# 言語間対応を考慮したバイリンガルサブワード分割

西田 祥人 [1]　二宮 崇 [2]　後藤 功雄 [2]
[1] 愛媛大学　[2] 愛媛大学大学院理工学研究科
{nishida@ai.cs., ninomiya@cs., goto.isao.fn@}ehime-u.ac.jp

## 概要

本研究では EM アルゴリズムを用いて，潜在変数として対訳サブワード間の対応関係を学習する手法を提案する．従来のニューラル機械翻訳では，言語ごとに独立して学習されたサブワード分割モデルを用いることが一般的である．しかし，従来手法は対訳関係を考慮していないため，原言語と目的言語で語内部の分割が不整合となり，翻訳モデルの学習を妨げる可能性がある．提案手法では，対訳コーパスからサブワード文対間のサブワードの対応関係を直接モデル化することで，言語間の分割の一貫性を向上させる．複数の機械翻訳タスクにおける評価実験の結果，提案手法が複数の言語対で翻訳精度を改善することを確認した．

## 1　はじめに

ニューラル機械翻訳（以下，NMT）では，事前に定義された語彙に依存しているため，翻訳時の原言語文に低頻度語や未知語が含まれると，翻訳性能が低下する．この語彙問題に対処するため，バイトペア符号化（以下, BPE）[1] や，ユニグラム言語モデル [2] に基づくサブワード分割が広く用いられており，これらの手法では各言語ごとに独立して分割モデルを学習している．また，複数言語のコーパスを統合して単一の分割モデルを学習する多言語 SentencePiece も用いられている [3].

しかし，これらの手法は対訳文対に基づく対応関係を直接的にモデル化するものではなく，対訳関係が反映されない．その結果，原言語と目的言語の間で語内部の分割が対応しなくなり，翻訳モデルの学習を妨げる可能性がある．例えば，日英翻訳において，“nonextended” と “延長されなかった” の対訳文があり，“nonextended” は “no next end ed”，“延長されなかった” は “延長 されな かった” と分割されたとする．ここで，NMT モデルが “next” を “延長” として学習してしまった場合，正しい翻訳結果が得られなくなる可能性がある．この問題に対処するため，対訳関係を考慮したサブワード分割 [4, 5, 6] が提案されている．出口ら [4] のサブワード分割は原言語側と目的言語側のうち，サブワード列のトークン数が多い方にトークン数を揃えるものであり，トークン長が近いとはいえ単語内部の分割が言語間で一貫する保証はない．Hiraoka ら [5] のバイリンガルサブワード分割は NMT モデルの学習が必要となり，サブワード分割および機械翻訳モデルの学習に大きなコストを要する．松井ら [6] のバイリンガルサブワード分割は対訳文対のサブワードの対応関係からサブワード列を取得するものであるが，対応関係がアライメントツールによって与えられているため，機械翻訳機の性能がアライメントツールに依存してしまう．

本研究では，潜在変数として対訳サブワード間の対応関係を学習することで，単語アライメントツールを使用せずにサブワード分割を行う新たなサブワード分割手法を提案する．提案手法では，ユニグラム言語モデルである SentencePiece [7] を用いて，対訳コーパス中の原言語文と目的言語文のサブワード分割候補をそれぞれ取得し，各サブワード列対のサブワードの対応関係をアライメント確率として学習する．ここで，この確率モデルにおけるアライメント確率は隠れ状態となるため，潜在変数付き確率モデルの学習に使われる EM アルゴリズムを用いる．ユニグラム言語モデルによる生起確率とアライメント確率を掛け合わせ，確率が最も大きくなるサブワード列対を訓練データとして使用する．翻訳時には目的言語文が存在しないため，目的言語側サブワードでアライメント確率の周辺化を行い，同様に確率が最も大きくなるサブワード列を評価データとして使用する．

