ニューラル機械翻訳のための
言語知識に基づくマルチタスク事前学習

Zhuoyuan Mao† Raj Dabre‡ Fabien Cromieres† Haiyue Song† 中尾 亮太† 黒橋 禎夫†
† 京都大学 大学院情報学研究科
‡ 国立研究開発法人 情報通信研究機構
{zhuoyuanmao, song, fabien, nakao, kuro}@nlp.ist.i.kyoto-u.ac.jp, raj.dabre@nict.go.jp

1 はじめに

特定の言語によらない MASS(MAsked Sequence to Sequence pre-training)[8] などの Seq2Seq 事前学習は、大規模な単言語コーパスを活用することにより、ニューラル機械翻訳（NMT）の低資源言語ペアでの翻訳精度を大幅に向上させた。また、言語知識を利用すれば NMT の精度をさらに改善することができる [6]。日本語は高精度の構文解析器 [5] が開発された言語であるため、事前学習の段階で言語知識を NMT モデルに注入することにより、低資源言語ペア翻訳に非常に役立つことが期待できる。

本稿は、日本語をソース言語またはターゲット言語とする NMT モデルにおいて、MASS の代わりに JASS (JApanese-specific Sequence to Sequence) という新たな事前学習タスクを提案する。JASS は、日本語文節を基に Masked Language Model (MLM) と並べ替え (Reordering) タスクを同時に行う事前学習手法である。また、文節に基づく MLM と並べ替えを一つの損失関数にしたモデルも提案する。ASPEC の日本語－英語や JaRuNC の日本語－ロシア語の翻訳タスクにおいて提案した事前学習モデルを評価する。

2 関連研究

低資源言語ペア翻訳の手法については、現在主に多言語転移学習 (cross-lingual transfer)、データ拡張 (data augmentation) 、事前学習 (pre-training) の研究が活発に行われている。本稿の提案手法は、事前学習に属する。

図 1 は現在低資源言語翻訳における state-of-the-art となっている事前学習手法の MASS の Seq2Seq 構造を示す。

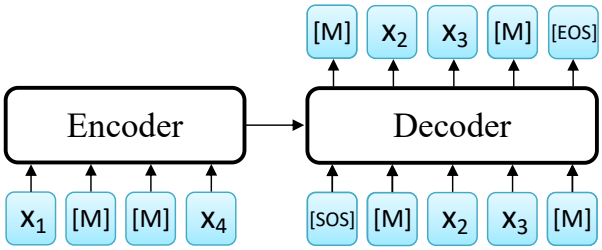

図 1: MASS の Seq2Seq 構造。x_i がトークン、x_2, x_3 が Mask ／予測される連続トークン。BMASS においては、x_2, x_3 は日本語の文節を表す。

3 JASS

MASS から着想を得ており、JASS(JApanese-specific Sequence to Sequence pre-training) という言語駆動型マルチタスク事前学習を提案する。JASS は、BMASS(Bunsetsu-based MAsked Sequence to Sequence pre-training) と BRSS(Bunsetsu Reordering Sequence to Sequence pre-training) の二つの手法の組み合わせである。また、JASS の他に、BMRSS(Bunsetsu-based Masking and Reordering Sequence to Sequence pre-training) という BMASS と BRSS の混合モデルを提案する。図 2 に本稿で提案する全てのモデルの例を示す。

3.1 BMASS

日本語単言語データの構文解析を活用した Seq2Seq 事前学習の BMASS を提案する。MASS は、文のランダムな部分を Mask した後、文脈を考慮して予測することによって NMT モデルを訓練する。これに対して BMASS は、いくつかの文節を Mask して予測するこ

図 2: MASS, BMASS, BRSS, BMRSS の例

とで、日本語の句構造知識を事前に学習できることが期待される。

3.2 BRSS

日本語のような SOV 言語を英語などの SVO 言語に翻訳する際、語順の並び替えを事前に行うことで翻訳精度が高まる [1, 9]。この考え方を基に、語順の並び替えタスクを事前学習の段階にて行うモデルを提案する。BRSS は、図 2 のように、文節ごとに SOV-SVO の模擬的な語順並び替えを行う。各文節内のトークンは変更しない。SVO–SOV の語順並び替えパターンを BRSS.F とし、 SOV–SVO を BRSS.R とする。

