自然言語処理：言語資源・意味解析（レクチャーシリーズ「人工知能の今」第6回）

レクチャーシリーズ「人工知能の今」〔第６回〕

自然言語処理
—言語資源・意味解析—

Natural Language Processing: Language Resources and Semantic Processing

松林　優一郎　東北大学大学院教育学研究科
Yuichiroh Matsubayashi　Department of Education, Tohoku University.
y.m@tohoku.ac.jp

浅原　正幸　人間文化研究機構国立国語研究所
Masayuki Asahara　National Institute for Japanese Language and Linguistics, Japan.
masayu-a@ninjal.ac.jp

Keywords: natural language processing, syntax, semantics, language resources.

1. はじめに

本稿では，深層学習全盛期の現在に紹介しておくべき自然言語処理の基礎的な知識として言語資源と呼ばれる自然言語処理の根底を支えているデータ群と，意味解析に関する基礎技術について解説する．

自然言語処理は主に人間の書き言葉を対象にした人工知能技術である．この分野では 1990 年代に文章を集積したコーパスと呼ばれるデータが利用され始めて以来，統計的な手法によって計算モデルを構築する方法が技術開発の主流となってきた．2010 年代に入ると，昨今の深層学習ブームの例に漏れず，いくつかのニューラルネットワーク手法の成功によりその主たるアプローチが大きく変化した．特に重要な転換点となったのは，それまで {0, 1} で離散的にその生起*1 を表現していた単語の情報を，実数値ベクトルを使った連続的な表現で表す手法が成功を収めたこと（単語分散表現の獲得 [Bojanowski 17, Mikolov 13a, Pennington 14]）と，文や文章のように単語が系列としてもつ情報を同じく実数値ベクトルで表現するための計算モデルを学習で構築できるようになったこと（単語系列に対する表現学習の実現 [Cho 14, Collobert 11, Mikolov 10, Socher 11]）である．

2000 年代の自然言語処理といえば，形態素解析・構文解析・意味解析などの個別のタスクごとに，複雑な計算手法の組合せや人手による特徴量工学（feature engineering）によって効果的な解析モデルの探索を行う方法論が主流であったが，今や解析時に利用する特徴量は単語レベルの情報から単語系列の内側に隠された文法的な構造情報に至るまで，ほとんどのものは深層学習内の表現学習によって自動的に獲得されている．各タスクの解析に使われるニューラルネットワークでは，解析対象となる文章の単語列あるいは文字列をそのまま入力値として用い，最終的に得たいラベルや構造を出力とする end-to-end と呼ばれる方式で学習と推定が行われ，この技術でつくられた解析器は自然言語処理のおおむねすべてのタスクで従来法をはるかにしのぐ解析精度を実現してきた．このような背景のもと，自然言語処理研究者の興味の対象は，これまで取り組まれてこなかったより難解な問題を定式化する方向，およびネットワーク全体の設計によって言語構造の機械学習にどのような帰納バイアスをもたせるべきかを探索する方向に転換しつつある．

一方で，end-to-end の解析手法ではモデルパラメータが本質的に何を意味するものかが自明ではなくなり，計算過程がブラックボックス化する．深層学習によるモデル獲得は，解きたいタスクを定義し，訓練時の入力から出力を再現するモデルを学習するアプローチであるが，なぜそのモデルが知的なタスクを解けるのか，言語の意味や構造についてどのくらい理解できているのかについては検証が難しい，あるいは検証自体を重視しないアプローチである．結果としてこれらのモデルでは，従来の意味解析で明確化されてきた，言語のどのような語彙的・構造的特徴を捉えているかの部分がわかりづらくなった．

しかしながら，社会的な要請として深層学習におけるその本質の部分の明確化は重要視されている．深層学習の解釈性・説明性の検証はプロービング（probing）と呼ばれ，学習済みのモデルに暗に含まれる言語の構造を，既存の言語資源が規定する個別のタスクとの関連から示されることが求められている．こうした要請の中では，これまで自然言語処理分野で構築されてきた言語資源や，基礎的な解析レイヤの種類，過去に発見されてきた言語解析のための有用な特徴量を理解しておくことが，

*1 出現の有無.

表1 分類語彙表の構造

レコードID	番号見出し	類	部 門	中項目	分類項目	分類番号	段落番号	小段落番号	語番号	見出し
001946	01838	体	関 係	存 在	成 立	1.1220	14	01	03	国 立
030548	29140	体	活 動	言 語	言 語	1.3101	03	01	01	国 語
022620	21486	体	主 体	社 会	社寺・学校	1.2630	15	01	01	研究所

挙動の理解に対する有用な手掛かりとなり得る．End-to-end モデル以前の自然言語処理では，文分割，単語分割，構文解析，語義曖昧性解消，意味構造解析，談話解析といったように，低レベルの言語解析レイヤから最終的な意味理解に至るまで，階層的に解析の手順を積み上げてきた．このパラダイムは，解析の各段階において言語の構造や意味を汎用的・普遍的に解析することを目指す，いわば現在の深層学習で行われているものとは対極的なアプローチである．しかし，こうした問題の切分けと，言語機能に対する一般性をもった問題の定式化は，言語理解に関する普遍的な解析能力の獲得とその評価にあたっては不可欠な要素である．本稿では，そのような立場から，言語のもつ一般的な構造や意味を捉える要素技術と，その構築の要となる資源である言語資源について解説する．

本稿の前半では，言語資源について解説する．上述の階層的に解析を積み上げる自然言語処理のパラダイムでは，それぞれ中間レイヤの処理結果としてある種の言語解析理論に基づいた合理的で整合性のある構造を割り当てる必要がある．ここで説明する言語資源と呼ばれるデータ群は，実際の文章や語彙辞書といった言語データに対して分析の具体例を与えることによって，これらの中間レイヤの出力に関して合理的な正解の構造を規定し，解くべき問題自体を定式化する役割を果たす．ここで与えられた実世界の生の言語データに対する分析例が各解析レイヤの教師あり機械学習の訓練・評価用データとして利用される．おのおのの言語資源が作成される過程では，その資源で興味の対象とする解析内容について，それに関わる言語理論が実世界のバラエティに富んだデータに照らして頑健な解析を与えられるレベルとなるよう，緻密に集積・構築・改善されてきた．

したがって，こうした資源に記述されている知識は，言語に関する非常に高度でかつ汎化された知識の結晶体である．本稿では，2章で，単語の語義の情報などを階層的に付与した『分類語彙表』[国立 04] に関連するデータと単語埋込み･事前学習モデルについて示す．3章では，『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）[Maekawa 14] に対するアノテーション（付加情報）について，統語情報（3･1節）・語義情報（3･2節）・文脈情報（3･3節）の三つに分けて紹介する．4章では，どのように言語資源を構築するか，作業環境・基準の策定手法について示す．言語資源の構築作業は統語・意味情報処理において，問題の定式化に直結する重要な作業であり，その構築過程そのものが研究の対象とされている．また，近年言語資源構築に利用されつつあるクラウドソーシングについても触れる．

本稿の後半では，基礎的な解析レイヤの一例として，意味役割解析，共参照・照応解析と呼ばれる意味解析技術を取り上げる．これらの基礎的な解析レイヤの処理結果は，言語データから汎用性の高い基本的な構造を取り出すために役に立つ．また，新たに作成した自然言語処理応用技術の性能を検証する過程において，そのモデルが言語の基本的な性質を理解しているかを確かめるための道具としても利用できる．5章では，意味役割付与，6章では共参照・照応解析と呼ばれる意味解析の基礎技術について，それぞれ最新のモデル，開発に利用されるデータ，過去に培われてきた解析のための主要な特徴量について紹介する．ここで取り上げるような細分化された言語の基礎解析技術は，それぞれが汎用的・普遍的な言語解析を目的として設計されたものである．したがって，個別のタスクの出力値に最適化された深層学習モデルと比べ，既存の情報処理システムに新たに自然言語処理の機能を取り込む際のコンポーネントの単位としても使い勝手が良い単位となっている．

2. 語 彙 資 源

本章では，日本語の統語・意味情報関連の語彙資源として，単語の語義の情報などを階層的に付与した『分類語彙表』[国立研究所 04] に関連するデータと，単語埋込み・事前学習モデルについて示す．

2･1 『分類語彙表』と関連データ

『分類語彙表』[国立研究所 04] は，「語を意味によって分類・整理したシソーラス（類義語集）」である．1964年に初版が出版され，2004年に増補改訂版が出版された．この増補改訂版の元データが研究用途に公開されている*2．

表1に分類語彙表の構造を示す．各語を一意に決める「レコードID番号」と「見出し語」を一意に決める見出し番号が付与されている．表中7列目の「分類番号」が

*2 https://pj.ninjal.ac.jp/corpus_center/goihyo.html もしくは https://github.com/masayu-a/wlsp からダウンロードできる．クリエイティブ・コモンズ表示–非営利–継承 3.0 非移植（CC BY-NC-SA3.0）だが，別途有償契約することで営利目的にも利用可能．

表 2　WLSP-Familiarity：単語親密度情報

分類番号	見出し	知っている	書　く	読　む	話　す	聞　く	生　産	受　容	書　記	音　声	生産–受容	書記–音声
1.3360	祭　り	1.74	0.75	0.98	1.05	1.14	1.79	2.13	1.73	2.19	− 0.34	− 0.46
1.3360	お祭り	2.04	0.86	1.16	1.39	1.44	2.25	2.60	2.01	2.83	− 0.35	− 0.82
1.3360	祭　事	1.27	− 0.32	0.26	− 0.38	0.04	− 0.71	0.30	− 0.07	− 0.34	− 1.00	0.27
1.3360	祭　祀	0.03	− 1.09	− 1.02	− 1.15	− 1.08	− 2.23	− 2.10	− 2.10	− 2.23	− 0.14	0.12
1.3360	神　事	1.08	− 0.93	− 0.28	− 0.39	0.02	− 1.32	− 0.25	− 1.20	− 0.37	− 1.06	− 0.83

語義のラベルに相当するもので，ピリオドより左が統語のラベル，ピリオドより右が意味のラベルである．統語のラベルとして「類」があり，体（1）・用（2）・相（3）・他（4）の 4 種類が設定されている．意味のラベルは，分類番号をピリオドの右 1 桁によって決められる「部門」，2 桁によって決められる「中項目」，4 桁によって決められる「分類項目」がある．「段落番号」,「小段落番号」は，類義語を示すグループを規定する．「語番号」は小段落番号以下で一意に単語を決める番号である．

近年，分類語彙表データに対するさまざまな拡張が進められている．以下ではさまざまな拡張情報について示す．

- UniDic- 分類語彙表対応表 [近藤 20]
- 反対語情報 [荻原 19]
- 単語親密度情報 [Asahara 19]
- 代表義・派生関係 [加藤 19c, 山崎 17]
- 分類語彙表–岩波語義対応表 [呉 19]

