画像説明文生成に向けた物体間の関係の認識

村岡 雅康 † Sumit Maharjan† 齋藤 真樹 ‡
山口 光太 ‡ 岡崎 直観 † 岡谷 貴之 ‡ 乾 健太郎 †
東北大学大学院 情報科学研究科 †
東北大学大学院 工学研究科 ‡
{muraoka, sumit, okazaki, inui}@ecei.tohoku.ac.jp†
{msaito, kyamagu, okatani}@vision.is.tohoku.ac.jp‡

1 はじめに

近年，画像から説明文を生成する研究が大きな注目を集めている．例えば，多層畳み込みニューラルネットワーク (CNN) と再帰ニューラルネットワーク (RNN) を組み合わせ，深層学習の枠組みで説明文生成を行う研究が多数提案されている [4, 7, 8, 9, 15]．これらの研究では，CNN を用いて入力画像の抽象的な素性を取り出し，取り出した素性を RNN 言語モデルでデコードすることで説明文生成を行う．大量の訓練事例を活用して様々な説明文を生成できるため，一見すると画像中の物体間の関係やシーンを認識しているように見える．しかしながら，これらの end-to-end の手法は内部が多層ニューラルネットワーク（DNN）で複雑化しているため，実際の挙動の解析が難しい．また，同様の理由から誤生成を修正する戦略が立てられない．

物体認識が人間にせまる精度で行えるようになった今，画像理解に向けた次のステップは物体間の関係の認識であると考える．例えば図 1 の画像を目にした時，人間は「男性がスケートボードに乗っている」,「スケートボードはテーブルの上にある」などと説明できる．物体間の関係を認識することで，画像説明文の質を向上させるだけでなく，物体間の関係を考慮した物体認識や SVO タプルによる画像の意味的な検索なども実現できる．

ところが，物体間の関係認識を行う研究は少ない．Elliott ら [6]，Kong ら [10]，Lin ら [11] は物体間の関係として位置関係を表す関係 (*close_to* や *on_top_of* など) を数種類あらかじめ人手で定義し，物体間の関係をそれらのいずれかに分類している．しかし，物体間の関係は位置関係だけではなく，*stand_at* や *throw* など，状態や動作を表す関係もある．また，Aditya ら [1] は画像に説明文が付与されたデータセットから様々な関係を抽出し，物体を頂点，物体間の関係を辺とするグラフを構築している．この方法は画像間の関係を説明文のみから収集するため，物体間の関係として不適切なものが取れてしまう可能性がある．

The skateboarder is putting on a show using the picnic table as his stage.
A skateboarder pulling tricks on top of a picnic table.
A man riding on a skateboard on top of a table.
A skate boarder doing a trick on a picnic table.
A person is riding a skateboard on a picnic table with a crowd watching.

図 1: MSCOCO データセット

こうした背景のもと本研究では，物体間の関係を表しうる表現の収集と，画像中の物体の関係認識に取り組む（図 2）．本研究では MSCOCO [12][1] と呼ばれる画像に説明文と物体の位置情報が付与されたデータを用いる．このデータには，画像中の物体を表す矩形とラベル，及び説明文が付与されているが，画像中の物体と説明文中の参照表現との対応までは付与されていない．そこで，統計的機械翻訳に基づくアライメント手法で，画像中の物体と説明文中の参照表現を自動的に対応づける．

説明文の構文解析とアライメントの結果に基づき，物体間の関係を説明しうる多様な表現を自動的に獲得する．これにより，先行研究 [6, 10, 11] で導入されていた物体間の関係の定義は不要となる．さらに，アライメント結果を物体間の関係認識の訓練データと見なし，画像中の 2 つの物体が与えられた時に，その関係を説明しうる表現を推定する分類器を構築する．既存研究 [6, 10, 11] は物体間の関係として単一のラベルを仮定していたが，一般的には物体間には複数の関係が成立し，それらには依存関係がある．例えば，ある物体間において関係 *ride_on* が成立する際，同時に関係 *ride* や関係 *on* も成立する可

