JaSPICE：日本語における述語項構造に基づく画像キャプション生成モデルの自動評価尺度

和田唯我　兼田寛大　杉浦孔明
慶應義塾大学
{yuiga,k.kaneda,komei.sugiura}@keio.jp

概要

画像キャプション生成タスクでは，生成文の品質が適切に評価されることが重要である．しかし，BLEU や METEOR のような n-gram に基づく自動評価尺度は人間による評価との相関が高くないことが報告されている．そのため英語においては，人間による評価との相関が高い SPICE 等が提案されてきたが，日本語においてはそのような自動評価尺度が存在しない．そこで本論文では，日本語のキャプションに対してシーングラフに基づく評価を行う自動評価尺度 JaSPICE を提案する．実験の結果，提案尺度はベースライン尺度ならびに機械翻訳による英訳文から算出された SPICE と比較して，人間による評価との相関係数が高いことを確認した．

1　はじめに

画像キャプション生成は，視覚障害者の補助，画像に関する対話生成，画像に基づく質問応答など，幅広く研究が行われ，様々な用途で社会応用されている [1–3]．本研究分野においては，生成文の品質が適切に評価されることが重要である．

一方，n-gram に基づく自動評価尺度は人間による評価との相関が高くないことが報告されている [4]．そのため英語においては，人間による評価との相関が高い SPICE [4] 等の自動評価尺度が提案されているものの，日本語を含めた全ての言語においてそのような自動評価尺度が存在するわけではない．したがって，日本語による画像キャプション生成において，人間による評価との相関が十分に高い自動評価尺度が構築されれば有益である．

SPICE は英語による画像キャプション生成タスクにおける標準的な尺度であり，シーングラフに基づいた評価を行う．ここで，SPICE は Universal Dependency (UD) [5] を用いてシーングラフを生成する．しかし，UD では基本的な依存関係しか抽出できず，日本語における「A の B」[6] などの複雑な関係への対処が不十分である（「金髪の男性」など）．また，生成文を英訳して SPICE を適用することも考えられるが，すべての問題設定に対応できるわけではない．例えば，TextCaps [7] では画像中の単語を翻訳することは必ずしも適切ではない．以上のように，SPICE を日本語へ直接適用することは難しい．

そこで，本論文では日本語による画像キャプション生成手法における自動評価尺度 JaSPICE を提案する．JaSPICE は係り受け構造と述語項構造から生成されたシーングラフに基づくため，複雑な依存関係を抽出できる．図 1 に提案手法の流れを示す．図のように，画像に対する参照文群とモデルの生成文を入力として，生成文がどの程度適切であるかを示す JaSPICE 値を計算する．

既存手法との違いは，日本語における画像キャプション生成モデルを評価できる点，係り受け構造と述語項構造に基づきシーングラフを生成する点，および同義語集合を自動評価に用いた点である．係り受け構造と述語項構造をシーングラフに反映させることで，参照文群と生成文に対し適切なシーングラフを生成することができると期待される．また，同義語集合を用いることで，表層表現の不一致による評価値の低下を避けることができるため，人間による評価との相関が高まることが期待される．

提案手法[1)]における新規性は以下の通りである．

- 日本語による画像キャプション生成タスクにおける自動評価尺度 JaSPICE を提案する．
- UD を用いる SPICE とは異なり，JaSPICE では係り受け構造と述語項構造に基づき，日本語の文からシーングラフを生成する．
- 同義語を利用したグラフの拡張手法を導入する．

1）https://github.com/keio-smilab23/JaSPICE

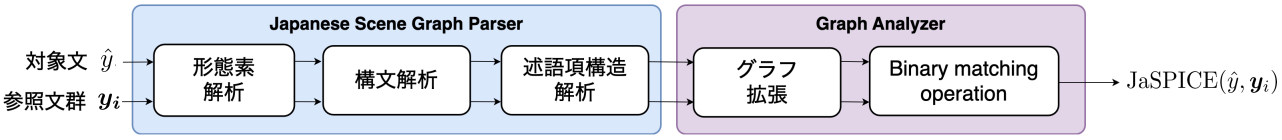


図 1: 提案手法の流れ

2 問題設定

本論文では，日本語での画像キャプション生成に対する自動評価を扱う．画像キャプション生成モデルにおける自動評価尺度は，人間による評価に近いことが望ましい．具体的には評価値と人間による評価との相関係数が高いことが望ましい．

