事前学習した単語分散表現を利用したマルチモーダル機械翻訳

平澤 寅庄　山岸 駿秀　松村 雪桜　小町 守
首都大学東京
{tosho.hirasawa@, yamagishi-hayahide@ed., matsumura-yukio@ed.,
komachi@}tmu.ac.jp

1 はじめに

マルチモーダル機械翻訳は、翻訳する際に原言語文と紐付けられた言語以外の情報も用いる技術のことである。近年、機械翻訳は Encoder-Decoder と呼ばれる系列変換モデルや Attention 機構の登場により、従来の統計的機械翻訳を上回る性能を達成している [1]。本研究では、マルチモーダル機械翻訳に画像の情報を利用し、文に含まれる曖昧性を解消することで、翻訳の性能のさらなる向上を目指す。この技術の発展により、ニュースや映像などのマルチモーダルな情報を扱う産業での応用が望まれる。

マルチモーダル機械翻訳の研究では、主に 2 つのアプローチがある。1 つは画像特徴量を利用するモデルの研究であり、マルチタスク学習 [7] や画像特徴量を翻訳システムに組み込む方法 [3, 4] が提案されている。一方で、マルチモーダル機械翻訳に利用できるデータセットは小さいため、外部リソースを利用する研究も行われている。このアプローチでは、画像を含まない対訳コーパス [7, 9] や逆翻訳で作成した疑似コーパス [11] を追加のリソースとして利用する手法が提案されている。

しかし、マルチモーダル機械翻訳において、事前学習した**単語分散表現**を外部リソースとして利用する先行研究はあまりない。リソースの乏しい言語対でのニューラル機械翻訳において、事前学習された単語分散表現を Encoder に組み込むことで性能が向上するが、Decoder に組み込んでも性能が向上しないことが確認されている [17]。Kumar ら [12] は、事前学習した単語分散表現を Decoder で予測する手法を提案し、従来のニューラル機械翻訳と同等以上の性能を達成し低頻度語の翻訳精度を向上させており、単語分散表現をより効果的に利用しているといえる。

本研究では、単語分散表現を積極的に活用する Kumar らの手法をマルチモーダル機械翻訳に導入し、事前学習された単語分散表現の有効性を示した。本研究の貢献は以下の通りである。

1. 単語分散表現を予測する新しいマルチモーダル機械翻訳のモデルを提案した。
2. 事前学習された単語分散表現が Decoder の性能を向上させることを確認した。

2 単語分散表現の予測によるマルチモーダル機械翻訳

本研究では、マルチモーダル機械翻訳に Kumar ら [12] の手法を導入し、事前学習された単語分散表現を組み込む。単語分散表現の効果を確認しやすくするため、ベースとなるモデルには、シンプルなマルチタスク学習型のモデルである IMAGINATION [7] を使用する。

IMAGINATION では機械翻訳と潜在共有空間構成の 2 つのタスクを学習する。後者では、文と画像を同一の潜在共有空間にマッピングするとき、ある文とそれに対応する画像の距離が近くになるように学習する。2 つのタスク間で Encoder を共有する。

マルチタスク学習に使用する損失関数は、タスクごとの損失関数の線形補間として与える。

$$J = \lambda J_{\mathrm{T}}(\theta, \phi_{\mathrm{T}}) + (1-\lambda) J_{\mathrm{V}}(\theta, \phi_{\mathrm{V}}) \qquad (1)$$

θ は共有された Encoder のパラメータで、ϕ_{T} および ϕ_{V} はそれぞれ、翻訳タスクおよび潜在共有空間の構成タスク特有のパラメータである。また、$J_{\mathrm{T}}(\theta, \phi_{\mathrm{T}})$ および $J_{\mathrm{V}}(\theta, \phi_{\mathrm{V}})$ はそれぞれ、翻訳タスクおよび潜在共有空間の構成タスクに関する項である。λ は補間係数である。

2.1 機械翻訳タスク

機械翻訳の基本的な構造は、Bahdanau ら [1] と同じであるが、Decoder の出力層で単語の生成確率では

なく、単語分散表現を予測し、事前学習された単語分散表現から最も近い単語をシステム出力とする点が異なる [12]。

$$\hat{\boldsymbol{e}}_{\boldsymbol{j}} = \tanh(\boldsymbol{W}_{\mathbf{o}}\boldsymbol{s}_{\boldsymbol{j}} + \boldsymbol{b}_{\mathbf{o}}) \quad (2)$$

$$\hat{y}_j = \underset{w \in \mathcal{V}}{\operatorname{argmin}}\{d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(w))\} \quad (3)$$

