様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２７　日現在

機関番号：　１４６０３
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０５３

研究課題名（和文）複合表現を考慮した構文解析手法に関する研究

研究課題名（英文）Research on syntactic parsing with multiword expressions

研究代表者
進藤　裕之（Shindo, Hiroyuki）

奈良先端科学技術大学院大学・情報科学研究科・助教

研究者番号：２０７３４７８４

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：自然言語処理における従来の構文解析では，単語間の関係に基づく，いわゆる構成的な意味の解析が中心であり，慣用句や複合語など，複数の単語で一つの意味を表す非構成的な現象の多くは見過ごされてきた．そこで本研究は，これらの「複合表現」を同定し，言語の意味を正しく認識できる構文解析手法と，大規模に収集した複合表現を構文木上に効率良くアノテーションを行う方法を開発した．さらに，構文解析と複合表現を同時に解析することにより，双方の性能向上に寄与することを示す．

研究成果の概要（英文）：In natural language processing, there is a lot of research on syntactic parsing exploiting word-based compositionality, that is, a meaning of phrase or sentence can be computed from the meaning of the words and the way they are combined. On the other hand, non-compositional expressions such as idiomatic constructions has been overlooked until now. In this work, we developed a method for syntactic parsing with identifying multiword expressions, and an efficient way to annotate multiword expressions on syntax trees. Further, we show that joint learning of syntactic parsing and identification of multiword expressions improves the performance of both tasks.

研究分野：　自然言語処理

キーワード：　構文解析　複合表現　複単語表現　アノテーション

１．研究開始当初の背景

最新の日英翻訳システムに“胸を打つ話”と入力すると,“Story that hit the chest”と誤訳される．これは，計算機が“胸を打つ”という慣用句を同定できなかったことが原因である．典型的な機械翻訳システムでは，まず入力文の構文解析を行い，文全体の統語構造（単語間の修飾・非修飾関係や主語・述語関係など）を決定してから，対象言語へ変換する．そのため，構文解析の段階で慣用句を正しく同定できなければ，翻訳結果は誤りとなる可能性が高い．

これまでに申請者らは，言語データから数百万種類の文法パターンを統計的に自動獲得することによって，構文解析の高精度化を実現する手法を発展させてきた．現在でも，ベンチマークデータを用いて構文解析のさらなる高精度化を目指す研究が盛んに行われている．

しかしながら，従来の構文解析（特に英語やヨーロッパ言語）では，慣用句や複合名詞など，複数の単語が一つの意味を表す「複合表現」の存在は見過ごされることが多かった．主な原因として以下の二点が挙げられる．

(1)　包括的な複合表現の辞書や，複合表現の注釈を付与した一定規模のコーパスがほとんど存在しない．一言で複合表現といっても，慣用句，固有表現（人名や地名），複合名詞など様々な種類が存在し，量も膨大である．そのため，首尾一貫した指針に基づく辞書やコーパスの構築には多大なコストを要する．

(2)　辞書やコーパスが整備されたとしても，従来の単語単位の構文解析手法では，あらゆる複合表現を漏れなく同定できない．例えば，英語の慣用表現である“bring ～ up”は,”bring”と”up”が文中で非連続に出現するため，離れた複数の単語を効率的に探索し，真偽を判別する新たな手法が必要となる．

２．研究の目的

上記の背景をもとに，本研究では，文中の複合表現を高精度に同定する手法を確立する．そして，複合表現の同定と，文全体の統語構造とを同時に決定する構文解析手法を開発する．具体的には，以下のことを明らかにするのが目的である．

(1)　複合表現の種類に応じて，その同定に必要な言語資源（辞書，コーパス，知識ベースの種類や入手可能性）を明らかにし，実際にデータを収集する．

(2)　入力文に含まれる複合表現の位置と種類を同定する手法を開発する．特に，従来手法では解決が困難である，”bring ～ up”などの非連続パターンや，派生パターン（“a kind of”,“kinds of”,“a new kind of”）を正しく判別するための統計モデル，パラメータの学習方法，実際の同定アルゴリズムを考案する．

