非テキストデータと接続可能なテキスト解析・推論技術の研究開発

宮尾　祐介

本稿では，自然言語テキストと画像やデータベースなどの非テキストデータをつなぐ理論・技術に関する研究について紹介する．具体的には，大規模データベースに対して自然言語で質問応答を行う研究と，画像に対して自然言語に基づく意味表現を認識する研究について概説する．

キーワード：自然言語処理，データベース検索，画像処理，意味解析

1. はじめに

本稿では，科学技術振興機構 (JST) の CREST ／さきがけ事業「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」において推進しているさきがけ研究「非テキストデータと接続可能なテキスト解析・推論技術の研究開発」について，これまでの研究成果と今後の展望を述べる．

人間の知的活動を模倣できるコンピュータシステムを実現することは，人類の夢の一つであり，古くから多くの研究が行われてきた．数多くの失敗を乗り越えて，現在は，自然言語処理，機械学習，形式論理，画像処理，音声処理，知識表現，データマイニングなど，特定のメディアやタスクにフォーカスし，解くべき問題をなるべく小さく切り取ることで，各分野において大きな成果を挙げている．一方で，個別タスクにおいては人間に匹敵する精度に達するまでになっているものの，それによって人間の生活が飛躍的に便利になったり，人間の労働を完全に置き換えてしまうといった実感はない．人間の知的能力と現在の技術の間には，まだ大きなギャップがある．

人間の知識は，自然言語を用いて表出・伝達され，蓄積されていく．日々の社会活動においても，自然言語によるコミュニケーションが人間の知的活動を支えていることは間違いない．一方，人間は，自然言語テキストで記述された知識だけでなく，画像・映像といった視覚データ，センサーデータや統計データなどの数値・時系列データ，データベースなど，さまざまなメディアで蓄積された情報を統合的に理解しながら，社会活動やコミュニケーションを行っている．たとえば，写真を見たらそこに何が映っているか，何を行っているか，といった情報を得ることができ，その内容に関する質問に答えることができる．たとえば科学研究においては，実験結果の数値データを見ることで，何が起きているのかを理解することができる．このように，人間は非テキストデータと自然言語を行ったり来たりしながら，新たな知識を生み出し，次の行動を行う，ということを繰り返している．しかし，異なるメディアの情報をどのように理解し，新たな知識として蓄積し，行動につなげていくのか，今のところまったく未知の領域である．

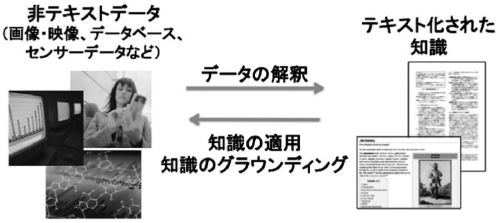

図 1　自然言語テキストと非テキストデータをつなぐ

本研究では，このように異質なメディアの情報を横断的に理解・活用するための基盤技術として，非テキストデータ（画像・映像，データベース，センサーデータなど）と，自然言語テキストとをつなぐ理論・技術の研究開発を掲げた（図 1）．本稿では，これまでの研究成果の中から，データベースと自然言語テキストをつなぐ技術（2 節）と，画像データと自然言語テキストをつなぐ技術（3 節）について紹介する．4 節では，これ以外の研究活動と，今後の展望について述べる．

この研究は，JST CREST ／さきがけの研究領域

みやお　ゆうすけ
国立情報学研究所
〒 101–8430 東京都千代田区一ッ橋 2–1–2
独立行政法人科学技術振興機構，さきがけ
〒 332–0012 埼玉県川口市本町 4–1–8
yusuke@nii.ac.jp

質問文: `Which river does the Brooklyn Bridge cross?`
答え: `http://dbpedia.org/resource/East_River`

図 2　データベースに対する質問応答の例

```
SELECT DISTINCT ?uri WHERE {
    res:Brooklyn_Bridge dbo:crosses ?uri
}
```

