The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

係り受け誤り埋め込み表現のクラスタリングによる複数のドメイン間での追加訓練の効果の比較分析

Comparative Analysis of the Effect of Additional Training between Multiple Domains by Clustering of the Embeddings of Parsing Errors

原 拓也 *1 Takuya Hara　松崎 拓也 *1 Takuya Matsuzaki　横野 光 *2 Hikaru Yokono　佐藤 理史 *1 Satoshi Sato

*1名古屋大学工学研究科 Graduate School of Engineering, Nagoya University
*2（株）富士通研究所 Fujitsu Laboratories Ltd.

We propose a method for studying the effect of additional training of Japanese dependency parsers across multiple domains from a bird's-eye view. We collected the parsing errors of a parser before and after additional training using target domain data. We then conducted cluster analysis of the parsing errors represented as dense real vectors, which are obtained from the internal states of the parser. Through quantitative and qualitative analysis of the clusters, we could grasp the types and the numbers of the parsing errors across multiple target domains.

1. はじめに

機械学習における課題の一つにドメイン適応問題がある。多くの場合、訓練に用いたデータのドメイン（以下、訓練ドメイン）と、適用の対象となるデータのドメイン（以下、評価ドメイン）が異なるとタスクの精度が低下する。精度の低下の原因は訓練ドメインと評価ドメインに含まれている文の構造や単語、あるいは正解の分布が異なっているからである。この精度の低下を防ぐという課題をドメイン適応問題という。

本研究では、ドメイン適応問題の対象となるタスクとして係り受け解析を取り上げる。係り受け解析とは文を構成している文節同士の係り受け関係（修飾・被修飾）を解析するタスクである。係り受け解析の例を図 1 に示す。図の中の四角が文節を表しており、矢印が係り受け関係を表している。それぞれの矢印は係り文節（修飾文節）から受け文節（被修飾文節）へと伸びている。

ドメイン適応問題は、評価ドメインデータを用いて追加訓練を行うことである程度解決できることが多い。評価ドメインデータの量は多くの場合限られており、それらを効率的に用いるための方法に関する研究は、多数存在する。例えば Daumé III と Marcu [Daumé III 06] は、追加訓練を行う際の評価ドメインデータへの適切な重み付け方法を提案した。また Daumé III [Daumé III 07] はドメイン固有の情報とドメインに依存しない情報を分けて訓練を行えるように素性を設計した。他にも、効率的な学習を行うために、次にアノテーションすべきデータを逐次的に見つける方法（アクティブラーニング）に関する研究も多数ある。

上で述べた研究では、適用する評価ドメインそれぞれに対し、追加訓練用データを用意する必要がある。しかし、評価ドメインのデータを多数のドメインについて用意することはコストの点から現実的でない。一方、追加訓練がなぜ有効なのかは、追加訓練によって解消された誤りや、解消されなかった誤りを一つ一つ見ていてもよくわからないことが多い。これに対し、追加訓練の効果を俯瞰的に分析することで、より少ないアノテーションで、効果的に適応を行うための知見が得られると期待できる。そこで、本研究では係り受け解析器のドメイン適応を対象に、追加訓練の効果を俯瞰的に分析するための手法として、係り受け誤りのクラスタリングを提案する。

連絡先: hara.takuya@h.mbox.nagoya-u.ac.jp

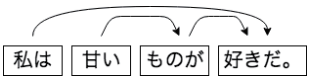

図 1: 係り受け解析の例

クラスタリングを行うためには係り受け誤りの間の類似性を数量的に捉える必要がある。そのため、本研究では係り受け解析器の内部状態を用いて係り受け誤りを埋め込み表現（密ベクトル表現）に変換する。

本論文では特に、複数の評価ドメインの間での追加訓練の効果の違いについて分析する。この分析は、評価ドメインデータを用いた追加訓練の前後における係り受け誤りを埋め込み表現に変換し、それらに対しクラスタリングを行い、得られた結果を複数ドメイン間で比較することで行われる。本論文では訓練ドメインとして新聞記事、評価ドメインとして中学校理科教科書及び特許文書を用いてクラスタリングを行った。クラスタに含まれる誤りの数の分布や、クラスタごとの誤りの性質を調べることで、ドメインごとの特徴を比較・分析することができた。

