Vol.2010-DBS-151 No.38
2010/11/13

格フレームを考慮した Web 検索スニペット解析による動作関係抽出

白川真澄[†1]　中山浩太郎[†2]　荒牧英治[†2]
原　隆浩[†1]　西尾章治郎[†1]

筆者らはこれまで，Wikipedia から抽出した情報に Web 全体の情報を組み合わせることにより，固有表現間の様々な動作関係を定義した大規模な辞書を構築する手法を提案してきた．しかし，実際に辞書構築を行うという観点から見た場合，関係抽出の精度や網羅性に加えて，処理時間も重要な指標となる．そこで本稿では，形態素解析を用いて，Web 検索スニペットから固有表現間の関係を表す動詞を抽出する際に，同一格フレームに存在するか否かを判別することにより，高精度を維持しつつ，網羅性向上および処理時間削減を目指す．

Case-frame-based Verb Relation Extraction using Web Search Snippet Analysis

Masumi Shirakawa,[†1] Kotaro Nakayama,[†2]
Eiji Aramaki,[†2] Takahiro Hara[†1]
and Shojiro Nishio[†1]

We proposed a method for building a large-scale dictionary which defines various relations among named entities described by verbs (predicates) by combining information extracted from Wikipedia and the Web. However, from the aspect of building a dictionary, it is important to consider not only the precision and the recall but also the processing time. In this paper, we aim at improving the coverage and cutting down the processing time while maintaining the high precision. Our method extracts verbs which describe relations between named entities from Web search snippets by judging whether both of a named entity pair belong to a same case frame using morphological analysis.

1. はじめに

近年，意味を考慮した処理が可能な次世代 Web の実現に注目が集まっており，セマンティック Web をはじめ，オントロジ，人工知能，自然言語処理など，様々な方面からこのような Web の実現に向けて研究が行われてきた．特に，世の中に存在するエンティティ間の関係を定義した辞書（以降，関係辞書と呼ぶ）は，次世代 Web を体現する様々なアプリケーションの基盤リソースとして必要とされている．実際に公開されている関係辞書として，概念辞書（上位下位関係や同位関係などの，ドメインに依存しない関係を定義した辞書）である WordNet[9)]，EDR 電子化辞書[⋆1]などがあり，検索エンジンの高度化[5)] や医療用 QA システム[17)] といったアプリケーションに利用されている．しかし，これらの辞書は，一般的なエンティティ間の関係を定義した概念辞書（上位オントロジ）であり，様々なアプリケーションの要求として，固有表現間の関係を定義した辞書が必要とされてきた．そのため，大規模な Web 百科事典である Wikipedia をマイニングすることで，固有表現間の関係抽出を試みる研究が，過去に数多く行われてきた．Wikipedia は，記事の網羅性や即時性，密で多様なリンク構造，質の高いリンクテキスト，URL による語彙の一意性など，知識抽出のコーパスとして有利な特徴を数多く持っている[11)]．実際に，DBPedia[1)] や YAGO[15)] などの研究によって，高精度で固有表現間の関係抽出が可能であることが示されてきた．しかし，これらの研究では，抽出した関係数やエンティティ数を主張する一方，得られる関係としては形式的なものが多く，関係の種類についてはあまり注目していない．

そこで，筆者らの先行研究では，Wikipedia から得られる情報に，Web の情報を組み合わせることで，固有表現間の多種多様な関係を定義した大規模な関係辞書構築を目指してきた[14)]．具体的には，Wikipedia から抽出した大規模連想シソーラス（Wikipedia シソーラス[10)]）を利用して関連のある固有表現ペアを取得した後，格助詞を付与した Web 検索クエリによって大幅に絞り込んだ Web ドキュメントから，形態素解析や係り受け解析などの自然言語処理技術を用いて動作関係（動詞と格助詞による関係）を抽出する手法を提案した．先行研究では，Wikipedia シソーラスを利用して，関係抽出の入力となる固有表現ペアを

†1 大阪大学
Osaka University
†2 東京大学
The University of Tokyo
⋆1 http://www2.nict.go.jp/r/r312/EDR/

