4W-06

講演動画の画像言語マッチングに基づくマルチモーダル機械翻訳

寺面 杏優† 大塚 琢生‡ 梶原 智之‡ 二宮 崇‡
愛媛大学工学部工学科† 愛媛大学大学院理工学研究科‡

1 はじめに

マルチモーダル機械翻訳[1]は、画像や音声など、テキスト以外のモダリティの情報を併用する機械翻訳である。画像や動画からは、テキストには含まれない視覚的な情報が得られるため、曖昧性のあるテキストに対して情報を補完し、翻訳品質を改善できると期待されている。

本研究では、講演動画の英語字幕を日本語に翻訳する課題に取り組む。この状況では、講演の発表資料の画像から有益な情報が得られるため、画像を併用するマルチモーダル機械翻訳によって、テキストのみの機械翻訳よりも翻訳品質の改善が見込める。しかし、講演動画から取得できる画像の中には、図1の左に示すように、講演者のみが写る画像など、字幕テキストと直接関連のない画像も含まれる。このような画像を入力すると、翻訳品質の改善は期待できない。

そこで本研究では、画像を併用するマルチモーダル機械翻訳の性能改善のために、講演字幕と時間的に対応する複数の画像の中から、テキストに最も適した画像を選択する手法を提案する。画像ベクトルと言語ベクトルの余弦類似度に基づく埋込ベースの手法および生成した画像キャプションと入力文の単語一致率に基づく生成ベースの手法を提案し、特に前者の手法によって翻訳品質が改善できることを示す。

2 提案手法

本研究では、図1に示すように、講演字幕の音声認識または書き起こしの英語文と、講演動画の中で所与の英語文と時間的に対応する複数枚の画像が与えられる設定を考える。我々が取り組むマルチモーダル機械翻訳では、英語文に加えて1枚の画像を入力し、日本語訳を出力する。図1の例では、講演者のみが写る左の画像や講演会場の全体が写る右の画像を選択する場合よりも、発表資料が写る中央の画像を選択する場合に、より高い翻訳品質が得られると期待する。

Multimodal Machine Translation Based on Image-Text Matching of Lecture Videos
† Ayu Teramen (teramen@ai.cs.ehime-u.ac.jp)
‡ Takumi Ohtsuka (ohtsuka@ai.cs.ehime-u.ac.jp)
‡ Tomoyuki Kajiwara (kajiwara@cs.ehime-u.ac.jp)
‡ Takashi Ninomiya (ninomiya@cs.ehime-u.ac.jp)
Ehime University

音声認識	when you do that here's what I can promise you're going to be the 800-pound gorilla in the forest
書き起こし	So when you do that, here's what I can promise: You're going to be the 800 pound gorilla in the forest.
参照翻訳	手書きの手紙の力によって、森にいる体重 300 キロの ゴリラのような存在になれるでしょう。

図1：講演のマルチモーダル対訳コーパスの例

入力の英語文に適した画像を選択するために、画像と言語の間の意味的な類似度を推定する。具体的には、Vision&Language 基盤モデルの BLIP[1][2]に基づく以下の2つの手法を提案する。

- **埋込ベース**：BLIP を用いて画像ベクトルと言語モデルをそれぞれ取得し、ベクトル間の余弦類似度が最高となる画像を選択する。
- **生成ベース**：BLIP を用いて画像に対する説明文を生成し、画像説明文と入力文の間の BLEU が最高となる画像を選択する。

3 実験設定

3.1 マルチモーダル機械翻訳

本研究では、画像を用いるマルチモーダル機械翻訳モデルとして、画像領域と単語の関連を注意機構で捉える Selective Attention モデル[2][3]を採用した。これは4層128次元の Transformer[4]に、Vision Transformer[5] (vit_tiny_patch16_384)からの画像特徴を融合したモデルである。最適化には RAdam を使用し、バッチサイズを 4,096 トークン、学習率を 1e-4 として訓練した。検証用データにおけるクロスエントロピー損失が 10 回改善されなくなったときに訓練を停止した。

3.2 データセット

講演動画の英日マルチモーダル対訳コーパスとして、我々が構築した TAIL[3][6]を用いた。これは、TED の講演動画の書き起こし英日対訳コーパスに対して、文単位で音声認識の英語文や時間的に対応する画像を付与したものである。図1のように、1文あたり3枚ずつの画像がある。

[1] https://github.com/salesforce/BLIP
[2] https://github.com/libeineu/fairseq_mmt
[3] https://github.com/EhimeNLP/TAIL

