語義曖昧性解消に着目した 英日マルチモーダル機械翻訳の評価セット構築と分析

佐藤郁子 [1]　平澤寅庄 [1]　金輝燦 [1]　岡照晃 [1]　小町守 [2]
[1] 東京都立大学　[2] 一橋大学
{sato-ayako, hirasawa-tosho, kim-hwichan, teruaki-oka}@ed.tmu.ac.jp,
mamoru.komachi@r.hit-u.ac.jp

概要

マルチモーダル機械翻訳 (MMT) は, 画像を用いた文脈補完によって単語の曖昧性を解消することを目的としている. しかし, 既存 MMT モデルによる品質改善は限定的で, その理由に評価ベンチマークの制限が挙げられる. 既存 MMT 評価セットに含まれるほとんどの原文は明瞭であり, 翻訳に画像を必要としないため, 視覚情報の効果を正確に評価できない. そこで, 本研究では正しい翻訳を行うために画像が必要な英日 MMT のベンチマークを提案する. 既存 MMT モデルを本データセットで評価した結果, 翻訳品質のわずかな向上が確認された. この結果から, 既存 MMT モデルは画像を必要とするシナリオにおいても画像を十分に活用できていないことがわかる.

1　はじめに

自然言語処理とコンピュータビジョンの融合が注目を集めている. マルチモーダル機械翻訳 （MMT）は, その融合の一分野であり, 視覚情報を利用して翻訳品質を向上させることが提案されている. 機械翻訳（MT）モデルが曖昧な文を翻訳する場合, 文脈だけでは十分な情報が得られないことがある. そこで MMT はより正確な翻訳のために視覚情報を追加し入力文の文脈情報を補完する. しかし, 既存の MMT システムは, 翻訳品質を効果的に向上できていない [1, 2]. このような結果の原因として, モデル構造, 学習データ, 評価データが考えられる. 本研究では, 評価データの質の影響に注目する.

MMT の標準的なベンチマークは, Flickr30K データセット [3] から英語のキャプションをドイツ語 [4], フランス語 [5], チェコ語 [2], 日本語 [6] に翻訳することで構築されている. 英語のキャプションは画像を曖昧性なく詳細に記述しているため, 正確な翻訳を生成するために視覚情報で補完する必要がないものがほとんどである [7]. したがって, このようなベンチマークは MMT における語義曖昧性解消への画像の寄与を評価するのには適していない.


En: This is a photo of a **seal**.　En: This is a photo of a **seal**.
Ja: これは**封**の写真である。　Ja: これは**アザラシ**の写真である。
図 1: 英日翻訳における視覚的文脈による曖昧性解消の例.

より正確な評価に向けて, 原文に語義曖昧性を明示的に挿入した評価データセットがいくつか存在する [8, 9]. Futeral ら [9] は, フランス語に翻訳した際に異なる表層となるような複数の語義を持つ英単語をもとに曖昧性解消指向のベンチマークを提案した. これらの研究はアルファベット言語が対象であり, 文化的距離の遠い非アルファベット言語はあまり注目されていない. そこで我々は, 画像が曖昧性解消の手がかりとなる場合のみを含む英日 MMT 評価セットを構築した. 具体的には, WordNet [10] を用いて複数の語義を持つ英単語を抽出し, 画像を与えることで語義が判別できる 250 ペアを手動で選択した. また, 先行研究 [9] に倣って複数の訳語候補があり得る語義曖昧性を含む文を作成し, その語義に対応する画像を ImageNet [11] から収集した （図 1）.

また, 既存の MMT モデルを我々のデータセットで評価し, 語義曖昧性解消の能力を評価した. さらに, 出力の定性的分析を行った. その結果, MMT システムはテキストのみのシステムよりも翻訳品質が若干向上する程度であり, ほとんどの曖昧な単語を曖昧性解消して翻訳することはできなかった. この結果は, MMT システムが画像を取り込むことができないのは, 評価データによるものではなく, モデル構造や学習データによるものであることを示している.

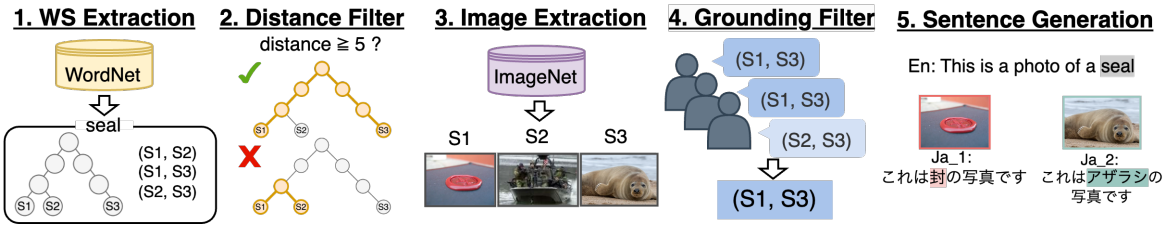


図 2: データセット構築の概要. S1, S2, S3 は語義 1, 語義 2, 語義 3 を表す.

