The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ニューラル機械翻訳におけるコーパスフィルタリングに関する固有表現に注目した分析

Corpus Filtering Focusing on Named Entities for Neural Machine Translation

本間広樹 Hiroki Homma　山岸駿秀 Hayahide Yamagishi　松村雪桜 Yukio Matsumura　小町守 Mamoru Komachi

首都大学東京
Tokyo Metropolitan University

Some parallel corpora include sentences that disturb learning of machine translation systems. By removing such noisy sentences like containing many out-of-vocabulary from the training corpus, it is expected to makes better translations. In this paper, we focus on the sentences containing named entities because most of the named entities fall into out-of-vocabulary due to low-frequencies. We propose two kinds of filtering methods, using byte pair encoding and using named entity recognition. By removing noisy sentences from a training corpus on Japanese-English language pair, BLEU scores improve statistically significantly by 0.5 points in both proposed methods. Analysis revealed that both our methods overcome the mistakes such as suffix of the noun, determiner, and sentence lengths.

1. はじめに

2013 年に提案されたニューラル機械翻訳 [Kalchbrenner 13] は現在，コンピュータを用いて自動的に翻訳を行う機械翻訳の主流の手法となっている．自然言語処理の分野では学習用のデータに含まれない単語を未知語と呼び，機械翻訳システムなどの言語生成システムでは一般的に，未知語を生成できないという問題がある．さらに，ニューラル機械翻訳は計算量の関係から実際に使用する語彙サイズを絞る必要があり，既知の単語であっても低頻度のものは未知語と同様に扱われてしまう．これらの，システムが扱うことのできない単語を out-of-vocabulary (OOV) と呼ぶ．

一般的な単語は，構成性のある単語と構成性のない単語の 2 種類に分類される．ここで構成性のある単語とは，語義を持つサブワードに分割することが可能な単語を指す．例えば *unkingly* という単語は構成性があり，*un*（〜でない），*king*（王），*ly*（〜らしい）の 3 つのサブワードに分割された場合，「王らしくない」という意味を推測できる可能性がある．逆に構成性のない単語は語義のあるサブワードに分割できない単語を指す．例えば *Maldives* という単語は構成性を持たず，このような単語が OOV であった場合は元の意味を復元するのは困難である．

このように，単語をさらに小さなサブワードという単位に分割することで OOV の問題を解決する手法として Byte Pair Encoding (BPE) モデル [Sennrich 16] が提案された．この手法では OOV を高頻度の細かいサブワード，最も細かいもので文字単位に分割することで，語彙サイズの制限による OOV の問題を解消する．

また，機械翻訳の手法の多くはモデルの学習のために原言語文と目的言語文のペアを集めたパラレルコーパスを必要とする．特にニューラル機械翻訳においては他の手法と比較して大規模なデータを要する．現在公開されているパラレルコーパスの多くは，複数の言語で書かれた論文や新聞記事などの中から，対応の取れる文対を機械的に集めることで対訳文としている．しかし，正しい対訳になっている文対であってもその文対の必要性について検討されていない場合も多く，これらのコーパスは翻訳モデルの学習に悪影響を及ぼす文を含有していることが考えられる．このような文を本稿では以降ノイズ文と呼ぶ．パラレルコーパスにさまざまなノイズを与えてその影響を調べた研究 [Khayrallah 18] ではニューラル機械翻訳は以前の主流な機械翻訳手法である統計的機械翻訳に比べてノイズに弱いことがわかった．ノイズ文を学習データから取り除くことで，より精度の高い機械翻訳システムが構築されることが見込まれる．

