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

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

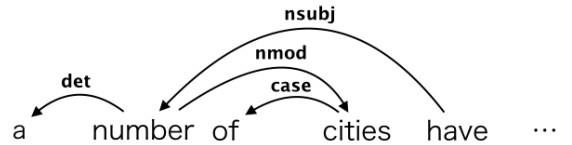
1 はじめに

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 は「単語境界を越える特異的な表現」としても知られ [1]，以下に示す様に様々な種類の MWE が存在する [1]．

MWE 種別	例
複合機能語	a number of, even though, after all
句動詞	look up, take over
複合名詞	customer service, traffic light
固有表現	New York, United Nations

MWE はしばしば意味的な非構成性を持つため，MWE の認識は情報検索 [2]，意見マイニング [3] など多くの応用タスクで重要な役割を果たしている．これらの応用タスクでは MWE を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があるため，単語ベースの依存構造よりも MWE ベースの依存構造を用いる方が望ましいと考えられる．単語ベースの依存構造では MWE の範囲は表現されない（図 1a）のに対して，MWE ベースの依存構造では MWE が単一トークンとして表現され，MWE としての品詞（以下，全体品詞）を考慮する事ができる（図 1b）．MWE の中でも特に複合機能語は，しばしば構成単語の品詞列から期待されるものとは異なる文法的性質を持つため，MWE の全体品詞を構成単語の品詞列から想定するのは難しい場合がある．例えば “by and large” の全体品詞は副詞であるが，これを構成単語の品詞列（IN，CC，JJ）から推定する事は難しい．

MWE ベースの依存構造コーパスの例としては，フランス語 MWE の大規模なアノテーションを施した French Treebank [4] が挙げられる．一方，英語 MWE については Penn Treebank に MWE アノテーションが付与されていなかった事もあり，利用可能なコーパスはこれまで限られていた．Schneider ら [5] は，English Web Treebank [6] 上に MWE の注釈付けを行ったが，これは依存構造解析器の訓練データとしては比較的小規模である [2]．また彼らの MWE アノテーションは句構造との整合性が保証されていない．つまり，MWE の範囲は単一の非終端ノードが支配する範囲に対応するとは限らないという問題点がある．



(a) 単語ベースの依存構造

(b) MWE ベースの依存構造

図 1: 単語ベースの依存構造と MWE ベースの依存構造．後者では MWE（図中の “a number of”）は単一トークンとして表現されており，依存構造中で限定詞として働く．

我々はこれまでに，複合機能語を考慮した MWE ベースの依存構造コーパスを構築してきた [7]．本研究では，これを固有表現にまで拡張し，より網羅性の高い MWE ベースの依存構造コーパスを構築する．我々の用いた英語の OntoNotes コーパス [8] では，固有表現と句構造木がそれぞれ独立にアノテートされており，固有表現の範囲と句構造木との整合性が保証されているわけではない．例えば，図 2 の句構造木において固有表現アノテーションは “Board of Investment” となっ

[1] 本研究では複数の単語からなる固有表現も広義の MWE として扱う．

[2] 約 3800 文中に約 3400 回 MWE が出現している．

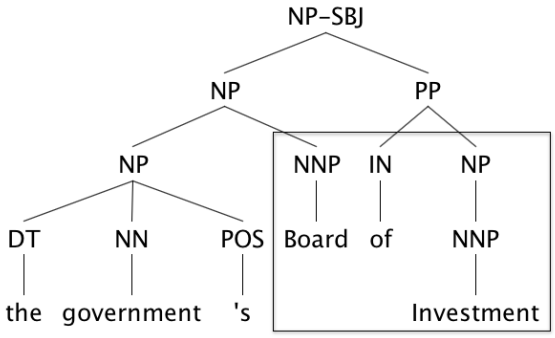


図 2: 固有表現と句構造木が不整合になっている例．図中の矩形は固有表現の範囲を示す．

ており，いずれの非終端ノードの範囲にも対応していない．したがって本研究では，固有表現の範囲が句構造木と整合しない場合，句構造木を局所的に修正する事によって固有表現を単一の部分木にまとめ，固有表現の範囲と句構造木との整合性の問題を解決する．

また，本研究で構築したコーパスを学習データとして，MWE ベースの依存構造解析を行う手法を提案する．本研究の依存構造コーパスでは，複合機能語と固有表現が単一トークンとなっている．そのため，実際に文の解析を行う場合には，MWE の範囲や種類の同定（MWE 認識）と，MWE を一トークンとする依存構造解析の双方を行う必要がある．そこで我々は両者の同時解析を行う Joint+supertagging モデルを提案する．提案手法は条件付き確率場（CRF）で予測した MWE の範囲を追加素性とした依存構造解析を行うため，CRF からの誤り伝播を緩和し，かつ文全体の情報を考慮する事ができる．また，複合機能語はほぼ辞書で網羅できるが，固有表現は高い生産性を持つため，依存構造解析時に用いる辞書制約は複合機能語についてのもののみとする．