2 関連研究

MMT を評価するためのデータセットを以下に示す. Caglayan ら [12] は入力文の一部をマスクすることで文脈情報を意図的に制限し, 画像の効果を分析している. この分析から, マスクされた入力文では MMT モデルが MT モデルを上回り画像が有効であることが示された. しかし, マスク入力は実用的な MMT システムには適していない. 本研究では, 入力文をマスクせず語義の曖昧性に着目することで, 翻訳候補が複数存在する設定を提案する.

Lala らは視覚的文脈やテキスト文脈が翻訳にどの程度寄与するかを調査するために, Multimodal Lexical Translation Dataset [8] を構築した. このデータセットは視覚的文脈に限定されておらず, 画像で表現できない単語も含まれているため, MMT における視覚的文脈の寄与を評価するには不向きである. そこで, 視覚的文脈のみによる曖昧性解消のための高品質な MMT 評価データセットを構築する.

CoMMuTE [9] は, 視覚的文脈によって訳語が決定される曖昧文からなる英仏データセットである. 各事例は, 曖昧な英文, 2 つの翻訳可能な対象文, 対象文に対応する 2 つの画像から構成される. 彼らは, 曖昧な英文を 談話評価データ [13] から 29 収集し, さらに 21 の文を自作し, 合計 50 の文を作成した. 英仏翻訳で発生する語義の曖昧さは, 英日翻訳でも同じ単語で発生する. しかし, 単にフランス語から日本語に翻訳するだけでは, 評価データセットが比較的小さくなってしまう. そこで, より効率的にデータサイズを拡張するために, WordNet を用いた曖昧語抽出法を提案する.

3 データセット構築

3.1 語義ペア候補の自動選定

このステップでは, WordNet から画像で表現しやすい名詞を抽出する. WordNet には動植物など専門的な名詞も多く含まれるが, 使用範囲が限定されるため学習データに含まれる可能性は低く, より出現頻度の高い一般的な単語の抽出を目指す.

Step 1: Word-senses extraction from WordNet 以下の条件に従って, WordNet から多義名詞と木構造語義を抽出する. (1) 名詞の長さが 10 文字以下（一般的な単語を抽出するため）. (2) 物理的実体に属する（画像で表現できる語義を抽出する）. 次に, 抽出した語義から語義対を作成する.

Step 2: Distance Filter 語義間の距離は 2 つの語義ノードを結ぶ辺の数として定義される. 距離が 5 未満の語義ペアは除外する. フィルタリング後の単語数は 725 であり, 各単語の語義対の平均数は 2.07 である. 各単語について, 距離の降順に語義対をソートする.

Step 3: Image Extraction from ImageNet 各語義に対応する画像を ImageNet から取得する. どちらかのノードに対応する画像がないペアは削除される.

3.2 語義ペアの人手アノテーション

自動的に抽出された語義ペアの中から, 各単語に適切なペアを手動で選択する. さらに, 選択されたペアに対応する画像が不適切な場合は, 画像を置き換える. アノテーションは, 日本語を母語とし, コンピュータサイエンスの修士課程に在籍する 3 名が担当した. すべてのアノテーターは, 語義対のリストから同じ順序で語義対を選択する.

Step 4: Grounding Filter 単語と意味のペアをチェックし, 両方の意味が一般的で画像で表現できるペアを選択する. 適切なペアがない場合, 対象語は除外される. 図 2 は, 除外すべき不適切な語義ペアの例である.

各アノテータはそれぞれ 194, 123, 158 のペアを選択した. 選択された語義ペアの全ペア中の一致率は, Fleiss の Kappa 値で計算すると 0.256% であり, おおよそ一致している. 少なくとも 1 人に選択されたペ

X 動植物名は一般的ではない　X 日本語に翻訳すると同じ単語になる　X 画像から語義を判断するのが困難

図 3: 人手によるアノテーションで除外された事例とその理由.

単語数	画像数	データサイズ	平均語義間距離
250	500	500	9.38

表 1: データセットの統計情報.

アは合計で 197 組ある[1)].