[1] http://mscoco.org/

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

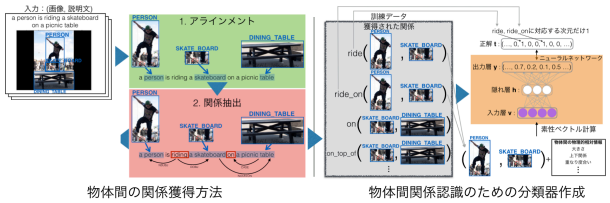

図 2: 提案手法

能性が高い．この問題に対処するため，本研究は隠れ層を持つニューラルネットワークで関係認識の分類器を構築する．

2 提案手法

本研究の目的は説明文が付与された画像のコーパスから物体間に成立しうる関係表現を獲得すること，および物体間の関係認識のための分類器を構築することである．本研究ではこれを次の 3 つの処理に分解して行う．

1. アライメント：画像中の物体と説明文中の言語表現の対応付け
2. 関係抽出：説明文中における物体間の関係の抽出
3. 分類器作成：物体間関係認識のための分類器の作成

以下では，利用するコーパスの説明に続けて，これらの処理を説明する．

2.1 MSCOCO データセット

MSCOCO[12] とは Microsoft 社が Creative Commons Attribution 4.0 License [2] および Flickr Terms of Use [3] 遵守のもと無料で公開しているデータセットである．このデータセットは 123,287 画像からなる．各画像には図 1 のように最低 5 文が人手で付与されている．さらに MSCOCO のデータセットは，人間が画像を説明する際によく用いる 90 種類のカテゴリに属する物体を人手で同定している．同定された物体には図 1 の青線のような矩形 (=**b**ounding **box**, bbox)(x, y, w, h) が付与される．ここで x および y は bounding box の左上の座標，w は x 軸方向の幅，h は y 軸方向の高さをそれぞれ表す．このように画像中の物体のカテゴリと位置を特定するタスクを物体認識と呼ぶ．本研究では物体認識済みのデータセットを用い，物体認識と関係認識の問題を切り分ける．

画像に付与された説明文には画像中の物体間の関係が記述される．例えば，図 1 の 3 番目の文から中央の男性 (a man) とスケートボード (a skateboard) は *ride_on* の関係があると読み取れる．したがって，この画像と説明文が組になったデータセットから物体間の関係としてよく用いられる様々な関係を大量に獲得することができる．

[2] https://creativecommons.org/licenses/by/4.0/legalcode

[3] https://info.yahoo.com/legal/us/yahoo/utos/utos-173.html

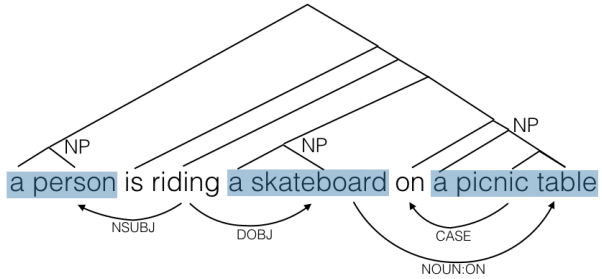

図 3: Stanford CoreNLP の出力結果

2.2 アライメント

物体のカテゴリは 90 種類に限定されているが，説明文中ではある物体を参照するのに様々な表現が用いられる．そのため画像中の物体とそれらの説明文中における表現との対応関係を求める必要がある．図 1 では，物体 **PERSON** は説明文中で **man, person, skateboarder, skate boarder** という表現で参照されている．

本研究ではこれをカテゴリ名の集合から説明文への翻訳問題とみなし，統計的機械翻訳における単語アラインメント手法である IBM Model [2] を用いて求める．IBM Model の翻訳確率を用いることで，P(説明文中の単語 w| 物体カテゴリ名 c) を，画像中に登場する物体のカテゴリ名の集合とその画像の説明文が並置された訓練データから学習する．例えば，以下のようなカテゴリ名の集合と説明文のペアが現れた時，

