B4IM2044

修士論文

言語と画像を統合した物体間関係の理解に関する研究

村岡 雅康

2016 年 3 月 25 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (情報科学) 授与の要件として提出した修士論文である.

村岡 雅康

審査委員：

乾 健太郎 教授　　(主指導教員)
北村 喜文 教授
大町 真一郎 教授
岡崎 直観 准教授　(副指導教員)

言語と画像を統合した物体間関係の理解に関する研究*

村岡 雅康

内容梗概

画像中の物体の種類と位置を特定する物体認識が人間に迫る精度で行えるようになりつつある今，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．ところが，そもそも物体間の関係認識を行う研究は少なく，また，それら既存研究には限られた関係しか扱うことができない，画像に付与された説明文のみから関係事例を抽出しているため物体間の関係として不適切な関係も抽出してしまうなどの課題がある．これらの背景のもと本研究では，物体間の関係事例の収集と画像中の物体の関係認識に取り組む．本研究では，物体間の関係を表現しうる様々な関係を大量に獲得するために，画像に説明文と物体の種類・位置情報が人手で付与されたデータセットを用いる．また，適切な物体間の関係を抽出するために画像中の物体と説明文中の参照表現との対応関係を求める．さらに，抽出した関係事例を用いて物体間関係認識器を構築する．この時，関係認識器の素性として物体間の物理的な相対情報を利用する．評価実験では，物体間の物理的相対情報が関係認識に有用であることを示す．

キーワード

自然言語処理，関係抽出，関係認識，言語と画像の統合

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B4IM2044, 2016 年 3 月 25 日.

目 次

図 目 次

表 目 次

1　はじめに

1.1　背景

画像処理の究極の目標は画像の理解，すなわち，画像に描かれた世界を記述することである [6]．具体的には画像中の物体およびその位置，向き，数量，色，形状，物体間の関係，外観などを認識し記述する．例えば人間であれば，頭や胴体，腕，足などの位置，姿勢，表情，視線，動作などを認識することが目標である．画像処理および画像理解には文字認識や医療用画像の認識，交通シーン理解，3D モデルの復元，監視，バイオメトリクス認証など様々な応用が考えられる．特に近年，入力画像の説明文を自動生成する画像説明文生成の研究が盛んに行われている [7, 8, 9, 10, 11]

画像説明文生成が盛んに行われるようになったきっかけは深層学習 [12, 13, 14] と呼ばれる多層のニューラルネットワークを用いて入力画像から直接説明文を生成する手法である．深層学習は機械学習の一種であり，大量の訓練データさえあればよく (ここでは画像と説明文のペア) [2]，画像から説明文を生成するのに必要な特徴量の設計および獲得は学習の過程でシステムが自動的に行う．その中には，人間の直感からは到底思いつかない，もしくは人間には理解できない，定義が困難な特徴量なども含まれる．深層学習により，表現力豊かな説明文を生成することが可能になった (図 1)．

図 1 の説明文を見ると，システムは正しく画像理解しているような印象を受ける．しかしながら，システムの内部は多層ニューラルネットワークで複雑化しており，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 (図 2 に図 1 と同じ手法による誤生成の例を示す)．

画像説明文生成の研究が盛んに行われている一方で，画像に描かれた物体を当てるカテゴリ認識と呼ばれるタスクもある．説明文生成が画像全体の理解であるのに対して，カテゴリ認識は画像の一部分に焦点を当てた局所的な理解と捉えることができる．このカテゴリ認識の精度 (誤識別率) は 2015 年 12 月の時点で，

[2] インターネット上に Flickr(`https://www.flickr.com/`) などの画像投稿サービスが登場したことで，大量の画像と説明文のペアが容易に入手可能なったことが深層学習の研究を加速させた要因の一つである．

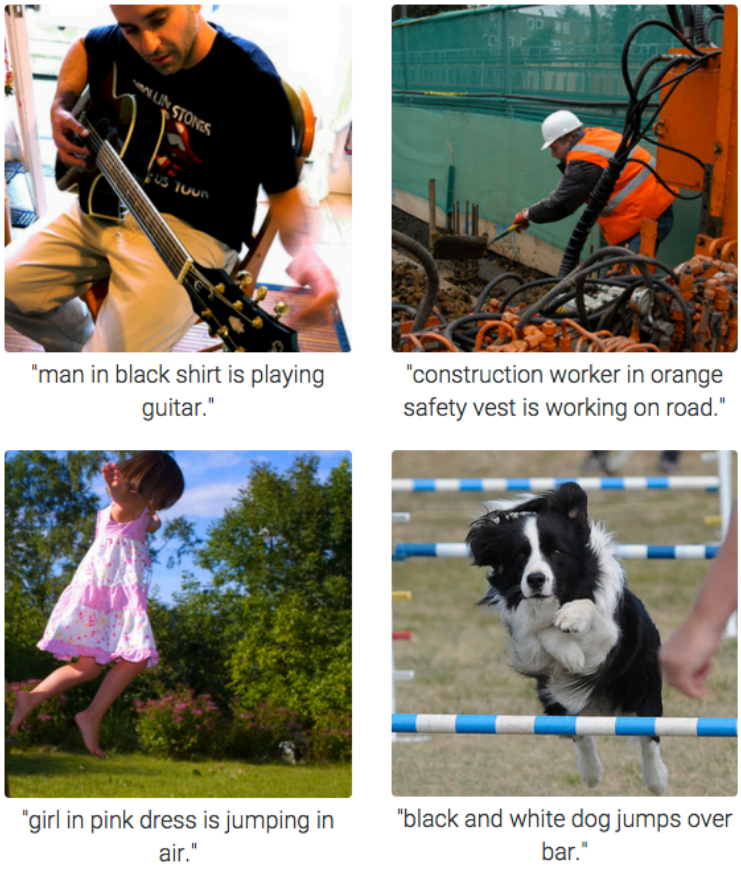


図 1: 深層学習を用いたシステムによる画像説明文の生成例 [4]

Microsoft Research Asia が提案した 152 層からなる深層ニューラルネットワーク [15] の 3.6%で，ILSVRC と呼ばれるコンペティションにおいて最高精度を達成した [3]．Russakovsky ら [16] は同タスクの人間の誤識別率は 5.1%と報告しており，すでに計算機が人間を上回る精度に至っている．

また，カテゴリに加えて物体の画像中における位置を特定する問題を物体検出と呼ぶ．ILSVRC ではカテゴリ認識の他に物体検出のコンペティションも行っており，2015 年の最高精度はカテゴリ認識と同じく Microsoft Research Asia の提案した手法で誤識別率 9.0%である．

[3] http://image-net.org/challenges/LSVRC/2015/results

[4] Karpath および Fei-Fei による研究 [9] のプロジェクトページ (http://cs.stanford.edu/people/karpathy/deepimagesent/) より引用

図 2: 深層学習を用いたシステムによる画像説明文の誤生成例 [4]

1.2 本研究の目的

このように物体の認識が人間に迫る精度で行えるようになりつつある今，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 3 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 (主語，動詞，目的語) の 3 つ組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 および Vries [3]，Kong ら [4]，Lin ら [5] 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*，*in_front_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*，*eat* など，状態や動作を表す関係もある．また，Aditya ら [17] 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する手法を提案しているが，この手法は物体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．

こうした背景のもと本研究では，物体間の関係事例の収集と，画像中の物体の関係認識に取り組む (図 4)．本研究において物体間の関係事例とは，ある具体的な

図 3: 男性がテーブルの上でスケートボードに乗っている画像

2 つの物体 o_1, o_2 間に関係 r があるとき，その 3 つ組 $r(o_1, o_2)$ のことを指す．物体間の関係事例の収集を行うのは，そもそも物体間関係認識器を学習するための訓練データがなく，それを作成する必要があるためである．本研究では MSCOCO [18][5] 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 (3 章で詳述する)．このデータには，画像中の物体を表す矩形とラベル，およ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ン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付ける．

説明文の構文解析とアラインメントの結果に基づき，物体間の関係事例を自動的かつ大量に獲得する．これにより，既存研究 [3, 4, 5] で導入されていた物体間の関係の定義は不要となる．また，アラインメントを取ることで適切な物体間の関係事例の獲得が可能となる．さらに，獲得した関係事例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推定する分類器を構築する．既存研究 [3, 4, 5] 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

[5] http://mscoco.org/

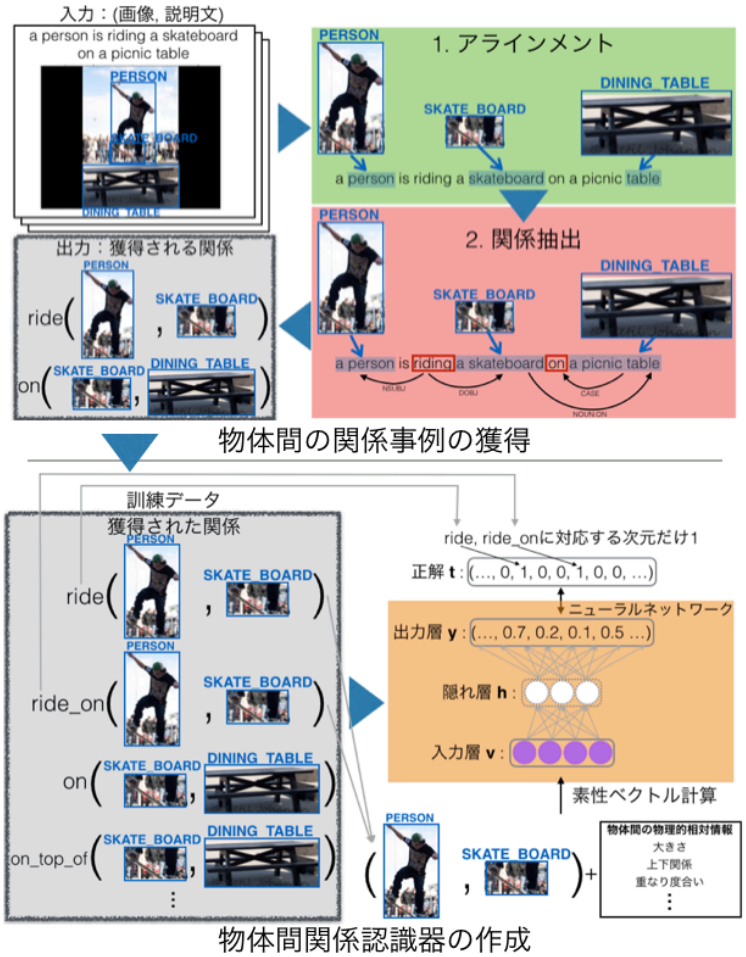


図 4: 提案手法

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．

以下に本研究の貢献をまとめる．

- 物体間の関係を表現しうる様々な関係表現の獲得
- 言語情報と画像情報の統合
- 物体間関係認識器の構築

1.3　本論文の構成

本論文の構成は以下の通り．2 章で本研究に関連のある，物体間の関係認識および深層学習を用いた研究を概観する．3 章では，物体間関係の事例獲得のために用いるデータセットについて説明する．具体的な物体間の関係事例の獲得方法を 4 章で述べ，続く 5 章にて，獲得した関係事例を訓練データとし，物体間関係認識器の学習を行う．6 章で物体間関係認識器の評価実験を行い，結果を考察する．最後に 7 章で本研究の総括を行う．