UniDic- 分類語彙表対応表 [近藤 20] は，形態素解析用辞書 UniDic の語彙素の情報と分類語彙表の見出し語との対応表である*3．表記揺れも含めたすべての可能な見出し語の多対多対応を整備した．表 3 に例を示す．UniDic の語彙素番号（LemmaID）に対して，対応する分類語彙表情報を付与している．形態素解析器 MeCab の出力と対照することで，可能なすべての語義のラベルが付与できる．Windows 利用者は対応表が同梱された ChaMame*4 を利用されたい．

表 3　UniDic- 分類語彙表対応表

分類語彙表情報	LemmaID	語彙素
1.1000, 体 - 関係 - 事柄 - 事柄, 1.1000-01-01-01	37877	者
1.1000, 体 - 関係 - 事柄 - 事柄, 1.1000-01-01-02	28990	の
1.1000, 体 - 関係 - 事柄 - 事柄, 1.1000-01-01-03	33543	分
1.1000, 体 - 関係 - 事柄 - 事柄, 1.1000-01-01-04	38347	奴

分類語彙表の語義のグループには類義語のみならず反対語が含まれる．この分類語彙表内の反対語の情報付与 [荻原 19] が進められている*5．同一の小段落番号・段落番号をもつ語のグループから反対語と思われるものを作業者により抽出したうえで，クラウドソーシングを用いて一般の日本語話者が反対語とみなすかどうかの評定値を付与した．反対語の認定においては，文脈において置換え可能かどうかの評定も行った．さらに反対語のタイプとして「両極」,「相補」,「視点」,「程度」,「変化」,「部分全体」,「2 側面」,「典型（2 値）」,「慣用」などの情報を付与している．表 4 に例を示す．評定値は反対語でない程度と，文脈中で置換可能かどうか（格交代しないかどうか）などを評価した．

表 4　WLSP-Antonym：反対語情報評定値

評定値 反対語でない	評定値 置換化	対象 A	対象 B	分　類
0	0.95	メリット	デメリット	相　補
0.15	0.5	未　熟	円　熟	両　極
0.15	0.45	行　く	帰　る	視　点
0.2	0.5	穿　く	脱　ぐ	変　化
0.3	0.55	掛け算	割り算	典　型
0.1	0.5	水　平	垂　直	2 側面
0.15	0.65	うまい	まずい	程　度
0.05	0.55	名を上げる	名を汚す	慣　用

分類語彙表の見出し語には単語親密度の情報 [Asahara 19] が別途付与されている*6．見出し語を書記刺激*7 として，クラウドソーシングにより以下の 5 観点の評定値を収集した．

- 「知っているか」（KNOW）
- 「書いているものにあらわれるか」（WRITE）
- 「読んでいるものにあらわれるか」（READ）

*3 https://github.com/masayu-a/WLSP2UniDic
*4 https://ja.osdn.net/projects/chaki/releases/70372
*5 https://github.com/masayu-a/WLSP-antonym
*6 https://github.com/masayu-a/WLSP-familiarity
*7 実験協力者の言語を書き言葉として画面にその書字形を呈示すること．

表 5 代表義・派生関係の検討

語彙素	分類番号		用例数	日国上位語義	代表義判定 [山崎 17]	派生関係 [加藤 19c]
若 い	3.1660	相–関係–時間–新旧・遅速	**42**			3.5701 から部分→全体
若 い	3.5701	相–自然–生命–生	17	○	○	
元 々	3.1030	相–関係–真偽–真偽・是非	**14**	○		
元 々	3.1642	相–関係–時間–過去	8		○	
も う	3.1670	相–関係–時間–時間的前後	48	○	○	
も う	3.1920	相–関係–量–程度	**72**			
近 い	3.1110	相–関係–類–関係	10			3.1911 から具体→抽象
近 い	3.1130	相–関係–類–異同・類似	**22**			3.1911 から転用
近 い	3.1643	相–関係–時間–未来	4			3.1911 から具体→抽象
近 い	3.1911	相–関係–量–遠近	**22**	○	○	

●「話すときにあらわれるか」（SPEAK）
●「聞くときにあらわれるか」（LISTEN）

これらの評定値はクラウドソーシングの作業者ごとに分布の偏りがある．分布の偏りの問題を緩和するために Bayesian Linear Mixed Model により統計処理したデータが公開されている．親密度の情報に基づき，語彙数判定テストが構築できるほか，複数の観点の差を取ることにより，見出し語の使用域が認定できる．例えば，WRITE + READ − SPEAK − LISTEN の値を計算することにより，書き言葉偏重（正の値）・話し言葉偏重（負の値）*8 などがわかる．WRITE − READ + SPEAK − LISTEN の値を計算することにより，生産過程偏重（正の値）・受容過程偏重（負の値）*9 などがわかる．表 2 に例を示す．

分類語彙表の異なり語 81 250 語のうち 17.8％は，一つの見出し語に対して複数の語義ラベルが付与されている多義語である．多義語の場合に，どの語義が代表義であるかという情報が付与されている [山崎 17]．代表義の認定にあたっては，コーパス中の頻度・具体性・類義語数などに基づいて作業が行われた．また，相の類を中心に代表義からの派生関係の調査も進められている [加藤 19c]．表 5 に代表義・派生関係の検討について示す．

分類語彙表には語釈文がない．この問題を解決するために『岩波国語辞典第五版タグ付きコーパス 2004』*10 に含まれる語釈文との対応表 [呉 19] が整備されている *11．表 6 に例を示す．

上に述べた，分類語彙表に関連する語彙資源の一部は Web 上の辞書ツール Cradle から利用できる *12．

表 6 岩波国語辞典–分類語彙表対応表

岩波国語辞典	分類語彙表
かくれる，隠れる，07730.0.1，1-1，姿が見えない状態になる．ものの陰で外から見えなくなる．「月が隠れる」人目につかないようにする．ひそむ．「人込みに隠れる」，	079081，75561，A，用，活動，心，見る，2.3091，24，07，01，隠れる，隠れる，かくれる，るれくか，，
かくれる，隠れる，07730.0.2，1-2，官職につかずに野にいる．「隠れた人材」，	067671，65148，A，用，関係，存在，出没，2.1210，11，01，01，隠れる，隠れる，かくれる，るれくか，，

2·2 単語埋込みと事前学習モデル

自然言語処理において語義を扱う基本的な言語資源として単語埋込み（word embedding）がある．単語埋込みは各単語の意味をベクトル表現化する技術で，コサイン類似度による 2 単語の意味的な近さの計算や，ベクトルの加減算による語義の合成・分解が可能になる．古くは，分布意味論 *13 において，単語の共起をモデル化する単語文書行列を構成していた．単語文書行列は一般に疎な行列であるため，特異値分解により，密な単語共起ベクトルを構成する手法が用いられてきた．2000 年代は，pLSA（Probabilistic Latent Semantic Analysis）[Hofmann 99]，LDA（Latent Dirichlet Allocation）[Blei 03] など，トピックモデルに基づいて推定されていた．2010 年代に入り，人工神経回路に基づく単語埋込みの研究が進展し，Word2Vec *14 [Mikolov 13a, Mikolov 13b]，GloVe *15，fastText *16 [Bojanowski 17] など，さまざまなモデルが提案された．日本語における主な訓練済み単語埋込みデータ資源の一覧を表 7 に示す．これらのデータを用いることにより，単語に実数値ベクトル表

＊8 書き言葉を WRITE，READ，話し言葉を SPEAK，LISTEN とする．
＊9 生産過程を WRITE，SPEAK，受容過程を READ，LISTEN とする．

＊10 https://www.gsk.or.jp/catalog/gsk2010-a/
＊11 https://github.com/masayu-a/WLSP2iwanami
＊12 https://cradle.ninjal.ac.jp/
＊13 Distributional Semantics：同じ文脈に出現する傾向にある語が，似た意味をもつという仮定にもとづいた意味論．
＊14 https://github.com/tmikolov/word2vec
＊15 https://nlp.stanford.edu/projects/glove/
＊16 https://fasttext.cc/

表 7　日本語の訓練済み単語埋込みデータ（多言語も含む）

摘　要	URL
Kyubyoug's Data	https://github.com/Kyubyong/wordvectors
Polyglot	https://sites.google.com/site/rmyeid/projects/polyglot
@Hironsan's Data	https://qiita.com/Hironsan/items/513b9f93752ecee9e670
fastText 157 languages [Grave 18]	https://github.com/facebookresearch/fastText/
wikipedia2vec	https://wikipedia2vec.github.io/wikipedia2vec/pretrained/
日本語 Wikipedia エンティティベクトル [鈴木 16]	http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
白ヤギコーポレーション	https://github.com/shiroyagicorp/japanese-word2vec-model-builder
朝日新聞単語ベクトル [田口 17]	https://cl.asahi.com/api_data/wordembedding.html
NWJC2vec [Asahara 18]	http://nwjc-data.ninjal.ac.jp/
ワークスアプリケーションズ [真鍋 19]	https://www.worksap.co.jp/nlp-activity/word-vector/

表 8　日本語の訓練済み BERT 事前学習データ

摘　要	URL
BERT with SentencePiece for Japanese text [Kikuta 19]	https://github.com/yoheikikuta/bert-japanese
hottoSNS-BERT	https://www.hottolink.co.jp/blog/20190311-2
BERT 日本語 Pretrained モデル（京都大学）[柴田 19]	http://nlp.ist.i.kyoto-u.ac.jp/「NLP リソース」

現を付与することができ，ベクトルのノルムにより単語が出現する文脈的な近さを得ることができる．分かち書きの単位，訓練データ（サイズ・ドメイン），訓練に用いたパッケージなど，それぞれに特色があり，用途に応じて選ぶ必要がある．

また，文脈を考慮した単語埋込みを学習する ELMo[*17] [Peters 18] が提案された．従来の単語埋込みは，表層形が同じ語に対して，常に同じベクトル表現を与えるものであった．ELMo の場合は，表層形が同じ語に対して，その文脈に応じて異なるベクトル表現を与えるため，多義語（同表記異義語）の問題が扱えるようになった．

さらに，自然言語処理の事前学習モデル（pre-training）である BERT が提案された [Devlin 19]．BERT は「空白穴埋め問題」と「文の隣接課題」[*18] を解くモデルを，転移学習（transfer learning）により，他の似た課題に転用できる．BERT の公開サイト[*19] に多言語訓練済みが公開されているほか，日本語の訓練済み事前学習モデルが公開されている（表 8）．これらのモデルを用いて，ELMo と同様に文脈を考慮した単語埋込みも得ることもできる[*20]．

