ニューラルネットワークによる日本語述語項構造解析の素性の汎化

松林 優一郎　　乾 健太郎
東北大学
{y-matsu, inui}@ecei.tohoku.ac.jp

1 はじめに

述語項構造は、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。例えば次の文、

[太郎] は [手紙] を書いた。

では、「書く」という表現が述語であり、「太郎」と「手紙」という表現がこの述語の項である。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事で、複雑な文構造・文章構造を持った文章において「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。

本稿では、日本語の述語項構造解析器をニューラルネットワーク（NN）モデルを用いて設計し、従来用いられてきた基本的な素性である統語関係パス及び単語の共起に関する素性を分散表現を用いて汎化する。これにより、述語項構造解析における主要な課題の一つである、特徴の複雑な組み合わせにおける疎データ問題を解消し、解析精度の向上を図ると共に、導入した分散表現による特徴量が、従来用いられてきた $\{0,1\}$ の二値による表現（二値素性と呼ぶ）を十分に代替可能であるか検証する。

述語項構造解析では、述語と項の間の関係を捉えるための主要な手段として、述語から項にたどり着くまでの係り受け関係を表現する統語関係パスといった素性や、述語と項の単語の共起性をとらえるための単語の字面による組み合わせ素性が従来より用いられてきた。[4]

一方で、述語と項の関係には、述語毎に個別の振る舞いが見られ、一般性の高い事例のみでなく、述語毎の個別のルールを覚えることが重要であることが分かっており [1]、したがって、項と隣接する単語や、係り受け木の親や子といった単語との組み合わせや、係り受けパス上の中間の単語情報を保持した、より詳細化された統語関係パスなど、複雑な組み合わせ構造が素性として提案され、解析精度の向上に寄与してきた。しかしながら、膨大かつロングテールの頻度分布を持つ述語についての個別のルールや、単語の共起に関する情報を単純な組み合わせ情報で学習することは、組み合わせ爆発による学習データの低頻度問題から限界があり、どのようにしてこれらの複雑な情報を一般化するかが、述語項構造解析における課題の一つであった。

近年、分散表現による単語情報の汎化が注目を浴び始め、述語項構造解析への適用も進んでいる [2]。また、統語パスを従来の記号的表現から埋め込みベクトルに汎化するといった方法が研究されてきており [7]、複雑な統語構造の効率的な一般化が期待されている。一方で、これらの埋め込み表現が、従来の二値素性による表現を十分に代替する能力があるかについては、十分な検証がされていない。例えば、慣用表現のように、特定の単語列が現れた場合に字義どおりとは異なる意味や機能になるものなど、直接的にパターンを記録して学習する従来の二値素性が有利に働く事例が存在する可能性も否定できない。

そこで我々は、従来の二値素性の組み合わせを使う解析モデルを NN アーキテクチャを用いて拡張し、日本語の述語項構造解析において、単語埋め込みと統語パス埋め込みの二つの分散表現を導入し、これらの汎化性能を明らかにする。具体的には、多層 NN モデルの入力層に、単語や統語パターンを直接覚え込むような従来の二値素性と、単語埋め込みベクトルと統語パス埋め込みベクトルの一方、あるいは双方を用いるモデルを設計し、これらを比較することで、埋め込み表現による二値素性の代替可能性について分析する。

実験では、(1) 単語埋め込みと統語関係パス埋め込みによる素性の汎化により、従来の二値素性による単語やパスの情報を完全に置き換えることができるか、(2) 単語埋め込み、統語関係パス埋め込み、従来の二値素性の分類性能に対する寄与の検証、及び (3) 既存研究との精度比較を行い、結果として、単語埋め込みと統語関係パス埋め込みによる素性の汎化は、少なくとも二値素性で学習した場合に比べて同程度の性能を実現できることが分かった。また、従来の二値素性の組み合わせを使うモデルを NN アーキテクチャによって拡張することで、従来のモデルからの大幅な性能向上が認められた。

2 問題設定

本研究では、述語項構造解析の研究で利用される主要なコーパスの一つである NAIST テキストコーパス (NTC) [3] の 1.5 版の注釈の仕様にもとづいて問題を定める。ただし、日本語では、項の省略が起こった場合に述語とその暗黙の項が異なる文に現れる場合があり、項が述語と同一文内に現れる場合に対して、異なる文に現れる場合では述語と項の間の統語的な関係を用いることが出来ないなど、問題の性質が大きく異なる。このため、本研究では既存研究 [5, 6, 10] に習い、述語項構造のうち、特に述語と項が同一文内にある事例のみを解析対象とする。

