

# フィールドワークデータによるジンポー語機械翻訳

田口 智大<sup>1</sup> 倉部 慶太<sup>2</sup> 坂井 優介<sup>3</sup> Rita Seng mai Nbanpa<sup>4</sup>

<sup>1</sup>University of Notre Dame <sup>2</sup>東京外国語大学アジア・アフリカ言語文化研究所

<sup>3</sup>奈良先端科学技術大学院大学 <sup>4</sup>University of Washington

ctaguchi@end.edu kurabe@aa.tufs.ac.jp

sakai.yusuke.sr9@is.naist.jp sengmai@uw.edu

## 概要

本研究では、フィールドワークによって収集されたデータを用いたジンポー語の機械翻訳システムを開発する。ジンポー語は主にミャンマー北東部のカチン州でカチン人によって話される言語であり、言語資源が極めて限られているため、機械翻訳などの言語処理タスクの応用が未だに進んでいない。そこで、本論文では、言語学者がカチン州でのフィールドワークで収集した民話と、母語話者によるその翻訳を対訳データとして用いて、英語・ジンポー語の機械翻訳システムを開発する。実験の結果、フィールドワークに基づく対訳データは少量ながら、ジンポー語から英語への機械翻訳では、民話ドメインにて最大約 10.7 ポイント、会話ドメインにて最大約 7.9 ポイントの BLEU スコア上昇が確認された。この結果は、低資源機械翻訳におけるフィールドワークデータの活用の可能性と重要性を示している。

## 1 はじめに

ジンポー語 (Jinghpaw, Jinghpo, Jingpo, Jingpho, 景頗語; ISO639-2: kac) は、シナ・チベット語族チベット・ビルマ語派に属する言語であり、主にミャンマー北部カチン州などにおいて、少数民族であるカチン人によって話されている。当言語はカチン州における多様な民族の共通語としての役割を果たしており、話者は百万人程度と推定されている。ジンポー語の基本語順は SOV であり、格助詞や後置詞を用いて文法関係を表すなど、日本語と共通する文法的特徴を多く持つ [1]。正書法としてはラテン文字を用いる。高・中・低・下降の四種の声調が区別されるが、これらの声調と声門閉鎖音は正書法では書かれない。以下にジンポー語の文例を正書法表記、音韻表記、註解、日本語訳を含むグロスで示す。

(1) *Gwi gaw nyau hpe kawa ai.*

*g̀wi g̀aw ʔnau p<sup>h</sup>eʔ gawá ai*

犬 は 猫 を 噛む 終止形

‘犬は猫を噛んだ。’

ジンポー語はいわゆる低資源言語であり、言語処理技術の応用やサポートは進んでいないのが現状である。そこで、本研究では、言語学者、自然言語処理研究者、そして母語話者が協働して、フィールドワークの成果として編纂されたデータをもとに、機械翻訳システムを開発する。実験の結果、NLLB [2] をベースとして、フィールドワークの成果として出版されたデータを用いて訓練することで、ジンポー語・英語間の機械翻訳の精度が民話ドメインにおいて最大約 10.7 ポイント、会話ドメインにて最大約 7.9 ポイントの BLEU スコア上昇が見られた。

## 2 関連研究

### 2.1 低資源機械翻訳

深層学習技術の進展、計算資源の増大、そして膨大なデータの蓄積に伴い、機械翻訳をはじめとする多くの自然言語処理タスクは高い精度で解かれるようになった。しかしながら、多くのタスクの性能はデータの質と量に依存しており、機械翻訳の精度向上においては特に対訳文コーパスのサイズが重要な因子となる [3]。したがって、英語をはじめとする大言語の間での機械翻訳では高い性能が達成されていても、低資源言語をソースまたはターゲットとする翻訳では依然として苦戦することが報告されている [4, 5]。

近年では、言語対につき一つの機械翻訳モデルを用意するのではなく、任意の言語対 ( $L_1, L_2$ ) を一つのモデルで翻訳する多言語機械翻訳モデルの開発の研究が進んでいる。例えば、M2M-100 はサポート

されている 100 言語のうちの任意の言語間での機械翻訳を可能にし [6], NLLB はさらにこれを推し進めて 200 言語をサポートしている [2]. 加えて, 多くの自然言語処理タスクにおいて汎化性能を示している巨大言語モデルも機械翻訳を行うことが可能である [7, 8, 9]. 特に, 文脈内学習 [10] を用いた機械翻訳は, 極めて限られたデータしか存在しない言語の機械翻訳にも有効であることが示されてきた [11].

