意味的構成性を考慮した言語モデルの語彙削減

田村 鴻希 [1,3]　吉永 直樹 [2,3]
[1] 東京大学 [2] 東京大学生産技術研究所 [3] 東京大学デジタルオブザーバトリ研究推進機構
tamura-k@tkl.iis.u-tokyo.ac.jp　ynaga@iis.u-tokyo.ac.jp

概要

言語モデルの大規模化，多言語化に伴い，語彙も肥大化が顕著である．コンパクトかつ効果的な語彙とするため，事前トークン化 (pretokenization) やドメインでの低頻度語彙の削減が提案されているが，包括的かつ説明性の高い基準は見出されていない．本研究では，サブワード獲得手法のバイト対符号化 (BPE) で学習された言語モデルの語彙を対象に，頻度に加え意味を構成要素から計算可能かの指標の構成性を同時に考慮して語彙を削除する手法を提案する．提案手法を ModernBERT に対して適用し，言語理解タスクでの性能評価実験を通してタスクの学習データの頻度のみに基づいて削除した場合と比較し，効果を検証する．

1　はじめに

大規模言語モデル (LLM) は運用に潤沢な計算資源が必要であり，ユーザ環境下で運用する際には，モデルサイズが大きな問題となる．特に，近年のモデルでは，量子化など汎用的なモデル圧縮の対象とならない語彙サイズの肥大化 [1, 2] が顕著であり，その効率化は重要な研究課題といえる．

この点を踏まえ，言語モデルの語彙を最適化する手法が研究されている．語彙の最適化は，主に統計的にサブワード語彙を学習する際に，不要なものを語彙に含めないようにすることで行われる．不要な文字列を語彙から取り除くための主な手法としては，語彙を学習する前に正規表現などで分割境界を与え，これを跨ぐ文字列を語彙に含めないようにする事前トークン化 (pre-tokenization) が挙げられる．対照的に，事後的に非有用な文字列を語彙から除く手法として，低頻度のものをモデルから明示的に除く手法が存在する [3, 4]．ただし，前者は数字や記号の連続 [5] など，個々の記号が明確に意味を持つ場面には効果があるが，それ以外への影響は限定的であり，後者は出現頻度のみを参照するため，実際に取り除く語彙が無用かという視点は欠けている．

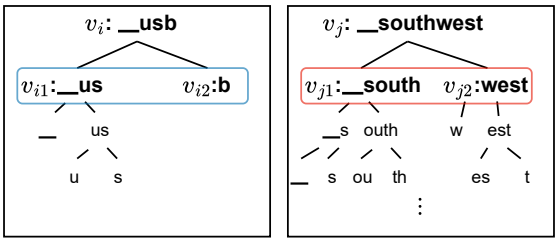

図 1　BPE トークンの構成性: 意味が非構成的な結合（左）と構成的な結合（右）．

本研究では，説明性の高い統一的なサブワード語彙の選択基準として，トークンの構成性（図 1）を考慮することを提案し、これを頻度と組み合わせた学習済みモデルのサブワード語彙削減手法を提案する．構成性は，複合語において構成要素の語の意味によって複合語全体の意味を説明可能であるかの数値的指標であり，構成性が低いほど個々の要素からは説明できない意味を持つといえる．同指標をサブワード語彙に転用し，各トークンについて値を算出することで，それぞれのトークンを構成要素で代替した際の，元のトークンの意味の説明可能性を数値化する．この構成性の値とバイト対符号化 (BPE) [6, 7] 自体のトークンの優先度であるトークン頻度と組み合わせた値を語彙中の全トークンに与え，その値を元にトークンの優先度を順位付ける．

実験では，事前学習済みモデルの ModernBERT [8] を対象に，同手順による順位に基づいて一定割合の語彙を削除し，GLUE 言語理解ベンチマーク [9] での影響を評価した．

2　提案手法

本研究では，構成性を考慮したスコアに基づき事前学習済みモデルの語彙の一部を削除する手法を提案する．元のモデルや BPE の語彙選択基準である頻度のみに基づき語彙を削除したモデルと比較して，削除基準に構成性を導入する効果を検証する．

本研究で扱う BPE 語彙の性能には，高頻度のトークンを収録していることが強く寄与するため，頻度と組み合わせる形で基準となるスコアを設ける．具体的には，微調整タスクの学習データセットでのトークン頻度と個々のトークンの構成性を，スコア付けに適する形に補正したのち掛け合わせた値が低い順に削除する．具体的に，構成性と頻度の補正後の値をそれぞれ構成性スコア，頻度スコアと呼称し，構成性，頻度の両スコアの積をトークンのスコアとする．スコア付け，および語彙の削除手順を示すため，以下で，本研究で扱う語彙とトークンの説明，構成性の計算手法および補正手法，トークン頻度の補正手法，具体的な削除手順を順に述べる．