既存研究の設定に準じ、以下のとおり問題の入出力を定める。入力として、文と、その文の正解の形態素情報、文節係り受け構造、述語の位置が与えられる。解析器は、与えられたそれぞれの述語に対し、文中からその述語のガ、ヲ、ニ格に対応する項をそれぞれ高々一つ出力する。項は、項の範囲として適切な文字列の最右の形態素で表現される。

開発・評価セットとして、多くの既存研究で利用されている平ら [9] の分割に習い、1 月 1 日〜11 日までのニュース記事と 1〜8 月の社説記事を訓練データ、1 月 12 日〜13 日のニュース記事と 9 月の社説記事を開発データ、1 月 14 日〜17 日のニュース記事と 10〜12 月の社説記事を評価データとして用いる。評価は、システムが出力した項の位置とラベルが、NTC に項として示されている共参照クラスタ内の形態

素のいずれかと一致すれば正解、しなければ不正解とし、適合率、再現率、F 値を求めることで行う。

3 解析モデル

先に述べたとおり、我々は NN の機構を利用し、日本語の述語項構造解析において低頻度の問題に直面している以下の二つの特徴量を分散表現で置き換える。

- 述語や項候補単語との共起に関する特徴量
- 語彙的に詳細化した統語関係パス

我々の目的は、これらを分散表現化した特徴量が従来の二値素性をどの程度効果的に置き換えるかを検証することであるので、それ以外の部分については古典的な二値の素性を用いるモデルにおいて最高精度を達成している 松林 & 乾 [5] の手法を参考に設計し、実験で上記二つの特徴量を従来どおり二値素性として利用する場合と比較する。

解析は、入力文中の各述語に対し、同一文内の各形態素、および文中に項がないことを表現する仮想的な形態素「NONE」を項候補とし、述語と項候補のペアに対してこのペアがガ、ヲ、ニの格関係となるかどうかのスコアを計算することで行う。具体的には、我々のモデルは入力文と文中の各述語が与えられたとき松林 & 乾 [5] の手法に従い以下の 3 ステップにより文中の各述語の項を特定する。

1. 訓練データ内の統計から項となることが稀な品詞を持つ項候補を枝刈りする。
2. 述語と項候補のペアに対して { ガ, ヲ, ニ, 無 } の多値分類を行うモデルを学習し、各候補について、それぞれのラベルに対するスコアを求める。
3. 各述語の { ガ, ヲ, ニ } のラベルについて、項候補から最もスコアの高いものを一つずつ選び、出力する。ただし、ラベルの適合率と再現率を調整するために、{ ガ, ヲ } のラベルのそれぞれについて実数値の修正項を定めておき、NONE のスコアには修正項を加算しておく。[*1]

本研究と松林 & 乾 [5] のモデルの本質的な違いは、述語と項候補のペアに対してラベルのスコアを与える部分に、以下で説明する NN モデルを導入している点である。

我々の NN モデルの概観を図 1 に示す。このモデルは大きく分けて二段階の構成となっている。一段目は、先に述べた二つの特徴量を分散表現として埋め込む部分であり、

- 統語関係パス埋め込み
- 述語と項候補の単語埋め込み

の二つのモジュールから成る。このうち、「統語関係パス埋め込み」は従来の統語関係パスおよびパスの中間の単語文字列や品詞を考慮した統語関係パスを代替する埋め込みベクトルを構成する部分であり、「述語と項候補の単語埋め込み」は従来の述語や項候補単語を表す二値素性を代替するための単語ベクトルを構成する部分である。この二つの埋め込みの詳細は次の節で述べる。

