低資源言語の機械翻訳における類似言語データへの言語固有の単語レベルノイズ注入

濱田祥希　秋葉友良
豊橋技術科学大学
{hamada.shoki.ew, akiba.tomoyoshi.tk}@tut.jp

塚田元
愛知産業大学
tsukada@asu.ac.jp

概要

近年のニューラル機械翻訳（NMT）は、高資源言語に対して高精度な翻訳を実現している。一方で、低資源言語の翻訳モデルの構築は、双方向の翻訳においてデータ量の観点から困難である。本研究では、低資源言語の単言語データから抽出した語彙および出現頻度情報を活用し、低資源言語に関連する高資源言語の対訳データに対して言語固有の単語レベルノイズを注入したものを学習データとして活用する手法を提案する。

実験では、(i) ヒンディー語 ↔ 英語の対訳データに対して、マガヒー語のノイズを注入する設定、および (ii) ベンガル語 ↔ 英語の対訳データに対してアッサム語のノイズを注入した。また、少量の低資源言語の対訳データが利用できる条件における効果も調査した。その結果、提案手法は低資源言語の対訳データが利用できない条件で、すべてのベースラインを上回った。さらに、ノイズ注入に用いる単言語データの規模を増加させることで性能が向上し、ドメインデータの追加による性能向上も確認した。

1　はじめに

世界には約 7,000 の言語が存在するが、ニューラル機械翻訳（Neural Machine Translation; NMT）の学習に必要な対訳コーパスが十分に整備されている高資源言語（High-Resource Language; HRL）は、そのごく一部に限られる。一方、対訳データが乏しい言語においても、単言語データは比較的豊富に存在し、対訳データと比べて容易に取得できることが知られている。

このように、世界中の多くの言語は低資源言語（Low-Resource Language; LRL）に分類され、単言語データのみが利用可能である場合も少なくない。近年では、単言語データは逆翻訳 [1] や大規模言語モデル（LLM）の継続事前学習など、様々な手法において広く活用されている。本研究では、低資源言語の単言語データを活用し、低資源言語と関連する高資源言語の対訳データに対して言語固有の単語レベルノイズを注入することで学習データを拡張し、これを用いて低資源言語の翻訳精度の向上を目指す。

CharSpan [2] は、低資源言語と高い語彙類似度を持つ高資源言語の対訳データに対して、文字レベルのノイズを注入することで、低資源言語から英語の翻訳精度を向上させた。この手法は低資源言語の文字情報をノイズとして活用している一方で、低資源言語の語彙情報 (例: 語彙リストや出現頻度) は十分に活用できていない。

そこで本研究では、低資源言語の単言語データから抽出した語彙リストおよび出現頻度に基づき、高資源言語の対訳データに単語レベルのノイズを注入する手法を提案する [3]。低資源言語の対訳データの有無を考慮した実験により、本手法は特に低資源言語の対訳データが利用できない設定において、文字レベルノイズを用いる既存手法と比較して翻訳性能が向上することを確認した。また、本手法は同一語群に属する低資源言語の翻訳にも一定の有効性を示した。さらに、ノイズ注入に用いる単言語データの規模を拡大することで、翻訳性能が向上する傾向を確認した。加えて、テストデータのソース文を単言語データに含めることで、さらなる性能向上が得られることも確認した。

2　関連研究

対訳データに対するノイズ注入が翻訳モデルの多様性や頑健性に与える影響を検討した研究はいくつか存在するが、言語間転移に対する影響については十分に調査されていない。Word Dropout [4] は単語埋め込みの一部をランダムにゼロベクトルへ置き換える手法である。SwitchOut [5] は、対訳データに対

