視覚翻訳言語モデルを用いた英日マルチモーダル機械翻訳

平澤寅庄　小町守
東京都立大学大学院
hirasawa-tosho@ed.tmu.ac.jp　komachi@tmu.ac.jp

概要

本論文では、視覚翻訳言語モデル（Visual Translation Language Model; VTLM）が英日マルチモーダル機械翻訳に適応できるかどうかを検証する。事前学習には Conceptual Captions、マルチモーダル機械翻訳には Flickr30K Entities および Flickr30K Entities JP を使用した。実験の結果、VTLM で事前学習を行うことで、BLEU でおよそ 3 ポイントの改善を達成した。一方で、機械翻訳の訓練事例が豊富な場合には、事前学習を行わないモデルと同程度の性能にとどまった。分析の結果、事前学習で獲得したサブワード分割器が機械翻訳の訓練事例を適切に分割できない例があることが判明した。コードおよび事前学習済みのモデルは https://github.com/toshohirasawa/VTLM_EnJa で公開している。

1　はじめに

マルチモーダル機械翻訳（Multimodal machine translation; MMT）では、翻訳文を生成する際に原文だけではなく、画像などの原文に紐付けられた非言語情報を用いる。画像や動画を持つマルチモーダル対訳コーパスのリリースにより、様々なマルチモーダル機械翻訳モデルが提案されている。

自然言語処理では近年、自己教師あり学習を用いて大規模な単言語コーパスで訓練されたニューラル言語モデルを様々なタスクに応用する研究が盛んに行われており、ニューラル言語モデルが獲得した言語や常識の知識が、各々のタスクに有用であることが示されてきた。機械翻訳でも言語モデルの利用が盛んに研究されており、翻訳品質の向上に有用であることが示されている [1, 2]。

また、対訳コーパスを用いた言語モデルの改善手法も提案されている。多言語の対訳コーパスを使用して訓練された翻訳言語モデル（Translation Language Modeling; TLM）は、分類モデルや機械翻訳モデルの精度を向上させることが報告されている [3]。TLM をマルチモーダルな入力を取るように拡張させた視覚翻訳言語モデル（Visual Translation Language Modeling）は、画像を用いる英独・英仏翻訳で翻訳品質を向上させている [4]。

本研究では VTLM が英日マルチモーダル翻訳でも適用可能かどうかを検証する。VTLM の事前学習には画像・説明文のデータセットである Conceptual Captions [5] に、オンライン自動翻訳サービスで翻訳文を付与した疑似マルチモーダル対訳コーパスを、マルチモーダル機械翻訳には画像つき英日対訳コーパスである Flickr30K Entities JP [6] を用いた。

実験の結果は大規模な疑似マルチモーダル対訳コーパスで事前学習した VTLM が、英日マルチモーダル機械翻訳に有用であることを示した。一方で、機械翻訳タスクの訓練事例が増えると、事前学習を行わなず機械翻訳タスクに特化した機械翻訳モデルと同程度にとどまった。分析の結果、事前学習で獲得したサブワード分割器が必ずしも機械翻訳の訓練事例を適切に分割できているわけではないことが分かった。

2　VTLM を用いた英日マルチモーダル機械翻訳

本研究では Caglayan et al. [4] の研究を英日マルチモーダル機械翻訳に適用する。

2.1　VTLM の事前学習

VTLM では、cross-lingual に加え、cross-modal な分散表現を獲得するように学習を行う。具体的には、cross-lingual な分散表現を学習する Translation Language Modeling (TLM) を拡張し、マスクされた画像特徴量のラベルを予測する masked region classification を行うことで、cross-modal な分散表現を学習する。

VTLM はまず、m トークンの原文 $\mathbf{s} = [s_1, \ldots, s_m]$、

表 1 事前学習およびマルチモーダル機械翻訳モデルの訓練に使用したデータセットの事例数

データセット	画像	英語文	日本語文
Conceptual Captions	3M	3M	-
Flickr30K Entities	29K	29K	-
Flickr30K Entities JP	-	-	29K

n トークンの翻訳文 $\mathbf{t} = [t_1, \ldots, t_n]$ および事前学習された画像認識モデルにより抽出された o 個の要素を持つ画像特徴量 $\mathbf{v} = [v_1, \ldots, v_o]$ を結合し、入力 $\mathbf{x}$ とする：

$$\mathbf{x} = [\mathbf{v} : \mathbf{s} : \mathbf{t}] \tag{1}$$

訓練では $\mathbf{x}$ の要素のうち、ランダムに選択された 15% がマスクされる。マスクに選択された原文・翻訳文の要素は TLM に倣い、80% は [MASK] トークンに置き換え、10% はランダムなトークンへの置き換え、残り 10% は変更せずに保持する。同様に、マスクに選択された画像特徴量のうち、80% は [MASK] トークンの埋め込み表現に置き換え、10% はランダムな埋め込み表現に置き換え、残り 10% は変更せずに保持する。

