# 画像キャプション生成及び物体認識を用いた駄洒落文ランキング手法の評価

浅野歴 [1,2] 谷津元樹 [1] 森田武史 [1,2] 江上周作 [2] 鵜飼孝典 [2,3] 福田賢一郎 [2]
[1] 青山学院大学 [2] 産業技術総合研究所 [3] 富士通株式会社
c5622198@aoyama.jp {yatsu,morita}@it.aoyama.ac.jp
ugai@fujitsu.com {s-egami,ken.fukuda}@aist.go.jp

## 概要

ソーシャルロボットが画像入力より得られる周辺状況に基づいた発話を行う際，駄洒落等のユーモアを付加できれば，ユーザはロボットに対しより高い親近感を持ちうると考えられる．本研究では既存のキャプション生成モデルや物体検出を用いて画像から画像内容に即した語を抽出し，駄洒落文を収集したコーパスから画像の描画内容に即した駄洒落文を選択するためのランキング手法を提案する．

## 1 はじめに

世界規模の少子高齢化の進行と，その結果である高齢者単身世帯の急激な増加は看過できない問題となっている．そのため，人手を要しない人工知能技術による支援は不可欠となる．同技術の代表格となりうるのが，人間らしい振る舞いの可能な対話知能である．その具体例として，言語的ユーモアを交え笑いを通して不安を軽減する能力および視覚入力としての画像からの周辺状況の推定能力が挙げられる．さらに，両者を統合したユーモアの表現能力は人間そのものに近いといえる．

本研究ではユーモア性を含む言語表現として駄洒落に注目した．画像キャプション生成とは，画像を入力として，その内容について記述した自然言語文を出力する問題である [1]．この問題は画像認識および自然言語テキスト生成という２つのタスクを同一のシステムに統合して解く必要がある．

駄洒落を含む画像キャプション生成には，画像とその描画内容に即した駄洒落文のペアを集めたデータセットが必要となる．しかし，現在，そのようなデータセットは存在しておらず，画像の描画内容に即した駄洒落文を手動で作成するコストは高く，データセットの作成には時間を要する．そのため，本研究では画像から駄洒落を含むキャプションを生成する前段階として，既存のキャプション生成モデルや物体検出を活用し，画像に合った駄洒落データベース [2] 内の駄洒落文を選択することを目的とする．

## 2 関連研究

荒木らの研究 [2] では前述した背景においてユーモア理解・生成技術の基盤となるデータセットを作成することを目的として，67,000 件の駄洒落文を収録したコーパスを構築している．本研究では，この駄洒落データベース内に存在する事例より，入力画像が描写している内容に対して最適な駄洒落文を選択する．Chandrasekaran らの研究 [3] では，画像から英語の駄洒落（語呂合わせ）を含むキャプション生成を課題として，画像キャプション生成とコーパス検索に基づく二つの手法を提案している．画像キャプション生成に基づく手法では，画像キャプション生成モデルを用いて得られた単語を，事前に用意した語呂合わせが成り立つ単語や発音が近い単語のリストに含まれる単語に強制的に絞り込み，強制的に絞り込まれた単語を使い画像キャプションを生成する．コーパス検索に基づく手法では，一般的な画像キャプション生成のモデルを用いて得られた単語と，事前に用意した語呂合わせが成り立つ単語や発音が近い単語の両方が含まれる文をコーパスから検索する．

関連研究は，画像から英語の駄洒落を含むキャプション生成を課題としているのに対して，本研究では，駄洒落データベースから画像に合った日本語の駄洒落文を選択することを目的としている．

## 3 提案手法の構成

本研究手法では，最初に画像を入力として，キャプション生成と物体検出を行う．キャプション生成に関しては，日本語画像キャプションデータセット STAIR Captions[4] と MS COCO[5] データセットの画像を用いて日本語キャプション生成モデルを学習