大量に用意することで，大規模な関係辞書構築の実現を目指している．また，格助詞を付与した Web 検索クエリによって解析対象となるテキストを大幅に絞り込むことにより，既存の統計的な解析を基にした手法では無視されがちであった関係について，精度よく抽出することを図っている．その結果，既存の辞書には定義されていないような，多種多様な動作関係が抽出できることを確認できた．

しかし，実際に関係辞書を構築しようとした場合，関係抽出の精度や網羅性だけでなく，処理時間についても考慮する必要がある．高精度を維持でき，現実的な処理時間を達成しつつ，出来る限り多くの関係を抽出できるような手法が望ましいと考える．先行研究では，Web ドキュメントをダウンロードし，得られたテキストに対して係り受け解析および形態素解析を用いる手法を採用していた．しかし，Web ドキュメントをダウンロードする方法は，膨大な処理時間がかかるという問題があった．また，先行研究で用いていた係り受け解析による手法は，高い精度で動作関係を抽出できる一方，網羅性については，改善の余地があると考えられる．そこで本研究では，Web 検索スニペットを解析対象とし，形態素解析を用いて，同一格フレームに存在するか否かを判別することで固有表現間の関係を表す動詞を抽出する手法を提案する．提案手法により，高精度の維持，現実的な処理時間の達成，抽出する関係の網羅性の拡大を目指す．

2. 関 連 研 究

関係抽出の研究に関して，Wikipedia を対象とした関係抽出の研究が近年盛んに行われている．Wikipedia は，Wiki を用いて構築された大規模 Web 百科事典であり，記事の網羅性や即時性，密で多様なリンク構造，質の高いリンクテキスト，URL による語彙の一意性など，知識抽出のコーパスとして有利な特徴を数多く持っている[11]．このような特徴を持つ Wikipedia を解析し，固有表現間の大規模な関係抽出を行った例として，DBpedia[1] や YAGO[15] が挙げられる．DBpedia[1] では，Wikipedia の記事に定義されているインフォボックスと呼ばれる構造化データに着目し，インフォボックスから情報を抽出して RDF (Resource Description Framework) に変換するという簡潔な手法により，多言語にわたる大規模な関係抽出を行っている．YAGO[15] では，Wikipedia に合わせて構成されたルールベースに基づく推論とヒューリスティクス（経験則）をもとに，一般語を対象とした概念辞書である WordNet[9] のクラスに Wikipedia のカテゴリをマッピングすることで，高精度で数百万規模の関係を抽出している．これらの研究では，抽出した関係数やエンティティ数を主張する一方，得られる関係としては形式的なものが多く，関係の種類についてはあまり注目していない．

また，Web 上の膨大なドキュメントを解析対象とする関係抽出も過去に数多く行われてきた．その中でも，ブートストラッピングとよばれる関係抽出手法[3] が近年注目を集めている．関係抽出におけるブートストラッピングとは，少量のシード（種）と呼ばれる入力データを基に，統計的な手法を用いてエンティティペアと関係の取得を漸増させる手法である．また，入力データを必要としない手法として，Hasegawa らの研究[4] や Danushka らの研究[2] では，固有名詞が複数出現する文から周辺の表現を取得し，クラスタリング[4] または共クラスタリング[2] を行うことで，固有名詞のペアと，それらの関係の抽出を行っている．Yan らの研究[18] では，Wikipedia の情報の質と Web の網羅性という利点を活かし，Web のみから関係抽出を行った場合と比較して，精度および網羅性の向上を達成している．

これらの研究では，統計的な情報に基づく解析が手法の根幹となっており，母数の多い固有表現ペアや関係について，高精度で関係を抽出できることが示されている．しかし，Web 上において出現頻度が低い固有表現ペアや関係については，あまり重要視していない．本研究では，固有表現間の多種多様な関係抽出を目的としているが，実質的には，Web 上において出現頻度が低い固有表現ペアや関係に対して，精度よく関係を抽出することが課題であるといえる．そのため，先行研究[14] では，関連のある固有表現ペアに，格助詞を付与したクエリを生成して Web 検索を行い，関係抽出に有効と思われるセンテンスのみを解析対象として絞り込む手法を採用した．

