英語の動詞系複単語表現コーパスの構築

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
{kato.akihiko.ju6, shindo, matsu}@is.naist.jp

1 はじめに

複単語表現（MWE）とは，統語構造あるいは意味構造上の単位として取り扱う必要のある複数の単語のまとまりである．MWE には複合名詞（例：**green signal**），複合機能語（例：**a lot of**）を始めとして様々なものが含まれる．

依存構造解析と MWE 認識は別々の NLP タスクとして扱う事もできるが，意味解析などの下流タスクにとっては，単語ベースの依存構造よりも，MWE を単一のノードとする依存構造の方が好ましいと考えられる．実際，英語やフランス語において，MWE を考慮した依存構造解析が行われている [4, 6, 10]．

本稿では，各種の MWE のなかでも特に，英語の動詞系 MWE（VMWE）に着目する（ 表 1）．VMWE の正確な認識は，意味解析などの下流タスクで重要である．たとえば句動詞は，動詞と particle が組み合わさることで，動詞の元の語義とは異なる語義を持ちうる（例：**put .. off**）．このため，意味的な非構成性を有する VMWE の認識は，正しく意味解析を行う上で必須となる．また，VMWE は出現が一般に非連続であったり，活用変化を伴うため（例：**take .. off, took .. off**），VMWE の認識には，固有表現などの連続 MWE の認識とは異なるアプローチが求められる．これらの理由により，VMWE アノテーションを施した言語資源の開発は重要である．

既存の VMWE コーパスとしては，English Web Treebank [3] 上に MWE を注釈した Schneider らのコーパス [12] が挙げられる．しかしながら，彼らのコーパスは VMWE の出現数（1,444 回）と種類数（1,155 種）の点で比較的小規模であり，VMWE 認識モデルの訓練データとしては十分でないという問題点がある．

そこで本稿では英語の Ontonotes コーパス [11] の Wall Street Journal（WSJ）部分全体をカバーする大規模な VMWE アノテーションに取り組む．その際，アノテーションを効率的に行うために，VMWE の統語的な性質を利用して，コーパス中の VMWE の候補に対してフィルタリングを行う（ 2 章）．また，フィルタリングを通過する候補数も 1,000 事例を超える規模であるため，VMWE アノテーションを，MWE の語義曖昧性解消タスクとして定式化し，クラウドソーシングを利用する（ 2 章）．

カテゴリ	例
Verb-particle constructions	pick up, take over
Prepositional verbs	look for, base on
Light verb constructions	make a decision
Verb-noun(-preposition)	take care (of)
Semi-fixed VMWEs	make one's way

表 1: 英語の動詞系複単語表現（VMWE）の主要カテゴリの一覧

本稿で実施した VMWE アノテーションの概要を以下に示す．各ステップの詳細は 2 章を参照されたい．

1. 英語の Wiktionary [1] を用いた VMWE の辞書構築
2. Ontonotes 上での VMWE の出現箇所の候補抽出
3. Gold の依存構造木を用いたフィルタリング
4. クラウドソーシングによるアノテーションの収集

上述の手順で得られた VMWE アノテーション（出現数：7,833，種類数：1,608 種）には，句動詞，軽動詞構文（Light verb constructions），Semi-fixed VMWE 等，各種の VMWE が含まれる．構築した VMWE 辞書およびコーパスアノテーションは `https://github.com/naist-cl-parsing/verbal_mwe_annotations` で公開を予定しており，英語 VMWE の認識モデルや，英語 MWE を考慮した依存構造解析モデルの研究に利用することができる．

[1] https://en.wiktionary.org

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

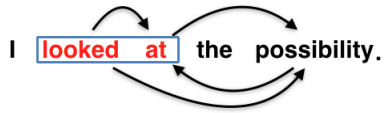


図 1: 機能語ヘッド（上）および内容語ヘッド（下）の依存構造木．両方の木に共通するエッジは省略している．図中の枠は VMWE（**look at**）を示す．VMWE の候補を依存構造の部分木として抽出するためには，機能語ヘッドの方が内容語ヘッドよりも適している．