16 種類の機械翻訳タスクにおける評価実験の結果，13 種類のタスクにおいて提案手法の BLEU が従来手法を上回る性能を達成した．

## 2 従来のサブワード分割手法

本節では，提案手法の前提となるユニグラム言語モデルに基づいたサブワード分割法 [2] について説明する．ユニグラム言語モデルでは，各サブワードが独立に生起すると仮定し，サブワード列の生起確率 $P_{\mathrm{U}}(\boldsymbol{x})$ を次式により表す．

$$P_{\mathrm{U}}(\boldsymbol{x}) = \prod_{i=1}^{I} P(u_i) \qquad (1)$$
$$\forall i \quad u_i \in V, \quad \sum_{u \in V} P(u) = 1$$

ただし，$\boldsymbol{x} = (u_1, \cdots, u_i, \cdots, u_I)$ はサブワード列であり，各 $u_i$ はサブワード集合 $V$ の要素である．サブワードの生起確率 $P(u)$ は，次式に表す周辺尤度 $L_{\mathrm{lm}}$ を最大化するように EM アルゴリズムによって推定される．

$$L_{\mathrm{lm}} = \sum_{n=1}^{N} \log P(X_n) = \sum_{n=1}^{N} \log \left( \sum_{\boldsymbol{x} \in S(X_n)} P_{\mathrm{U}}(\boldsymbol{x}) \right) \qquad (2)$$

ただし，$N$ は訓練データに含まれる対訳文対の数，$X_n$ はその $n$ 番目の原言語文または目的言語文であり，$S(X_n)$ は $X_n$ に対して生成可能なサブワード列の候補集合を表す．

モデルの学習後，文 $X$ に対する生起確率が最大となるサブワード列は次式によって算出される．

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in S(X)}{\operatorname{argmax}} P_{\mathrm{U}}(\boldsymbol{x}) \qquad (3)$$

また，$k$-best 分割候補も同様に $P_{\mathrm{U}}(\boldsymbol{x})$ に基づいて算出可能であり，提案手法ではこれらを用いてサブワード分割候補の集合を構築する．

## 3 提案手法

本節では，提案手法となる対訳文対のサブワードの対応関係を学習するサブワード分割法について説明する．サブワード文対の確率モデルの定義を与え（3.1 節），EM アルゴリズムによるアライメント確率の更新式を導出し（3.2 節），訓練データ及び評価データに対するサブワード分割を行う（3.3, 3.4 節）．

### 3.1 提案手法の確率モデル

原言語文 $X$ と目的言語文 $Y$ が与えられたときの，提案手法におけるサブワード分割のための確率モデルを次の式で定義する．

$$\begin{aligned} P(X,Y) &= \sum_{\boldsymbol{x} \in S(X)} \sum_{\boldsymbol{y} \in S(Y)} \sum_{a \in A(\boldsymbol{x},\boldsymbol{y})} P_{\mathrm{M}}(\boldsymbol{x},\boldsymbol{y},a) \\ &\approx \sum_{k,l} \sum_{a \in A(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)})} P_{\mathrm{M}}(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(l)},a) \end{aligned} \qquad (4)$$

ただし，$X$ に対するサブワード列の候補集合 $S(X)$ のうち，サブワード列の生起確率 $P_{\mathrm{U}}(\boldsymbol{x})$ が高い top-$K$ 個をそれぞれ $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(k)}, \cdots, \boldsymbol{x}^{(K)}$, $Y$ に対するサブワード列の候補集合 $S(Y)$ のうち，サブワード列の生起確率 $P_{\mathrm{U}}(\boldsymbol{y})$ が高い top-$L$ 個をそれぞれ $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(l)}, \cdots, \boldsymbol{y}^{(L)}$ とする．ここで，$A(\boldsymbol{x},\boldsymbol{y})$ は原言語文のサブワード列 $\boldsymbol{x}$ の各サブワードと目的言語文のサブワード列 $\boldsymbol{y}$ の各サブワードとの全ての可能なアライメント集合を表す．さらに，$a \in A(\boldsymbol{x},\boldsymbol{y})$ は一つの具体的なサブワードアライメントを表す．$P_{\mathrm{M}}$ は原言語文のサブワード列 $\boldsymbol{x}$ と目的言語文のサブワード列 $\boldsymbol{y}$，およびそれらの間のアライメント $a$ に対する確率モデルであり，次の式で定義する．

$$P_{\mathrm{M}}(\boldsymbol{x},\boldsymbol{y},a) = P_{\mathrm{U}}(\boldsymbol{x}) P_{\mathrm{U}}(\boldsymbol{y}) \prod_{(u,v) \in a} \alpha_{uv} \qquad (5)$$