日本語の関係抽出に関しては，語彙の区切りや格助詞の存在など，言語構造の違いが大きいため，日本語に特化した解析手法によって関係抽出を試みる研究が行われている．本研究と同様に，固有名詞間の様々な関係を抽出する手法として，数原らの手法[16] では，話題語（固有名詞）によるブログでの検索スニペットを解析対象とし，それぞれの語が係る動詞を抽出した後，動詞を軸に語を結合することで，動作関係（動詞と格助詞による関係）を抽出している．また，河原らの格フレーム辞書[6] では，膨大な量の新聞記事を解析し，名詞・格助詞・動詞の共起情報を基に名詞のクラスタリングを行うことで，名詞間の動作関係を抽出している．本研究では，日本語においては格助詞が関係抽出に有効であると考え，格フレームの考え方を取り入れた関係抽出手法を提案している．

3. 動作関係抽出のプロセス

本研究では，Wikipedia シソーラスを用いて関連のある固有表現ペアを取得した後，それらをクエリとした Web 検索によって解析対象となるセンテンスを絞り込み，各センテン

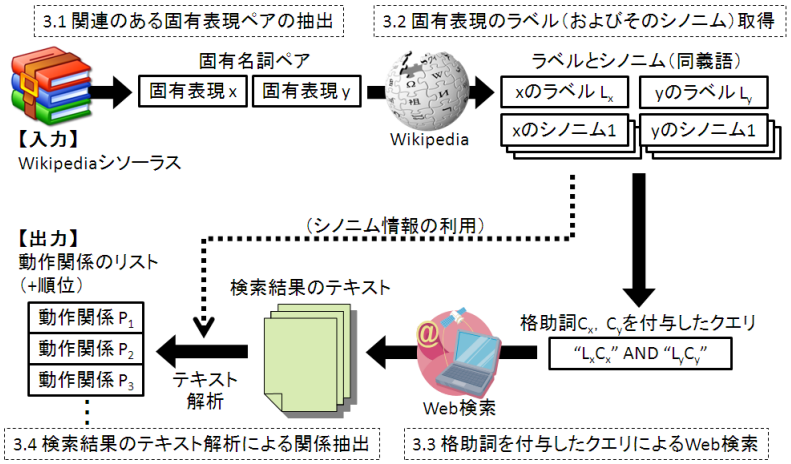

図 1 動作関係抽出のプロセス
Fig. 1 Processes of verb (predicate) relation extraction

スを解析することで動作関係（動詞と格助詞による関係）を抽出する．具体的には，図 1 に示すプロセスによって固有表現ペア間の動作関係抽出を試みる．以下では各プロセスについて説明する．なお，これらのプロセスは筆者らが先行研究[14] で提示したものである．

3.1 関連のある固有表現ペアの抽出

動作関係抽出の入力として，Wikipedia シソーラス[10] を用いる．すなわち，Wikipedia シソーラスを用いて関連のある固有表現ペア x，y を取得する．Wikipedia シソーラスは，Wikipedia で定義されているエンティティ（記事）間の関連度を，「エンティティ x，エンティティ y，関連度」という形式で定義した辞書である．Wikipedia では，特に固有表現に関する記事が数多く存在していることから，Wikipedia シソーラスについても同様に，固有表現間の関連度を豊富に定義している．Wikipedia シソーラスを用いることで，何らかの関係が存在しうる固有表現ペアを推測できるため，3.3 節において，Web 検索を用いて関係を抽出すべき固有表現ペアを絞り込める．