本論文で使用する用語を以下のように定義する．

- **正解キャプション**：画像に対してアノテータが付与したキャプション．
- **述語項構造**：文中の述語とその項の関係を表現する構造 [8]．
- **シーングラフ**：画像内の物体同士の意味的関係を表現したグラフ．詳しくは 3.1 節にて述べる．

画像キャプション生成モデルにおける自動評価尺度は，i 番目の画像に対してモデルの生成するキャプション $\hat{y}_i$ と，画像に対する正解キャプション $\{y_{i,j}\}_{j=1}^{N}$ を入力として，$\{y_{i,j}\}_{j=1}^{N}$ に対して $\hat{y}_i$ が適切であるかの評価値を計算する．ここで，N は y_i あたりの正解キャプション数を示す．

本自動評価尺度の評価には人間の評価との相関係数（Pearson/Spearman/Kendall の相関係数）を使用する．本論文では，日本語の画像キャプションに対する自動評価を前提とする．ただし，本論文の議論の一部は，他言語に応用可能であると考えられる．

3 提案手法

本論文では，日本語における画像キャプション生成のための自動評価尺度 JaSPICE を提案する．JaSPICE は SPICE [4] を拡張した自動評価尺度であり，日本語のキャプションに対してシーングラフに基づく評価を行うことが可能である．本評価尺度は SPICE を拡張した手法だが，主語の補完や同義語によるノードの追加など，SPICE では扱わない要素も考慮しており，本論文の議論の一部は，他の自動評価尺度に対しても応用可能であると考えられる．

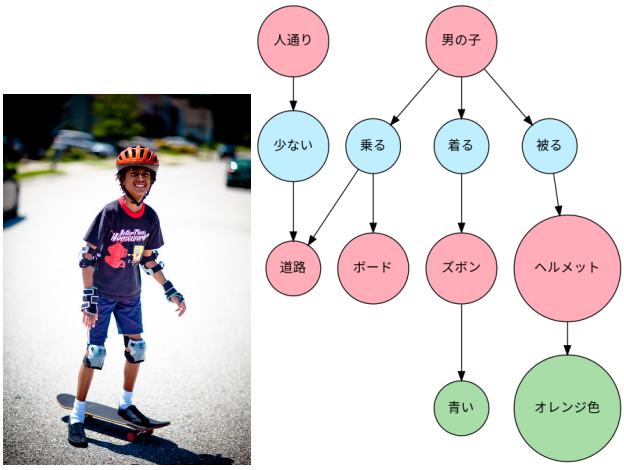
本提案手法と SPICE の主な違いは以下の通りである．

- UD [5] を用いる SPICE とは異なり，JaSPICE は係り受け構造と述語項構造に基づきルールベースでシーングラフを生成する．
- JaSPICE はヒューリスティックなゼロ照応解析と同義語を利用したグラフの拡張を行う．

本提案手法は，図 1 のように Japanese Scene Graph Parser (JaSGP) と Graph Analyzer (GA) に分けられる．

3.1 シーングラフ

シーングラフはキャプション y に対して $G(y) = \mathcal{G}$ <$O(y), E(y), K(y)$> で表される．ここで，$O(y)$ は y に属する物体の集合，$E(y)$ は物体同士の関係の集合，また $K(y)$ は属性を持った物体の集合である．C，R，A をそれぞれ物体，関係，属性の全体集合とすると，$O(y) \subseteq C$，$E(y) \subseteq O(y) \times R \times O(y), K(y) \subseteq O(y) \times A$ である．



(a) (b)