3.3 BMRSS

BMRSS は、BRSS に加えて BMASS のようにいくつかの文節を Mask して予測する。この事前学習タスクは、表現能力の高いニューラルネットワークにとっても困難なタスクであるから、効果はあまり期待できないはずである。

3.4 マルチタスク事前学習

いくつかのタスクを交代で事前学習することをマルチタスク事前学習と呼ぶ。本研究の実験では、いくつかのタスクの組み合わせを試す。詳細は次章参照。特に、BMASS と BRSS との組み合わせは日本語構文解析を利用した日本語専用の事前学習タスクであるため、JASS と命名する。また、日本語以外の言語に対して事前学習を MASS で行う。

4 翻訳実験

全ての実験は OpenNMT[1] [4] を用いて行った。

4.1 データセット

提案した事前学習モデル BMASS,BRSS,BMRSS とそれぞれのマルチタスクの組み合わせを評価するのに使用するデータセットは、単言語データは WMT において公開されている News Crawl と Common Crawl[2]、対訳データは日英対訳コーパス ASPEC[7] と

[1] https://github.com/OpenNMT/OpenNMT-py
[2] http://www.statmt.org/wmt19/translation-task.html

	言語	データセット	サイズ
単言語	Ja	Common Crawl	22M
	En	News Crawl	22M
	Ru	News Crawl	22M
対訳	Ja-En	ASPEC-JE	1M
	Ja-Ru	JaRuNC	12K
	Ja-En	JaRuNC	42K
	Ru-En	JaRuNC	84K

表 1: データ概要

JaRuNC[2] を用いて、表 1 に記載されたデータサイズで実験する。

4.2 事前学習

下記のモデルが評価される事前学習のモデルの対象となる。

1. MASS: [8] と同様のセッティング。

2. BMASS: MASS のようにランダムにトークンを Mask するのではなく、日本語の文節をスパンとしていくつかの文節内の全てのトークンを Mask。

3. BRSS: 日本語文を用いて SVO–SOV (BRSS.F) と SOV–SVO (BRSS.R) を実験。

4. JASS(MASS+BRSS): BMASS と BRSS を含んだマルチタスク事前学習。

5. MASS+BMASS: MASS と BMASS を含んだマルチタスク事前学習。

6. MASS+BRSS: MASS と BRSS を含んだマルチタスク事前学習。

7. MASS+JASS: BMASS、BRSS と MASS を含んだマルチタスク事前学習。

8. BMRSS: BMASS と BRSS を一つの損失関数にした事前学習。

9. MASS+BMRSS: BMRSS と MASS を含んだマルチタスク事前学習。

4.3 Fine-tuning

Fine-tuning は、日本語－英語 (Ja-En)、英語－日本語 (En-Ja)、日本語－ロシア語 (Ja-Ru)、ロシア語－日本語 (Ru-Ja) の四つの方向で実験する。下記の設定で Fine-tuning する：

1. Ja-En & En-Ja: ASPEC 対訳コーパスから 3K〜1M 対訳文をサブサンプルし、Fine-tuning する。

2. Ja-Ru & Ru-Ja: [2] のように、JaRuNC 対訳コーパスを基に単方向（UNI）と双方向 (M2M) の翻訳モデルを Fine-tuning する。M2M モデルを構築する際、[3] のように各文頭に翻訳目標言語を表示する言語トークンを加える。