## 2.2 ジンポー語と自然言語処理

上述したような多言語機械翻訳の発展に伴って, ジンポー語を含むより多くの言語をカバーした対訳文のデータセットが公開されている. NLLB データセット<sup>1)</sup>は現状公開されているジンポー語の対訳データのなかで最大であり, 約 100 万文の対訳文を含む [2]. その多くは聖書の公開対訳データをソースとしていると見られる. しかしながら, NLLB データセットの対訳文はウェブクロウリングに基づいて自動的にアラインメントが取られている影響で, 部分的にアラインメントのミスによる対訳の不一致や, 自動言語識別のミスによる他言語の混入などが見られる.

FLORES [12] は, 多言語機械翻訳の性能を統一的に評価するために開発されているデータセットで, Wikinews<sup>2)</sup>や Wikibooks<sup>3)</sup>などをソースとする英文を原文として, ジンポー語を含む 200 言語以上に翻訳されている. このデータセットは専ら評価のために開発されているため, 訓練データは存在せず, 公開されている検証データ (dev) と擬似テストデータ (devtest), ならびに非公開のテストデータ (test) から構成されている.

ジンポー語をサポートしている機械翻訳モデルとして, パラメータが公開されているものは前述の NLLB モデルが挙げられる. これは, 前述の同名の NLLB データセットをもとに訓練されている. また, パラメータ非公開のモデルとしては, 2024 年に Google 翻訳がジンポー語を含む 110 言語のサポートを新たに開始した<sup>4)</sup>. なお, GPT4 [13] などの大規模言語モデルも多少ながらジンポー語の知識を有すると考えられるが, どの程度の知識があるのかはまだ定量的に検証されていない.

1) <https://opus.nlpl.eu/NLLB/corpus/version/NLLB>

2) <https://www.wikinews.org>

3) <https://www.wikibooks.org>

4) <https://blog.google/products/translate/google-translate-new-languages-2024>

表 1 本実験で用いるデータセットの概要. 数値の単位は文.

|           | 訓練      | 検証    | テスト   | 合計      |
|-----------|---------|-------|-------|---------|
| PARADISEC | 38,113  | 755   | 743   | 39,611  |
| 辞書例文      | 3,151   | —     | —     | 3,151   |
| 会話文       | —       | —     | 222   | 222     |
| NLLB      | 121,081 | —     | —     | 121,081 |
| FLORES    | —       | 997   | 1,011 | 2,008   |
| 合計        | 162,345 | 1,752 | 1,976 | 166,073 |

ジンポー語の機械翻訳モデルを開発するにあたって障壁となるのが, ジンポー語の属するチベット・ビルマ語派のどの言語も低資源であるため, 類似言語からの転移学習 [14] の有効性が見込めない点である. また, 上質な対訳データが極めて限られている点も精度向上の妨げとなっている. 一方で, ジンポー語のデータは, NLLB データセットの他に, ジンポー語を専門とする言語学者とジンポー語母話者によって作成された民話データ (PARADISEC) [15, 16] や, 文法書 [1], 辞書 [17], 読本 [18] が出版されている.

## 3 実験

本節では, ジンポー語機械翻訳モデルを訓練する上で使用した実験設計について述べる.

