非構成性を指向したサブワード語彙獲得手法と機械翻訳への応用

田村 鴻希[1,a)]　吉永 直樹[2,b)]

概要：深層学習に基づく言語モデルでは，基本的にテキストをサブワードにトークン化して言語理解や生成を行うが，この過程ではモデルが単語，句，文といった言語表現の意味を構成的に計算できることが暗黙のうちに仮定されている．一方で，複単語表現や固有名詞のような非構成的な言語表現は構成的に意味を捉えられないため，これらをモデルが扱うためにはパラメタに知識として保持する必要がある．本研究では，これらの非構成的な言語表現を独立した語彙とすることで構成的な計算の影響を受けないようにすべく，その手段として，サブワード語彙の獲得手法のうちバイト対符号化 (BPE) に焦点を当て，頻度に加えて語彙の非構成性を考慮して語彙獲得を行う手法を実装する．実験では，ASPEC 日英対訳コーパスの各文について同手法によるトークンを利用して分割し，学習した機械翻訳モデルを通して，提案手法の有効性を検証する．

1. はじめに

Transformer [1] をはじめとしたニューラル言語モデルでは，基本的に入力テキストを部分文字列であるトークンの系列に分割し，それぞれのトークンに埋め込みを与えてから処理を行う．特に，文字列の頻度や尤度といった特徴からトークン化用の語彙を教師なし学習で獲得するサブワード学習アルゴリズムは，手法が特定の言語に依存せず，語彙の教師データを必要としないこと，また高頻度の系列をまとめ上げることで系列長を短縮し処理速度を高め，さらにモデルの性能に寄与する [2] ことから多くの言語モデルで用いられている．

サブワードに基づくトークン化は，サブワードから構成的に単語や句の意味が計算できることを暗黙に仮定しているため，イディオムのような複単語表現や固有名詞など，非構成的な言語表現の処理に課題がある [3]．大規模言語モデルでは，これらの非構成的な表現については，モデルの内部パラメタを消費して知識として計算する可能性が示唆されているが [4]，構成するサブワードの意味を学習する際にノイズとなりうること，またモデルの解釈性の観点でも独立したトークンとして扱うことが望ましいと考えられる．しかしながら，既存のサブワード語彙の獲得手法は，

[1] 東京大学
The University of Tokyo
[2] 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo
[a)] tamura-k@tkl.iis.u-tokyo.ac.jp
[b)] ynaga@iis.u-tokyo.ac.jp

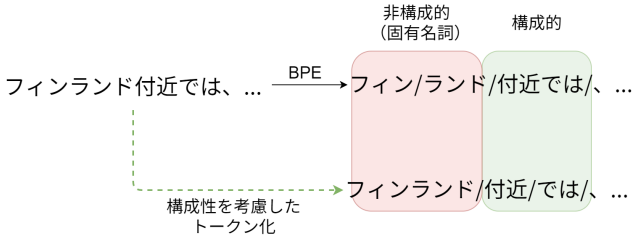

図 1 BPE アルゴリズムは頻度のみに基づいて語彙を定めるため，意味として独立な表現であるか（非構成的であるか）を考慮していない

意味の（非）構成性を陽に考慮せず，部分文字列の頻度に基づき語彙の獲得を行うため，単純にモデルの語彙サイズを大きくすると，高頻度で構成的な語彙を多く含むことになる．

そこで，本研究では，サブワードトークンの獲得手法のうちバイト対符号化 (BPE) に焦点を当て，頻度に加えて語彙の非構成性を考慮して語彙獲得を行う手法を検討する．具体的には，学習した BPE 語彙についてそれぞれ語彙の構成性を求め，構成性が高い語彙を頻度が低い順に目的の語彙サイズになるまで削除する．語彙の構成性については，トークン自身，およびその構成要素の埋め込みを用いる既存研究 [5] を利用する．