VTLM では、マスクされた入力 $\tilde{\mathbf{x}}$ から、マスクした位置の要素 $\hat{\mathbf{y}}$ を予測するように学習を行う。

$$\mathcal{L} = \frac{1}{|\mathfrak{X}|} \sum_{x \in \mathfrak{X}} \log \Pr(\hat{\mathbf{y}} \mid \tilde{\mathbf{x}}; \theta) \tag{2}$$

$\mathfrak{X}$ は全訓練データで、θ は VTLM モデルのパラメータである。

2.2 マルチモーダル機械翻訳モデルの学習

事前学習された VTLM の重みは、Transformer をベースにしたマルチモーダル機械翻訳モデルの初期化に用いられる。この際、VTLM は Transformer decoder にある cross-attention 機構を持たないため、初期化対象のモデルの embedding 層、self-attention 層、feed-forward 層を初期化する場合にのみ使用する。

初期化されたマルチモーダル機械翻訳モデルは、目的ドメインのマルチモーダル対訳コーパスを用いて追加の学習を行う。

3 実験設定

3.1 データセットと前処理

表 1 に実験で使用したデータセットの概要を示す。実験ではそれぞれのモデルを 1 回訓練し、その結果を報告した。

表 2 Flickr30K Entities JP の結果。“B” は BLEU、“C” は COMET のスコアである。‘—‘ は事前学習を行わずに Flickr30K Entities JP のみで訓練した場合の結果である。**太字**は最高性能を示す。

		検証		評価	
事前学習	MT	B	C	B	C
—	NMT	37.51	0.267	37.83	0.291
—	MMT	33.35	0.115	32.96	0.121
TLM	NMT	40.53	0.385	39.92	0.363
TLM	MMT	**41.28**	0.402	40.72	0.391
VTLM	NMT	41.01	0.411	**41.25**	0.401
VTLM	MMT	40.79	**0.420**	40.86	**0.407**

事前学習用データ VTLM の事前学習には Conceptual Captions を用いた。Conceptual Captions は画像と英語説明文で構成されるデータセットであり、日本語翻訳文を含まない。本研究では、『みんなの自動翻訳@TexTra®』[1] を用いて英語説明文を日本語に翻訳し、擬似的なマルチモーダル対訳コーパスを作成した[2]。英語文は Moses 、日本語文は MeCab [7](IPA 辞書) を用いてそれぞれ単語分割を行った。その後、BPE [8] を用いて、サブワード化した。BPE のサイズは 50,000 で、英語・日本語でサブワードを共有した。

マルチモーダル機械翻訳 マルチモーダル機械翻訳モデルの訓練・検証・評価には画像と英語説明文で構成される Flickr30K Entities [9] およびその翻訳である Flickr30K Entities JP [6] を用いた。データの切り分けは WMT 2018 の shared task に従い、訓練データ 29,000 事例、検証データ 1,014 事例、評価データ 1,000 事例に分割した。英語文および日本語文はそれぞれ Moses および MeCab（IPA 辞書）を用いて単語分割を行った。また、サブワード分割には事前学習用データで学習したサブワード分割器を使用した。評価には BLEU [10] および COMET [11] を使用した。

3.2 機械翻訳モデル

MMT マルチモーダル機械翻訳モデルには Caglayan et al. [4] で使用されたものと同じものを使

1) 「みんなの自動翻訳」は、国立研究開発法人情報通信研究機構の登録商標です（第 6120510 号）。「TexTra」は、国立研究開発法人情報通信研究機構の登録商標です（第 5398294 号）。

2) 翻訳には汎用 NT モデルを用いた。

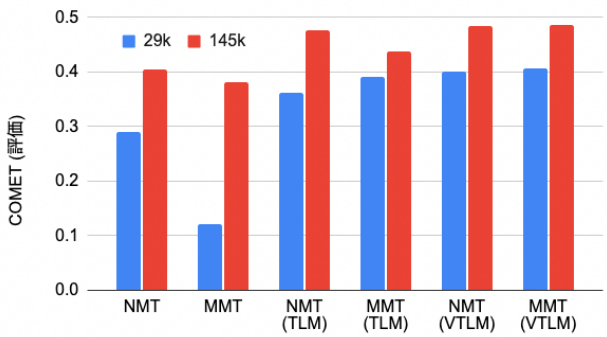

図 1　訓練データを増やしたときの各モデルの性能。括弧内は事前学習の手法である。括弧がないモデルは事前学習を行っていない。

用し、ハイパー・パラメータも同じものを使用した[3]。このモデルでは、画像特徴量と原文の埋め込み表現を結合したものを入力とし、翻訳文を出力するように訓練される。

NMT　画像を使用しない機械翻訳モデルとして、Transformer (base) [12] を使用した。事前学習を行わない場合は、モデルを機械翻訳タスクの訓練事例で学習を行った。事前学習を行う場合は、学習済みの TLM もしくは VTLM のパラメータでモデルを初期化し、機械翻訳タスクの訓練事例で追加の学習を行った。