2　関連研究

本章では本研究の目的である物体間の関係認識に関連する既存研究について述べ，本研究との相違点について明らかにする．また，後半では深層学習の問題点を指摘し，本研究で採用する方針の利点を述べる．

2.1　物体間の関係認識

Elliott および Vries [3] は Visual Dependency Representation(VDR) と呼ばれる，画像中の物体間の依存関係を表すグラフ構造を求めるために表 1 に示した 5 つの関係を定義し，画像中の物体間の関係をそれらのいずれかに分類している．Elliott および Vries は説明文生成の評価実験において，VDR を用いた手法は深層学習を用いた手法と同程度の性能であると報告している．このことから，物体間の関係を認識することは画像理解および説明文生成において有用であると言える．

表 1: Elliott および Vries[3] による物体間関係の定義

関係	定義
Beside	2 物体のなす角が 315° から 45° または 135° から 225°
Above	2 物体のなす角が 225° から 315°
Below	2 物体のなす角が 45° から 135°
On	2 物体の重なり度合いが 50% 以上
Surrounds	2 物体の重なり度合いが 90% 以上

また，Kong ら [4] は Markov Random Field(MRF) を用いて，物体を頂点，物体間の関係を辺としたグラフ構造を構築し，画像中の物体とそれを指し示す説明文中の表現とを結びつけるタスクに取り組んでいる．その際，物体間の関係として表 2 の 2 種類の関係を定義し，それらを MRF のエネルギー関数の素性として使用している．

表 2: Kong ら [4] による物体間関係の定義

関係	定義
close–to	2 物体間の距離が 0.5m 以下
on–top–of	一方の物体が他方の物体より上方にあり，かつ， 上にある物体の底面が下にある物体の上面と少なくとも 80% 重なる

Lin ら [5] は室内を描いた画像を説明する文章 (単文ではなく意味が一貫した複数の文) を生成するタスクにおいて画像から Scene Graph を構築している．Scene Graph とは Kong ら [4] の手法を拡張したもので，物体の属性 (例えば，色や大きさなど) もグラフの頂点で表すことで画像についてより詳しい説明文生成が可能となっている．Scene Graph 中では物体間の関係として表 3 の 8 種類の関係を用いている．

表 3: Lin ら [5] が用いた物体間関係

next–to, near, top–of, above, in–front–of, behind, to–left–of, to–right–of
ただし，上記の詳しい定義に関しては論文中で言及されていない

一方，Aditya ら [17] は画像に説明文が付与されたデータセットを用いて様々な関係を説明文から抽出する手法を提案している．物体間の関係を収集は説明文のみを用いて，以下の手順で行う．まず Stanford Parser と呼ばれるツールを用いて説明文の係り受け解析を行う．続いて K-parser と呼ばれるツールによりオントロジーおよび意味的情報を付与する．オントロジーとは boy の上位語は person，animal の下位語は dog，cat などのように意味的階層構造のことであり，意味的情報は動作主や受け手，起点，終点などの説明文中における単語の意味的役割を

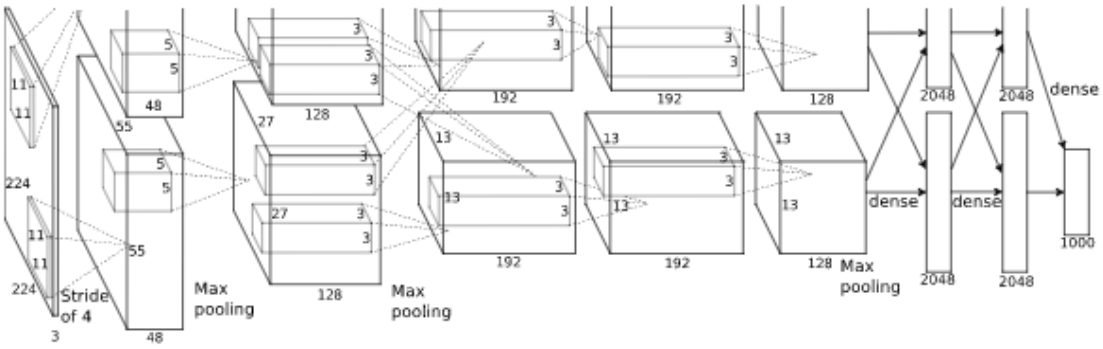


図 5: Krizhevsk ら [1] が用いた Convolutional Neural Network([1] より引用)

指す[6]．この結果，entity(物体) および event(関係) が頂点，辺は event から entity への有向辺で表されるグラフが作られる．

しかし，われわれ人間が物体間の関係を記述する際，表 1，2 および 3 に挙げた位置関係を表すものだけでなく，一般的に *look_at* や *throw*，*eat* など，状態や動作を表す関係も用いる．上で述べた既存研究 [3, 4, 5] ではこのような関係を扱うことができない．また，そのような関係を扱うためにあらゆる関係について，個別に定義することは現実的に不可能である．また，Aditya ら [17] の手法は説明文のみからグラフを構築してい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．

これらの問題に対して，本研究では画像に説明文が付与されたデータセットを用いて画像中の物体を指し示す説明文中の表現を特定し，その表現間に成り立つ関係のみを抽出することで人手による関係の定義は不要となり，また，人間が物体間の関係を記述するのに用いる表現を自動的かつ大量に獲得することが可能となる．

2.2 深層学習

2012 年に行われた ILSVRC のカテゴリ認識タスクにおいて，Krizhevsky らの SuperVision[1] が 2 位の手法に大差をつけ勝利したことで深層学習が注目を浴びるきっかけとなった．2 位以下の手法およびこれまでのカテゴリ認識の主流は，画像から取り出すことのできるカテゴリ認識に効きそうな素性 (例えば，色やエッジなど) を人手で設計し，それをカテゴリ認識器の入力として認識器を学習するという方法だった．それに対し，Krizhevsky らの手法はそのような素性設計を一切行わず，入力として画像をそのまま使用するというものだった．具体的には Convolutional Neural Network(CNN) と呼ばれるニューラルネットワークを多層にしたもの (図 5) で，出力層 (多次元ベクトル) の各次元がカテゴリのいずれかに対応し，その値が確率を表している．CNN の中間層は高次の層に向かうにつれ，単純な視覚的特徴からより複雑で意味をもつ特徴になることがわかっている [19]．また，学習済み CNN を出力層のみ付け替え別タスクに応用する転移学習においても CNN は高精度を出すことが知られている [20, 21]．

CNN 素性の画像説明文生成タスクへの応用では，基本的には画像から CNN 素性 (CNN の出力層) を取り出し，取り出した素性を言語モデルとして学習した Recurrent Neural Network(RNN) の初期値に用いるという方針をとる．RNN は自然言語や音声のように入力が可変長であるデータ (時系列データ) を扱う場合に用いるニューラルネットワークである．過去の情報を保持する仕組みとして，図 6 のように時刻 $t-1$ で計算された中間層の素性 CONTEXT$(t-1)$ を時刻 t の出力層を計算する際に用いるという特徴がある．この CNN と RNN を組み合わせた手法が近年多数提案され高精度を収めている [7, 8, 9, 10, 11]．深層学習による手法が主流になる以前の画像説明文生成は，類似画像に付与されている説明文を用いたり [22, 23]，テンプレートを用いる方法によるものであったため [24]，多様な説明文を生成できなかったのに対し，これらの手法は豊富な語彙を用いて様々な説明文を生成することが可能である．しかしながら，手法の内部は多層ニューラルネットワークで複雑化しており，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という問題がある．

[6] 詳しいラベル体系については `kparser.org` を参照されたい．

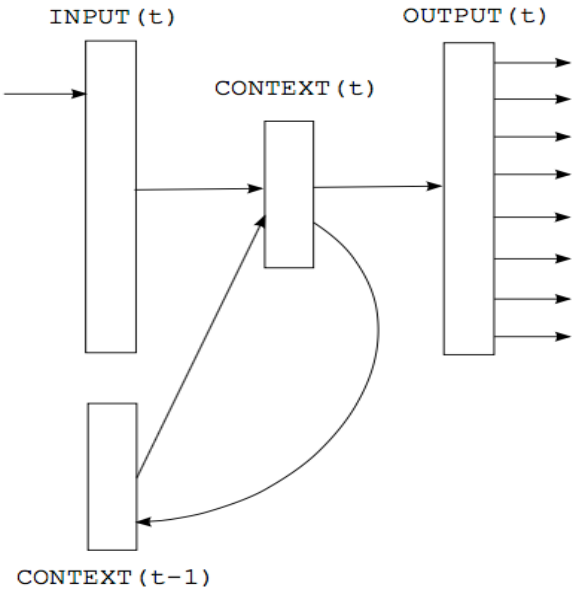


図 6: Reccurrent Neural Network([2] より引用)

このように内部がブラックボックスで挙動の解析が困難なシステムは，たとえ精度がよくても，画像を理解しているとは言えないだろう．そこで本研究では，まず画像中の物体を認識し，その後で物体間の関係を認識するという手順を踏む．また，関係認識器の構築の際，物体間の関係認識に有効と思われる素性を人手で設計することでその素性の関係認識に対する効果を検証することができる．このことにより，エラーの所在が特定されるだけでなく，関係認識に関する知見が得られる．本研究では物体認識と物体間関係認識に問題を切り分け，物体認識については正解が分かっているデータセットを用いることで物体間関係認識のみの評価を行う．

3 Microsoft Common Objects in Context

本研究で使用するデータセットについて述べる．

Microsoft Common Objects in Context[18](以降，MSCOCO と略記する) とは Microsoft 社が Creative Commons Attribution 4.0 License [7] および Flickr Terms of Use [8] 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 328,124 画像からなる．各画像には図 7 のように少なくとも 5 文，その画像を説明する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 (表 4) 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 7 に示した青線の矩形 (=**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．本研究ではこの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．

また，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 7 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用いられる様々な関係を大量に獲得することで物体間関係認識器の訓練データを作ることができる．

物体認識と関係認識の問題を切り分ける理由は以下の通りである．そもそも画像に説明文のみが付与されたデータセットは MSCOCO の他に多数存在するが [9]，それらを用いる場合は物体認識を人手で行うか物体認識を行う検出器が必要となる．物体間の様々な関係の獲得を行う本研究において，前者は大規模なデータセットに対して行う必要があるため非常にコストがかかる．一方，後者は物体認識のエラーが物体間の関係の獲得にも影響してしまうという問題がある．そこで本研究では MSCOCO データセットを用いることで物体認識と関係獲得に問題を切り分け，物体認識が正しく行われたと仮定する．その結果，関係認識に関する純粋な評価が可能となる．しかし，近年の深層学習を用いた物体認識器の精度向上は