モデル	Ja-En						En-Ja					
	3K	10K	20K	50K	200K	1M	3K	10K	20K	50K	200K	1M
Baseline	1.3	2.6	9.5	17.6	25.5	29.5	1.5	3.7	11.5	21.0	32.7	40.3
MASS	9.9	15.2	18.7	22.3	26.7	29.6	11.0	18.1	21.6	28.0	34.9	**41.2**
BMASS	8.5	14.3	18.6	22.3	26.8	29.7	9.8	17.3	21.1	27.1	34.9	40.5
BRSS.F	8.6	14.8	18.4	22.1	26.5	29.5	10.3	17.7	21.6	27.5	34.9	40.4
BRSS.R	7.4	13.7	17.3	21.8	26.6	29.5	9.8	16.9	20.8	26.3	33.9	40.3
JASS(MASS+BRSS.F)	10.1	15.6†	19.1†	22.9†	26.9	29.7	11.4	18.5†	22.2†	27.3	34.7	40.6
MASS+BMASS	8.7	14.2	18.2	22.2	26.7	29.8	10.0	17.4	21.5	27.2	**35.1**	40.7
MASS+BRSS.F	**10.8**†	16.0†	19.0	22.7	26.8	**29.9**	12.0†	19.0†	22.3†	27.3	34.7	41.1
MASS+JASS	10.7†	**16.1**†	**19.3**†	**23.2**†	**27.1**†	29.5	**12.6**†	**19.2**†	**23.0**†	**28.1**	34.8	40.8
BMRSS	9.6	15.4	18.4	22.5	26.6	29.5	11.6†	18.3	22.2†	27.1	34.7	40.5
MASS+BMRSS	9.9	15.5	18.6	22.1	26.8	29.7	11.5†	18.2	22.2†	26.8	34.8	40.6

表 2: ASPEC 日英対訳コーパスを用いて 3K〜1M 文をサブサンプルした低／高資源の模擬セッティングにおける BLEU スコア。† は、その BLEU スコアが $p < 0.05$ で MASS より統計的に有意であるものを表す。

モデル	UNI		M2M	
	Ja-Ru	Ru-Ja	Ja-Ru	Ru-Ja
Baseline	0.50	0.72	1.70	4.49
MASS	0.96	2.84	3.08	6.54
BMASS	0.97	2.77	**3.57**†	7.12†
BRSS	0.85	2.36	3.11	6.98†
JASS(MASS+BRSS)	**1.20**	3.08	3.36	**7.92**†
MASS+JASS	1.07	**3.45**†	3.02	7.06†

表 3: JaRuNC(真の低資源シナリオ) における BLEU スコア。†は、その BLEU スコアが $p < 0.05$ で MASS より統計的に有意であるものを表す。

4.4 実験結果

表 2 では日英対訳コーパス ASPEC を用いて低資源から高資源のいくつかの設定をシミュレートした。表 3 は日本語ロシア語翻訳の真の低資源シナリオである。これらから、事前学習を使用したほぼすべての設定で、事前学習を使用しない設定よりも高い BLEU スコアになることがわかる。これは、事前学習の重要性を示している。次に、 JASS が MASS より一般的に優れており、JASS+MASS の組み合わせがどちらよりも優れていることも示した。また、事前学習の効果は翻訳の方向やドメインによって変わることも明らかにした。BMRSS と MASS+BMRSS は、MASS より精度を向上させず、マルチタスクで適切な難しさで事前学習を行う重要性を示していると考えられる。

5 まとめ

本稿では、構文解析器からの情報を活用する事前学習手法である JASS（JApanese-specific Sequence to Sequence）を提案した。この手法は、日本語をソースまたはターゲットとして含む翻訳ペアに対する事前学習で使用でき、MASS などの言語に依存しない事前学習の代替手段となる。ASPEC の日本語－英語対訳および JaRuNC の日本語－ロシア語対訳の実験において、ほとんどの低資源シナリオで JASS が MASS よりも優れていることが示された。さらに、MASS と JASS の組み合わせは、個々の方法よりもはるかに優れた結果をもたらすことを示した。

参考文献

[1] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Two-stage pre-ordering for Japanese-to-English statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1062–1066, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[2] Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pp. 128–139, Dublin, Ireland, 19–23 August 2019.

[3] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 339–351, 2017.

[4] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pp. 22–28, 1994.

[6] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3868–3873, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. Overview of the 2nd workshop on Asian translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pp. 1–28, Kyoto, Japan, October 2015. Workshop on Asian Translation.

[8] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 5926–5936, 2019.

[9] Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig. Handling syntactic divergence in low-resource machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1388–1394, Hong Kong, China, November 2019. Association for Computational Linguistics.