3.3　画像特徴量

本研究では事前学習およびマルチモーダル機械翻訳タスクでともに ResNet101 をバックボーンとする Faster R-CNN モデル[4]を用いて、画像特徴量の抽出を行った。この画像認識モデルでは 1 つの画像に付き 36 個の領域について、画像特徴量および物体ラベルが抽出される。1 つの画像特徴量は 2,048 次元である。

4　結果

表 2 に結果を示す。VTLM で事前学習をした場合、事前学習を行わない場合に比べ、評価データにおいて NMT では +2.32 BLEU、MMT では +7.90 BLEU の性能向上が確認できた。一方で、TLM で事前学習した場合と比較すると改善は NMT で +0.33 BLEU, MMT で +0.16 であり、事前学習において画像を使用することの有効性は限定的であることが分かった。この傾向は [4] の結果と整合しており、

3)　詳細は https://github.com/ImperialNLP/VTLM を参照。
4)　https://github.com/tensorflow/models/blob/master/research/object_detection/samples/configs/faster_rcnn_inception_resnet_v2_atrous_oid_v4.config

表 3　異なるサブワード分割器によるモデル毎の評価データにおける COMET スコア。“サブワード分割” 列は “入力言語のサブワード分割 → 目的言語のサブワード分割” を表しており、“word” ではサブワード分割を行わない。“29K”/“145K” は 29,000/145,000 事例で訓練した結果をそれぞれ示す。

サブワード分割	モデル	29K	145K
general → general	NMT	0.2910	0.4046
	MMT	0.1206	0.3813
	NMT (VTLM)	0.4010	0.4836
	MMT (VTLM)	**0.4067**	0.4858
in-domain → word	NMT	0.3604	**0.5067**
	MMT	0.1752	0.4896

VTLM が文化的・言語的に近しい言語間（例えば、英独）だけではなく、遠い言語間でも有効であることが示せた。

また、COMET で評価した場合、VTLM で事前学習した MMT モデルが検証および評価で最高性能となった。

5　考察

5.1　MT タスクの訓練データ

一般に、MT タスクの訓練データが増えるにつれ、事前学習による改善が逓減する。ここでは、本実験設定において訓練データを増やした場合に最終的な MT モデルの性能を検証する。図 1 に訓練データを 29,000 事例から 145,000 事例に増やした場合の、各モデルにおける性能の変化を示した。訓練データを 145K 事例に増やした場合でも、VTLM を使った事前学習の有効性は確認できたが、29K 事例の場合に比べるとその効果が逓減していることが分かる。

また、サブワード分割は MT モデルの最終的な性能に大きな影響を与えることが知られている [13]。表 3 に、事前学習データで学習した general サブワード分割器を使用したモデルの性能と、SOTA モデルである Zhao et al. [14] に従い、入力言語側では機械翻訳の訓練事例のみで学習した in-domain サブワード分割器を使用し、目的言語側ではサブワード分割を行わないモデルの性能を示す。訓練事例が少ない場合（29K）は、general サブワード分割器を利用し、VTLM で事前学習したモデルの性能が良いことが確認できた。一方で、訓練事例が多い場合（145K）では、general サブワード分割器を利用し事前学習を

表 4　モデルごとの入力とシステム出力。括弧内にはシステム出力の COMET スコアである。

	src (in-domain): an oriental traveler awaits his turn at the curren@@ cy exchange . NMT: ...が、東洋風の交換会で、自分の順番を待っている。(−0.84) src (general): an oriental traveler awaits his turn at the currency exchange . MMT (VTLM): ...が、紙幣交換所で自分の番を待っている。(0.72) ref: 東洋人の旅行者が、両替所で順番を待つ。
	src (in-domain): girl wearing radio t-shirt has open mouth NMT: ラジオ T シャツを着た女の子が口を開けている。(0.33) src (general): girl wearing radio t@@ -@@ shirt has open mouth MMT (VTLM): ラジオの UNK シャツを着た女の子が口を開けている。(−0.77) ref: ＲＡＤＩＯの T シャツを着た若い女性が、口を開いている。