してランダムな単語置換を施すデータ拡張手法である。これらの手法はいずれも翻訳モデルの頑健性を向上させるが、言語間転移能力の向上は限定的である。

単言語データを活用するニューラル機械翻訳における代表的なデータ拡張手法として、逆翻訳 [1] がある。対訳データが乏しい場合に、目的言語の単言語データを逆方向の翻訳モデルで翻訳し、疑似的な対訳データを生成する手法である。さらに近年では、反復的逆翻訳 [6, 7, 8] が提案されている。この手法では、両言語側の単言語データを用いて、擬似対訳データの生成と両方向の翻訳モデルの更新を交互に繰り返す。

3 言語固有の単語レベルノイズ

本章では、低資源言語 (目的言語) と関連する高資源言語の対訳コーパスに対し、低資源言語の単言語コーパスを用いた単語レベルノイズを付加することで、低資源言語 → 英語翻訳モデルの頑健性を高める手法を提案する。図 1 に提案手法の概略図を示す。

高資源言語の言語対におけるソース側学習データ C^{src}_{HRL} に低資源言語の語彙情報を用いて、単語レベルノイズを注入し、ノイズ付き対訳データ C'_{HRL} を作成する。

まず、任意の文 $S = (x_1, x_2, \ldots, x_n)$ から単語 x_i をランダムに選択する。次に、幾何分布の成功確率を p とし、式 1 に従い編集距離 d を決定する。

$$d \sim P(k) \propto p(1-p)^{k-1} \quad (k = 1, 2, \ldots, K) \tag{1}$$

ここで K は最大編集距離を表す。その後、決定された編集距離 d に基づき、低資源言語の語彙集合 V_{LRL} から候補集合 $V(d, x_i)$ を抽出する。

$$V(d, x_i) = \{w | w \in V_{LRL}, ED(w, x_i) = d\} \tag{2}$$

ここで、$ED(\cdot,\cdot)$ はレーベンシュタイン距離を表す。次に、候補集合 $V(d, x_i)$ の中から低資源言語の単語 $\hat{w}$ を、低資源言語の単言語コーパスにおける単語出現頻度 $f(\cdot)$ に基づき、以下の確率分布に従って選択する。

$$\hat{w} \sim \frac{f(w)}{\sum_{v \in V(d,x_i)} f(v)} \tag{3}$$

以上の過程をまとめると式 4 の確率で単語 x_i から低資源言語の単語 $\hat{w}$ を決定し、置換する。

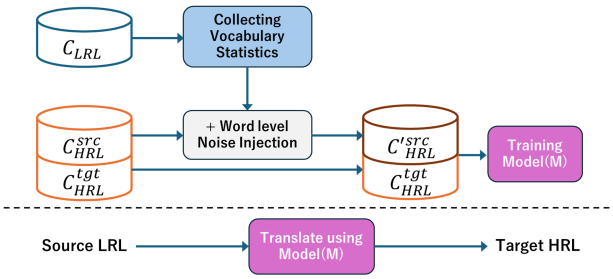

図 1 提案手法の概略図

HRL (HIN): यह प्रेक्षित दिशा यथार्थ दिशा नहीं होती है ।
ENG: This observed direction is not the actual direction.

Word Noise: यह प्रेक्षित दिया पदार्थ दिशा नहीं होती है #

図 2 ヒンディー語（HRL）に対する単語レベルノイズ注入の例。上段: ヒンディー語の文, 下段: マガヒー語の単言語データに基づくノイズ注入後の文

$$\hat{w} \sim P(w) = P(k = d) \frac{f(w)}{\sum_{v \in V(d,x_i)} f(v)} \tag{4}$$

単語 x_i の選択から単語 $\hat{w}$ への置換までを、各文においてノイズの割合が所定の値に達するまで繰り返す。図 2 に、ヒンディー語の文に対してマガヒー語の単言語データを用いて単語レベルノイズを注入した例を示す。