### 3.1 データ

実験で用いるデータセットは, PARADISEC のジンポー語民話データ, 『ジンポー語用例辞典』 [17] に記載されている例文, 『ジンポー語読本』 [18] の会話文, NLLB データセット, FLORES データセットである (表 1 参照). このうち, 辞書例文と会話文はサンプル数が少ないため, 前者を全て訓練データ, 後者を全てテストデータとして用いる. また, NLLB データセットは質が担保されていないことから, 文献 [19] の手法をもとに重複文および文長が二言語間で著しく異なる対訳文を取り除いた後, 残り全てを訓練データとして使用する. フィルタリング前の NLLB データセットの対訳文は 1,003,100 文, フィルタリング後は 121,081 文となった. また, FLORES データセットは本来評価用であるため, dev セットを検証データとして, devtest セットをテストデータとして用い, 訓練データには一切用いない. 各データセットのサンプルの例を表 4 に示す.

## 3.2 モデル訓練

今回の実験では、NLLB-600M<sup>5)</sup>とNLLB-1.3B<sup>6)</sup>の二つのモデルをベースラインとして、上述のジンポー語・英語の対訳文を用いてファインチューニングを行う。本実験では、特に、ファインチューニングに用いるデータの違いによる性能の差を比較することを目的とする。翻訳の言語方向は、ジンポー語から英語、および英語からジンポー語の両方向について、それぞれ別にファインチューニングを行う。訓練目標は、元のNLLBモデル同様、交差エントロピー誤差の最小化とする。どのモデル訓練においても、訓練サンプルのバッチサイズは8、学習率は0.0001、ウォームアップのステップ数は500、エポック毎の学習率乗数を0.9、最大エポック数は20とする。その他のハイパーパラメータはNLLBモデルのデフォルトの設定に従う。検証データおよびテストデータでの評価に用いる指標としては、BLEU[20]とChrF++[21]を用いる。最終的なテストでは、検証データのBLEUスコアが最も高いチェックポイントを評価に使用する。

## 4 結果

ジンポー語 → 英語の実験結果を表2に、英語 → ジンポー語の実験結果を表3に示す。全体的な傾向としては、1.3Bモデルの方がおおむねスコアが若干高いが、大きな差は特に見られなかった。また、ファインチューニングによって、BLEUスコアが民話ドメインで最大約10.7ポイント、会話ドメインにて最大約7.9ポイント上昇した。一方で、ファインチューニング後にFLORESベンチマークにおけるスコアが最大約8.5ポイント減少した。NLLB-1.3BモデルをPARADISECデータセットと辞書例文データセットを用いてファインチューニングしたモデルの評価データにおける出力結果の例を表4に示す。

## 5 議論

**FLORES ドメインの忘却** 結果で見たように、提案したデータを用いて訓練したモデルは、PARADISECと日常会話文のテストデータでは精度が改善したが、一方でFLORESのテストデータでは精度が悪化した。FLORESはニュースや科学的な内容を含んだドメインのデータセットであることか

5) <https://huggingface.co/facebook/nllb-200-distilled-600M>

6) <https://huggingface.co/facebook/nllb-200-1.3B>

ら、上記のデータセットによるファインチューニングによって、そのような専門的なドメインの翻訳に弱くなっていると考えられる。なお、このドメインの忘却は、ベースラインのモデル訓練で使用されたNLLBデータセットを用いても防がれなかった(表3, 2のP+D+N列を参照)。

**民話の文体の影響** 民話データを用いた訓練は自動評価指標のスコア上は有効であったことが示されたが、民話に特有の文体が民話以外のドメインの翻訳にも散見された。主な例は伝聞の終助詞の *da* (〜だそうだ、らしい) であり、実際に表4のFLORESの出力例でも確認できる。この問題を解決するには、より多様でバランスの良い訓練データが必要であると考えられる。

**文化に根ざした高品質な評価データセットの重要性** 今回評価に用いたFLORESデータセットは、ニュースなどの英文を原文として統一的に多言語に翻訳されているという性質上、固有名詞が頻出する傾向があるほか、英文で使用されている概念が対象言語の語彙に存在しないケースが存在する。そのため、FLORESのジンポー語の評価文においても、不自然な翻訳が散見された。また、FLORESの開発プロセスでは質の担保を目的とした自動チェックおよび人手のレビューが行われているものの、今回の実験では、FLORESジンポー語版中の誤字や語訳が確認された。多言語で統一された評価データであるFLORESは多言語機械翻訳のベンチマークとして有意義なものであるが、翻訳の質や英語中心主義的な文選定といった問題点を有している。民話や日常会話文といった、対象言語で実際に用いられる自然な文も用いた、多元的な評価を行うことが重要である。