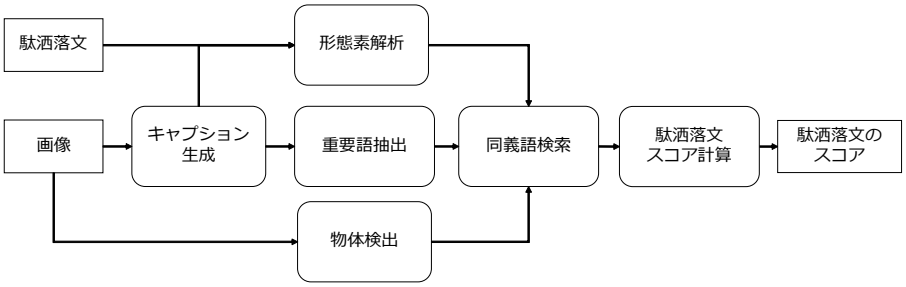


**図 1**　駄洒落文スコア計算の手順

し，このモデルに基づいて画像から日本語のプレーンキャプションを生成する．生成されたプレーンキャプションに対して重要語抽出と形態素解析を行う．重要語抽出では TermExtract[6] を用いてプレーンキャプション含まれる重要な単語を抽出する．形態素解析では MeCab[7] を使用しプレーンキャプションを形態素に区切る．物体検出では物体検出モデル YOLOX[8] を用いて入力画像から物体ラベルを抽出する．日本語 WordNet[9] を用いて，抽出した重要語，形態素，物体ラベルの同義語検索する．駄洒落文，重要語，形態素，物体ラベル，同義語を入力として，駄洒落文のスコアを計算し，ランキングを行い，入力画像の描画内容に即した駄洒落文を選択し，出力する．以下では，提案システムの構成要素の詳細について述べる．

## 3.1　画像キャプション生成

本研究では，Show and Tell モデル [10] を用いて日本語画像キャプションデータセット STAIR Captions[4] を学習した日本語画像キャプション生成モデルを用いる．日本語画像キャプション生成を用いる理由として，画像から駄洒落文を選択するときに日本語画像キャプション生成で生成された文と似た文を探すことで画像にあった駄洒落文を選択できると考えたからである．STAIR Captions は日本語での画像キャプション生成のために人手で構築された最大規模のデータセットであり，これを訓練データとして学習したモデルのキャプション生成精度が，同一の画像を使用する MS COCO[5] における英文キャプションに機械翻訳を行い，得られた文を訓練データとした場合を上回ることが判明している．このことから STAIR Captions は優れたデータセットであるといえるため，本研究では STAIR Captions を使用する．日本語画像キャプション生成モデルから得られたキャプションをプレーンキャプションと呼ぶ．

## 3.2　物体検出

物体検出とは画像や動画に写っている物体のラベル（物体ラベル）を検出するタスクである．物体検出を本研究に用いる理由は，画像に描写されている物体の呼称がプレーンキャプションの生成結果に含まれなかった場合でも，物体検出の結果に含まれていれば駄洒落文選択に利用できる語の数が増加する場合があるためである．本研究では物体検出を行うために，リアルタイム物体検出アルゴリズムである YOLO[11] の改良型である YOLOX[8] を用いる．

## 3.3　形態素解析

プレーンキャプションから形態素解析エンジン MeCab を用いて形態素を抽出する．形態素解析を本研究に用いることにより，プレーンキャプションから抽出した形態素と駄洒落文から抽出した形態素の共通部分を抽出し，画像の描画内容に即した駄洒落文を選択するためのスコアとして利用できる．

## 3.4　重要語抽出

生成されたプレーンキャプションから TermExtract[6] を用いて重要語を抽出することにより，プレーンキャプションに含まれる単語の重要度を求めることができる．重要度の高い語を多く含む候補文を優先することにより，より画像の内容に適した駄洒落文を選択できると考えられる．

## 3.5　同義語検索

プレーンキャプションから形態素解析により抽出された単語，物体検出により画像から抽出された物体ラベル，プレーンキャプションから重要語抽出により抽出された重要語の各同義語を，日本語 WordNet[9] を用いて検索する．これらの同義語集合に重みを与えることにより，画像の描写に関連性の高い駄洒落文をより多く比較の対象とすることができると考えられる．

### 3.6 駄洒落文選択

図 1 に，本研究が提案する駄洒落文スコア計算の手順を示す．駄洒落文選択は以下の 1 から 8 の手順で行う．式 (1) から式 (8) における記号の説明を表 1 に示す．

1. 式 (1) より駄洒落文に含まれるプレーンキャプションから抽出した形態素に重み付けをする．

$$f_t(c, d, w_t) = w_t \cdot |T_c \cap T_d| \quad (1)$$

2. 式 (2) より駄洒落文に含まれるプレーンキャプションから抽出した形態素の同義語に重み付けをする．

$$f_{ts}(c, d, w_{ts}) = w_{ts} \cdot \left|\{S_{t_c} \cap T_d \middle| t_c \in T_c\}\right| \quad (2)$$