2.1 予備知識: BPE

本研究では，バイト対符号化 (Byte-Pair Encoding; BPE) [6, 7] で語彙を学習し，事前学習された言語モデルを語彙削減の対象とする．BPE では，まず文字やバイト等の単位の集合を初期語彙として学習データを分割，トークン化する．追加語彙は，トークン列中で高頻度の 2-gram を逐次的に連結し，その度に新しい 1 トークンとして加えることで学習される．そのため，初期語彙以外のトークンはいずれも結合元のトークンを 2 つずつ持つ．

本研究では，この性質を利用し，初期語彙以外のトークンについて結合元の 2 トークンを利用して構成性を求める．また，この性質を再帰的に適用して，個々のトークンに初期語彙まで辿る木構造を図 1 のように与え，頻度の補正に用いる．

2.2 構成性の計算

まず，モデルの語彙中の個々のトークンの意味が構成要素から計算できるかの指標である構成性を算出する手法について述べる．本研究では，構成性の具体的な算出手法として，Cordeiro ら [10] の名詞複合語に対する算出手法を BPE サブワードへと転用した．同手法では，それぞれの複合語 w とその構成要素 w_1, w_2 についてそれぞれ埋め込み $\mathbf{E}$ を求め，構成要素の埋め込みの加重平均と複合語全体の埋め込みのコサイン類似度を測る．

本研究では，当手法を事前学習済みモデルの BPE 語彙 V 中のトークン v に適用する．個々のトークンの埋め込みには，対象の事前学習済みモデルの埋め込み層のトークン埋め込みを用いる．また，加重平均の重みは，Cordeiro らの実験において英語データで人手アノテーションとの相関係数が最も高い 0.5 に定める．ここで，v の結合元のトークンを v_1, v_2 とすると，v の構成性は以下のように計算できる．

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

同手法で得られた構成性は，頻度の分布と比べてとる値とその幅が小さく，頻度と直接掛け合わせる際に寄与する幅が限定的になる問題を抱える．また，本研究では非構成的なトークンを優先したいため，構成性が低いトークンが高いスコアを持つ必要がある．そのため，構成的なトークンに低い値を与え，また全体のスコアにおいてトークンの構成性を重視するための補正を行う．

具体的には，個々のトークンの構成性の補正に用いる関数を f とおき，構成性スコアを $\mathrm{cs}(v) = f(\mathrm{cp}(v))$ と定義する．f は，引数を $x \in \mathbf{R}, -1 \le x \le 1$ としたとき，(i) $f(x)$ が単調減少であり，構成性が低いほど高い値をとる (ii) $f(x) \ge 0$ を満たすように定める．この f を複数用意し，実験で比較を行う．

2.3 頻度スコア

次に，語彙の頻度スコアの計算手法について述べる．本研究では，トークン頻度として，事前学習済みモデルの微調整時の訓練データにモデルのトークン化器を適用し，文章をトークン列とした際の各トークンの出現頻度 $\mathrm{freq}(v)$ を用いる．

ここで，BPE による語彙は，専ら他のトークンの構成要素（中間トークン）として存在し，ほとんど出現しないトークンを含む [11]．図 1 右の中間トークン “outh” を例にとると，単体では意味を持たず，主に “south” のような他トークンの一部として存在する．よって，直接頻度をとると低い値をとり，削除対象となりやすい．しかし，BPE の結合は再帰的なため，同トークンを削除すると，より頻出と考えられる結合後の “south,” “southwest” のトークンを扱えなくなる問題が存在する．ただし，中間トークン自身の意味が構成的に計算できる場合にはその時点で削除することを指向する．

そこで，頻度スコア $\mathrm{fs}(v)$ として，自身を子要素に持つ親トークン集合の頻度スコアの最大値を参照する．これにより中間トークンには全ての親要素以上の頻度スコアを与え，削除時には全ての親トークンを削除する．なお，1 つのトークンは同時に親要素にも子要素にもなりうるため，親トークンの頻度は BPE の結合順の末尾から順に再帰的に参照する．

