英語の句動詞表現の同定とコーパス構築

駒井 雅之　　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
{komai.masayuki.jy4, shindo, matsu}@is.naist.jp

1 はじめに

Multiword expressions(MWEs) は2語以上から成る特異な解釈を要する表現として定義される [4]. MWEs の語彙と単一語からなる語彙は同程度の大きさとされ, 深い言語解析のためには MWEs の理解が不可欠である. しかし, 人手注釈の際に発生する膨大なコストのために, MWEs の情報が十分に注釈された大規模コーパスは, あまり存在しないのが現状である. 英語の MWEs は, 分離可能なものと, そうでないものとに大別され, 重藤ら [7] は fixed expressions と呼ばれる分離不可能な表現を Penn Treebank[1] に注釈したが, 分離可能な MWEs に関する注釈は十分に成されていない. そこで, 我々は分離可能な MWEs において代表的なものである句動詞 (Phrasal Verbs) に関して, Penn Treebank 上に効率よく注釈する方法を提案する. 句動詞は, 図 1 に示すように動詞と一つ以上のパーティクルから構成される表現である. 提案手法は, Penn Treebank に含まれる句構造情報を利用し, 機械的に判別可能な MWEs と, 人手による注釈が必要な MWEs とに分類する規則を構築し, Penn Treebank に注釈を行うものである. 提案法により, 人手注釈のコストを抑えつつも大規模なコーパス構築を可能にした.

また, 注釈の付与された MWEs を同定するタスクにおいては, IOB スタイルのタグを割り当てるチャンキングの問題として解いている手法が多い. チャンキングの問題として解くことで, 学習データに存在しない MWEs が検出できるという利点がある一方, 系列ラベリングの手法を利用しているため, 非連続な MWEs の解析には適しておらず, かつ高精度な検出は難しい. そこで, 系列ラベリングの問題としてではなく, 2 値分類問題として解くことで MWEs の検出の精度向上を図った. 分類問題として解く提案手法は, 規則に基づく手法と Schneider らの手法を, 適合率 · 再現率共に上回った.

[1] http://www.cis.upenn.edu/~treebank/

(a) We bring our computers up.

(b) She goes over the question.

(c) Someone goes over there.

図 1: 分離可能な bring up に関する正例 (a) と
分離不可能な go over に関する正例 (b) と負例 (c) の例

2 関連研究

MWEs のリソースに関してであるが, 重藤ら [7] は必ず連接する fixed expressions と呼ばれる表現を Penn Treebank に対して注釈した. その際, Wiktionary から抽出された fixed expressions の候補一覧と Penn Treebank が持つ句構造情報を利用し, 効率的に注釈を行う方法を提案している. また, Schneider ら [6] は English Web Treebank を対象に全ての MWEs を注釈した.

また, MWEs の同定タスクにおいて, 重藤ら [7] は彼らが注釈したコーパスを用いて, MWEs を条件付き確率場を使って検出する手法について述べている. Constant ら [2] は MWEs の事前検出や, 構文木の Reranking 時に MWEs を素性として導入することで, フランス語で書かれた文の構文解析精度の向上を実証したが, 事前検出には Ramshaw ら [3] が提案した入力系列に IOB スタイルのタグを割り当てるチャンキング問題として定式化し事前検出を行っている. この 2 つの手法の共通点として, 不連続の MWEs を検知することは困難であるという点があげられる.

分離可能な MWEs の同定を行う手法として, Schneider ら [5] は Constant らと同様に, 対象の入力系列に IOB スタイルのタグを割り当てることで英語 MWEs の検出を行っている. 彼らはタグの種類を {O o B b Ī ī Ĩ ĩ} と拡張し, Strong MWEs と Week MWEs の区別及び MWEs の構成単語間に語が含まれることにより, 不連続となった MWEs の検出を可能にした. この構成単語間の語のことを Gaps と呼ぶ. しかし,

表 1: 句動詞候補数について

	句動詞数
Wiktionary	793
他サイト	313
計	1106

IOB スタイルのタグ割り当てによる MWEs の検出は, 学習データに存在しない MWEs を検出できる可能性を生むというメリットがあるものの, 依然としてあまりに離れすぎている MWEs を検出することは困難であり, かつ高い精度での検出は難しい.

MWEs のリソースにおいて, 分離可能な表現が注釈された大規模なコーパスはほとんど存在しないため, 本研究では, Penn Treebank の句構造情報を利用して, 効率的に句動詞の注釈を行う方法を提案する. 3 章では提案手法を含めたコーパス構築法について述べ, 4 章では構築したコーパスを用いた句動詞の同定実験について述べる.

3 句動詞のコーパス構築

コーパス構築の手順は以下の通りである.