PERSON, SKATEBOARD — a man is riding a skateboard

PERSON, DONUT — a man who is eating a donut

IBM Model は P(man|PERSON) の確率が高くなるように学習を進める．最終的に得られた翻訳確率 $P(w|c)$ を用いて，説明文中の各単語 w に対して $P(w|c) \geq \alpha$ となるすべてのカテゴリを求める（α はパラメータ）[4]．

2.3 関係抽出

物体が対応付けられた説明文から関係の抽出を行う．例えば，以下の入出力を想定する．

入力例 (アラインメント結果)

a man/PERSON is riding a skateboard/SKATEBOARD on a picnic table/DINING_TABLE

出力例 (物体間の関係)

ride(a man/PERSON, a skateboard/SKATEBOARD),
on(a skateboard/SKATEBOARD,
a picnic table/DINING_TABLE)

本論文において，大文字はカテゴリ名を表し，’ / ’ の左側の単語または名詞句に右側の物体が対応することを意味する．また，*ride(a_man/PERSON, a skateboard/SKATEBOARD)* は PERSON カテゴリの物体

[4] 本研究では IBM Model の実装として GIZA++ [14] (https://github.com/moses-smt/giza-pp) を用いた．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 獲得した関係表現上位 10 件

物体間関係	事例数 (%)	物体間関係	事例数 (%)
on	19,666 (12.58)	of	4,096 (2.62)
in	14,300 (9.15)	next_to	3,974 (2.54)
with	13,047 (8.35)	ride	3,711 (2.37)
hold	5,136 (3.29)	sit_on	3,265 (2.09)
at	4,345 (2.78)	on_top_of	2,393 (1.53)

(説明文中では *a man*) が SKATEBOARD カテゴリの物体 (説明文中では *a skateboard*) と *ride* の関係にあることを意味する．

まず，Stanford CoreNLP [13][5] の構文解析結果を用いて，句構造の高さが h_{tree} 以下の名詞句のうち，高さ最大の名詞句を抽出する (図 3 の青色の部分)．その後 Stanford CoreNLP の係り受け解析結果から，物体が対応付けられている名詞 (句) の係り受けパスを辿り，以下の 2 種類のパタンで関係表現を抽出する．

- 動詞 [句](NSUBJ の子, DOBJ の子)
- 前置詞 [句](CASE の親の親，CASE の子)

特に，動詞句および前置詞句の関係を獲得するため，以下の点に注意する．

1. NSUBJ および DOBJ の親となっている動詞と COMPOUND:PRT[6] の関係にある単語を動詞句として統合する．
2. CASE[7] の親の親が動詞で (品詞タグで VB から始まるもの)，その動詞が直接目的語 (係り受けタグで DOBJ) を持たないときのみ，その動詞と前置詞を組み合わせたものを新しい関係とし，その動詞の主語 (係り受けタグで NSUBJ) を抽出する関係の第 1 引数とする．
3. *on_top_of* のような複合前置詞のうち一般的によく用いられる 58 種類を考慮する [8]．

結果として，合計 156,293 事例および 5,153 種類の関係が得られた (アラインメント時における翻訳確率の閾値 $\alpha = 0.64$，句構造の高さ $h_{tree} = 3$ として抽出を行った)．獲得した関係表現のうち，事例数の多い上位 10 件を表 1 に示す．前置詞 (*on* や *in* など) および空間的な関係 (*next_to* や *on_top_of* など) が多く見られる中，*hold* や *ride*，*sit_on* など動詞（句）で表現される関係も抽出されている．

2.4 分類器の構築

2.3 節で抽出した物体間の関係の事例を用いて，物体間関係認識のための分類器を構築する．2.3 節で抽出された関係の集合を R で表す．本研究では，2 つの物体 o_1, o_2 が与えられた時，ある関係 $r \in R$ が成立する確率 $P(+1|r, o_1, o_2)$ をモデル化する．ここで，2 物体間には複数の関係が成立しうる点に注意されたい．すなわち，ラベル間に依存関係があるマルチラベル分類問題である．物体 o_1, o_2 に関係 $r_i \in R$ が成立する確率をベクトル $\boldsymbol{y}$ の要素 y_i で表すと，関数 $\boldsymbol{y} = F(o_1, o_2)$ を求めたい．