二段目は、一段目で構成した埋め込みベクトルと従来利用されてきた二値素性の特徴ベクトルを結合したベクトルを入力に取る多層 NN となっている。松林 & 乾 [5] の結果より、従来の二値素性はそれらの 5 次までの詳細な組み合わせによって解析精度が向上することが分かっているため、多層に隠れ層を重ねることによって、素性の高次の組み合わせ構造を学習する。二段目のネットワークは feed forward ネットワークであり、形式的に、

$$g(h_n) = \mathrm{softmax}(h_n) \tag{1}$$
$$h_i(h_{i-1}) = \mathrm{ReLU}(\mathrm{BN}(W_i h_{i-1} + b_i)) \tag{2}$$
$$h_1(m) = \mathrm{ReLU}(\mathrm{BN}(W_1 m + b_1)) \tag{3}$$
$$m = u \oplus w_p \oplus w_a \oplus f(x) \tag{4}$$

と書ける。ここで、$x = (p, a, s)$ は入力である述語 p、項候補 a、文 s の三つ組であり、$u \in \mathbb{R}^{d_u}$ は統語パス埋め込みベクトル、$w_p, w_a \in \mathbb{R}^{d_w}$ はそれぞれ述語と項候補の単語埋め込みベクトル、$f(x) \in \mathbb{R}^{d_f}$ は二値素性ベクトルである。$m \in \mathbb{R}^{d_u+2\cdot d_w+d_f}$ は二段目のネットワークの入力層であり、式の中の $\oplus$ はベクトルの連結を表す。h_i は i 番目の隠れ層であり、n は隠れ層の数、$W_1 \in \mathbb{R}^{d_{h,i}\times(d_u+2\cdot d_w+d_f)}$ と $W_i \in \mathbb{R}^{d_{h,i}\times d_{h,i-1}}$ はそれぞれ隠れ層の重み行列であり、$b_i \in \mathbb{R}^{d_{h,i}}$ はバイアス項である。

各隠れ層には batch normalization と ReLU 活性化関数を適用する。$g : \mathbb{R}^{d_{h,n}} \to \mathbb{R}^4$ は softmax 層であり、最終的に各ラベルの出力値が得られる。隠れ層の次元 $d_{h,i}$ は $d_{h,1}$ の場合のみ探索し、$i > 1$ の場合は $d_{h,i} = d_{h,1}/2$ を利用する。ロス関数は categorical crossentropy であり、最適化手法には Adam を採用する。

■統語関係パスの埋め込み 近年、semantic role labeling や関係抽出タスクに利用する目的で、統語関係パスを分散表現化する手法が提案された。[8, 7] これらはいずれも LSTM を利用して単語や品詞、係り受けの向きとラベルの情報を係り受けの順にベクトルに埋め込む手法であった。本研究では、Roth ら [7] の手法に習い、これを日本語の係り受け構造の埋め込みに適用することを考える。

図 1 左のように、統語関係パスの埋め込みモジュールは、入力として述語と項候補、文の係り受け構造、各形態素の品詞が与えられた時、はじめに述語から項候補側に係り受けパスを辿り、この過程で経由した語の「品詞」「原形」「係り受けの向き」を順に並べた系列を作る。次にこの系列の t 番目の要素をベクトル表現 $e_t \in \mathbb{R}^{d_e}$ に変換する埋め込み層を適用する。結果として得られたベクトル系列を順に GRU に入力していき、系列のすべての要素を入力した後の出力 $h_{g,t} \in \mathbb{R}^{d_e}$ を統語関係パスの埋め込みベクトルとして、二段目の多層 NN の入力に渡す。

統語関係パスの埋め込みモジュールは、形式的に、

$$z_t = \sigma(W_z e_t + U_z h_{g,t-1} + b_z) \tag{5}$$
$$r_t = \sigma(W_r e_t + U_r h_{g,t-1} + b_r) \tag{6}$$
$$h_{g,t} = z_t \circ h_{g,t-1} + (1 - z_t)\sigma(W_h e_t + U_h(r_t \circ h_{g,t-1}) + b_h) \tag{7}$$

と書ける。それぞれの式で、$W, U \in \mathbb{R}^{d_e \times d_e}$ は層の重み行列であり、$b \in \mathbb{R}^{d_e}$ はバイアス項である。

■述語と項候補の単語埋め込み 単語の埋め込みモジュールは、単語の原形毎に専用のベクトル表現を保持する。単語の初期ベクトルとして、日本語 wikipedia の 2016 年 9 月 1 日時点での dump データを用いて作成した word2vec ベクトルを用いる。WikiExtractor[*2] により本文を抽出し、CaboCha 0.68 により形態素区切りを得たのち、word2vec [*3] を用い