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

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

[9] よく用いられるデータセットに Flickr8K[25]，Flickr30K[26]，YFCC[27] などがある．

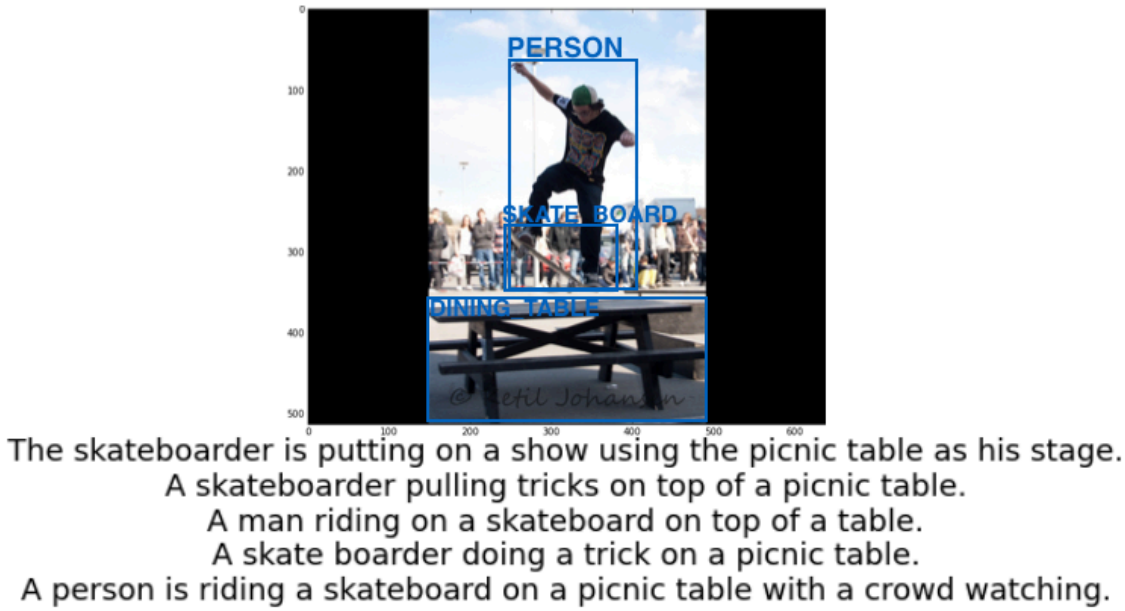


図 7: MSCOCO データセット

めざましく，将来的には実用的なレベルの物体認識器を用いて，MSCOCO 以外のデータセットも併用したより大規模な物体間関係の獲得も可能になると考えられる．

表 4: MSCOCO で使用されるカテゴリ一覧

PERSON	BICYCLE	CAR	MOTORCYCLE	BUS
TRAIN	TRUCK	BOAT	FIRE_HYDRANT	STOP_SIGN
PARKING_METER	BENCH	CAT	DOG	HORSE
SHEEP	ELEPHANT	BEAR	ZEBRA	GIRAFFE
UMBRELLA	HANDBAG	TIE	SUITCASE	SKIS
SNOWBOARD	SPORTS_BALL	KITE	BASEBALL_GLOVE	ANIMAL
SKATEBOARD	SURFBOARD	TENNIS_RACKET	WINE_GLASS	CUP
FORK	KNIFE	BOWL	BANANA	APPLE
SANDWICH	BROCCOLI	CARROT	HOT_DOG	PIZZA
CAKE	CHAIR	COUCH	POTTED_PLANT	DINING_TABLE
TOILET	TV	LAPTOP	REMOTE	KEYBOARD
CELL_PHONE	MICROWAVE	TOASTER	SINK	REFRIGERATOR
BOOK	VASE	SCISSORS	TEDDY_BEAR	HAIR_DRIVER
OUTDOOR	FOOD	INDOOR	APPLIANCE	VEHICLE
FURNITURE	ACCESSORY	ELECTRONIC	AIRPLANE	TRAFFIC_LIGHT
BIRD	KITCHEN	COW	BACKPACK	FRISBEE
BASEBALL_BAT	BOTTLE	SPOON	ORANGE	DONUT
BED	MOUSE	OVEN	CLOCK	TOOTHBRUCH

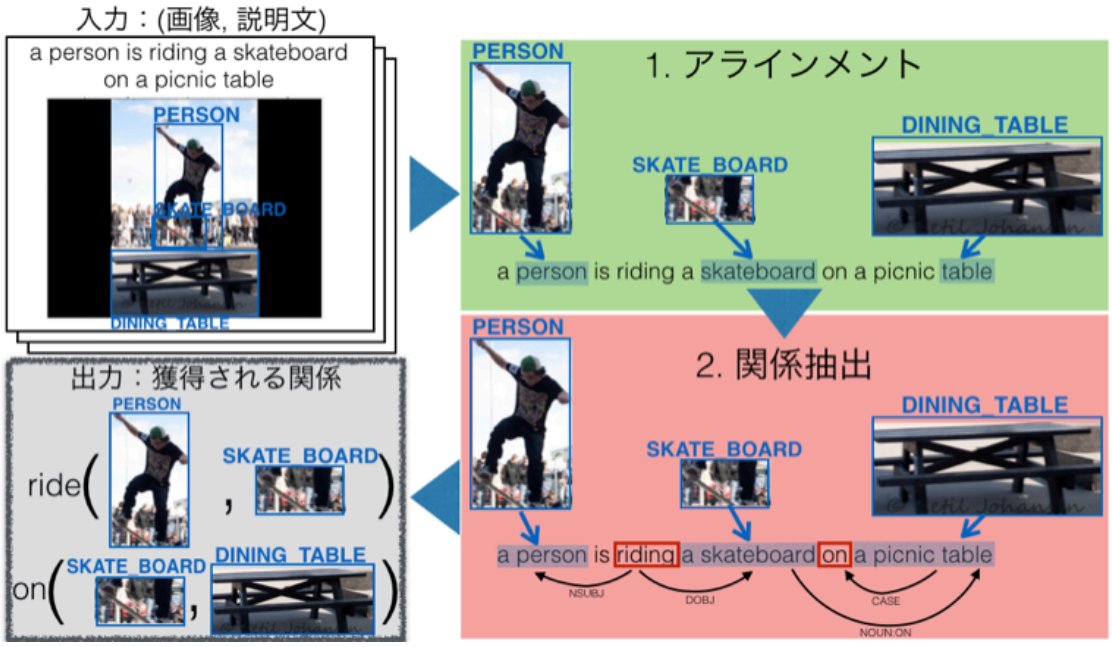


図 8: 物体間の関係獲得

4　物体間の関係獲得

物体間関係認識器の学習のためには大量の訓練事例が必要である．この訓練事例は入力が 2 つの物体，出力がその間の関係という構造を持った事例でなければならない．しかし，このようなデータセットは存在しないため，大量に存在し，容易に入手可能な画像に説明文が付与されたデータセットから物体間関係認識器の訓練事例として適当な事例を抽出し，この訓練データを作成する必要がある．

本章では物体間関係認識器の訓練データ作成のため，説明文が付与された画像のコーパスから物体 o_1, o_2 間に成立する関係 r の事例 $r(o_1, o_2)$ を獲得する手法について述べる．本研究ではこれを次の 3 つの処理に分解して行う．図 8 に概要を示す．

1. アラインメント：画像中の物体と説明文中の参照表現の対応付け

2. 説明文の言語解析：Stanford CoreNLP を用いた説明文の基本的な言語解析

3. 物体間関係の事例の抽出：係り受けパスを用いた物体間関係の事例の抽出

以下の各節でこれらの詳細を述べる．

4.1 アラインメント

MSCOCO データセットには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．物体のカテゴリは 90 種類に限定されている一方，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 7 では，画像中の 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．この対応関係を求める処理をアラインメントと呼ぶ．本研究ではこのアラインメントを行うのに次の 3 種類の手法を試行し，評価データに対して最も良い結果の手法を用いて関係事例の獲得を行う．

4.1.1 IBM Model を用いたアラインメント

カテゴリ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 [28] を用いることでアラインメントを取る．IBM Model は翻訳確率 $P($ 説明文中の単語 $w|$ 物体カテゴリ $c)$ を画像中に登場する物体のカテゴリの集合とその画像の説明文が対となっ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 eating a donut
- PERSON BENCH — a young man sitting on a bench

IBM Model は $P(\text{man}|\text{PERSON})$ 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）．

訓練データ作成時，各説明文に対し以下の処理を行った．

- 全ての文字を小文字に変換
- 文末以外の改行文字を削除
- 文末のピリオドを削除

また，本研究ではIBM Modelの実装としてGIZA++ [29] [10] を用いた．学習時のパラメータはデフォルト値を用いた．

4.1.2 単語ベクトルの類似度を用いたアラインメント

Mikolovら [30] は「任意の単語は，同じ文脈で出現する単語の分布からその単語の意味が推定可能である」とする分布仮説 [31] に基づき，次の双対数線形関数で表される対数尤度 $\mathcal{L}$ を最大化することで単語 w の意味を多次元実数ベクトルとして獲得した [11]．

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\sum_{-d\leq j\leq d, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

T はコーパスの総単語数，d は考慮する周辺単語の数 (文脈サイズ) である．この単語ベクトルは意味が似ている単語同士はその単語ベクトルもベクトル空間上で互いに近くに分布するという性質を持つ．

この性質を用いて，以下の条件を満たすすべてのカテゴリを説明文中の単語 w に対応付ける:

$$cos(\boldsymbol{w}, \boldsymbol{c}) = \frac{\boldsymbol{w}\cdot\boldsymbol{c}}{|\boldsymbol{w}||\boldsymbol{c}|} \geq \beta. \tag{2}$$

ここで，$\boldsymbol{w}$，$\boldsymbol{c}$ はそれぞれ説明文中の単語 w の単語ベクトル，物体のカテゴリ c の単語ベクトルであり，cos はコサイン類似度と呼ばれる．

本研究では，約1000億単語からなるGoogle News datasetを訓練データとして学習した300次元からなる単語ベクトルを使用した [12]．

[10] https://github.com/moses-smt/giza-pp

[11] 著者らはこのモデルを **word2vec** というツールとして公開している．https://code.google.com/archive/p/word2vec/

[12] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing より入手可能

4.1.3 WordNet の階層情報を用いたアラインメント

WordNet[32][13] は人手で整備されたシソーラスである．シソーラスとは似た意味を持つ単語を同義語としてまとめ上げ，それら同義語間を上位・下位関係の階層構造として関連付けた辞書である．例えば，mammal は dog の上位語であり，animal の下位語である．また，person の同義語として individual，someone などがある．

Jiang および Conrath[33] は WordNet の階層構造を利用して 2 単語 w_1，w_2 間の意味的類似度を次式で定式化した:

$$sim(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) + 2 * IC(LCS(w_1, w_2))}, \tag{3}$$

$$IC(w) = -\log p(w). \tag{4}$$

LCS(Least Common Subsumer) は WordNet の階層構造において 2 単語の共通の上位語のうち 2 単語から最も近い上位語を指す．$p(w)$ はコーパスにおける単語 w の出現確率であり，IC(Information Content) は情報量と呼ばれる．単語間の意味的類似度の計算法としてこの他にも複数の手法が提案されているが [34, 35, 36, 37, 38]，Fernando および Stevenson[39] は Jiang および Conrath の類似度が最も優れていることを示している．

そこで本研究でも (3) 式を用いて，説明文中のある単語 w に対して物体のカテゴリ c との類似度が,

$$sim(w, c) \geq \gamma \tag{5}$$

となるすべてのカテゴリをその単語に対応付ける．

各単語の情報量として本研究では Pedersen[40] が配布しているデータセットを用いた [14]．

物体のカテゴリを説明文中の単語に対応付けた例を以下に示す．

[13] http://wordnet.princeton.edu/

[14] http://wn-similarity.sourceforge.net/ より入手可能．Version 3.0 を使用．

- a man/PERSON is riding a skateboard/SKATEBOARD
- a man/PERSON eating a donut/DONUT
- a young man/PERSON sitting on a bench/BENCH