ただし，$\alpha_{uv}$ は原言語側サブワード $u$ と目的言語側サブワード $v$ が対応する確率である．松井ら [6] のバイリンガルサブワード分割では，$a$ はアライメントツールによって推定された，原言語文のサブワード列 $\boldsymbol{x}$ と目的言語文のサブワード列 $\boldsymbol{y}$ の間のアライメントを表す．一方，提案手法では原言語側サブワードと目的言語側サブワードの間の対応関係そのものを潜在変数として導入し，それらを直接学習する点で異なる．

### 3.2 アライメント確率 $\alpha_{uv}$ の学習

EM アルゴリズムを用いてアライメント確率 $\alpha_{uv}$ を求める．式 5 を使って $Q$ 関数を計算すると，次の式のようになる．

$$Q = \sum_{n,k,l} \sum_{a \in A(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)})} \frac{P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)},a) \log P_{\mathrm{M}}^{\mathrm{new}}(\boldsymbol{x}_n^{(k)},\boldsymbol{y}_n^{(l)},a)}{\sum_{k',l'} \sum_{a' \in A(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')})} P_{\mathrm{M}}^{\mathrm{old}}(\boldsymbol{x}_n^{(k')},\boldsymbol{y}_n^{(l')},a')} \qquad (6)$$

$\alpha_{uv}^{\mathrm{new}}$ に関して式 6 の $Q$ 関数を最大化することで，$\alpha_{uv}^{\mathrm{new}}$ の更新式を得る．

$$\alpha_{uv}^{\mathrm{new}} = \frac{\sum_{n,k,l} E_{nkluv}}{\sum_{u' \in V'_{\mathrm{src}}} \sum_{v' \in V'_{\mathrm{tgt}}} \sum_{n,k,l} E_{nklu'v'}} \qquad (7)$$

$$E_{nkluv} \approx \frac{\left(P_{\mathrm{U}}(\boldsymbol{x}_n^{(k)})P_{\mathrm{U}}(\boldsymbol{y}_n^{(l)}) \prod_{u\in\boldsymbol{x}^{(k)}} \sum_{v\in\boldsymbol{y}^{(l)}} \alpha_{uv}^{\mathrm{old}}\right) C_{nkluv}}{\sum_{k',l'} P_{\mathrm{U}}(\boldsymbol{x}_n^{(k')})P_{\mathrm{U}}(\boldsymbol{y}_n^{(l')}) \prod_{u\in\boldsymbol{x}^{(k')}} \sum_{v\in\boldsymbol{y}^{(l')}} \alpha_{uv}^{\mathrm{old}}} \quad (8)$$

ただし，$V'_{\mathrm{src}}$ は原言語側のサブワード集合，$V'_{\mathrm{tgt}}$ は目的言語側のサブワード集合，$C_{nkluv}$ は，$n$ 番目の文の $\boldsymbol{x}_n^{(k)}$ と $\boldsymbol{y}_n^{(l)}$ のサブワード文対にサブワード $u$ と $v$ が同時に出現する回数である．

### 3.3 訓練データのサブワード分割

訓練データの各文対 $X, Y$ に対して，アライメント確率に基づき対応関係が最大となるサブワード列 $\boldsymbol{x}^{(\hat{k})}, \boldsymbol{y}^{(\hat{l})}$ を次式に従って算出し，これをサブワード文対として採用する．

$$\hat{k},\hat{l} = \underset{k,l}{\operatorname{argmax}}\, P_{\mathrm{U}}(\boldsymbol{x}^{(k)})P_{\mathrm{U}}(\boldsymbol{y}^{(l)}) \prod_{(u,v)\in a} \alpha_{uv} \quad (9)$$