[*1] ニ格は事例が少なく、過剰適合の恐れがあるため修正項を定めない。

[*2] https://github.com/attardi/wikiextractor

[*3] https://code.google.com/archive/p/word2vec/

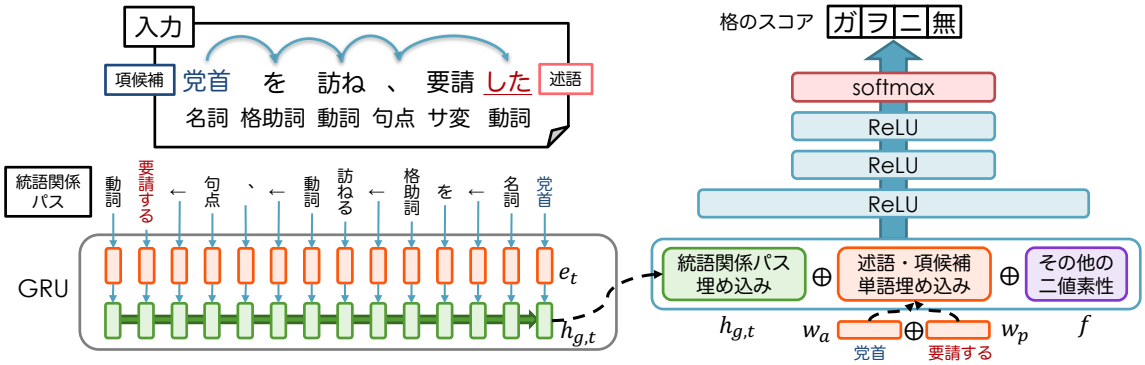

図 1　(述語, 項, 文) の三つ組に対して、格ラベルのスコアを与える NN モデルのネットワーク図

て単語ベクトルを得た。CaboCha の実行時には JUMAN 辞書を利用した。word2vec の実行時のパラメータとして “-cbow 1 -size 256 -window 10 -negative 10 -hs 0 -sample 1e-5 -threads 40 -binary 0 -iter 3 -min-count 10” を指定した。これらの単語ベクトルは多層 NN からの教師信号により更新する。

■二値素性　二値素性として、松林&乾 [5] で利用されているテンプレートの各要素にあたる素性を利用する。ただし、このうち、名詞クラスタと大規模文書データからの共起情報は単語埋め込みモジュールにおける word2vec の事前学習によって代替するため使用しない。具体的に利用する素性は、松林&乾 [5] の表 1 に示されるもののうち、以下の素性である。

述語に関する素性	述語の表層、原形、品詞、品詞細分類、活用形、サ変動詞汎化*4、格交替を起こす接尾文字列*5
項候補に関する素性	項候補の表層、原形、品詞、品詞細分類、固有名詞タグ、項候補主辞判定、文節文字列（文節先頭から主辞までの文字列)、項候補の助詞（文節の主辞以降に含まれる文字列)、項候補の右の語の原形、読み、品詞、品詞細分類
述語と項候補の関係に関する素性	項候補以外の係助詞*6、 項候補と述語の前後関係、項候補が述語と同一文節内にあるか、項候補と述語の形態素距離、文節係り受け距離、文節係り受けパス*7、 述語・助詞つき係り受けパス、主辞・助詞つき係り受けパス、述語の連体修飾 (隣接型)*8

表 1　利用する二値素性

組み合わせ素性については、多層 NN によって自動的に学習することを狙うため二値素性としては表現しない。また、訓練データでの発火回数が 10 回未満の素性は切り捨てる。

4　ハイパーパラメータ

各ハイパーパラメータは開発データでの F 値が最大となるよう値を定める。調整する値は、統語関係パスの埋め込みに利用する GRU の W, U の各 dropout 値 o_W, o_U、統語関係パスの埋め込みベクトルの次元 d_e、多層 NN の隠れ層の数 n と隠れ層の次元 d_h、Adam の学習係数 l、NONE の修正項 $t_{\mathrm{ga}}, t_{\mathrm{wo}}$ である。 それぞれについて、$o_W, o_U \in \{0.0, 0.1, 0.2, 0.3, 0.5\}$、$d_e \in \{32, 64, 128\}$、$n \in \{1, 2, 3, 4\}$、$d_h \in \{128, 256, 512, 1000, 2000\}$、$l \in \{0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$ を試行した。$t_{\mathrm{ga}}, t_{\mathrm{wo}}$ については、[-0.2, 2.0] の範囲を 0.01 刻みで探索した。