(3)　(2)で開発した手法と,申請者らがこれまでに開発してきた構文解析技術を統合させて，複合表現の同定と，文の統語構造とを同時に決定する構文解析手法を開発する．このことによって，文全体の文脈を考慮しなければ同定することが困難な複合表現を正しく同定できると期待される．具体的には，複合表現と統語構造の同時統計モデル，パラメータの学習方法，実際の構文解析アルゴリズムを考案する．

(4)　(3)の手法を計算機プログラムとして実装し，実際の言語データへ適用する．同時解析によって，文全体の文脈を考慮しなければ判別が困難な複合表現を正しく同定できることを明らかにする．例えば,“break the ice”は,文字通り「氷を割る」という意味と，慣用句として「話の口火を切る」という意味がある．これを正しく判別するためには，“break the ice”の主語や述語，修飾語などを正しく推定する必要がある．そのため，複合表現の同定と，文全体の統語構造とを同時に決定することで，双方の精度向上が期待できる．

３．研究の方法

本研究では，入力文に含まれる複合表現を漏れなく同定する手法を考案し，それを文全体の構文解析と統合することによって，文の非構成的な意味を正しく理解する言語解析プログラムを開発する．

具体的な研究項目は，以下の通りである．

(1)　複合表現の同定に必要な言語資源の調査と収集

あらゆる複合表現を漏れなく同定するために，複合表現の種類に応じて，利用すべき言語資源を明らかにする．申請者らの予備調査によれば，複合表現には少なくとも以下の三種類が存在し，それらの同定に必要な言語資源が異なる．

①“as well as”,“bring up”などの慣用表現は，今後も数量が大きく増加するものではないため，静的な辞書を構築する．現在，申請者らのグループでは,英語の Wiktionary データから，網羅的な慣用表現の辞書を既に構築しており，本研究で利用する．

②“Barack Obama”,“New York”などの固有表現は,Wikipedia や Freebase などの無償で利用できる知識ベースから，網羅的に取得することができる．

③“bus driver”,“olive oil”などの複合語や,“a new kind of”のように修飾語が挿入される複合表現は，次々に造語が生まれるため，これらの複合表現を辞書として固定しておくことは困難である．そのため，大規模な Web や新聞記事のデータからマイニング手法を用いて複合語を自動獲得し，それらを訓練データとして，未知の単語列が複合語であるかを判別する統計モデル，アルゴリズムを考案する．

(2)　入力文に含まれる複合表現を漏れなく同定する手法の考案

特に，複合表現の派生形や非連続形を同定できる手法を考案する．①や③の複合表現は，“a kind of”, “kinds of”, “a new kind of”のような派生形や，“bring ～ up”のような非連続形が存在するため，従来の方法では複合表現を漏れなく同定することが難しい．派生形や非連続形を漏れなく同定するために，複合表現内の単語の挿入や削除をモデル化し，入力文に含まれる複合表現の候補を効率的に列挙するアルゴリズムを考案する．

上記のアルゴリズムで列挙される複合表現の候補は，統計モデルでスコア付けされ，一定のスコア以上の候補を複合表現であると同定することができる．

上記手法は言語に依存しない汎用的なものであるが，まずは英語を対象とした評価実験を行う．英語の評価データを用いて同定の精度と再現率を評価し，単純なベースライン手法と比較して性能が向上することを確認する．単純なベースライン手法として，従来の固有表現抽出の研究で用いられる，派生形や非連続形を全く考慮しない手法を用いる．

(3)　複合語の同定と，文の統語解析を同時に行う手法の開発

複合表現の同定と，文全体の統語構造を同時に決定することで，文全体の文脈を考慮しなければ判別が困難な複合表現を正しく同定できることを明らかにする．具体的には，入力文に含まれる複合表現の全候補を，コンパクトなグラフ構造（ラティス）として保存し，必要に応じてそれらを展開しながら構文解析を行う効率的な手法を考案する．