また, 対応する語義を適切に表し, 特徴抽出に十分な品質の画像を確保するため, 123 枚の不適切な（解像度が著しく低い, 教師ラベルの語義が正しくない）画像を, Flickr から CC BY ライセンスで取得した代替画像に置き換える.

Step 5: Sentence Generation　対象単語はテンプレート文 “This is a/an/the [ ].” に挿入される. このテンプレート文は文章の文脈から語義が判断できないように曖昧さを維持した形式になっている. 各単語に 2 つの語義が使われているため, 最終的な文数は対象単語数の 2 倍となる. 構築したデータセットの統計量を表 1 に示す.

4　既存モデルにおける評価

4.1　実験設定

データ　評価には我々のデータセットを使用し, 学習と評価の両方に Flickr30k Entities-JP を使用した. Flickr30k Entities-JP は 29,000 の学習データ, 1,014 の検証データ, 1,000 の評価データがある. 英語は Multi30K task 1 [4] に従ってトークン化し, 日本語は MeCab を使って単語分割した. (IPA 辞書）を用いて単語分割を行った. サブワード分割は BPE を用いて行う.

モデル　MMT モデルと MT モデルを比較し, 画像の寄与を評価した. テキストベースの MT モデルとして Transformer-Tiny [15] を用いた. また, Transformer-based Attentive multimodal Transformer（Attentive）[16], Gated multimodal Transformer（Gated）[15], Visual Translation Language Modelling （VTLM）[17] を MMT モデルとして用いた. VTLM は Conceptual Captions データセットで事前学習されている. CoMMuTE の研究で提案されたモデルは, 大量のキャプションデータに対する事前学習が必要であり, 計算コストの点から本研究では使用しなかった. 画像特徴として CLIP [18], Vision Transformer [19], ResNet-50 [20] を用いた. MT と MMT モデルのアーキテクチャは, 層数を 4, 注意メカニズムのヘッド数を 4, 隠れ層の次元数を 256 とした.

評価指標　sacreBLEU [21] と COMET [22] を使用した. 訳文の本質的でない摂動（文末の変更など）の影響を軽減するため, 3 つの参照文を作成し, その平均値を報告する. さらに, モデルの曖昧性解消を評価するために, [8] で提案された指標も採用し, $\frac{C}{N}$ を計算する. ここで, C は出力中の対象単語が参照中の対象単語と正確に一致した回数, N はデータセットサイズである. この指標を本研究では Lexical Accuracy (LA) と呼ぶ.

4.2　結果

表 2 は, 既存の MMT モデルの自動評価における性能である. MMT モデルは MT モデルを上回り, 画像の寄与が示された. 特に, Attentive (RCNN) は顕著な改善を示し, このモデルが画像に対してより敏感であることを示唆している.

Flickr30k のスコアは, BLEU では −3.61 から 1.06 の範囲, COMET では -3.40×10^{-3} から 1.30×10^{-3} の範囲で改善した. 一方, 我々のデータセットは, BLEU で −1.50 から 2.30, COMET で 4.26×10^{-2} から 5.24×10^{-2} の範囲で改善した. 我々のデータセットは Flickr30k よりも大幅に改善し, 画像の寄与をより敏感に評価することができた.

4.3　分析

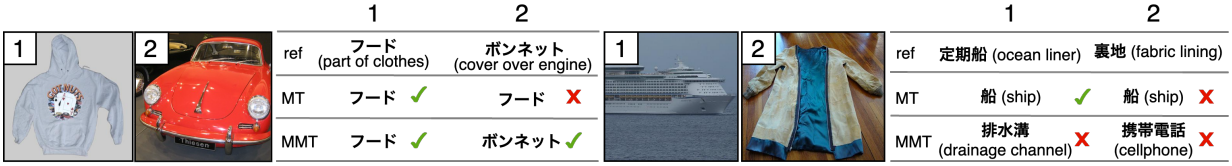
4.3.1　視覚情報による訳語変化

また, システムの出力を詳細に分析した. 図 4 は 2 つの出力例を示しており, MT モデルは Transformer-Tiny , MMT モデルは (a) VTLM (RCNN), (b) Attentive (RCNN) である. (a) の例では, MT モデルは両方の語

1)　データセットを拡張するために, CoMMuTE と Word-in-Context Dataset [14] から 53 の語義ペアを選択する. Step 2 の語義ペアを組み合わせることで, 最終的に 250 のペアを得る.