上記の例において，大文字はカテゴリを表し，'/' の左側の単語に右側の物体が対応することを意味する．

4.2 説明文の言語解析

本節では物体間の関係事例を獲得するために説明文に対して行う言語的な解析について説明する．

Stanford CoreNLP [41] [15] は自然言語の基本的な処理を行うツールである．具体的には単語の lemmatize(見出し語化)，文区切り，品詞タグ付け，構文解析，係り受け解析，固有名詞抽出，共参照解析などを行う．本研究ではこのうち単語の lemmatize(見出し語化)，品詞タグ付け，構文解析，係り受け解析，共参照解析の結果を利用する．

単語の lemmatize(見出し語化) とは単語の原形を求める処理である．これは例えば，*a man* ***riding*** *a skateboard* と *a man* ***rides*** *a skateboard* という 2 つの説明文がある場合でも，*ride* という一つの関係を獲得するために行う．

また，品詞タグ付けにより各単語には品詞が自動付与される．これは後述する関係事例の抽出時に利用する．Stanford CoreNLP は Penn Treebank タグセット [42] に基づく品詞タグ (付録 A) を出力する．

構文解析は入力文の文法的な関係を解析する処理である．これにより説明文の名詞句や動詞句などの統語構造 (句構造) が求められる．構文解析の結果はしばしば図 9 上のような木構造で表される．

構文解析が入力文の統語的な性質に注目した解析である一方，係り受け解析は単語間の関係に注目した解析である．この処理により，説明文中のメインとなる動詞やその動詞の主語や目的語となる単語など単語間の関係が求められる．図 9

[15] `http://stanfordnlp.github.io/CoreNLP/` 本研究では Version 3.5.2 を使用.

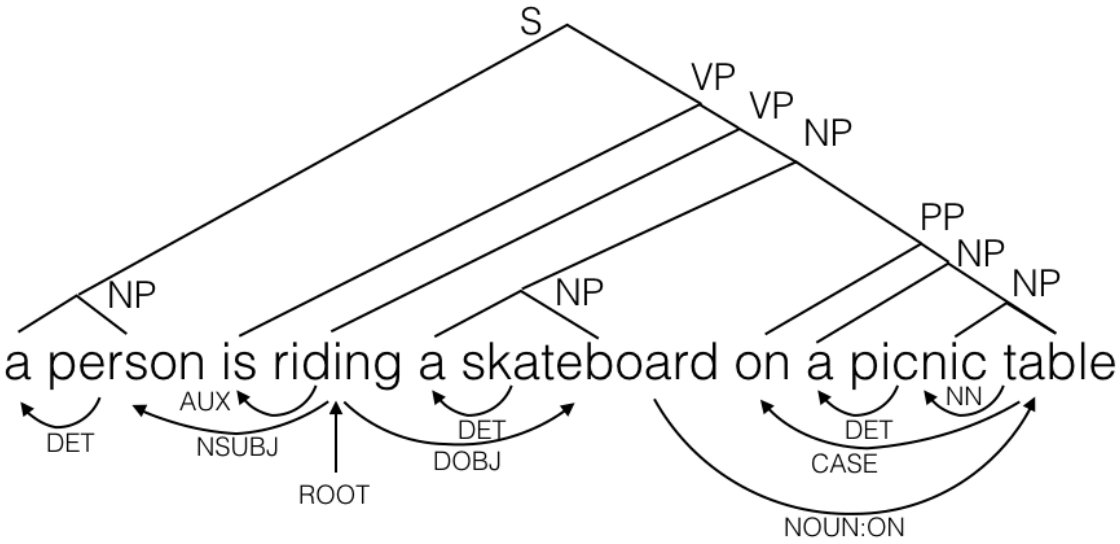


図 9: 構文解析 (上) および係り受け解析 (下) 結果

において，単語の下側にある矢印が係り受け解析結果である．動詞 ride の主語は person(NSUBJ，nominal subject 関係)，目的語は skateboard(DOBJ，direct object 関係) であることがわかる．この係り受け解析結果に基づき物体間の関係事例の抽出を行う (詳細は次節で述べる)．Stanford CoreNLP で使用される係り受けのタグは Universal Dependencies [16] と同一のものである (付録 B)．

共参照解析とは，代名詞の参照先を解析することである．例えば，説明文 *a bowl of soup has some carrots in* ***it*** 中の代名詞 *it* は *bowl* を指すことが共参照解析により明らかになる．

これらの結果を用いて，句構造の高さ (木構造の分岐点が句構造の高さを表す) が h_{tree} 以下の名詞句のうち，高さ最大の名詞句の抽出する (図 10 の青色の部分)．名詞句の抽出後，名詞句の内部の係り受けパスは削除し，1 つの名詞句を構成するいずれかの単語とその名詞句以外の単語との間に存在する係り受けのパスは名詞句全体から張られるように書き換える．上記の処理において名詞句を抽出したのは，後述する物体間関係の事例抽出において抽出候補を減らすためである．

[16] `http://universaldependencies.org/\#language`

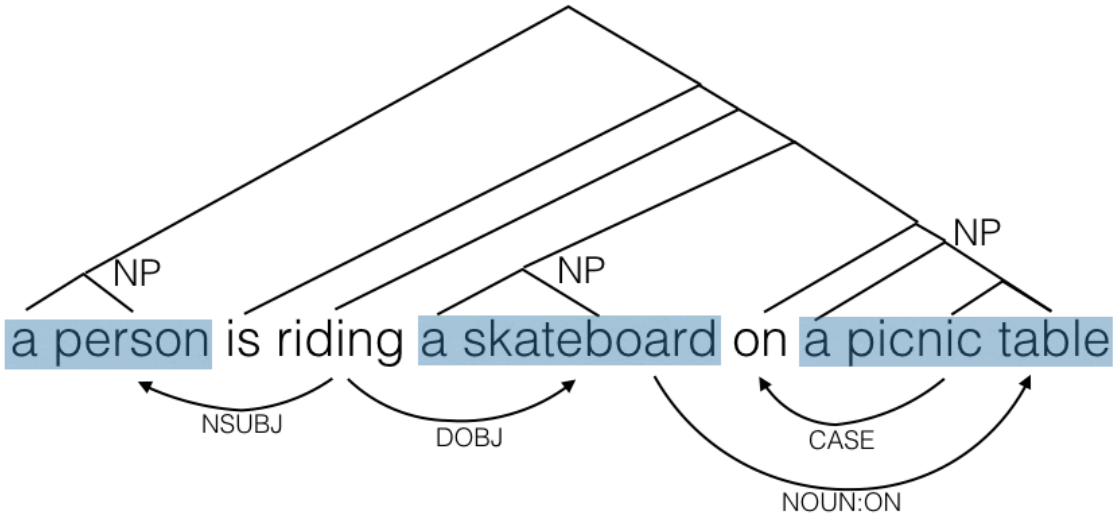


図 10: 名詞句の抽出

4.3 物体間関係の事例の抽出

4.2 節の結果を用いて，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事例の抽出を行う．

- 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 [17] の関係にある単語を動詞句として統合する．
2. CASE[18] の親の親が動詞 (品詞タグで VB から始まるもの) であり，その動詞が直接目的語 (係り受けタグが 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 (係り受けタグが NSUBJ) を関係事例の第 1 引数とする．

[17]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
[18]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

表 5: 獲得した関係表現上位 20 件

物体間関係	事例数 (%)	物体間関係	事例数 (%)
on	19,666 (12.58)	in_front_of	1,670 (1.07)
in	14,300 (9.15)	by	1,532 (0.98)
with	13,047 (8.35)	wear	1,413 (0.90)
hold	5,136 (3.29)	lay_on	1,341 (0.86)
at	4,345 (2.78)	near	1,315 (0.84)
of	4,096 (2.62)	to	1,185 (0.76)
next_to	3,974 (2.54)	sit_next_to	1,184 (0.76)
ride	3,711 (2.37)	ride_on	1,117 (0.71)
sit_on	3,265 (2.09)	sit_in	1,109 (0.71)
on_top_of	2,393 (1.53)	through	1,080 (0.69)

3. *on_top_of* のような複数単語で一つの前置詞として機能する複合前置詞のうち一般的によく用いられる複合前置詞 58 種類 (付録 C 参照) を一つの関係として考慮する [19].

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 $\alpha = 0.6$，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 20 件を表 5 に示す．既存研究 [3, 4, 5] でも扱われていた前置詞 (*on* や *in* など) および位置的/空間的関係 (*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．

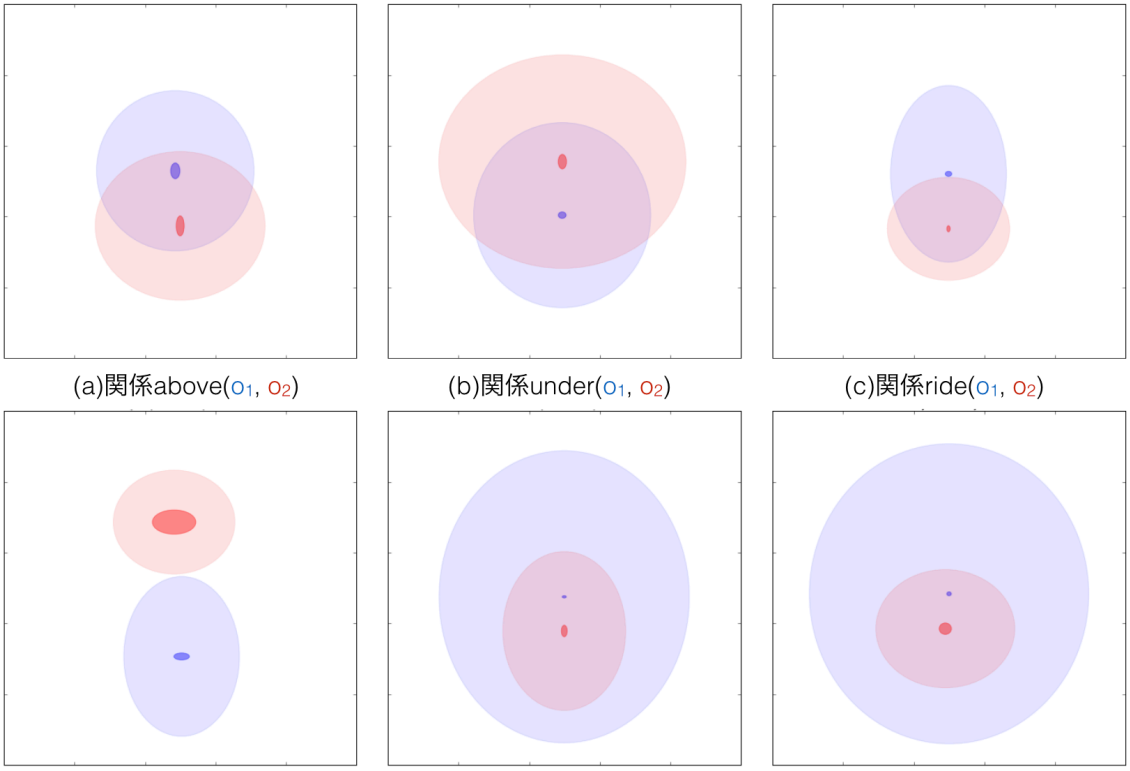


図 11: 物体間関係の可視化

4.4 物体間関係の可視化

獲得した物体間の関係事例 $r(o_1, o_2)$ のうち関係 r は言語表現である一方，o_1, o_2 は画像中の物体であり，各物体には画像中における位置座標 (=bounding box) が付与されている．関係が r となる事例に関して，この bounding box の平均および分散を求めることで関係 r が成立する 2 次元平面上での 2 物体の平均的な位置関係がわかる．これは言語または画像情報のみからは決して求めることのできないものであり，また，言語と画像を結びつける非常に興味深い知見である．