固有名詞などに代表される固有表現は一般的な語彙に比べて出現頻度が低い．このため，固有表現を多く含む文は計算量の制約で語彙サイズが制限されるニューラル機械翻訳において，多数の OOV を持つことになる．OOV の占める割合が高い文は，文としての意味が大きく欠けている場合が多く，入力文と参照文のどちらに用いても翻訳モデルの学習に対して悪影響を与えると思われる．また，前処理として BPE などのサブワード化を行う場合を考えると，出現頻度の低い固有表現はある程度の長さのある高頻度のサブワードで構成されていないことが多く，他の構成性のある単語よりも細かいサブワードに分割される．このように細かく分割された固有表現を多く含む文は翻訳結果の改善にほとんど寄与せず，逆に文長の増加や語義を持たないサブワードの増加により正しい翻訳を妨げる学習につながってしまうと考えられる．

これらのことから，固有表現を多く含む文を学習データから取り除くことで，より正しい翻訳が行われる機械翻訳システムが構築されることが見込まれる．この仮定のもと，本研究では 2 種類のコーパスフィルタリング手法を提案する．一つは固有表現抽出を用いたフィルタリング（NERF）で，文中に含まれる固有表現数が多い文を取り除く手法である．この手法は直接固有表現抽出ツールを用いる必要があるため，ツールが存在しない言語にも容易に適用できる手法として，BPE を用いたフィルタリング（BPEF）も考案した．これはサブワード化したときに文長が大きく増加する文，つまり細かいサブワードに分割される単語を多く含む文を取り除く手法である．

本研究では，これら 2 種類の手法でパラレルコーパスからノイズ文を取り除く実験とその分析を行った．日英翻訳の実験ではノイズ文を取り除くことで，どちらの手法でも機械翻訳の自動評価尺度である BLEU スコアが 0.5 ポイント改善した．

連絡先: 本間広樹，homma-hiroki@ed.tmu.ac.jp

同実験において人手評価も実施し，NERF に関してはスコアの向上が見られた．分析から，BLEU スコアを上昇させる要因として複数形の有無や冠詞の間違いなどが改善されたほか，文長がより正しく学習できていることが分かった．

2. フィルタリング手法

パラレルコーパス内に含まれるノイズ文を取り除く方法として，固有表現抽出を用いる手法と BPE を用いる手法の 2 種類を考える．

2.1 NERF: 固有表現抽出によるフィルタリング

固有表現抽出はテキストから固有表現を抽出するタスクである．現在では計算機によって大量のテキストから固有表現を自動的に抽出する研究が一般的であり，ツールとして様々なものが公開されている．固有表現抽出手法は数多く提案されているが，本研究では機械学習を用いた系列ラベリングによる固有表現抽出の結果を用いる．

固有表現抽出による学習データのフィルタリングは以下の流れで行う．

1. 固有表現抽出を用いて原言語側または目的言語側の学習データの各文で固有表現数をカウント
2. 固有表現数の多い上位 n 文の文番号を取得 *1
3. 2. で取得した文番号の文を原言語側と目的言語側の両方から削除

2.2 BPEF: BPE によるフィルタリング

Byte Pair Encoding (BPE)[Sennrich 16] は，低頻度の単語を高頻度のサブワードに分割することで機械翻訳の OOV の処理を行う手法であり，現在，ニューラル機械翻訳では一般的に使用されている．BPE の計算方法を以下に示す．

1. 単語分割された文の集合 *2 を入力
   “*Quickly, accurately, and, effectively*”
2. 文の集合に含まれる各単語に対して文字を要素としたリスト化
   $\left[Q,\ u,\ i,\ c,\ k,\ l,\ y,\ @\right]$, $\left[a,\ c,\ c,\ u,\ r,\ a,\ t,\ e,\ l,\ y,\ @\right]$,
   $\left[a,\ n,\ d,\ @\right]$, $\left[e,\ f,\ f,\ e,\ c,\ t,\ i,\ v,\ e,\ l,\ y,\ @\right]$ *3
3. リストの要素 2-gram の頻度を計算
   {*ly*: 3, *y@*: 3, *Qu*: 1, *ui*: 1, *ic*: 1, ... }
4. 最も高頻度の 2 要素を結合してリストを更新
   $\left[Q,\ u,\ i,\ c,\ k,\ \boldsymbol{ly},\ @\right]$, $\left[a,\ c,\ c,\ u,\ r,\ a,\ t,\ e,\ \boldsymbol{ly},\ @\right]$,
   $\left[a,\ n,\ d,\ @\right]$, $\left[e,\ f,\ f,\ e,\ c,\ t,\ i,\ v,\ e,\ \boldsymbol{ly},\ @\right]$
5. 語彙サイズを決定するパラメータ η 回だけ 3. と 4. を繰り返し実行
   $\left[Quick,\ ly@\right]$, $\left[accurat,\ ely@\right]$, $\left[and@\right]$, $\left[effectiv,\ ely@\right]$
6. 使用した全ての文字とサブワードを語彙として使用
   {*Q, u, i, c, k, ..., ly, ly@, ely@, ..., accurat, effectiv*}