Metric	Eval data	Transformer-Tiny	Gated		Attentive		VTLM
			CLIP	ResNet	CLIP	R-CNN	R-CNN
BLEU	Flickr30k	43.42	**43.48**	**44.12**	**44.48**	**43.99**	39.81
	Ours	29.40	**29.68**	**30.07**	**30.43**	**31.69**	27.90
COMET	Flickr30k	0.9679	0.9672	0.9673	**0.9688**	**0.9692**	0.9645
	Ours	0.8888	**0.9314**	**0.9344**	**0.9399**	**0.9381**	**0.9412**
LA	Ours	0.1900	**0.1960**	0.1860	**0.1960**	**0.1980**	**0.2200**

表 2: (M)MT モデルの結果. **太字**は, MT モデルを上回っていることを示す.



(a) src: This is a photo of a **hood**. (b) src: This is a photo of a **liner**.

図 4: いくつかの出力例. **太字**は対象単語を示す.

義を「フード」に翻訳した. 一方, MMT モデルは, 対応する画像を参照することで,「ボンネット」と区別することができた. しかし, MMT モデルが正しい訳語に変換できたのは 8 例だけであった. また, 目的語以外の単語が変化している例 (文末の変化, 読点の挿入など) も複数あった. これらの結果から, 自動評価スコアの向上は, 対象単語以外の変更によるトークン数の変化に大きく影響される可能性があることが示唆された.

翻訳品質が向上したのは 8 例だけであったが, 視覚情報が出力対象語に影響を与えたと思われる例も複数あった (図 4 の **liner** など). このような事例の数をモデルごとに付録 A の表 4 に示す. 画像によって訳語が変化した例のうち, 正しく訳されたのは 7% だけであった. つまり, 既存のモデルは視覚情報をわずかしか利用しておらず, 改善の余地がある.

4.3.2 対象単語の学習/検証データでの存在割合

本研究では英日翻訳の際に曖昧性が生じる単語 (対象単語) をもとに曖昧性のある文対を作成している. 4.3.1 節での分析の結果より, 翻訳品質の改善は限定的であることが示唆される. この原因として対象単語が学習データに含まれないためその単語についての学習が不十分であることが考えられる. そこで, 本節では対象単語が学習/検証データに含まれる割合を調査する.

表 3 の結果より, 英日双方に含まれる対象単語の割合は 0 であることがわかる. すなわち, 対象単語が評価データと同じ英日翻訳で使用されている文は学習/検証データには存在しないにもかかわらず, わずかな事例は正確に翻訳できているということである. 一方で, 91.6% の対象単語はソース側の学習データに含まれており, 69.0% の対象単語はターゲット側の学習データに含まれている. すなわち, ソース側とターゲット側それぞれで単語が画像中の物体とマッピングされている場合, 画像が仲介することでソース側の表層とターゲット側の語義を対応付けている可能性が考えられる.

モデルが語義と画像の対応関係を学習できているかについても検証を行い, 付録 B に示す.

	学習データ	検証データ
英語	0.916	0.208
日本語	0.690	0.218
英日	0.000	0.000

表 3: 全対象単語のうち学習/検証データに含まれる単語の割合. 英語はソース側に対象単語の表層が含まれる割合, 日本語はターゲット側に対象単語の語義が含まれる割合, 英日は対象単語が 1 つの文についてソースとターゲットの双方に含まれる割合を示す.

5 おわりに

MMT における視覚情報の寄与を正確に評価するために, 英日評価データセットを構築した. 本データセットで既存のモデルを評価した結果, 画像が翻訳品質を向上させるケースはわずかであることがわかった. つまり, MMT がうまく機能しない原因は, 画像を必要としない設定の評価データではなく, モデル構造や学習データに改善の余地があることがわかった.

謝辞

本研究は AAMT/Japio 特許翻訳研究会の助成を受けたものです.

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] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **Transactions of the Association for Computational Linguistics**, Vol. 2, pp. 67–78, 2014.

[4] 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.

[5] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In **Proceedings of the Second Conference on Machine Translation**, 2017.

[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**, 2020.

[7] Stella Frank, Desmond Elliott, and Lucia Specia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. **Natural Language Engineering**, 2018.

[8] Chiraag Lala and Lucia Specia. Multimodal lexical translation. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[9] Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2023.

[10] Christiane Fellbaum, editor. **WordNet: An Electronic Lexical Database**. MIT Press, 1998.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision (IJCV)**, Vol. 115, No. 3, pp. 211–252, 2015.

[12] 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.

[13] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, 2018.

[14] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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)**, 2019.

[15] 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.

[16] 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.

[17] 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**, 2021.

[18] 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.

[19] 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.

[20] 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.

[21] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, 2018.

[22] 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)**, 2020.