実験では，本手法によって獲得した語彙を用いたときの後段タスクの性能を測るため，本手法による語彙，および素の BPE による語彙を利用して，ASPEC 日英対訳コーパス [6] を用いてそれぞれ英日機械翻訳モデルを学習し，そ

© 1959 Information Processing Society of Japan

の翻訳精度を評価した．その結果，本研究で提案した，構成性を考慮したサブワード語彙を用いたとき，素の BPE と比べて翻訳精度が向上することを確認した．

本研究の貢献は，

- サブワード上でイディオム等の表現をよく扱うための指標として構成性を導入したこと
- 構成性の算出手法がサブワードアルゴリズムの BPE に適用可能であることを示し，また BPE 語彙へのフィルタリングという形で構成性を考慮したサブワードの獲得手法を提案したこと
- 機械翻訳を通して，同指標を用いて得られた語彙を利用することの影響を検証したこと

が挙げられる．

2. 提案手法

サブワード学習アルゴリズムの BPE [7], [8] は，学習データに含まれる文字の集合を初期語彙として，学習データにおいて最も高頻度の 2-gram を連結した文字列を語彙集合に追加し，この 2-gram を一つの単位として新しい最頻 2-gram の計算と語彙集合を逐次的に目的の語彙サイズまで繰り返すことで，学習データから語彙を獲得する．これにより，単語を語彙集合とする場合と比べて，未知語の問題を低減しつつ，小さい語彙サイズで系列長の観点で効率的なトークン化を行うことができる．その一方で，テキスト中にはイディオムや固有表現などの非構成的な複単語表現もまた存在し，モデルはこれらの表現の処理が困難である．そのため，本研究では，非構成的な表現を選択的に語彙に取り込むため，既存手法 [5] を用いて語彙の候補がそれぞれどの程度構成的かを定量化し，そのスコアを用いて語彙のフィルタリングを行う．

2.1 構成性の計算

トークンが構成的であるかのフィルタリングにあたって，それぞれのトークンについて，そのトークンが構成要素の意味と類似するかの度合いである構成性を算出する．BPE で学習した語彙 V では，それぞれのトークン $v \in V$ は 2-gram を併合することによって構成されるため，最小単位以外のトークンはいずれも構成元のトークンを 2 つずつ持つ．そのため，それぞれのトークンについて，併合後のトークン自身 $v_{1,2}$ および併合元のトークン v_1, v_2 の埋め込みを用いて類似度として構成性を算出する．具体的な算出手法として，Cordeiro ら [5] の名詞複合語に対する構成性の算出手法を BPE によるサブワードへと転用した．この手法では，計算に用いるそれぞれのトークンについて埋め込み $\mathbf{E}(v_{1,2}), \mathbf{E}(v_1), \mathbf{E}(v_2)$ を求め，構成元トークンの埋め込みの加重平均と併合後トークンの埋め込みのコサイン類似度を測る．

$$\mathrm{cp}(v_{1,2}) = \cos\left(\mathbf{E}(v_{1,2}), \beta\frac{\mathbf{E}(v_1)}{||\mathbf{E}(v_1)||} + (1-\beta)\frac{\mathbf{E}(v_2)}{||\mathbf{E}(v_2)||}\right) \tag{1}$$

なお，Cordeiro らの実験結果に従い，埋め込み手法およびハイパーパラメータ β はそれぞれ英語データにおいて人手アノテーションとの相関係数が高くなる設定である word2vec (skip-gram) および 0.5 に設定する．そのため，以下では上記の式は

$$\mathrm{cp}(v_{1,2}) = \cos\left(\mathbf{E}(v_{1,2}), 0.5\left(\frac{\mathbf{E}(v_1)}{||\mathbf{E}(v_1)||} + \frac{\mathbf{E}(v_2)}{||\mathbf{E}(v_2)||}\right)\right) \tag{2}$$