我々はベースラインと提案手法について，構築したコーパス上で実験を行った（4 章）．なお，本研究で構築したコーパスの前身である複合機能語ベースの依存構造コーパスは 2017 年 3 月現在，LDC（Linguistic Data Consortium）から配布されている [3]．

2 MWE ベースの依存構造コーパス

我々は Ontonotes Release 5.0（LDC2013T19）の Wall Street Journal 部分において，句構造木に複合機能語と固有表現のアノテーションを統合した上で，MWE ベースの依存構造へと変換を行った．我々はまず句構造木と複合機能語アノテーションを統合し（2.1），次に統合対象を固有表現アノテーションまで拡張した（2.2）

[3] https://catalog.ldc.upenn.edu/LDC2017T01

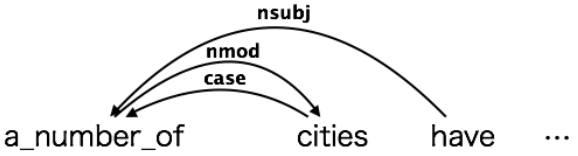


図 3: 単語ベースの依存構造において MWE 配下のノードを 1 つにまとめた結果，ループと複数の主辞が生じる例

ので順に説明する [4]．

2.1 複合機能語ベースの依存構造コーパス

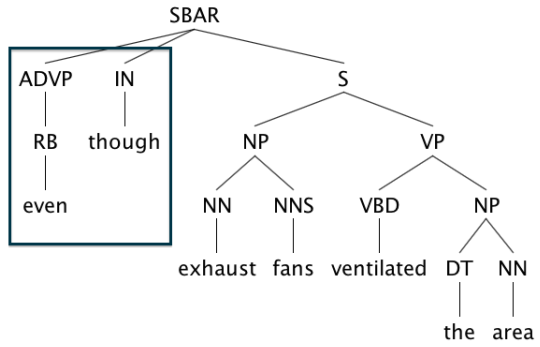
MWE ベースの依存構造を得るための最も単純な手段としては，単語ベースの依存構造において各 MWE を単一ノードにまとめる直接変換型のアプローチが考えられる．しかしこの手段を適用すると事例によっては木構造が得られない事がある．例えば図 1a で MWE を単一ノードにまとめると “number” → “cities” と “cities” → “of” のエッジによってループが生じる（図 3）．また 図 3 では “a number of” が “cities” と “have” という複数の主辞を持っており，根以外のノードはただ一つの主辞を持つという依存構造木の必要条件を満たしていない．さらに，MWE ベースの依存構造（図 1b）では “a number of” は限定詞であるため，“a number of cities” の主辞は “cities” となる．従って “have” の子ノードは “a number of” でなく “cities” となるが，これも上記の方法では得る事が出来ない．

上記の問題点を解決するために，我々はまず句構造木において MWE を単一の部分木にまとめ，その後で依存構造に変換するというアプローチを採用する．このアプローチであれば，上述したループや複数の主辞の発生を回避する事ができる．

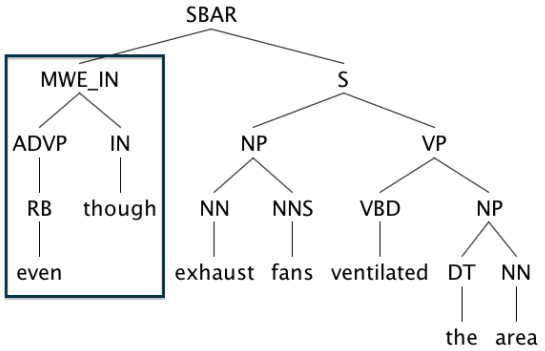
2.1.1 MWE の部分木へのまとめ上げ

以下では MWE を単一の部分木にまとめる手順について述べる．まず各 MWE の範囲が句構造木の非終端ノードに対応するかどうかを調査する．非終端ノードに対応しない MWE は（1）複数の連続した子ノードに対応するもの（“multiple contiguous children”），（2）それ以外（“crossing brackets”）に分類される [7, 10]. “multiple contiguous children” については MWE の範囲に対応するノード群をまとめる新しい非終端ノードを挿入し，シンボルとして MWE 全体品詞を持たせる（図 4a → 図 4b）．また “crossing brackets” についてはできるだけ元の木構造を残しつつ，MWE が単一の部分木になる様に句構造木を修正する（図 5）．なおマニュアルアノテーションは “crossing brackets” の一