3. コーパスとアノテーション

本章では BCCWJ に付与されているさまざまな統語・意味情報アノテーションデータについて触れる．これらのデータは，主に自然言語処理の研究者により，教師あり機械学習のための正解データとして構築された．ほかに，コーパス言語学の研究者により，言語運用の実態を解明するための統計分析用データとして構築されたものもある．3·1 節に文法関連を中心とした統語情報アノテーションについて示す．3·2 節に局所的な意味情報である語義情報アノテーションについて示す．3·3 節に大域的な意味情報である文脈情報アノテーションについて示す．

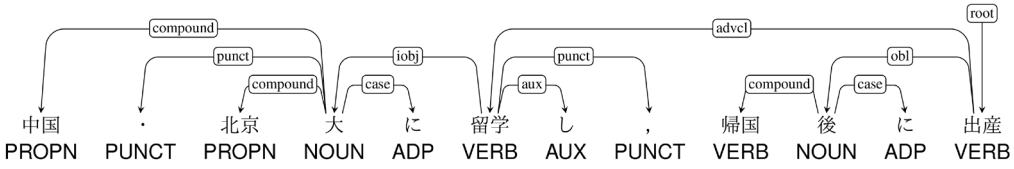
なお，本稿で示すアノテーションデータの多くは，BCCWJ の有償版契約者のみがアクセス可能なサイト[*21] から得ることができる．

3·1　統語情報アノテーション

以下ではアノテーションについて示す．

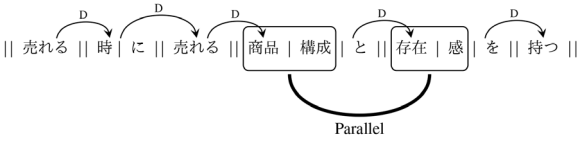
- 文節係り受け BCCWJ-DepPara [浅原 18]
- 単語係り受け Universal Dependencies [浅原 19]
- 述語項構造（表層格）BCCWJ-PAS [植田 15]
- 述語項構造（深層格）BCCWJ-PT [Takeuchi 15]

＊17 https://allennlp.org/elmo
＊18 2 文を入力とし，その 2 文が隣接するかどうかを判定する課題．
＊19 https://github.com/google-research/bert
＊20 https://github.com/google-research/bert/blob/master/extract_features.py を用いる．
＊21 https://chunagon.ninjal.ac.jp/ から認証後に BCCWJ の「データ配布」リンクをクリック．



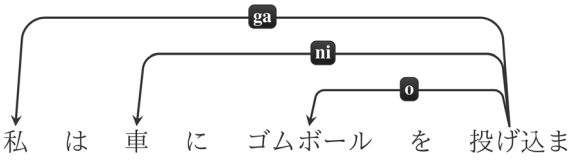
【BCCWJ: PN1c_00001 を一部改変】

図 1 単語係り受け UD Japanese BCCWJ



【BCCWJ: PB46_00066】

図 2 文節係り受け・並列構造アノテーション BCCWJ-DepPara



【BCCWJ:PM12_00020】

図 3 述語項構造 BCCWJ-PAS

●「れる・られる」用法 BCCWJ-AUX [小山田 12]
●助動詞用法 [加藤 19a]
●節境界 BCCWJ-ToriClause [Matsumoto 18]
●モダリティ BCCWJ-EME [松吉 10]
●否定の焦点 BCCWJ-NEG [松吉 14]

BCCWJ-DepPara [浅原 18] は文節単位の係り受け構造と並列構造を付与したものである．図 2 に例を示す．図中 ‖ が文節境界，| が（短単位）形態素境界を表す．【】内は BCCWJ のサンプル ID を表す．文節単位に係り受けを例文の上に矢印で示すとともに，形態素単位の並列構造の対応関係を例文の下に太い実線で示している．

Universal Dependencies（略称 UD）*22 は国際的な係り受けアノテーションの規格である．70 言語・100 以上のデータが整備・公開されている．言語横断的な自然言語処理で用いられるほか，言語類型論などの研究に用いることを目標として，各国でデータの整備が進められている．日本語においても整備が進められている．このうち．BCCWJ-DepPara の文節係り受けを単語単位の係り受けに変換した UD Japanese BCCWJ は世界で 3 番目の規模のデータである．図 1 に例を示す*23．単語の下に付与されているラベルは品詞情報である．矢印に付与されているラベルは係り受けのラベル情報である．

BCCWJ-PAS [植田 15] は表層格の述語項構造を付与したものである．具体的には名詞述語を含む述語相当句に，ガ格・ヲ格・ニ格相当の名詞句の情報を関係付けしたものである．図 3 に例を示す．述語項構造解析の一般的なタスク定義や具体的な解析手法については 5 章で概説する．BCCWJ-PAS は他に名詞句の外界照応・共参照の情報が付与されている．これについては後述する．

BCCWJ-PT [Takeuchi 15] は深層格の述語項構造を付与したものである*24．具体的には述語に対応する項に対して，動作主・対象などの 72 種類からなる意味役割情報を付与したものである．

格要素を付与するにあたって重要な情報として，受身などの格交代の助動詞の用法がある．BCCWJ のコアデータに対して，助動詞「れる・られる」の 4 用法（受身・尊敬・自発・可能）を付与したデータ [小山田 12] があるほか，コアデータの一部に対しては，使役を含むすべての助動詞の用法が付与されている [加藤 19a]．表 9 に，後に述べる分類語彙表ラベルアノテーションと助動詞の用法アノテーションデータの重ね合わせデータの例を示す．

係り受けは，重文・複文などの節間の関係を記述することが困難である．BCCWJ-ToriClause [Matsumoto 18] は，鳥取大学の鳥バンク*25 互換の節境界情報アノテーションを BCCWJ の新聞サンプル*26 に付与したものである．表表 10 に例を示す．最上位階層は補足節・名詞修飾節・副詞節・並列節の 4 種類に分かれており，それぞれ第 3 階層までの細分類が，節の意味に基づいて分類されている*27．

日本語はテンス（時制）・アスペクト（相）の表現よりもモダリティ（法性）が豊かな言語である．BCCWJEME は BCCWJ の一部に，モダリティ情報と事態の成立に関する情報を付与したものである [松吉 10]．具体的には，モダリティを表出する態度表明者・相対時・仮定された条件の有無・態度・真偽判断・価値判断などの情報が付与されている．さらにモダリティから発展して，BCCWJ-NEG は BCCWJ の新聞データに対して否定辞の焦点（スコープ）の情報を付与したものである [松吉 14]．

このように BCCWJ には重畳的に，文法に基づく統語

*22 https://universaldependencies.org/
*23 係り受け関係は，日本語においては従属部から主部方向へ矢印を書くが，国際的な通例としては主部から従属部方向へ矢印が書かれる．
*24 http://pth.cl.cs.okayama-u.ac.jp/
*25 池原悟氏による日本語重文と複文を対象とする意味類型パターン辞書．
*26 BCCWJ には新聞・書籍・雑誌などいくつかの媒体別に文章サンプルが収録されている．
*27 https://github.com/masayu-a/clause_pattern

表 9　助動詞の用法と BCCWJ-WLSP との重ね合わせデータ【BCCWJ:PM11_00055】

短単位書字形	語彙素番号	短単位分類語彙表	類	分類項目	機能語 / 内容語	助動詞	用　法
今	2460	1.1641	1：体	1641：関係–時間–現在	CONT		
に	28178		–	—-	FUNC		
も	37562		–	—-	FUNC		
舌なめずり	57139	1.3393	1：体	3393：活動–生活–口・鼻・目の動作	CONT		
し	19537	2.3430	2：用	3430：活動–行為–行為・活動	CONT		
そう	20934	3.1130	3：相	1130：関係–類–異同・類似	CONT	そうだ	様態
な	22916		–	—-	FUNC	だ	断定・指定
顔	5985	1.5601	1：体	5601：自然–身体–頭・目鼻・顔	CONT		
だっ	22916		–	—-	FUNC	だ	断定・指定
た	21642		–	—-	FUNC	た	過去（含経験）
.			–	—-	FUNC		

表 10　BCCWJ-ToriClause：節ラベルアノテーション【BCCWJ:PN3a_00003】

表層形	節ラベル
病院の	
収入が	
減る	MSd200：名詞修飾節：機能的表現：文末表現相当
わけで	FUb200：副詞節：因果関係：結果
,	
病院側が	
「入院日数を	
短縮しようと	FUf：副詞節：目的
努力すること	HSa100：補足節：名詞節：コト型
は	
間違いない」と	HSc100：補足節：引用節：直接引用

情報が付与されている．

3·2　語義情報アノテーション

本章では，BCCWJ に付与された，以下に示すさまざまな語義情報アノテーションデータについて解説する．具体的には以下の四つについて示す．

- ●岩波国語辞典語義アノテーション
- ●分類語彙表番号アノテーション BCCWJ-WLSP [加藤 19b]
- ●指標比喩情報アノテーション [Kikuchi 19]
- ●結合比喩情報アノテーション [Kato 19d]

BCCWJ には，先に述べた岩波国語辞典の語義の情報が付与されている [Okumura 11]．同データは，評価型ワークショップ SemEval-2010 [*28] の日本語全語義曖昧性解消（Japanese all words WSD）の訓練・評価データとして構築された．

BCCWJ-WLSP は，BCCWJ の新聞・書籍・雑誌データの一部に分類語彙表番号を悉皆付与したものである．先の岩波国語辞典と同様に all Words WSD の構成データとして用いることができる．岩波のデータと異なり，異なる見出し語間で，階層構造に基づく語義ラベル（分類語彙表の分類番号）を共有している．表 9 に例を示す．

語義のアノテーションにおいて，換喩・提喩を含む比喩表現が問題になる．機械処理においても，字義どおりからの語義の転換が行われるために重要な課題となる．現在，中村らの比喩分類 [中村 77] に基づいて，分類語彙表番号を手掛かりとした指標比喩（直喩）情報アノテーション [Kikuchi 19] や結合比喩（隠喩）情報アノテーション [Kato 19d] が進められている．表 11 に例を示す．

3·3　文脈情報アノテーション

統語・語義の情報のほかに，文脈情報についても整備されている．以下では，次のアノテーションについて解説する：

- ●共参照情報 BCCWJ-PAS [植田 15]
- ●外界照応情報 BCCWJ-PAS [植田 15]
- ●情報構造 BCCWJ-Infostr [宮内 18]
- ●時間的順序関係 BCCWJ-TimeBank

BCCWJ-PAS には，名詞句の共参照情報が付与されている．名詞句が指す実体が同一である場合に関係を付与する．図 4 に例を示す．下線が付与されているインデックス番号が同じ名詞句が，同じ実体を指す．共参照解析