図 3　図 2 の質問を表す SPARQL クエリ

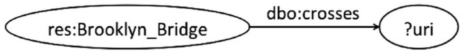

図 4　図 3 のクエリが表すグラフ

「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」に属している．しかし，本研究で扱うデータは必ずしも「ビッグ」ではない．この研究では，ビッグデータのサイズに着目するのではなく，ビッグデータの「多様性」に着目した．すなわち，ビッグデータはただ単にサイズが大きいデータなのではなく，多種多様なデータが入り乱れており，それらを統合的に利用して新たな情報や知識を得ることが本質的であると考えている．

近年では，大規模テキストからデータベースを自動構築する研究 [1]，画像に対して説明文を自動生成する研究 [2, 3]，画像に対して自然言語による質問応答を行う研究 [4, 5] など，異なるメディアをつなぐさまざまな解析技術の研究が行われている．これらの研究のほとんどは，特定の入出力（たとえば，画像とその説明文）を設定し，それらを直接つなぐことを目的としている．

本研究では，これらの研究のように異なるメディアを直接つなぐのではなく，異なるメディアに共通の「意味表現」を探求することを目指した．具体的には，自然言語テキストを解析することで得られる構文・意味構造を「共通の意味表現」とし，これを介してテキストデータとデータベース・画像データを接続するフレームワークを提案する．現段階では，本稿で紹介する研究はそれぞれ独立しており，たとえばデータベースと画像データをつなぐ具体的な研究は行っていない．しかし，さまざまなメディアから自然言語に基づく意味表現を高精度で得られる技術が確立されれば，それを介して非テキストデータ同士をつないだり，複数の非テキストデータ（たとえば画像とデータベース）とテキストデータを統合的に利用するシステムの開発につながると期待される．

2.　データベースに対する自然言語による質問応答

本研究は，DBPedia や Freebase といった大規模データベース（リンクトデータ）を知識源として用いて自然言語の質問に答えることを目的とする [6]．このタスクにおける代表的なコンペティションである Question Answering over Linked Data (QALD) [7] で提供されたデータの例を図 2 に示す．後述するように，類似したタスク設定でいくつかの研究コミュニティやデータセットがあるが，QALD では，質問文が自然言語テキストで与えられ，答えを Uniform Resource Identifier (URI) で返すことが求められる（答えが複数の場合もある）．

このタスクは，与えられたデータベース中（図 2 の場合は DBPedia）に質問に対する答えが存在することを前提としている．すると，データベースに対する検索クエリを適切に記述すれば，答えを得ることができる．たとえば，図 2 の例については，図 3 のようなクエリを作り，データベースに問い合わせればよい[1]．このクエリは，SPARQL という検索用言語で書かれている．詳細は省くが，`WHERE` 以下のブロックが検索条件を表している．各行は，リンクトデータの一つのトリプルを表しており，最初と最後の要素がデータベース中のノード（主語と目的語），真ん中の要素がそれらの関係（述語）を表している．また，`?` で始まるトークンは変数である．上の例では，`res:Brooklyn_Bridge` というノードと変数 `?uri` が `dbo:crosses` という関係でつながっていることを表している（図 4 参照）．`res:Brooklyn_Bridge` は Brooklyn Bridge を表すノードの URI，`dbo:crosses` は X が Y を `cross` する，という関係を表していることから，このクエリは，“Brooklyn Bridge が cross しているものは何？”という検索要求を意味している．

このように，適切な検索クエリを構築し，データベースに対する自然言語質問応答が実現できるのであれば，このタスクは，自然言語の質問文を検索クエリに変換する問題に帰着できる．上記の例では，“the Brooklyn Bridge” を `res:Brooklyn_Bridge` に，“does... cross” を `dbo:crosses` に，“which river” を `?uri` に変換し，これらを適切な順番で組み合わせれば，図 3 のクエリを得ることができ，正しい答えが得られる．

データベースに対する自然言語質問応答，あるいは，

[1] ここでは，簡単のため簡略化した URI を用いた．詳しくは，QALD のタスク説明を参照．

自然言語の質問文をデータベース検索クエリに変換するタスクは，独立した三つのコミュニティで研究が行われており，それぞれ異なる標準データセットが用いられている．一つは，自然言語処理コミュニティである．自然言語処理の分野では，自然言語テキストを解析して「意味表現」を得る意味構造解析が古くから研究されている．その一つとして，自然言語文をデータベースクエリに変換する研究が行われていた [8]．つまり，データベースクエリを「意味表現」とみなして意味構造解析の研究を行う，というモチベーションである．この研究の発展形として，Freebase に対する質問応答の研究が行われるようになった [9].