本研究では，1 層の隠れ層を持つニューラルネットワークで関数 F をモデル化する [9]．これにより，ニューラルネットワークの隠れ層がラベル間の依存関係を捉える中間表現となると期待される．物体 o_1, o_2 から計算される素性ベクトルを $\boldsymbol{v} \in \mathbb{R}^d$ とすると，関係の予測結果 $\boldsymbol{y} \in \mathbb{R}^{|R|}$ を次式で求める．

$$\boldsymbol{y} = \sigma(W_2 \boldsymbol{h} + b_2) \quad (1)$$

$$\boldsymbol{h} = \sigma(W_1 \boldsymbol{v} + b_1) \quad (2)$$

ここで W_1, b_1, W_2, b_2 はニューラルネットワークのモデルパラメータ，$\sigma(.)$ は（ベクトルの要素ごとの）シグモイド関数を表す．実際に関係を予測する際は，$\forall i : y_i \geq 0.5$ となる全ての関係 i を出力する．

物体 o_1, o_2 に対応する素性ベクトル $\boldsymbol{v} \in \mathbb{R}^d$ は，それぞれの物体に付与されているカテゴリ名と画像中での位置を表す bounding box から計算する．本研究では，物体 o_1, o_2 の面積および物体 o_1 の物体 o_2 に対する面積比，画像全体に対する 2 物体の合計面積の比，2 物体の重なり度合いの 4 つの素性を定義した．ここで，o_2 の bounding box は o_1 に対する相対座標に変換した．このような画像における物体間の物理的な相対情報は関係認識精度に大きく寄与すると期待できる．以上の素性にカテゴリ名および bounding box 自身を加えた全 6 種類の素性から $d = 193$ 次元の素性ベクトル $\boldsymbol{v}$ を作成し [10]，訓練事例 $(\boldsymbol{v}, \boldsymbol{t})$ とした．ここで，$\boldsymbol{t}$ は物体 o_1, o_2 間に成り立つ n 個の関係 $\{r_1, r_2, \cdots\ r_n\}$ の成立を表す n-hot ベクトルである．

このようにして作成した訓練データには出現頻度の低い物体間関係の事例が大量に含まれる．ニューラルネットワークの学習時には，それらはノイズとなるため，本研究では出現頻度が高いものから順に上位 80%の関係表現を求め，これに関連する訓練データを実験に用いた．その結果，65,063 事例，133 種類の関係からなる訓練データが得られた ($|R| = 133$)．

[5] `http://stanfordnlp.github.io/CoreNLP/` (Version 3.5.2)
[6] 動詞と結びついて句動詞を形成する不変化詞との間にできる関係
[7] 前置詞とその前置詞句内の名詞句の主辞との間にできる関係
[8] Stanford CoreNLP ではこのような複合前置詞を Multi Word Expression(MWE) として定め解析できる仕様となっている (MWE, NMOD:ON_TOP_OF などの係り受け関係が付与される) [3]
[9] 実装には Preferred Networks, Inc. が公開しているニューラルネットワーク用フレームワーク Chainer(`http://chainer.org/`) を用いた．
[10] o_1 および o_2 のカテゴリ名は 1-hot ベクトル (90 次元 × 2) で表現した．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 関係ラベル予測の精度

	適合率 [%]	再現率 [%]	F1[%]
隠れ層なしニューラルネットワーク			
カテゴリ名のみ	23.7	21.6	21.2
+相対位置素性	25.7	24.3	23.5
+面積領域素性	27.2	24.0	23.9
隠れ層ありニューラルネットワーク			
カテゴリ名のみ	27.6	24.5	23.7
+相対位置素性	26.8	24.0	23.6
+面積領域素性	**28.4**	**25.8**	**25.1**