以下、2 節で提案手法について説明する。3 節では実際のデータを用いた分析を行う。

2. 係り受け誤りのクラスタリング

2.1 係り受け解析器

本研究では、LSTM に基づく係り受け解析器の誤りを分析する。また、誤りの埋め込み表現を得るために、この解析器の内部状態を用いる。本研究で用いる解析器は松野ら [松野 18] のものをベースにした。松野らのモデルは大きく分けて、(i) 単語層 LSTM による単語表現の獲得、(ii) 文節層 LSTM による文節表現の獲得、(iii) 多層パーセプトロンによる次元圧縮、(iv)Biaffine 変換による係り受け関係スコアの計算の 4 段階に分けられる。本研究では (i) 及び (ii) はそのままで、(iii) 及び (iv) の代わりに多層パーセプトロンによる係り受け関係スコアの計算を行う。

具体的には以下のモデルを用いる。最初に文の各単語の品詞（4 レベル）・活用形・活用タイプ・基本形をそれぞれ埋め込み層に通し、それらを結合させることで単語分散表現とする。文の各単語の単語分散表現をまとめて系列にしたものを $\boldsymbol{x}^{word}$

として単語層 LSTM に入力する。この時 (i) の操作は以下の式のように行われる。

$$(\overleftarrow{\boldsymbol{y}}_i^{word}, \overrightarrow{\boldsymbol{y}}_i^{word}) = \mathrm{BiLSTM}^{word}(\boldsymbol{x}_i^{word}) \tag{1}$$

t 番目の文節が i 番目から j 番目の単語で構成されている時、文節層 LSTM への入力 $\boldsymbol{x}^{chunk}$ は以下のように計算される。

$$\overleftarrow{\boldsymbol{x}}_t^{chunk} = \overleftarrow{\boldsymbol{y}}_i^{word} - \overleftarrow{\boldsymbol{y}}_{j+1}^{word} \tag{2}$$

$$\overrightarrow{\boldsymbol{x}}_t^{chunk} = \overrightarrow{\boldsymbol{y}}_{i-1}^{word} - \overrightarrow{\boldsymbol{y}}_j^{word} \tag{3}$$

$$\boldsymbol{x}_t^{chunk} = \overleftarrow{\boldsymbol{x}}_t^{chunk} \oplus \overrightarrow{\boldsymbol{x}}_t^{chunk} \tag{4}$$

t 番目の文節に対応した文節表現を $\boldsymbol{y}_t^{chunk}$ とする。このとき (ii) の操作は以下の式のように計算される。

$$(\overleftarrow{\boldsymbol{y}}_t^{chunk}, \overrightarrow{\boldsymbol{y}}_t^{chunk}) = \mathrm{BiLSTM}^{chunk}(\boldsymbol{x}_t^{chunk}) \tag{5}$$

$$\boldsymbol{y}_t^{chunk} = \overleftarrow{\boldsymbol{y}}_t^{chunk} \oplus \overrightarrow{\boldsymbol{y}}_t^{chunk} \tag{6}$$

t 番目の文節が u 番目の文節に係る際（$t < u$）の係り受け関係スコア $s_{t,u}$ は、以下のように定義される。

$$s_{t,u} = \mathrm{MLP}(\boldsymbol{y}_t^{chunk} \oplus \boldsymbol{y}_u^{chunk} \oplus \boldsymbol{d}(t,u)) \tag{7}$$

$$\boldsymbol{d}(t,u) = [u - t = 1] \oplus [u - t \leq 5] \tag{8}$$

ただし $[\cdot]$ は内部の条件が真ならば要素が 1 のベクトルを、偽ならば要素が 0 のベクトルを表す。

この係り受け関係スコアモデルを用いて内元ら [内元 99] のアルゴリズムで訓練・解析を行う。内元らのアルゴリズムでは、後方の文節から順に係り先を決定する。その際、ある係り文節に対し、候補となる受け文節の集合の中から 1 つの文節を選択する必要がある。訓練時は、係り文節と正しい受け文節のペアの係り受け関係スコアが、係り文節とその他の受け文節の候補のペアの係り受け関係スコアよりも高くなるように訓練する。具体的には、ある係り文節 t に対し、正しい受け文節を u、候補文節集合を $\mathcal{C}_t$ とするとき、損失関数を以下のように定義する。

$$loss = -\sum_{u' \in \mathcal{C}_t} \log \frac{\exp(s_{t,u})}{\exp(s_{t,u}) + \exp(s_{t,u'})} \tag{9}$$