$\boldsymbol{s}_{\boldsymbol{j}}$、$\hat{\boldsymbol{e}}_{\boldsymbol{j}}$、$\hat{y}_j$ はそれぞれ各タイムステップ j における Decoder の隠れ状態、単語分散表現予測、システム出力で、$\boldsymbol{W}_{\mathbf{o}}$ および $\boldsymbol{b}_{\mathbf{o}}$ は出力層のパラメータである。また、$\mathcal{V}$ は出力言語側の語彙集合、w は出力言語の語彙集合に含まれる単語、$\mathbf{e}(w_k)$ は w_k に対応する事前学習された単語分散表現、$d(\boldsymbol{e}_{\boldsymbol{a}}, \boldsymbol{e}_{\boldsymbol{b}})$ は 2 つの単語分散表現 $\boldsymbol{e}_{\boldsymbol{a}}$ と $\boldsymbol{e}_{\boldsymbol{b}}$ の間の距離を表し、本研究では距離関数にコサイン類似度を使用する。

機械翻訳の損失関数には Lazaridou ら [13] が提案する Margin-based Ranking Loss を使用する。

$$J_T(\theta, \phi_{\mathrm{T}}) = \sum_{j}^{M} \max\{0, \gamma + d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(w_j^-)) - d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(y_j))\} \quad (4)$$

$$w_j^- = \underset{w \in \mathcal{V}}{\operatorname{argmax}}\{d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(w)) - d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(y_j)) \quad (5)$$

M は出力文の長さ、γ はマージンである[1]。w_j^- は負例であり、予測した単語分散表現と近く、正解の単語分散表現から遠いものが 1 つ選ばれる。

事前学習された単語分散表現は Encoder 埋め込み層と Decoder 埋め込み層の初期化、および Decoder の出力層に使用する。Encoder の埋め込み層は初期化ののち、学習データを使用して追加の学習を行うが、Decoder の埋め込み層と出力層はパラメータを固定し、学習を行わない。

2.2 潜在共有空間構成タスク

このタスクの Decoder では、共有している Encoder の隠れ状態 $\boldsymbol{h}_{\boldsymbol{i}}$ の平均を計算したのち、潜在共有空間における文の分散表現 $\hat{\boldsymbol{v}}$ を得る。

$$\hat{\boldsymbol{v}} = \tanh(\boldsymbol{W}_{\mathbf{v}} \cdot \frac{1}{N}\sum_{i}^{N} \boldsymbol{h}_{\boldsymbol{i}}) \quad (6)$$

N は入力文の長さ、$\boldsymbol{W}_{\mathbf{v}}$ はパラメータである。

損失関数には Max Margin Loss を使用し、対応する画像の特徴量 $\boldsymbol{v}$ との距離が近くなるように学習する。

$$J_V(\theta, \phi_{\mathrm{V}}) = \sum_{\boldsymbol{v'} \neq \boldsymbol{v}} \max\{0, \alpha + d(\hat{\boldsymbol{v}}, \boldsymbol{v'}) - d(\hat{\boldsymbol{v}}, \boldsymbol{v})\} \quad (7)$$

[1]実験では $\gamma = 0.5$ を使用した

α はマージンで、潜在共有空間における各点の離れ度合いを制御する[2]。

3 実験

3.1 データセット

実験では WMT17 Shared Task で公開された Multi30k [6] データセットを使用して学習、検証、評価した。評価指標には BLEU [16] と METEOR [5] を使用した。評価する言語対は、入力言語にフランス語、出力言語に英語とした。

画像の特徴量の抽出には Multi30k で提供されているスクリプト feature-extractor を使用した[3]。これにより、事前学習された ResNet-50 [10] により画像がエンコードされたのち、ネットワークの pool5 層にある隠れ状態（2048 次元）が特徴量として抽出される。