(1) 句動詞の候補収集

(2) Penn Treebank から事例獲得

(3) 各事例への注釈

3.1 句動詞の候補収集

まず句動詞の候補収集を行った. はじめに Wiktionary[2] の見出し語を収集し, ここから 2 語以上かつ動詞用法のものを抽出した. さらにいくつかの句動詞を人手で収集し, 合わせて句動詞の候補とした. 収集した句動詞数を表 1 に示す.

3.2 Penn Treebank から事例獲得

収集した各句動詞が, 出現した可能性のある文を Penn Treebank から抽出した. 具体的には, Penn Treebank の文中で, 句動詞の構成単語が順序通り共起した文を事例として抽出した. 注意すべき点として, 句動詞は分離可能な性質を持っているので, 句動詞間に Gaps を含むことで不連続となっている事例も抽出した.

[2]https://www.wiktionary.org/

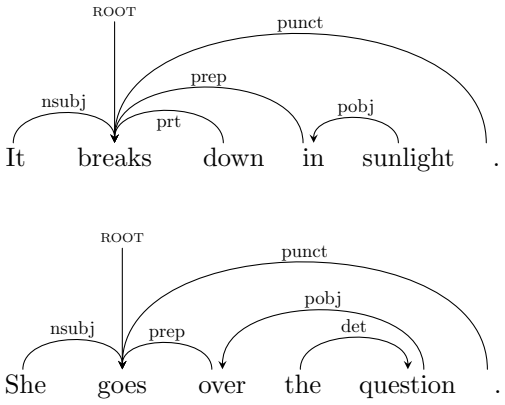

図 2: prt(break down) と prep(go over) の例

3.3 各事例への注釈

最後に Penn Treebank から抽出された各事例に対して, 正例か負例かを注釈した. 人手による注釈コストを削減するため, 本節では, Penn Treebank が持つ句構造情報を利用した句動詞のための注釈手法を提案する. 本研究で提案する手法は, はじめに Penn Treebank が持つ句構造情報を利用して機械的な注釈を行い, 曖昧な事例に対しては人手注釈を行うものである.

機械的な注釈には, 主に prt と prep という修飾ラベルの性質を利用した [1]. prt は句動詞間の修飾関係を表すラベルであり, prep は前置詞 (句) が動詞, 形容詞, 名詞または別の前置詞 (句) を修飾していることを意味するラベルである. prt と prep に関する例を図 2 に示す.

3.3.1 prt と prep の事例について

パーティクルが他の語を prt で修飾している事例においては, 事例間で直接係り受け関係下にあるものは正例と仮定し, そうでないものを負例とした. これは prt というラベルが持つ性質をそのまま利用している.

パーティクルが他の語を prep で修飾している事例においては, 句動詞表現を形成している場合と, 前置詞句による修飾を意味している場合がある. しかし, 事例間で直接係り受け関係にあることと, 連接していることの両方を満たす事例においては, その事例間で強い結びつきがあるだろうと仮定し, 正例とした. どちらか一方だけを満たす事例は曖昧な事例と仮定し, 人手による注釈を行った. どちらの条件も満たさない事例については負例とした.

表 2: 機械的注釈における規則. *は T または F を意味する.

直接係り受け	連接	ラベル	
T	*	prt	正例
F	*	prt	負例
T	T	prep	正例
F	T	prep	人手注釈
T	F	prep	人手注釈
F	F	prep	負例
T	*	*other*	人手注釈
F	*	*other*	負例

3.3.2 prt, prep 以外の事例について

句動詞となる事例は, パーティクルが prt か prep で他の語を修飾する可能性が高いが, それ以外のラベルであっても, 直接係り受け関係下にある事例においては正例の可能性もあるだろうと推測し, そのような事例に対しては人手による注釈をした. 一方, そうでない事例は負例とした. 以上の機械的注釈及び人手注釈の判断規則を表 2 に示す. なお, 表中の *other* は prt, prep 以外のラベルを指す. また, 図 2 における break down に関する事例と go over に関する事例は, それぞれ 1 行目と 3 行目の規則が使われている.

表 2 の 1 行目は, 事例間で直接係り受け関係があり, かつそのラベルが prt である場合, 機械的に正例と注釈していることを示している. この場合は連接か否かは考慮していない. しかし, 我々が組み上げたこの表 2 の枠組みだけでは, いくつかの事例間でラベルの衝突が起こってしまった. 具体的には catch up with と catch up のように, 一方の句動詞がもう一方の句動詞を完全に含んでいるような事例間において注釈の衝突が発生した. このように, ラベルの衝突が起こってしまった事例は再度人手での注釈を行った.

3.4 コーパス統計量

構築したコーパスに関する統計量を表 3 に示す. 収集した句動詞の候補一覧と Penn Treebank に存在する全 37015 文から 18069 事例が抽出された. そのうち 4823 事例が正例となり, 13246 事例が負例となった. また, 人手注釈数を 2895 事例に抑えることができた.

