依存構造情報を用いた柔軟な複単語表現の同定

森元 彩華　吉本 暁文　加藤 明彦　進藤 裕之　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科

{morimoto.ayaka.lw1, akifumi-y, kato.akihiko.ju6, shindo, matsu}@is.naist.jp

1 はじめに

複単語表現 (Multiword Expression, 以降 MWE) は，「単語境界 (またはスペース) を越えて特異な解釈を持つ表現」[7] であり MWE の認識は各種の言語解析で重要である．MWE は以下のカテゴリに分類される．

- Lexicalized phrases
  - 固定表現：完全に決められた語順や形のもの (例： *with respect to*)．
  - 半固定表現: 一部の単語が活用形等の語形変化するもの (例：*keep up with, kept up with*)．
  - 構文的に柔軟な表現: 構文の多様性の幅が広いもの (例えば MWE の中には，その内部の単語に修飾語を持つことが出来る)．
- Institutionalized phrases
  - 意味的に固有のフレーズ (例： *traffic light*)．

本研究では，他のタイプの MWE に比べてこれまで調査の少なかった英語の構文的に柔軟な MWE に焦点を当てる．MWE には，副詞・限定詞・前置詞・従属接続詞等の機能的な表現として文法的に振る舞うものが多くある．本研究では，これらの機能表現以外に，後述する駒井ら [4] により辞書構築されている句動詞以外の複単語動詞を考慮に入れる．複単語名詞については，多くが構文的に名詞句として振る舞うことから，構文的には問題ではないため考慮しない．MWE は特定の文法的機能を持ち，辞書拡張において重要であると考えられる．

本研究の目的は，構文的に柔軟な MWE を幅広く網羅的に構築し，依存構造内における MWE の構造を修飾語も含めて記述し，OntoNotes コーパス [5] の一部である Wall Street Journal において出現する MWE のアノテートを行うことである．

英語の MWE 辞書を構築する研究がこれまでにいくつか行われている．重藤ら [10]，駒井ら [4] は英語の固定的 MWE と句動詞の辞書を提示している．また彼らは， Penn Treebank において出現する固定的 MWE と句動詞の表現にアノテーションを行っている．人間が使用する多くの英語辞書には，MWE や熟語や慣用句が多数含まれている．しかし，NLP タスクに使用可能な英語の柔軟な MWE の包括的な辞書の構築はなされていない．

本研究での主な貢献は，MWE をリストアップしている複数の Web サイトを参照して，英語の構文的に柔軟な MWE 辞書を作成した点と，OntoNotes コーパスに含まれる MWE のアノテートを行い，表現内に出現する可能性のある修飾語を特定した点である．

我々の現在の研究では，OntoNotes において柔軟な使用方法の形式が出現しない表現が，現実世界においてどれだけ柔軟な MWE であるのかわからないという問題がある．よって，最初の貢献については現在も進行中である．しかし，2 つ目の貢献については，OntoNotes で柔軟な MWE のすべての出現にアノテート可能となるように設計している．以降の章では，まず関連研究について説明し，英語の柔軟な MWE をどのように収集したか，OntoNotes において出現する MWE に対してどのようにアノテートを行ったかを説明する．

2 関連研究

MEW コーパスのアノテーションは，英語では大規模には行われていない．一方フランス語では MWE の大規模なアノテーションコーパス [1] があり，これには 30,000 の MWE がアノテートされた 18,000 文が含まれている．英語では Schneider ら [9] が English Social Web コーパスに MWE をアノテートしたコーパスを作成した．このコーパスは構文的に柔軟な英語の MWE アノテーションをもつ唯一のコーパスであるが，3,800 文と規模は小さい．

英語における固定的 MWE や半固定的な MWE には，辞書の構築や大規模なコーパスに対するアノテートに関する研究がいくつかある．重藤ら [10] や加藤ら [2] は，OntoNotes の Wall Street Jounal に固定機能をもつ MWE をアノテートした．コーパスのサイズは 37,000 文であり，アノテーション付きの MWE の数は 6,900 である．前者の研究では英語での固定的な MWE の辞書を構築し，コーパス内のすべての MWE の範囲