二つ目は，セマンティックウェブやリンクトデータのコミュニティである．この研究分野では，機械可読な大規模データベースや，それを利用した推論・検索手法の研究が盛んに行われた．その結果，現在までに多数の大規模データベースや検索システムが構築され，さまざまな分野で利用されている．しかし，これらのデータベースを利用するためには，それぞれのデータベースのスキーマや構造，データベースで用いられているボキャブラリを知ったうえで，図 3 に示すようなクエリを書く必要がある．これは一般ユーザにとっては非常に高い壁であり，リンクトデータがより広く利用されるために，自然言語によるインタフェースが一つの可能性として期待されている [7].

もう一つが，生物情報学を始めとする，大規模データベースを利用するコミュニティである [10]．生物情報学では，研究データや研究成果をデータベースとして公開・共有するしくみが古くから運用されており，実際にこれらのデータベースを活用してさまざまな研究が行われている．しかし，生物情報学ではさまざまな種類の大規模データベースが構築されており，それらに効率的にアクセスすることは難しい．したがって，自然言語を利用したデータベースアクセスが解決策の一つとして模索されている．

これらの三つの研究分野が，同時発生的に類似した研究テーマを掲げていることは興味深い．これは，このタスクがいくつかの重要な側面をもっていることを表している．まず，自然言語に関する研究の立場からすると，このタスクは自然言語表現を実世界のデータにグラウンディングすることを目指していると見ることができる．前述のように，自然言語の意味構造解析は古くから研究が行われてきた．しかし，出力である「意味表現」は原理的に観測不可能であり，明確に定義することは難しい．しかし，データベースの検索というタスクを設定すれば，「意味表現」は自ずと決まるため，研究プログラムとして取り組みやすく，また評価や比較が容易にできる．

一方，セマンティックウェブ・リンクトデータから見ると，自然言語による質問応答は，大規模かつ複雑なデータベースに対する効率的なインタフェースの一つとみることができる．また，生物情報学などのデータベースを利用する立場からみると，データベースに対する自然言語質問応答は，研究活動を効率化・活性化するために必須の技術とみなされる．このように，データベースと自然言語をつなぐ技術は，基盤技術として，インタフェースとして，さらには実用アプリケーションとしても必要とされているのである．

以下では，自然言語処理の立場からみたタスク設定と，われわれの提案手法について概説する．上述のように，本タスクは，自然言語文を入力，データベースクエリを出力とするデータ変換問題ととらえることができる．学習データとして，図 2 のような質問文と答え，さらに図 3 のようなクエリが数百～数千組与えられる．したがって，与えられた学習データから，自然言語文からクエリへの変換ルールを自動的に得ることが問題の本質である．変換ルールは，2 種類必要である．一つは，単語・フレーズを URI に変換するルール，もう一つは，URI を組み合わせてクエリを構成するルールである．前者をリンキング問題，後者はクエリ構成問題と呼ぶ．

リンキング問題は，対象のデータベースが非常に大きい（数百万ノード以上）のに比べて学習データがごく小さいことから，教師あり学習は不可能である．したがって，自然言語表現とデータベース中の情報（URI の文字列やメタ情報など）との類似度を定義する手法や，学習データを用いて類似度計算モデルを学習する手法が一般的である．一方，クエリ構成問題は，自然言語のもつ構造とデータベースの構造との関係をモデル化する必要がある．自然言語処理においては，文の背後にある構文構造や意味構造を計算する手法が古くから研究されている．たとえば，図 2 の質問文に対して構文解析を行うと，図 5 のようなデータ構造が得ら