2 動詞系 MWE コーパスの構築

2.1 辞書構築と候補抽出

本稿ではまず英語の Wiktionary から 2 語以上からなる動詞を抽出することによって，VMWE の辞書を構築した [2]．ただし（1）Be 動詞と動詞でない単語からなる MWE（例：**be above, be with**），（2）助動詞は取り除いた．この結果，8,369 種の VMWE が得られた．次に Ontonotes 5.0 （LDC2013T19）の WSJ 部分（37,015 文）において，構築した辞書を用いて VMWE の出現箇所の候補を抽出した．候補抽出処理にあたっては，**take .. off** の様なギャップありの出現，動詞の活用，Semi-fixed MWE における可変部分（例： **someone, something, one's, oneself**）を考慮した．また，Ontonotes で与えられている Gold の品詞情報を用いて，動詞を含まない候補は除外した．ギャップありの出現に関しては，ギャップ中に他の動詞または句読点を含む候補を除去した．

本稿ではさらに，多くの VMWE が統語的に正則である事を考慮し，上記で得られた候補に対して，Ontonotes で与えられている Gold の構文情報を用いたフィルタリングを行った．具体的には Gold の句構造木を Stanford Basic Dependencies [8] に変換し，VMWE の構成単語のみで部分木をなす候補を抽出した．Stanford Basic Dependencies は機能語ヘッドであるため，Prepositional verb の抽出に適している．Prepositional verb は品詞タグの系列として "V IN" を有するため，その後には多くの場合，名詞句が来る．この場合，内容語ヘッドの dependency scheme（例：Universal Dependencies [9]）では MWE 中の動詞は名詞句のヘッドを子として持つため，VMWE の構成単語のみからなる部分木は得られない（図 1 下側）．一方，機能語ヘッドの場合には，MWE 中の動詞が直後の前置詞を子として持つため，部分木が得られる（図 1 上側）．

[2] 2 語以上からなるエントリの内，カテゴリとして"English_verbs" を有するものを選択した．

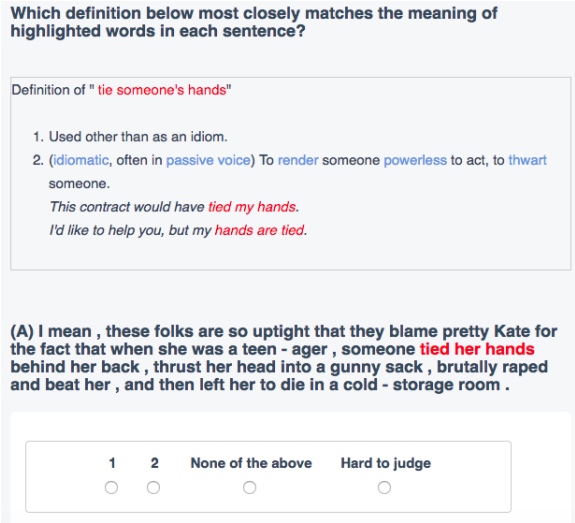


図 2: クラウドソーシングに用いた Web インタフェースのスクリーンショット．

句動詞については，Komai ら [7] が Ontonotes 上の句動詞の一部についてアノテーションを行っているため，Komai らの構築した辞書でカバーされている VMWE については彼らのコーパスアノテーションを採用した．一方，Komai らの辞書でカバーされていない句動詞については，上述の手順で得られた VMWE の各候補について以下の手順を行った．

1. 句動詞を Verb-particle construction（VPC） と Prepositional verb [2] に分類する．
2. 動詞から particle への dependency edge のラベルに応じて以下の処理を行う [7]．VPC については係り受けラベルが "prt" の場合にのみ正例とする．Prepositional verb については，係り受けラベルが "prep" で，かつ動詞と particle が連続している場合に正例とする．
3. 上記以外のケースについてはクラウドソーシングによるアノテーションを行う．

2.2 クラウドソーシングによる VMWE アノテーション

上記手順の結果，人手注釈が必要な VMWE の候補として 2,135 事例が得られた．これらの事例群に対してクラウドソーシングを行うために，VMWE アノテーションを，MWE の語義曖昧性解消タスクとして定式化する．実際にクラウドソーシングに用いた Web インタフェースを 図 2 に示す．アノテーターには，Wiktionary から抽出した VMWE の語義一覧と，