表 3: 事例統計量

	事例数
総正例数	4823
総負例数	13246
機械的注釈数	15174
人手注釈数	2895
計	18069

4 実験

構築したコーパスを 4:1 の比率に分割し, それぞれ学習データとテストデータとした上で, MWEs を検出する実験を行った. 具体的には, 品詞が付与された文を入力とし, その文に含まれる句動詞の同定を行う実験である. その際, 3 つの設定で実験を行った.

1. 規則に基づく手法 (ベースライン)
2. Schneider らの手法
3. 提案手法

4.1 実験設定

まず図 3 に示す規則に基づく手法で分類する実験を行った. 句動詞が分離可能な事例に対しては一つまでの Gaps を含む事例は正例と予測し, 句動詞が分離不可能な事例に対しては連接する事例にのみ正例と予測するような規則である. これを比較の際のベースラインとした.

次に Schneider ら [5] の手法で実験を行った. 彼らは 4 種類の IOB スタイルのタグ体系を示しているが, ここでは {`O o B b I i`} というタグ体系を用いる. ここで `OBI` は, それぞれ外側, MWEs の始まり, MWEs の内側を指し, `obi` は Gaps 中の外側, 始まり, 内側を指す. このタグ体系は, 彼らが提示しているタグ体系の中でも Gaps の存在を許し, 今回作成したコーパスにおいて句動詞の同定を可能にする必要最低限のタグ体系である.

最後に提案手法での実験を行った. 手法は, 各事例を 2 値分類問題と定式化し Support Vector Machine(SVM) で学習・分類を行うものである. SVM の実装には SVM_{light}[3] を用いた. 素性には表 4 に示すものを用いており, 依存構造の情報は用いていない.

[3] http://svmlight.joachims.org/

if $s = True$ then
    if $l \leq 1$ then *positive*
    else *negative*
else
    if $l = 0$ then *positive*
    else *negative*

図 3: ベースラインで用いた規則. s = 対象の句動詞が分離可能かどうか ($s = True$ なら分離可能), l = Gaps の長さ, *positive* は正例を, *negative* は負例を指す

表 4: 提案手法で用いた素性. v = 動詞, w_t = t 番目のパーティクルの単語, p_t = t 番目のパーティクルの品詞, G_{pos} = Gaps の品詞集合, l = Gaps の長さ

長さ 2 の句動詞に対する素性
vp_1, w_1p_1, vw_1, $l = 0$, $floor(l\ /\ 3)$, $\forall g\ (\in G_{pos})$
長さ 3 の句動詞に対する素性
vp_1p_2, $w_1p_1w_2p_2$, vw_1w_2, $l = 0$, $floor(l\ /\ 3)$, $\forall g\ (\in G_{pos})$

4.2 実験結果

実験結果を表 5 に示す. 2 値分類問題として解いた提案手法は, ベースラインや Schneider らの手法の精度を上回った. 系列ラベリングの手法は, 隣接する単語間の関係を主な特徴量として用いるので, 不連続の MWEs の検出には適しておらず, SVM の 2 値分類問題として定式化することで精度の向上を達成できた.

5 おわりに

句動詞に関するコーパス構築の流れとコーパス統計量, 及びその検出手法について述べた. Penn Treebank の句構造情報を利用することで, 人手コストを大幅に抑えた句動詞情報の注釈を実現した. また, MWEs 検出のための提案手法は, 規則に基づく手法やチャンキング問題として解く手法以上の検出精度を達成した.

今回は句動詞に焦点を当てて, コーパス構築及び分類実験を行ったが, 他の MWEs に関するコーパス, 特に構文的な表現に関するコーパスも十分なものが存在しないのが現状であり, 句動詞以外の表現に関するコーパスも構築する必要がある. また今回の実験では, 分類する際の素性は表層の情報に限定しており, 語が持つ意味情報等は用いなかった.

表 5: 実験結果

	適合率	再現率
ベースライン	82.90	92.70
Schneider+ ($iter = 5$, $\rho = 0$)	89.20	85.71
提案手法	**94.72**	**94.04**

今後は構文的表現に関するコーパス構築や, 語の意味情報などを導入した高精度な検出手法を考察していきたい.

参考文献

[1] Marie catherine De Marneffe and Christopher D. Manning. Stanford typed dependencies manual, 2008.

[2] Matthieu Constant, Anthony Sigogne, and Patrick Watrin. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 204–212, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[3] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. *CoRR*, Vol. cmp-lg/9505040, , 1995.

[4] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15, 2001.

[5] Nathan Schneider, Emily Danchik, Chris Dyer, and A. Noah Smith. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pp. 193–206, 2014.

[6] 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 the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[7] 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 *Proceedings of the 9th Workshop on Multiword Expressions*, pp. 139–144, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.