マルチモーダル機械翻訳における 画像・入力文間類似度と翻訳品質の相関の分析

佐藤郁子　平澤寅庄　金輝燦　岡照晃　小町守
東京都立大学
{sato-ayako, hirasawa-tosho, kim-hwichan}@ed.tmu.ac.jp,
{teruaki-oka, komachi}@tmu.ac.jp

概要

本研究では，画像・入力文間類似度が低い事例がマルチモーダル機械翻訳 (MMT) に及ぼす影響を調査する．MMT では，画像情報による入力文の曖昧性解消で，翻訳品質の改善を図る．先行研究ではシステムレベルでの翻訳品質改善は限定的であると報告されているが，画像情報の貢献度について事例レベルでの定量的な分析は行われていない．そこで本研究では，画像情報が有効な事例の自動識別を目的に分析を行う．画像・入力文間類似度が低い事例は画像情報が入力文を補完し，翻訳品質が改善することが期待できる．その仮説の下，画像・入力文間類似度と画像情報の貢献度の相関を分析し，MMT で翻訳品質が改善する事例で，負の相関を確認した．

1　はじめに

近年，機械翻訳 (MT) において，テキストだけでなく画像情報を利用して翻訳品質を改善するマルチモーダル機械翻訳 (MMT) が注目されている．MT モデルが曖昧性のある文を翻訳する際に，文脈だけでは情報が不足する場合があるという背景から，MMT では，画像情報を追加入力することで入力文の文脈情報を補完し，より正確な翻訳を行う．

システムレベルの評価では MMT モデルの画像情報の貢献はわずかのため，その有効性は議論されている [1, 2, 3]．一方，事例レベルでの画像情報の貢献は，参照文と出力文の比較によって定性的に評価されており，入力を用いた定量的な分析は行われていない．入力のみを用いた事例レベルの自動評価が可能になることにより，分析対象となる事例選択の効率化や MMT モデルの評価に適したデータ作成を実現することができる．

本研究では，画像情報が有効な事例の自動評価を目的として，定量的に翻訳品質を分析する．MMT では文脈情報の不足を画像で補完することから，有効な事例の画像には入力文にない情報が含まれており，その情報量の差分によって画像・入力文間類似度が低くなると考えられる．そこで，「画像情報が有効な事例では，画像・入力文間の類似度が低くなる」という仮説に基づき，英日翻訳の品質を分析した．仮説の概要を図 1 に示す．実験では，画像・入力文間類似度，画像情報の貢献度の計算を行い，それらの相関を確認することで仮説を検証する．

実験の結果より，MMT モデルで翻訳品質が改善する事例において，画像・入力文間類似度は画像情報の有効性と負の相関があることを確認した．また，画像・入力文間類似度が低い事例では，画像情報の入力がノイズになりうることを確認した．

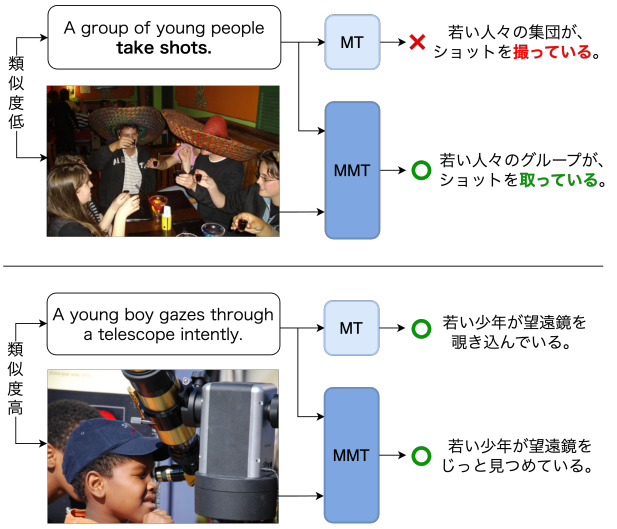

図 1: 画像情報が有効な事例（上）と有効でない事例（下）．画像・入力文間類似度（計算方法は 3 節を参照）は，上の事例が 0.292，下の事例が 0.388 である．上の事例では，“take shots” の「写真を撮る/ショットグラスを取る」という入力文の曖昧性を，画像情報により解消している．下の事例では，画像情報の有無に関わらず翻訳は正しい．