## 6 結論

本研究は、NLLBモデルを用いた英語・ジンポー語機械翻訳の開発を行なった。ジンポー語は上質な対訳文データの少ない低資源言語であるが、言語学者がフィールドワークに基づいて母語話者コミュニティと協力して編纂したデータを用いることで、民話・会話ドメインの翻訳精度が最大約10.7ポイント(BLEU)上昇した。この結果は、低資源の多言語機械翻訳におけるフィールドワークデータの活用の有効性を示した。加えて、対象言語の文化に基づいたドメインでの性能評価の重要性を論じた。

表2 ジンポー語 → 英語の実験結果. 数値の左側は BLEU スコアで, 右側は chrF++スコア. ベースラインはファインチューニング前の NLLB モデルである. D は辞書例文データセット (Dictionary), P は PARADISEC データセット, N は NLLB データセットを表す.

|           |                  | ベースライン                      | D             | P                    | P+D                         | P+D+N                       |
|-----------|------------------|-----------------------------|---------------|----------------------|-----------------------------|-----------------------------|
| NLLB-600M | PARADISEC (test) | 2.32 / 20.34                | 4.58 / 22.52  | <b>12.58</b> / 31.35 | 12.55 / <b>31.91</b>        | 12.86 / 31.91               |
|           | FLORES (devtest) | <b>12.77</b> / <b>35.47</b> | 9.73 / 31.44  | 5.91 / 27.34         | 6.04 / 27.12                | 6.88 / 30.03                |
|           | 日常会話文            | 17.35 / 32.42               | 18.08 / 34.43 | 22.41 / 39.78        | 22.30 / 40.38               | <b>23.35</b> / <b>41.67</b> |
|           | 平均               | 10.81 / 29.41               | 10.80 / 29.46 | 13.63 / 32.82        | 13.63 / 33.14               | <b>14.36</b> / <b>34.54</b> |
| NLLB-1.3B | PARADISEC (test) | 2.29 / 19.72                | 4.23 / 20.44  | 12.95 / 32.11        | 12.95 / 31.45               | <b>13.40</b> / <b>32.19</b> |
|           | FLORES (devtest) | <b>13.95</b> / <b>37.27</b> | 10.90 / 30.56 | 5.48 / 26.59         | 5.41 / 26.66                | 4.87 / 27.25                |
|           | 日常会話文            | 16.66 / 33.44               | 24.57 / 41.71 | 18.09 / 34.84        | <b>23.49</b> / <b>40.96</b> | 22.44 / 39.81               |
|           | 平均               | 10.97 / 30.14               | 13.23 / 30.90 | 12.17 / 31.18        | <b>13.95</b> / <b>33.02</b> | 13.57 / 33.08               |

表3 英語 → ジンポー語の実験結果. 数値の左側は BLEU スコアで, 右側は chrF++スコア. ベースラインはファインチューニング前の NLLB モデルである. D は辞書例文データセット (Dictionary), P は PARADISEC データセット, N は NLLB データセットを表す.