3.2 固有表現のラベル（およびそのシノニム）取得

自然言語処理による関係抽出においては，一つのエンティティ（または固有表現）に対して，少なくともそれを意味するラベル（代表となる表記）を認識できる必要がある．そこで，各固有表現 x，y に対して，ラベル L_x，L_y を決定するために，Wikipedia のリンクテキストを用いたシノニム（同義語）抽出手法を利用する[12]．Wikipedia では，ある記事中に他のエンティティを意味する語句が出現したとき，その語句をリンクテキストとして，そのエンティティ（記事）にリンクが張られる．したがって，あるエンティティに対して張られている全てのリンクについて，リンクテキストを解析することで，そのエンティティを意味する表記の集合を取得できる．このとき，出現頻度が相対的に高い表記ほど，そのエンティティの表記として適していると考えられるため，最も出現頻度の高いリンクテキストをラベルとして用いる．また，それ以外のリンクテキストはシノニムとして，テキスト解析時の固有表現の発見に利用可能である．

3.3 格助詞を付与したクエリによる Web 検索

固有表現ペア x，y の双方のラベル L_x，L_y に格助詞 C_x，C_y をそれぞれ付与した Web 検索クエリを生成し，固有表現のラベルと格助詞の組 L_xC_x と L_yC_y を共に含むテキスト（Web ドキュメントあるいはスニペット）を取得する．数原らが行った動作関係抽出[16] では，格助詞を付与せずに Web（ブログ）検索を行っているが，箇条書きやリストなどの箇所にクエリが適合し，動詞が出現するテキストをほとんど取得できず，動作関係を抽出できないケースが存在すると述べられている．そこで，クエリ生成の時点で格助詞を付与することで，固有表現ペアの動作関係を抽出するのに適したテキストを絞り込む．なお，筆者らの先行研究[14] において，Web 検索クエリに格助詞を付与する手法の有効性を確認している．

3.4 検索結果のテキスト解析による関係抽出

各クエリによって Web 検索を行い，検索結果の上位のテキストを取得した後，句点やピリオドなどの区切り文字によってセンテンス単位に分割する．その後，自然言語処理技術を用いて，各センテンスを解析することで，固有表現ペアの関係を表す動詞を抽出する．抽出した動詞 V とそのとき各固有表現に付属していた格助詞の組合せ C_x，C_y を一つの動作関係 P として，動作関係のリストを出力する．抽出回数に基づき，動作関係の順位付けを行うことも可能である．なお，ここで用いる動作関係抽出手法については次章で詳しく説明する．

4. 動作関係抽出手法

本研究では，実際に関係辞書構築を行うため，高精度を維持し，現実的な処理時間を達成しつつ，出来る限り多くの関係を抽出できる手法の構築を目指す．

4.1 既存手法の問題点

先行研究では，Web ドキュメントをダウンロードし，得られたテキストに対して係り受

け解析および形態素解析を用いる手法を採用していた．しかし，Web ドキュメントをダウンロードする方法は，処理時間の点で問題がある．実際，Ohshima らの研究[13] では，Web 検索スニペットを用いて関連語の抽出を行っているが，Web ドキュメントを取得する場合と Web 検索スニペットを用いる場合について比較を行っており，処理時間に約 20 倍もの差が発生している．また，係り受け解析による手法については，高精度の係り受け解析器を用いた場合，高い精度で動作関係を抽出できるが，網羅性については，改善の余地があった．これは，係り受け解析による関係抽出では，固有表現ペアの双方が同じ動詞に係る場合のみ，その動詞を抽出することになるが，このようなケースが少ないためである．数原らの研究[16] では，係り受け解析における網羅性の問題に対して，固有名詞ペアの片方が係る動詞から関係を抽出する手法を提案している．各固有名詞ごとに，それが係る動詞を抽出した後，動詞を軸にして固有名詞を結合することで，動作関係を抽出している．数原らは，ブログで話題に上がっている固有名詞ペアをテストデータとして用いた評価実験により，関係抽出の精度を維持しつつ，網羅性の向上に成功している．しかし，数原らが用いた話題語ペアは，Web 検索におけるヒット件数が多いものであるため，Web 上での出現回数が少ない固有表現ペアに対しては，動詞を結合するのに十分な動詞数が抽出できず，精度を維持することが難しいと考えられる．

4.2 提 案 手 法

本研究では，同一格フレームに存在するか否かを判別することで固有表現間の関係を表す動詞を抽出する手法を提案する．また，前節の問題点を踏まえ，Web 検索スニペットを解析対象とする．

