シーングラフに基づく画像キャプション生成モデルの自動評価と解析

Automatic Evaluation and Analysis of Image Captioning Models Based on Scene Graphs

田中 励雄 *1　和田 唯我 *1　杉浦 孔明 *1
Reo Tanaka　Yuiga Wada　Komei Sugiura

*1慶應義塾大学
Keio University

Image captioning studies rely heavily on automatic evaluation metrics such as BLEU and METEOR, which are based on n-grams. However, these metrics have shown poor correlation with human evaluations, leading to the proposal of alternative metrics such as JaSPICE. JaSPICE has only been validated for a general image captioning task without an error analysis. In this paper, we analyze JaSPICE for a fetching instruction generation task and identify its errors for an image captioning task. We conducted experiments on STAIR Captions and PFN-PIC datasets and JaSPICE outperformed the baseline metrics on the correlation coefficient with human evaluation.

1. はじめに

画像キャプション生成は幅広く研究が行われており，生活支援ロボットにおいては指示文付与タスクへと応用されている [Magassouba 19, Ogura 20, Kambara 21]．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おり [Anderson 16]，シーングラフに基づく自動評価尺度として，SPICE [Anderson 16] や JaSPICE [和田 23] が提案されている．しかし，JaSPICE は一般的な画像キャプション生成タスクにおいてのみ人間による評価との相関が検証されており，指示文付与タスクにおいてはその有効性が検証されていない．したがって，指示文付与タスクにおける生成文と人間による評価との相関を解析することは有益である．また，[和田 23] では JaSPICE の失敗例についての解析が行われておらず，エラーを分析することも有用であるといえる．

そこで，本論文では，指示文付与タスクにおける JaSPICE と人間による評価との相関係数を解析し，STAIR Captions [Yoshikawa 17] における JaSPICE のエラーを分析する．本論文における貢献は以下の通りである．

- 指示文付与タスクにおける JaSPICE と人間による評価との相関係数を解析する．
- STAIR Captions における JaSPICE のエラー分析を行う．

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 [Matsubayashi 18]．
- **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．

連絡先: 田中励雄，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，lixiong0218@keio.jp

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．

3. JaSPICE

JaSPICE は SPICE [Anderson 16] 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．JaSPICE は，Japanese Scene Graph Parser (JaSGP) と Graph Analyzer (GA) の二つのモジュールから構成される．

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．

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．また，JaSPICE ではゼロ代名詞の影響を軽減するため，ヒューリスティックな方法でゼロ照応解析を行う．

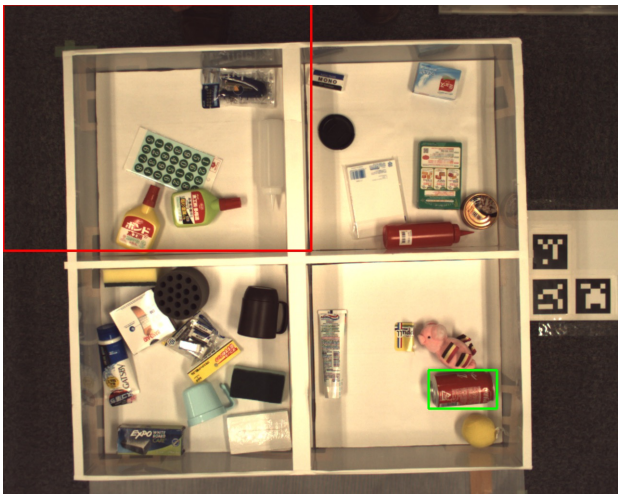
3.3 Graph Analyzer

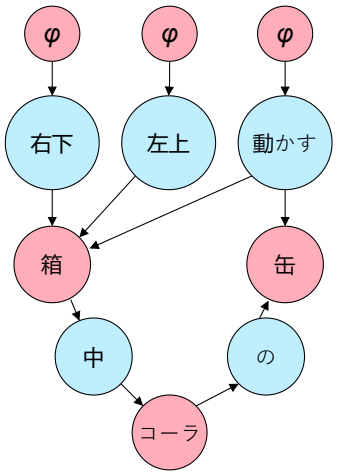
GA における入力は $\{y_{i,j}\}_{j=1}^{N}$ から得られた $\{G(y_j)\}_{j=1}^{N}$ と $\hat{y}$ から得られた $G(\hat{y})$ である．まず，GA では同義語によるノードの追加を行い，$\{y_{i,j}\}_{j=1}^{N}$ に対する $\{G(y_{i,j})\}_{j=1}^{N}$ について，これらを 1 つの $G(\{y_{i,j}\}_{j=1}^{N})$ へと統合する．具体的には，