図 2: 画像と対応するシーングラフの一例

図 2 に画像とシーングラフの例を示す．図 2(b) は図 2(a) の説明文「人通りの少なくなった道路で，青いズボンを着た男の子がオレンジ色のヘルメットを被り，スケートボードに乗っている．」から得られたシーングラフである．ピンク，緑，水色のノードはそれぞれ物体，属性，関係を表し，矢印は依存関係を表す．

3.2 Japanese Scene Graph Parser

JaSGP における入力は日本語キャプション $\hat{y}$ であり，出力は入力されたキャプション $\hat{y}$ に対するシーングラフ $G(\hat{y})$ である．まず，形態素解析器，構文解析器，述語項解析器より，$\hat{y}$ から述語項構造と係り受け構造が取り出される．次に，述語項構造と係り受け構造から 10 種類の格を抽出し，抽出した格よりルールベースでシーングラフ $\mathcal{G}$ <$O(\hat{y}), E(\hat{y}), K(\hat{y})$> を生成する．ここで，10 種類の格とはガ格, ヲ格, ニ格, ト格, デ格, カラ格, ヨリ格, ヘ格, マデ格, 時間格 [9] である．

後述の通り，提案する自動評価尺度では $E(\cdot)$ を使用するが，日本語ではゼロ代名詞が存在する場合，すなわち関係 Rel <o, R, o'> のうち物体 o が欠損している場合がある．したがって，提案手法では次のようにヒューリスティックな方法でゼロ照応解析を行う．いま，物体 o_2 と o_3 が関係 R によって接続されているとする．述語項構造と係り受け構造より述語に対する主語が特定できない $\mathcal{R} =$ Rel <$?, R, o_2$> が存在する場合，o_2 と接続している別の関係 Rel <o_3, R', o_2> から $\mathcal{R}$ における主語を o_3 へと決定する．

3.3 Graph Analyzer

GA における入力は $\{y_{i,j}\}_{j=1}^{N}$ から得られた $\{G(y_j)\}_{j=1}^{N}$ と $\hat{y}$ から得られた $G(\hat{y})$ である．まず，GA では次のように同義語によるノードの追加を行う．すなわち，$o \in O(\hat{y})$ の同義語集合 $S(\hat{y})$ を生成し，$O(\hat{y})$ と $S(\hat{y})$ の和集合 $O'(\hat{y})$ を用いて新たに $G'(\hat{y})$ を定義する．ここで，同義語集合には日本語 WordNet [10] を用いた．

次に，$\{y_{i,j}\}_{j=1}^{N}$ に対する $\{G(y_{i,j})\}_{j=1}^{N}$ について，これらを 1 つの $G(\{y_{i,j}\}_{j=1}^{N})$ へと統合する．具体的には，$\mathcal{G}$ <$O(y_{i,j}), E(y_{i,j}), K(y_{i,j})$> について，$\mathcal{G}$ <$\{O(y_{i,j})\}_{j=1}^{N}, \{E(y_{i,j})\}_{j=1}^{N}, \{K(y_{i,j})\}_{j=1}^{N}$> を $G(\{y_{i,j}\}_{j=1}^{N})$ とする．$T(G(x))$ を $T(G(x)) := O(x) \cup E(x) \cup K(x)$ と定義すると，$T(G'(\hat{y}))$ と $T(G(\{y_{i,j}\}_{j=1}^{N}))$ から適合率 P, 再現率 R, および F1 値 F_1 を次のように計算する．

$$P(\hat{y}, \boldsymbol{y_i}) = \frac{|T(G'(\hat{y})) \otimes T(G(\{y_{i,j}\}_{j=1}^{N}))|}{|T(G'(\hat{y}))|}$$

$$R(\hat{y}, \boldsymbol{y_i}) = \frac{|T(G'(\hat{y})) \otimes T(G(\{y_{i,j}\}_{j=1}^{N}))|}{|T(G(\{y_{i,j}\}_{j=1}^{N}))|}$$

$$\mathrm{JaSPICE}(\hat{y}, \boldsymbol{y_i}) = F_1(\hat{y}, \boldsymbol{y_i}) = \frac{2 \cdot P(\hat{y}, \boldsymbol{y_i}) \cdot R(\hat{y}, \boldsymbol{y_i})}{P(\hat{y}, \boldsymbol{y_i}) + R(\hat{y}, \boldsymbol{y_i})}$$