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

VMWE の候補の構成単語が強調表示された文が提示される．各 VMWE は 1 つの literal sense と，一般に複数の non-literal sense に相当する一連の定義文を持つ．ただし，Wiktionary で literal sense に相当する定義文が記載されていない場合，“Used other than as an idiom” という選択肢を補った．アノテーターはこれらの情報に基づき，文中の強調表示された単語群の意味として，どの語義が最も適合するかを選択する．ただし提示された語義のいずれにも適合しない場合は “None of the above” と回答する事ができる．また，どの語義に適合するか判断が難しい場合には “Hard to judge” と回答する事ができる．

クラウドソーシングは CrowdFlower [3] を用いて行い，アノテーターには以下の要件を課した．(1) CrowdFlower で *Level 3* contributor に属している [4]，(2) 英語を公用語とする国に居住する，(3) 著者が正解を作成したテスト問題において 70 %以上の正解率を達成する．

アノテーションの効率を向上させるために，アノテーターには同一の VMWE に関する複数の事例（最大 5 事例）を提示した．VMWE の各事例について，3 名のアノテーターからアノテーションを収集し，総費用は$1,016 USD となった．

本稿では収集したアノテーションに基づき，以下の手順を用いて各候補が VMWE の正例かどうかを決定した．

1. 3 名全員が同一の語義を選択した場合（67.1 %），その語義が literal sense であれば負例，それ以外であれば正例とする．
2. 3 名のいずれも literal sense を選択しなかった場合（9.0 %），正例とする．
3. 上記のいずれにも該当しない場合（23.8 %），当該の文において VMWE の候補に最も適合する語義を，著者らが人手で選択し，その語義が literal sense であれば負例，それ以外であれば正例とする．

2.3 包含と重複の解決

上記手順で得られた VMWE の出現箇所について，Komai ら [7] のアノテーションとの間で包含あるいは部分重複の関係にあるものを調査した結果，159 事例が包含関係，40 事例が部分重複の関係にある事が分かった．まず包含関係にある事例群については，より広い方のスパンを採用した．たとえば “come at” と “come at a price” に相当する 2 つのスパンが包含関係にある場合，後者のみを残した．次に部分重複の関係にある事例群については，Cambridge Dictionary [5]，The Free Dictionary [6] の双方に記載されている新たな VMWE が得られる場合，2 つの VMWE のスパンのマージを行った．たとえば，“take the reins” と “take over” の出現をマージする事によって “take over the reins” という新たな VMWE の出現が得られる．また，上記 40 事例の内，一部の事例でアノテーションミスに由来する擬似的な部分重複が見られたため，これらを修正した．この結果，部分重複に相当する事例数は最終的に 11 事例となった [7]．

構成単語数	2	3	4	≥ 5	合計
VMWE の出現数	7,067	597	138	31	7,833
VMWE の種類数	1,235	270	80	23	1,608

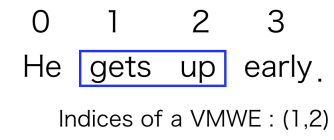
表 2: 構成単語数別に示した VMWE の出現数および種類数．

ギャップ数	0	1	2	合計
VMWE の出現数	6,855	968	10	7,833

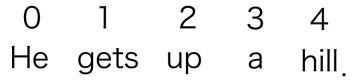
表 3: ギャップ数別に示した VMWE の出現数および種類数．

3 構築したリソースについて

本アノテーションの結果，Ontonotes の WSJ 部分において 1,608 種／ 7,833 回の VMWE の出現を注釈す



(a) 正例（non-literal usage）



(b) 負例（literal usage）

図 3: VMWE（**get up**）の正例と負例の模式図．

[3] https://www.crowdflower.com

[4] CrowdFlower では “the smallest group of most experienced, highest accuracy contributors” と規定されている．

[5] http://dictionary.cambridge.org

[6] http://idioms.thefreedictionary.com

[7] これらは真に複数の VMWE の部分重複となっており，たとえば以下の文で “look back” と “look .. on .. as” のスパンは部分重複している．“He may be able to **look back on** this election **as** the high-water mark of far-left opposition.”