行ったとしても、in-domain サブワード分割器を利用し機械翻訳タスクの訓練事例のみで訓練したモデルと同程度であった。

5.2　サブワード分割

そこで、それぞれのサブワード手法が性能の変化に与える影響を調べた。まず、general サブワード分割器および in-domain サブワード分割器でそれぞれサブワード分割した評価データ間の編集距離を計算し、MMT (VTLM) と in-domain サブワード分割を使用する NMT モデルの性能差との相関係数を算出した。その結果、入力言語（英語）の編集距離と性能差には優位な相関はなく、出力言語（日本語）の編集距離と性能差には弱い相関（0.12）が認められた。

また、表 4 では、in-domain サブワード分割器を使用したモデル（NMT）と general サブワード分割器を使用し VTLM で事前学習したモデル（**MMT (VTLM)**）のいくつかのシステム出力を示す。上の例では、in-domain サブワード分割器は “currency” を “currency@@ cy” と分割しており、その結果、正しく翻訳できていない。一方で、general サブワード分割器は分割を行っておらず、“紙幣” と翻訳できている。下の例では、逆に general サブワード分割器が “t-shirt” を “t@@ -@@ shirt” と分割したため、正しく翻訳できていない。これは、各サブワード分割器が学習した訓練事例での出現頻度に依存しており、実際 “t-shirt” は Conceptual Captions では一度も出てこず、Flickr30K Entities では 1,181 回出現する。このように、十分に機械翻訳の訓練事例に出現する単語については、分割を行わず、機械翻訳タスクで単語単位の埋め込み表現を学習することで、翻訳精度を向上させると考えられる。

6　関連研究

マルチモーダル機械翻訳は Multi30K データセット [15] が提案されていこう、様々な手法が提案されてきた。一方で、Multi30K は小規模なデータセットであるため、Multi30K データセット以外の資源を用いてマルチモーダル機械翻訳モデルの性能を向上させる試みが、当初から行われてきた。Grönroos et al. [16] は字幕ドメインの対訳コーパスでマルチモーダル機械翻訳を事前学習することで、Multi30K における翻訳性能を大幅に改善した。Hirasawa et al. [17] は単言語コーパスから学習した単語分散表現や（視覚）言語モデルをマルチモーダル機械翻訳モデルに統合することで、翻訳精度を改善した。

7　おわりに

本研究では、視覚翻訳言語モデルがマルチモーダル英日翻訳に適用できるかどうかを検証した。実験の結果、大規模な疑似マルチモーダル対訳コーパスで学習した VTLM は英日翻訳の品質を向上させることが示された。一方で、機械翻訳の訓練事例が増えるに連れ、その効果は逓減していき、機械翻訳の訓練事例のみで訓練した機械翻訳モデルと同程度になることが分かった。

本研究では、VTLM の有効性を制限する一因として、事前学習データで学習したサブワード分割器が機械翻訳の訓練事例を分割するのに最適ではないことを示した。今後は、事前学習で学習したサブワード分割器を機械翻訳タスクの訓練でサブワードの更新や分割の微調整することで、機械翻訳タスクの訓練事例からも適切に学習できるような手法を検証していきたい。

参考文献

[1] Kenji Imamura and Eiichiro Sumita. Recycling a pre-trained BERT encoder for neural machine translation. In **Proceedings of the 3rd Workshop on Neural Generation and Translation**, pp. 23–31, Hong Kong, November 2019. Association for Computational Linguistics.

[2] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. Incorporating BERT into neural machine translation. In **International Conference on Learning Representations**, 2020.

[3] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. **arXiv preprint arXiv:1901.07291**, 2019.

[4] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. Cross-lingual visual pre-training for multimodal machine translation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1317–1324, Online, April 2021. Association for Computational Linguistics.

[5] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4204–4210, Marseille, France, May 2020. European Language Resources Association.

[7] Taku Kudo. MeCab: Yet another part-of-speech and morphological analyzer, 2006. http://taku910.github.io/mecab/.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **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.

[9] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30K Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. **International Journal of Computer Vision**, Vol. 123, No. 1, pp. 74–93, 2017.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz 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.

[13] Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In **Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)**, pp. 1017–1024, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[14] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. TMEKU system for the WAT2021 multimodal translation task. In **Proceedings of the 8th Workshop on Asian Translation (WAT2021)**, pp. 174–180, Online, August 2021. Association for Computational Linguistics.

[15] 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, Berlin, Germany, August 2016. Association for Computational Linguistics.

[16] 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 **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 603–611, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[17] Tosho Hirasawa, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. Pre-trained word embedding and language model improve multimodal machine translation: A case study in Multi30K. **IEEE Access**, Vol. 10, pp. 67653–67668, 2022.