本手法は，格フレーム辞書[6] で用いられている格フレーム（および格構文）の考え方を取り入れている．ここでいう格フレームとは，格助詞を伴う名詞句を一つの格として，連続する格とその直後の動詞のまとまりを意味している．本手法は，ある動詞の格フレーム内に，注目している固有表現ペアが収まっているか否かが，その動詞によって固有表現ペアの関係を表せるか否かに直接影響している，という考えに基づいている．例えば，同じセンテンス中に二つの固有表現が存在しており，その二つが同じ動詞の格フレームに収まっている場合，その動詞と各固有表現に付属する格助詞（動作関係）によって，二者の関係を表現できる可能性が高い．また，その二つが同じ動詞の格フレームに収まっていない場合，その動詞は二者の関係を表現するのに相応しくない可能性が高いと判断できる．このように，格フレーム判定は動作関係抽出に直接的であり，あるセンテンスに関係抽出対象となる二つの固有表現が共に出現する場合，高精度且つ高網羅性を維持した関係抽出が期待できる．

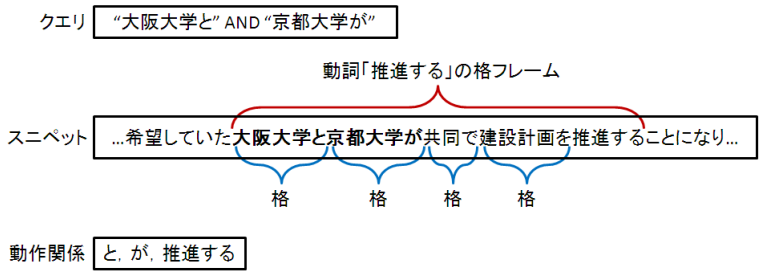

図 2 格フレーム判定を用いた動作関係抽出の例
Fig. 2 Example of verb (predicate) relation extraction using case frame judgment

以下では格フレーム判定を用いた関係抽出手法について説明する．まず，3.3 節において，固有表現 x，y のラベル L_x，L_y にそれぞれ格助詞 C_x，C_y を付与した Web 検索クエリ「“L_xC_x” AND “L_yC_y”」を生成し，検索結果の上位のスニペットを取得する．なお，ダブルクォーテーション “” はフレーズ検索（ダブルクォーテーションで囲われた語句を必ず検索結果およびスニペットに含める）を意味している．その後，3.4 節において，スニペットの中のセンテンスに対し，名詞句を $N(s)_i$，格助詞を C_i として，以下の格フレームの判定式を用いて動作関係 P を抽出する．

$$L_xC_x, L_yC_y \in \left(\sum_{i=1}^{k} N(s)_i C_i \right) V \tag{1}$$

上式は，固有表現のラベルと格助詞から成る二つの格 L_xC_x，L_yC_y が共に，一つの動詞 V から成る格フレーム内に収まっているかどうかを判定している．上式の格フレームの判定式を満たす場合，動詞 V と格助詞 C_x，C_y の組合せを，以下の式のとおり，動作関係 P として抽出する．

$$P = (C_x, C_y, V) \tag{2}$$

全てのセンテンスについて解析を行った後，抽出回数の多い順に動作関係を出力する．

格フレーム判定の具体例を図 2 に示す．クエリ「“大阪大学と” AND “京都大学が”」による検索結果のスニペットの中に，「大阪大学と京都大学が共同で建設計画を推進する」という格フレームが存在しており，その中に二つの格「大阪大学と」「京都大学が」が含まれ

