低資源言語のための辞書を用いた言語間語彙転移

坂上 温紀 [1]　Justin Vasselli[1]　井手 佑翼 [1]
坂井 優介 [1]　Yingtao Tian[2]　上垣外 英剛 [1]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学　[2]Sakana AI
sakajo.haruki.sd9@naist.ac.jp　tian@alantian.net
{vasselli.justin_ray.vk4,ide.yusuke.ja6,sakai.yusuke.sr9,kamigaito.h,taro}
@is.naist.jp

概要

事前学習済みモデルの性能を新たな言語、特に低資源言語に転移させる際、事前学習済みモデルのサブワード埋め込みを用いて目的言語のサブワード埋め込みを初期化する手法が知られている。それらは目的言語のコーパスを用いるが、低資源言語の中には十分なコーパスが存在しない場合が多い。その一方で、多くの言語には対訳辞書が存在している。そこで本研究では、対訳辞書に基づくサブワード埋め込みの初期化の手法を提案する。実験の結果、既存の手法では課題となっていた言語に対して提案手法は有効であることが明らかとなった。

1　はじめに

自分が知らない言語で書かれた書物を読む際、辞書を引き、自分がよく知る言語の単語と関連づけることで、その内容を知ることができる。同様に、言語モデルを新たな言語に対応させるときには、文脈内学習や継続事前学習など、様々な手法がある。

その中の一つに、事前学習済みモデルの語彙を目的言語に転移させるという手法がある。既存の手法はパラレルコーパスを用いて原言語と目的言語の対応関係を取ったり [1]、あるいは原言語と目的言語の語彙の重なりに着目したりして、語彙を転移させる [2, 3, 4, 5]。しかし、これらの手法は、目的言語が低資源言語の場合、原言語とのパラレルコーパスが存在していないという問題に直面する [6, 7]。

そこで本研究では、この問題を解決するために、対訳辞書を基盤とした新たな手法を提案する。本手法は、目的言語のトークナイザーの訓練と、二言語間でのサブワードの対応付け、それを用いた目的言語のサブワードの埋め込みの初期化からなる。

実験は RoBERTa [8] を原モデルとして、7 つの言語を用いて行った。実験の結果、提案手法は用いるデータが少量にもかかわらず、一部の言語においては従来手法よりも良い性能を収めた。特に従来手法では転移先での性能が限られていた言語においては、提案手法が有効であることが示された。本研究の手法は、対訳辞書という多くの低資源言語で利用可能な言語資源を用いることで効果的に事前学習済みモデルを他の言語に適応させるものであり、低資源言語の広範なタスクにまで既存の言語モデルの恩恵を届ける可能性を開くものである。

2　関連研究

事前学習済みモデルを新たな言語に適用させるにはいくつかのアプローチがある。その中の一つに、原言語のサブワード埋め込みを用いて目的言語のサブワード埋め込みを初期化するアプローチがある。WECHSEL [2] は各言語の単言語コーパスを用いて獲得した静的埋め込みと、原言語・目的言語の対訳辞書を用いて、目的言語のサブワード埋め込みを初期化する手法を提案した。。一方で、FOCUS [3] は対訳辞書を用いず、原言語と目的言語のサブワードの重なりに注目してサブワード埋め込みを初期化する。さらに、FOCUS はサブワードの初期化のあとに Language-Adaptive Pretraining (LAPT) を行う。UniBridge [4] は語彙サイズの探索を行った後、統語と意味の両方に注目してサブワード埋め込みを初期化する。FOCUS と同様にサブワードの重なりに注目して初期化をするが、サブワードが重なることが稀な言語に対しては単言語コーパスから得られる静的埋め込みを用いて類似度を考慮して初期化する。また、UniBridge は LAPT に MLM 損失と KL divergence を用いている。Trans-Tokenizer [1] は対訳コーパスを用いて単語単位のアライメントを取ることで目的言語のサブワードを初期化する。

WECHSEL は静的埋め込みを作るためのある程度の規模の単言語コーパスと原言語と目的言語の二言語辞書が必須という点で低資源言語への適用に課題があった。FOCUS と UniBridge はサブワードの重なりを用いるため、原言語として用られる言語と異なる文字を用いていたり、サブワードの重なりがほとんどなかったりするときにはやはり限界がある。Trans-Tokenizer は原言語と目的言語の対訳コーパスを用いるため、それが存在しない低資源言語には適用できないという課題がある。

3 提案手法

本研究では、目的言語の単語（エントリー）とそれに対する英語による定義からなる辞書を用いた新たな手法を提案する。提案手法は目的言語のトークナイザーの訓練、二言語間のサブワードの対応付け、目的言語のサブワード埋め込みの初期化からなる。