＊28 http://semeval2.fbk.eu/semeval2.php?location=tasks#T15

表11 比喩情報アノテーション例

前文脈	比喩指標要素	後文脈
」とよばれ，うやまわれてきた．三原山の噴火が，観光客をよび，島を豊かにしてきたのだ．その三原山が，十五年前，五百年ぶりに大噴火をおこした．十カ所以上の火口から，溶岩がふきだしたのだ．流れ出た溶岩の跡が，幅百メートル，長さ三キロメートルもの大きさで，今も残っている．その姿は，	まるで	巨大な黒ヘビのようだ．ここを訪れた観光客は，当時のすさまじさにふれて，息をのむ．ゴツゴツとした黒い地面が広がり，ところどころにまっ黒にこげた木の切れはしが落ちている．火口だった場所は，すり鉢のように深くえぐれている．けれど，今はそこに緑の木がのび，時の流れを感じさせる．大噴火のとき，島民一万人全員を島の外に

比喩表現該当部			
流れ出た溶岩の跡が，幅百メートル，長さ三キロメートルもの大きさで，今も残っている．その姿は，まるで巨大な黒ヘビのようだ．			
被喩辞 流れ出た溶岩の跡	喩　辞 巨大な黒ヘビ	結　合 跡が蛇	備　考 まるで・よう
類型A 跡（1.1720）	付属語A ハ	類型B ヘビ（1.5503）	付属語B ノヨウ
種　別 擬　生		比喩指標要素 F-1-1・K-9-1	

★プレゼントはビートルズ$_1$

米CNNテレビのプロデューサー，ウェンディ・ウィットワースさん$_2$の50歳の誕生日，夫のラルフさん$_3$が“プレゼント”したのは何と，元ビートルズ$_1$のポール・マッカートニー$_4$のプライベートコンサート$_5$だった．

誕生パーティーで夫$_3$がポール$_4$の登場を告げると，彼$_4$の大ファンのウェンディさん$_2$は思わず涙．

約1時間半のコンサート$_5$の終幕近くには，ポール$_4$はウェンディさん$_2$をステージに上げ，ビートルズ$_1$の「バースデイ」を演奏したという．

ポール$_4$は，報酬の100万ドルを反地雷の慈善団体に寄付するとしている．

図4 BCCWJ-PAS：共参照情報アノテーション【BCCWJ：PN3b_00001】

今日は「後ろのヤツ」とはしゃべら$_{ガ:外界一人称}$んかった．まあ授業に集中し$_{ガ:外界一人称}$てましたからね．っていうよりは英語とプログラミングが移動なんでね．プロは考え$_{ガ:外界一人称}$てる暇すら与え$_{ニ:外界一人称}$てくれないよ．でも今日は指の調子が良かったので，早く打ち終わる$_{ガ:外界一人称,ニ:外界一般}$ことが出来$_{ガ:外界一人称}$ました．今日は部活があったけど1時半で帰れた$_{ガ:外界一人称,ニ:外界一般}$のよ〜ん．そいじゃ，まったあしたぁ．

図5 BCCWJ-PAS：外界照応情報アノテーション【BCCWJ：OY08_00189】

の一般的なタスク定義や具体的な解析手法については6章に概説しているので，そちらを参照されたい．

日本語は省略が多い言語である．BCCWJ-PASは，述語項構造をアノテーションする際に，名詞句が陽にあらわれず外界の実体を指す場合に，述語の項構造として外界照応情報を付与している．図5に例を示す．外界一人称は指示対象が筆者であることを表す．外界一般は指示対象がテキスト中に言及されていない筆者もしくは読者以外の何かであることを表す．ほかに外界二人称（指示対象が読者）というラベルがある．

BCCWJ-Infostrは，BCCWJの新聞データに名詞句の情報構造を付与したものである[宮内 18]．書き手の観点からの情報の新旧である情報状態と読み手の観点からの情報の新旧である共有性が付与されている．また関連する情報として，翻訳時の冠詞選択に影響を与える定性や特定性の情報が付与されている．

BCCWJ-TimeBankは，BCCWJの新聞データに対して事態が起こった時間的順序関係を付与したものである[Asahara 13]．英語における時間情報アノテーション基準TimeML [Pustejovsky 03]に準拠し，Allenの範囲代数に基づいて，二つの事態の時間的順序関係が付与されている．

4. 言語資源の構築手法

4·1 作 業 環 境

言語資源を構築する環境は付与する情報の構造に基づいて準備する必要がある．以下では典型的な作業過程と必要な作業環境構築について示す．

文字列に対する範囲選択・ラベル付与：多くのアノテーションは文字列の範囲選択とラベル付けによる．古くはXMLエディタなどにより，マウスドラッグなどによる範囲選択が行われてきた．ラベルの統制はXMLのDTDやXML Schemaにより行う．図6にXML EditoroXygen[*29]によるBCCWJ-TimeBankのタグ付け環境を示す．近年ではブラウザ上で範囲選択をするものもある．

単語列に対する範囲選択・ラベル付与：単語境界や句構造境界を前提として範囲選択する場合には，文字列に

＊29 https://www.oxygenxml.com/

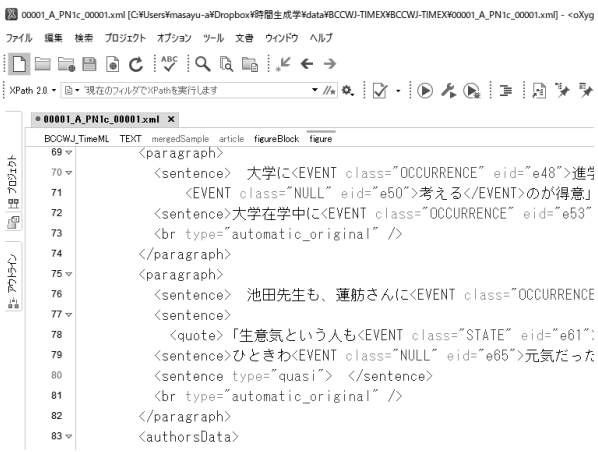


図 6　XML Editor oXygen による BCCWJ-TimeBank のタグ付け

2	B-DATE - Absolute or relative dates or periods.
5	I-DATE - Absolute or relative dates or periods.
日	I-DATE - Absolute or relative dates or periods.
も	
楽しみ	
に	
さ	
れ	
て	
ください	
。	

図 7　スプレッドシートによる範囲のタグ付け

対する範囲選択ではなく，スプレッドシートの表形式で範囲選択する．例えば，単語列に品詞情報などが付与されたものを見ながら，可能なラベルをスプレッドシート上のリストを構築し，リストから選択してアノテーションを進める．複数の単語からなる場合には，ラベル名に “B”（要素の開始位置）・“I”（要素の内部）を付加して範囲指定する．図 7 にスプレッドシートによる範囲のタグ付け例を示す．この例では「25 日」が日付表現であることを表すために，日付表現の開始位置である ”B-DATE” と日付表現の内部である “I-DATE” が付与される＊30．

関係付与：述語項構造・共参照など，複数の範囲の関係を付与するタスクにおいては，二つ以上の範囲選択を行うような GUI が必要である．有向・無向関係，推移律など関係の制約に基づいた作業環境が用いられる．

係り受け木：係り受け木は，単語間もしくは文節間の関係付与の拡張であるが，閉路がない，もしくは平面グラフ上で辺が交差しないなど，アノテーション基準上の制約がある場合が多い．このため，基準に応じて独自の GUI を構成する場合が多い．図 8 に，コーパス管理システム ChaKi.NET による BCCWJ-DepPara と BCCWJ-PAS の重ね合わせたデータの可視化を示す．左下に文節係り受け関係が，右上に述語項構造関係が確認できる．矢印をドラッグすることで，係り受け関係や述語項構造関係の修正が可能である．

＊30　実作業においてはスプレッドシートの色付け作業で行い，作業後に機械可読形式の B，I による範囲認定に変換することもある．

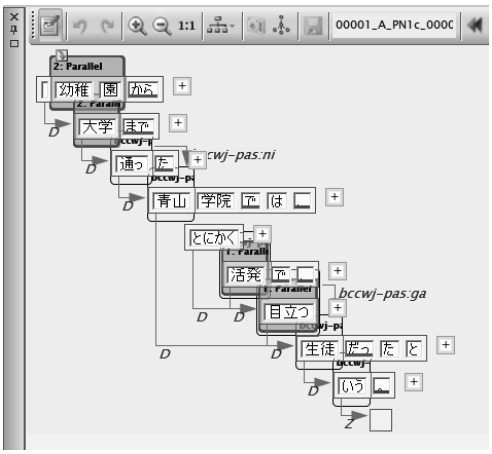


図 8　ChaKi.NET による BCCWJ-DepParaPAS のタグ付け

4·2　基準の策定

現実社会の課題に対処するためには，どういった問題を解決するのかを定義しながら仕様・ガイドラインを作成し，実作業をしながら修正するという作業を行う必要がある．この基準の策定について，言語学的な知識を有するデータ構築者と機械学習の知識を有するモデル構築者の協働作業フローが提案されている．

MATTER サイクル [Pustejovsky 06] は，機械学習に基づく解析器の構成を前提としたアノテーション整備手法である．次に示す作業を繰り返し実施することによりアノテーション基準を策定するとともに，アノテーションを構築する．

（1）Model Guideline：アノテーション仕様を認定し，ガイドラインを作成する
（2）Annotate：実際にアノテーションを行う
（3）Train：アノテーションに基づいて，機械学習に基づく解析器を訓練する
（4）Test：アノテーションに基づいて，機械学習に基づく解析器をテストする
（5）Evaluate：解析器の結果に基づきアノテーションを評価する
（6）Revise：アノテーションを修正する

図 9 に MATTER サイクルを示す．

機械学習に基づく解析器を構成する前の段階では，Train と Test のステップを飛ばして，アノテーションを進める．これを MAMA サイクルと呼ぶ．

他のアノテーションサイクルとして CASCADES サイクル [Bunt 13] がある．これはアノテーションを形式言語としてみた場合に，その意味論を与えるものである．主に 4·1 節に示したような定型的な作業環境では形式化

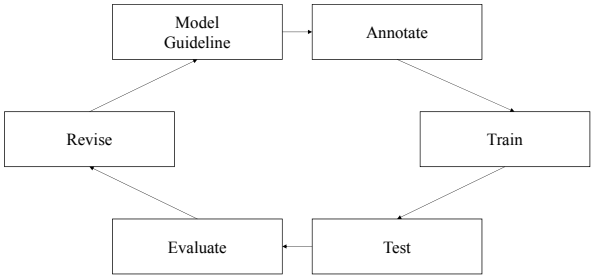


図 9 MATTER サイクル

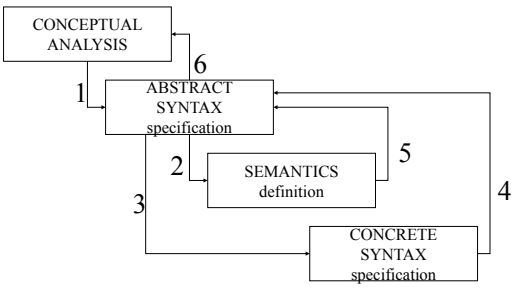


図 10 CASCADES サイクル

困難な，文脈を見るようなアノテーションに用いられ，以下の四つのステップからなる．

（1）Conceptual Analysis：アノテーションの問題として捉えるべき情報の概念について形式化する．

（2）Abstract Syntax：基本的な概念のリストの形で，アノテーションで表現する概念を明確化する．より形式的な仕様を集合論的な構造で形式化する．

（3）Semantics：Abstract Syntax による構造に対して形式意味論に基づく計算論的な仕様を与える．

（4）Concrete Syntax：Abstract Syntax によって定義される構造に対して，どのような表現を与えるかを仕様策定する．特に仕様付けされる表現形式が完全であるかについて検討する．