ここで，$\otimes$ は 2 つのシーングラフのうち一致している組を返す演算子である．GA では $\mathrm{JaSPICE}(\hat{y}, \boldsymbol{y_i})$ を出力とし，この値を JaSPICE 値と定義する．

4 実験

4.1 実験設定

JaSPICE を既存の自動評価尺度と比較評価するため，JaSPICE 値と人間による評価との相関係数を用いた評価実験を行う．

$s_J^{(i)}$ を i 番目のキャプションに対する JaSPICE 値，$s_H^{(i)}$ を i 番目のキャプションに対する人間による評価とする．このとき，N 対の $\{(s_J^{(i)}, s_H^{(i)})\}_{i=1}^{N}$ に対する相関係数（Pearson, Spearman, Kendall の相関係数）を評価に用いる．また，実験設定の詳細は付録 A.2 に記載する．

4.2 実験結果

表 1 に提案尺度ならびにベースライン尺度と，人間による評価との相関係数を示す．ここでベースライン尺度には，画像キャプション生成において標準的な尺度である BLEU [11], ROUGE [12], METEOR [13], CIDEr [14] を用いた．表 1 より，JaSPICE は Pearson, Spearman, Kendall の相関係数において，それぞれ $0.501, 0.529, 0.413$ であり，ベースライン尺度を上回った．

表 2 に JaSPICE および SPICE と人間による評価との相関係数を示す．ここで，$\mathrm{SPICE}_{\mathrm{trm}}$ は JParaCrawl [15] で訓練した Transformer の出力した英訳文，$\mathrm{SPICE}_{\mathrm{service}}$ は一般的な機械翻訳システム (DeepL) の出力した英訳文を用いて算出した SPICE 値である．JaSPICE は Pearson, Spearman, Kendall の相関係数において，それぞれ $0.501, 0.529, 0.413$ であり，$\mathrm{SPICE}_{\mathrm{trm}}$ と比較して $0.01, 0.013, 0.01$ ポイント上回った．同様に，$\mathrm{SPICE}_{\mathrm{service}}$ と比較して，JaSPICE はそれぞれ $0.013, 0.014, 0.011$ ポイント上回った．

図 3 に提案尺度の成功例を示す．図は入力画像と $\hat{y}_i$「眼鏡をかけた女性が青い携帯電話を操作している」に対するシーングラフである．図における $y_{i,1}$ は「女性が青いスマートフォンを片手に持ってい

表 1: 自動評価尺度と人間による評価との相関係数

自動評価尺度	Pearson	Spearman	Kendall
BLEU [11]	0.296	0.343	0.260
ROUGE [12]	0.366	0.340	0.258
METEOR [13]	0.345	0.366	0.279
CIDEr [14]	0.312	0.355	0.269
JaSPICE	**0.501**	**0.529**	**0.413**

表 2: JaSPICE および SPICE と人間による評価との相関係数

自動評価尺度	Pearson	Spearman	Kendall
JaSPICE	**0.501**	**0.529**	**0.413**
$\text{SPICE}_{\text{trm}}$	0.491	0.516	0.403
$\text{SPICE}_{\text{service}}$	0.488	0.515	0.402

る」であり，$\text{JaSPICE}(\hat{y}, \boldsymbol{y}_i) = 0.588$, $s_H^{(i)} = 5$ であった．テスト集合において，この JaSPICE 値は上位 0.02%の値であるため，図の例において提案尺度は人間による評価に近い値を出力しているといえる．