$G(\{y_{i,j}\}_{j=1}^{N}) := \mathcal{G}$ <$\{O(y_{i,j})\}_{j=1}^{N}, \{E(y_{i,j})\}_{j=1}^{N}, \{K(y_{i,j})\}_{j=1}^{N}$>

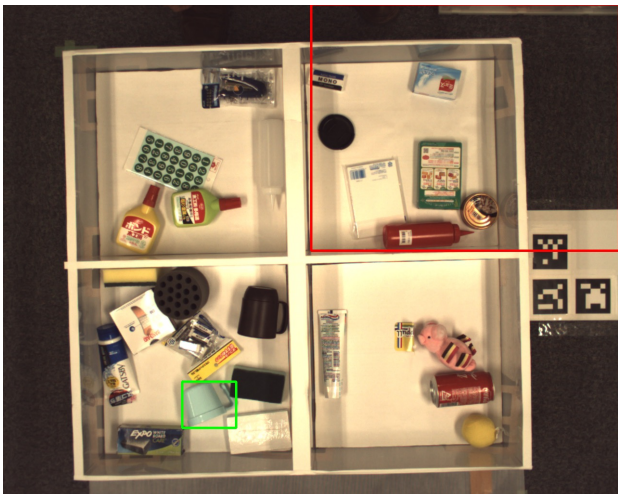
とする．$T(G(x))$ を $T(G(x)) := O(x) \cup E(x) \cup K(x)$ と定義すると，$T(G'(\hat{y}))$ と $T(G(\{y_{i,j}\}_{j=1}^{N}))$ から適合率 P, 再現率 R, および F1 値 F_1 を次のように計算する．

(a)





(b)



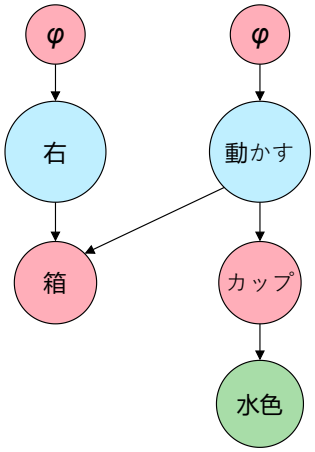


図 1: 成功例における入力画像とシーングラフ．ϕ はゼロ代名詞を表す．(a) $\hat{y}_i$: 「右下の箱の中のコーラの缶を，左上の箱に動かしてください」, JaSPICE$(\hat{y}, \boldsymbol{y}_i) = 0.870$ (b) $\hat{y}_j$:「水色のカップを、右上の箱に動かしてください」, JaSPICE$(\hat{y}, \boldsymbol{y}_j) = 0.385$

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

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

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

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 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．

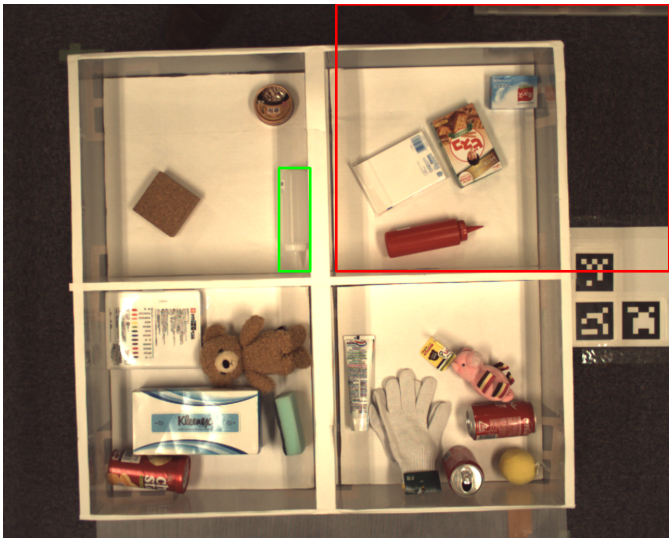
本研究では，日本語の指示文生成において標準的なコーパスである PFN-PIC [Hatori 18] および，日本語の画像キャプション生成において標準的なコーパスである STAIR Captions [Yoshikawa 17] を用いた．本実験では PFN-PIC および STAIR Captions を訓練集合，検証集合，テスト集合に分割した．PFN-PIC におけるそれぞれの集合は 81087，8774，898 個のサンプルを含み，STAIR Captions におけるそれぞれの集合は 413915, 37269, 35594 個のキャプションを含む．