2 関連研究

Elliott [4] は，MMT モデルにおいて，入力文と対応する画像，対応しない画像をそれぞれ追加入力させたときの性能を分析している．この分析では，3 つのモデルのうち 2 つでは，対応する画像を入力した時に性能改善が見られた．Caglayan ら [5] は，入力文にノイズを加え，画像情報の有効性を分析している．この分析では，入力文中の名詞をマスクして文脈情報が限定的な場合，MT モデルより MMT モデルの方が性能が高く，画像情報が有効であることがわかった．しかし，不一致画像の追加入力や入力文劣化は，MMT システムを実際に利用する際の設定に適していない．本研究では，入力を操作しない実験設定で，定量的な翻訳品質の分析を行う．

Hatami ら [6] は，WordNet の語義数に基づいて入力文中の全ての名詞の曖昧度スコアを計算し，MT, MMT モデルの翻訳品質との相関を分析している．この分析では，どちらのモデルも，入力文の曖昧度スコアが高いほど翻訳品質が低くなり，MMT モデルの方が全体の性能が良くなることから，入力文の曖昧性解消に画像情報が有効であることがわかった．しかし，彼らの曖昧度スコアは入力文のみを用いるため，追加入力される画像情報の有効性評価には適していない．そこで本研究では，画像情報の有効性評価のための指標として，画像情報を考慮した画像・入力文間類似度を提案する．

3 分析手法

本研究では入力文 x，参照文 y，画像 v の三つ組のテストデータ $D = \{(x_1, y_1, v_1), ..., (x_{|D|}, y_{|D|}, v_{|D|})\}$ を用いて仮説を検証する．ここで $|D|$ はテストデータのサイズを表す．分析手順は以下の通りである．

手順 1. テストデータ内の各入力文とそれに対応する画像間の類似度集合 $S = \{s_1, ..., s_{|D|}\}$ を求める．ここで画像・入力文間類似度 s_i は，画像 v_i と入力文 x_i をそれぞれ共有潜在空間にエンコードして得たベクトル $\mathbf{v}_i$ と $\mathbf{x}_i$ の内積である．

手順 2. 画像・入力文間類似度の降順にテストデータを並べ替え，N 文ずつ $M(= \lfloor \frac{|D|}{N} \rfloor)$ 組のサブセット $\{D_1, ..., D_M\}$ に分割する．

手順 3. サブセット D_i の (a) 画像・入力文間類似度 s_{D_i} 及び (b) 画像情報の貢献度 c_{D_i} を計算する．

(a) 画像・入力文間類似度 s_{D_i} は，各サブセット内の画像・入力文間類似度の平均とする．

(b) 画像情報の貢献度 c_{D_i} は，評価指標 E を用いて以下のように定義する．

$$c_{D_i} := E(Y_{D_i}^{\mathrm{MMT}}, Y_{D_i}) - E(Y_{D_i}^{\mathrm{MT}}, Y_{D_i}) \quad (1)$$

ここで Y_{D_i} はサブセット D_i に含まれる参照文の集合であり，$Y_{D_i}^{\mathrm{MMT}}$ と $Y_{D_i}^{\mathrm{MT}}$ はそれぞれ MMT, MT モデルの出力文の集合である．

手順 4. 各サブセットの類似度 s_{D_i} と貢献度 c_{D_i} の相関を確認する．負の相関であれば，「画像情報が有効な事例では，画像・入力文間の類似度が低くなる」という仮説が立証される．

4 実験

4.1 実験設定

データ 本研究は英日翻訳による実験で評価した．英語は Flickr30K Entities [7]，日本語は Flickr30K Entities JP [8] を用いる．Multi30K task1 [9] に従い，学習データ 29,000 事例，検証データ 1,014 事例，テストデータ (test2016) 1,000 事例に分割し， 画像とそれに対応したキャプションを用いる．Flickr30K Entities，Flickr30K Entities JP には各画像につきキャプションが 5 文あり，本実験では 1 文目を使用する．テストデータ 1,000 事例を 50 文ずつ 20 組のサブセットに分割し，サブセットごとに BLEU [10] で評価を行った．英語は Multi30K task1 に従ったトークナイズを行い，日本語は MeCab [11] (IPA 辞書) で単語分割した．BPE [12] でサブワード分割し，語彙サイズは英日共有で 8,000 とした．