解析時は、それぞれの係り文節に対する受け文節候補集合のうち、係り受け関係スコアが最も高くなるものを選択する。

2.2 係り受け誤りの収集法

本研究では評価ドメインデータによる追加訓練を行う前後で、解析結果の誤り方がどのように変化するかを観察する。この目的のため、以下のように訓練データ及び解析対象のデータが異なる 3 タイプの誤り例を収集した。

(1) 訓練ドメイン誤りの収集　最初に、訓練ドメインのデータの半分を用いてベースモデルを作る。訓練ドメインのデータの残り半分をベースモデルで解析した時の誤りを収集する。

(2) 追加訓練前の評価ドメイン誤りの収集　評価ドメインデータの全てをベースモデルで解析した時の誤りを収集する。

(3) 追加訓練後の評価ドメイン誤りの収集　評価ドメインデータに対し 4 分割交差検証を行う。ただし、ベースモデルの多層パーセプトロン（式 (7)）の部分のみを、ベースモデルのパラメータ値を初期値として、評価ドメインデータを用いて訓練する（その結果を以降では追加訓練済みモデルと呼ぶ）。その結果得られた誤りを収集する。

追加訓練において多層パーセプトロンのみを訓練することで、すべてのモデルについて文節の埋め込み表現は共通となる。これにより係り受け誤りの埋め込み結果（2.3.1 項）を比較することができる。なお、本研究では追加訓練済みモデルで訓練ドメインデータを解析した際の誤りは用いない。理由は本研究の目的が追加訓練による評価ドメイン誤りの変化の観察であり、訓練ドメイン誤りの変化には興味がないためである。

2.3 係り受け誤りの分析

係り受け誤りを俯瞰的に観察するための手法としてクラスタリングを行いたい。クラスタリングを行うためには、誤りの間に距離を定義する必要がある。そこで最初に係り受け誤りを埋め込み表現（ベクトル）に変換し、ユークリッド距離を誤り間の距離とする。

2.3.1 係り受け誤り埋め込み表現

係り受け誤り埋め込み表現（以降、係り受け誤り表現）は係り受け誤りの 3 要素である係り文節・正しい受け文節・誤った受け文節とそれらの間の距離素性を埋め込んだものである。

2.1 項で述べた係り受け解析器には、中間層として文節表現 ((6) 式の $\boldsymbol{y}_t^{chunk}$) がある。文節表現は文節及び周辺文脈から、係り受け解析に必要な情報を抽出したものだと考えることができる。そこで係り文節・正しい受け文節・誤った受け文節に対応する文節表現を結合し、さらに距離素性を結合したものは係り受け誤り全体の特徴を表現するものと考えることができる。係り文節を t, 正しい受け文節を u, 誤った受け文節を $\tilde{u}$ として、係り受け誤り表現 $\boldsymbol{e}_{t,u,\tilde{u}}$ は以下のように作成する。

$$\boldsymbol{e}_{t,u,\tilde{u}} = \begin{cases} \boldsymbol{y}_t^{chunk} \oplus \boldsymbol{y}_u^{chunk} \oplus \boldsymbol{y}_{\tilde{u}}^{chunk} \oplus \boldsymbol{d}(t,u) \oplus \boldsymbol{d}(t,\tilde{u}) & (u < \tilde{u}) \\ \boldsymbol{y}_t^{chunk} \oplus \boldsymbol{y}_{\tilde{u}}^{chunk} \oplus \boldsymbol{y}_u^{chunk} \oplus \boldsymbol{d}(t,\tilde{u}) \oplus \boldsymbol{d}(t,u) & (u > \tilde{u}) \end{cases}$$

上式は、二つの受け文節候補 u と $\tilde{u}$ のうち、正解がどちらであるかとは無関係に、文中での出現順に従って埋め込み表現を構成することを表す。これにより、係り受けアノテーションされていないデータについても、例えば係り受け関係スコア 1 位と 2 位のペアを用いて、同様に埋め込み表現を作成できる。