### 3.4 評価データのサブワード分割

評価データに対するサブワード分割では，目的言語文が存在しないため，アライメント確率を次式のように目的言語側サブワードで周辺化することによって，原言語側サブワードの確率を求める．

$$\alpha'_u = \sum_{v\in V_{\mathrm{tgt}}} \alpha_{uv} \quad (10)$$

評価データの各文 $X$ に対して，次式に従ってサブワード列 $\boldsymbol{x}^{(\hat{k})}$ を求め，サブワード文として採用する．

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P_{\mathrm{M}'}(\boldsymbol{x}^{(k)}) \quad (11)$$

$$P_{\mathrm{M}'}(\boldsymbol{x}) = P_{\mathrm{U}}(\boldsymbol{x}) \prod_{u\in\boldsymbol{x}} \alpha'_u \quad (12)$$

## 4 評価実験

提案手法の有効性を検証するため，6 種類の異なる言語対（en-ja, ja-zh, en-de, en-hi, en-id, en-th）を対象に，従来手法（ユニグラム言語モデル）との比較による機械翻訳の評価実験を行った．また，en-ja については，データ分布の異なる 3 種類のデータセットを使用し，同様に従来手法との比較実験を行い，データセット間での性能差を分析した．

### 4.1 データセット

en-ja のデータセットには WAT Asian Scientific Paper Excerpt Corpus（以下，ASPEC）[8] 英日・日英翻訳タスク, 京都フリー翻訳タスク（以下，KFTT）[9] , Wikimatrix v1（以下，Wikimatrix）[10] を使用した．ja-zh については，ASPEC 日中・中日翻訳タスクを使用した．en-de については，訓練データに WMT18 News Commentary v13（以下，WMT18）[1] ，検証データに WMT17 testsets, 評価データに WMT18 testsets を使用した．en-hi と en-id については，Wikimarix を使用した．en-th については，大規模英語-Thai 並列コーパス（以下，scb-mt-en-th）[11] を使用した．また，Wikimatrix と scb-mt-en-th の検証データには flores200 [2] [12] dev set，評価データには flores200 devtest set を使用した．具体的なデータセットの構成を A 節の表 4 に示す．

### 4.2 実験設定

ユニグラム言語モデルによるサブワード列候補集合を得るために，SentencePiece [7] を使用した．ユニグラム言語モデルの学習は，原言語側と目的言語側で独立して行い，辞書サイズはどちらも 16,000 とした．サブワードの候補数は原言語側と目的言語側それぞれ，ユニグラム言語モデルによるサブワードの生起確率が高い上位 10 通り ($K$=$L$=10) とした．従来手法と提案手法を用いてサブワード分割を行い，それぞれの分割結果で訓練した NMT モデルの性能を比較した．

NMT モデルには Fairseq [13] を使用し，Transformer base [14] モデルを使用した．全ての NMT モデルにおいて，パラメータの最適化には Adam [15] , 学習率は 1e-4, バッチサイズは 128 とし，その他のパラメータは Fairseq のデフォルトの値を使用した．学習は 30 エポックで終了させ，各エポックのモデルの内，検証データ上で最も SacreBLEU [16] の精度が高かったものを利用して評価データの翻訳を行った．

翻訳性能の評価には，SacreBLEU と COMET scores [3] [17] を使用した．SacreBLEU の flores200 のトークナイズには flores200 を，日本語のトークナイズには ja-mecab [18] を，中国語のトークナイズには zh を，英語とドイツ語のトークナイズには 13a を使用した．実験はランダムシードを変えて 3 度行い，その平均を実験結果とした．

1） https://www.statmt.org/wmt18/translation-task.html
2） https://github.com/facebookresearch/flores/tree/main/flores200
3） https://huggingface.co/Unbabel/wmt22-comet-da