3.1 トークナイザーの訓練

辞書のエントリーを用いて目的言語の byte-level BPE トークナイザー [9] を訓練する。

3.2 サブワードの対応付け

本研究では、サブワード単位で対応付けと単語単位で対応付けという 2 種類の対応付けの手法を提案する。２つに共通して、辞書のエントリーを用いて訓練したトークナイザーと、転移元のモデル（原モデル）のトークナイザーを用いて、辞書のエントリーと定義の両方をサブワード分割する（図 1a）。このエントリーと定義のペアの集合を対訳コーパスとみなし、Trans-Tokenizer [1] と同様に fast_align [10] を用いてサブワード単位、もしくは単語単位で目的言語・原言語のペアの確率と頻度を得る（図 1b）。この中から出現頻度の高いペアのみを抽出し、目的言語のサブワードと原言語のサブワードの対応表を作成する。

サブワード単位の対応付け　目的言語・原言語サブワード対応表を基にして、目的言語のサブワードと原言語のサブワードの 1 対多の対応表を作成する（図 1c 上）。目的言語のトークナイザーが持つサブワードのうち、対応する原言語のサブワードが設定されないものには原言語のモデルの UNK トークンを割り当てる。

単語単位の対応付け　Trans-Tokenizer と同様に単語内のすべてのサブワードを原言語・目的言語間に対応させたもの (多対多対応) と、順番通りにサブワードを対応させたもの (順序保存対応) を用意する（図 1c 下）。ここでも原言語のサブワードが紐づいていないサブワードには原言語のモデルの UNK トークンを割り当てる。このとき、それぞれの方法によって対応付けられるサブワードペアの頻度は単語ペアの頻度の半分とする。ここで、原言語と目的言語の単語ペア (w_s, w_t) に対するサブワード列を $S = (s_1, \ldots, s_m), T = (t_1, \ldots, t_n)$ とし、$m = |S|, n = |T|$ とおく。また、この単語ペアの頻度を $c(w_s, w_t)$ とする。多対多対応では、原言語と目的言語の単語内のすべてのサブワードを互いに対応付ける。これから得られるサブワードペア (s, t) の頻度は次のように定義される。

$$\forall s_i \in S \quad \forall t_j \in T : c_{\text{all}}(s, t) = c(w_s, w_t) \tag{1}$$

順序保存対応では、原言語と目的言語のサブワード列の順序を保持しながら対応付けを行う。まず以下のようにサブワード列を拡張する。

$$S' = (\underbrace{s_1, \ldots, s_1}_{n/\gcd(m,n)\text{ 個}}, s_2, \ldots, s_2, \ldots, s_m, \ldots, s_m)$$

$$T' = (\underbrace{t_1, \ldots, t_1}_{m/\gcd(m,n)\text{ 個}}, t_2, \ldots, t_2, \ldots, t_n, \ldots, t_n)$$

これから得られるサブワードペア (s, t) の頻度は次のように定義される。

$$\forall i \in \{1, \ldots, \text{lcm}(m, n)\} : c_{\text{in}}(s, t) = \frac{c(w_s, w_t)}{m/\gcd(m, n)} \tag{2}$$

まとめると、単語単位での対応付けを行うときのサブワードペア (s, t) の頻度は次のように定義される。

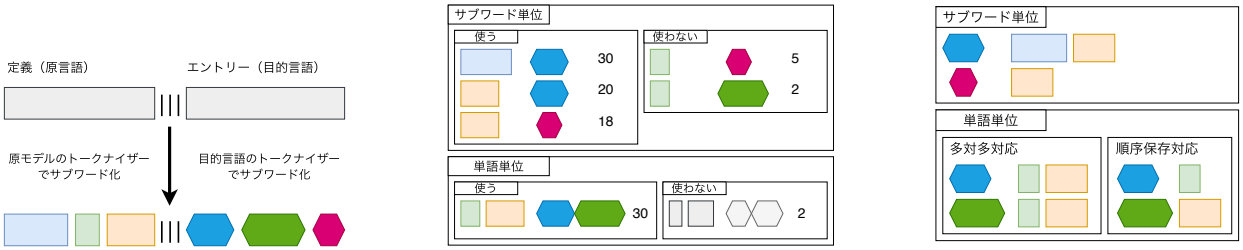
$$c(s, t) = \frac{1}{2}(c_{\text{all}}(s, t) + c_{\text{in}}(s, t)) \tag{3}$$