$$\mathbf{E}(v) = \mathrm{word2vec}(v, \text{skip-gram}) \tag{3}$$

として扱う．

2.2 語彙の獲得手順

最終的な目的数よりも多く学習した BPE 語彙 V について，構成性の閾値 c を上回るトークンを最終的な目的数 n に達するまで頻度の低い順に削除することで非構成的な表現を選択的に語彙に取り込む．具体的な過程としては，まず，構成性を算出する過程で個々のトークンの埋め込みが必要となるため，予め通常通りの BPE アルゴリズムによって語彙の候補 V を獲得する．次に，この語彙候補について，BPE 学習時のコーパスを利用してそれぞれのトークン v について word2vec 埋め込み $\mathbf{E}(v)$ を学習し，最小単位でないものについては上記の式に従って構成性を算出する．この語彙候補を BPE の適用順にソートしたのち，最小単位，構成性の閾値を上回る部分集合，閾値以下の部分集合の 3 集合に分ける．すべての最小単位および閾値以下の部分集合を語彙として登録したのち，構成性の閾値を上回る部分集合の語彙に含まれるトークンを最終的な目的数に達するまで BPE の適用順に登録し，再び BPE の適用順にソートし直す．これにより，最終的な目的数のトークンが含まれ，かつ非構成的な表現を優先的に取り込んだ語彙 $V'(|V'| = n)$ を構築する．

3. 実験設定

提案手法によって獲得された，非構成的表現を多く含む語彙を用いることによる効果を検証するため，本実験では，提案手法によるフィルタリングを適用した BPE，および直接最終的な語彙サイズのもとで学習した BPE をそれぞれトークンとして採用した機械翻訳モデル間での比較を行う．

3.1 翻訳モデル

本実験では，6 層のエンコーダおよび 6 層のデコーダからなる Transformer [1] を翻訳モデルとして使用し，事前学習は行わず，データセットから直接モデルを学習した．

© 1959 Information Processing Society of Japan

表 1　純粋な BPE と，提案手法を通して非構成的語彙を除くためのフィルタリングをかけたときに得られた語彙をそれぞれ用いた翻訳モデルの推論精度.

	BLEU	COMET
BPE	39.82	89.47
+フィルタリング	**40.19**	**89.65**

3.2　語彙獲得

提案手法，ベースラインの両方において，サブワード学習アルゴリズムとして BPE を使用する．BPE には SentencePiece [9] の実装[*1]を用いたが，提案手法を実装する過程で，それぞれのサブワードが 2-gram としてのそれぞれの構成要素を出力できるよう改変した．なお，本実験では，イディオムなどの粒度の大きい表現を獲得するため，一般的な設定とは異なりスペースによる事前分割を行わない．本実験では，翻訳前後の両言語について，フィルタリング前の語彙サイズを $|V| = 48000$，最終的な語彙サイズを $n = 32000$，また構成性の閾値について，$c = 0.6$ と固定して実験を行った．また，SentencePiece を用いる際，文字のカバー率を日本語では 0.9995，英語では 1.0 とした．word2vec の学習時の設定については Cordeiro らの先行の実装に倣った．具体的なパラメタは付録の表に示す．

3.3　データセット

機械翻訳モデルの学習，および評価には，ASPEC 日英コーパス [6] を用いた．なお，学習データには，train データのうち上位 200 万文対を用いた．

3.4　評価指標

翻訳精度を評価するため，指標として BLEU [10] および COMET [11] を用いた．なお，BLEU は SacreBLEU [12] に実装されたもの[*2]を用い，COMET のモデルには `wmt22-comet-da` を用いた[*3]．以下，表に示す評価値には，異なるシードのもとで行う 3 回の試行の平均値を用いる．

4.　実験結果

表 1 に, 実験結果を示す．表の結果では, BLEU, COMET 両指標において，提案手法を用いたモデルの推論精度は元の BPE の推論精度と比較してわずかに高い．このことは，頻度に加えて構成性を考慮して語彙を学習することは有益であることを示している．