図 10 に CASCADES サイクルを示す．図中の矢印に付与された番号の順に検討が進められる．

以上，さまざまなアノテーションの作業サイクルについて述べた．言語は離散的な構造で表現するために，何を同一視し，何を区別するかといった判断が求められる．アノテーション自身も形式言語であるため，上記サイクルでは，形式言語としての健全性と完全性の観点から検討をする．さらに，作業者が判断可能かだけでなく，現在の技術で識別可能かといった観点も，実応用においては考慮の対象となる．

4·3 クラウドソーシングの利用

近年では短期間で大量にデータを構築するためにクラウドソーシングが用いられる．具体的にはマイクロタスク型クラウドソーシングと呼ばれる 1 作業当たり数円〜数十円程度の謝金を支払い，アノテーション作業を大人数の作業者に依頼する．クラウドソーシングでは，専門家でない方が作業を行うために，作業基準をわかりやすくすることが求められる．また，作業の統制が困難なために，ワーカを選別するためのタスクが行われることが多い．

統語や語義の捉え方は人によって異なるために単一のラベル付与が困難である．しかしながら，統語・意味レベルのアノテーションにおいては，構造の整合性が問われる．複数のアノテータにおける異なる構造をどのように一つの構造に集約するかといった検討が行われている [Nguyen 17, Paun 19]．一方で，統語・意味レベルの構造は，本質的に画一的な構造に集約できないとし，作業者間の揺れ（分散）を考慮したうえでそれをモデル化するという考え方もある [Asahara 19]．教師ありの識別学習で認定するにあたっては単一の正解の構造が必要になるが，認知科学・社会言語学的な分析においては，一般的な言語運用者の総体的なアノテーションが求められる．そういった状況においては，質問紙調査的にアノテーションの分布を収集することが行われる．

5. 意味役割解析

文の意味を構造的に解きほぐして理解する技術の一つとして，意味役割解析，述語項構造解析という技術が知られている．述語項構造とは，文章内の述語とその項の間の関係を規定する構造である．例えば図 11 の文では，「尋ねる」や「答える」という表現が述語であり，「記者団」や「首相」という表現が述語「尋ねる」の項である．このように文章中の要素を述語との関係によって構造的に整理することで，複雑な構造をもった文章において「誰が，何を，どうした」のような文章理解に重要な情報を抽出することができる．またここでは，文法上の主語・目的語関係だけではなく解析対象の述語と意味上のつながりのある語句を取り出すことが解析の目的となっている．例えば，日本語のように省略のある言語では，図 11 のように明示的な格関係（ガ，ヲ，ニなどの格助詞で表されるような文法関係）を伴っていない「首相に尋ねる」や「記者団に答える」のような意味関係を捉えることは極めて重要であり，このような関係を解析の対象としているのが意味役割・述語項構造解析である．

項構造解析は意味処理における基礎解析に分類され，ここで取り出された結果は文章内容の分析に直接利用することができる．また，要約や翻訳など，より下流の処理内容を調整する用途で用いられることもある．例えば要約では，項構造解析により主要な項と周辺的な情報を

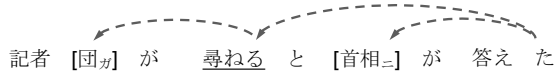


図 11 述語項構造解析の例．
下線は説明のために現在注目している述語．かぎ括弧は述語の項の主辞．下付き文字は格関係．破線は統語係り受け関係

与える項を取り出し，不要な項を削除することで主要な意味内容を保持したまま文章量を削減することができる [Hardy 18, Liao 18]．また，翻訳では，日本語で頻出する省略の内容を正しく特定することで英語への翻訳時に必要となる代名詞の種別を決定するといった使われ方をする場合がある [Kudo 14]．

5·1　意味役割付与コーパス

述語に関する意味構造の解析を典型的な教師あり学習の枠組みで行おうとすれば，我々はまず文中の語がそれぞれどのような意味的な構造をもつのかをあらかじめ理解する必要がある．文にこのような語の間の意味関係を付与し分析を加えたデータとして，英語では FrameNet [Ruppenhofer 06]，PropBank [Kingsbury 02]，OntoNotes [Hovy 06]，Abstract Meaning Representation [Banarescu 13]，日本語では京都大学テキストコーパス [黒橋 97]，NAIST テキストコーパス [飯田 10]，BCCWJ-PAS [植田 15]，京都大学ウェブ文書リードコーパス [Hangyo 12] などが知られている．

FrameNet，PropBank，OntoNotes の三つのコーパスでは，文中の語（主に述語）が**フレーム**と呼ばれる特定の項構造をもつという考えに基づいて語の間の意味関係が規定されている．例えば，図 12 の例のように FrameNet，PropBank は動詞 *sell*（の特定の語義）にそれぞれ述語の項構造を定義している．このフレームはそれぞれのコーパスで独自の名前をもっており，その内部構造として特定の意味的な役割をもったいくつかのスロット（意味役割スロット）をもっている．この意味役割スロットは，一般に文中に存在する語句で埋められるものである．これらのコーパスでは，コーパス内に出現する述語に対して語義ごとにフレームが定義され，それが辞書のような形で編纂されている．このフレーム辞書をもとに，図 13 のように実際の文にアノテーションが行われている．英語のコーパスが意味役割を中心としたアノテーションを行っているのに対して，日本語のコーパスでは図 11 にあげたように，ガ，ヲ，ニといった，格助詞の表層的な記号をそのまま意味ラベルとして利用するアノテーションの方法が主に用いられている．この違いが現れた理由としては，日本語においては，述語に対して役割をもつ語（項と呼ばれる）の省略が頻繁に起こるという性質のほか，副助詞「は」，「も」などが使用されている場合や，連体修飾の関係にある場合など，構文上の係り受け情報だけからでは表層的な格関係自体が自明でない場合があるため，述語とその項の位置関係や表層的な格関係を明らかにすることそのものが意味解析における第一の目標とされたことがある．データの規模としては，ベンチーマークデータとして広く利用されている PropBank，NAIST テキストコーパスの文数がそれぞれ約 44 000 文，約 40 000 文であり，項構造解析の学習・評価用データとしては同程度の量が妥当と見られている．

sell.v	PropBank	FrameNet
Frame	sell.01	Commerce_sell
Roles	0 (Seller)	Seller
	1 (Thing Sold)	Goods
	2 (Buyer)	Buyer
	3 (Price Paid)	Money
	4 (Benefactive)	Recipient
		...

図 12　PropBank と FrameNet における動詞 *sell* に対するフレーム定義

5·2　最新のモデル

自然言語処理の多くのタスクで end-to-end の解析手法が成功したことを受けて，意味役割解析の最新のモデルでも同様の手法が取られている．

このタスクでは，形式的な入力として，単語区切りが与えられた文 $\mathrm{w} = w_1, ..., w_n$ と，事前に特定した述語の位置 $\mathrm{p} = p_1, ..., p_q$ が与えられる．解析器は，与えられたそれぞれの述語に対し，その述語に対して意味的な関係をもつ文中の区間を特定し，意味的関係の種類を表すラ

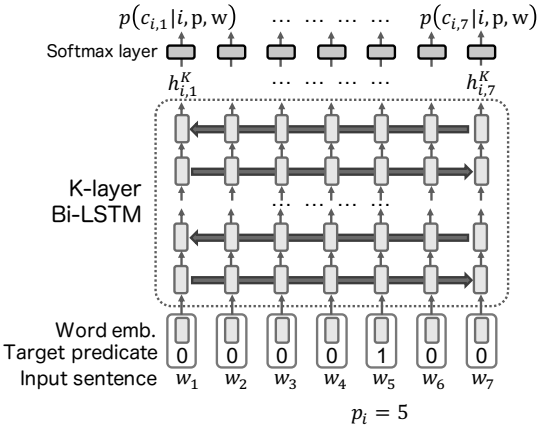


図 14　多層双方向 LSTM を用いた標準的な意味役割解析モデル

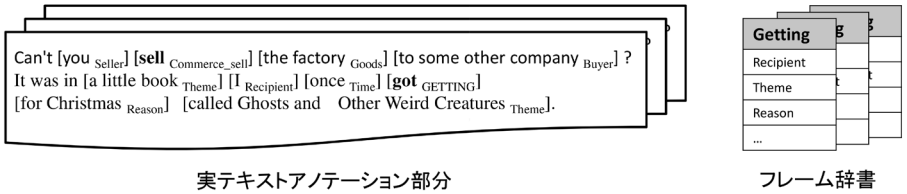


図 13　意味役割付与コーパスの概要図

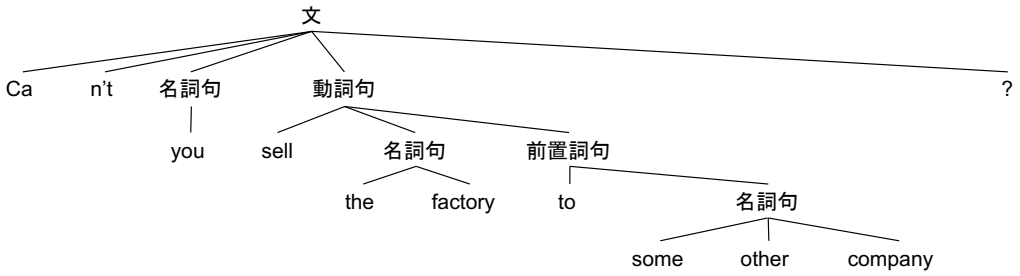


図 15　句構造文法に基づく文の構文構造分析の例

ベル（意味役割ラベル）とともに出力する．この区間特定問題は，文中のそれぞれの単語に対して意味役割名付きの BIO ラベル*31 を割り当てる問題として定式化される．