2.3.2 クラスタリング

本研究では、係り受け誤り表現を以下の手順で分析する。

1. 訓練ドメイン誤りの埋め込み表現を k-means 法でクラスタリングする。
2. 評価ドメイン誤りそれぞれの埋め込み表現 e について、1. で得られたクラスタのうち、e からクラスタ重心までの距離が最も近いものに e を割り当てる。

この手順は数の多い訓練ドメイン誤りを類型ごとにグループ化した後、評価ドメイン誤りのそれぞれを、近い類型をもつクラスタへと分類することを意図している。

3. 実験

この節では最初に実験に用いたデータの統計について述べ、次に係り受け誤りの収集に関する詳細を述べる。最後にクラスタリングの結果とその分析について述べる。

3.1 利用データ

訓練ドメインデータとして新聞記事（京都大学テキストコーパス [黒橋 97]）を用いた。また、評価ドメインデータとして理科教科書及び特許文書を用いた。理科教科書データは東京書籍「新編新しい科学（中学 1 年生、中学 2 年生）」のテキス

表 1: データの統計

データ	文数	文節数	文節数／文
新聞記事	38,400	372,130	9.7
理科教科書	2,460	22,923	9.3
特許文書	953	11,157	11.7

表 2: 係り受け誤りの収集結果

対象	モデル	誤り文節数	精度
新聞記事	ベースモデル	13,878	91.39
理科教科書	ベースモデル	2,429	88.13
	教科書訓練済みモデル	1,671	91.83
特許文書	ベースモデル	1,465	85.64
	特許訓練済みモデル	1,143	88.80

ト、特許文書は高分子化合物に関する特許の実施例の箇所にそれぞれ係り受け情報を付与したものである。それぞれのデータの統計値を表 1 に示す。表に示すように、本論文における実験では、理科教科書及び特許文書の量は新聞記事の 10 分の 1 未満と少量である。

3.2 係り受け誤りの収集

2.1 項で述べた係り受け解析器の訓練は以下のように行った。新聞記事の全体の 50 %をベースモデルの訓練データ、残りの 50 %を開発データ兼評価データとして用いた。学習データの中で出現数が 1 回だった単語は UNK トークンに置き換えた。単語の品詞（4 レベル）、活用形、活用タイプはそれぞれ 20 次元に、単語の基本形は 200 次元に埋め込んだ。単語層・文節層 LSTM は隠れ層 1 層（300 次元）のものを用いた。多層パーセプトロンは隠れ層 1 層（300 次元）のものを用いた。訓練アルゴリズムは学習係数 0.01 の AdaGrad で、ミニバッチサイズは 32 文、重み減衰係数は 0.00001、勾配の大きさは 5 に制限して訓練を行った。またエポック数は最大 32 エポックだが、開発データに対する損失が最も少ないものを最終結果とした。実装には Chainer [Tokui 15] を用いた。追加訓練は理科教科書・特許文書それぞれについて 4 分割交差検証の要領で行った。*1

次に 2.2 項で述べた方法で係り受け誤りを収集した。その結果を表 2 に示す。ベースモデルによる新聞記事に対する解析精度は、松野ら [松野 18] と同程度であった。適応先ドメインデータによる追加訓練を行うことで精度が向上していることがわかる。ベースモデルによる解析精度が理科教科書と特許文書で 2.5 ポイントほどの差があるので、特許文書は理科教科書より解析が難しい（新聞記事との違いが大きい）と考えられる。

3.3 クラスタリング結果と分析

2.3 項で述べた方法で新聞記事、理科教科書、特許文書それぞれの係り受け誤りを埋め込み表現に変換した。またクラスタ数を 30 としてクラスタリングを行った。さらに追加訓練前後における誤りの傾向の変化を調べるため、各クラスタに含まれていた追加訓練前の誤りの数・追加訓練後の誤りの数・追加訓練によって解消できた誤りの数を調べた。その結果を図 2（理科教科書）、図 3（特許文書）に示す。グラフの縦軸は各クラスタに含まれる新聞記事誤りの数に対する評価ドメインデータ誤りの比（百分率）を表している。つまり、この比が相対的に高い値になるクラスタは、評価ドメインに特有の誤りの類型に対応すると考えられる。グラフから、ほぼ全てのクラスタにおいて、追加訓練によって誤りの数が減少していることがわかる。またクラスタ間で誤りの分布は一様ではなく偏っており、ドメインごとに固有の分布になっていることもわかる。本論文では、クラスタ内の誤りの傾向が分かりやすかったクラスタ 13 とクラスタ 24 についてその特徴を述べる。以降では係り受け誤りの例を表す時、係り文節を下線、正しい受け文節を ○ のついた波線、誤った受け文節を × のついた波線で表す。また誤り例の通し番号の最初の文字が「N」ならば新聞記事を、「T」ならば理科教科書を、「P」ならば特許文書を表す。