3. 式 (3) より駄洒落文に含まれるプレーンキャプションから抽出した重要語に重み付けをする．

$$f_{it}(c, d, w_{it}) = w_{it} \cdot |IT_c \cap T_d| \quad (3)$$

4. 式 (4) より駄洒落文に含まれるプレーンキャプションから抽出した重要語の同義語に重み付けをする．

$$f_{its}(c, d, w_{its}) = w_{its} \cdot \left|\{S_{t_c} \cap T_d \middle| t_c \in IT_c\}\right| \quad (4)$$

5. 式 (5) より駄洒落文に含まれる画像から抽出した物体ラベルに重み付けをする．

$$f_o(i, d, w_o) = w_o \sum_{o \in O_i \cap T_d} CS_o \quad (5)$$

6. 式 (6) より駄洒落文に含まれる画像から抽出した物体ラベルの同義語に重み付けをする．

$$f_{os}(i, d, w_{os}) = w_{os} \sum_{o \in \{o \in O_i | s \in S_o \wedge s \in T_d\}} CS_o \quad (6)$$

7. 式 (7) より式 (1) から式 (6) により求めた重みの合計値を駄洒落文のスコアとして計算する．

$$\begin{aligned} f_s(i, d, \boldsymbol{w}) = & f_t(IC_i, d, w_t) + f_{ts}(IC_i, d, w_{ts}) + \\ & f_{it}(IC_i, d, w_{it}) + f_{its}(IC_i, d, w_{its}) + \\ & f_o(i, d, w_o) + f_{os}(i, d, w_{os}) \end{aligned} \quad (7)$$

8. 式 (8) よりスコアの高い上位 n 件の駄洒落文とそのスコアを出力する．

$$\begin{aligned} & \text{DajareSearch}(i, n, \boldsymbol{w}) = \{(d_1, s) : d_1 \in D \wedge \\ & |\{d_2 \in D \; s = f_s(i, d_1, w) < f_s(i, d_2, w)\}| < n\} \end{aligned} \quad (8)$$

本研究では $w_t : 50$，$w_{ts} : 25$，$w_{it} : 100$，$w_{its} : 50$，$w_o : 500$，$w_{os} : 250$ とする．図 3 に入力[1]及び表 3 に出力の例を示す．

1) 出典:https://www.flickr.com/photos/mjk4219/2118125339/

**表 1** 式 (1) から式 (8) における記号の説明

| 記号 | 説明 |
|---|---|
| $i$ | 画像ファイル |
| $d$ | 駄洒落文 |
| $c$ | 画像キャプション |
| $n$ | 取得する駄洒落文とスコアの数 |
| $D$ | 駄洒落文のセット |
| $\boldsymbol{w}$ | スコアの重みベクトル<br>$\boldsymbol{w} = (w_t, w_{ts}, w_{it}, w_{its}, w_o, w_{os})$ |
| $w_t$ | 形態素スコアの重み |
| $w_{ts}$ | 形態素の同義語スコアの重み |
| $w_{it}$ | 重要語スコアの重み |
| $w_{its}$ | 重要語の同義語スコアの重み |
| $w_o$ | 物体ラベルスコアの重み |
| $w_{os}$ | 物体ラベルの同義語スコアの重み |
| $IC_i$ | 画像ファイル $i$ のキャプション |
| $T_{text}$ | $text$ 中のトークンのセット |
| $IT_{text}$ | $text$ 中の重要語のセット |
| $S_{token}$ | $token$ の同義語セット |
| $CS_o$ | 画像ラベル $o$ の確信度を 10 倍し，小数点以下を切り捨てた値 |

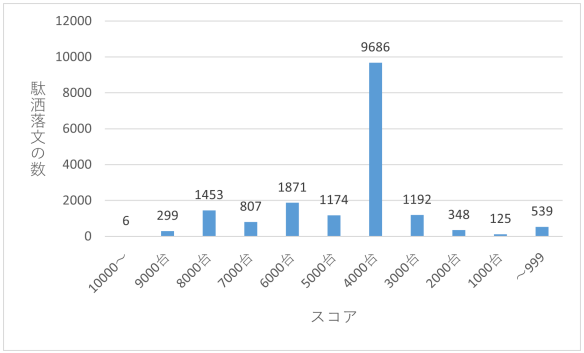


**図 2** 駄洒落文 17500 文におけるスコアの度数分布

## 4 実験と評価

### 4.1 実験方法

MSCOCO 内の画像 3500 枚を対象に各画像に対し，スコアの高い上位 5 件の駄洒落文とそのスコアとその内訳，画像に対するプレーンキャプション，物体ラベル，同義語を取得した．本実験で得られた 17500 文の駄洒落に対しての分析を表 2 に示す．1 行目にはスコアの要因となる 6 つの指標を示し，2 行目には 17500 文の内，スコアが使われた駄洒落文の数を示す．3〜8 行目には 1 文の駄洒落文に対して，含まれていた単語の数の最大，最小，平均，中央値，最頻値，分散である．図 2 は 17500 文の駄洒