にアノテートを行った．後者は，MWE の機能性に応じて文の依存構造のアノテートを修正している．MWE の特定のタイプである句動詞については，同じコーパスに駒井ら [4] によってアノテートされた．この研究では，37,000 文の中に 22,600 回出現する句動詞にアノテートを行っている．また，多言語の解析と言語リソースにおけるマルチテキスト表現の問題に特化した PARSEME (PARSing and Multi-word Expressions) プロジェクト[1]がある．このプロジェクトの包括的な説明は，Savary ら [8] が示している．加えて，Treebanks における MWE の詳細な調査は Rose ら [6] が行っている．

3 English Flexible MWE の収集

英語の柔軟な MWE の候補を収集するために，我々は英語学習サイトである ALL IN ONE[2] と Weblio[3]の英単語辞典の索引を調査して収集を行った．両サイトは，辞書・例・有用な表現等，英語学習者にとって有益な情報を提供している．さらに，English Wiktionary[4]の品詞が動詞・形容詞・副詞・前置詞・接続詞であるエントリーも候補として収集を行った．

結果として，上記のサイトから 16,339 の MWE 候補を収集した．次に Web 上での 1 から 5gram の全出現頻度がまとめられている Web 1T 5-gram(LDC2006T13)[5]を使用して，頻度により候補の足切りを行った．閾値を設け，上位 3,000 件の表現を得た．次に，以前の研究によって構築された辞書に基づいて，固定的な MWE または句動詞として既に知られている MWE を削除した．その結果，2,928 個の表現が最終的な候補として得られた．この段階では，それらが実際に柔軟な MWE であるかどうかは分かっていない．更に，候補となる MWE が柔軟な MWE である事が既知であっても，それがどれほど柔軟であるのか，それがどのような変化を持つのかは未知である．

4 依存構造による MWE の同定とアノテーション

4.1 概要と目標

本研究の目的の 1 つは，柔軟性の程度の情報を用いて英語の柔軟な MWE の辞書を構築することである．本研究では表現内の変更に関する柔軟性にのみ焦点を当てる．MWE の例として，“a number of” は，“a growing number of” または “a very large number of” のように使用される．これらの出現と統語的な一貫性を見て，“a number of” が表現の中にある “number” を修飾する単語を含むことが可能であると推測することが出来る．我々は，候補の MWE の正しい構文構造を知るために，OntoNotes Release 5.0 (LDC2013T19) を，Satanford dependency[6]に基づき，全ての句構造木を依存構造木に変換した．句構造木ではなく依存構造木を使用した理由は，Penn Treebank における句構造が統一されていないためである．同じ MWE またはそれらを含む句において，異なる句構造または句の名前がアノテートされている．それらを依存構造木に変換することで構造が統一される．

もう 1 つの目的は，OntoNotes の MWE の全ての出現に固定的または柔軟な形式の MWE としてアノテーションを付けることである．MWE の完全一致の出現のみだけではなく，MWE の単語の間に単語が入り得る場合に対してアノテートを行った．このとき，依存情報を用いて MWE の出現を半自動的にアノテートした．いくつかの同じ形式の MWE でも字句通りの使用が可能である場合がある．したがって，公開する前に全てのアノテーション結果について手動でチェックを行う．以下の節では MWE 候補の半自動アノテート方法について説明する．

4.2 MWE を含む最小の依存構造の抽出

この節では候補となる MWE を含む依存構造を抽出する方法について説明する．句構造木を Stanford dependency 形式の依存構造木に変換した OntoNotes の Wall Street Journal を使用した．以下の段階を踏み，MWE を含む依存構造の抽出を行う．