図 4: 物体間の関係認識の動作例

3 実験

本節では，2.4 節で作成した分類器およびその学習時に使用した素性の有用性の評価実験を行う．

3.1 実験設定

分類器であるニューラルネットワークの隠れ層は 150 次元とし，モデルパラメータは $\mathcal{N}(0, \sqrt{1/d})$ で初期化した．ただし，d はその層への入力ベクトルの次元である．損失関数としてクロスエントロピー誤差，最適化手法として AdaGrad [5] を用いた．また，学習率は $0.1/1.1^{l-1}$(ただし，l はエポック数) とした．開発セットでの最高精度が 10 エポック以上にわたり変化しなくなった時点を学習終了とみなした (early stopping)．

評価データは MSCOCO データセットから無作為に選んだ 50 画像について，説明文の関係表現の中から正解を選ぶことで作成した．結果，51 事例および 31 種類の関係からなる評価データが得られた．

3.2 物体間関係の予測

結果を表 2 に示す．比較手法として，カテゴリ名のみおよびカテゴリ名と 2 物体 o_1, o_2 の相対位置素性のみを用いて学習したニューラルネットワークの結果も示す．全般的に精度が低いのは，説明文に現れない関係表現を予測しても正解とならないためである．しかしながら，カテゴリ名および相対位置素性に加えて面積領域素性を追加することで精度向上が見られた．また，どの素性を用いて学習したかにかかわらず隠れ層ありニューラルネットワークの方が優れた結果となった．このことから本研究で用いた面積領域素性およびニューラルネットワークの隠れ層が物体間関係の識別に有用であると言える．

図 4 に分類器の予測結果の例を示した．PERSON と CELL_PHONE 間に成立する関係の正解は *on* と *talk_on* である．この画像に対して，カテゴリ名のみ，およびカテゴリ名と相対位置素性を用いて学習した分類器は *on* の関係表現しか予測できなかった．一方，面積領域素性も用いて学習した分類器は *on* および *talk_on*，*hold* と予測した．関係 *hold* は正解の関係ラベルには無かったが，正しい関係表現と言える．

4 結論

本稿では，画像説明文生成の重要なサブタスクである物体間関係認識に取り組んだ．具体的には物体間の関係を表す表現の獲得，および分類器の構築を行った．評価実験より，関係認識のためには物体のカテゴリ情報や相対位置素性に加え，2 物体の面積比や重なり度合いなどの物理的な相対情報が有用であることが分かった．今後は，本研究で作成した物体間関係認識器を用いて，説明文生成などのタスクに取り組みたいと考えている．

謝辞 本研究は，文部科学省科研費 (15H01702, 15H05318) から部分的な支援を受けて行われた．

参考文献

[1] S. Aditya, Y. Yang, C. Baral, C. Fermüller, and Y. Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *CoRR*, Vol. abs/1511.03292, , 2015.

[2] P. F. Brown, V. J. Della Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *CL*, Vol. 19, No. 2, pp. 263–311, 1993.

[3] M-C. de Marneffe and C. D. Manning. *Stanford typed dependencies manual*. Stanford University, 2008. Revised for the Stanford Parser v. 3.5.2 in April 2015. `http://nlp.stanford.edu/software/dependencies_manual.pdf`.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[5] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, Vol. 12, pp. 2121–2159, 2011.

[6] D. Elliott and A. de Vries. Describing images using inferred visual dependency representations. In *ACL-IJCNLP*, pp. 42–52, 2015.

[7] J. Johnson, A. Karpathy, and F-F. Li. DenseCap: Fully convolutional localization networks for dense captioning. *CoRR*, Vol. abs/1511.07571, , 2015.

[8] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[9] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, Vol. abs/1411.2539, , 2014.

[10] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014.

[11] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Generating multi-sentence lingual descriptions of indoor scenes. In *BMVC*, 2015.

[12] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[13] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, pp. 55–60, 2014.

[14] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *CL*, Vol. 29, No. 1, pp. 19–51, 2003.

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

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.