3.2 モデル

実装には NMTPY[4]の v3.0.0 を使用した。

共有 Encoder の入力層は 128 次元で、隠れ層は双方向 GRU で 256 次元である。機械翻訳システムの Decoder の入力および出力層で使用する単語分散表現は 300 次元で、隠れ層は 256 次元である。また、潜在共有空間は 2048 次元である。

入力文と参照文は Multi30k データセットのスクリプト task1-tokenize.sh を使用し前処理を行った。語彙サイズはすべての入力言語および出力言語で 10,000 とした。

損失関数の補間係数には $\lambda = 0.5$ を使用した。最適化手法には Adam を使用し、学習率は 0.0004 である。勾配は 1.0 でクリッピングし、ドロップアウト率は 0.3 に設定した。

3.3 単語分散表現

モデルに使用する単語分散表現の事前学習には FastText [2] を使用する。Wikipedia および Common Crawl から学習した単語分散表現はオンラインで公開されている[5][8]。これらの単語分散表現は skip-gram [14] を使って学習されており、次元は 300 である。

[2]実験では $\alpha = 0.1$ を使用した
[3]https://github.com/multi30k/dataset
[4]https://github.com/lium-lst/nmtpytorch
[5]https://fasttext.cc/

	dev	test	
Model	BLEU	BLEU	METEOR
NMT	50.83	51.00±.37	42.65±.12
IMAG+	51.03	51.18±.16	42.80±.19
Proposed	52.20	52.90±.07	43.70±.11

表 1: Multi30k の評価

未知語の単語分散表現には、事前学習に使用したコーパスに含まれるがモデルの学習データに含まれない単語分散表現の平均を使用する。また、Mu ら [15] の手法にしたがって、事前学習された単語分散表現を前処理する。まず、全語彙の単語分散表現の平均をゼロになるよう調整（センタリング）したのち、主成分分析を行い、単語分散表現から主成分 5 つに相当する成分を取り除く。

4 結果

表 1 は Multi30k データセットで実験を 3 回行った結果である。dev、test にそれぞれ検証と評価データセットでの結果を示す。NMT は画像を使わず単語の生成確率を予測する機械翻訳システム、IMAG+は IMAGINATION [7] を再実装したマルチモーダル機械翻訳システムの結果である。提案手法は NMT と IMAGINATION にくらべ、BLEU スコアでそれぞれ+1.90 と+1.72 の改善を確認した。

5 考察

事前学習された単語分散表現が、機械翻訳システムへ与える影響を確認するため、表 2 では、埋め込み層をランダムで初期化した場合、および、Decoder の埋め込み層を固定しない場合の結果を示した。random は一様分布、fasttext は FastText で事前学習した単語分散表現で初期化した。Fixed は Decoder 埋め込み層のマッピング行列を固定するかどうかを表す。

Decoder 埋め込み層を固定しない場合、IMAGINATION と同等以下の性能となる。このことから、単語分散表現の予測に基づくマルチモーダル機械翻訳システムにおいては、Decoder の埋め込み層を事前学習された単語分散表現に固定することが重要である。

また、すべての埋め込み層を一様分布で初期化する場合と比べ、Decoder 側を FastText で初期化した場合は BLEU スコアで+1.84 の改善が見られ、Encoder 側を FastText で初期化した場合の改善（+0.55 ポイント）と比べ、大きな改善が確認できる。これは、マルチタスク学習を行うことにより、同じ学習データを使用したとしても、Encoder 側は Decoder 側に比べ、十分に学習することができ、事前学習された単語分散表現の効果が低下するためである。

Encoder	Decoder	Fixed	BLEU	METEOR
fasttext	fasttext	Yes	52.90	43.70
random	fasttext	Yes	52.06	43.23
fasttext	random	No	50.77	42.44
random	random	No	50.22	41.97
fasttext	fasttext	No	50.29	42.25
random	fasttext	No	49.69	41.68