**表 1** BLEU による自動評価の結果

| | ASPEC | | | | WMT18 | | Wikimatrix | | | | | | scb-mt-en-th | | KFTT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | en-ja | ja-en | ja-zh | zh-ja | en-de | de-en | en-hi | hi-en | en-id | id-en | en-ja | ja-en | en-th | th-en | en-ja | ja-en |
| 従来手法 | 27.2 | 27.0 | 35.4 | 28.9 | 21.4 | 21.7 | 22.0 | 19.9 | **41.9** | **36.5** | 19.3 | 20.9 | 28.3 | 17.2 | 22.0 | 20.9 |
| 提案手法 | **27.6** | **27.5** | **35.5** | **29.2** | **22.0** | **21.8** | 22.0 | **20.2** | 41.6 | 36.0 | **20.3** | **21.2** | **28.8** | **17.3** | **22.8** | **21.3** |

**表 2** COMET による自動評価の結果

| | ASPEC | | | | WMT18 | |
|---|---|---|---|---|---|---|
| | en-ja | ja-en | ja-zh | zh-ja | en-de | de-en |
| 従来手法 | **0.8882** | 0.8182 | 0.8675 | 0.9049 | 0.6482 | 0.6650 |
| 提案手法 | 0.8880 | **0.8195** | **0.8680** | **0.9055** | **0.6517** | **0.6676** |

| | Wikimatrix | | | | | |
|---|---|---|---|---|---|---|
| | en-hi | hi-en | en-id | id-en | en-ja | ja-en |
| 従来手法 | **0.6215** | 0.7403 | **0.8735** | **0.8395** | 0.8321 | 0.8037 |
| 提案手法 | 0.6196 | **0.7439** | 0.8721 | 0.8375 | **0.8354** | **0.8064** |

| | scb-mt-en-th | | KFTT | |
|---|---|---|---|---|
| | en-th | th-en | en-ja | ja-en |
| 従来手法 | 0.7576 | **0.7519** | 0.8102 | **0.7576** |
| 提案手法 | **0.7629** | 0.7509 | **0.8137** | 0.7554 |

**表 3** 提案手法の翻訳が改善した例

| | 分割結果 | 翻訳結果 |
|---|---|---|
| 正解データ | quilibrious<br>interval disorder | 平衡間隔失調 |
| 従来手法 | _ qui lib r ious<br>_ interval _ disorder | 巧 妙 な 区間 障害 |
| 提案手法 | _ qui lib ri ous<br>_ interval _ disorder | 平衡 間隔 障害 |

## 4.3 実験結果

表 1 に BLEU による自動評価の結果を示す．実験結果から分かる通り，提案手法は 16 種類の機械翻訳タスクのうち，13 種類のタスクにおいて，従来手法より性能が改善されている．また，分かち書きされていない言語を含む機械翻訳タスクにおいては，すべてのタスクにおいて従来手法を上回る性能改善が確認された．特に，分かち書きされた言語から分かち書きされていない言語方向への翻訳の性能が大幅に改善されていた．これは，提案手法によって分かち書きされていない言語において，従来のような不自然な分割が解消されたことで，対訳文対のサブワードが正しい対応関係を学習できるようになったからと考えられる．

表 2 に COMET による自動評価の結果を示す．COMET においては，16 種類の機械翻訳タスクのうち 10 種類のタスクにおいて，提案手法が従来手法を上回る結果が得られたもの，大幅な性能改善は確認されなかった．BLEU の性能が向上した一方，COMET では大幅な改善が見られなかったことから，文全体の意味的な品質には影響を与えず，語彙選択の改善に寄与したと考えられる．

## 4.4 分析

表 3 に提案手法を適用することによって翻訳が改善した例を示す．従来手法では分割がうまくいっておらず間違った翻訳がされているが，提案手法では分割が改善し，より正解に近い翻訳となっている．

訓練データの中で，従来手法と提案手法で分割が一致する割合を確認したところ，原言語側では一致率が高い一方，目的言語側では低い傾向が見られた．具体的な数値について，B 節の表 5 に従来手法と提案手法で分割が一致する割合を示す．これは，式 8 における $\prod_{u \in \boldsymbol{x}^{(k)}} \sum_{v \in \boldsymbol{y}^{(l)}} \alpha_{uv}^{\text{old}}$ が，原言語側と目的言語側で非対称であるためである．探索空間が上位 $K$ 個の分割候補に制限されるため，原言語側の分割は主にユニグラム言語モデルの尤度によって決定される．その結果，モデルは高確率な原言語側トークンを優先しつつ，目的言語側は原言語側トークンに合わせて分割を柔軟に変えるためであると考えられる．