$s_J^{(i)}$ 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．

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．

PFN-PIC での評価および STAIR Captions における評価には，それぞれ，モデルの出力した各キャプション，また $\{y_i\}$ と $\{y_{\text{rand}}\}$ を含む合計 1920 個および 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．

PFN-PIC における評価に用いるモデルは，指示文付与タスク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 [Xu 15]，ORT [Herdade 19], CRT [Kambara 21] である．また STAIR Captions におけ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として，SAT [Xu 15], ORT [Herdade 19], $\mathcal{M}^2$-Transformer [Cornia 20],



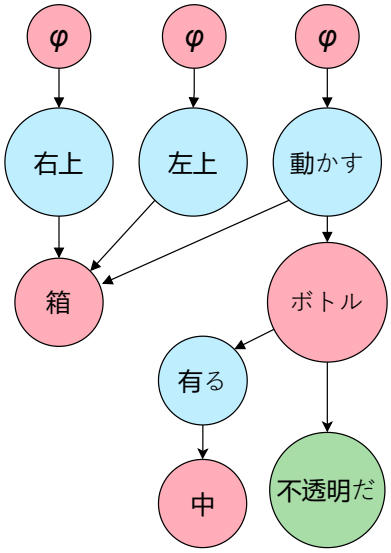


図 2: 失敗例における入力画像とシーングラフ．$\hat{y}_j$「左上の箱の中にある白くて不透明なボトルを，右上の箱に移してください」，JaSPICE$(\hat{y}, \boldsymbol{y}_j) = 0.09$

表 1: PFN-PIC における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 [Papineni 02]	0.484	0.466	0.352
ROUGE [Lin 04]	0.500	0.474	0.365
METEOR [Banerjee 05]	0.423	0.457	0.352
CIDEr [Vedantam 15]	0.416	0.462	0.353
$\text{SPICE}_{\text{service}}$	0.416	0.418	0.316
$\text{SPICE}_{\text{trm}}$	0.427	0.420	0.317
JaSPICE	**0.572**	**0.587**	**0.452**

DLCT [Luo 21], ER-SAN [Li 22], $\text{ClipCap}_{\text{mlp}}$ [Mokady 21], $\text{ClipCap}_{\text{trm}}$, および 3 種類の Transformer [Vaswani 17] を使用した．ここで，$\text{ClipCap}_{\text{mlp}}$, $\text{ClipCap}_{\text{trm}}$ 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 [Anderson 18] 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 [Morishita 20] で訓練した Transformer，および一般的な機械翻訳システム [*1] を用いた．

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 [Papineni 02], ROUGE [Lin 04], METEOR [Banerjee 05], CIDEr [Vedantam 15] を用い，$\text{SPICE}_{\text{trm}}$, $\text{SPICE}_{\text{service}}$ はそれぞれ，JParaCrawl [Morishita 20] で訓練した Transformer の出力文，および一般的な機械翻訳システムの英訳文を用いて算出した SPICE 値を指す．

表 1 より，JaSPICE は Pearson, Spearman, Kendall の相関係数において，それぞれ 0.572, 0.587, 0.452 であり，ベースライン尺度を上回った．また $\text{SPICE}_{\text{trm}}$ と比較して 0.145, 0.167, 0.135 ポイント上回った．同様に，$\text{SPICE}_{\text{service}}$ と比較して，JaSPICE はそれぞれ 0.156, 0.169, 0.136 ポイント上回った．