3.3 サブワード埋め込みの初期化

ここで、目的言語のあるサブワード t に対応する原言語のサブワードの集合を A_t、原言語のあるサブワード s 、目的言語のあるサブワード t に対する埋め込みの集合をそれぞれ E_s^S 、E_t^T とする。A_t に属するあるサブワード s の A_t における相対頻度は以下のように定義される。

$$f(s|t) = \frac{c(s, t)}{\Sigma_{x \in A_t} c(x, t)} \tag{4}$$

原言語のサブワード $s \in A_t$ の埋め込み $\boldsymbol{e}_s^S \in E^S$ を用いて、目的言語のサブワード t の埋め込み $\boldsymbol{e}_t^T \in E^T$



(a) 辞書のトークナイズ　(b) サブワードの確率と頻度　(c) サブワードの対応付け

図 1: 辞書を用いたサブワードの対応付け。サブワード単位・単語単位の対応に基づいて目的言語のサブワードに対応する原言語のサブワードの集合を構築する。ここでは出現頻度の閾値を 10 とした。

は以下のように対応する原言語のサブワードの埋め込みの加重平均として初期化される。

$$\boldsymbol{e}_t^T = \sum_{s \in A_t} f(s|t) \cdot \boldsymbol{e}_s^S \tag{5}$$

4　実験設定

英語のデータで事前学習された RoBERTa [1] を原モデルとして実験を行う。また、トークナイザーの訓練時には最終的なサブワード数が 1000 になるように指定した（付録 A を参照）。ドイツ語、日本語、古英語、ウイグル語、サンスクリット語、クメール語、満洲語を用いて、NER タスクで評価する。これらの言語とデータセットで用いられる言語は表 1 の通りである[2]。辞書データは、ドイツ語と日本語については WECHSEL [2] で使用されている MUSE [11] の目的言語 – 英語の二言語辞書[3]を用いる。ウイグル語とサンスクリット語とクメール語に関しては WECHSEL [2] で用いられた Wiktionary の目的言語 – 英語辞書[4] を用いる。満洲語については満洲語 – 英語辞書 [12] から抽出したデータ[5] を用いる。満洲語の辞書は複数語からなるエントリーや定義を含むが、その他の辞書は目的言語の単語 1 つに対して英単語 1 つが割り当てられている。

サブワード埋め込みの初期化の手法の性能を比較するために、FOCUS [3] を用いても実験する。FOCUS は単言語コーパスを用いるものであるので、Wikipedia のデータ[6]を用いる。ただし満洲語のデータは適当なデータがないため行わなかった。各言語のトークナイザーはこのデータセットから最大 10K の項目で訓練したものを用いる。各辞書に含まれる語彙数とテキストデータの総語数は表 2 の通りである。

表 1: 各言語とデータセットで使用されている文字

言語名	文字
ドイツ語	ラテン文字
日本語	漢字・かな・カナ
古英語	ラテン文字
ウイグル語	アラビア文字
サンスクリット語	デーヴァナーガリー
クメール語	クメール文字
満洲語	ラテン文字

表 2: 各言語における辞書の含まれる語彙数とテキストデータの語数

言語名	語彙数	語数
ドイツ語	101,997	1237M
日本語	25,969	119M
古英語	7,592	0.4M
ウイグル語	1,131	3M
サンスクリット語	5,282	3M
クメール語	5,656	2M
満洲語	21,620	–

転移させたモデルはファインチューニングをし、NER の性能で評価する。ここでは転移したモデルに対する LAPT は行わない。詳細は付録 B に記載する。NER データセットは WikiANN [13, 14] と ManNER [15] を用いる。

5　結果と分析

実験の結果は表 3 の通りである。転移元のモデルである RoBERTa の結果と提案手法の結果を比べると、アライメントの単位がサブワードか単語かによらず、サンスクリット語、クメール語、満洲語においては提案手法のほうが良い結果となった。サブワード単位で対応付けをする場合と単語単位で対応付けをする場合を比べると、クメール語と満洲語で

1）https://huggingface.co/FacebookAI/xlm-roberta-base/
2）満洲語は本来満洲文字で表記されるが、ここでは既存のデータセットに倣いラテン文字で転写されたものを扱う。
3）https://github.com/facebookresearch/MUSE
4）https://github.com/CPJKU/wechsel/tree/main/dicts
5）https://github.com/tyotakuki/manchuvocabdata
6）https://huggingface.co/datasets/graelo/wikipedia

表 3: WikiANN と ManNER に対する micro F1 スコア。Ours(word) は 単語単位 でアライメントをした結果を表し、 Ours(subword) は サブワード単位でアライメントをしたときの結果を表す。