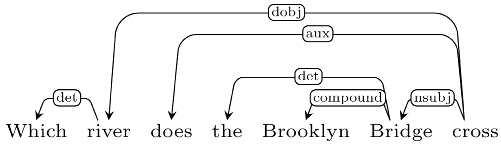

図 5　図 2 の質問文の構文解析結果

表 1　Universal Dependencies の依存関係ラベル

`nsubj`	主語
`dobj`	直接目的語
`nmod`	修飾語
`compound`	複合語
`det`	限定詞
`aux`	助動詞
`case`	前置詞

れる．これは，ノードが単語，エッジが単語間の関係を表す木構造であり，依存構造木と呼ばれる[2]．この例では，“cross” の主語が “Brooklyn Bridge” であり，目的語が “which river” であることがわかる．これは図 4 のグラフとほぼ等価な構造である．したがって，クエリ構成問題では，自然言語の構文・意味構造を参照しながら，リンキング結果を組み合わせていくことが必要となる．

われわれは，リンキングとクエリ構成を同時に最適化する手法を提案した [6]．リンキング問題は文中のコンテキストに大きく依存するため，精度 100%を達成することは本質的に困難である．たとえば，「宮崎」という文字列は，人名，県名，都市名，組織名，会社名，商品名など，リンキング先の候補は無数に考えられる．そこで，まず各単語・フレーズに対し，リンキング先の候補 top-N を列挙する．すると，残る問題は，これらの多数の候補の中から，各単語・フレーズの正しいリンキング先を選び，かつそれらの正しい組み合わせを求めることになる．この問題は，図 5 の木構造を図 4 の木構造に変換する問題として定式化できる．そこで，提案手法では，木構造変換モデルとして tree-to-tree transducer [11] を採用した．

Tree-to-tree transducer は，入力の木構造を走査しながら木構造を出力する枠組みであり，機械翻訳などで利用されている．既存研究では，tree-to-tree transducer の変換ルールは，ノード間の変換ルール（リンキング規則）と木構造構成ルールを共にデータから学習する手法が用いられていた [12]．この手法は，学習データが膨大な機械翻訳（しばしば数百万文ペア以上の学習データが用いられる）などの応用では可能であるが，データベースに対する質問応答ではリンキングルールをデータから学習することはほぼ不可能である．そこで，類似度に基づくリンキングモデルと tree-to-tree transducer を組み合わせて，全体のクエリが最適化されるようにリンキング結果の選択とクエリ構成を行う手法を提案した．単純には，リンキング結果のすべての可能な組み合わせを列挙し，それがデータベースクエリとして有効（データベースの構造と整合する）かどうかをチェックすればよい．しかし，これは明らかに組み合わせ爆発を起こすため，動的計画法とビームサーチを組み合わせ，効率的に最適な組み合わせを求める手法を開発した．

[2] ここでは，Universal Dependencies というプロジェクトで採用されている構文構造を例として用いている．エッジのラベルについては表 1 を参照．

表 2　データベース質問応答の実験結果

	Top-1 精度	Top-10000 精度
提案手法	.64	.78
ビームサーチなし	.53	.65
リンキングなし	.00	.01

表 2 に，実験結果を示す．ここでは，自然言語処理分野において広く用いられている Free917 データセットを用いた [13]．これは，Freebase を対象とした質問応答データセットである．Top-1 精度は，システムが出力したクエリのトップ一つを用いたときの質問応答の正解率である．Top-10000 精度は，システムが出力したクエリのトップ 10000 を用いて，どれか一つでも正解が得られた割合である．Top-10000 精度は，各システムで到達できる最大精度の近似値とみなすことができる．ビームサーチなしは，貪欲法でクエリを計算するシステムであり，リンキングなしは，類似度に基づくリンキングモデルを用いず，リンキングの変換ルールも学習データから学習したシステムである．実験結果から，リンキングを学習データから得るのは非現実的であることが明らかである．また，ビームサーチを利用した探索が有効であり，精度を有意に向上させることがわかる．