図 1 に，PFN-PIC において JaSPICE が成功した二つの例を示す．ここで，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．図 1 (a) は入力画像と $\hat{y}_j$「右下の箱の中のコーラの缶を，左上の箱に動かしてください」に対するシーングラフである．ここで，入力画像における緑色の枠線と赤色の枠線は，それぞれ移動物体と目標領域を示している．図 1 (a) における $y_{i,1}$ は「コーラの缶を，左上のケースに動かしてちょうだい」であり，JaSPICE$(\hat{y}, \boldsymbol{y}_i) = 0.870$, $s_H^{(i)} = 5$ であった．この JaSPICE 値は上位 0.3%であるため，一つ目の例において，提案手法による評価は人間による評価に近いといえる．図 1 (b) は $\hat{y}_j$「水色のカップを、右上の箱に動かしてください」に対するシーングラフである．図 1 (b) における $y_{j,1}$ は「左下の箱の中にある水色のカップを、右上の箱に動かしてください」であり，JaSPICE$(\hat{y}, \boldsymbol{y}_j) = 0.385$, $s_H^{(j)} = 5$ であった．この JaSPICE 値は上位 8%であるため，二つ目の例においても提案手法による評価は人間による評価に近いといえる．

JaSPICE の失敗例を $f(s_H^{(i)}, s_J^{(i)}) \geq \theta$ を満たすサンプルと定義する．ここで，$f(s_H^{(i)}, s_J^{(i)}) := \left| \frac{s_H^{(i)}}{\max_i s_H^{(i)}} - \frac{s_J^{(i)}}{\max_i s_J^{(i)}} \right|$ とし，本論文では $\theta = 0.5$ とした．図 2 に，PFN-PIC における失敗例の一例を示す．図 2 は入力画像と $\hat{y}_j$「左上の箱の中にある白くて不透明なボトルを，右上の箱に移してください」に対するシーングラフである．図 2 における $y_{j,1}$ は「白い半透明の円筒ボトル容器を右隣のボックスに動かしてください」であり，$s_H^{(j)} = 5$ であったのに対して，JaSPICE$(\hat{y}, \boldsymbol{y}_j) = 0.09$ であった．図 2 の例では，$f(s_H^{(i)}, s_J^{(i)}) = 0.897 \geq \theta$ であり，$f(s_H^{(i)}, s_J^{(i)})$ が $\{f(s_H^{(i)}, s_J^{(i)})\}_{i=1}^N$ において上位 0.3%の値であったため，人間による評価と JaSPICE 値が乖離しているといえる．図 2 の例が失敗した原因として，以下に示す 2 つが考えられる．

(i) 表層の不一致：$y_{i,1}$ において「ボトル容器」と表現しているのに対し，$\hat{y}_j$ では「ボトル」と表現しており，表層の一部は一致するが，完全には一致しない形態素を用いている．

*1 https://deepl.com

(ii) 存在領域を示す情報の欠如：$\hat{y}_j$ では対象物体が「左上の箱の中にある」ことを示しているのに対して，$y_{i,1}$ には対象物体の存在領域に関する情報が欠如している．

4.3 STAIR Captions におけるエラー分析

本節では STAIR Captions においてエラー分析を行う．STAIR Captions の失敗例については $\theta = 1$ とし，テスト集合に失敗例は 130 サンプル含まれていた．表 2 に失敗例の分類を示す．130 個の失敗例のうち 100 個を調査し，次の 5 種類に大別した．

表 2: 失敗例の分類結果

説明	サンプル数
(i)	46
(ii)	20
(iii)	18
(iv)	10
(v) その他	6

(i) $\hat{y}$ と $\boldsymbol{y}_i$ における単語の粒度の違い：画像中に含まれるある物体，関係，属性に対して，$\boldsymbol{y}_i$ が下位語を使用する一方，$\hat{y}$ は上位語を使用している場合を指す．例えば，$\hat{y}$ が「皿に料理が盛られている」という文であるのに対して，$y_{i,1}$ が「皿に肉が盛られている」という文であるような場合である．この場合，下位語である「肉」に対して $\hat{y}$ は「料理」と上位語を用いて表現しており，表層表現の不一致により不適切な JaSPICE 値が出力される．

(ii) 注目領域の相違：$\boldsymbol{y}_i$ が注目している領域と，$\hat{y}$ が注目している領域が異なる場合を指す．

(iii) 表層の一部は一致するが，完全一致はしない形態素を含む文の比較：例えば，$\hat{y}$ が「テニスラケット」，$y_{i,1}$ が「テニス」を含む文である場合に，$T(G'(\hat{y}))$ と $T(G(\{y_{i,j}\}_{j=1}^{N})$ とで一致する組の数が減少することで，不適切な JaSPICE 値が出力される．