4 実験設定

4.1 データセット

本研究では、インド・アーリア語群に属する言語を対象として実験を行った。インド・アーリア語群は、インド国内で最も話者数が多いヒンディー語を中心に構成されており、豊富な対訳コーパスが利用可能である一方、多くの低資源言語を含む点から、本研究の評価対象として適していると考えられる。高資源言語としてヒンディー語（Hi）、低資源言語としてマガヒー語（Mag）を用いた。マガヒー語を選定した理由は、対象とする低資源言語の中で、ヒンディー語との語彙類似度が最も高いためである [2]。ヒンディー語 ↔ 英語の対訳データに対して、マガヒー語の単言語データから得られた語彙情報を用いたノイズ注入を行った。さらに、マガヒー語固有の単語レベルノイズが、同一語群に属する他の低資源言語の翻訳性能に与える影響についても検証した。

次に、低資源言語の対訳データが利用可能な設定における評価のため、WMT25 Shared Task が提供する英語 ↔ アッサム語（Asm）および英語 ↔ マニプ

表 1　X → En の実験結果 (対訳: Hi ↔ En, 単言語: Mag)

	Mag		Awa		Bho		Hne		Mai		Npi		San	
Models	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
w/o noise	16.46	44.1	16.78	43.4	10.17	36.3	15.00	41.9	8.96	35.8	4.81	26.0	2.42	19.7
CharSpan	22.72	50.0	21.01	47.7	13.30	40.1	**21.72**	48.8	13.90	**41.9**	**7.42**	**31.5**	3.10	**23.5**
Word-Level noise	**24.33**	**51.2**	**21.13**	**48.0**	**14.67**	**40.5**	21.70	**48.9**	**14.07**	41.7	7.03	29.8	**3.76**	22.0

表 2　X → En の実験結果 (対訳: Bn ↔ En, 単言語: Asm)

	Asm		Mni	
Models	BLEU	chrF	BLEU	chrF
w/o noise	5.49	25.2	**1.29**	**18.9**
CharSpan	10.44	36.1	0.75	18.3
Word-Level noise	**12.92**	**38.1**	0.65	17.3
Word-Level noise (w/ test)	**12.97**	**38.7**	—	—
w/o noise + parallel	19.07	44.4	9.8	34.8
CharSpan + parallel	**21.46**	**47.4**	11.97	**38.8**
Word-Level noise + parallel	21.44	46.9	**12.12**	37.3

リ語（Mni）の対訳データを使用した [9, 10] 。実験では、アッサム語およびマニプリ語の対訳データをいずれも学習データに含めて実験を行った。また、この設定では文字体系を一致させるため、ベンガル語（Bn）↔ 英語の対訳データに対して、アッサム語の単言語データを用いたノイズ注入を行った。表 5 に、本研究で使用したコーパスの概要を示す。

4.2　データ前処理

英語データについては、まず Unicode 正規化（NFKC）を行い、`sacremoses` によるトークナイズを適用した。次に、文頭および固有名詞における大文字・小文字の表記揺れを低減するため truecasing を適用し、最後に `subword-nmt` を用いて Byte Pair Encoding (BPE)[11, 12] を学習・適用した。BPE のマージ回数は 16,000 とした。

ヒンディー語、ベンガル語、マガヒー語を含むインド諸言語についても、英語と同様に Unicode 正規化（NFKC）および `sacremoses` によるトークナイズを行い、BPE（16,000 マージ）を適用した。ただし、これらの言語に対しては truecasing は適用しなかった。

4.3　実験条件

単語レベルノイズ注入では、最大編集距離 K を 5 に設定し、編集距離 d は成功確率 $p = 0.5$ の幾何分布からサンプリングした。各文に対して、置換によって変更された文字数の割合が 10% に達するまでノイズを注入した。