4.3 Ablation study

以下の 2 つの条件を Ablation study に定めた．

(i) UD を用いたグラフ解析器を使用した場合：JaSGP を UD を用いたグラフ解析器に置き換えることで，JaSPICE の性能への影響を調査した．

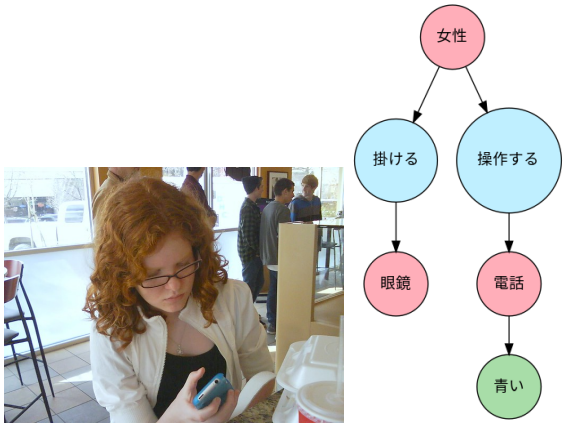
(ii) 同義語によるグラフ拡張を行わない場合：同義語によるグラフ拡張を行わないことで，JaSPICE の性能への影響を調査した．

JaSPICE は $T(G'(\hat{y}))$ と $T(G(\{y_{i,j}\}_{j=1}^{N}))$ をもとに一致する組を調べるため，一致する組がない場合 JaSPICE 値が 0 になることがある．そのため，上記の ablation 条件において，相関係数および $\text{JaSPICE}(\hat{y}, \boldsymbol{y}_i) = 0$ であったサンプル数 M についても調査した．

表 3 に ablation study の結果を示す．条件 (ii) と (iv) を比較すると Pearson, Spearman, Kendall の相関係数において，それぞれ $0.102, 0.139, 0.104$ ポイント下回った．また，M については 84 サンプル下回った．このことから JaSGP の導入が最も性能に寄与していると考えられる．同様に，条件 (i) と (iv)，条件 (iii) と (iv) から，グラフ拡張の導入も性能に寄与していることが確認できる．

表 3: Ablation study (P: Pearson, S: Spearman, K: Kendall)

条件	Parser	グラフ拡張	P	S	K	M
(i)	UD		0.398	0.390	0.309	1465
(ii)	UD	✓	0.399	0.390	0.309	1430
(iii)	JaSGP		0.493	0.524	0.410	1417
(iv)	JaSGP	✓	**0.501**	**0.529**	**0.413**	**1346**



(a) (b)

図 3: 成功例における画像とシーングラフ

5 おわりに

本論文では，日本語での画像キャプション生成に対する自動評価尺度を提案した．本研究の貢献を以下に示す．

- 日本語の画像キャプション生成に対する自動評価尺度 JaSPICE を提案した．
- UD [5] を用いる SPICE [4] とは異なり，係り受け構造と述語項構造に基づくルールベースのグラフ解析器 JaSGP を提案した．
- 同義語を考慮した評価を行うため，同義語を利用したグラフの拡張を導入した．
- JaSPICE はベースライン尺度，ならびに機械翻訳による英訳文から算出された SPICE と比較して，人間による評価との相関係数が高いことを示した．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[1] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In **ECCV**, pp. 417–434, 2020.

[2] Julia White, Gabriel Poesia, Robert Hawkins, et al. Open-domain Clarification Question Generation Without Question Examples. In **EMNLP**, pp. 563–570, 2021.

[3] Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. CapWAP: Image Captioning with a Purpose. In **EMNLP**, pp. 8755–8768, 2020.

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In **ECCV**, pp. 382–398, 2016.

[5] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, et al. Universal Stanford dependencies: A cross-linguistic typology. In **LREC**, pp. 4585–4592, 2014.

[6] 黒橋禎夫, 酒井康行. 国語辞典を用いた名詞句「A の B」の意味解析. 情報処理学会研究報告. NL, 自然言語処理研究会報告, Vol. 129, pp. 109–116, 1999.

[7] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, et al. TextCaps: a Dataset for Image Captioning with Reading Comprehension. In **ECCV**, pp. 742–758, 2020.