(iv) 評価者による誤り：人間による評価と生成文の質とが乖離している場合を指す．例えば，$\hat{y}_i$ が「紙コップの隣にバナナが置いてある」という不適切な生成文 (「バナ」は誤表記) に対して，$S_H^{(i)} = 5$ と付与された場合である．

表 2 より， ボトルネックは $\hat{y}$ と $\boldsymbol{y}_i$ における単語の粒度の違いによる失敗と言える．そのため，上位語と下位語の関係を考慮したモデルを導入することでボトルネックを軽減することができると考えられる．

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を扱った．本論文における貢献は以下の通りである．

- JaSPICE が PFN-PIC において，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．
- STAIR Captions [Yoshikawa 17] における JaSPICE のエラー分析を行った．

謝辞

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

参考文献

[Anderson 16] Anderson, P., Fernando, B., Johnson, M., and Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation, in *ECCV*, pp. 382–398 (2016)

[Anderson 18] Anderson, P., He, X., et al.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, in *CVPR*, pp. 6077–6086 (2018)

[Banerjee 05] Banerjee, S., et al.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in *IEEvaluation@ACL*, pp. 65–72 (2005)

[Cornia 20] Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R.: Meshed-Memory Transformer for Image Captioning, in *CVPR*, pp. 10578–10587 (2020)

[Hatori 18] Hatori, J., Kikuchi, Y., Kobayashi, S., et al.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, in *ICRA*, pp. 3774–3781 (2018)

[Herdade 19] Herdade, S., Kappeler, A., Boakye, K., and Soares, J.: Image Captioning: Transforming Objects into Words, in *NeurIPS*, Vol. 32, pp. 11137–11147 (2019)

[Kambara 21] Kambara, M., et al.: Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions, *IEEE RAL*, Vol. 6, No. 4, pp. 8371–8378 (2021)

[Li 22] Li, J., Mao, Z., Fang, S., and Li, H.: ER-SAN: Enhanced-Adaptive Relation Self-Attention Network for Image Captioning, in *IJCAI*, pp. 1081–1087 (2022)

[Lin 04] Lin, C.: ROUGE: A Package For Automatic Evaluation Of Summaries, in *ACL*, pp. 74–81 (2004)

[Luo 21] Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.-W., et al.: Dual-Level Collaborative Transformer for Image Captioning, *AAAI*, Vol. 35, No. 3, pp. 2286–2293 (2021)

[Magassouba 19] Magassouba, A., Sugiura, K., and Kawai, H.: Multimodal Attention Branch Network for Perspective-Free Sentence Generation, in *CORL*, pp. 76–85 (2019)

[Matsubayashi 18] Matsubayashi, Y. and Inui, K.: Distance-Free Modeling of Multi-Predicate Interactions in End-to-End Japanese Predicate-Argument Structure Analysis, in *COLING*, pp. 94–106 (2018)

[Mokady 21] Mokady, R., Hertz, A., and Bermano, A.: ClipCap: CLIP Prefix for Image Captioning, *arXiv preprint arXiv:2107.06912* (2021)

[Morishita 20] Morishita, M., Suzuki, J., and Nagata, M.: JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus, in *LREC*, pp. 3603–3609 (2020)

[Ogura 20] Ogura, T., 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)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.: Bleu: a Method for Automatic Evaluation of Machine Translation, in *ACL*, pp. 311–318 (2002)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Polosukhin, K., et al.: Attention is all you need, in *NeurIPS*, Vol. 30, pp. 5998–6008 (2017)

[Vedantam 15] Vedantam, R., Zitnick, L., and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, in *CVPR*, pp. 4566–4575 (2015)

[Xu 15] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., et al.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in *ICML*, pp. 2048–2057 (2015)

[Yoshikawa 17] Yoshikawa, Y., Shigeto, Y., and Takeuchi, A.: STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset, in *ACL*, pp. 417–421 (2017)

[和田 23] 和田唯我, 兼田寛大, 杉浦孔明：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, 言語処理学会第 29 回年次大会 (2023)