表1：マルチモーダル機械翻訳の結果（BLEU）

	音声認識	書き起こし
画像なし	3.94	4.73
ランダム	7.07	8.97
埋込ベース	**7.30**	**9.48**
生成ベース	6.96	8.90

本データセットは、約 10 万文対の訓練用データと 2,669 文対の検証用データおよび 2,371 文対の評価用データからなる。本研究では、訓練用データに対して以下の 2 つの前処理を実施した。

- 前処理 1：本データセットには 1 文あたり 3 枚ずつの画像が含まれているが、3 枚が同じ画像である場合には提案手法が効果を発揮しない。このような事例を除外するために、BLIP を用いて画像ベクトル間の余弦類似度を計算し、3 枚の画像が互いに 0.7 以上の類似度を持つ事例を除外候補として抽出した。
- 前処理 2：上で抽出された除外候補には、全ての画像が入力文と関連する場合としない場合の両方が含まれる。そこで、埋込ベースの提案手法と同様に画像と言語の間の余弦類似度を計算し、3 枚の画像と入力文の最高の類似度が低い順に事例を除外した。

前処理 1 によって約 6 万件が除外候補として抽出され、その半分にあたる約 3 万件を前処理 2 で除外し、残った 7 万文対を訓練に用いた。その他の前処理として、英語には MosesTokenizer[4]、日本語には MeCab[5]（IPADIC）を用いてそれぞれ単語分割を行い、その後 fastBPE[6]による語彙サイズ 16,000 のサブワード分割を行った。

3.3 比較手法

以下の 2 手法をベースラインとする。各モデルは、ランダムシードを変更して 3 回ずつ訓練し、BLEU による評価の平均値を用いて議論する。

- **画像なし**：テキストのみを用いる機械翻訳。
- **ランダム**：無作為に選択された画像を入力するマルチモーダル機械翻訳。

4 実験結果

表 1 に実験結果を示す。画像を用いないベースライン機械翻訳に比べて、画像を用いるマルチモーダル機械翻訳は手法によらず一貫して高い翻訳品質を達成した。マルチモーダル機械翻訳の中では、ノイズを含む音声認識テキストおよびノイズを含まない書き起こしテキストの両方の入力英語文に対して、提案手法のうち埋込ベースの手法が最高の翻訳品質を達成した。

表2：画像選択の人手評価（正解率）

ランダム	埋込ベース	生成ベース
0.495	**0.785**	0.410

表 2 に、無作為抽出した 200 事例に対する画像選択の性能を示す。本分析では、入力文と意味的に関連する画像を適切に選択できた割合を人手評価した。なお、全ての画像が入力文と関連する事例や全てが関連しない事例を除いて 200 事例を評価したことに注意されたい。翻訳品質に関する表 1 の実験結果と同じく、埋込ベースの提案手法が最高性能を達成した。入力文に関連する画像を選択することでマルチモーダル機械翻訳の品質を改善できることが示唆される。

5 おわりに

本研究では、講演動画中の画像を参照しつつ英語字幕を日本語訳するマルチモーダル機械翻訳に取り組んだ。画像と言語の埋込ベースの類似度推定に基づき、入力文に時間的に対応する 3 枚の画像の中から 1 枚を選択した結果、画像を用いない機械翻訳やランダムに画像を選択するマルチモーダル機械翻訳のベースラインと比較して、BLEU による翻訳品質の改善を確認できた。

謝辞　本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究（課題番号：22501）により得られたものです。

参考文献

[1] Umut Sulubacak, Ozan Caglayan, Stig-Arne Gronroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jorg Tiedemann. Multimodal Machine Translation through Visuals and Speech. Machine Translation, Vol.34, pp.97–147, 2020.
[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proc. of ICML, pp.12888-12900, 2022.
[3] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. On Vision Features in Multimodal Machine Translation. In Proc. of ACL, pp.6327–6337, 2022.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Proc. of NIPS, pp.5998–6008, 2017.
[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proc. of ICLR, 2021.
[6] 寺面杏優, 近藤里咲, 梶原智之, 二宮崇. 講演動画の言語横断字幕生成のための英日マルチモーダル対訳コーパスの構築. 言語処理学会第 30 回年次大会, 2024.

[4] https://github.com/moses-smt/mosesdecoder
[5] https://taku910.github.io/mecab/
[6] https://github.com/glample/fastBPE