Vol.2010-DBS-151 No.38
2010/11/13

表 1 評価に用いた格助詞の組合せ
Table 1 Combinations of case particles for evaluation

ガ格 + ヲ格	ガ格 + ニ格	ガ格 + デ格
デ格 + ヲ格	デ格 + ニ格	カラ格 + ニ格
カラ格 + ヘ格	カラ格 + マデ格	ト格 + ガ格

ているため，このときの動詞「推進する」と格助詞の組合せ「と」「が」を動作関係として抽出している．

5. 性能評価

提案手法である，格フレーム判定による動作関係抽出手法について，精度，網羅性，処理時間の三つの観点から評価を行った．

5.1 評価環境

評価方法として，関連のある固有表現ペアに対して，提案手法を用いて動作関係（動詞と格助詞からなる関係）を抽出し，人手による評価を行った．具体的には，Wikipedia シソーラスから，リンクテキスト数が 100 未満のノイズとなるエンティティを取り除き，ランダムに選んだ 100 の固有表現について，それぞれ最も関連度の高い固有表現を取得し，100 の固有表現ペアを選出した．その後，表 1 に示す格助詞の組合せ（9 通り × 2）を付与したクエリを生成して Web 検索を行い，スニペットを 50 件ずつ取得した．なお，表 1 の格助詞の組合せは，準備実験により決定した．準備実験では，数百程度の固有表現ペアにそれぞれ格助詞を付与したときの Web 検索のヒット件数を平均し，表 1 の組合せが，ある程度，動作関係を抽出できる可能性がある組合せであることを確認している．各手法によって得られた関係のうち，抽出回数が三位までの関係を，筆者の一人が下記に挙げる判定の基準に基づいて，正解，部分正解，不正解に分類した．なお，判定の公平性を保つため，各手法によって抽出した関係を混在させてから判定を行った．判定の基準として，固有表現ペアと動作関係のみで関係が理解もしくは推測できるものを正解，関係の理解・推測に補足情報が必要なものを部分正解，その他を不正解とした．また，ある関係が噂として一般に認知されている場合，その関係が事実とは異なっていても正解であるとみなした．

比較手法として，動作関係抽出に係り受け解析を用いた場合において，Web 検索スニペット，Web ドキュメント[14] をそれぞれ解析対象として関係抽出を行う手法，および，動詞を軸に各固有表現が係る動詞を結合する数原らの手法[16] を採用した．なお，本評価において，係り受け解析には CaboCha[7]，形態素解析には MeCab[8] を用いた．

評価指標には，抽出した関係の適合率，何らかの関係を抽出した固有表現ペア数，処理時間の三つを用いた．なお，抽出した関係の適合率は，何らかの関係を抽出した固有表現ペアに対して，固有表現ペアごとに三位までの関係について適合率（正答率）を計算した後，それらを平均した値を採用している．抽出した全ての関係について適合率を算出しないのは，固有表現ペアごとに抽出した関係数が異なっており，関係を多く抽出できた固有表現ペアの適合率に大きく影響を受けるためである．また，各固有表現ペアに対して正答率を算出する際，部分正解を含む場合と含まない場合に分けて適合率を算出した．

5.2 評価結果

評価結果を表 2 に示す．まず，同じ Web 検索スニペットを解析対象として，提案手法である格フレーム判定を用いた手法と係り受け解析を用いた手法を比較すると，提案手法は関係を抽出できた固有表現ペア数が約 1.4 倍と大きく上回っている上に，部分正解を除いた場合の適合率も優っている．これは，ある動詞が固有表現ペアを含む格フレームを成しているか否かの判定が，その動詞が固有表現ペアの関係を表すか否かの判別に直接影響しているためであると考えられる．係り受け解析を用いた場合，ある動詞の格フレーム内に固有表現ペアが収まっていてもその動詞を関係として抽出しないケースがある一方，格フレーム内に固有表現ペアが収まっていなくても動作関係を抽出するケースが存在するが，後者は抽出難度が高いため，結果的に精度および関係を抽出した固有表現ペア数が劣っていると考えられる．すなわち，格フレーム判定による動作関係抽出は，関係抽出が容易なセンテンスからの取りこぼしを極力防ぐことによって精度，網羅性が向上したといえる．処理時間に関しては，提案手法が少し劣っているが，大幅な差ではない．したがって，動作関係抽出のパフォーマンスとしては，係り受け解析を用いるよりも，格フレーム判定を用いたほうが優れているといえる．