*1 新聞記事・理科教科書は文単位で分割し、訓練に用いたが、特許文書は文書単位で分割した。その理由は特許文書には同じ文書内では似た文が連続するという特徴があり、それらが訓練データと評価データに分かれると、汎化性能を正しく評価できないからである。

表 3: 評価データにおける「…/係り文節/…/名詞＋と/用言」という文型を含む文の数（出現数）と正しい係り先

データ	出現数	正しい係り文節	
		名詞＋と	用言
新聞記事	594	378	216
理科教科書	274	36	238
特許文書	12	5	7

クラスタ 13 クラスタ 13 に属する典型的な新聞記事誤りと理科教科書誤りを以下に示す。

N-13：
立証負担が/軽く、/証拠も/そろえやすい/ことから/東京地検は/「製造の/企て」に/○ 絞り込んだと/× 聞く。

T-13：
ドルトンは，/19世紀の/初めごろ,/物質は/それ以上/分割する/ことの/できない/小さな/粒子で/× できていると/○ 考えた。

クラスタ 13 に属する誤りは、正しい受け文節と誤った受け文節のいずれか一方が引用を表す助詞「と」を含み、他方が直後の動詞であるパターンが多かった。理科教科書ではその中でも誤った受け文節が助詞「と」を含んでおり、正しい受け文節がその直後であるパターンの方が多かった。評価データにおける、この文型の出現数を表 3 に示す。表から、新聞記事では「名詞＋と」の文節にかかる割合がその直後の文節に係る割合の 2 倍程度だが、理科教科書では用言文節にはるかに大きく偏っていることがわかる。これは理科教科書では説明文（「…を/…と/いう」）が非常に多いからである（238 個中 213 個）。ここから、図 2 に見られるようにクラスタ 13 に属する多数の理科教科書誤りのうち、多くが追加訓練で解消されたのは、頻出する説明文の構文を学習したからだと考えられる。またクラスタ 13 に属する特許文書誤りの数が少なかったのは、この文型の出現回数の少なさが原因であることがわかる。

クラスタ 24 クラスタ 24 に属する典型的な新聞記事誤り、理科教科書誤り及び特許文書誤りを以下に示す。

N-24：
ところが/阪神大震災では/一番/重要な/情報収集が/遅れ、/自衛隊の/初動問題に/× 代表されるように/首相官邸だけでなく/政府全体の/対応が/後手に/○ 回った。

T-24：
図1のように，/光源から/出た/光は/四方八方に/広がり,/私たちの/目に/直接/× 届くか,/何かに/当たって/はね返って/私たちの/目に/○ 届く。

P-24：
次いで、/圧力を/常圧から/１３．３ｋＰａに/し、/加熱槽温度を/１９０℃まで/１時間で/○ 上昇させながら、/ 発生する/フェノールを/反応容器外へ/× 抜き出した。

