訓練データ中の頻度バイアスを解消する前処理の提案

神戸隆志 [1]　鈴木潤 [1,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
takashi.kambe.r8@dc.tohoku.ac.jp
{jun.suzuki,kentaro.inui}@tohoku.ac.jp

概要

訓練データ中の単語頻度の偏りは，深層ニューラルネットワークに基づく自然言語処理の性能に悪影響を及ぼすことが知られている．そのため，頻度の偏りによる影響を解消しようとする試みが報告されている．本研究では，その取り組みの一つとして，文の前処理の段階で高頻度語の頻度を減らし，訓練データ中の頻度分布を均一に近づけ，頻度バイアスを解消する手法を提案する．本稿では，提案法の有効性を機械翻訳タスクにて検証する．具体的には，提案法により前処理を施した訓練データを用いて構築した翻訳モデルと，従来通りの訓練データで構築した翻訳モデルの翻訳性能を比較する．

1　はじめに

深層ニューラルネットワークに基づく自然言語処理では，文を単語などの何かしらの基準で分割し，その分割単位に基づいて処理を行う．近年は，文を分割する単位として，サブワードを用いることが主流である．本稿では，議論を簡単にするために，単語やサブワードなど，どのような基準で分割されたかによらず分割後の各要素をトークンと表記する．いずれの場合も文を構成するトークンは少数の高頻度トークンと多数と低頻度トークンから構成されるという特徴がある．これは，サブワードなどを用いたとしても，程度の差はあれどいわゆる Zipf の法則に概ね従う分布になるためである．このトークン頻度の偏り（以下，頻度バイアスと表記）が自然言語処理のモデルに影響を与え，様々なタスクの精度が低下する現象が報告されている [1]．また，この頻度バイアスの悪影響を緩和することで自然言語処理タスクの性能向上を目指す研究が報告されている [1, 2, 3]．

本研究は，これらの先行研究と同様に，頻度バイアスを緩和する方法を考案することを目的とする．ただし，従来の方法とは違い，利便性が高く扱いやすい方法の確立を目指す．その観点から，文の前処理の段階で高頻度トークンの頻度を減らし，訓練データ中のトークン頻度分布をできるだけ均一に近づける方法を提案する．より具体的には，訓練データ中の高頻度トークンを複数の新しいトークンに分類し，高頻度トークンの出現頻度を均す操作を行う．前処理により偏りを解消することができれば，モデルの訓練方法の変更などを必要としないため，基本的にあらゆる自然言語処理タスクで利用可能という利点となる．

本研究では，機械翻訳のデータを用いた実験により提案法の有効性を検証する．具体的には，提案する前処理を用いて頻度バイアスを軽減した場合に，従来法に相当する頻度バイアスの軽減をしない場合と比較して性能が向上するかを実験的に確かめる．

2　関連研究

2.1　サブワード分割

サブワードは単語よりも小さな単位で文を表現する単位であり，近年の自然言語処理において広く用いられている単位である．サブワードを使用する利点として，語彙数を自由に設計して固定することができる点や，未知単語をサブワードの組み合わせとして表現できるといった点が挙げられる．自然言語処理におけるサブワードの構築方法として Byte Pair Encoding (BPE) [4] や Unigram Language Model [5] などが挙げられる．しかし，いずれの方法においても，高頻度単語はそのままサブワードとして語彙に含められ，低頻度単語はより小さなサブワードの組み合わせとして表現されやすいという傾向がある．この傾向から，サブワード単位の語彙も単語単位の語彙と同様に，高頻度サブワードと低頻度サブワードに偏る．

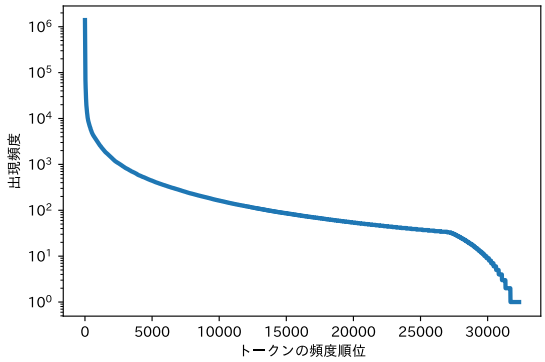


図 1 4.1 節で使用する英文データに対して BPE を 32,000 回適用した場合のトークン頻度分布

2.2 頻度バイアス

近年の自然言語処理は，単語やサブワードなどのトークンを埋め込み表現と呼ばれるベクトルとして表現し，そのベクトルを処理していく手法が一般的である．埋め込み表現の特徴として，似た意味の単語には近いベクトルが割り当てられることが期待されるが，前述の頻度バイアスにより意味ではなく頻度を反映するような埋め込み表現が構築されてしまうことが報告されている [1, 2]．具体的には，高頻度語同士が近くなるように，また低頻度語同士が近くなるような埋め込み表現が学習され，たとえ意味の近い単語であっても頻度の違いにより埋め込み表現が遠ざかる現象が起こってしまう．この埋め込み表現における現象は様々な自然言語処理のタスクの精度を低下させることが報告されており，頻度バイアスを解消しようとする試みが研究されている [1, 2, 3]．