[19]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)．`http://nlp.stanford.edu/software/dependencies_manual.pdf`

図 11 に特に興味深いと思われる関係を示す[20]．図 11 中の濃い楕円の中心は物体の座標の平均，長軸と短軸は座標の分散を表す．薄い楕円の長軸と短軸はそれに物体の bounding box の縦幅および横幅の平均を加えたものである．例えば，関係 above/under は 2 物体が図 11(a) または (b) のような位置関係のときに用いられることがわかる．その他の関係もそれぞれ意味的もしくは直感に合う位置関係となっていると言える．above/under のような位置関係を表す関係は「関係 above は物体 o_1 が物体 o_2 の上方/下方にあり，かつ 2 つの物体が接触していない場合にのみ成立」のように予めルールを決めておくことで未知の物体間でも関係の認識が可能である．しかし，wear や eat など動作を表す関係は簡単なルールを定義することは容易ではなく，また全ての関係についてルールを定めることは不可能である．そのため，本研究のように画像に説明文が付与されたデータセットから物体間の関係を自動的かつ大量に収集することで図 11 の (c)〜(f) のような動作を表す関係の 2 物体の位置関係を統計的に求めることができるのは本研究の貢献の一つであると言える．

[20]bounding box の平均・分散を求める際，bounding box の値は $[0, 1]$ に正規化し，o_2 は o_1 の相対座標に変換した．

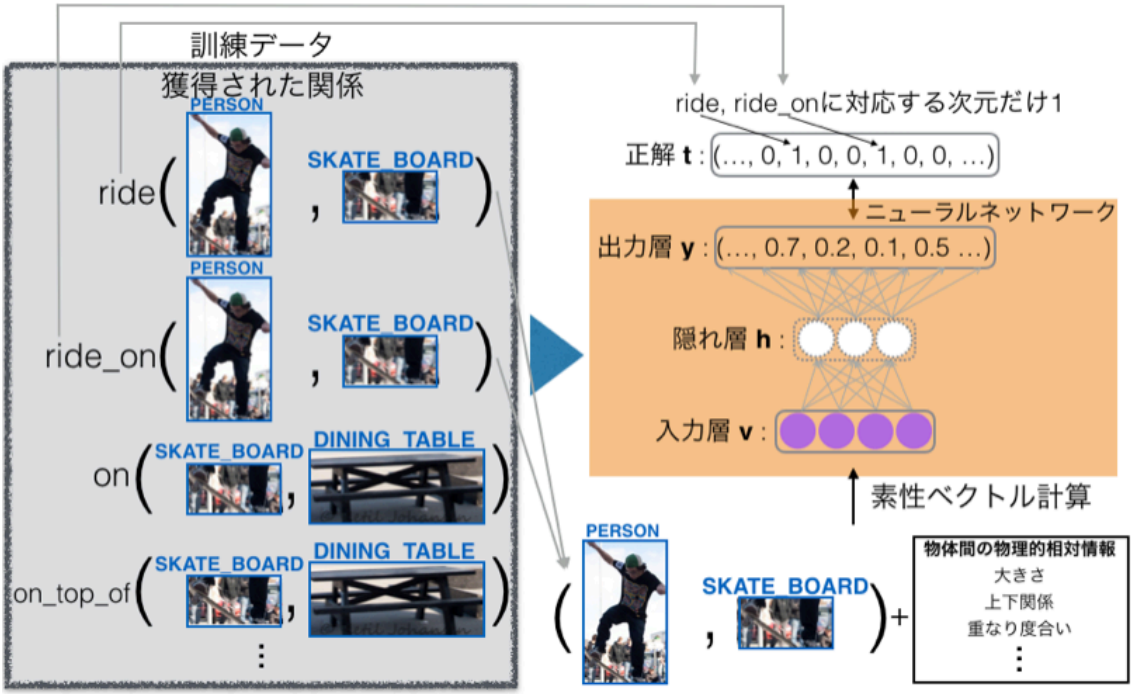


図 12: 物体間関係認識器の作成

5　物体間の関係認識器の作成

4 章で獲得した物体間関係の事例を用いて，物体間関係認識器を構築する．図 12 に概要を示す．本研究では認識器として 1 層の隠れ層を持つニューラルネットワークを使用する．以下では物体間関係認識器としてのニューラルネットワークについて説明した後，訓練データ作成およびニューラルネットワークの学習について述べる．

5.1　物体間関係認識器としてのニューラルネットワーク

4 章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，マルチラベル分類問題である．また，例えばある物体間に関係 $ride\_on$ が成り立つ時，

同時に関係 $ride$ や on も成り立つ可能性が高いことから，ラベル間には依存関係があると考えられる．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[21]．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:

$$\boldsymbol{y} = \sigma(W_2\boldsymbol{h} + b_2), \tag{6}$$

$$\boldsymbol{h} = \sigma(W_1\boldsymbol{v} + b_1). \tag{7}$$

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 $\frac{1}{1+\exp(-x)}$ を表す．ニューラルネットワークへの入力 $\boldsymbol{v}$ の詳細は次節で述べる．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．ニューラルネットワークの構造の詳細については付録 D.1 に示す．

5.2 訓練データ作成

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．具体的な計算式を表 6 に示す．ただし，コーパス中の画像はサイズが不均一であるため，各 bounding box の値はその画像のサイズで正規化処理を行った $(0 \leq x_o, y_o, w_o, h_o \leq 1)$．また，$o_2$ 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，人間が物体間の関係を認識する際にも考慮している情報であり，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

[21]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/`) Version 1.5.1 を用いた．

表 6: 面積領域素性

2 物体 o_1, o_2 の面積	$w_{o_1}h_{o_1}$，$w_{o_2}h_{o_2}$
o_1 の o_2 に対する面積比	$w_{o_1}h_{o_1}/w_{o_2}h_{o_2}$
画像全体に対する 2 物体の合計面積の比	$S_{o_1o_2} - S_{overlap}$
2 物体の重なり度合い	$S_{overlap}/(S_{o_1o_2} - S_{overlap})$

ただし，x_o, y_o, w_o, h_o は物体 o の bounding box を表す．
また，$S_{o_1o_2} = w_{o_1}h_{o_1} + w_{o_2}h_{o_2}$,
$S_{overlap} = (\min(x_{o_1} + w_{o_1}, x_{o_2} + w_{o_2} + x_{o_1}) - \max(x_{o_1}, x_{o_2} + x_{o_1}))$
$*(\min(y_{o_1} + h_{o_1}, y_{o_2} + h_{o_2} + y_{o_1}) - \max(y_{o_1}, y_{o_2} + y_{o_1}))$.

性ベクトル $\boldsymbol{v}$ を作成し [22]，ニューラルネットワークの訓練事例 $(\boldsymbol{v}, \boldsymbol{t})$ 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 100 回以上の関係表現を求め，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 ($|R| = 133$)．

5.3 物体間関係認識器の学習

物体間関係認識器であるニューラルネットワークの学習について述べる．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 $\mathcal{N}(\mathbf{0}_d, \sqrt{1/d}\boldsymbol{I}_{d\times d})$ で初期化した．ただし，$\mathcal{N}(m, s)$ は平均 m，分散 s の正規分布であり，$\mathbf{0}_d$，$\boldsymbol{I}_{d\times d}$ はそれぞれ d 次元の 0 ベクトル，$d \times d$ の単位行列である．また，d はその層への入力ベクトルの次元数を表す．損失関数としてクロスエントロピー誤差を用いた．ニューラルネットワークの出力を $\boldsymbol{y}$，正解を $\boldsymbol{t}$ とするとクロスエントロピー誤差

[22] 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．

は次式で表される:

$$-\sum_{i=1}^{R} t_i \log y_i. \tag{8}$$

最適化手法は AdaGrad [43] を用いた．また，学習率は $0.1/1.1^{l-1}$(ただし，l は訓練データの反復回数) とした．開発セットでの最高精度が反復回数 10 回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．このような学習終了の方法を早期打ち切り (early stopping) と呼び，ニューラルネットワークが過学習するのを防ぐ．ニューラルネットワークの学習の詳細については付録 D.2 を参照．

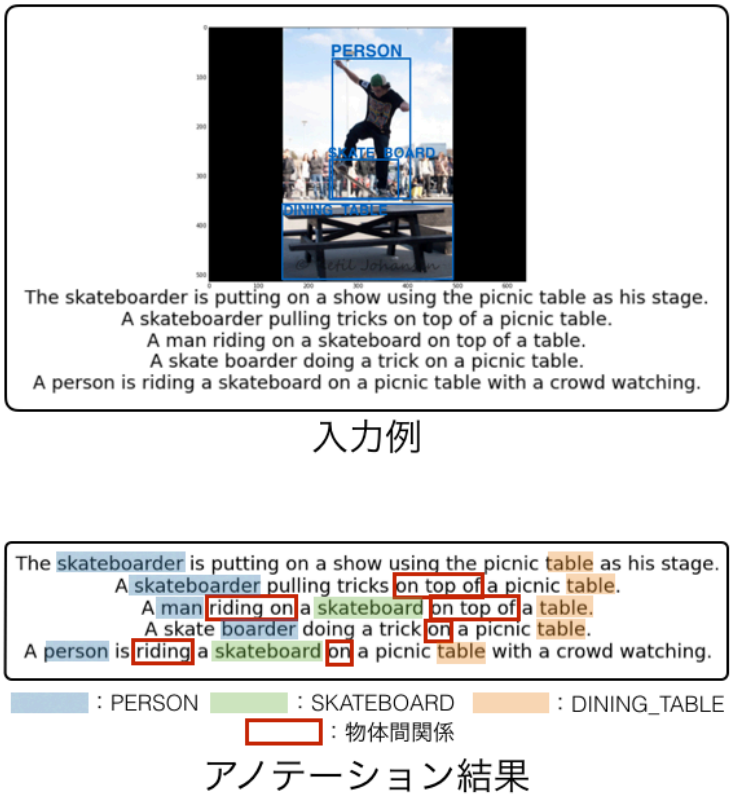


図 13: アラインメントおよび物体間関係のアノテーション例

6 評価実験

本章では以下の 3 つの評価実験を行う.