ここで，リストの要素 2-gram とはリストに含まれる要素を隣同士で結合したものである．例えば，リスト $[a,\ b,\ c]$ の要素 2-gram は $\{ab,\ bc\}$ となる．BPE によって使用する語彙サイズが削減され，低頻度の単語は複数のサブワード *4 に分割されるようになる．BPE によってサブワード化された文の例を以下に示す．

サブワード化前（5 トークン）
HDPE における メルトフラクチャー 特性 。

サブワード化後（8 トークン）
HDPE における メルト フラ クチャ ー 特性 。

BPE の結合回数を語彙サイズの制限以下にすることで全ての単語をサブワードで表現できるため，OOV がなくなる．その一方で，構成性のない単語は高頻度の大きなサブワードに分割できず，そのような単語を多く含む文は，とても細かく分割されたサブワードの増加と文長の増加により学習に悪影響を与える文になる可能性がある．つまり，BPE の適用で文長が大きく増加した文を学習データから取り除くことで，NERF と同様により正しい翻訳が行われる機械翻訳システムが構築されることが見込まれる．この仮定のもと BPEF では BPE の適用で文長が大きく増加した文を学習データから取り除く．

BPE を用いた学習データのフィルタリングは以下の流れで行う．

1. 原言語側または目的言語側の学習データの各文に対し，BPE を適用する前のトークン数と適用した後のトークン数をそれぞれ w と t としてカウント
2. $t/w > \theta$ を満たす文の文番号を取得
3. 2. で取得した文番号の文を原言語側と目的言語側の両方から削除

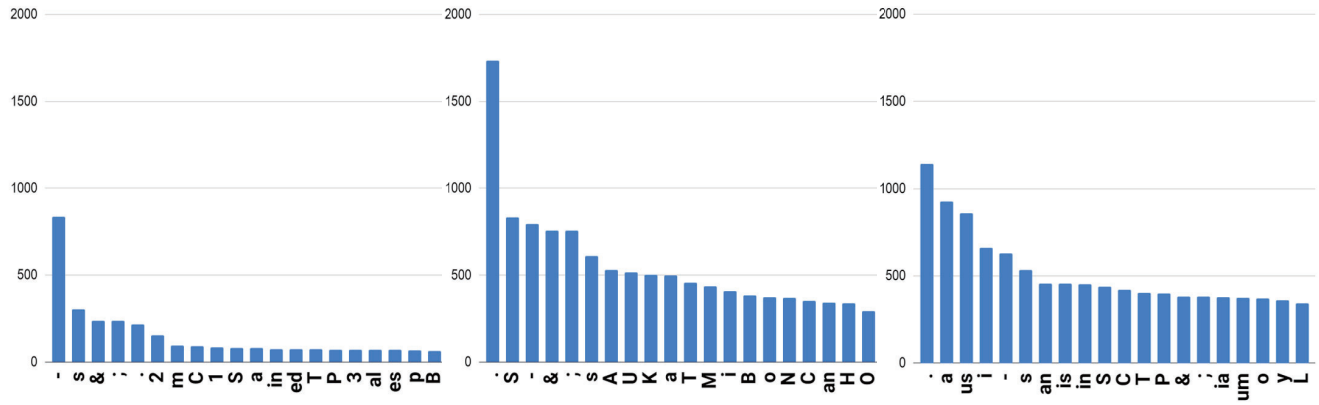
ここで θ は削除する文長比のしきい値を表す．ここでいう文長比は BPE を適用する前後の文長の増加率である．これにより，単語単位とサブワード単位で文長が大きく異なるノイズ文を削除できる．