また，数原らの手法である，動詞を軸に各固有表現が係る動詞を結合する手法についてみると，全手法中，最も多くの固有表現ペアについて関係を抽出できている．これは，一つのセンテンスに固有表現ペアの片方のみが出現するケース（片方の固有表現が省略された形）が多いためである[16]．一方，適合率に関しては，他の手法よりも劣っている．これは，別のコンテキストで用いられた動詞を結合することにより，誤った関係を導いてしまったためであると考えられる．実際に抽出された関係を見ると，特に同位関係にある固有表現ペアについて，誤った関係が多く抽出されている傾向があった．この理由として，同位関係にある固有表現は，それぞれ別のコンテキストにおいて同じ動詞を用いることが多いためであると考えられる．なお，数原の手法に関しては，提案手法に対して排他的な技術ではないため，

表 2 評価結果
Table 2 Evaluation result

評価指標	係り受け解析（ドキュメント）[14]		係り受け解析（スニペット）		係り受け解析&結合（スニペット）[16]		格フレーム判定（スニペット）	
	部分正解なし	部分正解あり	部分正解なし	部分正解あり	部分正解なし	部分正解あり	部分正解なし	部分正解あり
抽出した関係の適合率	0.668	0.849	0.663	0.875	0.513	0.662	0.744	0.882
関係を抽出した固有表現ペア数	33 ペア		34 ペア		61 ペア		47 ペア	
処理時間	77,771 秒		1,817 秒		1,764 秒		2,077 秒	

今後，手法の組合せを検討する予定である．

係り受け解析を用いた手法において，Web ドキュメントをダウンロードして全文を解析した場合とスニペットのみを解析した場合について比較すると，前者は Web ドキュメントをダウンロードするために膨大な時間を費やしてしまうため，全体としての処理時間が 40 倍以上となっている．一方，適合率や，関係を抽出できた固有表現ペア数に関しては，ほぼ同程度となっている．まず，適合率が向上していない理由として，本研究では統計的なアプローチを手法の根幹としていないことが挙げられる．統計的なアプローチでは，通常，解析対象となるテキストが多くなるほど，ノイズを除去しやすくなり，精度向上が期待できる．しかし，本研究では，固有表現ペアを Web 検索クエリとして，正しい関係を抽出できそうなセンテンスのみに解析対象を絞り込む，すなわち，ノイズを除去するのでなく避ける，というアプローチを採用している．Web ドキュメントをダウンロードすることにより，ノイズを取得してしまう可能性を増大させていることが，適合率に影響を与えていると考えられる．また，Web ドキュメントを解析対象とした場合，関係を抽出できた固有表現ペア数は増加しなかったが，実際のデータをみると，固有表現ペア当たりの関係数は上回っていた．このことから，何らかの関係を抽出できる固有表現ペアに対しては，Web ドキュメントを用いることで，より多くの関係を抽出できる可能性がある一方，関係を抽出できる見込みのない固有表現ペアに対しては，Web ドキュメントを用いてもあまり効果がないことが分かる．

以上の結果から，提案手法である格フレーム判定を用いた手法は，係り受け解析を用いた手法と比較して，処理時間をほとんど増加させることなく，より多くの動作関係を，高精度で抽出できることが確認できた．また，Web ドキュメントをダウンロードする方法を用いずに，Web 検索スニペットを解析対象とすることで，精度や抽出できる関係数にあまり影響を与えることなく，処理時間を削減できた．

6. おわりに

本稿では，Web 検索スニペットと形態素解析を用い，テキスト解析時に同一格フレームに存在するか否かを判別することにより，関係抽出の精度や網羅性だけでなく，処理時間を考慮した手法を提案した．本研究では，Wikipedia の情報の質と Web の網羅性という利点を活かし，両者の情報を組み合わせることで，固有表現間の様々な関係を定義した大規模な関係辞書構築を目指している．そのため，高精度を維持し，現実的な処理時間を達成しつつ，抽出した関係数の拡大を図る必要があった．

評価結果より，格フレーム判定による手法が，係り受け解析による手法と比べて，ほぼ同程度の処理時間で，1.4 倍程度の数の固有表現ペアに対して 88%の精度（部分的に関係を表しているものを含む）で動作関係を抽出できることを確認した．