**表 2**　画像 3500 枚の入力により得られる駄洒落文 17500 文の各要素の集計

| | 形態素 | 形態素の同義語 | 重要語 | 重要語の同義語 | 物体ラベル | 物体ラベルの同義語 |
|---|---|---|---|---|---|---|
| 使用した文の数 | 16734 | 3263 | 8204 | 2579 | 16515 | 3212 |
| 最大 | 10 | 4 | 3 | 4 | 3 | 3 |
| 最小 | 0 | 0 | 0 | 0 | 0 | 0 |
| 平均 | 3.198 | 0.231 | 0.538 | 0.167 | 1.108 | 0.225 |
| 中央値 | 3 | 0 | 0 | 0 | 1 | 0 |
| 最頻値 | 3 | 0 | 0 | 0 | 1 | 0 |
| 分散 | 3.039 | 0.280 | 0.389 | 0.188 | 0.209 | 0.264 |

**図 3**　入力画像の例

**表 3**　駄洒落文の出力結果

| 駄洒落文 | スコア |
|---|---|
| 猫がベッドで寝ころぶ | 4850 |
| 猫カフェにいるのは，猫か！へぇ〜 | 4800 |
| ミケ猫には，眉間がある | 4775 |
| 古民家にいる猫見んか？ | 4750 |
| タマ（猫？）が玉にじゃれる | 4750 |

落文のスコアの度数分布である．スコアの中央値は 4750 となった．

### 4.2　考察

平均を見たときにプレーンキャプションの形態素が最大の 3.198 となった．これは重要語や物体ラベル，同義語にはない助詞の「は，が，を」などが含まれているので他に比べて相対的に多い．2 番目に平均が高い物体ラベルは本研究では物体ラベルの重み，つまり $w_o$ を 1 番高く設定しているため，物体ラベルを含んでいる駄洒落文が多く選択されたと考える．それぞれの同義語に関しては，重みを低く設定しているが同義語が寄与している駄洒落文が少なからずあることがあることが分かった．しかし，同義語などを提案手法に用いたことによって画像に対して適切な，ユーモアのある駄洒落文を選択できているかの評価ではない．図 2 ではスコアが 4000 台の駄洒落文の数が 1 番多くなった．これは，本研究では物体ラベルの重みを 500 と高く設定しており，物体の確信度が 0.8 を超えると，その物体ラベルを含んでいる駄洒落文のスコアが 4000 を超えるからだと考える．

## 5　まとめと課題

本研究では既存のキャプション生成モデルや物体検出を活用し画像に合った駄洒落文を選択するためのランキング手法を提案した．提案手法では，STAIR Captions データセットより学習した日本語キャプション生成モデルよりプレーンキャプションを生成し，得られたキャプションから重要語および他の形態素を抽出し，画像からの物体検出を行い画像に写っている物体の名前（物体ラベル）を得た．このようにして得られた単語に対して，同義語取得を行った．物体ラベル及び重要語を重視して重み付けをした．出力として駄洒落データベースにおいて重みの和が最大となる駄洒落文を選択した．実験では，駄洒落文のスコアの要因となっている 6 つについて分析を行った．その結果どの指標も駄洒落文選択に寄与していることが分かった．

今後の課題として，駄洒落文選択における重みベクトルを機械学習により求めることが挙げられる．また，同義語などを提案手法に入れたことによってよりよい駄洒落文が選択できているかの評価．さらには，本研究の提案手法は，駄洒落データベースを参照しているため，出力する駄洒落文が限定されており，新しい駄洒落文を生成できないことも課題である．今後は，キャプション生成手法を用いて駄洒落データベース内の駄洒落文だけでなく多種多様の駄洒落文を生成できるようにしたい．さらに，画像に写っている人の動作や位置関係に関する動詞を検出可能にすることで，より画像の描写内容に即した駄洒落文の選択や生成が可能になると考えられる．

## 謝辞

本研究は JSPS 科研費 JP21K12007 の助成を受けたものです．本研究成果の一部は，国立研究開発法人新エネルギー・産業 技術総合開発機構 (NEDO) の委託業務 (JPNP20006) の結果得られたものです．

## 参考文献

[1] 牛久祥孝. 画像に関連した言語生成の取組み. 人工知能, Vol. 34, No. 4, pp. 483–491, 2019.