|           |                  | ベースライン                      | D             | P                           | P+D                  | P+D+N                       |
|-----------|------------------|-----------------------------|---------------|-----------------------------|----------------------|-----------------------------|
| NLLB-600M | PARADISEC (test) | 3.67 / 25.32                | 4.61 / 26.77  | 10.99 / 32.92               | 11.40 / 33.87        | <b>11.43</b> / <b>34.35</b> |
|           | FLORES (devtest) | <b>9.68</b> / <b>34.43</b>  | 8.06 / 31.84  | 1.65 / 21.76                | 2.50 / 24.17         | 3.94 / 24.02                |
|           | 日常会話文            | 13.05 / 39.61               | 17.10 / 41.36 | 18.90 / 44.28               | <b>21.75</b> / 44.82 | 21.17 / <b>45.68</b>        |
|           | 平均               | 8.8 / 33.12                 | 9.92 / 33.32  | 10.51 / 32.99               | 11.88 / 34.29        | <b>12.18</b> / <b>34.68</b> |
| NLLB-1.3B | PARADISEC (test) | 4.15 / 25.94                | 4.48 / 26.05  | 12.19 / 35.04               | 12.30 / 35.75        | <b>12.78</b> / <b>35.87</b> |
|           | FLORES (devtest) | <b>10.97</b> / <b>36.00</b> | 6.12 / 29.68  | 2.26 / 24.09                | 3.74 / 27.45         | 3.60 / 25.16                |
|           | 日常会話文            | 16.95 / 43.02               | 20.17 / 44.34 | <b>24.41</b> / <b>48.21</b> | 22.75 / 48.01        | 23.01 / 47.82               |
|           | 平均               | 10.69 / 34.99               | 10.26 / 33.36 | 12.95 / 35.78               | 12.93 / 37.07        | <b>13.13</b> / <b>36.28</b> |

表4 NLLB-1.3B をもとに, PARADISEC データセットと辞書例文データセットを用いてファインチューニングしたモデルの出力例. kac はジンポー語, eng は英語を指す. PARADISEC と FLORES は検証データ (dev) の出力である.

|           |          |   |
|-----------|----------|---|
| PARADISEC | kac (正解) | Dai wa she wa du langai lu gap ai shi gaw.  |
|           | eng (正解) | One day, he could hunt a boar.  |
|           | eng→kac  | Yang gaw wa n du langai mi lu gap ai da.  |
|           | kac→eng  | He got a wild boar.   |
| 会話文       | kac (正解) | Ndai langu si i?  |
|           | eng (正解) | Is this a banana?   |
|           | eng→kac  | Ndai langu si n re i?   |
|           | kac→eng  | This banana plant?  |
| FLORES    | kac (正解) | Dai nbungli gau ai wa hpe Squadron a ning baw Dilokrit Pattavee ngu maram chye lu ai.           |
|           | eng (正解) | The pilot was identified as Squadron Leader Dilokrit Pattavee.                                  |
|           | eng→kac  | Dai kaw na gaw dai kaw na gaw pagawp gaw n dai nbungli woi ai dilokrit patta u ngu ai wa re da. |
|           | kac→eng  | The captain was also known as a bat.  |

## 謝辞

本研究はアメリカ国立科学財団 (NSF) 研究費 No. BCS2109709, および JSPS 科研費 24K03887 の助成を受けたものです。

## 参考文献

- [1] 倉部慶太. ジンポー語文法入門. 東京外国語大学アジア・アフリカ言語文化研究所, 2020.
- [2] NLLB Team, Marta R. Costa-jussà, 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, Searley 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, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.
- [3] Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. Small data, big impact: Leveraging minimal data for effective machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2740–2756, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [4] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. **Computational Linguistics**, Vol. 48, No. 3, pp. 673–732, September 2022.
- [5] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [6] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.
- [7] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.
- [8] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [9] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study, 2023.
- [10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [11] Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book, 2024.
- [12] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 522–538, 2022.
- [13] OpenAI. Gpt-4 technical report, 2024.
- [14] Toan Q. Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. In Greg Kondrak and Taro Watanabe, editors, **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 296–301, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.
- [15] Keita Kurabe. Kachin folktales told in jinghpaw, 2013.
- [16] Keita Kurabe. Kachin culture and history told in jinghpaw, 2017.
- [17] 倉部慶太. ジンポー語用例辞典. 東京外国語大学アジア・アフリカ言語文化研究所, 2020.
- [18] 倉部慶太. ジンポー語読本. 東京外国語大学アジア・アフリカ言語文化研究所, 2020.
- [19] Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. NAIST-NICT WMT’23 general MT task submission. In Philipp Koehn, Barry Haddow, Tom Koehn, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 110–118, Singapore, December 2023. Association for Computational Linguistics.
- [20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [21] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.