モデル 本実験では, 画像・入力文間類似度の指標がモデルに依存しないことを確認するため，複数の MMT モデルで実験を行う．テキストベースの MT モデルは， Transformer-Tiny [13] を用いた．MMT モデルには， Transformer ベースの Attentive multimodal Transformer [14]，Gated multimodal Transformer [13] を用いた．Attentive multimodal Transformer には，テキストと画像の文脈ベクトルの連結手法が flat, hierarchical, serial, parallel の 4 種類がある．画像特徴量には Vision Transformer [15] ベースの事前学習済み画像分類モデル CLIP [16]，Faster R-CNN [17], ResNet-50 [18] を用いた．特徴量の数は， CLIP 及び ResNet-50 は 1， Faster R-CNN は 36 である[1)]．MT モデルと MMT モデルのアーキテクチャは，レイ

1) 本研究ではこれらのモデルのうち事前実験で BLEU スコアの高かった上位 3 つを使用する．事前実験で用いたすべてのモデルの BLEU スコアは付録 A に記載する．

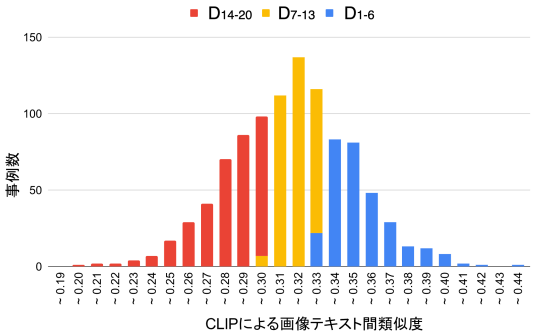

図 2: テストデータにおける画像・入力文間類似度の度数分布 (D_{1-6}, D_{7-13}, D_{14-20} については 4.3 節参照).

表 1: 類似度順のサブセット D_{1-20} 及び D_{1-13} における類似度と貢献度の相関係数.

モデル	画像特徴量	D_{1-20}	D_{1-13}
Attentive-hierarchical	CLIP	0.3715	−0.3083
Attentive-hierarchical	Faster R-CNN	0.0857	−0.6108
Gated	ResNet-50	0.1141	−0.3455

ヤー数を 4 層，注意機構のヘッド数を 4 個，隠れ層の次元数を 256 とし，ミニバッチサイズを最大 4,096 トークンに設定した．最適化手法としては Adam [19] を使用した．

類似度計算　画像と入力文のエンコードには CLIP を用いる． CLIP では，インターネット上で収集した 4 億組の画像・自然言語テキストデータに対して，どのキャプションがどの画像に対応するかを予測する事前学習を行う．テキストのエンコードには Transformer [20]，画像のエンコードには ResNet-50 ベースのモデルまたは Vision Transformer ベースのモデルが使われる．本実験では，Vision Transformer ベースの ViT-B/32 モデルを使用した．テストデータに対して，CLIP で計算した画像・入力文間類似度の度数分布を図 2 に示す．

4.2　結果

全サブセットでの評価　テストデータ D のサブセットを画像・入力文間類似度の高い順から $D_1, D_2, ..., D_{20}$ とし，全サブセット（D_{1-20}）における類似度と貢献度の相関係数を表 1 の 3 カラム目に示す．D_{1-20} では類似度と貢献度は正の相関であり，D_{1-20} においては必ずしも仮説が正しいとは言えない．我々は画像・入力文間類似度が過度に低い事例では入力文の意味に合わない画像情報がノイズになると考え，これを確かめるため各サブセットの類似度と貢献度の分布を調べた．その結果を図 3 に示す．縦軸は MT モデルと MMT モデルの BLEU を