今後の課題として，さらに精度，網羅性を向上させるため，クラスタリング手法や数原らの手法[16] との組合せを検討している．また，実際に動作関係辞書の構築を進める予定である．このような関係辞書を構築し，公開することは，関連する様々な研究分野の発展において重要であると考える．例えば，関係辞書を基盤リソースとしたアプリケーションの研究開発を行う場合，公開された関係辞書を用いることで，意味情報を扱うサービスの構築が容易になることが期待される．

謝辞 本研究の一部は，科学研究費補助金基盤研究 C(20500093)，および科学研究費補助金基盤研究 B(21300032) の助成によるものである．ここに記して謝意を表す．

参 考 文 献

1) Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. G.: DBpedia: A Nucleus for a Web of Open Data, *Proceedings of International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp.722–735

(2007).
2) Bollegala, D.T., Matsuo, Y. and Ishizuka, M.: Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web, *Proceedings of International Conference on World Wide Web (WWW)*, pp.151–160 (2010).
3) Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: Unsupervised Named-entity Extraction from the Web: An Experimental Study, *Artificial Intelligence*, Vol.165, No.1, pp.91–134 (2005).
4) Hasegawa, T., Sekine, S. and Grishman, R.: Discovering Relations among Named Entities from Large Corpora, *Proceedings of Meeting on Association for Computational Linguistics (ACL)*, pp.415–422 (2004).
5) Hemayati, R., Meng, W. and Yu, C.: Semantic-based Grouping of Search Engine Results using WordNet, *Proceedings of Joint International Conferences on Asia-Pacific Web Conference and Web-Age Information Management (APWeb/WAIM)*, pp.678–686 (2007).
6) 河原大輔，黒橋禎夫：格フレーム辞書の漸次的自動構築，自然言語処理， Vol.12, No.2, pp.109–131 (2005).
7) 工藤　拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌， Vol.43, No.6, pp.1834–1842 (2002).
8) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.230–237 (2004).
9) Miller, G.A.: WordNet: A Lexical Database for English, *Communications of the ACM (CACM)*, Vol.38, No.11, pp.39–41 (1995).
10) 中山浩太郎，原　隆浩，西尾章治郎：Wikipedia マイニングによるシソーラス辞書の構築手法，情報処理学会論文誌， Vol.47, No.10, pp.2917–2928 (2006).
11) 中山浩太郎，原　隆浩，西尾章治郎：人工知能研究の新しいフロンティア: Wikipedia，人工知能学会誌， Vol.22, No.5, pp.693–701 (2007).
12) Nakayama, K., Hara, T. and Nishio, S.: A Thesaurus Construction Method from Large Scale Web Dictionaries, *Proceedings of IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pp.932–939 (2007).
13) Ohshima, H. and Tanaka, K.: High-speed Detection of Ontological Knowledge and Bi-directional Lexico-Syntactic Patterns from the Web, *Journal of Software*, Vol.5, No.2, pp.195–205 (2010).
14) 白川真澄，中山浩太郎，荒牧英治，原　隆浩，西尾章治郎：Web 検索を用いた関連のある概念間の関係抽出手法，第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010) (2010).
15) Suchanek, F.M., Kasneci, G. and Weikum, G.: YAGO: A Core of Semantic Knowledge, *Proceedings of International Conference on World Wide Web (WWW)*, pp.697–706 (2007).
16) 数原良彦，戸田浩之，櫻井彰人：ブログ記事を用いた複数話題語間の動作関係抽出手法，電子情報通信学会論文誌 D， Vol.J91-D, No.3, pp.619–627 (2008).
17) Terol, R.M., Martinez-Barco, P. and Palomar, M.: A Knowledge Based Method for the Medical Question Answering Problem, *Computers in Biology and Medicine*, Vol.37, No.10, pp.1511–1521 (2007).
18) Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z. and Ishizuka, M.: Unsupervised Relation Extraction by Mining Wikipedia Texts using Information from the Web, *Proceedings of Annual Meeting on Association for Computational Linguistics, International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pp.1021–1029 (2009).