表 1　語彙の異なる設定での，ModernBERT (base) モデルの GLUE タスクの推論結果

$\frac{\lvert V'\rvert}{\lvert V\rvert}$	f(x)	CoLA	MNLI-m	MRPC	QNLI	QQP	RTE	SST2	STSB	avg
1/4	$1-x$	59.92	**87.91**	91.70	**92.70**	90.13	82.79	**95.18**	**91.50**	86.48
	$\frac{x'}{1-x'}$	58.89	87.84	91.19	92.59	**90.16**	83.03	94.72	**91.50**	86.24
	only freq	**60.54**	87.85	**91.88**	92.65	90.13	**83.39**	94.95	91.44	**86.60**
1/2	$1-x$	58.09	**88.36**	**91.70**	**93.03**	90.20	83.15	95.03	91.70	86.41
	$\frac{x'}{1-x'}$	**60.93**	88.29	91.57	93.02	90.27	83.15	**95.11**	**91.73**	**86.76**
	only freq	59.66	88.34	91.35	93.01	**90.41**	**83.75**	95.03	91.63	86.65
1	original	59.16	88.50	91.52	93.16	90.32	84.36	95.15	91.92	86.76

2.4　構成性・頻度スコアを用いた語彙削除

対象のモデルに含まれる BPE 語彙中のトークン $v \in V$ について，上述の頻度スコアと構成性スコアの積 $\mathrm{ts}(v) = \mathrm{ts}(v) \cdot \mathrm{cs}(v)$ が低い順に削除する．また，削除時には削除したトークンを子要素に持つ親トークン全体を同時に削除する．

3　実験設定

本節では，事前学習済みモデルの ModernBERT [8] を対象に，上記の手順に従って $\mathrm{ts}(v)$ の低いトークンを一定割合削除したモデルの性能を比較する．同モデルは，トークン化器として LLM の OLMo [12] (v1) のものを流用しており，語彙は BPE アルゴリズムにより Web コーパスから学習されている．

3.1　評価タスク

本実験では，手法を適用したモデルの評価のためのデータセットとして，言語理解ベンチマークである GLUE [13] を用いて実験した結果を記載する．同ベンチマークは，受容性（CoLA），感情分析（SST-2），言い換え（MRPC，QQP），テキスト類似度（STS-B），および自然言語推論（NLI; MNLI，QNLI，RTE，WNLI）のデータセットを含む．本実験では，適用先の ModernBERT の論文内の実験と同様に，WNLI を除いた dev データでの結果[1)]を記載する．

3.2　構成性の考慮手法

実験で比較する構成性スコアの関数 f を示す．

- $f(x) = 1 - x$: 最も単純に，構成性の値が高いときに小さい値をとるように正負を反転させる．
- $f(x) = \frac{x'}{1-x'}$: 構成性の値が高いトークンでより低く，低いトークンでより高い値をとる補正[2)]．

1)　GLUE の test データの評価用ラベルは非公開．
2)　x' は x の集合 X のうち最小値が ϵ，最大値が $1-\epsilon$ となるようにスケーリングしたもの．実験では，$\epsilon = 10^{-4}$ とした

3.3　学習設定

本実験では，既存の実験設定に従い，RTE, MRPC, STS-B の 3 データセットでは MNLI で微調整したモデルを初期値として学習を行う．データセットごとのハイパーパラメタとして，ModernBERT の論文内で base モデルで最も dev データでの評価指標が最良となった学習率，weight decay を採用する．学習エポック数は，最大値を適用先モデルにおいて微調整を行ったときのものと揃え，2 エポック連続で評価指標の値が向上しなければ学習を打ち切った．具体的なハイパーパラメタは付録の表 4 に示す．以下の節で示す評価値には，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，ModernBERT (base) モデルの GLUE ベンチマークでの実験結果を示す．結果から，構成性を考慮した設定がベンチマーク内の各種タスクにおいて頻度のみに従って語彙を削除したベンチマークと比較して削除基準として有益であるかは曖昧であり，ベンチマーク全体に対して有益とは言えない．また MRPC と QQP，MNLI, QNLI と RTE のように，同じ類型のタスクの間でも最良の設定は異なり，それ以外のタスクについては設定の語彙サイズが変われば最良の $f(x)$ が異なる点からも曖昧性が伺える．

4.1　結果の分析

本研究では，元のモデルの語彙の一部を削除した設定で学習を行い，その結果の違いを比較した．しかし，BPE では一部のトークンが全体の出現頻度の多くを占め，データセットの内容次第ではトークン化の結果がそれらの高頻度トークンのみで占められる可能性が存在する．そこで，元モデルと比較対象

表 2　頻度のみの設定とトークン列が異なる事例での，頻度のみからの差分

$\frac{\lvert V'\rvert}{\lvert V\rvert}$	f(x)	MNLI-m	QNLI	QQP	SST2
1/4	$1-x$	0.23	-0.02	0.36	0.69
	$\frac{x'}{1-x'}$	-0.06	-0.20	-0.15	0.38