表 2: 埋め込み層別の単語分散表現の効果

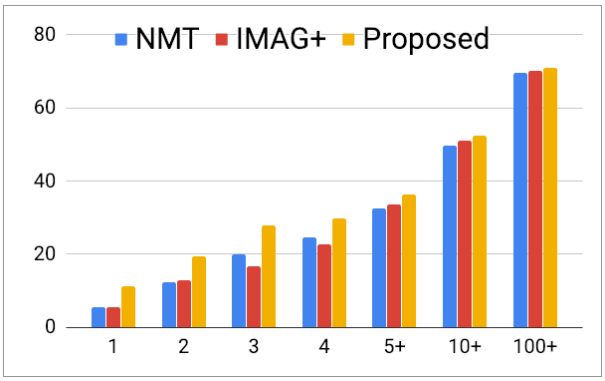

図 1: 単語の出現頻度ごとの F 値

次に、事前学習された単語分散表現が、システム出力の傾向に与える影響を確認するため、図 1 では、単語の出現頻度ごとの F 値を示した。ここでいう F 値とは、単語が出現した参照文のインデックスを正解ラベル、単語が出現したシステム出力のインデックスを予測として計算したものである。提案手法はベースラインと比較し、低頻度語で顕著な改善が見られた一方、高頻度語になると改善の効果が限定的であることが分かった。

6 先行研究

近年行われているマルチモーダル機械翻訳の研究は概ね、画像特徴量を利用する研究と外部リソースを利用する研究の 2 つに分けられる。

画像特徴量を利用するモデルの研究では、ResNet [10] などの画像処理技術で画像特徴量を抽出し、機械翻訳システムへの組み込みやマルチタスク学習を行う。画像特徴量を機械翻訳システムへ組み込むこれらの試みは一定の性能改善を達成したが、ベースとなる機械翻訳のモデルに依存しており、いまだ確立された手法が存在しない。マルチタスク学習は、文と画像の関係を学習するモデルが多く使用されている。Elliott ら [7] は機械翻訳タスクと文から画像特徴量を再構成するタスクを同時学習する IMAGINATION と呼ばれるアプローチを提案した。これらのタスクでは Encoder を共有しており、マルチタスク学習することで、性能向上が期待できる。

外部リソースを利用する研究では、対訳コーパスを利用することが盛んに行われている。Grönroos ら [9] は、画像データを含まない対訳コーパスを組み合わせて使用することで、WMT 2018 Multimodal Shared Task で最高性能を達成している。しかし、単言語コーパスから学習できる知識を利用する研究はほとんど行われていない。

事前学習された単語分散表現を従来のニューラル機械翻訳に組み込む研究は多い。Qi ら [17] は、事前学習された単語分散表現を対訳コーパスが乏しい言語対に適用することで、大きく性能が向上することを報告している。また、事前学習された単語分散表現をより効果的に利用するニューラル機械翻訳システムも提案されている。Kumar ら [12] は Decoder の出力層で、単語の生成確率ではなく、単語分散表現を予測する手法が提案し、従来のニューラル機械翻訳と同等以上の性能を達成するとともに、低頻度語の翻訳精度を向上させた。

7 おわりに

本研究では、事前学習された単語分散表現を効果的に利用する手法を、マルチモーダル機械翻訳に導入し、事前学習された単語分散表現の利用がマルチモーダル機械翻訳の性能を改善することとともに、事前学習された単語分散表現が Decoder の改善に有効であることを示した。

今後は、マルチモーダル機械翻訳に最適な単語分散表現を得られる単言語コーパスの研究や、未知語を構成するための手法、画像の情報を事前学習された単語分散表現に統合することを検討していきたい。

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *TACL*, Vol. 5, pp. 135–146, 2017.

[3] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. LIUM-CVC submissions for WMT17 multimodal translation task. In *WMT*, pp. 432–439, 2017.

[4] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pp. 1913–1924, 2017.

[5] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, 2014.

[6] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74, 2016.

[7] Desmond Elliott and Àkos Kádár. Imagination improves multimodal translation. In *IJCNLP*, Vol. 1, pp. 130–141, 2017.

[8] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *LREC*, 2018.

[9] Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. The MeMAD submission to the WMT18 multimodal translation task. In *WMT*, pp. 603–611, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

[11] Jindřich Helcl, Jindřich Libovický, and Dusan Varis. CUNI system for the WMT18 multimodal translation task. In *WMT*, pp. 616–623, 2018.

[12] Sachin Kumar and Yulia Tsvetkov. Von Mises-Fisher loss for training sequence to sequence models with continuous outputs. In *ICLR*, 2019.

[13] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL-IJCNLP*, pp. 270–280, 2015.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

[15] Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *ICLR*, 2018.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

[17] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *NAACL*, pp. 529–535, 2018.