類似高資源言語の対訳を利用した低資源言語の機械翻訳

学籍番号:221827　可知　拓実　指導教員:秋葉　友良

Abstract: low-resource languages (LRLs) lack sufficient parallel data, limiting the development of high-quality machine translation systems. Building on prior noise-injection methods such as CharSpan and language-specific word-level noise, this study proposes a length-controlled word-level noise technique that adjusts the maximum edit distance based on the target word length. We construct five variants of Hindi–English training data and evaluate them on Magahi→English translation. The model using word-level noise with the maximum edit distance set to the full word length achieves the best performance across major evaluation metrics. These results indicate that tailoring edit-distance of noise to word length is effective for improving translation accuracy in low-resource settings.

1.　はじめに

世界には 7000 種類以上の言語が存在するが、その大半は十分な対訳データを持たない低資源言語（LRL）であり、高資源言語（HRL）に自然言語処理資源が偏っていることが指摘されている (1) [1]. 対訳データが不足する LRL では高精度な機械翻訳モデルの構築が困難である一方、多くの場合単言語データは存在する. そこで本研究では、LRL の単語リストを活用し、HRL (A→B) の並列データのうち、A に対して語彙的に類似する LRL のノイズを注入する先行研究を基盤とし、文字レベルのノイズを導入する CharSpan [2] や、単語レベルで言語固有ノイズを注入する手法 [3] を発展させ、LRL→HRL 翻訳精度の向上を目指す.

2.　実験手法

2.1 ノイズ付与手法

先行研究では、LRL と語彙的に近い HRL のデータに対し、文字スパンを対象としたノイズを付与する CharSpan が提案されている [2]. また、AkibaNLP-TUT により、LRL の単語リストと頻度情報を用いて 言語固有の単語レベルノイズを注入する手法 が提案されている [3].

本研究では、この単語レベルノイズ手法を基盤とし、編集距離 d の最大値を単語長に応じて変動させるモデルを新たに導入し、計 5 条件のデータで翻訳モデル（hi2en）を学習した.

1. 任意の文xに対して，ランダムに単語のインデックスx_iを選ぶ.

2. 選ばれた単語に対して，編集距離 $d \in \{1,2,3,4,5\}$ を決定する．dは成功確率 p = 0.5 の幾何分布に従い決定される.

3. 編集距離 dに一致する低資源言語の単語リストから出現頻度に基づき，置換する単語を選ぶ.

4. ノイズが全体の 10%に到達するまで 2~3 を繰り返し行う.

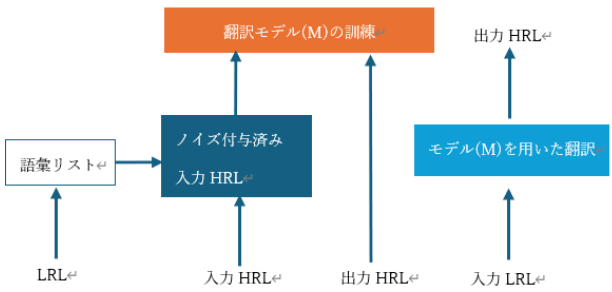

図 1 ：LRL 単語レベルノイズ付与手法の図解

本研究では、d の決定方法を拡張し、以下 5 条件のデータで fairseq 翻訳モデル（hi2en）を学習した

A：ノイズなし

B：CharSpan による文字ノイズ

C：先行研究による単語レベルノイズ

D：対象単語長の 半分を d の最大値としたノイズ

E：対象単語長そのものを d の最大値としたノイズ

2,2　データ

使用したデータとして、以下を用いた.

・ヒンディー語-英語対訳データ（train:samanantar, 2531427 文、dev: FLORES-200, 1000 文）

・マガヒー語-英語対訳データ（test: wat2025, 1012 文）

それぞれデータに対して、英語については NFCK 正規化、トークナイズ、truecasing、BPE を、ヒンディー語・インド諸言語に対しては NFCK 正規化、トークナイズ、BPE を前処理として行った.

3.　実験結果

	BLEU	chrF	ter	COMET	BERTscore
A w/o noise	15.82	45.7	72.44	0.6694	0.381219
B Charspan	19.55	49.4	65.4	0.7208	0.472106
C wordlevel	22.12	52.6	58.98	0.7594	0.534375
D d_max=0.5L	18.86	49.0	64.84	0.7241	0.478497
E d_max=L	23.29	53.1	58.46	0.7589	0.543039

table1:マガヒー→英語翻訳における各モデルの評価方法別スコア

表 1 に、5 条件で学習したモデルの Magahi→English 翻訳性能を示す。

単語長に基づき編集距離の最大値を設定した E：d_max = L が、BLEU・chrF・TER・BERTScore で最も高い性能を示した。COMET においても、先行研究（C）と同等の性能を維持しており、性能低下は見られなかった。

4.　結論

本研究では、LRL の単語をノイズとして HRL の対訳データに導入し、翻訳モデルを作成した。その際、対象となる単語文字数に基づいて編集距離の最大値を設定することが翻訳精度を向上させる要因であると示された. 本提案手法は、対訳データが乏しい LRL の機械翻訳モデル改善に有効である.

参考文献

(1) Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World

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

(3) Shoki Hamada, Tomoyoshi Akiba, and Hajime Tsukada. 2025. AkibaNLP-TUT: Injecting Language-Specific Word-Level Noise for Low-Resource Language Translation