3.　画像に対する意味解析

本研究は，与えられた画像に対して，その内容を表す意味表現を求めることを目的とする [14]．たとえば，図 6 の画像では，象が草むらの中で立っていることがわかる．最近盛んに研究が行われている画像の説明文生成タスクでは，図下側のような，画像の意味内容を表す自然言語文を自動生成することを目的としている [2, 3]．自然言語と画像をつなぐ研究としてこれは非常に興味深いタスクである．しかし，一つの文で表現できることは限られており，画像の意味内容を利用するタスク（画像に対する自然言語検索など）では，説明文生成技術をそのまま応用することはできない．たとえば，図 6 の画像は象が歩いているように見ること

An elephant standing on rocks and grass next to a fence.
The elephant walks toward the rope perimeter fence.
An elephant with hay on its back walking towards a wire fence.
An elephant walks around a fenced area with grass.
An elephant stands in an enclosure near a fence.

図 6　画像と説明文の例（Microsoft COCO データセットより）

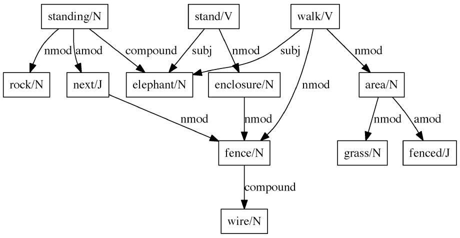

図 7　図 6 の説明文から得られる意味表現

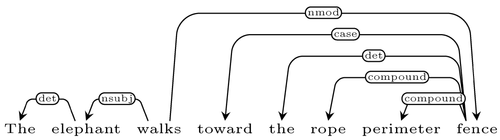

図 8　図 6 の説明文の構文解析の例

もできるため，「象が歩いている」画像を検索したら，ヒットすべきである．しかし，この画像に対して「象が草むらで立っている」という説明文を生成したら，説明文としては正しくても，前述のような検索要求には答えられない．したがって，一文で表現できる内容を超えて，画像の意味内容をできるだけ表現でき，さらにさまざまなタスクに応用することができる「意味表現」を得ることが必要である．

ただし，先に述べたように，「意味表現」を明確に定義することは難しい．特に，画像の意味内容の表現方法について確立した理論はない．そこで，本研究では，自然言語テキストを出発点として，画像理解のための意味表現を定義する．具体的には，図 6 のような画像と説明文のデータを利用し，説明文データに対して自然言語解析技術を応用して意味表現を得る（図 7）．すると，画像の意味解析は，入力画像に対して図 7 のような意味表現を出力するタスクとして定義することができる．

本研究では，以下のプロセスで画像に対する意味表現データを得る．

1. 画像に対する説明文データ（図 6 下）に対し，構文解析器を適用する．これにより，各文に対して図 8 のような構文木が得られる．
2. 構文木から内容語（名詞，動詞，形容詞など）のみを抽出する．
3. 依存関係ラベルを，意味関係ラベルに変換する．たとえば，受身文の主語は，「目的語」とする．
4. 同じ画像に対する複数の文から得られた意味表現をマージする．図 6 下の五つの説明文からは，図 7 のような意味表現グラフが得られる．

本提案手法により，画像と説明文のペアのデータがあれば，画像と意味表現のペアのデータを得ることができる．このデータは，画像を入力，意味表現を出力と考えると，画像に対して意味表現を計算するモデルの学習データとして利用することができる．入力画像に対して図 7 のようなグラフ構造を自動認識する手法はさまざま考えられるが，現在は一般物体認識で高精度を達成している Convolutional Neural Networks (CNN) を拡張したモデルで，入力画像に対して複数のラベルを認識することができるモデルを用いている [15]．既存研究は，単語をラベルとして出力するものであるが，モデルはそのままで，依存関係をラベルとして出力するモデルを学習した．

図 9 に，提案手法による意味表現認識結果を示す．図の中央にノード giraffe があり，それに関係してさまざまな依存関係（stand, walk, tall など）が認識されている．また，walk や stand の修飾語として grassy, green, open など，さまざまな形容詞が認識されている．提案手法により，一文で表現するのは難しい画像の意味内容が適切に表現されることがわかる．一方，この例では認識に失敗している関係も多くみられる．たとえば，area はどの動詞とも関係していないが，実際には stand や walk の修飾語として認識されるべきである．また，green, open などの形容詞も，field だけでなく area の修飾語としても認識されるべきである．依存関係の認識モデルは，まだまだ改良の余地が大きい．