3. 実験

3.1 実験設定

学習コーパスには科学技術論文の概要から作成された日英パラレルコーパスである Asian Scientific Paper Excerpt Corpus (ASPEC)[Nakazawa 16] を用いた．文の数は学習用，開発用，評価用それぞれ 977,367, 1,790, 1,812 文である．日本語側の単語分割には MeCab*5 (辞書: IPADic Ver. 2.7.0) を用い，英語側のトークン化には Moses*6 に付属する tokenizer.perl を用いた．さらに，OOV の処理として日本語側と英語側に BPE を適用した．これは OOV の問題を解決するための BPE の適用であるため，フィルタリングの有無に関わらず全ての実験で適用する．このとき，英語側と日本語側を合わせたコーパスに対して BPE を適用した．BPE の実装には SentencePiece*7 を使用した．また，すべての実験において BPE 化の結合回数

*1 n は削除文数を決定するパラメータ
*2 この例では 1 文からなる集合を扱う．
*3 @は単語の終わりを表すトークン
*4 ここでサブワードとは単語がサブワード化によって複数に分割されたときのそれぞれのトークンのことを指し，もとの単語まで結合されたようなサブワードは含まない
*5 http://mecab.sourceforge.net/
*6 http://statmt.org/moses/
*7 https://github.com/google/sentencepiece/



[1] ランダムな 4,030 文　[2] NERF　[3] BPEF

図 1: 削除した文中の文字 1-gram と文字 2-gram のサブワードの上位 20 件

表 1: 改善した翻訳結果の一例（置換により BPE をもとに戻した文）

例 1	入力文	感知用と出力用の 2 基のコイル, 増幅器, 及び位相シフト回路からなる, 位相シフト磁気センサーシステムを開発している。
	参照訳	here was developed a phase shift magnetic sensor system composed of two sets of coils , amplifiers , and phase shifts for sensing and output .
	ベースライン	the phase shift magnetic sensor system consists of two **coil** , **amplifier** , and phase shift circuit , which **consists of two coils** for sensing and output .
	NERF	a phase shift magnetic sensor system consists of two ***coils*** , amplifier , and phase shift circuits for sensing and output .
	BPEF	a phase shift magnetic sensor system consists of two ***coils*** , ***amplifiers*** and phase shift circuits for sensing and output .
例 2	入力文	その結果, 電気光学特性として,V10=11.8V,V90=18V を得た
	参照訳	the electro - optical property obtained was V10 = 11.8V and V90 = 18V .
	ベースライン	as a result , we obtained V10 = 11.8 V , V90 = 18V as **an** electrooptic property .
	NERF	as a result , *the* electrical optical properties were obtained with V10 = 11.8 V and V90 = 18V .
	BPEF	as a result , V10 = 11.8V , V90 = 18V were obtained as *the* electrical optical properties .

は $\eta = 32,000$ とした．各フィルタリング手法において学習データ内の語彙サイズと文数は表 2 のようになった．ここで，削除文数のパラメータは BPEF に合わせて $n = 4,030$ としている．

ベースラインは，エンコーダデコーダモデルにアテンション機構を付加したモデル [Luong 15] をフィルタリング前のデータで学習させたものである．学習は 20 epoch 繰り返し，その中で開発データで測った BLEU スコアが最も高かったものを用いて評価した．両言語の単語ベクトルには学習用データを用いて word2vec[*8] を学習させて得た分散表現を用いた．モデルの各種パラメータは，埋め込み層の次元数及び隠れ層の次元数が 512，レイヤー数が 2 とした．また，最適化手法には AdaGrad を用いて初期学習率は 0.01 とし，ドロップアウト率 0.2 でドロップアウトを行う設定とした．