最新の手法では，解析モデルは典型的に多層双方向 LSTM もしくは Transformer 機構 [Vaswani 17] を用いた end-to-end のニューラル解析モデルとして実現される [He 17, Ouchi 18, Zhou 15]．図 14 にその一例を示す．このモデルは入力として，単語列 $\mathrm{w} = w_1, ..., w_n$ と，解析対象となる述語の位置 p_i を取り，入力された単語列に対応するラベルの確率 $p(c_{i,1}|i, \mathrm{p}, \mathrm{w}), ..., p(c_{i,n}|i, \mathrm{p}, \mathrm{w})$ を出力する．ここで $c_{i,t}$ は述語 w_{p_i} に対する単語 w_t の意味役割ラベルを表すものである．入力層では，入力の各単語を表現するベクトルと，解析対象の述語の位置をバイナリ値として表現したものが結合される．次に，これらのベクトル列が多層双方向 RNN に入力される．一般に，意味役割解析では奇数層を順方向，偶数層を逆方向とし，一方向の RNN の出力を逆方向の RNN の入力に用いる形式の双方向 RNN が用いられている．その後，多層双方向 RNN の最終層における各時刻 t の出力 $h^K_{i,t}$ を softmax 層で確率 $p(c_{i,t}|i, \mathrm{p}, \mathrm{w})$ を表す多次元ベクトルに変換し，最も高い確率をもつラベルを述語 p_i に対する単語 w_t の役割ラベルとして出力する．最終的な確率値の出力はラベル間の依存関係を考慮して conditional random field でモデル化される場合もある．

5·3　意味解析における手掛かり

深層学習時代の意味役割解析モデルは極めてブラックボックス的であるが，それ以前の特徴量工学の時代には解析に利用するためのさまざまな特徴量が考案されてきた．中でもとりわけ重要視されていた特徴量として，大きく二つのものがある．一つは，述語とその項の間の文法上の関係を構文木上の位置関係として表現するものである．一般に，ある述語に対して意味的に関わりをもつ語句は，その述語に対して主語・目的語といった構文構造上の決められた位置に来ることが多い．例えば，図 15 は図 13 の最初の文 “Can't you sell the company to some other company?” の構文木を示したものであるが，この文の述語 *sell* に対して意味役割をもつ句は，それぞれ主語（動詞句の右にある名詞句），目的語（動詞句内の最初の名詞句），前置詞句の位置にある．このそれぞれの句の位置は，今注目している動詞 *sell* から “V↑VP↑S↓NP”（主語：you）や “V↑VP↓NP”（目的語：the factory）のように構文木を上下方向にたどることで構文構造上の相対位置として表現できる．このような構文木上の軌跡によって構造上の位置関係を表現した特徴量を「構文パス」と呼ぶ．

もう一つの重要な情報は，語句の共起に関する情報である．例えば通常動詞であるような語を使った複合名詞（例えば，理解不足＝理解が不足する，理解完了＝理解を完了する）や連体修飾句（例えば，食べる人＝人が食べる，食べるもの＝ものを食べる）などでは，構文構造による手掛かりの一部が失われてしまい，役割の曖昧性解消が行いづらくなってしまう．こういった場合に利用できるのは「ある語がある述語に対して特定の意味役割をもつもっともらしさ」＝「(語，述語，統語関係）のタプルの共起情報」である．この三つ組の共起情報は「述語の選択選好性」と呼ばれる．ただし，このような語の組合せに関する共起情報は，数百万語という規模の記号をもつ言語において，たかだか 4 万文の学習データから統計を得ることが不可能なため，一般には数百万～数十億文といったより大規模なデータであらかじめその統計値を計算する方法が取られる．選択選好は，項の省略現象が頻出する日本語では，省略された語の予測にとって特に重要な情報として知られており，統計値の獲得手法においてもさまざまな研究がなされてきた [Iida 07, Imamura 09, 小町 10, Sasano 11]．先の 2·2 節で述べた単語埋込みの手法もこうした単語共起情報の獲得手法の一つといえる．

深層学習による解析手法が取られるようになった現在においても，これらの情報を明示的に利用することで解析精度を向上させる試みは続いている．例えば，[Roth 16, Shwartz 16] では，LSTM を利用して統語パスの各中継点に位置する単語の情報（のベクトル表現）を順に埋め込んでいくことによって，統語パス全体を一つのベクトル表現として埋め込む手法が提案され，この表現を

*31　意味役割区間の先頭の語に「B- 役割名」，区間が継続する語に「I- 役割名」，意味役割区間の外側の語に「O」というラベルを割り当てる方式．

用いることにより，意味役割解析や名詞間の意味的関係を抽出するタスクの解析精度が向上することが確認されている．また [He 17] では，図 14 のような end-to-end の解析器の出力に対して，そのラベルの区間が構文構造上の句や節と一致するように制約を明示的に設けることで性能の改善が見られることも報告されている．[Shibata 16] は，大規模な付加情報なしのコーパス（生コーパス）に構文解析を走らせて得た述語項構造を用い，項構造内の一つの項を他の項から予測する確率モデルを学習し，これをニューラルネットワークベースのモデルに組み込むことによって項構造の解析精度を向上させることができることを示している．さらに，これらの従来有効とされてきた特徴を深層学習モデルの内部パラメータが同様に捉えているかを検証する試みも見られる．

このように，過去に発見されてきた有効な特徴量をどのようにして自然な形で深層学習モデルの中に組み込むかという試行錯誤や，文内部の構文的・意味的な構造を深層学習モデルが妥当な形で理解できているかを検証する試みは，ブラックボックス的な学習が行われる end-to-end モデルを真の意味で言語を理解できるモデルに近づけていくうえで重要だと考えられている．

6. 共参照・照応解析

文章では，そこで主題にしている人物や事柄，テーマについて何度も言及されることが一般的である．読み手はこのような同一の人物・ものに対する言及をまとめ上げることで文章全体の論旨やストーリーを理解することができる．しかし，これらの言及の中には同一のものを指すのに「総理大臣」と「首相」のように異なる表現を用いる場合がある．また，「彼」のように代名詞によって前方の表現を肩代わりさせる場合もある．このような表面上異なる文字列で同じ物事を指す表現をまとめる処理として，共参照・照応解析と呼ばれる技術がある．

共参照解析は，文章中で同一のものやこと，人物を指している表現を同定するタスクである．例えば次の文，

(1) ［木村氏$_i$］は稽古で［彼$_i$］の弟子に対して実技の披露を行った．

では，「木村氏」と「彼」は同一人物である．共参照解析はこのような表現をクラスタとしてまとめ上げる．

照応解析は，代名詞など，文章中にある他の言語表現に依存して指示するものが変わる要素（照応詞）の指示先（先行詞）を特定するタスクである．共参照解析と照応解析は，その定義から一部の問題を共有している．例えば，上の例の「木村氏」と「彼」という表現の間の共参照関係は，代名詞の照応解析ともみなすことができる．一方で，次の例は共参照関係と照応関係の違いを明確に説明する．

(2) 太郎は［車$_i$］を買った．次郎も［それ$_j$］を買った．

この例の場合，「車」と「それ」は照応関係にあるが，共参照関係にはない．こうした区別はあるものの，共参照・照応解析はいずれも文章中の表現が実世界のどのような物事を指し示すかということがその興味の対象となる．ただし日本語の場合，共参照・照応解析のうち，省略された内容を対象とするものに関しては空項解析あるいはゼロ照応解析と呼ばれており，慣例的に述語項構造解析の範ちゅうで解かれるのが一般的である．

[Ng 10] の分類に基づけば，共参照・照応関係の解析手法は大別して三つに分けられる．一つ目は，文章中の任意の二つの表現について，それらが共参照関係にあるかを 2 値分類する手法で，mention-pair モデルと呼ばれる [Bengtson 08, Ng 02]．二つ目は，現在解析対象とする表現に先行する任意の表現について明示的に関連度を順位付けする手法で，mention-ranking モデルと呼ばれる [Clark 16, Lee 17, Zhang 19, Wiseman 15]．三つ目は，解析を終えた表現どうしをまとめてクラスタをつくっておき，新たに解析対象となる表現はこれらのいずれかのクラスタ（ただし，自分自身を唯一の要素とする新しいクラスタを含む）に分類するという手法であり，entity-mention モデルと呼ばれる [Clark 15, Haghighi 10, Wiseman 16]．

共参照・照応関係が付与されたコーパスとして代表的なものとして，英語では MUCdataset[*32]，ACE dataset[*33]，OntoNotes [Hovy 06]，GENIA corpus [Ohta 02]，日本語では述語項構造コーパスと同様の京都大学テキストコーパス，NAIST テキストコーパス，BCCWJ-PAS，京都大学ウェブ文書リードコーパスなどが知られている．

6·1 最新のモデル

最新の代表的な手法として，[Lee 17] や [Zhang 19] で採用されているモデルを紹介する（図 16）．このタスクでは，入力として，単語区切りされた T 語からなる文章 D が与えられる．ここで文章中のあらゆる単語区間を考えるとその数は

$$N = \frac{T(T+1)}{2}$$

である．この区間を先頭位置，末尾位置の順で優先順位をつけて昇順に並べたものを $\{s_1, ..., s_N\}$ とすると，共参照解析タスクは，それぞれの区間 s_i に対し，$Y(i) = \{s_1, ..., s_{i-1}, \varepsilon\}$ の中から対応する先行詞を一つ選択する分類問題として定式化される．ここで ε は，対応する先行詞がないことを表す擬似候補である．

共参照・照応解析においても，深層学習ベースの手法の多くでは文章内の各語の文脈付き表現を学習する

*32 https://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html
*33 https://catalog.ldc.upenn.edu/LDC2014T18

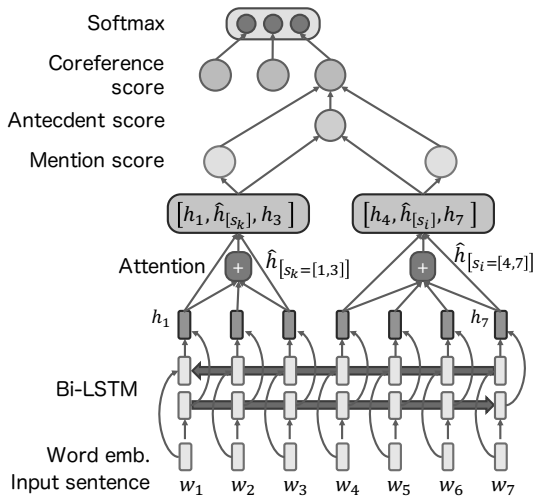


図 16 [Lee 17, Zhang 19] で採用されている共参照解析モデル

ために意味役割解析の 5 章で示したのと同様の双方向 LSTM が典型的に用いられる．LSTM によって各語のベクトルが文脈を考慮した中間表現に変更されると，次に，それらの変更されたベクトルの重み付き和によって文の特定区間の表現を構成的に計算する．重みの計算にはこれらのベクトルをスカラ値に写像した後にそれらの値の softmax を取る，注意機構 [Bahdanau 15] と呼ばれる機構が用いられる．ここで各語に対する中間表現の softmax を取る意味は，言語の性質として，ある単語区間の意味を中心的に支配する「主辞」と呼ばれる語が存在する，という一般的事実をモデル化していると考えることができる．次に，ある区間 s_i とその先行詞となり得る区間 s_k（$k < i$）のペアに対して，計算した区間表現のペアを用いて，それらの区間が同一の対象を指しているかをスコア化する（coreference score）．この際，それぞれの区間表現自身が実世界の何かを支持する表現になっているか（mention score）と，二つの表現が同じ意味を表していそうか（antecedent score）という二つの意図をもったスコアを計算し，この組合せで最終的な coreference score を計算する．最後に，すべての先行詞候補との間で算出した coreference score について softmax を取り，最もスコアの高かったものを先行詞として選択する．ただし，一般にすべての区間を先行詞候補とすると N の数が大きくなるため，あらかじめ mention score の低い区間を解析から除外する方法が取られる．