GRU のパラメータは、GRU の出力したベクトルのみを多層 NN の入力としたモデルを用いて調整し、$d_e = 64$、$o_W = o_U = 0.2$ を得た。その他のモデルでは、GRU のパラメータをこの値に固定して、その他のパラメータのみを調整した。

5　実験

実験では、次の三つのことを確かめる。

(1) 単語埋め込みと統語関係パス埋め込みによる素性の汎化により、従来の二値素性による単語やパスの情報を完全に置き換えることができるか。
(2) 単語埋め込み、統語関係パス埋め込み、従来の二値素性の分類性能に対する寄与
(3) 既存研究との精度比較

(1) と (2) を検証するために、我々は二種類の二値素性ベクトルを用意する。一つは、ベースラインである松林&乾 [5] で利用されている基本素性について、3 節で挙げた全てを利用するもの（「B」で表記）であり、もうひとつは、このうち単語表層、原形と、全ての統語パス素性を省いたもの（「$\mathrm{B}_{\mathrm{excl}}$」で表記）である。これらを利用して、「統語関係パス埋め込み」「単語埋め込み」「二値素性」の三つの要素のうち、任意の組み合わせを二段目の多層 NN に入力し、各要素の効果を比較する。各モデルは、入力する情報に応じて記号の組み合わせで表記する。二値素性はその使い分けに応じて「B」「$\mathrm{B}_{\mathrm{excl}}$」で表記するほか、統語関係パス埋め込みを利用するものは「P」を表記し、述語と項の単語埋め込みを利用するものは「W」を表記する。

表 2 に、比較結果を示す。まず、三つの要素全てを入力した PWB と $\mathrm{PWB}_{\mathrm{excl}}$ を見ると、これらは同程度の性能を示している。このことから、特定の単語の共起や、中間の単語情報を含んだ詳細化された統語関係を特定の記号パタンとして直接おぼえるような従来の二値素性による方法と比べても、分散表現の組み合わせによる表現が、同様の情報を上手く汎化して捉えられているものと解釈できる。

次に特筆すべき点として、単語ベクトルの性能が挙げられる。単語ベクトルはそれ単体（W モデル）では当然ながら分類性能は低いものの、単語ベクトルを利用して組み合わせ素性を組み上げるモデルである二種類の PWB モデルと二種類の WB モデルは共に肉迫している。このことから、単語ベクトルは、述語項構造解析のように述語や項ごとの素性の複雑な共起が重要な情報になるタスクにおいて、組み合わ

	DEV	TEST					
モデル	F1 (%)	F1 (%)	Pre. (%)	Rec. (%)	d_h	n	l
PWB	82.46	82.78	85.75	80.00	2000	3	0.0002
PWB_{excl}	82.52	82.72	85.76	79.88	2000	3	0.0002
WB	82.54	82.48	84.42	80.62	2000	3	0.0005
WB_{excl}	82.41	82.21	84.49	80.05	2000	3	0.0002
PB	82.09	81.88	84.11	79.76	2000	3	0.0001
PB_{excl}	82.00	82.00	84.98	79.22	1000	3	0.001
PW	73.44	73.50	84.04	65.30	1000	2	0.0005
B	81.91	81.76	84.40	79.29	1000	3	0.0005
B_{excl}	78.81	78.77	81.51	76.21	1000	3	0.005
P	72.14	72.15	82.40	64.17	128	2	0.0005
W	33.72	33.74	32.37	35.23	512	2	0.0002

表 2　提案モデル間の精度比較: 多層 NN に入力する情報に応じて記号の組み合わせでモデル名を表記。統語関係パス埋め込みを利用するものは「P」、単語埋め込みを利用するものは「W」、二値素性は全て使うものを「B」、単語表層やパス素性を省いたものを「B_{excl}」で表記。

