subword Segmentation for Neural Machine Translation. In Proc. of COLING 2020, pp. 4287–4297.

[6] 出口祥之，内山将夫，田村晃裕，二宮崇，隅田英一郎．（2021）．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. 26, No. 2, pp. 632–650.

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

[8] T. Kudo and J. Richardson.（2018）. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proc. of EMNLP 2018: System Demonstrations, pp. 66–71.

[9] D. P. Kingma and J. Ba. (2015). Adam: A Method for Stochastic Optimization. In Proc. of ICLR 2015. arXiv:1412.6980.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. (2016). Rethinking the Inception Architecture for Computer Vision. In Proc. of CVPR 2016, pp. 2818–2826.

[11] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes and J. Dean. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144.

[12] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proc. of ACL 2002, pp. 311–318.

[13] P. Koehn. (2004). Statistical Significance Tests for Machine Translation Evaluation. In Proc. of EMNLP 2004, pp. 388–395.

[14] I. Provilkov, D. Emelianenko and E. Voita. (2020). BPE-Dropout: Simple and Effective Subword Regularization. In Proc. ACL 2020, pp.1882–1892.

[15] C. Cherry, G. Foster, A. Bapna, O. Firat and W. Macherey. (2018). Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In Proc. of EMNLP 2018, pp. 4295–4305.

[16] D. Ataman, M. Negri, M. Turchi and M. Federico. (2017). Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. arXiv preprint arXiv:1707.09879.

[17] D. Ataman and M. Federico. (2018b). An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. In Proc. of AMTA 2018 (Volume 1: Research Papers), pp. 97–110.

[18] M. Huck, S. Riess and A. Fraser. (2017). Target-side Word Segmentation Strategies for Neural Machine Translation. In Proc. of WMT 2017, pp. 56–67.

[19] D. Ataman and M. Federico. (2018a). Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In Proc. of ACL 2018 (Volume 2: Short Papers), pp. 305–311.