の各スコアで語彙を削除したモデルでトークン列が異なるエントリの割合を付録の表 3 に提示する．表からは，特に語彙サイズ=1/4 の設定において，トークン化が異なる文章の割合は $f(x) = \frac{x'}{1-x'}$ で補正をかけた導入，$f(x) = 1-x$ での単純な導入，頻度のみの設定，の順に大きく，構成性の考慮幅の大きさに沿うことがわかる．一方で，語彙サイズ=1/2 の設定では，特にデータセットの大きい MNLI, QNLI, QQP を除いてこの傾向が崩れている．これは，語彙の量に対してデータセットの学習データに現れるトークンの種類が少なく，設定間で語彙にほとんど差が生まれないことが原因と考えられる．

次に，構成性を導入することでトークン化の傾向に差が生まれているかを示すため，同じ語彙サイズで頻度のみを基準とした設定との間で，同様にトークン化の結果が異なるエントリの割合を提示する．

表 2 に，トークン化の差がモデルの性能の差に繋がっているかを測るため，この両者の間でトークン化の結果が異なるエントリのみを対象としてそれぞれ評価指標での結果を示す．なお，当分析は，表 3 において想定する割合の傾向が見られた設定の，語彙サイズ 1/4 の MNLI, MRPC, QNLI, QQP の 4 データセットに対して行った．表 2 の結果からは，単純に語彙を追加した設定と比較して，f(x) = $1-x$ の設定ではやや推論性能が高い傾向がみられる．一方，f(x) = $\frac{x'}{1-x'}$ では 4 データセットのうち 3 データセットで推論性能が低下している．f(x) = $1-x$ での結果は，トークンの構成性を削除基準に組み込んだ場合でも推論に悪影響がないか，よい影響を及ぼす可能性が考えられる．ただし，構成的なトークンをより削除する f(x) = $\frac{x'}{1-x'}$ では推論性能が低下傾向のため，中頻度のトークンを削除することの影響は無視できないといえる．当分析で使用した文章の全データ中の割合は，付録の表 5 に記す．

5　関連研究

本節では，本研究で取り組んだ内容に関連する研究として，モデルにとって効果的なトークンの範疇を探る研究，また，本研究で利用した構成性に関連する研究についてそれぞれ説明する．

5.1　サブワード語彙の最適化

非効果的なトークンを削減する研究としては，中間トークンを取り除くために語彙から削除 [14] したり，構成要素にフォールバック [15] する手法がある．また，語彙中に含まれるが実際の推論には現れない記号列などのトークンを特定する研究 [16] も行われている．一方で，モデルが扱うトークンの範囲を拡張する手法として，BPE の学習中に，一定の制約の下でスペースを跨ぐことを許可する研究 [17, 18] が挙げられる．また，本研究のようにトークンの構成的な構造を利用するものとしては，語の屈折による派生語を，独立したトークン埋め込みではなく語根と派生形の組み合わせで表現する研究 [19] が存在する．ただし，この研究は英語の語全体を纏め上げたトークンを対象としており，本研究のように BPE 語彙全体を対象としたものではない．

5.2　単語の意味の構成性

語句の意味の敵構成性を算出する研究には，word2vec などの事前学習済み埋め込みから構成性を算出する研究 [10]，BERT の中間層から複合語かを判定する分類器を抽出する [20] 研究がある．また，関連して，イディオムなどの動詞句を検出する分類器を学習する研究 [21] や，その学習の過程で明示的に構成性の計算を行った研究 [22] が挙げられる．

6　結論

効率的な語彙体系を導入することで事前学習済みモデルの性能を保ちながら語彙を削減することを目的に，語彙の選択基準となるスコアの設計に構成性を導入することを試みた．実験では，構成性と頻度の両要素からなるスコアを用いて事前学習済み言語モデル ModernBERT の語彙の一部を削除し，GLUE ベンチマークを用いて言語理解タスクでの推論性能を比較した．実験の結果，GLUE ベンチマーク全体を通した性能の良化は確認できなかったが，データ量の多いデータセットにおいては，頻度のみを基準に語彙を削除した場合と比較して性能が向上する傾向のある設定の存在を確認した．

謝辞

本研究は，東京大学デジタルオブザーバトリ研究推進機構の支援を受けて行われた．

参考文献

[1] Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 114147–114179. Curran Associates, Inc., 2024.

[2] Sho Takase, Ryokan Ri, Shun Kiyono, and Takuya Kato. Large vocabulary size improves large language models. In **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 1015–1026, Vienna, Austria, July 2025. Association for Computational Linguistics.