- 画像中の物体のカテゴリ名と説明文中の参照表現とのアラインメントの評価 (4.1 節),
- 説明文からの物体間関係事例の抽出の評価 (4.3 節),
- 物体間の関係認識器の評価 (5 章)

6.1 評価データ作成

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，説明文を見ながら人手で物体のカテゴリ情報および関係を付与することで作成した．正解データを付与するこの作業をアノテーションと呼ぶ．

例えば，図13上のような各物体のgrounding boxが付与された画像および説明文が与えられたとする．第3文中のmanは物体PERSONを，skateboardは物体SKATEBOARDを，tableは物体DINING_TABLEをそれぞれ表すことが画像および説明文から判断できる．また，それらの物体間の関係として，*ride_on* および *on_top_of* が同説明文中で用いられている．これらの情報を図13下のように説明文に付与していく．

アノテーションの結果，55事例および32種類の関係からなる評価データが得られた．

6.2 評価指標

実験において用いる評価指標について説明する．

以下に述べる実験において，測るべき指標はシステムの出力の正しさおよび網羅性である．システムの網羅性とは，システムが人手でつけた正解のうちどの程度カバー(出力)出来ているかを示す指標である．そこで本研究では，評価指標として適合率，再現率，F値[44]を用いる．それぞれの値は次式から求められる．

$$\text{適合率} = \frac{\text{正解した事例数}}{\text{システムが予測した事例数}} \tag{9}$$

$$\text{再現率} = \frac{\text{正解した事例数}}{\text{評価データの事例数}} \tag{10}$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{11}$$

適合率がシステムの出力の正しさを測る指標である．また，再現率がシステムの網羅性を測る指標である．それらの調和平均であるF値が高いほど，そのシステムが優れていることを表す．これらの指標は機械翻訳におけるアラインメントおよび情報検索の評価指標として一般的に用いられている[45, 46, 47]．

6.3 アラインメント結果

表 7: アラインメント結果

手法	適合率 [%]	再現率 [%]	F 値 [%]
IBM Model	**88.0**	**74.3**	**80.6**
単語ベクトル	76.9	67.5	71.9
WordNet	73.8	56.3	63.8

4.1 節において提案した，IBM Model，単語ベクトルおよび WordNet を用いた物体のカテゴリ名と説明文中の参照表現とのアラインメント手法の評価結果を表 7 に示す．

IBM Model を用いたアラインメントが最も優れた結果となった．特に，適合率が 88.0% と高く，システムの出力には誤りが少ないことがわかる．アラインメントの結果が次の物体間関係事例の抽出に影響するため，ここでは適合率が高いことが重要であるが，IBM Model はそれを満たしているといえる．この結果より，物体間の関係事例抽出は IBM Model を用いて行ったアラインメント結果を用いる．

表 8: IBM Model によるアラインメントのエラーの上位 5 件

誤ってカテゴリを出力したもの上位 5 件	合計 38 事例 (19 種類)
単語 tennis にカテゴリ TENNIS_RACKET を付与	6 事例
単語 bike にカテゴリ BICYCLE を付与	4 事例
単語 surf にカテゴリ SURFBOARD を付与	4 事例
単語 wine にカテゴリ WINE_GLASS を付与	3 事例
単語 wave にカテゴリ SURFBOARD を付与	3 事例
正解のカテゴリを出力できなかったもの上位 5 件	合計 84 事例 (49 種類)
カテゴリ DINING_TABLE が付与されるべき単語 table	15 事例
カテゴリ PERSON が付与されるべき単語 snowboarder	4 事例
カテゴリ BED が付与されるべき単語 bedroom	3 事例
カテゴリ CAR, TRUCK, MOTORCYCLE が付与されるべき単語 cars	3 事例
カテゴリ TV が付与されるべき単語 screen	3 事例

IBM Model によるアラインメントのエラー分析を行った結果，エラーとして画

像中の物体を参照していない単語に誤ってカテゴリを付与したものが 38 事例 (19 種類)，および，カテゴリが付与されるべき単語に付与できなかったものが 84 事例 (49 種類) あった．それぞれのエラーに関して，事例数が多かった上位 5 件を表 8 に示す．

IBM Model が誤って出力したものに関しては正解しているようにみえるものがあるが，これらの単語はいずれも tennis court や tennis ball，wine grass のように複合名詞の一部であるため，誤りとなっている．これらの原因は IBM Model が文脈を考慮せずにカテゴリを付与することにある．一方，正解のカテゴリを出力できなかったものの原因として，IBM Model がカテゴリを付与するための閾値 $\alpha = 0.64$ が高過ぎることが考えられる．例えば翻訳確率 $P(\text{table}|c)$ を確率の高い上位 3 カテゴリを順に見てみると，$P(\text{table}|\text{DINING\_TABLE}) = 0.4$，$P(\text{table}|\text{CUP}) = 0.16$，$P(\text{table}|\text{FORK}) = 0.1$ となっている．このような確率分布になるのは，単語 table が使われる説明文が付与された画像には物体 CUP や FORK が出てくることが多いためである．

6.4 関係事例抽出結果

表 9: 物体間の関係事例抽出結果

	適合率 [%]	再現率 [%]	F 値 [%]
IBM Model の出力を用いた抽出	29.23	34.13	31.49
正解のアラインメントを用いた抽出	49.54	64.07	55.88

4.3 節で述べた説明文からの物体間の関係事例の抽出手法を評価データに適応した結果を表 9 に示す．IBM Model の出力を用いた抽出はアラインメントを IBM Model で行ったのに対し，正解のアラインメントを用いた抽出は入力として評価データに付与されたアラインメントを用いて 4.3 節で述べた手法を適用した結果である．

正解のアラインメントを用いた場合でもF値で5割程度と低い結果となった．この原因を調査するため，間違った事例を個別に検証した．正解のアラインメントを用いた抽出結果のエラー分析を行った結果を表10に示す．

表 10: 関係抽出のエラーの原因

エラーの原因	事例数
係り受け解析ミス	22事例
動詞＋前置詞の前置詞のみによるエラー	18事例
正解	12事例
意味的に不適切	10事例
A of BのBに物体がアラインされているが主辞はA	9事例
原因不明	8事例
その他	23事例

エラーの原因として最も多いのがStanford CoreNLPによる係り受け解析のミスによる抽出エラーであった．これは例えば，説明文a skateboarder putting on a show using a picnic table. のshowとuseの間に係り受け関係ができてしまうような場合である．本来はuse(a skateboarder/PERSON, a picnic table/DINING_TABLE)という関係事例を抽出したいがこの場合にはそれができない．

また，ride_on_top_of(a man/PERSON, a skateboard/SKATEBOARD) のような関係事例を抽出した時，同様にon_top_of(a man/PERSON, a skateboard/SKATEBOARD)も抽出するようにしている．動詞+前置詞の前置詞のみによるエラーは，これがlook_atやstate_atのような場合に，関係atも抽出してしまうことによるエラーである．

さらに，正解事例の取りこぼしが12事例見つかった．これにより表9の結果の改善を見込める．

一方，意味的に不適切なエラーは係り受け解析および関係事例抽出方法は正しいが，抽出された関係が不適切な事例を指す．例えば，説明文the giraffe is being kept by itself indoors. からkeep_by(the giraffe/GIRAFFE, itself/GIRAFFE)のよ

うな事例が抽出されることによるエラーである．

a group of people や a couple of birds のような名詞句は people や birds に物体がアラインされるが，主辞が group と couple であるため，動詞 (主語，目的語) のパタンでは動詞 (people/PERSON，目的語) や動詞 (birds/BIRD，目的語) のような関係事例を抽出することができない．

また，a skateboarder pulling tricks on top of a picnic table. のような説明文の場合，on_top_of(a skateboarder/PERSON, a picnic table/DINING_TABLE) は抽出するが，pull_on_top_of は pull に目的語 trick があるため，抽出しないようにプログラムを組んでいる．しかし，実際，pull_on_top_of を関係とする事例が抽出されているため，さらなる原因究明が必要である．

6.5 物体間関係認識結果

表 11: 関係ラベル予測の精度

	適合率 [%]	再現率 [%]	F 値 [%]
隠れ層なしニューラルネットワーク			
カテゴリ名のみ	31.8	23.0	25.9
+相対位置素性	32.7	24.5	27.1
+面積領域素性	34.2	24.5	27.5
隠れ層ありニューラルネットワーク			
カテゴリ名のみ	36.4	27.0	29.6
+相対位置素性	37.3	26.4	29.8
+面積領域素性	**37.6**	**30.9**	**32.6**

5 章で作成した物体間関係認識器を用いて評価データに含まれる事例に対して物体間の関係を予測した結果を表 11 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．ここで，相対位置素性とは 2 物体の相対座標を表し，面積領域素性は表 6 で表される素性である．

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．

表 12: 物体間関係認識のエラーの原因

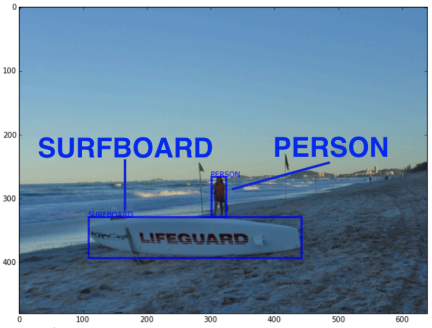
エラーの原因	事例数
データ不足	17 事例
追加情報が必要	8 事例
原因不明	5 事例
解決困難	3 事例

精度が低い原因を究明するため，面積領域素性まで用いたモデルにおける個々の事例のエラー分析を行った．結果を表 12 にまとめた．また，各エラーの具体例を図 14 に示す．

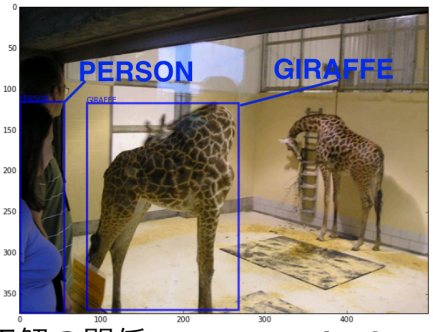
図 14(a) は物体 PERSON と物体 SURFBOARD の間の関係として in_front_of と near が正解となっている．一方，物体間関係認識器の出力は ride と on である (括弧内は関係認識器の確信度)．このような結果が得られるのは 2 物体の位置関係が ride と on を許容していることと，訓練データ中に物体 PERSON と物体 SURFBOARD の間の関係として ride(1,134 事例) と on(1,314 事例) が多く，in_front_of(19 事例) と near(8 事例) が少ないためである．このようなエラーに対して訓練データを増やすことで対処することはできず (訓練データを増やしたとしてもデータ数が少ない事例が必ず出てくるため)，それ以外の対処法を考える必要がある．

追加情報が必要なエラーとして図 14(b) のような事例がある．物体 PERSON と物体 GIRAFFE は画像中における位置関係的には隣り合っているが，実際は間にガラスの仕切りがある．この場合，正解の関係を出力するためには PERSON の向きや視線，この画像が撮影された場所情報などが必要であると考えられる．

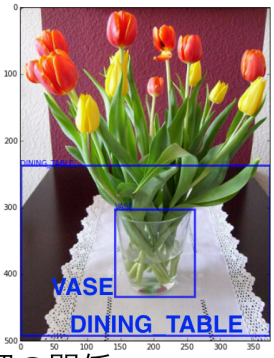
図 14(c) では関係認識器は with を出力しているが，訓練データ中に物体 VASE