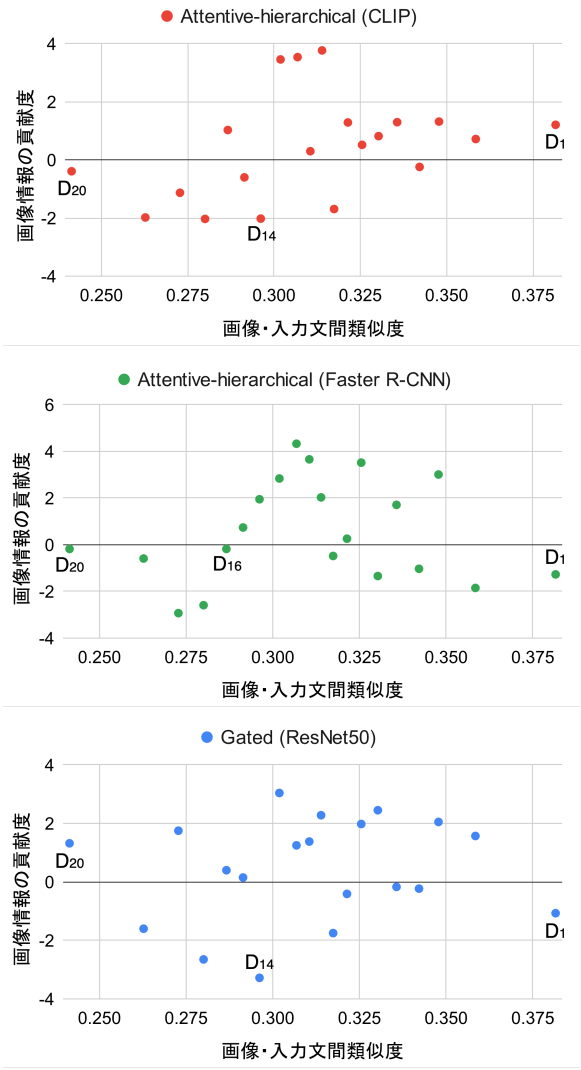

図 3: 類似度順のサブセット D_{1-20} における画像情報の貢献度の分布.

用いた画像情報の貢献度，横軸は各サブセットの画像・テキスト間類似度である．Attentive-hierarchical (CLIP) と Gated (ResNet-50) では，画像・入力文間類似度が過度に低い 0.3 未満のサブセット D_{14} から MT モデルと MMT モデルの翻訳品質が逆転するサブセットが多くなる．同様に Attentive-hierarchical (Faster R-CNN) では, サブセット D_{16} から MT モデルと MMT モデルの翻訳品質が逆転している. 従って，画像・入力文間類似度が過度に低いサブセットでは画像情報がノイズになることがわかる．

ノイズとなるサブセット以外での評価　画像情報により翻訳品質が改善するという前提のもと分析を行うため，画像情報がノイズとなりうるサブセットを取り除いた実験を行った．翻訳品質が逆転するサ

表 2: 画像・入力文間類似度が 0.3 未満の事例（上段）及び 0.3 以上の事例（中段），高い事例（下段）．

 類似度：0.289	SRC	A group of people in white shirts stand on a **grass pitch** covered with many dark balls.
	MT	白いシャツを着た人々の集団が、たくさんの暗いボールで覆われた**芝生**の上に立っている。
	MMT	白いシャツを着た人々のグループが、たくさんの暗い色のボールで覆われた**芝生の投球をして**立っている。
	REF	白いシャツを着た人々の集団が、**芝生のグラウンド**で立っており、そのグラウンドには、そこら中にたくさんの黒っぽい色のボールがころがっている。
 類似度：0.308	SRC	A man and a woman **holding two young boys** are sitting on a park bench, posing for a photograph.
	MT	男性と女性が、公園のベンチに座って写真のためにポーズを取っている。
	MMT	男性と女性が公園のベンチに座って、**二人の幼い男の子を抱いて**写真のためにポーズを取っている。
	REF	**2 人の幼い少年を抱きかかえている**女性と男性が公園のベンチに座って、写真撮影のためにポーズを取っている。
 類似度：0.368	SRC	Two young men in black and white t-shirt are walking down the street with a girl in a pink shirt and patterned leggings.
	MT	黒と白のＴシャツを着た二人の若い男性が、ピンクのシャツと柄のレギンスを身につけた少女と一緒に通りを歩いている。
	MMT	黒と白のＴシャツを着た二人の若い男性が、ピンクのシャツと柄のレギンスを着た少女と一緒に通りを歩いている。
	REF	黒と白のＴシャツを着ている二人の若い男性は、ピンクのシャツと柄のあるレギンスを着ている女性と、通りを歩いている。