# 5 まとめ

本研究では，EM アルゴリズムを用いて，潜在変数として対訳サブワード間の対応関係を学習する，新たなサブワード分割手法を提案した．実験の結果，提案手法を用いることで従来の分割手法と比較して翻訳性能が改善し，その有効性が確認された．今後の課題として，本手法の多言語への拡張が挙げられる．現在は 2 言語間（バイリンガル）での学習に限られているが，これを多言語間のサブワード分割へと適用することで，提案手法の有効性を検証したい．

## 謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）および JSPS 科研費 JP24K15071 の助成を受けたものです.

## 参考文献

[1] 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.

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

[3] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 726–742, 2020.

[4] Hiroyuki Deguchi, Masao Utiyama, Akihiro Tamura, Takashi Ninomiya, and Eiichiro Sumita. Bilingual subword segmentation for neural machine translation. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4287–4297, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Joint optimization of tokenization and downstream model. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 244–255, Online, August 2021. Association for Computational Linguistics.

[6] 松井大樹, 二宮崇, 田村晃裕. バイリンガルサブワード分割のための EM アルゴリズム. 言語処理学会 第 29 回年次大会 発表論文集, pp. 1469–1473, 2023.

[7] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[8] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association.

[9] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[10] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1351–1361, Online, April 2021. Association for Computational Linguistics.

[11] Lalita Lowphansirikul, Charin Polpanumas, Attapol T Rutherford, and Sarana Nutanong. scb-mt-en-th-2020: A large english-thai parallel corpus. 2020.

[12] James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. No language left behind: Scaling human-centered machine translation. 2022.

[13] 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. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[15] Diederik P Kingma. Adam: A method for stochastic optimization. 2014.

[16] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[17] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[18] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

# 付録

## A　データセットの詳細

**表 4**　データセットの構成

| | 訓練 | 検証 | 評価 |
|---|---|---|---|
| ASPEC (en-ja) | 1,000,000 | 1,790 | 1,812 |
| ASPEC (ja-zh) | 672,315 | 2,090 | 2,107 |
| WMT18 (en-de) | 284,246 | 3,004 | 2,998 |
| Wikimatrix (en-hi) | 231,459 | 997 | 1,012 |
| Wikimatrix (en-id) | 1,019,170 | 997 | 1,012 |
| Wikimatrix (en-ja) | 851,706 | 997 | 1,012 |
| scb-mt-en-th (en-th) | 988,259 | 997 | 1,012 |
| KFTT (en-ja) | 440,288 | 1,166 | 1,160 |

## B　分析の詳細

**表 5**　従来手法と提案手法で分割が一致する割合 (%)

| | ASPEC | | | | WMT18 | | scb-mt-en-th | |
|---|---|---|---|---|---|---|---|---|
| | en-ja | ja-en | ja-zh | zh-ja | en-de | de-en | en-th | th-en |
| 原言語側 | 98.4 | 98.1 | 98.1 | 97.4 | 99.1 | 98.5 | 97.5 | 95.3 |
| 目的言語側 | 29.8 | 82.6 | 24.2 | 28.9 | 70.6 | 33.4 | 18.6 | 66.7 |

| | Wikimatrix | | | | | | KFTT | |
|---|---|---|---|---|---|---|---|---|
| | en-hi | hi-en | en-id | id-en | en-ja | ja-en | en-ja | ja-en |
| 原言語側 | 70.9 | 92.8 | 85.7 | 87.7 | 95.3 | 97.9 | 98.6 | 98.5 |
| 目的言語側 | 22.9 | 52.0 | 54.6 | 33.0 | 31.5 | 36.8 | 38.0 | 65.7 |