4.1　翻訳事例の分析

提案手法の適用前後でトークン分割が変わることによって起こる翻訳結果の違いを，事例として表 2 に示す．なお，事例 1，2 は，文ごとの COMET 評価値において提案手法と BPE での差を求め，その差が最も大きい 100 文から適当な例を抽出したものである．事例 1 では提案手法の評価値が高い例から，事例 2 では BPE の評価値が高い例から抽出した．まず，事例 1 では，いずれも「瘀血（おけつ）」に正しい漢字を充てられていない誤りは存在するが，両者の翻訳の違いを見ると，「東洋医学」の部分に差があり，原文を比較すると提案手法を適用した側では”oriental”が 1 トークンとして扱われているため，これが訳の差を生んでいるものと考えられる．”oriental”の語自体は”orient”と”al”から構成できる点で構成的と捉えることができるが，BPE 側のトークン列を見ると”ori”と”ental”から構成されており，これらのトークンから”oriental”の意味を直接導くことはできず，BPE としては非構成的な例と考えられる．事例 2 では，提案手法では”complaints”に対応する訳の「愁訴」を生成できていない．また，この「愁訴」は元の BPE 側のみに含まれるトークン，すなわち構成性が高いとして除去したものである．この語は専ら医学の分野で用いられる用語であり，それぞれの文字自体は医学に関連する意味を持たないため，構成性は低く本来削除するべきでない語と考えられる．しかし，「愁」が「愁訴」として出現する例が多いと仮定すると，「愁」は医学用語と共起することが多いため，「愁訴」と近い埋め込みを持つことで構成性が高いと判断されたと考えられる．

4.2　系列長への影響

BPE をはじめとしたサブワードの学習アルゴリズムでは，高頻度な文字列をトークンとしてまとめることで，語彙サイズの小ささと系列長の短さのバランスを取っている．ここで，提案手法では，素の BPE と比較したとき，語彙サイズを変えずに高頻度のトークンの一部を除去して非構成的な低頻度のものに置き換えているため，元の BPE と比較したとき，系列長の面でこのバランスが崩れている可能性がある．具体的には，文の平均トークン長が増加することが考えられる．そこで，両手法での入出力文の平均トークン長を計測した．

表 3 からは，日英いずれの言語でも元の BPE の方が平均トークン長が短く，提案手法では概して細かくトークン化される傾向にある．このことは，提案手法による語彙は文字列の圧縮効率の面では元の BPE に劣ることを示す．ただし，いずれの言語でも平均トークン長の増加幅は 1 以下と小さく，提案手法を用いることがモデルの処理速度に与える影響は小さいといえる．

4.3　入出力それぞれでの効果

表 1 に示した結果は入出力の両言語に提案手法を適用したときのものであるが，機械翻訳として使用した Transformer モデルはエンコーダ，デコーダを繋げた構造となっ

[*1] https://github.com/google/sentencepiece
[*2] https://github.com/mjpost/sacrebleu
[*3] https://huggingface.co/Unbabel/wmt22-comet-da

© 1959 Information Processing Society of Japan

表 2 BPE および提案手法のフィルタリングを適用した語彙を持つ翻訳モデルによる翻訳例.
文字列はトークンごとにスペースで分割，元の文章のスペースは␣の文字で代替する.
** **で囲まれたトークンは，その語彙にのみ含まれるトークンを指す.

事例 1		
BPE	原文	␣The ␣action ␣mechanism ␣depend ed ␣to ␣the ␣sever ity ␣of ␣the ␣” d irty ␣blood ␣syndrom ” ␣as ␣call ed ␣in ␣ori ental ␣medicine .
	翻訳文	␣ 作用機序 は 「 お 血 症候群 」 の 重症度 に依存する 。
提案手法	原文	␣The ␣action ␣mechanism ␣depend ed ␣to ␣the ** ␣severity ␣of ␣the** ␣” d irty ␣blood ␣syndrom ” ␣as ␣call ed ␣in ** ␣oriental** ␣medicine .
	翻訳文	␣その 作用機序 は 東 洋 医学 といわれる 「 お 血 症候群 」 の 重症 度に 左右 される 。
事例 2		
BPE	原文	␣F oot ␣complaints ␣are ␣recognized ␣in ␣8 ␣nurses .
	翻訳文	␣8 名の 看護婦 に 足 部 **愁訴** を認めた 。
提案手法	原文	␣F oot ** ␣complaints ␣are** ␣recognized ␣in ␣8 ␣nurses .
	翻訳文	␣8 名の 看護婦 に フ ッ ト レ ッ ク を認めた 。