[3] Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. Efficient Multilingual Language Model Compression through Vocabulary Trimming. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 14725–14739, Singapore, February 2023. Association for Computational Linguistics.

[4] Itay Nakash, Nitay Calderon, Eyal Ben-David, Elad Hoffer, and Roi Reichart. Adaptivocab: Enhancing LLM efficiency in focused domains through lightweight vocabulary adaptation. In **Second Conference on Language Modeling**, 2025.

[5] Aaditya K. Singh and D. J. Strouse. Tokenization counts: The impact of tokenization on arithmetic in frontier LLMs, February 2024.

[6] Philip Gage. A new algorithm for data compression - Document - Gale General OneFile, 1994.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[10] Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised Compositionality Prediction of Nominal Compounds. **Computational Linguistics**, Vol. 45, No. 1, pp. 1–57, March 2019.

[11] Kaj Bostrom and Greg Durrett. Byte Pair Encoding is Suboptimal for Language Model Pretraining. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 4617–4624. Association for Computational Linguistics.

[12] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024.

[13] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, January 2018. Association for Computational Linguistics.

[14] Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 16587–16604, Miami, Florida, USA, January 2024. Association for Computational Linguistics.

[15] Haoran Lian, Yizhe Xiong, Jianwei Niu, Shasha Mo, Zhenpeng Su, Zijia Lin, Hui Chen, Jungong Han, and Guiguang Ding. Scaffold-BPE: Enhancing Byte Pair Encoding for Large Language Models with Simple and Effective Scaffold Token Removal. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 39, No. 23, pp. 24539–24548, April 2025.

[16] Sander Land and Max Bartolo. Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 11631–11646, Miami, Florida, USA, January 2024. Association for Computational Linguistics.

[17] Alisa Liu, Jonathan Hayase, Valentin Hofmann, Sewoong Oh, Noah A. Smith, and Yejin Choi. SuperBPE: Space Travel for Language Models, April 2025.

[18] Craig W. Schmidt, Varshini Reddy, Chris Tanner, and Yuval Pinter. Boundless Byte Pair Encoding: Breaking the Pre-tokenization Barrier. In **Second Conference on Language Modeling**, August 2025.

[19] Yuval Reif, Guy Kaplan, and Roy Schwartz. Vocab Diet: Reshaping the Vocabulary of LLMs with Vector Arithmetic, October 2025.

[20] Filip Miletic and Sabine Schulte im Walde. A Systematic Search for Compound Semantics in Pretrained BERT Architectures. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1499–1512, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[21] Ziheng Zeng and Suma Bhat. Idiomatic Expression Identification using Semantic Compatibility. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1546–1562, 2021.

[22] Kazuma Hashimoto and Yoshimasa Tsuruoka. Adaptive joint learning of compositional and non-compositional phrase embeddings. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 205–215, Berlin, Germany, August 2016. Association for Computational Linguistics.

A 付録

付録として、本文中に記載のない表を記載する。

表 3 実験設定ごとの，original モデルとトークン化の異なる文章の割合

$\frac{\lvert V'\rvert}{\lvert V\rvert}$	f(x)	CoLA	MNLI-m	MRPC	QNLI	QQP	RTE	SST2	STSB
1/4	$1-x$	21.57	71.78	85.78	94.01	48.93	94.22	46.90	64.93
	$\frac{x'}{1-x'}$	21.57	73.16	86.27	94.69	50.44	94.22	47.48	64.80
	only freq	23.20	70.68	86.76	93.74	48.14	94.58	46.44	64.93
1/2	$1-x$	13.14	26.93	59.80	60.97	15.82	68.59	21.56	64.93
	$\frac{x'}{1-x'}$	13.14	28.18	59.80	61.94	16.45	68.59	21.56	64.80
	only freq	18.40	26.53	62.26	61.25	15.56	71.12	29.01	64.93

表 4 データセットごとのハイパーパラメタ設定

	lr	wd	最大エポック
CoLA	8e-5	1e-6	10
MNLI-m	5e-5	5e-6	3
MRPC	5e-5	5e-6	10
QNLI	8e-5	5e-6	10
QQP	5e-5	5e-6	10
RTE	5e-5	1e-5	3
SST2	8e-5	1e-5	3
STSB	8e-5	5e-6	10

表 5 頻度のみの設定とトークン列が異なる事例の割合 (%)

f(x)	MNLI-m	QNLI	QQP	SST2
$1-x$	23.44	38.95	11.22	5.50
$\frac{x'}{1-x'}$	38.24	59.14	19.57	9.98