BPEF では，対象を原言語側とし，サブワード化と同様に語彙サイズを決定するパラメータを $\eta = 32,000$ とした BPE を用いた．削除文数を決定するパラメータ θ の値は，$\theta = 1.2$ としたときと $\theta = 1.5$ としたときの結果を比較する予備実験 [*9] を行い，$\theta = 1.5$ がより有効であったため，本実験ではこのパラメータを用いる．NERF では，対象を目的言語側とし，固有表現抽出の実装に Stanford の CoreNLP（Ver. 3.9.2）[*10] を使用した．固有表現の数は原言語側と目的言語側とで等しいと考えられるため，どちらを対象としても問題ないとみなした．

各手法の評価には自動評価と人手評価の双方を用いた．自動評価は，BLEU スコアと METEOR スコア及び参照訳の文長との二乗誤差によって行った．BLEU スコアは n-gram 適合率に基づいて翻訳の精度を評価する評価尺度で，翻訳の流暢性などを表し，METEOR スコアは類義語，語幹正規化，並べ替え情報などを考慮した評価尺度で，翻訳の妥当性などを表す．また，BLEU スコアと METEOR スコアは機械翻訳の仮説検定ツール [*11] を用いて算出し，ベースラインに対する統計的有意差の有無も合わせて調べた．人手評価は 2 人の評価者によって行い，各手法の出力結果からランダムに抽出した 100 文に対し，流暢性と妥当性をそれぞれ 1 から 3 の 3 段階で評価し，その算術平均を用いた．数字が大きいほど良い評価である．

3.2 実験結果

表 3 に実験結果を載せる．BLEU スコア及び METEOR スコアにおける「*」は有意水準 0.05 で有意差があることを示す．NERF と BPEF の両方でベースラインより BLEU スコアが 0.5 程度改善した．また，METEOR スコアが NERF で 0.1

*8 https://radimrehurek.com/gensim/models/word2vec.html
*9 本実験と異なり，ドロップアウトを行わず，レイヤー数を 1 とした．また，BPE 化の語彙サイズは $\eta = 32,000$ 及び $\eta = 64,000$ とし，英語側と日本語側で合わせず別々に行った．その他の設定は本実験と同様である．
*10 https://stanfordnlp.github.io/CoreNLP/
*11 https://github.com/jhclark/multeval

表 2: 各手法で使用する学習データの語彙サイズと文数

フィルタリング手法	語彙サイズ		文数
	原言語	目的言語	
ベースライン	26,578	19,755	977,367
NERF	26,526	19,754	973,337
BPEF	26,453	19,753	973,337

表 3: 各手法の自動評価及び人手評価

手法	BLEU	METEOR	流暢性	妥当性
ベースライン	23.3	29.7	2.49	2.31
NERF	*23.8	29.8	2.54	2.35
BPEF	*23.8	*30.0	2.48	2.30

程度，BPEF で 0.3 程度改善した．人手評価では，流暢性と妥当性の双方において BPEF はベースラインとほぼ変わらず，NERF は 0.05 程度改善した．フィルタリングにより翻訳結果を向上させることができている．さらに，各手法における出力の参照文に対する文長の二乗誤差は，ベースライン，NERF，BPEF それぞれ，33.9，30.9，31.5 であった．

4. 分析と考察

4.1 削除した文の分析

削除した文中の文字 1-gram と文字 2-gram のサブワードの上位 20 件を図 1 に示す．NERF と BPEF の特徴を見るために学習データから抽出したランダムな 4,030 文と合わせて比較する．固有表現抽出と BPE によるフィルタリングでフィルタリングされたサブワードの数に注目すると，ランダムな 4,030 文に比べて大幅に多いことがわかる．これは，削除した文の中には BPE を適用したときに 1 文字か 2 文字の細かいサブワードが出るような細かい分割が多く行われていることによる．どちらにおいても “.” が多いのは，*the U.S.A.* など，固有表現に “.” を含む固有名詞が多いからであると考えられる．