表 3 提案手法と純粋な BPE それぞれの語彙でトークン化したときの１文あたり平均トークン長

	入力文 (En)	出力文 (Ja)
BPE	**22.05**	**20.35**
+フィルタリング	22.61	20.59

表 4 提案手法を入出力の言語の片側のみに適用したときの結果

	BLEU	COMET
BPE	39.82	89.47
+両言語フィルタリング	**40.19**	**89.65**
英語（入力）のみ	39.85	89.58
日本語（出力）のみ	39.88	89.57

ており，それぞれ入力言語に対する言語理解，出力言語での言語生成を担当するため，片側のみに提案手法を適用したモデルを学習することで，言語理解，言語生成についてそれぞれ提案手法の効果を測る．表 4 にその結果を示す．表より，提案手法は入出力の片側のみに適用した場合でも元の BPE での翻訳精度を上回るが，いずれも両言語に適用したときの翻訳精度は下回っており，入出力それぞれに適用することによる効果は共存するといえる．

5. 関連研究

本節では，本研究で試みた非構成的なサブワードのフィルタリングに関連する研究として，モデルにとって有用な語彙を与えるための研究，ニューラル言語処理において非構成的な表現と強く関連する複単語表現を扱った研究をそれぞれ説明する．

5.1 サブワード語彙の獲得

サブワードを語彙として用いるモデルの性能向上や推論速度向上のため，効率的なサブワードの学習方法や，語彙として適切なサイズや単位を探る研究が行われている．既存のサブワードのアルゴリズムを拡張することでモデルの性能を高めることを目的とした研究としては，タイプミスなどによる細かい表現への頑健性を高める研究 [13] や後段タスクのドメインに適した語彙を与える研究 [14] が存在する．また，特に本研究のように有用でないと考えられる語彙を除去する方向での研究としては，モデルには含まれるが実際に出現することが考えづらい語彙を検出する研究 [15] や多言語モデルにおいて特定の言語での処理を行うのに必要ない語彙を除去する研究 [16] が挙げられる．本研究ではこれらの研究と同様にサブワード語彙の一部を削除するが，これらの研究が大規模なコーパスで学習するために不要な語彙が入るという仮定のもとで行われているのに対し，本研究はサブワードアルゴリズム自体の特徴を理由に収録される語彙を対象とする点で異なる．

適切な語彙サイズを探る研究としては，Xu ら [17] は，モデルの性能を高めるトークンを最適輸送によって選択的に語彙として含めることでモデルの性能を高めるコンパクトな語彙の学習手法を提案した．また，語彙サイズを増やし，モデルが扱える語彙を直接増やすことで多言語モデル [18] や大規模言語モデル [19] の性能が向上することが報告されているが，多言語モデルの例では語彙サイズを増やし続けたとき性能が向上し続けるわけではないこと，また語彙サイズの増大に伴いデコーダの softmax 関数の計算コストが増大するなど，このような場合でも語彙の選択は重要である．

また，サブワードの単位に焦点を当てた研究としては，Wang ら [20] はサブワードの最小単位をバイト単位まで縮めることで，サブワード語彙が任意の言語表現を受け付けるようにしたほか，後述するように複単語表現を単位として扱えるようにした研究 [21], [22], [23], [24] が存在する．

© 1959 Information Processing Society of Japan

本研究も，これらの研究と同様，単語以下の狭義の「サブワード」に限定せず，トークンの単位として，後述の複単語表現のような意味表現を含めることを志向したものである．

5.2 複単語表現を用いたニューラル言語処理

イディオム，複合語といった，複数の語でありながら独立した意味を持つ表現を対象とした研究として，表現の検出やモデル内部での処理の分析，また，本研究での構成性と同様に，これらの表現を後段タスクで利用する研究が存在する．具体的なタスクへの応用としては，Otani らによる，多言語埋め込みにおいて言語間で粒度の揃った表現対を獲得するために単語に加えて複単語表現を単位として扱う研究 [25] が存在する．また，機械翻訳への応用として，Zaninello と Birch による，複単語表現のリストを与えた機械翻訳の研究が存在する [21].