翻訳モデルには Transformer アーキテクチャ [13] を採用した。エンコーダおよびデコーダはいずれも 6 層とし、最適化には Adam [14]（$\beta_1 = 0.9, \beta_2 = 0.98$）を使用した。初期学習率は 5×10^{-4} とし、学習率スケジューラには Inverse Sqrt Decay を用いた。勾配爆発を防ぐため、勾配クリップノルムを 1.0 に設定した。ドロップアウト率は 0.2、バッチあたりの最大トークン数は 8,000 とした。学習はエポック数の上限を設けず、検証損失が 5 回連続で改善しない場合に終了した。

5　結果

5.1　実験結果

表 1 に、ヒンディー語 ↔ 英語の対訳データおよびマガヒー語の単言語データを用いた設定における、インド諸言語から英語への翻訳結果を示す。表中の値は、翻訳の自動評価指標である BLEU および chrF のスコアである。w/o noise は学習データにノイズを注入しないベースライン、CharSpan は先行研究、Word-Level noise は提案手法である。

マガヒー語から英語への翻訳において、提案手法は BLEU スコアで w/o noise を 7.87 ポイント、CharSpan を 1.61 ポイント上回り、最も高い性能を達成した。これは、文字や語彙を破壊せずに、低資源言語の語彙情報を保持したままランダム性のあるノイズを注入できたことが、翻訳性能の向上に寄与したと考えられる。一方、その他の低資源言語においても、提案手法は w/o noise を大きく上回る性能を示したが、CharSpan と比較すると同程度の性能向上にとどまった。これらの結果は、提案手法が同一語群の低資源言語に対して一様に大きな効果をもたらすわけではないものの、低資源言語の単言語データから語彙情報を効果的に活用できていることを示唆している。

表 2 に、ベンガル語 ↔ 英語の対訳データおよびアッサム語の単言語データを用いた設定における、低資源言語から英語への翻訳結果を示す。表の上段は、低資源言語の対訳データを使用しない設定であり、下段の「+parallel」は、WMT25 Shared Task で配布された両言語の対訳コーパスを学習データに追

表 3　ノイズ注入のためのアッサム語単言語データのサブセット。

# Sentences	Vocabulary Size
(full data + test) 164,639	182,727
(full data) 163,627	180,810
120,000	154,461
100,000	139,751
50,000	93,797
10,000	33,620

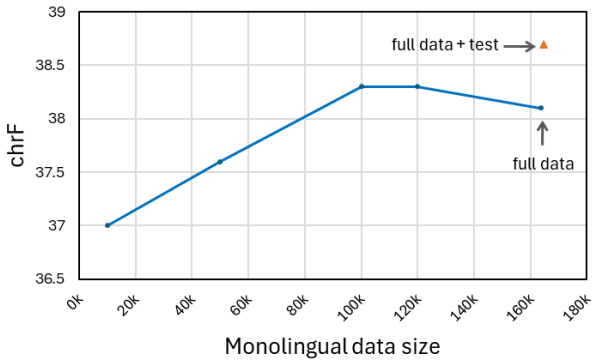

図 3　アッサム語 → 英語翻訳における、アッサム語単言語データの規模とドメインが chrF スコアに与える影響。

加した設定を表す。提案手法はアッサム語から英語への翻訳において、低資源言語の対訳データを使用しない設定では、BLEU および chrF の両方において w/o noise および CharSpan を大きく上回った。一方、低資源言語の対訳データを使用した設定では、提案手法は w/o noise を上回ったものの、CharSpan と同程度の性能向上にとどまった。これは、対訳データの追加によりベースラインの翻訳性能がすでに向上している場合、ノイズ注入による追加的な性能向上が限定的になることを示している。さらに、マニプリ語においては、低資源言語の対訳データの有無により改善の度合いが異なり、提案手法は、ノイズを含まないベースラインが一定の翻訳能力を持つ場合に有効であることを示している。