6·2 共参照・照応解析の手掛かり

共参照・照応解析に用いられてきた特徴量は，大別して四つのグループに分けられる．一つ目は，解析対象の表現と先行詞の文法的特性を取り出すもので，単数形・複数形などの語の文法上の特徴の一致や，構文構造上で同一の文法機能（主語・目的語など）をもっているか，構文情報から計算される表現間の距離などさまざまなものが用いられてきた．二つ目は，解析対象表現の意味的な性質を捉えるもので，周辺に出現する動詞や文脈語の類似度，意味役割解析で用いられていたのと同様の語の共起に関する情報，オントロジーやシソーラスを用いて計算する単語間の意味の類縁性や包含関係などを考慮に入れる．三つ目は，文脈情報を扱うもので，解析対象の表現と先行詞候補の間の距離や，間に挟まれている対立候補の数，それぞれの候補が文脈上どれくらい際立っているかに関するスコアを算出して用いる方法が取られた．四つ目は，Wikipedia などの百科事典的な資源から世界知識を獲得し，これを解析に利用するものである．これは，特に地名，人名，会社名などの固有名詞に関して，文章内から得られる情報に加えて，言い換えや，強く関連する語を獲得する方法として使われた．

このように，深層学習以前の共参照・照応解析においては非常に多彩な情報が解析の特徴量として用いられてきたが，現在の end-to-end の解析手法においてはこれらのほとんどは用いられていない．一方で，オントロジーや世界知識の利用など一部の知識については，意味役割解析の 5 章で言及したのと同様に，深層学習モデルへの自然な統合方法を開発する研究の方向性は十分に考えられる．実際に，[Zhang 19] では，深層学習モデルの出力に加えて，知識データベースを検索して得た関連知識を埋め込んだ表現を考慮して解析を行うことで，固有名詞に関する共参照解析の精度が向上することが示されている．今後，このような深層学習モデルへの過去の研究遺産の統合が進み，多様な知識を取り込んだモデルの開発が進んでいくと期待される．

7. お わ り に

本稿では，深層学習全盛期の現在に紹介しておくべき自然言語処理の基礎的な知識として，自然言語処理を支える言語資源データと意味解析に関する基礎技術について解説した．言語資源は，言語の語彙や統語に関する豊富な知識を与えるものであり，また，言語解析の出力に関して扱いに優れた汎用的な表現形式を与えるものであった．自然語処理技術を使った新しいタスクを構築する際は，ここで述べたテキストアノテーションのサイクルを辿って教師データを構築することにより，問題の定式化や正確な教師データの作成を効率的かつ厳格に行うことができる．5 章の意味解析では，言語データから汎用性の高い基本的な構造を取り出すための基礎解析技術の例として，意味役割解析と照応・共参照解析を取り上げた．こうした基礎解析の出力結果は，テキストデータを対象とした既存の情報処理システムに新たに自然言語処理の機能を取り込む際のコンポーネントの単位として使い勝手が良い．また，最新のモデルと過去に培われてきた解析のための主要な特徴量の対比において，ブラックボックス的な学習が行われる end-to-end モデルの検

証や改善の手掛かりとして，これまでの特徴量工学で培われてきた知見が生かされている例を示した．現在主流となっている深層学習によるモデル開発環境においては，これまで自然言語処理分野で構築されてきた言語資源や，基礎的な解析レイヤにおける解析の知識が，個別の応用技術を構築・改善するうえでの有用な手掛かりとなるだろう．

謝　辞

本稿で紹介した研究のうち，著者らがその研究に従事したものの一部は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものと，JSPS 科研費挑戦的研究（萌芽）18K18519，基盤研究（A）17H00917，新学術領域研究 18H05521，基盤研究（C）18K00634，19K00591，19K00655 の助成を受けたものです．

◇ 参 考 文 献 ◇

[Asahara 13] Asahara, M., Yasuda, S., Konishi, H., Imada, M. and Maekawa, K.: BCCWJ-TimeBank: Temporal and event information annotation on Japanese text, *Proc. 27th Pacific Asia Conf. on Language, Information, and Computation* (*PACLIC 27*), pp. 206-214（2013）

[Asahara 18] Asahara, M.: NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus', *Terminology: Int. Journal of Theoretical and Applied Issues in Specialized Communication,* Vol. 24, No. 2, pp. 7-25（2018）

[Asahara 19] Asahara, M.: Word familiarity rate estimation by bayesian linear mixed model, *Proc. Aggregating and Analysing Crowdsourced Annotations for NLP*（*AnnoNLP*）, pp. 6-14（2019）

[浅原 18] 浅原正幸，松本裕治:『現代日本語書き言葉均衡コーパス』に対する文節係り受け･並列構造アノテーション，自然言語処理，Vol. 25, No. 4, pp. 331-356（2018）

[浅原 19] 浅原正幸, 金山 博, 宮尾祐介, 田中貴秋, 大村 舞, 村脇有吾, 松本裕治：Universal Dependencies 日本語コーパス，自然言語処理，Vol. 26, No. 1, pp. 3-36（2019）

[Bahdanau 15] Bahdanau, D., Cho, K. and Bengio, Y.: Neural machine translation by jointly learning to align and translate, *3rd Int. Conf. on Learning Representations*（*ICLR*）, pp.1-15（2015）

[Banarescu 13] Banarescu, L., Bonial, C., Cai, S., Georgescu, M.,Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N.: Abstract meaning representation for sembanking, *Proc. 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178-186（2013）

[Bengtson 08] Bengtson, E. and Roth, D.: Understanding the value of features for coreference resolution, *Proc. 2008 Conf. on Empirical Methods in Natural Language Processing,* pp. 294-303（2008）

[Blei 03] Blei, D., Ng, A. and Jordan, M.: Latent dirichlet allocation, *J. of Machine Learning Research,* pp. 1107-1135（2003）

[Bojanowski 17] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching word vectors with subword information, *Trans. of the Association for Computational Linguistics,* Vol. 5, pp. 135-146（2017）

[Bunt 13] Bunt, H.: A methodology for designing semantic annotations, Technical Report, TiCC TR 2013-001, Tilburg University（2013）

[Cho 14] Cho, K., Merriënboer, van B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing*（*EMNLP*）, pp. 1724-1734（2014）

[Clark 15] Clark, K. and Manning, C. D.: Entity-Centric Coreference Resolution with Model Stacking, *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*（*Vol. 1: Long Papers*）, pp. 1405-1415（2015）

[Clark 16] Clark, K. and Manning, C. D.: Deep reinforcement learning for mention-ranking coreference models, *Proc. 2016 Conf. on Empirical Methods in Natural Language Processing,* pp. 2256-2262（2016）

[Collobert 11] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P.: Natural language processing（almost）from scratch, *J. Machine Learning Research,* Vol. 12, pp. 2493-2537（2011）

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*（*Long and Short Papers*）, pp. 4171-4186（2019）

[Grave 18] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. and Mikolov, T.: Learning word vectors for 157 languages, *Proc. 11th Int. Conf. on Language Resources and Evaluation*（*LREC 2018*）, pp. 3483-3487（2018）

[Haghighi 10] Haghighi, A. and Klein, D.: Coreference resolution in a modular, entity-centered model, *Human Language Technologies: 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics,* pp. 385-393（2010）

[Hangyo 12] Hangyo, M., Kawahara, D. and Kurohashi, S.: Building a diverse document leads corpus annotated with semantic relations, *Proc. 26th Pacific Asia Conf. on Language, Information, and Computation,* pp. 535-544（2012）

[Hardy 18] Hardy, H. and Vlachos, A.: Guided neural language generation for abstractive summarization using abstract meaning representation, *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing,* pp. 768-773（2018）

[He 17] He, L., Lee, K., Lewis, M. and Zettlemoyer, L.: Deep semantic role labeling: Whatworks and what's next, *Proc. 55th Annual Meeting of the Association for Computational Linguistics*（*Vol. 1: Long Papers*）, pp. 473-483（2017）

[Hofmann 99] Hofmann, T.: Probabilistic latent semantic indexing, *Proc. 15th Conference on Uncertainty in Artificial Inteligence,* pp. 289-296（1999）

[Hovy 06] Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. and Weischedel, R.: OntoNotes: The 90% Solution, *Proc. Human Language Technology Conf. of the NAACL, Companion Volume: Short Papers,* pp. 57-60（2006）

[Iida 07] Iida, R., Inui, K. and Matsumoto, Y.: Zero-anaphora resolution by learning rich syntactic pattern features, *ACM Trans. on Asian Language Information Processing,* Vol. 6, No. 4, pp. 1:1-1:22（2007）

[飯田 10] 飯田 龍，小町 守，井之上直也，乾健太郎，松本裕治：述語項構造と照応関係のアノテーション：NAIST テキストコーパス構築の経験から，自然言語処理，Vol. 17, No. 2, pp. 25-50（2010）

[Imamura 09] Imamura, K., Saito, K. and Izumi, T.: Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution, *Proc. ACL-IJCNLP 2009 Conf. Short Papers,* pp. 85-88（2009）

[加藤 19a] 加藤 祥，浅原正幸，山崎 誠：『現代日本語書き言葉均衡コーパス』新聞・書籍・雑誌データの助動詞に対する用法情報付与，日本語学会 2019 年度春季大会予稿集（2019）

[加藤 19b] 加藤 祥，浅原正幸，山崎 誠：分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ，日本語の研究，Vol. 15, No. 2, pp. 134-141（2019）