付録

A 視覚情報による訳語変化

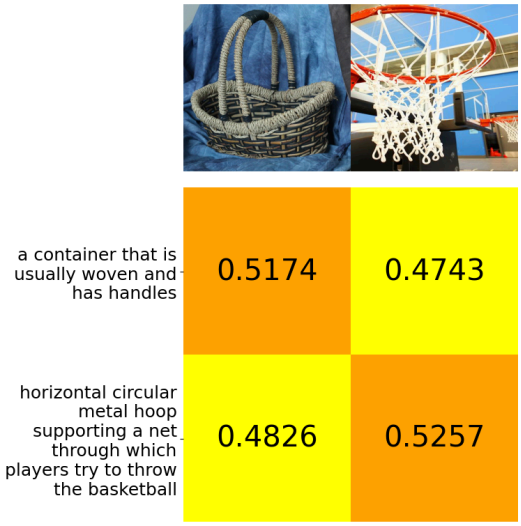
Model	Correct	Incorrect
Gated (CLIP)	0	2
Gated (ResNet-50)	0	3
Attentive (CLIP)	1	3
Attentive (Faster R-CNN)	2	42
VTLM (Faster R-CNN)	5	56

表 4: 全 MMT モデルにおける視覚情報によって翻訳が変わった事例の数.

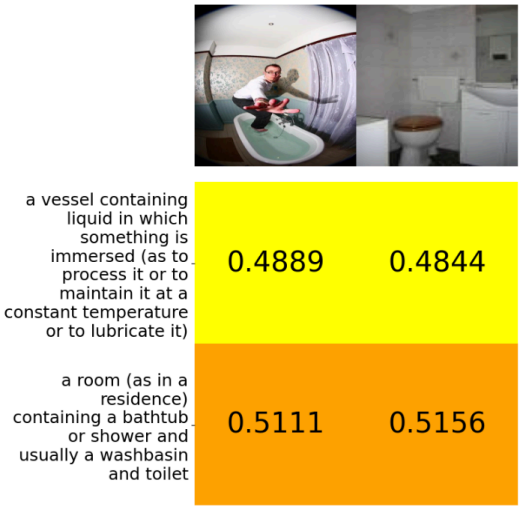
B 画像の識別能力評価

既存 MMT モデルにおいて画像情報の効果が限定的である原因として，エンコードしているモデル自体が語義と画像を結び付けられていない可能性が挙げられる．モデルが語義と画像の対応関係を学習できているか検証するため, 画像の識別能力を調査する．検証では, 対象単語の語義ペアごとに, 各画像についてどちらの語義を表しているかを 2 値分類し, その正解率を報告する. 分類のモデルには CLIP [18] を使用する.

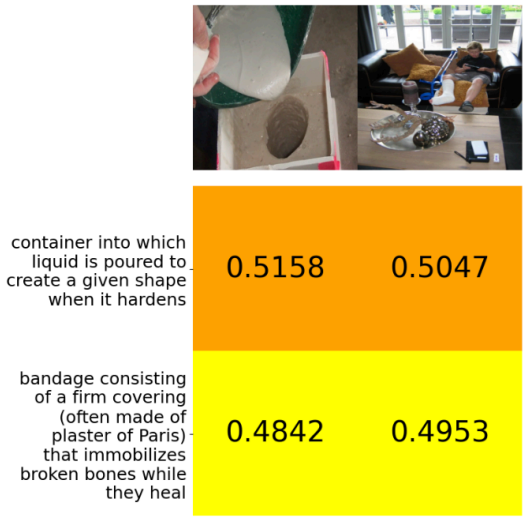
検証の結果, 画像 2 値分類の正解率は 92.4% で, 500 語義のうち誤分類は 38 語義のみであった．すなわち, モデルはほとんど全ての画像と語義の対応付けを正しく行なっていると言える. 分類結果の一部を図 5 に示し, 定性分析を行う．basket （かご / バスケットゴール, 図 5a) の例では, 画像と対応する語義が正しく分類されている. 一方で, cast （型 / ギプス, 図 5b) と bath （浴槽 / バスルーム, 図 5c) の例では誤った分類結果となっている．cast の例は, 画像がギプス以外の物体が多く画像のエンコード自体が困難なことが原因として考えられる．bath の例は, 浴槽自体がバスルームに包含されるものであるという関係から対応語義を選択することが困難であったと考えられる.



(a) basket （かご / バスケットゴール）



(b) bath （浴槽 / バスルーム）.



(c) cast （型 / ギプス） の例.

図 5: 画像の 2 値分類における分類結果.