3 頻度バイアスを軽減する前処理

2.2 節の頻度バイアスによる問題は，訓練データに出現するトークンの頻度が偏っていることに起因する問題である．そこで本研究では，図 1 の様な訓練データ中の偏ったトークン頻度分布に対して，できる限り均一な頻度分布に近づける前処理を実施した上でモデルを学習する方法を考案する．提案する前処理では，高頻度トークンを複数のトークンに分類して均一な頻度分布に近づける．提案法は，高頻度トークンを一定のルールに従って分類し頻度を均一に近づけることから**分類ルール**と呼ぶ．以下，分類ルールの詳細を説明する．

3.1 基本的な着想

英語における高頻度トークンである “the” を例にすると，分類ルールでは訓練データ中の “the” を “the0”, “the1”, “the2”,... といった複数のトークンに置き換えることで “the” の頻度を減少させる．しかし無作為に訓練データ中の “the” を複数のトークンに分割すると，全く同じ文に対するトークン列でも異なるものになる可能性がある．よって，何かしらの基準に基づいて分類を行う必要がある．そこで本研究では，分類するトークンを中心とする trigram を基準に分類する．まず，訓練データ中の最も頻度が大きいトークンを取得し，そのトークンを中心とする trigram を全て列挙する．次に列挙した trigram の中で最も頻度が大きい trigram の中心の高頻度トークンをその trigram に対応するトークンとして分類する．例えば “on the other” という trigram が “the” を中心とする最も高頻度な trigram だった場合，訓練データ中の “on the other” として出現する “the” は “the0” に置き換え，“the” とは異なるトークンとして扱う．ここで，分類ルールは頻度分布をできるだけ均一に近づけることを目的としているため，極端な低頻度 trigram に対応した低頻度トークンを生成してしまうことは望ましくない．そこで，出現頻度に閾値を設定し，新たに分類したトークンがその閾値を下回るような分類は行わないこととする．また，複数の trigram に対して 1 回の分類を行うことも可能とし，例えば “on the basis”, “in the case” として出現する “the” を “the1” として分類するといった操作を許容し，新たに生成されるトークンの頻度が閾値を下回らないようにする．図 2 に分類ルールによって高頻度トークンが複数のトークンに分類される例を示す．

3.2 アルゴリズム

以上の分類の考え方を踏まえ，具体的な処理手順は以下の通りである．

1. 出現頻度の大きいトークンの出現頻度を減らし頻度分布を均一に近づけるため，訓練データ中の最も高頻度なトークンを取得する．
2. 1 で取得したトークンを中心とする訓練データ中の trigram を列挙し，trigram の出現頻度が大きい順にソートする．
3. 頻度の大きい trigram から順に頻度を足し合わせていき，設定した閾値を超えた時点で，それ

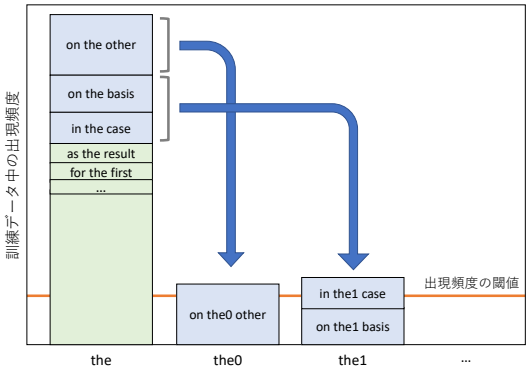


図 2　英語における高頻度トークンである “the” を分類する様子: 一回目の分類では，“the” を中心とする trigram のうち最も高頻度である “on the other” の “the” を “the0” として分類する．二回目の分類では “on the basis” の “the” を “the1” として分類しようと試みるが，その場合 “the1” の出現頻度が閾値以下になってしまうため，次点で高頻度な “in the case” も使用して分類することで “the1” の出現頻度が閾値以上になる様にする．

らの trigram の中心に出現するトークンは新たなトークンとして分類する．

4. 1-3 を目的の語彙数になるまで繰り返す．

上記の 1-3 の手順によって新たに分類されたトークンが 1 種類増えるため，分類ルールを 1 回適用することで語彙数が 1 だけ増える．従って，分類ルール適用後の訓練データの語彙数は，分類ルール適用前の訓練データ中の語彙数と分類ルールの適用回数を足し合わせた値となる．分類ルールの適用回数は自由に設定できるため，目的とする語彙数に調整することが可能である．