本手法は，正解データ（画像に対してすべての正しい依存関係を与えたデータ）を構築することが困難なため，認識された意味表現の精度を直接評価すること

表3　ビデオ検索タスクの実験結果

	MEDTEST13		MEDTEST14	
	10Ex	100Ex	10Ex	100Ex
ベースライン	.2420	.4101	.1707	.3449
ベースライン ＋ 意味表現素性	.2584	.4244	.1853	.3571

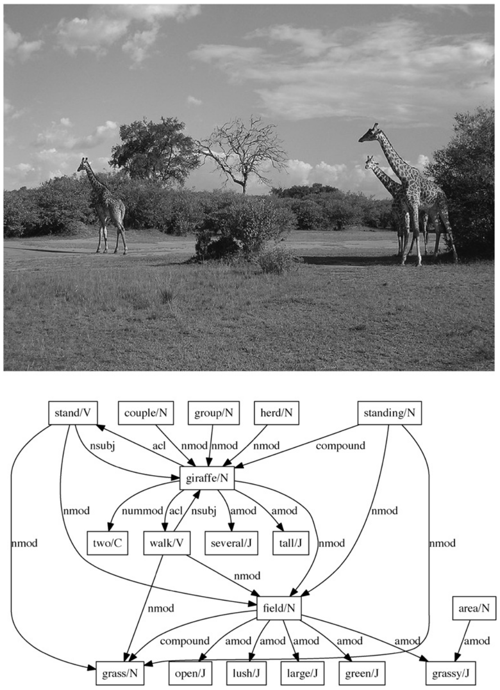

図9　画像に対する意味解析結果の例

が難しい．たとえば，図7の意味表現は人間が記述した正しい説明文から得たものであるため，正解データとみなせると思われるかもしれない．しかし，前述のように説明文にはすべての正しい情報が含まれるわけではない．たとえば，図7の例では stand と area の間に依存関係が存在しないが，これは明らかに認識すべき関係である．したがって，説明文の正解データから得た意味表現をそのまま正解データとして精度評価を行うことはあまり意味がない．実際，異なる説明文から得た意味表現同士の一致率は，20％程度である．

そこで，本研究では，応用アプリケーションに対する貢献度で意味解析の有効性を評価した．表3は，TRECVID というビデオ検索のコンペティション [16] のデータを用いたビデオ検索精度である．このタスクは，与えられたクエリ（たとえば birthday party）に対して，内容が一致するビデオを検索することを目的とする．MEDTEST13，MEDTEST14 は，それぞれ 2013 年と 2014 年のコンペティションの評価データである．10Ex と 100Ex は，学習データとして各クエリにつき正解ラベル（クエリに一致あるいは不一致）つきビデオを 10 個与える設定と 100 個与える設定である．もちろん，10Ex のほうが学習データが小さいため，難しい設定である．表3より，ベースライン（画像や音声の特徴量を用いたモデル）に対して，意味解析から得られた特徴量を加えたモデルは，いずれの設定においても精度を向上させることがわかる．間接的な評価ではあるが，本実験により，画像に対して意味表現を計算する手法の有効性が示された．

4. おわりに

本稿では，テキストデータと非テキストデータをつなぐ技術として，データベースに対する自然言語質問応答と，画像に対する意味解析の研究を紹介した．これらの研究は，本プロジェクトで掲げた「非テキストデータとテキストデータをつなぐ理論・技術」というテーマの一部にすぎない．本プロジェクトでは，映像データに対する意味解析のためのリソース構築 [17]，自然言語テキスト間の含意関係認識 [18]，株式市場データに対する意味解析の研究なども併せて推進している．これらの研究は，今のところそれぞれ独立して，自然言語テキストとさまざまなメディアとをつなぐための意味解析技術を模索している段階である．しかし，自然言語テキストを基にした構文・意味表現を核とすることはすべてに共通している．将来的には，テキストと個々のメディアをつなぐだけでなく，複数のメディアを横断した意味解析やその応用，あるいは自然言語処理において盛んに研究されている意味推論技術をほかのメディアに応用する研究など，多様な研究の展開が期待される．