本研究と同様にサブワードとこれらの表現を組み合わせる研究としては，Kumar と Thawani による機械翻訳の研究 [22] があり，単語内 BPE に加えて頻度順に 2 語，3 語，および 1 語を挟んだ 2 語による skip-gram を追加できるようにした設定でそれぞれ機械翻訳モデルを学習しているほか，Chirkova と Troshin による研究 [23] ではコード生成を対象として単語を跨いだ設定とそうでない設定の BPE の比較を行っているが，これらの研究では単語を跨いだ表現を含んだ BPE では精度は向上しないか悪化することが報告されている．一方で，Liu らの研究 [24] では，汎用大規模言語モデルを元にドメイン特化モデルを構築する際に，unigram によるサブワードを用いたドメイン特化語彙で複単語表現を含めて追加し，これは性能の向上を報告している．

本研究では，これらの研究で扱う複単語表現の持つ非構成性を踏まえ，サブワードに基づくニューラル機械翻訳の語彙にこれらの表現のような非構成的な表現を含めることを志向したものである．

6. 結論

本研究では，頻度のみによって計算される BPE に対して，非構成的な表現を選択的に取り込んだ際の影響を，英日機械翻訳モデルの学習，評価を通して確かめた．また，これらの表現を取り込む手段として，構成的な表現をフィルタリングで除去するための手法を提案し，また実装した．実験の結果，BPE に非構成的な表現を選択的に取り込むことは機械翻訳モデルにおいて若干の性能向上をもたらすことを確かめた．今後は，翻訳対を増やし実験で得られた傾向の普遍性を確かめ，また異なるフィルタリングの閾値，異なる語彙サイズで実験することによって非構成的な表現をどれだけ含めることが最適なのかを確かめたい．また，より精度の高い語彙を抽出するため，より適した構成性の算出手法を検討したい．

謝辞 本研究は，東京大学生産技術研究所特別研究経費および JSPS 科研費 JP21H03494 の助成を受けたものである．

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention Is All You Need, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc. (2017).

[2] Wolleb, B., Silvestri, R., Vernikos, G., Dolamic, L. and Popescu-Belis, A.: Assessing the Importance of Frequency versus Compositionality for Subword-based Tokenization in NMT, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C. and Moniz, H., eds.), Tampere, Finland, European Association for Machine Translation, pp. 137–146 (2023).

[3] Dankers, V., Lucas, C. and Titov, I.: Can Transformer Be Too Compositional? Analysing Idiom Processing in Neural Machine Translation, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 3608–3626 (online), DOI: 10.18653/v1/2022.acl-long.252 (2022).

[4] Feucht, S., Atkinson, D., Wallace, B. C. and Bau, D.: Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 9727–9739 (2024).

[5] Cordeiro, S., Villavicencio, A., Idiart, M. and Ramisch, C.: Unsupervised Compositionality Prediction of Nominal Compounds, *Computational Linguistics*, Vol. 45, No. 1, pp. 1–57 (online), DOI: 10.1162/coli_a_00341 (2019).

[6] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S., eds.), Portorož, Slovenia, European Language Resources Association (ELRA), pp. 2204–2208 (2016).

[7] Gage, P.: A New Algorithm for Data Compression - Document - Gale General OneFile (1994).

[8] 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 Lin-

© 1959 Information Processing Society of Japan

guistics, pp. 1715–1725 (online), DOI: 10.18653/v1/P16-1162 (2016).