4　実験

本研究では，機械翻訳タスクを用いて分類ルールの効果の検証する．ベースラインとして BPE [4] でサブワード分割を行った訓練データを用意し，そのデータに対して分類ルールを適用する場合としない場合の 2 種類の訓練データを作成する．それらのデータを用いて翻訳の学習を行い，翻訳精度を比較することで分類ルールの効果を検証する．

4.1　実験設定

データ：　WMT2022[1] の英日翻訳データに対して参照文不要の COMET [6] を適用し，計算されたスコアの上位 1,500,000 文対を訓練データとして使用した．評価データは，WMT2020 および WMT2021 のニュース翻訳タスクの評価データを使用した．以下，WMT2020 の評価データを `newstest20`，WMT2021 の評価データを `newstest21` と表記する．

トークン分割方法：　サブワード構築のための BPE のマージ回数は，16,000 回または 32,000 回とする．本稿では，マージ回数 16,000 の BPE を $\texttt{BPE}(m = 16k)$，マージ回数 32,000 の BPE を $\texttt{BPE}(m = 32k)$ と表記する．分類ルールは，BPE を 16,000 回適用した入力側のデータに対して 8,000 回または 16,000 回を適用した．また分類ルールにおける最低頻度の閾値は 800 とした．これによって，出現頻度が 800 より小さくなるトークンは新たに生成されないことに注意されたい．本稿では，BPE を 16,000 回適用した後に分類ルールを 8,000 回適用したものを $\texttt{BPE}(m = 16k)\texttt{+CRULE}(c = 8k)$，BPE を 16,000 回適用した後に分類ルールを 16,000 回適用したものを $\texttt{BPE}(m = 16k)\texttt{+CRULE}(c = 16k)$ と表記する．

翻訳モデル：　翻訳モデルには，現在最もよく用いられるモデルである Transformer [7] を採用した．また，その実装は，fairseq [8] を用いた．

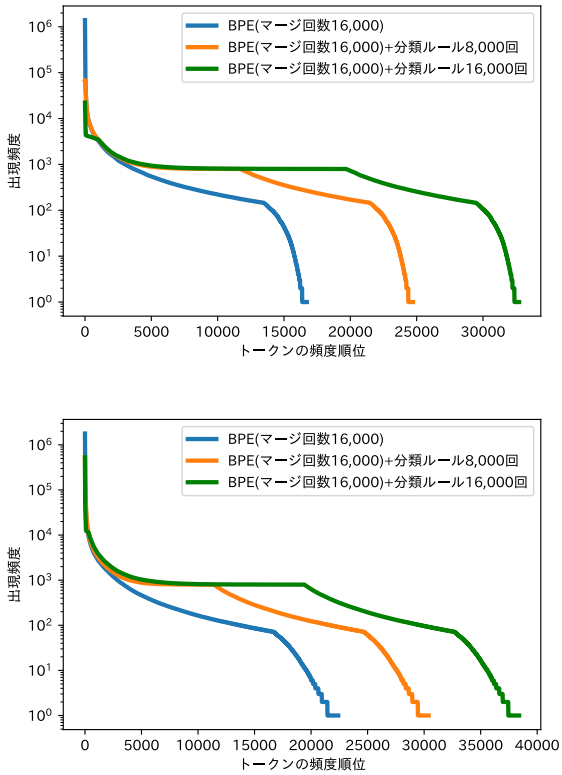
評価指標：　翻訳結果の評価には，機械翻訳の評価指標として広く用いられる BLEU スコアを用いた．BLEU スコアの計算には，sacreBLUE [9] を使用した．

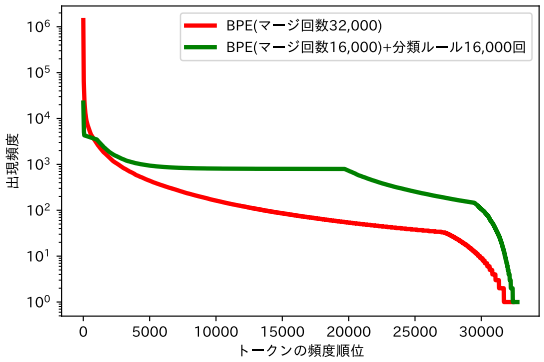
4.2　分類ルールによる頻度分布の変化

分類ルールによる頻度分布の変化を図 3 に示す．ここで，縦軸は対数スケールの出現頻度であることに注意されたい．図 3 から，分類ルールによって高頻度トークンが複数の低頻度トークンに分割されることで，より均一な頻度分布に近づいているいることが確認できる．ただし，分類ルール適用後も高頻度のまま残ってしまうトークンが存在する．これは，分類に使用した trigram が既に高頻度である場合に，現状の trigram に基づく分類ルールではこれ以上分割して頻度を減らすことができないためである．