1. MWE の候補毎に，MWE を含む OntoNotes のすべての文を抽出する．例えば，“number” の場合は “a”, “ number”, “of” の順序で全てを含む文を抽出する．この処理では，MWE を含む可能性のあるすべての文を取得する．
2. 全文を Stanford 依存構造木に変換する [3] [7]．
3. 各文に対して，MWE 内の全ての単語を含む最小の依存構造木を抽出する．抽出された部分木の例を図 1 に示す．
4. 部分木内には，MWE を構成しない単語や部分木が含まれる場合がある．図 1 の “division heads” から成る部分木がその例である．このような場合，部分木のルートだけを残し，その他の全ての子ノードを削除する．次に，MWE を構成しない

[1] http://typo.uni-konstanz.de/parseme/
[2] http://www.allinone-english.com/A13E/phrases-table-A-K.html, /phrases-table-L-Z.html
[3] http://ejje.weblio.jp/cat/dictionary/eidhg
[4] https://en.wiktionary.org/wiki/Wiktionary:Main Page
[5] https://catalog.ldc.upenn.edu/ldc2006t13
[6] http://nlp.stanford.edu/software/stanford-dependencies.shtml
[7] 変換コマンドのオプション：“-conllx -basic -makeCopulaHead - keepPunct ”

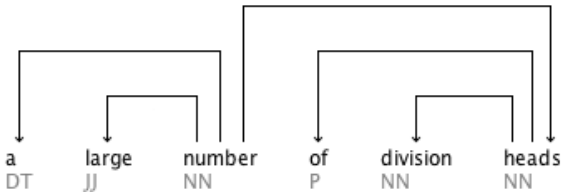

図 1: “a number of” を含む最小部分木

すべての単語を品詞ラベルに置き換える．図 1 の例では，MWE の柔軟な使用法は “*a JJ number of NN*” となる．これを図 2 に示す．

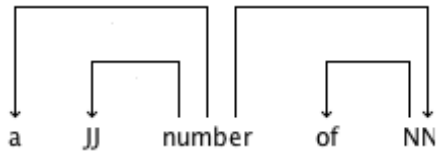

図 2: “a number of’ の柔軟な使用法の表現

図 3, 4, 5 は “a number of” の異なる形式の出現結果を示した図である．

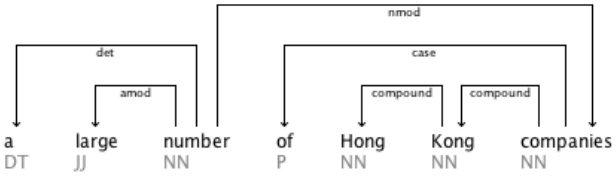

図 3: “a large number of ...” における最小部分木

上記の手順の後，表現内の修飾語の存在の違いを表すための同形木を得る．例えば図 3 の木は JJ を “number” の修飾語として含む．これらの木から図 6 に示す依存構造木を柔軟な MWE として取得することが出来る．図中の*1 は今回の例の場合では JJ または空要素として定義されるが，最終的には {ϵ, JJ, NN, VBG, VBN} と定義されるワイルドカードであり，*2 は {NN, SYM} のような品詞タグを持つ単語が対応する．

各候補の MWE について，上記の手順を実行して部分木を取得する．部分木のいくつかは，固定的な MWE とされる形式である可能性もある．得られた部分木において MWE 内の単語のみから構成されるものを，固定的な MWE の依存構造とみなす．MWE 以外の単語が含まれるような依存構造木を固定的な MWE の依存構造木と比較し，MEW 内の単語における構造が同形である場合は，それらを柔軟な MWE とみなす．

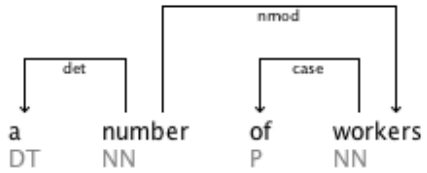

図 4: ”a number of workers ...” における最小部分木

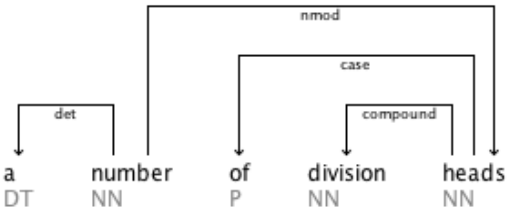

図 5: ”a number of division heads ...” における最小部分木

固定的な MWE と同形でない依存構造木を持つ使用例の場合は負例とみなす．このとき負例とは，MWE 内の単語は全て依存構造木に含むが，MWE として使用されていないものを指している．