	ドイツ語	日本語	古英語	ウイグル語	サンスクリット語	クメール語	満洲語
RoBERTa	89.61	75.33	**62.39**	**38.73**	51.48	27.58	73.52
XLM-R	**90.27**	**81.28**	37.59	28.30	48.85	34.78	65.32
FOCUS (XLM-R) w/o LAPT	89.42	78.40	32.10	7.69	31.90	24.71	–
UniBridge (XLM-R) [4]	–	–	43.20	–	42.67	29.74	–
Ours(word)@10 w/o LAPT	79.65	71.62	41.87	31.46	53.50	38.05	**92.41**
Ours(subword)@10 w/o LAPT	80.01	71.57	44.02	37.96	**59.11**	32.16	91.81
Ours(subword)@2 w/o LAPT	79.64	71.57	38.23	34.96	42.08	**40.18**	90.61

は単語単位のほうが性能が高いことがわかる。これについては第 5.1 節で論じる（付録 C も参照）。目的言語のサブワードの埋め込みの初期化に用いるための、原言語のサブワードとのペアの出現頻度の閾値の違いを見ると、閾値を 2 回に設定する場合はいずれの言語においても 10 回設定したときより性能が低下している（付録 D を参照）。

5.1 言語の特徴と性能の関係

まず、転移元の RoBERTa と比較したときの性能と各言語の特徴を考察する。性能が向上したサンスクリット語、クメール語、満洲語の 3 言語に共通する特徴として、英語と系統的に遠いということが挙げられる。また、サンスクリット語とクメール語については英語と全く異なる文字体系を持つ。一方で英語と系統的にも異なっており、異なる文字を扱っている日本語では RoBERTa のほうが性能が良い。この理由としては日本語の文字種の多さが考えられる。文字の種類が多いということは、辞書という限られた言語資源を用いてトークナイザーを訓練するだけではサブワードやその埋め込みは十分な表現力を得ることができない。このため日本語に対する提案手法の性能は限られていたと推察される。ウイグル語も日本語と同じように英語とは系統的に異なる言語であるが、 RoBERTa のほうが性能が良い。学習に用いた辞書の規模が小さいことは理由の一つであると考えられる。

次にサブワード単位の対応付けと単語単位の対応付けを比較する。単語単位のほうが性能が良い言語はクメール語と満洲語であった。この 2 言語に共通し、また日本語以外の他の言語とは共通しないような特徴として名詞の格変化がある。クメール語と満洲語に関しては名詞に格変化が存在しないため、単語単位でも十分にサブワードの表現が対応付けられ、またサブワード単位よりもノイズが入りにくいという利点があると考えられる。一方で名詞の格変化がある言語については、単語単位の場合は下流タスクの中で名詞の格変化に対応できない。これは辞書のエントリーには多くの場合、無標の表層形のみが含まれるからである。日本語もクメール語や満洲語と同様に名詞の格変化が存在しない言語であるが、たとえば漢字の組み合わせによる名詞の変化を弱い格変化と考えれば、日本語における単語単位の性能低下を説明できる。NER タスクで評価しているため、名詞の検知に影響を及ぼす名詞の曲用の有無によって性能が変化するのは理にかなっている。

以上の結果から、転移の性能と系統論的な差異や形態統語論的な特徴には関係性が見られる。

5.2 従来手法との比較

サブワード埋め込みの初期化の手法の一つである FOCUS と比較する。テキストデータを用いた LAPT を行わない場合に関しては提案手法の性能が一貫して FOCUS を上回っている。また、多くの低資源言語に対して高い性能を誇る UniBeidge と比較しても、提案手法のほうが優れていることがわかる。表 2 に示されるように、提案手法で用いるデータは少量であることから、たとえデータが少なかったとしても、辞書データは言語モデルの転移に有益であると言える。

6 おわりに

本稿では二言語辞書を用いたサブワードの埋め込み転移の新たな手法を提案した。実験の結果、提案手法は少ないデータでより多くのデータを用いる従来手法の性能を上回ることが示された。これは提案手法が低資源言語の言語モデルの性能を向上させるのに大いに役立つことを意味する。

参考文献

[1] François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP. In **First Conference on Language Modeling**, 2024.

[2] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13440–13454, Singapore, December 2023. Association for Computational Linguistics.

[4] Trinh Pham, Khoi Le, and Anh Tuan Luu. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3168–3184, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Benjamin Minixhofer, Edoardo Ponti, and Ivan Vulić. Zero-Shot Tokenizer Transfer. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, November 2024.