図 4 は，BPE を 32,000 回適用した訓練データのトークン頻度分布と，BPE を 16,000 回適用し，更に分類ルールを 16,000 回適用した訓練データのトークン頻度分布を比較したものである．分類ルールを適用することで，BPE を継続して適用するより均一な分布に近くなっていることが確認できた．

1）　https://www.statmt.org/wmt22/translation-task.html





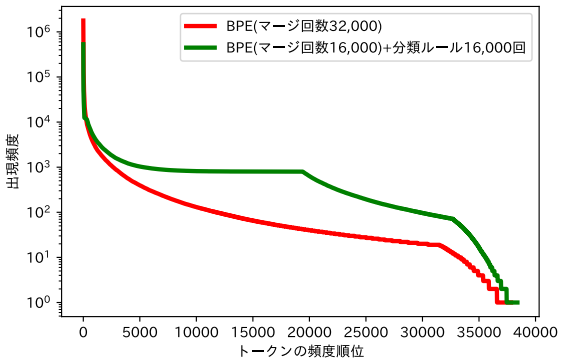


図 3　分類ルール適用によるトークン頻度分布の変化 (上：英語，下：日本語)

図 4　BPE と分類ルールで語彙数を統一した場合の比較 (上：英語，下：日本語)

表 1　日英翻訳における BLEU スコア

	newstest20	newstest21
BPE($m = 16k$)	**17.2**	17.6
BPE($m = 32k$)	16.9	17.8
BPE($m = 16k$)+CRULE($c = 8k$)	16.8	**18.0**
BPE($m = 16k$)+CRULE($c = 16k$)	16.5	17.2

表 2　英日翻訳における BLEU スコア

	newstest20	newstest21
BPE($m = 16k$)	**18.0**	**19.5**
BPE($m = 32k$)	**18.0**	19.0
BPE($m = 16k$)+CRULE($c = 8k$)	17.8	18.9
BPE($m = 16k$)+CRULE($c = 16k$)	**18.0**	18.3

4.3　機械翻訳の翻訳品質

表 1 に日英翻訳における BLEU スコアを示す．同様に，表 2 に英日翻訳における BLEU スコアを示す．日英翻訳では，newstest20 においては BPE(マージ回数 16,000) の設定が最も優れており，newstest21 においては分類ルールを 8,000 回適用した設定が最も優れている結果となった．英日翻訳では，newstest20 においては，BPE(マージ回数 16,000)，BPE(マージ回数 32,000)，分類ルール 16,000 回の設定が最も優れており，newstest21 においては BPE(マージ回数 16,000) の設定が最も優れていた．しかし，いずれの設定においても，最も優れている設定とそれ以外の設定における BLEU スコアの差は小さい．このことから，提案法である分類ルールを適用することによって普遍的に翻訳精度が向上するという結果は得られなかった．

5　おわりに

本研究では，前処理の段階で高頻度トークンの頻度を均すことで頻度バイアスを解消する手法を提案した．実験では，分類ルールによって訓練データ中のトークン頻度分布が均一に近づくことが確認できた．また，機械翻訳タスクにおける一部の評価データにおいて提案法が優れていることが確認できたが，提案法による普遍的な翻訳精度の向上を確認することはできなかった．今後の展望として，分類ルールを用いて学習した翻訳モデルの埋め込み層の観察など，翻訳精度以外の観点からも分類ルールの効果を検証する必要があると考えている．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである．

参考文献

[1] Chengyue Gong, Tao Qin, Di He, Liwei Wang, Xu Tan, and Tie Yan Liu. Frage: Frequency-agnostic word representation. **Advances in Neural Information Processing Systems**, Vol. 2018-Decem, pp. 1334–1345, 2018.

[2] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie Yan Liu. Representation degeneration problem in training natural language generation models. In **7th International Conference on Learning Representations, ICLR 2019**, 2019.

[3] Sangwon Yu, Jongyoon Song, Heeseung Kim, Seong-Min Lee, Woo-Jong Ryu, and Sungroh Yoon. Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings. In **ArXiv**, pp. 29–45, 2022.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers**, Vol. 3, pp. 1715–1725. Association for Computational Linguistics (ACL), 2016.

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

[6] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. **EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference**, pp. 2685–2702, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 2017-Decem, pp. 5999–6009, 2017.

[8] 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. **NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Demonstrations Session**, pp. 48–53, 2019.

[9] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference**, Vol. 1, pp. 186–191. Association for Computational Linguistics (ACL), 2018.