ブセットはモデルによって異なるが分析しやすくするため，全てのモデルに共通の画像・入力文間類似度の閾値を設けた．具体的には, 画像・入力文間類似度が 0.3 より低いサブセット D_{14-20} は画像情報がノイズになると仮定し，それらを除いたサブセット D_{1-13} で実験行った．サブセット D_{1-13} における類似度と貢献度の相関係数を表 1 の 4 カラム目に示す．D_{1-13} では類似度と貢献度は負の相関であるため，仮説が立証された．すなわち，画像がノイズとなり MMT モデルで翻訳品質が低下する事例を除いた場合，画像・入力文間類似度が小さいほど画像情報の貢献度が大きくなるといえる．これらの実験結果より，画像情報によって翻訳品質が改善する事例において，画像・入力文間類似度は，画像情報の貢献度を測る指標として適当であるといえる．

4.3 翻訳結果の比較

本節では，画像情報の有効性評価における画像・入力文間類似度の妥当性について分析する．画像・入力文間類似度を 3 段階（0.3 未満：D_{14-20}，0.3 以上：D_{7-13}，高い：D_{1-6}）に分け，それぞれにおける MT と MMT の翻訳結果を表 2 に示す．画像・入力文間類似度は，上段の事例が 0.289，中段の事例が 0.308，下段の事例が 0.368 であり，MMT は Gated (ResNet-50) の出力である．類似度が 0.3 未満の上段の事例では，MT が正しい翻訳をしている一方で，MMT は pitch（競技場）を「投球して」と誤訳しており，画像情報がノイズになっていることがわかる．これは，画像中の人物やボールの個数が多く，画像情報のエンコードが難しいためだと考えられる．類似度が 0.3 以上の中段の事例では，MT で抜け落ちている「二人の幼い男の子を抱いて」という情報が MMT の出力では補完されていることがわかる．これは，画像情報によって入力文がより適切にエンコードされたためだと考えられる．一方，類似度の高い下段の事例では，MT，MMT どちらの翻訳も正しく出力されており，画像情報が不要な事例であることがわかる．

5 おわりに

本研究では，画像・入力文間類似度に着目した MMT の翻訳品質の分析を行った．英日翻訳の実験結果より，MMT モデルで翻訳品質が改善する事例において，画像・入力文間類似度と画像情報の貢献度に負の相関があることがわかった．また，画像・入力文間類似度が一定の値より低い事例では画像情報がノイズとなりうることがわかった．今後の課題として，画像情報がノイズとなる事例の自動判別や，より翻訳品質差との相関が高い評価指標の獲得に取り組みたい．

参考文献

[1] 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**, 2018.

[2] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, 2018.

[3] Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. Sheffield submissions for WMT18 multimodal translation shared task. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, 2018.

[4] Desmond Elliott. Adversarial evaluation of multimodal machine translation. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, 2018.

[5] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2019.

[6] Ali Hatami, Paul Buitelaar, and Mihael Arcan. Analysing the correlation between lexical ambiguity and translation quality in a multimodal setting using wordnet. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop**, 2022.

[7] 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.

[8] 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**, 2020.

[9] 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**, 2016.

[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 on Association for Computational Linguistics**, 2002.

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

[12] 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**, 2016.

[13] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing**, 2021.

[14] Jindřich Libovický, Jindřich Helcl, and David Mareček. Input combination strategies for multi-source transformer decoder. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, 2018.

[15] 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 **Proceedings of the 9th International Conference on Learning Representations**, 2021.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **Proceedings of the 38th International Conference on Machine Learning**, 2021.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In **Advances in Neural Information Processing Systems 28**, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition**, 2016.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference on Learning Representations**, 2015.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems 30**, 2017.

付録

A　全モデルの BLEU スコア

事前実験で用いた全モデルの BLEU スコアを表 3 に示す．

表 3: テストデータに対する各モデルの BLEU. 但し，ResNet-50 については Gated multimodal Transformer では average pool，Attentive multimodal Transformer では last layer を使用した．**太字**は BLEU スコア上位 3 つを示す．

モデル	画像特徴量	BLEU
Transformer-Tiny	-	34.96
Attentive-flat	CLIP	34.71
Attentive-flat	Faster R-CNN	35.06
Attentive-flat	ResNet-50	35.27
Attentive-hierarchical	CLIP	**35.70**
Attentive-hierarchical	Faster R-CNN	**35.52**
Attentive-hierarchical	ResNet-50	34.31
Attentive-parallel	CLIP	35.15
Attentive-parallel	Faster R-CNN	34.19
Attentive-parallel	ResNet-50	34.47
Attentive-serial	CLIP	34.76
Attentive-serial	Faster R-CNN	34.50
Attentive-serial	ResNet-50	35.06
Gated	CLIP	34.84
Gated	ResNet-50	**35.37**