4.2 出力結果の考察

表 1 に出力例を載せる．例 1 のベースラインの出力を見ると，本来複数形として出力するべき単語を単数形として出力している翻訳ミスや，同じフレーズを繰り返し出力している翻訳ミスが見受けられる．それに対して NERF や BPEF ではこれらの箇所に改善が見られた．また，例 2 のベースラインの出力を見ると，本来 *the* と出力するべき部分を *an* と誤って出力してしまっているが，どちらの提案手法でも正しく *the* と出力できていることがわかる．

翻訳改善の理由として，複数形や冠詞以外の意味を持つ，固有表現がサブワード化されて出現した *s* や *an* を含む文を一部取り除くことで，*s* や *an* の複数形や冠詞としての意味をモデルがより正しく学習できたことが考えられる．今回の実験結果では BLEU も METEOR も向上しているが，METEOR の改善が小さいことがわかる．METEOR スコアは妥当性を測ることのできる評価尺度であるため，妥当性より流暢性のほうが改善していると考えられる．人手評価でも NERF において流暢性のほうがわずかに上がり幅が大きいことが確認できる．妥当性が低い出力結果を見ると，*strut-screw* を *but screw* や *snut screw* と誤って出力している例があった．これは *strut* が *str* と *ut* に分かれ，*str* のほうだけうまく翻訳ができていなかったものであり，言語による BPE のかかり方の違いも関わってくると考えられる．

また，参照文との文長の二乗誤差を見ると，どちらの提案手法でもベースラインに比べて小さくなっており，出力する文長が参照文に近づいていることがわかる．このことから，固有表現を多く含む文，つまり BPE をかけたときにトークン数が増加する文を取り除いたことで，翻訳モデルの出力文長がより正しく学習できたことが考えられる．

5. 関連研究

これまでにも機械翻訳におけるパラレルコーパスのフィルタリングに関する研究が行われている．例えば外れ値検出アルゴリズムを用いてパラレルコーパスをフィルタリングする研究 [Kaveh 11] や，文対の意味の違いを識別することに焦点を当てたフィルタリングの研究 [Carpuat 17] がある．また，パラレルコーパスに様々なノイズを与えてその影響を調べる研究 [Khayrallah 18] も行われている．Khayrallah らはニューラル機械翻訳が統計的機械翻訳に比べてノイズによる悪影響を受けやすいことを示している．

6. おわりに

本論文では，ニューラル機械翻訳におけるコーパスフィルタリングに関する分析について報告した．固有表現抽出を用いる方法と BPE を用いる方法の 2 種類で学習コーパスのフィルタリングを行った．フィルタリングを行わない場合とともに日英翻訳の実験を行ったところ，BLEU スコアの 0.5 ポイントの上昇を確認した．この理由を分析したところ，複数形の有無や冠詞の間違い，誤った文長で出力される問題などが一部改善されたことが分かった．今後は異なる言語対での実験を行い，言語ごとの有効性の違いを検証していきたい．

参考文献

[Carpuat 17] Carpuat, M., Vyas, Y., and Niu, X.: Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation, in *WNMT* (2017)

[Kalchbrenner 13] Kalchbrenner, N. and Blunsom, P.: Recurrent Continuous Translation Models, in *EMNLP* (2013)

[Kaveh 11] Kaveh, T., Shahram, K., and Jia, X.: Parallel corpus refinement as an outlier detection algorithm, in *MT Summit* (2011)

[Khayrallah 18] Khayrallah, H. and Koehn, P.: On the Impact of Various Types of Noise on Neural Machine Translation, in *WNMT* (2018)

[Luong 15] Luong, T., Pham, H., and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, in *EMNLP* (2015)

[Nakazawa 16] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, in *LREC* (2016)

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, in *ACL* (2016)