[8] Yuichiroh Matsubayashi and Kentaro Inui. Distance-Free Modeling of Multi-Predicate Interactions in End-to-End Japanese Predicate-Argument Structure Analysis. In **COLING**, pp. 94–106, 2018.

[9] 河原大輔, 笹野遼平, 黒橋禎夫, 橋田浩一. 格・省略・共参照タグ付けの基準, 2005. https://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0/doc/rel_guideline.pdf.

[10] Francis Bond, Hitoshi Isahara, Sanae Fujita, et al. Enhancing the Japanese WordNet. In **Workshop on Asian Language Resources**, pp. 1–8, 2009.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In **ACL**, pp. 311–318, 2002.

[12] Chin Lin. ROUGE: A Package For Automatic Evaluation Of Summaries. In **ACL**, pp. 74–81, 2004.

[13] Satanjeev Banerjee, et al. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In **IEEvaluation@ACL**, pp. 65–72, 2005.

[14] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In **CVPR**, pp. 4566–4575, 2015.

[15] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In **LREC**, pp. 3603–3609, 2020.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In **ICML**, pp. 2048–2057, 2015.

[17] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image Captioning: Transforming Objects into Words. In **NeurIPS**, Vol. 32, pp. 11137–11147, 2019.

[18] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In **CVPR**, pp. 10578–10587, 2020.

[19] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, et al. Dual-Level Collaborative Transformer for Image Captioning. **AAAI**, Vol. 35, No. 3, pp. 2286–2293, 2021.

[20] Jingyu Li, Zhendong Mao, et al. ER-SAN: Enhanced-Adaptive Relation Self-Attention Network for Image Captioning. In **IJCAI**, pp. 1081–1087, 2022.

[21] 加藤駿弥, Chenhui Chu, 黒橋禎夫. 抽象度を制御可能なエンティティレベルの画像キャプション生成. 言語処理学会第 28 回年次大会, pp. 1349–1354, 2021.

[22] Motonari Kambara, et al. Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions. **IEEE RAL**, Vol. 6, No. 4, pp. 8371–8378, 2021.

[23] Tadashi Ogura, Aly Magassouba, Komei Sugiura, Tsubasa Hirakawa, et al. Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder-Decoder Network. **IEEE RAL**, Vol. 5, No. 4, pp. 5945–5952, 2020.

[24] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. Multimodal Attention Branch Network for Perspective-Free Sentence Generation. In **CORL**, pp. 76–85, 2019.

[25] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. **arXiv preprint arXiv:2107.06912**, 2021.

[26] Tsung Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, et al. Microsoft COCO: Common Objects in Context. In **ECCV**, pp. 740–755, 2014.

[27] Peter Young, et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. **TACL**, Vol. 2, pp. 67–78, 2014.

[28] Piyush Sharma, Nan Ding, et al. Conceptual captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In **ACL**, pp. 2556–2565, 2018.

[29] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In **ACL**, pp. 417–421, 2017.

[30] Takashi Miyazaki and Nobuyuki Shimizu. Cross-Lingual Image Caption Generation. In **ACL**, pp. 1780–1790, 2016.

[31] Ron Mokady, Amir Hertz, and Amit Bermano. ClipCap: CLIP Prefix for Image Captioning. **arXiv preprint arXiv:2107.06912**, 2021.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention is all you need. In **NeurIPS**, Vol. 30, pp. 5998–6008, 2017.

[33] Peter Anderson, Xiaodong He, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In **CVPR**, pp. 6077–6086, 2018.

[34] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating Automatic Metrics for Image Captioning. In **EACL**, pp. 199–209, 2017.