		直接係り受け			文内ゼロ照応			同一文節内		
モデル	ALL (%)	ガ	ヲ	ニ	ガ	ヲ	ニ	ガ	ヲ	ニ
PWB_{excl}	82.72	90.7	94.8	61.7	52.7	37.5	4.26	47	29	67.7
B	81.76	89.6	94.3	59.7	48.5	34.2	2.26	47	21	61.8
MI14	80.9	87.8	94.0	63.7	49.0	27.7	25.7	–	–	–
OU15 [6]	79.23	88.13	92.74	38.39	48.11	24.43	4.80	–	–	–
大内 16 [10] DRGM	81.22	88.66	93.95	66.50	51.57	38.06	9.44	–	–	–

表 3　既存研究との精度比較（TEST セット F 値）

せ爆発による低頻度問題を解消する意味で恩恵が大きいものといえる。

次に、既存研究との比較結果を表 3 に示す。まず、我々の手法のベースラインである松林&乾 [5] と比べて、このモデルの基本素性のみを利用した B モデルが F 値で 0.8 ポイントの改善を示している。松林&乾 [5] は人手による精緻な素性設計により最高性能を達成したものであったが、この結果から、我々が当時人手で設計していたものよりもさらに細やかな特徴の組み合わせが性能に大きく寄与していると言える。また、我々の提案手法のうち最高性能を達成する PWB_{excl} モデルは、ベースラインに比べて 1.8 ポイントの改善を示した。これは、入力に正解の統語情報を利用するモデルの中では最高性能を達成するモデルである。

大内 16 は品詞や係り受け情報などの正解の統語構造を使用しておらず、入力が異なるため直接的な性能比較は出来ないが、この結果から、統語構造から得られる詳細な構造がとらえられれば、十分な精度改善が見込めることが分かる。

加えて、OU15 や大内 16 は複数の述語とその項ラベルの間の関係を考慮したモデルで、そのことによって性能が改善している。我々のモデルは複数の項ラベル間の情報は利用していない。したがって、これらの手法を組み合わせることでさらなる性能改善が期待できる。

6　おわりに

本稿では、日本語の述語項構造解析器をニューラルネットワーク（NN）モデルを用いて設計し、従来用いられてきた基本的な素性である統語関係パス及び単語の共起に関する素性を分散表現を用いて汎化し、これにより、述語項構造解析における主要な課題の一つである、特徴の複雑な組み合わせにおける疎データ問題を解消し、解析精度の向上を図ると共に、導入した分散表現による特徴量が、従来用いられてきた $\{0, 1\}$ の二値による表現（二値素性と呼ぶ）を十分に代替可能であるか検証した。結果として、単語埋め込みと統語関係パス埋め込みによる素性の汎化は、少なくとも二値素性で学習した場合に比べて同程度の性能を実現できることが分かった。また、従来の二値素性の組み合わせを使うモデルを NN アーキテクチャによって拡張することで、従来のモデルからの大幅な性能向上が認められた。

謝辞

本研究は JSPS 科研費 JP15K16045、および JP15H01702 の助成を受けたものです。

参考文献

[1] Alan Akbik and Yunyao Li. K-srl: Instance-based learning for semantic role labeling. In *COLING 2016*, pp. 599–608, 2016.
[2] Nicholas Fitzgerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Semantic Role Labeling with Neural Network Factors. *EMNLP*, pp. 960–970, 2015.
[3] 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.
[4] Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, Vol. 34, No. 2, pp. 145–159, 2008.
[5] 松林優一郎, 乾健太郎. 統計的日本語述語項構造解析のための素性設計再考. 言語処理学会第 20 回年次大会発表論文集, pp. 360–363, 2014.
[6] Hiroki Ouchi, Hiroyuki Shindo, Kevin Duh, and Yuji Matsumoto. Joint Case Argument Identification for Japanese Predicate Argument Structure Analysis. In *ACL-IJCNLP*, pp. 961–970, 2015.
[7] Michael Roth and Mirella Lapata. Neural Semantic Role Labeling with Dependency Path Embeddings. In *ACL*, No. 2002, pp. 1192–1202, 2016.
[8] Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving Hypernymy Detection with an Integrated Pattern-based and Distributional Method. In *ACL*, pp. 2389–2398, 2016.
[9] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A japanese predicate argument structure analysis using decision lists. In *EMNLP*, pp. 523–532, 2008.
[10] 啓樹大内, 裕之進藤, 裕治松本. 深層リカレントニューラルネットワークを用いた日本語述語項構造解析. 情報処理学会研究報告 自然言語処理（NL）2016-NL-229, 第 21 巻, pp. 1–8, 2016.