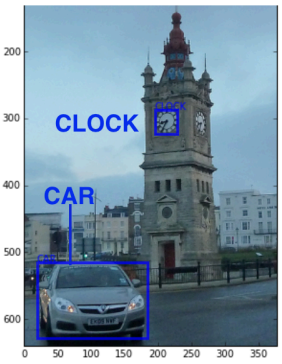
正解の関係：**in_front_of**, **near**
関係認識器の出力：ride[80.4%], on[74.8%]
(a) データ不足



正解の関係：**stare_at**, **look_at**
関係認識器の出力：なし
(b) 追加情報が必要



正解の関係：**on**
関係認識器の出力：with[81.5%]
(c) 原因不明



正解の関係：**look_at**
関係認識器の出力：なし
(d) 解決困難

図 14: 物体間関係認識器のエラー例

と物体 DINING_TABLE の間の関係として with は存在しない．この原因を突き止めるためにはさらなる調査が必要である．

人間はある物体を説明する際，”a clock looks at a car.”のように無機物に対して動作を表す動詞を述語として用いることがある (図 14(d))．しかし実際，clock に視線はなく，このような無機物の物体に対して視線情報を自動的にであれ，人間が付与するのであれ非常に困難である．これらのエラーの対処については今後の課題とする．

7　おわりに

本論文では画像理解に向けた物体間関係認識に取り組んだ．具体的には，MSCOCO [18] 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セットを使用し，物体間の関係事例の獲得，および物体間関係認識器の作成を行った．

MSCOCO には画像中の物体と説明文中との参照表現との対応関係までは付与されていない．この対応関係を求めるために，本論文では統計的機械翻訳における単語アラインメント手法の IBM Model を用いた．この対応関係と説明文の係り受け情報から物体間の様々な関係事例を大量に獲得した．この方法は既存手法 [3, 4, 5] のような関係の定義は不要であり，また，アラインメントをとることで Aditya ら [17] のように物体間の関係として不適切な事例を抽出することないというメリットがある．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．抽出した関係の中には *hold* や *ride*，*sit_on* など，人間が物体間の関係を記述するのに用いる多様な関係が確認できた．また，ある関係に関してそのフィラーである 2 物体に付与された bounding box の平均および分散を求めることで，その関係が成立する 2 次元平面上での 2 物体の平均的な位置関係を可視化した．可視化の結果，簡単なルールとして定義することが困難な関係 (wear や eat など動作を表すもの) についても，平均的な 2 物体の位置関係が求められ，それらは直感に合うものであることを確認した．

続いて獲得した物体間関係の事例を用いて，物体間関係認識器を作成した．ある物体間には同時に複数の関係が相互に依存して成立しうると考えるのが自然であるため，本研究では認識器として 1 層の隠れ層を持つニューラルネットワークを用いた．また，物体間の認識には物体間の物理的な相対情報が有用であるだろうと期待の下，認識器の入力素性として，物体のカテゴリ名および bounding box に加え，面積比や重なり度合いなどの物体間の相対情報を用いた．作成した認識器の評価実験の結果，関係認識のためにはニューラルネットワークの隠れ層および，物体間の相対情報が有用であることが確認できた．

今後の課題としてより正確な評価を行うためにクラウドソーシングを用いた評価用データの作成を行う必要がある．また，本研究で作成した物体間関係認識器を用いて画像説明文生成を行い，深層学習を用いた既存手法と比較することも興

味深い．一方，(物体 1，関係，物体 2) の 3 項組からそれらを満たす画像の検索を行うことも実用上有用であり，これに向けて，本研究で収集した大量の物体間関係事例を用いて，画像検索システムを構築しその精度を調査することも今後の課題である．

謝辞

本研究を通して終始，適切なご指導ご助言をいただき，厳しくも温かく見守って下さった指導教員の乾健太郎教授に心より深く感謝致します．同じく本研究を通して終始研究の相談に親身に乗っていただき，本論文の執筆に関して懇切丁寧なご指導いただいた指導教員の岡崎直観准教授に心より深く感謝致します．

また，本論文の審査過程において貴重なご助言を賜った本学 大町真一郎教授および北村喜文教授に深く感謝致します．

本研究を進めるにあたり，特に画像処理分野に関する多くのご助言・技術提供していただいた本学 工学研究科岡谷研究室 岡谷貴之教授および山口光太助教，博士後期課程３年 齋藤真樹氏ならびに同研究室の皆様に感謝申し上げます．

日頃より研究方針に関する数々のご指導ご助言をいただいた乾・岡崎研究室の松林優一郎研究特任助教および，研究会やその他様々な機会での議論においてたくさんのアドバイス・アイディアの提供をしていただいた同研究室の皆様に感謝致します．また，研究に専念できるよう研究室の環境づくりや事務処理等，様々な面で多大なサポートをしていただいた八巻智子秘書，成田順子技術補佐員，菅原真由美秘書に感謝致します．

最後になりますが，これまであらゆる場面において支えたくれた家族と友人に感謝します．ありがとうございます．

参考文献

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[2] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, Vol. 2, p. 3, 2010.

[3] Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 42–52, Beijing, China, July 2015. Association for Computational Linguistics.

[4] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Generating multi-sentence natural language descriptions of indoor scenes. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 93.1–93.13. BMVA Press, September 2015.

[6] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-

term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. Densecap: Fully convolutional localization networks for dense captioning. *CoRR*, Vol. abs/1511.07571, , 2015.

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. June 2015.

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

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, Vol. 61, pp. 85–117, 2015.

[13] Yoshua Bengio. Deep learning of representations: Looking forward. In *Statistical language and speech processing*, pp. 1–37. Springer, 2013.

[14] Yoshua Bengio Ian Goodfellow and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, Vol. abs/1512.03385, , 2015.

[16] 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 recog-

nition challenge. *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

[17] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermüller, and Yiannis Aloimonos. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *CoRR*, Vol. abs/1511.03292, , 2015.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zürich, September 2014.

[19] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[20] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, Vol. abs/1310.1531, , 2013.

[21] Ali Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813, 2014.

[22] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pp. 15–29. Springer, 2010.

[23] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pp. 1143–1151, 2011.

[24] Gaurav Kulkarni, Visruth Premraj, Vicente Ordonez, Sudipta Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara Berg. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 35, No. 12, pp. 2891–2903, 2013.

[25] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pp. 853–899, 2013.

[26] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pp. 67–78, 2014.

[27] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *CoRR*, Vol. abs/1503.01817, , 2015.

[28] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[29] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[31] Zellig S Harris. Distributional structure. *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

[32] Christiane Fellbaum. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pp. 665–670, Oxford, 2005. Elsevier.

[33] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of 10th International Conference on Research in Computational Linguistics, ROCLING ' 97, 1997.

[34] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pp. 805–810, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[35] Claudia Leacock and Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database.*, chapter 13, pp. 265–283. MIT Press, 1998.

[36] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pp. 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[37] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pp. 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[38] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[39] Samuel Fernando and Mark Stevenson. A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, pp. 45–52. Citeseer, 2008.

[40] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pp. 38–41. Association for Computational Linguistics, 2004.

[41] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[42] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[43] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, 2011.

[44] Allen Kent, Madeline M Berry, Fred U Luehrs, and James W Perry. Machine literature searching viii. operational criteria for designing information retrieval systems. *American documentation*, Vol. 6, No. 2, pp. 93–101, 1955.

[45] Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3*, pp. 1–10. Association for Computational Linguistics, 2003.

[46] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[47] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Em-*

pirical Methods in Natural Language Processing, pp. 1535–1545. Association for Computational Linguistics, 2011.

[48] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Vol. 7 of *IJCAI'07*, pp. 2670–2676, 2007.

[49] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534. Association for Computational Linguistics, 2012.

[50] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. Patty: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1135–1145. Association for Computational Linguistics, 2012.

[51] Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. Open information extraction with tree kernels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 868–877. Association for Computational Linguistics, 2013.

[52] Andrea Moro and Roberto Navigli. Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, IJCAI'13, 2013.

[53] Sho Takase, Naoaki Okazaki, and Kentaro Inui. Fast and large-scale unsupervised relation extraction. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, 2015.

[54] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Vol. 5, No. 4, pp. 115–133, 1943.

[55] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, Vol. 2, No. 3, pp. 183–192, 1989.

[56] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, Vol. 5, No. 3, p. 1, 1988.

付録

A　品詞タグ一覧

Stanford CoreNLP の品詞タグ付けにおいて付与される品詞タグおよびその説明の一覧を表 13 に示す．全 36 種類の品詞タグに加え，句読点や特殊記号に関する 12 種類のタグが存在する．説明文中から物体間の関係事例を獲得する際に使用する．単語自身ではなく品詞を見ることで物体間の多様な関係を獲得することが可能となる．

表 13: Stanford CoreNLP で使用される品詞タグ一覧

品詞タグ	説明	品詞タグ	説明
CC	Coordinating conjunction	TO	*to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential *there*	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present participle
IN	Preposition/subordinating conjunction	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sing. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sing. present
JJS	Adjective, superlative	WDT	*wh*-determiner
LS	List item marker	WP	*wh*-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	*wh*-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PRP$	Possessive pronoun	)	Right bracket character
RB	Adverb	”	Straight double quote
RBR	Adverb, comparative	’	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol (mathemathical or scientific)	”	Right close double quote

B　係り受けタグ一覧

Stanford CoreNLP の係り受け解析で付与されるタグおよびその説明の一覧を表 14 に示す．説明文中から物体間の関係事例を獲得する際に使用する．自然言語文から様々な (関係) 知識を獲得するタスクは Open Information Extraction[48] と呼ばれ，近年は係り受け構造などの単語間の依存構造を利用した手法が主流となっている [49, 50, 51, 52, 53]．

表 14: Stanford CoreNLP で使用される係り受けタグ一覧

係り受けタグ	説明	係り受けタグ	説明
ACL	clausal modifier of noun	EXPL	expletive
ACL:RELCL	relative clause modifier	FOREIGN	foreign words
ADVCL	adverbial clause modifier	GOESWITH	goes with
ADVMOD	adverbial modifier	IOBJ	indirect object
AMOD	adjectival modifier	LIST	list
APPOS	appositional modifier	MARK	marker
AUX	auxiliary	MWE	multi-word expression
AUXPASS	passive auxiliary	NAME	name
CASE	case marking	NEG	negation modifier
CC	coordination	NMOD	nominal modifier
CC:PRECONJ	preconjunct	NMOD:NPMOD	noun phrase as adverbial modifier
CCOMP	clausal complement	NMOD:POSS	possessive nominal modifier
COMPOUND	compound	NMOD:TMOD	temporal modifier
COMPOUND:PRT	phrasal verb particle	NSUBJPASS	passive nominal subject
CONJ	conjunct	NSUBJ	nominal subject
COP	copula	NUMMOD	numeric modifier
CSUBJ	clausal subject	PARATEXIS	parataxis
CSUBJPASS	clausal passive subject	PUNCT	punctuation
DEP	dependent	REMNANT	remnant in ellipsis
DET	determiner	REPARANDUM	overridden disfluency
DET:PREDET	predeterminer	ROOT	root
DISCOURSE	discourse element	VOCATIVE	vocative
DISLOCATED	dislocated elements	XCOMP	open clausal complement
DOBJ	direct object		

C Stanford CoreNLP で解析可能な複合前置詞

Stanford CoreNLP が Multi Word Expression(MWE) として定め，1 トークンとして係り受け解析できるようになっている複合前置詞 58 種類を表 15 に示す．表 15 にある on_top_of や in_front_of などの複合前置詞は物体間の関係を表すのに有用である．しかし，4.3 節において示した物体間関係事例抽出のためのパタンでは獲得することができず，また，簡単なルールとして定義することが難しい．そのため説明文中に表 15 にある複合前置詞のいずれかが完全一致する単語列が存在した場合，それを物体間関係の抽出対象する．

表 15: Multi Word Expression として解析される複合前置詞

according_to	as_per	compared_to	instead_of	preparatory_to
across_from	as_to	compared_with	irrespective_of	previous_to
ahead_of	aside_from	due_to	next_to	prior_to
along_with	away_from	depending_on	near_to	pursuant_to
alongside_of	based_on	except_for	off_of	regardless_of
apart_from	because_of	exclusive_of	out_of	subsequent_to
as_for	close_by	contrary_to	outside_of	such_as
as_from	close_to	followed_by	owing_to	thanks_to
by_means_of	in_case_of	in_place_of	on_behalf_of	with_respect_to
in_accordance_with	in_front_of	in_spite_of	on_top_of	in_addition_to
in_lieu_of	on_account_of	with_regard_to		

D　フィードフォーワードニューラルネットワーク

物体間関係認識器として使用したニューラルネットワークの構造およびその学習について詳しく述べる.