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.

POS pattern	連続	非連続（ギャップあり）	VMWE の例と頻度
V_IN	3,071	260	base on : 142 look for : 86 focus on : 77 go to : 70
V_RP	2,081	229	set up : 62 take over : 49 point out : 47 turn out : 43
V_RB	547	116	go back : 17 come back : 17 do well : 15 go down : 13
V_NN	280	167	take place : 41 do business : 27 take effect : 26
V_DT_NN	114	45	take a look : 13 make a decision : 8 pave the way : 5

表 4: VMWE の構成単語の品詞タグの系列と出現数の関係（出現数の上位 5 パターンのみ記載）

ることができた．得られた VMWE の，構成単語数およびギャップ数に関するヒストグラムをそれぞれ 表 2, 表 3 に示す．また，VMWE の構成単語の品詞タグの系列を出現数の降順にソートしたものを 表 4 に示す．この表から分かる様に，本アノテーションには句動詞，軽動詞構文等，各種の VMWE が含まれる．また，複数の品詞タグ系列で非連続な出現が存在する．

本コーパスアノテーションの成果物は，VMWE の正例（non-literal usage）における，構成単語群の文中の位置（token index）の系列である．これを用いると， 図 3 のように，コーパス中の VMWE の候補群において，正例と負例を区別することができる．

4 関連研究

MWE アノテーションを有するコーパスとしては，まず French Treebank [1] が挙げられる．French Treebank はフランス語の句構造コーパスであり，フランス語 MWE を考慮した依存構造解析 [4] のデータセットとしてしばしば用いられている．

また，英語 MWE については重藤ら [13] が，英語の Wiktionary から複合機能語 [8] を抽出し，Ontonotes 上での複合機能語のアノテーションを行っている．加藤ら [5, 6] は，MWE を部分木としてまとめ上げる事によって，重藤らの複合機能語アノテーションと Ontonotes の固有表現アノテーションを句構造木に統合している．彼らはこの手順で得たデータセット（LDC2017T16）を用いて，英語 MWE を単一のノードとする依存構造解析の実験を行っている．

5 おわりに

本稿では，英語の動詞系 MWE の，Ontonotes コーパスの WSJ 部分全体への大規模アノテーションに取り組んだ．この結果，句動詞，軽動詞構文，Semi-fixed VMWE 等，各種の VMWE を含むアノテーション（出現数 : 7,833，種類数 : 1,608 種）を得ることができた．

[8] MWE が全体として前置詞，接続詞，限定詞，代名詞，副詞のいずれかの働きをするものを指す．

参考文献

[1] Anne Abeillé, Lionel Clément, and François Toussenel. *Building a Treebank for French*. Springer Netherlands, 2003.

[2] Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2), 2009.

[3] Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. English web treebank. *Technical Report LDC2012T13, Linguistic Data Consortium*, 2012.

[4] Marie Candito and Matthieu Constant. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of ACL*, 2014.

[5] Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. Construction of an english dependency corpus incorporating compound function words. In *Proceedings of LREC*, 2016.

[6] Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. English multiword expression-aware dependency parsing including named entities. In *Proceedings of ACL*, 2017.

[7] Masayuki Komai, Hiroyuki Shindo, and Yuji Matsumoto. An efficient annotation for phrasal verbs using dependency information. In *Proceedings of PACLIC*, 2015.

[8] Marie-Catherine Marneffe and D. Christopher Manning. The stanford typed dependencies representation. In *Coling : Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 2008.

[9] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, 2013.

[10] Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. Joint dependency parsing and multiword expression tokenization. In *Proceedings of ACL and IJCNLP*, 2015.

[11] Sameer S. Pradhan, Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. Ontonotes: A unified relational semantic representation. In *Proceedings of ICSC*, 2007.

[12] 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.

[13] Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. *Proceedings of the 9th Workshop on Multiword Expressions*, chapter Construction of English MWE Dictionary and its Application to POS Tagging. 2013.

Copyright(C) 2018 The Association for Natural Language Processing.
All Rights Reserved.