5.2　単言語データの規模による影響

本研究で用いたノイズ注入手法は、出現頻度に基づいて低資源言語の語彙から置換候補を抽出する。単言語コーパスの規模を拡大すると、より有益な候補を選択できる可能性が高まる。この効果を検証するため、ノイズ注入に用いる単言語コーパスの文数を段階的に変化させ、翻訳性能を評価した。表 3 にアッサム語の単言語コーパスから作成された各サブセットの文数と語彙数を示す。なお、データのばらつきによる影響を排除するため、各サブセットは包含関係を保つように構築した。full data + test は、アッサム語の単言語データに FLORES-200 からアッサム語のテスト文を追加したサブセットである。各サブセットについて BPE を学習し、表 2 と同一の設定でモデルを学習し、FLORES-200 で評価した。

図 3 はアッサム語のサブセットを用いてノイズ注入を行った設定における、アッサム語から英語への翻訳結果を示している。縦軸は翻訳の自動評価指標である chrF、横軸は単言語データの文数である。結果として、単言語データの文数を増加させるにつれて chrF は向上し、約 10 万文付近で最大値に達した。一方で、16 万文まで増加させると、chrF はわずかに低下する傾向が見られた。この結果は単言語データ規模を拡大することで、ノイズ注入時により多様な置換候補を選べるようになり、翻訳性能の向上につながる可能性を示唆している。さらに、テストデータのソース文をノイズ注入用の単言語データに追加した場合、chrF スコアで full data から 0.6 ポイントの向上が見られた。これは、評価データと同一ドメインの文を単言語コーパスに含めることで、置換候補の分布が評価データの分布に近づき、ドメイン適応の観点から追加の改善が得られたことを示唆している。

6　おわりに

本研究では、低資源言語の単言語データから抽出した語彙および頻度情報に基づき、低資源言語と関連する高資源言語の対訳データに言語固有の単語レベルノイズを注入する手法を提案した。インド・アーリア語群を対象とした実験により、提案手法は特に低資源言語の対訳データが利用できない条件において、w/o noise および文字レベルノイズに基づく既存手法を上回る翻訳性能を示した。また、単言語データ規模の拡大に伴って性能が向上する傾向や、評価データと同一ドメインの文を単言語コーパスに含めることによる追加的な改善も観察された。

今後の課題として、事前学習済み翻訳モデルへの適用可能性を検証し、提案手法が強いベースラインに対しても有効に機能する条件を明らかにする。さらに、言語間距離（文字体系/語彙類似度/形態的類似性など）と性能改善の関係を体系的に分析し、本手法が効果を発揮しやすい言語対やデータの特徴を特定することが重要である。

謝辞

本研究は JSPS 科研費 23K11118 の助成を受けたものです。

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. 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. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Kaushal Maurya, Rahul Kejriwal, Maunendra Desarkar, and Anoop Kunchukuttan. CharSpan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 294–310, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[3] Shoki Hamada, Tomoyosi Akiba, and Hajime Tsukada. AkibaNLP-TUT: Injecting language-specific word-level noise for low-resource language translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, **Proceedings of the Tenth Conference on Machine Translation**, pp. 1259–1264, Suzhou, China, November 2025. Association for Computational Linguistics.

[4] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. **Advances in neural information processing systems**, Vol. 29, , 2016.

[5] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 856–861, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Tomohiro Morita, Tomoyosi Akiba, and Hajime Tsukada. A study on unsupervised adaptation of neural machine translation with bidirectional back-translation (in Japanese). In **IPSJ SIG Technical Report**, Vol. 2018-NL-238, pp. 1–5, 12 2018.

[7] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda, editors, **Proceedings of the 2nd Workshop on Neural Machine Translation and Generation**, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data. In **Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence**, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[9] Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. Findings of the WMT 2023 shared task on low-resource Indic language translation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, **Proceedings of the Eighth Conference on Machine Translation**, pp. 682–694, Singapore, December 2023. Association for Computational Linguistics.

[10] Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. Findings of WMT 2024 shared task on low-resource Indic languages translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, **Proceedings of the Ninth Conference on Machine Translation**, pp. 654–668, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] 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.