[2] 荒木健治, 内田ゆず, 佐山公一, 谷津元樹. 駄洒落データベースの構築及び分析. 人工知能学会第 2 種研究会ことば工学研究会資料 SIG-LSE-B702-3, pp. 13–24, 2017.

[3] Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. Punny captions: Witty wordplay in image descriptions. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 770–775. Association for Computational Linguistics, 2018.

[4] 吉川友也, 重藤優太郎, 竹内彰一. Stair captions: 大規模日本語画像キャプションデータセット. 言語処理学会第 23 回年次大会 (NLP2017), 2017.

[5] T.-Y. Lin, M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context. In **European Conference on Computer Vision (ECCV 2014)**. Springer, Cham, 2014.

[6] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)**, pp. 230–237, 2004.

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. In **arXiv preprint arXiv:2107. 08430**, 2021.

[9] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki. Development of japanese wordnet. In **the 6th Edition of its Language Resources and Evaluation Conference (LREC 2008)**, Marrakech, 2008.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In **CVPR 2015**, pp. 3156–3164, 2015.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In **CVPR 2016**, 2016.

# 付録






| **駄洒落文** | バスケットの試合に馬場助っ人 | 電車の中で大便したことについて、代弁してくれる人はいま | バスケットの試合に馬場助っ人 | 快男子が怪談してる！ | サラウンドの音楽に合わせて皿運動！ | みなさまの気体に固体て液体です |
|---|---|---|---|---|---|---|
| スコア | 13700 | 13300 | 11600 | 250 | 300 | 300 |
| $f_t$ | 100 | 200 | 100 | 100 | 200 | 200 |
| $f_{ts}$ | 0 | 0 | 0 | 50 | 0 | 0 |
| $f_{it}$ | 100 | 100 | 0 | 0 | 100 | 100 |
| $f_{its}$ | 0 | 0 | 0 | 100 | 0 | 0 |
| $f_o$ | 13500 | 13000 | 11500 | 0 | 0 | 0 |
| $f_{os}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| プレーンキャプション | 赤い服を着た女性が馬に乗っている | 道路にたくさんの車が走っている | 道路の脇に自転車が停められている | 男性がスケートボードを持っている | 白い皿の上にホットドッグが乗っている | 赤いスーツケースの中に白い液体が入っている |
| 物体ラベル | バス：0.96, 馬：0.93, 人： 0.91 | 人：0.91, 電車：0.91, 車：0.86 | 人：0.92, 馬：0.90, バス：0.60 | 信号機：0.89 | ホットドッグ：0.97 | スーツケース：0.96 |
| 同義語 | | | | 男子，男 | | |

**図 4**　本実験で得られた 17500 の駄洒落文の内スコアが高いトップ 3 とスコアの低いトップ 3 の画像と駄洒落文，スコア，スコアの内訳，プレーンキャプション，物体ラベル，同義語を示したものを左からスコアが高い 1 位，2 位，3 位，スコアの低い 1 位，2 位，3 位の順に示す．$f_t$：形態素のスコア，$f_{ts}$：形態素の同義語のスコア，$f_{it}$：重要語のスコア，$f_{its}$：重要語の同義語のスコア，$f_o$：物体ラベルのスコア，$f_{os}$：物体ラベルの同義語のスコアである．スコアの低いトップ 3 を見るとスコアが低い要因として，物体ラベルによるスコアがないことが要因として挙げられる．これは，物体ラベルとして「信号機，ホットドッグ，スーツケース」を検出できているが駄洒落データベース内にその単語が含まれる駄洒落文がないことでスコアが低い駄洒落文が選ばれている．スコアの高いトップ 3 では物体ラベルのスコアが極端に高くなっている．これは，スコアの高い 1 位を例に見ると，物体ラベルとして「バス，馬，人」が得られているが駄洒落文との部分一致でスコアを加算しているので「バスケット，馬鹿」など「バス，馬」との関係がない形態素が含まれている駄洒落文のスコアが高くなってしまい，適切ではない駄洒落文が選択された．