2,928 個の MWE リストに対して OntoNotes コーパスにおける 37,015 文で検索を行った結果，出現した MWE の種類は 1,817 個であった．得られた部分木に関して文全体の木と異なるものは 26,358 個，ユニークな部分木は 14,146 個であった．結果を表 1 にまとめる．

4.3 MWE の自動アノテーション

本章では MWE のアノテートにより得られた結果について示す．使用した OntoNotes コーパスの Wall Street Journal データにおける MWE の出現数と，各 MWE のユニークな部分木の数から柔軟な MWE であるかどうかを判定する．得られた結果について分析するに当たり，出現回数が 1 回となる MWE は分析が不可能なため除外している．結果として，固定的な用法がされている固定的な MWE として 1,194 件，柔軟な MWE として 1,704 件がアノテート結果として得られた．このとき 11,248 件は負例であった．表 2 に示す．

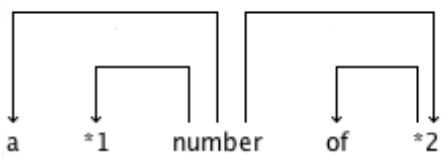

図 6: “a number of” の柔軟な MWE の表現

表 1: OntoNotes から抽出した MWE と依存構造木数

MWE の種類	1,871
得られた部分木	26,358
得られた固有な部分木	14,146

表 2: 固定的な MWE と柔軟な MWE の結果数

label	count
固定的な MWE	1,194
柔軟な MWE	1,704
負例	11,248

5 おわりに

本研究では，柔軟な MWE の辞書構築と OntoNotes へのアノテーションについて述べた．柔軟な MWE を抽出するにあたり，依存構造木における MWE を含む最小部分木の一致から半自動的に判定を行った．

今後は，辞書構築におけるカバレッジ性の向上のために Gigaword を用いて，そこで出現する MWE のアノテーションと柔軟な MWE の同定を行う．また，アノテーションの有無において，構文解析の精度が変化するかどうかの確認を行う．更に，抽出は行ったが人手での判定が必要となることから判定を保留としている柔軟であるかもしれない MWE について，人手でのアノテーションを行い辞書の拡張を行う．加えて，今回まとめた MWE に対して品詞の付与を行い，こられらの MWE が出現する依存構造木に対して加藤ら [11] のように木構造の修正を行う．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[1] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for french. In *Treebanks*, pp. 165–187. Springer, 2003.

[2] Kato Akihiko, Shindo Hiroyuki, and Matsumoto Yuji. Construction of an english dependency corpus incorpo- rating compound function words. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1667–1671, 2016.

[3] Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Association for Computational Linguistics, 2008.

[4] Masayuki Komai, Hiroyuki Shindo, and Yuji Matsumoto. An efficient annotation for phrasal verbs using dependency information. *In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC), Posters*, pp. 125–131, 2015.

[5] Sameer S Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, Vol. 1, No. 04, pp. 405–419, 2007.

[6] Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. Mwes in treebanks: From survey to guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[7] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 1–15. Springer, 2002.

[8] Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, et al. Parseme–parsing and multiword expressions within a european multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, 2015.

[9] Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T Mordowanec, Henrietta Conrad, and Noah A Smith. Comprehensive annotation of multiword expressions in a social web corpus. *In Proceedings of LREC-2014*, pp. 455–461, 2014.

[10] Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. Construction of english mwe dictionary and its application to pos tagging. In *Proc. of the 9th Workshop on Multiword Expressions*, pp. 139–144, 2013.

[11] 加藤明彦, 進藤裕之, 松本裕治. 固有表現と複合機能語を考慮した MWE ベースの依存構造コーパス構築と解析. 言語処理学会第 23 回年次大会 (NLP2017), 2017.