A 付録

A.1 関連研究

画像キャプション生成の研究は盛んに行われており [16–21]，生活支援ロボットへの応用も行われている [22–24]．画像キャプション生成に関するサーベイ論文である [25] では，画像キャプション生成モデルや標準的なデータセット，および評価尺度についての包括的な総括がなされている．

画像キャプション生成における代表的な尺度として，BLEU [11], ROUGE [12], METEOR [13], CIDEr [14] が挙げられる．また英語における代表的な尺度としては，これらに加え SPICE [4] が挙げられる．

英語の画像キャプション生成タスクにおける標準的なデータセットとして，MS COCO [26] や Flickr30K [27]，CC3M [28] がある．また日本語の画像キャプション生成タスクにおける標準的なデータセットとして，MS COCO に日本語のキャプションが付与された STAIR Captions [29] や YJ Captions [30] がある．

A.2 実験設定

本研究では，日本語の画像キャプション生成において標準的なコーパスである STAIR Captions [29] を用いた．本実験では STAIR Captions を訓練集合，検証集合，テスト集合に分割し，それぞれの集合は 413915, 37269, 35594 個のキャプションを含む．

人間による評価は，与えられた 1 枚の画像と，対応するキャプションの組に対して，キャプションの適切さを 5 段階で評価したものである．ここで，人間による評価はクラウドソーシングサービスを用いて 100 人の評価者から収集した．

評価には，モデルの出力した各キャプション，また $\{y_i\}$ と $\{y_{\mathrm{rand}}\}$ を含む合計 21227 個のキャプションを使用した．ここで，y_i は i 番目の画像に対する $\{y_{i,j}\}_{j=1}^{5}$ のうち 1 つを無作為に抽出したキャプションであり，y_{rand} は全画像における正解キャプションのうち 1 つを無作為に抽出したキャプションである．

評価に使用するモデルは，画像キャプション生成において標準的なモデルを採用した．使用したモデルは SAT [16], ORT [17], $\mathcal{M}^2$-Transformer [18], DLCT [19], ER-SAN [20], $\mathrm{ClipCap}_{\mathrm{mlp}}$ [31], $\mathrm{ClipCap}_{\mathrm{trm}}$, および 3 種類の Transformer [32] である．ここで，

(a) (b)

図 4: 失敗例における画像とシーングラフ

$\mathrm{ClipCap}_{\mathrm{mlp}}, \mathrm{ClipCap}_{\mathrm{trm}}$ はそれぞれ, ClipCap において Mapping Network を MLP, Transformer としたものであり，また使用した 3 種類の Transformer は Bottom-up Feature [33] を入力に用いた 3, 6, 12 層からなる．

また上記実験に加えて，日本語で学習したモデルの出力文を機械翻訳で英訳し，英訳文から算出した SPICE 値と人間による評価との相関係数を計算する．ここで，機械翻訳には JParaCrawl [15] で訓練した Transformer，および一般的な機械翻訳システム[2)] を用いた．

[4] では，自動評価尺度の評価のため $\{(s_S^{(i)}, s_H^{(i)})\}_{i=1}^{N}$ に対する相関係数と，モデルごとの平均値に対する相関係数 $\{(\bar{s}_S^{(j)}, \bar{s}_H^{(j)})\}_{j=1}^{J}$ が提示されている．ここで $s_S^{(i)}$ を i 番目のキャプションに対する SPICE 値とし，J をモデルの個数とする．しかし，一般に J は極めて小さいため，相関係数を計算する上で適切であるとは言えない．そのため前者は後者より適切であると考えられ，実際 [34] もキャプションごとの相関係数のみを評価に用いている．したがって，本論文ではモデルごとの平均値を用いた相関係数の計算は行わず，$\{(s_J^{(i)}, s_H^{(i)})\}_{i=1}^{N}$ に対する相関係数を評価に用いる．

A.3 失敗例

図 4 に提案尺度の失敗例を示す．図は入力画像と $\hat{y}_k$「皿に料理が盛られている」に対するシーングラフである．図における $y_{k,1}$ は「パンにハムときゅうりとトマトとチーズが挟まっている」であり，$s_H^{(k)} = 5$ であったのに対し，$\mathrm{JaSPICE}(\hat{y}, \boldsymbol{y}_k) = 0$ であった．図 4 の例では，$y_{k,1}$ が「パン」や「ハム」という語を用いているのに対して，$\hat{y}$ が「料理」というそれらの上位語を用いており，表層表現の不一致から低い値が出力されている．

2) https://deepl.com