[12] Philip Gage. A new algorithm for data compression. **C Users J.**, Vol. 12, No. 2, p. 23–38, February 1994.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

表 4　インド諸言語の言語コード

言語コード	言語名	語族	語群
Hi	Hindi	Indo-European	Indo-Aryan
Mag	Magahi	Indo-European	Indo-Aryan
Awa	Awadhi	Indo-European	Indo-Aryan
Bho	Bhojpuri	Indo-European	Indo-Aryan
Hne	Chhattisgarhi	Indo-European	Indo-Aryan
Mai	Maithili	Indo-European	Indo-Aryan
Npi	Nepali	Indo-European	Indo-Aryan
San	Sanskrit	Indo-European	Indo-Aryan
Bn	Bengali	Indo-European	Indo-Aryan
Asm	Assamese	Indo-European	Indo-Aryan
Mni	Manipuri	Sino-Tibetan	Tibeto-Burman

A　本研究で使用したデータセット

表 5　本研究で使用したコーパスの概要

文字体系	コーパス	言語	用途	文数
Devanagari	Samanantar	英語 – ヒンディー語	学習 / 検証（HRL 対訳）	2,531,426 / 1000
	Panlingua Language Processing LLP	マガヒー語	単言語（ノイズ注入）	148,606
	FLORES-200	インド諸言語	評価	1,012
Bengali	Samanantar	英語 – ベンガル語	学習（HRL 対訳）	8,604,580
	WMT25 Shared Task	英語 – アッサム語	学習 / 検証（LRL 対訳）	53,003 / 997
		英語 – マニプリ語	学習 / 検証（LRL 対訳）	22,690 / 997
	FLORES-200	アッサム語・マニプリ語	検証 / 評価	997 / 1,012
	Community 2017	アッサム語	単言語（ノイズ注入）	63,627
	Wikipedia 2021			100,000

B　文字レベルノイズに使用した文字リスト

Language	Script	Candidate Alphabets
Magahi	Devanagari	'म', 'न', 'व', 'ध', 'क', 'र', 'ल', 'स', 'य', 'त', 'ष', 'ट', 'उ', 'च', 'ब', 'घ', 'ण', 'प', 'ग', 'ठ', 'द', 'ह', 'भ', 'अ', 'ए', 'औ', 'थ', 'ओ', 'श', 'ई', 'ज', 'ख', 'आ', 'ड', 'इ', 'छ', 'फ', 'ढ', 'झ', 'ञ', 'ऐ', 'ा', 'ि', 'े', 'ं', 'ु', 'ृ', 'ो', 'ी', 'ृ', 'ू', 'ौ', 'ृ', 'ै', 'ँ'

図 4　CharSpan のノイズに用いた、マガヒー語の文字リスト

Language	Script	Candidate Alphabets
Assamese	Bengali	'ঁ', 'ং', 'ঃ', 'অ', 'আ', 'ই', 'ঈ', 'উ', 'ঊ', 'ঋ', 'এ', 'ঐ', 'ও', 'ঔ', 'ক', 'খ', 'গ', 'ঘ', 'ঙ', 'চ', 'ছ', 'জ', 'ঝ', 'ঞ', 'ট', 'ঠ', 'ড', 'ঢ', 'ণ', 'ত', 'থ', 'দ', 'ধ', 'ন', 'প', 'ফ', 'ব', 'ভ', 'ম', 'য', 'ল', 'শ', 'ষ', 'স', 'হ', '়', 'া', 'ি', 'ী', 'ু', 'ূ', 'ৃ', 'ে', 'ৈ', 'ো', 'ৌ', '্', 'ৎ', 'ড়', 'ঢ়', 'য়', '০', '১', '২', '৩', '৪', '৫', '৬', '৭', '৮', '৯', 'ৰ', 'ৱ'

図 5　CharSpan のノイズに用いた、アッサム語の文字リスト