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[7] Meng Fang and Trevor Cohn. Model transfer for tagging low-resource languages using a bilingual dictionary. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 587–593, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.

[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, **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.

[10] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[11] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. **arXiv preprint arXiv:1711.00043**, 2017.

[12] Jerry Norman. **A comprehensive Manchu-English dictionary**. Harvard-Yenching Institute Monograph Series. Harvard University, Asia Center, Cambridge, MA, April 2013.

[13] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[14] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 151–164, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Sangah Lee, Sungjoo Byun, Jean Seo, and Minha Kang. ManNER & ManPOS: Pioneering NLP for endangered Manchu language. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 11030–11039, Torino, Italia, May 2024. ELRA and ICCL.

A　サブワードのタイプ数と性能

ここでは代表例としてドイツ語を取り上げる。表 4 からわかる通り、サブワードのタイプ数が増えたとしても性能の向上は限定的である。このことから、サブワードのタイプ数の増加はモデルの表現力を高めるのにはあまり役立たないことがわかる。

表 4: サブワード単位でアライメントを取る場合におけるトークナイザーがもつサブワードのタイプ数と性能の関係

サブワードのタイプ数	1000	30000
F1 スコア	80.01	80.36

B　実験の詳細設定

ファインチューニングに用いたパラメータは表 5 の通りである。実験には NVIDIA RTX3090、NVIDIA A6000 を用いた。

表 5: ファインチューニングのパラメータ

ハイパーパラメータ	値
seed	42
epoch	5, 25
learning rate	5e-5
batch size	64

C　辞書の種類と性能の関係

目的言語の単語一つに対して原言語の単語一つが定義される辞書では、単語単位での対応付けを行うときに、高頻度になることがないのは明らかである。満洲語以外の言語で、サブワード単位での対応付けを行ったほうが単語単位より性能が良いのはこのためである。

D　高頻度ペアに含まれるサブワードとその網羅率

表 6 はサブワード単位を用いた場合における、原言語のサブワードとのペアが高い頻度で出現した目的言語のサブワード、(UNK タグではない原言語のサブワードの埋め込みを用いて埋め込みが初期化された目的言語のサブワード) が学習した目的言語のトークナイザーの語彙に占める割合と、トークナイザーの学習に用いた辞書のエントリーをトークナイズしたときのすべてのサブワードに占める割合を示したものである。

サブワード埋め込みの初期化に用いる目的言語・原言語ペア出現頻度の閾値を 10 に設定した場合（threshold@10) を見ると、サンスクリット語とクメール語においては、埋め込みが UNK トークン以外で初期化されたサブワードが、辞書のエントリーのサブワード全体に対して他の言語より高い割合を占めていることがわかる。これらのサブワードが原言語のサブワードと強く対応付けられていることはサブワード単位での言語間の何らかの共通性の存在を示唆するものであるように思える。

表 3 と併せて見ると、転移先の性能が向上したサンスクリット語、クメール語、満洲語は、少なくとも辞書のエントリーのサブワード全体に対しては、より多くのサブワードが原言語のサブワードを用いて初期化されていることがわかる。目的言語のサブワードの埋め込みに用いる頻度の閾値が 10 の場合に関しては、辞書のエントリーのサブワード全体に対する有効なサブワードの割合が一定の割合を超えていることが、転移による性能向上に必要であるように思われる。一方で、閾値を 2 に設定したときはより多くの目的言語のサブワードの埋め込みが原言語のサブワードを用いて初期化されているが、クメール語を除けば下流タスクの性能は低下している。このことから、低頻度のペアを用いた埋め込みの初期化は多くの場合で転移先のモデルに悪影響を及ぼすと言える。

表 6: 原言語のサブワードとのペアが高い頻度で出現した目的言語のサブワードが目的言語のトークナイザーの語彙（タイプ頻度）に占める割合 (%) と、トークナイザーの学習に用いた辞書のエントリーをトークナイズしたときのすべてのサブワード（トークン頻度）に占める割合 (%)。threshold@10 は 10 回以上観測されたペア、threshold@2 は 2 回以上観測されたペアの目的言語のサブワード。

	ドイツ語	日本語	古英語	ウイグル語	サンスクリット語	クメール語	満洲語
threshold@10 / vocab	71.20	17.90	0.80	0.00	0.4	6.70	21.50
threshold@10 / subwords	90.72	36.99	2.54	0.00	26.13	81.24	45.65
threshold@2 / vocab	97.70	81.90	54.00	1.60	2.60	20.90	88.3
threshold@2 / subwords	99.92	92.62	68.80	2.55	58.97	90.70	96.95