[加藤 19c] 加藤 祥，田邊 絢，浅原正幸，古宮嘉那子，新納浩幸：

多義語の語義分布と語義間の派生関係調査の試み—相の類を中心に，言語処理学会第 25 回年次大会発表論文集，pp. 347-350（2019）
[Kato 19d] Kato, S. and Asahara, M.: Exploring metaphorical expressions in Japanese Newspaper-Article Corpora, *15th Int. Cognitive Linguistics Conference*（*ICLC-15*）（2019）
[Kikuchi 19] Kikuchi, R., Kato, S. and Asahara, M.: Collecting figurative expressions using indicators and semantic tagged Japanese corpus, *15th Int. Cognitive Linguistics Conference*（*ICLC-15*）（2019）
[Kikuta 19] Kikuta, Y.: BERT Pretrained model trained on Japanese Wikipedia Articles, `https://github.com/yoheikikuta/bert-japanese`（2019）
[Kingsbury 02] Kingsbury, P. and Palmer, M.: From TreeBank to PropBank, *Proc. 3rd Int. Conf. on Language Resources and Evaluation*（*LREC*）, pp. 1989-1993（2002）
[国立 04] 国立国語研究所 編：国立国語研究所資料集 14『分類語彙表—増補改訂版—』，大日本図書（2004）
[小町 10] 小町 守，飯田龍，乾健太郎，松本裕治：名詞句の語彙統語パターンを用いた事態性名詞の項構造解析，自然言語処理，Vol. 17, No. 1, pp. 141-159（2010）
[近藤 20] 近藤明日子，田中牧郎：「分類語彙表番号— UniDic 語彙素番号対応表」の構築，国立国語研究所論集（2020）（to appear）
[Kudo 14] Kudo, T., Ichikawa, H. and Kazawa, H.: A joint inference of deep case analysis and zero subject generation for Japanese-to-English statistical machine translation, *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*（*Vol. 2: Short Papers*）, pp. 557-562（2014）
[黒橋 97] 黒橋禎夫，長尾 真：京都大学テキストコーパス・プロジェクト，人工知能学会第 11 回全国大会論文集，pp. 58-61（1997）
[Lee 17] Lee, K., He, L., Lewis, M. and Zettlemoyer, L.: End to-end neural coreference resolution, *Proc. 2017 Conf. on Empirical Methods in Natural Language Processing,* pp. 188-197（2017）
[Liao 18] Liao, K., Lebanoff, L. and Liu, F.: Abstract meaning representation for multi-document summarization, *Proc. 27th Int. Conf. on Computational Linguistics,* pp. 1178-1190（2018）
[Maekawa 14] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y.: Balanced corpus of contemporary written japanese, *Language Resources and Evaluation,* Vol. 48, pp. 345-371（2014）
[真鍋 19] 真鍋陽俊，岡 照晃，海川祥毅，髙岡一馬，内田佳孝，浅原正幸：複数粒度の分割結果に基づく日本語単語分散表現，言語処理学会第 25 回年次大会（NLP 2019）発表論文集，pp. 1407-1410（2019）
[Matsumoto 18] Matsumoto, S., Asahara, M. and Arita, S.: Japanese clause classification annotation on the 'Balanced Corpus of Contemporary Written Japanese', *Proc. Asian Language Resources 13*（*ALR13*）, pp. 1-8（2018）
[松吉 10] 松吉 俊，江口 萌，佐尾ちとせ，村上浩司，乾健太郎，松本裕治：テキスト情報分析のための判断情報アノテーション，信学論（D），Vol. J93-D, No. 6, pp. 705-713（2010）
[松吉 14] 松吉 俊：否定の焦点情報アノテーション，自然言語処理，Vol. 21, No. 2, pp. 249-270（2014）
[Mikolov 10] Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J. and Khudanpur, S.: Recurrent neural network based language model, *11th Annual Conf. of the Int. Speech Communication Association,* pp. 1045-1048（2010）
[Mikolov 13a] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *Proc. Workshops at ICLR,* pp. 1-12（2013）
[Mikolov 13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems 26*（*NIPS*）, pp. 3111-3119（2013）
[宮内 18] 宮内拓也，浅原正幸，中川奈津子，加藤 祥：『現代日本語書き言葉均衡コーパス』への情報構造アノテーションとその分析，国立国語研究所論集，Vol. 16, pp. 19-33（2018）
[中村 77] 中村 明：比喩表現の理論と分類，国立国語研究所報告，No. 57，秀英出版（1977）
[Ng 02] Ng, V. and Cardie, C.: Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution, *COLING 2002: 19th Int. Conf. on Computational Linguistics,* pp. 1-7（2002）
[Ng 10] Ng, V.: Supervised noun phrase coreference research: the first fifteen years, *Proc. 48th Annual Meeting of the Association for Computational Linguistics,* pp. 1396-1411（2010）
[Nguyen 17] Nguyen, A. T., Wallace, B., Li, J. J., Nenkova, A. and Lease, M.: Aggregating and predicting sequence labels from crowd annotations, *Proc. 55th Annual Meeting of the Association for Computational Linguistics*（*Vol. 1: Long Papers*）, pp. 299-309（2017）
[荻原 19] 荻原亜彩美，森山奈々美，浅原正幸，加藤 祥，山崎 誠：『分類語彙表』に対する反対語情報付与，言語処理学会第 25 回年次大会発表論文集，pp. 1061-1064（2019）
[Ohta 02] Ohta, T., Tateisi, Y. and Kim, J.-D.: The GENIA Corpus: An annotated research abstract corpus in molecular biology domain, *Proc. 2nd Int. Conf. on Human Language Technology Research, HLT'02,* pp. 82-86（2002）
[Okumura 11] Okumura, M., Shirai, K., Komiya, K. and Yokono, H.: On SemEval-2010 Japanese WSD Task, 自然言語処理，Vol. 18, No. 3, pp. 293-307（2011）
[小山田 12] 小山田由紀，柏野和佳子，前川喜久雄：助動詞レル・ラレルへの意味アノテーション作業経過報告，第 2 回コーパス日本語学ワークショップ予稿集，pp. 59-69（2012）
[Ouchi 18] Ouchi, H., Shindo, H. and Matsumoto, Y.: A span selection model for semantic role labeling, *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing,* pp. 1630-1642（2018）
[Paun 19] Paun, S. and Hovy, D., eds.: *Proc. 1st Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*（2019）
[Pennington 14] Pennington, J., Socher, R. and Manning, C.: Glove: Global vectors for word representation, *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing*（*EMNLP*）, pp. 1532-1543（2014）
[Peters 18] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep contextualized word representations, *Proc. NAACL*（2018）
[Pustejovsky 03] Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A. and Katz, G.: TimeML: Robust specification of event and temporal expressions in text, *Proc. 5th Int. Workshop on Computational Semantics*（*IWCS-5*）, pp. 337-353（2003）
[Pustejovsky 06] Pustejovsky, J.: Unifying Linguistic Annotations: A TIMEML case study, *Proc. Text, Speech, and Dialogue Conference*（2006）
[Roth 16] Roth, M. and Lapata, M.: Neural semantic role labeling with dependency path embeddings, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*（*Vol. 1: Long Papers*）, pp. 1192-1202（2016）
[Ruppenhofer 06] Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. and Scheffczyk, J.: FrameNet II: Extended theory and practice, *Berkeley FrameNet Release,* Vol. 1（2006）
[Sasano 11] Sasano, R. and Kurohashi, S.: A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames, *Proc. 5th Int. Joint Conf. on Natural Language Processing,* pp. 758-766（2011）
[Shibata 16] Shibata, T., Kawahara, D. and Kurohashi, S.: Neural network-based model for Japanese predicate argument structure analysis, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*（*Vol. 1: Long Papers*）, pp. 1235-1244（2016）
[柴田 19] 柴田知秀，河原大輔，黒橋禎夫：BERT による日本語構文解析の精度向上，言語処理学会第 25 回年次大会（NLP 2019）発表論文集，pp. 205-208（2019）
[Shwartz 16] Shwartz, V., Goldberg, Y. and Dagan, I.: Improving hypernymy detection with an integrated path-based and distributional method, *Proc. 54th Annual Meeting of the*

Association for Computational Linguistics（*Vol. 1: Long Papers*）, pp. 2389-2398（2016）
[Socher 11] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y. and Manning, C. D.: Semi-supervised recursive autoencoders for predicting sentiment distributions, *Proc. 2011 Conf. on Empirical Methods in Natural Language Processing,* pp. 151-161（2011）
[鈴木 16] 鈴木正敏，松田耕史，関根聡，岡崎直観，乾健太郎：Wikipedia 記事に対する拡張固有表現ラベルの多重付与，言語処理学会第 22 回年次大会（NLP 2016）発表論文集，pp. 797-800（2016）
[田口 17] 田口雄哉，田森秀明，人見雄太，西鳥羽二郎，菊田洸：同義語を考慮した日本語単語分散表現の学習，情処学研報，第 233 回自然言語処理研究会，2017-NL-233，pp. 1-5（2017）
[Takeuchi 15] Takeuchi, K., Ueno, M. and Takeuchi, N.: Annotating semantic role information to Japanese balanced corpus, *Proc. MAPLEX-2015*（2015）
[植田 15] 植田禎子，飯田 龍，浅原正幸，松本裕治，徳永健伸：『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション，第 8 回コーパス日本語学ワークショップ予稿集，pp. 205-214（2015）
[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is all you need, *31st Conf. on Neural Information Processing Systems*（*NIPS*）, pp. 5998-6008（2017）
[Wiseman 15] Wiseman, S., Rush, A. M., Shieber, S. and Weston, J.: Learning anaphoricity and antecedent ranking features for coreference resolution, *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and 7th Int. Joint Conf. on Natural Language Processing*（*Vol. 1: Long Papers*）, pp. 1416-1426（2015）
[Wiseman 16] Wiseman, S., Rush, A. M. and Shieber, S. M.: Learning global features for coreference resolution, *Proc. 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 994-1004（2016）
[呉 19] 呉 佩珣，近藤森音，森山奈々美，荻原亜彩美，加藤 祥，浅原正幸：『分類語彙表』と『岩波国語辞典第五版タグ付きコーパス 2004』の対応表，言語資源活用ワークショップ 2019 発表論文集（2019）
[山崎 17] 山崎 誠，柏野和佳子：『分類語彙表』の多義語に対する代表義情報アノテーション，言語処理学会第 23 回年次大会発表論文集，pp. 302-305（2017）
[Zhang 19] Zhang, H., Song, Y., Song, Y. and Yu, D.: Knowledgeaware pronoun coreference resolution, *Proc. 57th Annual Meeting of the Association for Computational Linguistics,* pp. 867-876（2019）
[Zhou 15] Zhou, J. and Xu, W.: End-to-end learning of semantic role labeling using recurrent neural networks, *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conf. on Natural Language Processing*（*ACL-IJCNLP*）, pp. 1127-1137（2015）

2019 年 12 月 2 日　受理

著　者　紹　介

松林　優一郎（正会員）

2010 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．国立情報学研究所，東北大学，理化学研究所を経て，現在，東北大学大学院教育学研究科准教授．自然言語処理，特に意味解析・教育応用の研究に従事．情報処理学会，言語処理学会，ACL 各会員．

浅原　正幸

1998 年京都大学総合人間学部卒業．2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．現在，人間文化研究機構国立国語研究所コーパス開発センター教授．