B　類似度と入力文長の関係

各サブセットにおける BLEU スコアを図 4 に示す．画像・入力文間類似度と BLEU には負の相関があり，類似度が低いほど BLEU が高くなることがわかる．そこで，各サブセットにおける翻訳難易度の調査を行った．具体的には，各サブセットごとに入力文の平均トークン数を求め，画像・入力文間類似度との相関を分析した．結果を図 5 に示す．入力文の平均トークン数と画像・入力文間類似度には正の相関があり，入力文の文長が短い事例では，画像・入力文間類似度が低くなることがわかる．分析結果より，画像・入力文間類似度は文長に大きく左右されることがわかった．より高精度な画像情報の有効性評価のためには，入力文の文長を考慮した指標の獲得が必要である．

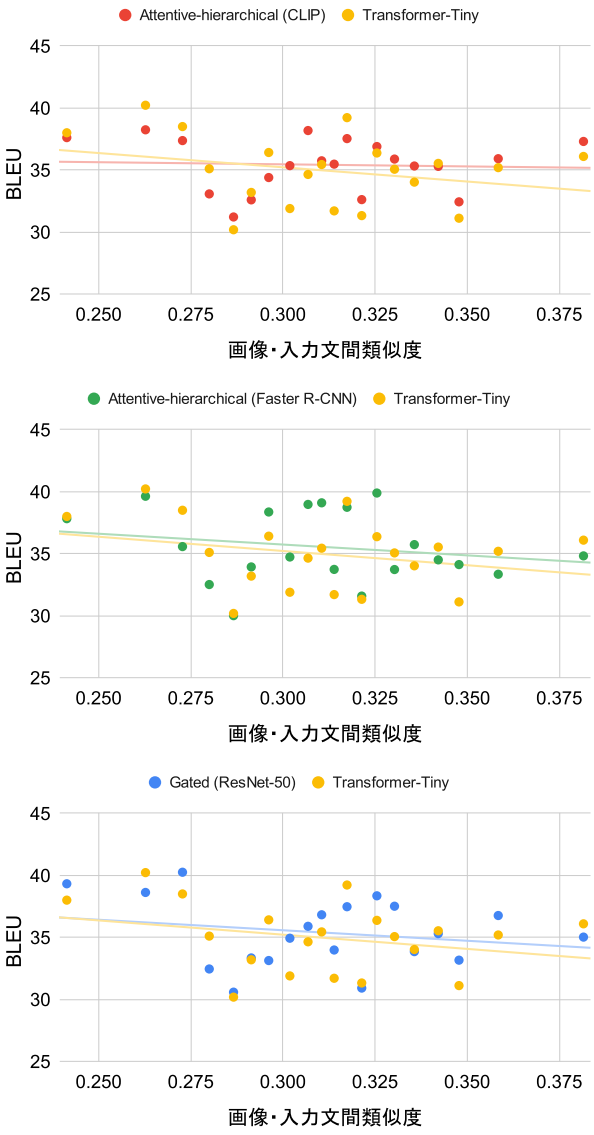

図 4: 類似度順サブセット D_{1-20} における MT モデルと MMT モデルの BLEU.

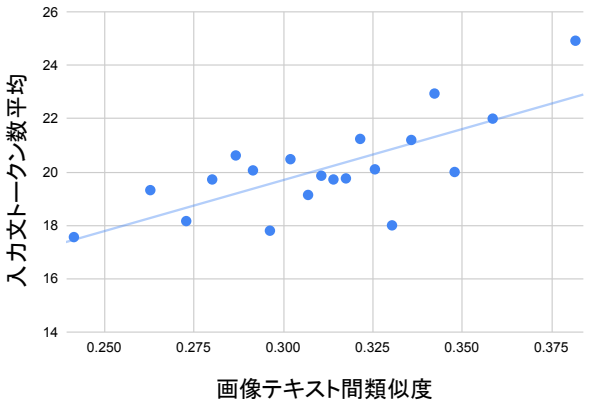

図 5: 画像・入力文間類似度と各サブセットの入力文の平均トークン数.