[4] 複合機能語に関するアノテーションとしては Shigeto ら [9] によるものを用いた．



(a) 修正前の句構造木（MWE に関する部分のみを記載）．図中の矩形は MWE の範囲を示す．



(b) 修正後の句構造木．MWE（“even though”）が部分木にまとめられている．

図 4: “multiple contiguous children” における句構造木の修正

部の事例のみで必要となる．詳細は加藤ら [7] を参照されたい．

2.1.2 依存構造への変換

依存構造を得る準備として，上記で得られた MWE に対応する部分木を単一のトークンにまとめる．このトークンの品詞タグは MWE の全体品詞であり，その表層形は，MWE の全ての構成単語の表層形をアンダースコアで連結したものである．最後に句構造を依存構造に変換する（図 1b）[5]．

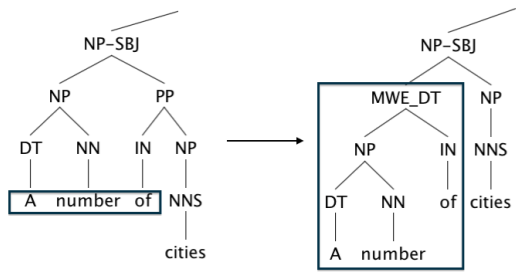


図 5: “crossing brackets” における句構造木の修正

[5]本研究では Stanford Dependency を採用した．

表 1: コーパス統計量

文数	37,015
MWE の出現数	27,949
MWE の異なり数	12,670

全体品詞	NNP	RB	IN	その他	Total
出現数	20,992	3,796	2,424	737	27,949

表 2: MWE の範囲と句構造木の整合性についてのヒストグラム

MWE 種別	ケース名	出現数
複合機能語	非終端ノードに対応	3,466
	“multiple contiguous children”	1,663
	“crossing brackets”	1,799
固有表現	非終端ノードに対応	18,625
	“multiple contiguous children”	2,252
	“crossing brackets”	144

2.2 固有表現アノテーションの句構造木への統合

2.1 と同様の手法を用いて Ontonotes 5.0 で提供されている固有表現アノテーションの句構造木への統合を行った [6]．ただし “crossing brackets” に分類される事例の内，句構造木を修正するよりも固有表現スパンを拡張する方が適切であると考えられる事例については後者の方針を取った [7]．

2.3 コーパス構築についてのまとめ

表 1 に構築したコーパスの統計量を示す．NNP を全体品詞に持つ MWE は固有表現であり，RB や IN 等，機能表現に相当する全体品詞を持つ MWE は複合機能語である．また，MWE の範囲と句構造木との関係を 2.1.1 で述べた基準で分類して得られたヒストグラムを 表 2 に示す．

3 モデル

以下ではベースラインであるパイプラインモデルと Joint モデル，および提案手法である Joint+supertagging モデルについて概略を述べる．詳細は加藤ら [11] を参照されたい．

[6]固有表現のタイプは PERSON / NORP / FACILITY / ORGANIZATION / GPE / LOCATION / PRODUCT / EVENT / WORK OF ART / LAW / LANGUAGE に限定し，数値・数量表現 / 日付・時間表現等は対象外とした．また，本研究では MWE に着目しているため，1 単語のみからなる固有表現は対象外とした．

[7]例としては “Peter and [Edward Bronfman]” の様に，並列句の一部のみに固有表現アノテーションが付与されているケースが挙げられる．この場合，並列句全体への固有表現スパンの拡張を行った．

表 3: テストセットに対する実験結果

	依存構造解析		MWE 認識	
Model	**UAS**	**LAS**	**FUM**	**FTM**
Pipeline	91.39	89.42	91.40	91.32
Joint	91.15	89.18	89.03	88.79
Joint+supertagging	91.50	89.51	92.75	92.60

3.1 パイプラインモデル

本モデルでは CRF で予測した MWE の範囲と全体品詞に基づいて，MWE を単一トークンとする依存構造を ArcEager 法に基づく遷移ベースの解析器によって推定する．

3.2 Joint モデル

MWE の範囲と依存構造を同時に推定するために，本モデルでは MWE ベースの依存構造（ 図 1b）中の MWE を Head-initial な部分木に変換した依存構造をデータ表現として用いる [8]．そして遷移ベースの解析器を用いて従来の単語ベースの依存構造解析を行う．この際，誤った依存構造木の推定を避けるため，遷移履歴と複合機能語辞書に基づく制約を用いる [9]．

3.3 Joint+supertagging モデル

本モデルでは，CRF で予測した MWE の範囲と全体品詞を Joint モデルの追加素性（以下，MWE 素性）として用いる．このため，パイプラインモデルで発生しうる系列ラベリングからの誤り伝播を緩和し，かつ文全体の情報を考慮できないという Joint モデルの欠点を緩和できると期待される．依存構造解析で用いるデータ表現と制約は Joint モデルと同一である．