自然言語テキストと多様なメディアの情報を行ったり来たりして新たな知識を紡ぎ出すという，人間があたり前に行っている活動をコンピュータで再現するまでにはまだ遠い道のりが残っている．しかし，この研究は従来の自然言語処理や画像処理の研究を超えて，人間の智に迫る研究テーマとなりうると確信している．また，本研究の重要な側面として，ほかの分野の研究者との共同研究が必要不可欠であるということが挙げら

れる．この点では，JST さきがけの枠組みの中でこの研究を行うことができたのは幸いである．今後も，この研究を通じて得たさまざまな研究者とのつながりを得て，この困難な研究テーマに立ち向かっていきたいと考えている．

謝辞　本研究は，JST さきがけの支援を受けたものである．

参考文献

[1] H. Ji and R. Grishman, "Knowledge base population: Successful approaches and challenges," In *Proceedings of ACL:HLT 2011*, pp. 1148–1158, 2011.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and L. C. Zitnick, "Microsoft COCO: Common objects in context," In *Proceedings of ECCV 2014*, pp. 740–755, 2014.

[3] P. Young, A. Lai, M. Hodosh and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, **2**, pp. 67–78, 2014.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "Vqa: Visual question answering," In *Proceedings of ICCV 2015*, pp. 2425–2433, 2015.

[5] M. Malinowski and M. Fritz, "Towards a visual turing challenge," In *Proceedings of NIPS 2014 Workshop on Learning Semantics*, 2014.

[6] P. Martnez-Gmez and Y. Miyao, "Rule extraction for tree-to-tree transducers by cost minimization," In *Proceedings of EMNLP 2016*, pp. 12–22, 2016.

[7] C. Unger, C. Forascu, V. Lopez, A.-C. N. Ngomo, E. Cabrio, P. Cimiano and S. Walter, "Question answering over linked data (QALD-5)," In *Working Notes of CLEF 2015*, 2015.

[8] Y. W. Wong and R. Mooney, "Learning for semantic parsing with statistical machine translation," In *Proceedings of HLT-NAACL 2006*, pp. 439–446, 2006.

[9] J. Berant, A. Chou, R. Frostig and P. Liang, "Semantic parsing on Freebase from question-answer pairs," In *Proceedings of EMNLP 2013*, pp. 1533–1544, 2013.

[10] K.-S. Choi, C. Unger, P. Vossen, J.-D. Kim, A.-C. N. Ngomo and T. Mitamura (eds.), *Open Knowledge Base and Question Answering Workshop*, 2016.

[11] K. Knight and J. Graehl, "An overview of probabilistic tree transducers for natural language processing," In *Computational Linguistics and Intelligent Text Processing*, pp. 1–24, 2005.

[12] J. Graehl and K. Knight, "Training tree transducers," In *Proceedings of HLT-NAACL 2004*, 2004.

[13] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," In *Proceedings of ACL 2013*, pp. 423–433, 2013.

[14] S. Phan, Y. Miyao, D.-D. Le and S. Satoh, "Video event detection by exploiting word dependencies from image captions," In *Proceedings of COLING 2016*, pp. 3318–3327, 2016.

[15] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick and G. Zweig, "From captions to visual concepts and back," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton and G. Quénot, "Trecvid 2014 – an overview of the goals," *TRECVID 2014*, pp. 1–52, 2014.

[17] 城綾実，牧野遼作，坊農真弓，高梨克也，佐藤真一，宮尾祐介，"異分野融合によるマルチモーダルコーパス設計—各種アノテーション方法と利用可能性について—，" 言語処理学会第 21 回年次大会発表論文集，pp. 561–564, 2015.

[18] R. Tian, Y. Miyao and T. Matsuzaki, "Logical inference on dependency-based compositional semantics," In *Proceedings of ACL 2014*, pp. 79–89, 2014.