[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 (online), DOI: 10.18653/v1/D18-2012 (2018).

[10] 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 on Association for Computational Linguistics*, ACL '02, USA, Association for Computational Linguistics, pp. 311–318 (online), DOI: 10.3115/1073083.1073135 (2002).

[11] Rei, R., Stewart, C., Farinha, A. C. and Lavie, A.: COMET: A Neural Framework for MT Evaluation, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B., Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 2685–2702 (online), DOI: 10.18653/v1/2020.emnlp-main.213 (2020).

[12] Post, M.: A Call for Clarity in Reporting BLEU Scores, *Proceedings of the Third Conference on Machine Translation: Research Papers* (Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M. and Verspoor, K., eds.), Brussels, Belgium, Association for Computational Linguistics, pp. 186–191 (online), DOI: 10.18653/v1/W18-6319 (2018).

[13] 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 (online), DOI: 10.18653/v1/2020.acl-main.170 (2020).

[14] Hiraoka, T., Takase, S., Uchiumi, K., Keyaki, A. and Okazaki, N.: Optimizing Word Segmentation for Downstream Task, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Association for Computational Linguistics, pp. 1341–1351 (online), DOI: 10.18653/v1/2020.findings-emnlp.120 (2020).

[15] Land, S. and Bartolo, M.: Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Al-Onaizan, Y., Bansal, M. and Chen, Y.-N., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 11631–11646 (2024).

[16] Ushio, A., Zhou, Y. and Camacho-Collados, J.: Efficient Multilingual Language Model Compression through Vocabulary Trimming, *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 14725–14739 (online), DOI: 10.18653/v1/2023.findings-emnlp.981 (2023).

[17] Xu, J., Zhou, H., Gan, C., Zheng, Z. and Li, L.: Vocabulary Learning via Optimal Transport for Neural Machine Translation, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Association for Computational Linguistics, pp. 7361–7373 (online), DOI: 10.18653/v1/2021.acl-long.571 (2021).

[18] Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., Zettlemoyer, L. and Khabsa, M.: XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 13142–13152 (online), DOI: 10.18653/v1/2023.emnlp-main.813 (2023).

[19] Takase, S., Ri, R., Kiyono, S. and Kato, T.: Large Vocabulary Size Improves Large Language Models (2024).

[20] Wang, C., Cho, K. and Gu, J.: Neural Machine Translation with Byte-Level Subwords, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9154–9160 (online), DOI: 10.1609/aaai.v34i05.6451 (2020).

[21] Zaninello, A. and Birch, A.: Multiword Expression Aware Neural Machine Translation, *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S., eds.), Marseille, France, European Language Resources Association, pp. 3816–3825 (2020).

[22] Kumar, D. and Thawani, A.: BPE beyond Word Boundary: How NOT to Use Multi Word Expressions in Neural Machine Translation, *Proceedings of the Third Workshop on Insights from Negative Results in NLP* (Tafreshi, S., Sedoc, J., Rogers, A., Drozd, A., Rumshisky, A. and Akula, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 172–179 (online), DOI: 10.18653/v1/2022.insights-1.24 (2022).

[23] Chirkova, N. and Troshin, S.: CodeBPE: Investigating Subtokenization Options for Large Language Model Pretraining on Source Code (2023).

[24] Liu, S., Deng, N., Sabour, S., Jia, Y., Huang, M. and Mihalcea, R.: Task-Adaptive Tokenization: Enhancing Long-Form Text Generation Efficacy in Mental Health and Beyond, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 15264–15281 (online), DOI: 10.18653/v1/2023.emnlp-main.944 (2023).

[25] Otani, N., Ozaki, S., Zhao, X., Li, Y., St Johns, M. and Levin, L.: Pre-Tokenization of Multi-word Expressions in Cross-lingual Word Embeddings, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B., Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 4451–4464 (online), DOI: 10.18653/v1/2020.emnlp-main.360 (2020).

© 1959 Information Processing Society of Japan