4 実験設定と結果および考察

実験設定は加藤ら [11] と基本的に同一であるが，CRF で利用する辞書素性については複合機能語辞書に加えて固有表現辞書を用いた [10]．実験結果を 表 3 に示す [11]．まず依存構造解析については，Joint+supertagging モデルが Joint モデルよりも 0.35 ポイント，パイプラインモデルよりも 0.11 ポイント高い UAS 値を示した．この結果は，Joint+supertagging モデルで用いた MWE 素性が，MWE の周辺や MWE 内部の依存構造の推定に有効である事を示唆している．次に MWE 認識については，Joint+supertagging モデルがパイプラインモデルを FUM 値で 1.35 ポイント上回った．一方，Joint モデルはパイプラインモデルよりも 2.37 ポイント低い FUM 値を示した．これらの結果は，依存構造解析器による MWE 認識で MWE 素性が有効である事と，MWE 素性無しでは CRF による MWE 認識に及ばない事を示唆している．

5 結論

本研究では英語の複合機能語と固有表現を各々単一トークンとする依存構造コーパスを Ontonotes 上に構築した．また MWE ベースの依存構造解析のためのモデルを構築して上記コーパスを用いた実験を行い，MWE 認識において提案モデルがベースラインを上回る F 値を示す事を確認した．

[8] Head-initial な部分木では，MWE の全ての後続単語は先頭単語を親に持ち，MWE の範囲と全体品詞が係り受けラベルとして表現されている．これは Universal Dependency [12] の複合機能語の表現形式に類似している．

[9] 複合機能語辞書としては Shigeto ら [9] が Wiktionary から網羅的に作成したものを用いた．

[10] 固有表現辞書は Wikipedia のタイトル一覧からストップワードを除いて作成した．Wikipedia のタイトル一覧は Wikipedia Dump（https://dumps.wikimedia.org/enwiki/20160920）から，またストップワードのリストは Stopword List（http://members.unine.ch/jacques.savoy/clef/englishST.txt）に 's を加えて作成した．

[11] 依存構造解析の精度は句読点を除いたトークンに関するものである．UAS / LAS はラベルなし / ラベルあり正解率である．FUM は MWE の範囲に関する F 値，FTM は MWE の範囲と全体品詞に関する F 値である．

参考文献

[1] Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A. and Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP, *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing, London, UK, UK, Springer-Verlag, pp. 1–15 (2002).

[2] Newman, D., Koilada, N., Lau, H. J. and Baldwin, T.: Bayesian Text Segmentation for Index Term Identification and Keyphrase Extraction, *COLING*, pp. 2077–2092 (2012).

[3] Berend, G.: Opinion Expression Mining by Exploiting Keyphrase Extraction, *IJCNLP*, Asian Federation of Natural Language Processing, pp. 1162–1170 (2011).

[4] Abeillé, A., Clément, L. and Toussenel, F.: Building a Treebank for French, *Treebanks : Building and Using Parsed Corpora*, Springer, pp. 165–188 (2003).

[5] Schneider, N., Onuffer, S., Kazour, N., Danchik, E., T. Mordowanec, M., Conrad, H. and A. Smith, N.: Comprehensive Annotation of Multiword Expressions in a Social Web Corpus, *LREC* (2014).

[6] Bies, A., Mott, J., Warner, C. and Kulick., S.: English Web Treebank, *Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.* (2012).

[7] Kato, A., Shindo, H. and Matsumoto, Y.: Construction of an English Dependency Corpus incorporating Compound Function Words, *LREC* (2016).

[8] Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A. and Weischedel, R. M.: OntoNotes: A Unified Relational Semantic Representation, *IEEE-ICSC*, pp. 517–526 (2007).

[9] Shigeto, Y., Azuma, A., Hisamoto, S., Kondo, S., Kouse, T., Sakaguchi, K., Yoshimoto, A., Yung, F. and Matsumoto, Y.: Construction of English MWE Dictionary and its Application to POS Tagging, *NAACL MWE Workshop*, pp. 139–144 (2013).

[10] Finkel, R. J. and Manning, D. C.: Joint Parsing and Named Entity Recognition, *NAACL*, Association for Computational Linguistics, pp. 326–334 (2009).

[11] Kato, A., Shindo, H. and Matsumoto, Y.: 複単語表現を考慮した英語の依存構造解析モデリング, **情報処理学会研究報告**, *Vol.2016-NL-229, No.23, pp.1-8. December 2016.*, IPSJ.

[12] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N. and Lee, J.: Universal Dependency Annotation for Multilingual Parsing, *ACL*, Association for Computational Linguistics, pp. 92–97 (2013).