４．研究成果

複単語表現のコーパス構築に関しては，機能表現と動詞に関して実施した．具体的には，大規模かつ網羅的な複単語表現をWiktionaryから収集し，それをOntoNotesコーパス（構文木のアノテーションが含まれるテキストコーパス）にアノテーションを行った．このとき，全てのアノテーションを人手で行うことは多大なコストを要するため，構文木の情報（単語の依存関係）に基づいて，複単語表現の辞書にマッチした句の候補に対して，確実に複単語表現であるもの，確実に複単語表現ではないもののフィルタリングを行った．フィルタリングのためのルールは論文で公開している．そして，残りの候補に対して，人手で複単語表現であるかどうかを判別するということを行い，アノテーションのコストを大幅に軽減することを実現した．機能語，動詞の双方において，このアプローチが適用可能であることがわかった．

複単語表現と構文解析の同時解析手法については，いくつかのバリエーションを考案し，性能を評価した．一つは，複単語表現の予測をあらかじめCRFで行い，その情報を構文解析の特徴量として導入する手法である．この場合，複単語表現の解析誤りがあったとしても，構文解析は影響を受けにくいことが期待される．もう一つは，複単語表現の情報を単語の依存構造へ変換して，全体を従来の依存構造解析と同じ手法として解く方法である．この場合，複単語表現は，先頭の単語を主辞とする依存構造とすることで，既存の依存構造解析を使用できるという利点があるが，複単語表現が一つの意味を持つ句であるという情報を上手く利用できないという欠点もある．評価実験の結果，複単語表現を系列ラベリング（ＣＲＦ）として事前に予測し，その結果を特徴量として構文解析を行う方法が最も性能が高いことがわかった．したがって，少なくとも我々の実験の範囲内では，複単語表現は１つのトークンとして扱う方が構文解析の観点からは望ましいといえる．また，同様の手法を用いて，固有表現と構文解析の同時解析も行った．この場合，固有表現の範囲と，構文木の句の範囲は整合していない（固有表現の範囲が構文木の句になっていない）というケースがある．そのため，固有表現の範囲と構文木の句の範囲を合致させるルールを考案し，固有表現，複単語表現，構文木の全ての情報を含むコーパスを構築した．また，本研究で構築した言語資源は，LDC（https://www.ldc.upenn.edu/）で公開した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

①　Learning Distributed Representations of Texts and Entities from Knowledge Base
Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji
Transactions of the Association for Computational Linguistics, Vol 5, 2017, pp. 397-411
https://transacl.org/ojs/index.php/tacl/article/view/1065/257
査読あり

②　Transition-Based Dependency Parsing Exploiting Supertags
Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto
IEEE Transactions on Audio, Speech and Language Processing, Volume 24, Issue 11, 2016, pp. 2059-2068
DOI: 10.1109/TASLP.2016.2598310
査読あり

〔学会発表〕（計　4 件）

①　English Multiword Expression-aware Dependency Parsing including Named Entities
Akihiko Kato, Hiroyuki Shindo and Yuji Matsumoto
association for computational linguistics,

2017
査読あり

② Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation
Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda and Yoshiyasu Takefuji
The SIGNLL Conference on Computational Natural Language Learning, 2016
査読あり

③ Construction of an English Dependency Corpus incorporating Compound Function Words
Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto
Language Resources and Evaluation Conference, 2016
査読あり

④ An Efficient Annotation for Phrasal Verbs using Dependency Information
Masayuki Komai, Hiroyuki Shindo and Yuji Matsumoto
Pacific Asia Conference on Language, Information and Computation, 2015
査読あり

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

○取得状況（計　0 件）

６．研究組織
(1)研究代表者
進藤　裕之（SHINDO, Hiroyuki）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：20734784

(2)研究分担者
該当無し

(3)連携研究者
該当無し

(4)研究協力者
該当無し