D.1　ネットワークの構造

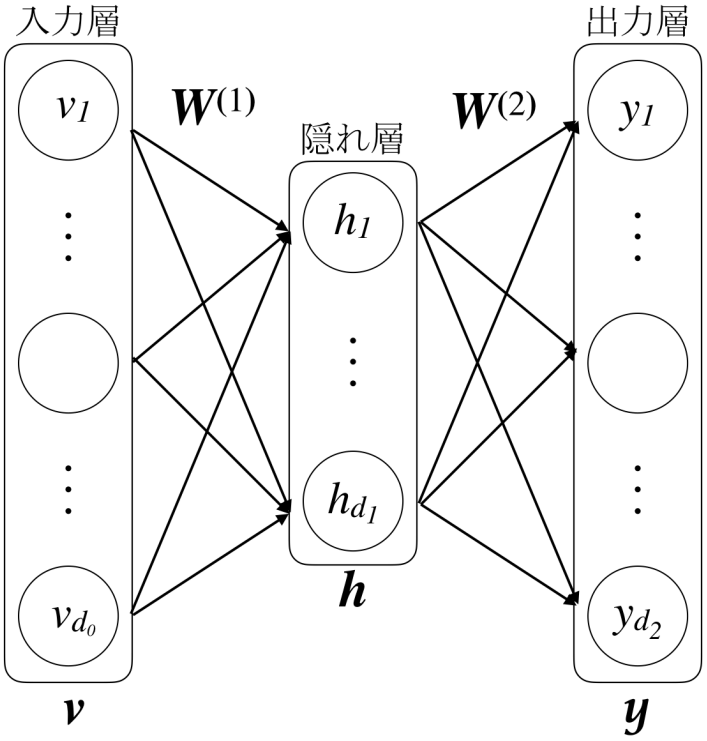


図 15: フィードフォーワードニューラルネットワーク

本論文で使用したニューラルネットワークはフィードフォーワードニューラルネットワークと呼ばれ，入力層から出力層へ一方向に信号が伝播する (図 15)．図中の白抜きの丸印 (ユニットと呼ばれる) は実数値をとり，縦一列に並んだユニットは層を形成する (数学的には多次元実数ベクトル)．入力層より高次の層の各ユニットは，その層より前の層からの重み付き和で求められる．例えば，図 15 の

隠れ層の 1 番目のユニット h_1 への入力は

$$w_{1,1}^{(1)}v_1 + w_{1,2}^{(1)}v_2 + \cdots w_{1,d_0}^{(1)}v_{d_0} + b_1 = \sum_{k=1}^{d_0} w_{1,k}^{(1)}v_k + b_1 \tag{12}$$

と表される (図 15 中では簡単のためバイアス項は省略している)．実際に h_1 を求めるには，上式に活性化関数 $\sigma(.)$ を適用する (詳しくは後述する)．これは神経細胞 (ニューロン) を数学的にモデル化した McCulloch-Pitts Neuron Model[54] から着想を得ている．1 層以上の隠れ層を持ち，非線形な活性化関数を用いたニューラルネットワークは，隠れ層に十分な数のユニットがあれば，有界閉集合上に定義域を持つ任意の連続関数を任意の精度で近似できる能力を持つことが証明されている [55]．すなわち，次節で述べる訓練データを再現するような関数をニューラルネットワークを用いて表せると期待できる．

入力層において信号 $\boldsymbol{v} \in \mathbb{R}^{d_0}$ を受け取ると，ニューラルネットワークはモデルパラメータ $\boldsymbol{W}^{(1)} \in \mathbb{R}^{d_1 \times d_0}$(重み行列) および $\boldsymbol{b}^{(1)} \in \mathbb{R}^{d_1}$(バイアス項) を用いて以下の線形変換を行う:

$$\boldsymbol{h}_{in} = \boldsymbol{W}^{(1)}\boldsymbol{v} + \boldsymbol{b}^{(1)}. \tag{13}$$

これに活性化関数と呼ばれる非線形関数 (シグモイド関数が一般的である) $\sigma(.)$ をベクトルの各要素ごとに適用することで隠れ層 $\boldsymbol{h} \in \mathbb{R}^{d_1}$ を得る:

$$\boldsymbol{h} = \sigma(\boldsymbol{h}_{in}). \tag{14}$$

同様にモデルパラメータ $\boldsymbol{W}^{(2)} \in \mathbb{R}^{d_2 \times d_1}$，$\boldsymbol{b}^{(2)} \in \mathbb{R}^{d_2}$ を用いて線形変換を行い，目的に応じた出力関数を適用することで出力層 $\boldsymbol{y} \in \mathbb{R}^{d_2}$ を得る:

$$\boldsymbol{y} = \sigma(\boldsymbol{W}^{(2)}\boldsymbol{h} + \boldsymbol{b}^{(2)}) \tag{15}$$

本研究では活性化関数および出力関数としてロジスティックシグモイド関数 $\sigma(x) = \frac{1}{1+\exp(-x)}$ を用いた．

ニューラルネットワークがある入力に対して所望の出力を得るために，モデルパラメータ $\theta = \{\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}\}$ を適当な値に選ぶ必要がある．これをニューラルネットワークの学習と呼び，一般的には以下で述べる誤差逆伝播法 [56] を用いて学習する．

D.2 学習

ある入力 $\boldsymbol{v}_i$ に対して正解の出力 $\boldsymbol{t}_i$ が既知である事例が T 個あるとする．これを訓練データと呼ぶ．この訓練データの入力 $\boldsymbol{v}_i$ からニューラルネットワークを用いて計算される出力 $\boldsymbol{y}_i$ と正解 $\boldsymbol{t}_i$ の誤差を適当な損失関数 $J(\theta)_i$ を用いて表すとき，

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{T} J(\theta)_i \tag{16}$$

を求めることが学習の目標である．具体的には，損失関数 $J(\theta)_i$ の勾配 $\nabla J(\theta)_i$ を用いて，以下の更新式よりモデルパラメータを更新する．

$$\theta_{i+1} \leftarrow \theta_i - \alpha \nabla J(\theta)_i \tag{17}$$

学習率 α はハイパーパラメータである．各 $\boldsymbol{W}^{(l)}$(ただし $l = 1, 2$) に関する勾配は次式より求められる．

$$\nabla J(\theta)_i = \frac{\partial J(\theta)_i}{\partial \boldsymbol{W}^{(l)}} = \boldsymbol{\delta}^{(l)} \boldsymbol{h}^{(l-1)} \tag{18}$$

$$\boldsymbol{\delta}^{(l)} \equiv \frac{\partial J(\theta)_i}{\partial \boldsymbol{h}_{in}^{(l)}} \tag{19}$$

また，連鎖律 (chain rule) より，

$$\boldsymbol{\delta}^{(l)} \equiv \frac{\partial J(\theta)_i}{\partial \boldsymbol{h}_{in}^{(l)}} = \frac{\partial J(\theta)_i}{\partial \boldsymbol{h}_{in}^{(l+1)}} \frac{\partial \boldsymbol{h}_{in}^{(l+1)}}{\partial \boldsymbol{h}_{in}^{(l)}} \tag{20}$$

であり，

$$\boldsymbol{h}_{in}^{(l+1)} = \boldsymbol{W}^{(l+1)} \boldsymbol{h}^{(l)} = \boldsymbol{W}^{(l+1)} \sigma(\boldsymbol{h}_{in}^{(l)}), \tag{21}$$

$$\frac{\partial \boldsymbol{h}_{in}^{(l+1)}}{\partial \boldsymbol{h}_{in}^{(l)}} = \boldsymbol{W}^{(l+1)} \sigma'(\boldsymbol{h}_{in}^{(l)}) \tag{22}$$

から

$$\boldsymbol{\delta}^{(l)} = \boldsymbol{\delta}^{(l+1)} \boldsymbol{W}^{(l+1)} \sigma'(\boldsymbol{h}_{in}^{(l)}) \tag{23}$$

が得られる．すなわち，出力層における勾配が求められれば，それ以前の層のモデルパラメータの勾配は出力層側から連鎖的に求められる [23]．

[23] 入力信号が入力層から出力層に伝わるのに対して，誤差はこれと逆向きに伝播することがこの学習方法が誤差逆伝播法と呼ばれる所以である．

発表文献一覧

受賞一覧

- 2014 年 3 月 総長賞 受賞
- 2014 年 3 月 言語処理学会第 20 回年次大会 若手奨励賞 受賞
- 2014 年 3 月 言語処理学会第 20 回年次大会 優秀賞 受賞
- 2014 年 3 月 第 3 回サイエンス・インカレ 独立行政法人科学技術振興機構理事長賞 受賞

国際会議論文

1. Masayasu Muraoka, Sonse Shimaoka, Kazeto Yamamoto, Yotaro Watanabe, Naoaki Okazaki and Kentaro Inui. Finding The Best Model Among Representative Compositional Models. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC), pp.65-74, December 2014. http://www.aclweb.org/anthology/Y14-1010

国内会議・研究会論文

1. 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．画像説明文生成に向けた物体間の関係の認識．言語処理学会第 22 回年次大会．
2. 村岡雅康，島岡聖世，山本風人，渡邉陽太郎，岡崎直観，乾健太郎．係り受け関係を用いた句ベクトルの生成．言語処理学会第 20 回年次大会．pp.1055-1058，2014 年 3 月．http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/A7-3.pdf

3. 島岡聖世，村岡雅康，山本風人，渡邉陽太郎，岡崎直観，乾健太郎．ガウス分布による単語と句の意味の分布的表現．言語処理学会第 20 回年次大会，pp.1051-1054，2014 年 3 月．http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/A7-2.pdf

4. 島岡聖世，村岡雅康．大規模言語データから単語間の意味的類似性と意味の広がりを自動学習する確率的言語モデル．第 3 回サイエンス・インカレ　– 学生による自主研究の祭典–．p.26，2014 年 3 月．http://www.science-i.jp/archive/2013/report/summary/pdf/summary_all.pdf