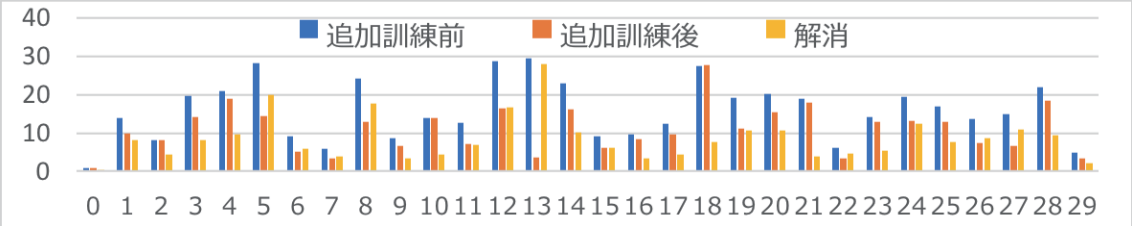

図 2: 追加訓練による、各クラスタに含まれる理科教科書誤りの割合の変化

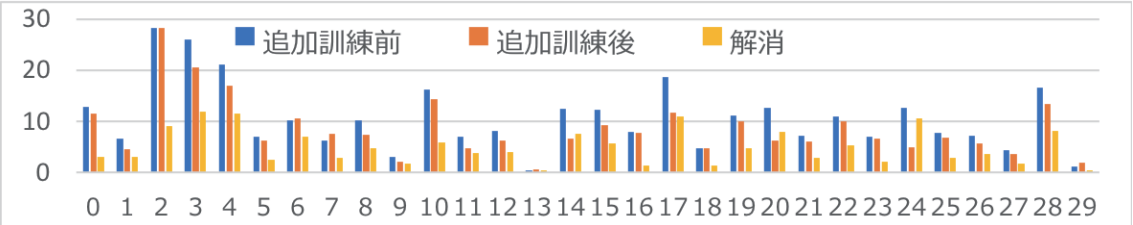

図 3: 追加訓練による、各クラスタに含まれる特許文書誤りの割合の変化

表 4: クラスタ 24 に属する係り受け誤りにおける誤った係り先の位置の分布

データ	ベースモデル		追加訓練済み	
	前	後	前	後
新聞記事	280	398	—	—
理科教科書	83	2	5	37
特許文書	13	59	12	7

クラスタ 24 に属する誤りは係り文節と正しい受け文節、誤った受け文節がすべて用言文節（節の末尾）であるパターンが多かった。3 つの節 A, B, C がこの順番で並んでいる時、意味的な構造は ((A B) C) と (A (B C)) の 2 パターン存在する [*2]。この 2 つの構造のうち、どちらが正しいか判定するためには、本来は節間の関係性を認識する必要があり、難しい判断を必要とする場合も多いと考えられる。しかし、図 2 と図 3 から、追加訓練によって、いずれのドメインでも比較的多くの誤りが解消されていることがわかる。その理由は、ドメインごとに出現しやすい構造に偏りがあるからだと考えられる。実際にそれぞれのドメインでクラスタ 24 に属する誤りについて、正しい受け文節と誤った受け文節のどちらが前方にあるかを調べた。結果を表 4 に示す。理科教科書データの解析では、追加訓練前のモデルは正しい受け文節よりも前方に存在する文節を選択する傾向があったが、追加訓練によりその傾向が改善されたことがわかる。また特許文書は理科教科書と対照的な結果である。ここから、理科教科書は T-24 のような (A (B C)) という形の階層的な並列関係を取る場合が多く、節 A はより遠くの文節 C に係ることが多いと考えられる。一方、特許文書は P-24 のような手続きを表す文を多数含み、3 つ以上の節が単純に並列された構文が多いと考えられる。

4. おわりに

本研究では係り受け解析器の内部状態を用いて係り受け誤りを埋め込み、クラスタリングすることで、追加訓練の効果を、複数のドメインに渡って俯瞰的に分析した。結果として追加訓練によって解消される誤りのタイプをいくつか特定でき、さらに、解析が難しい文の特徴のドメインごとの差を比較することができた。今後の課題としては、解消されない誤りのタイプとその原因の特定、重心から誤りへの距離と誤りの性質の関係の調査、ドメイン適応技術への応用などがあげられる。

*2 T-24 では「図 1 のように，…広がり，」、「私たちの目に直接届くか，」、「何かに…目に届く。」という 3 つの節に分けられるが、最初の節は後ろの 2 つの両方に意味的に係るので、(A (B C)) のパターン。P-24 は節が順番に接続しているので、((A B) C) のパターン。

謝辞

本研究では、JSPS 科研費 16H01819 で開発された教科書アノテーションデータの提供を受けた。ここに記して謝意を表する。また本研究は名古屋大学と富士通研究所の共同研究の成果である。

参考文献

[Daumé III 06] Daumé III, H. and Marcu, D.: Domain adaptation for statistical classifiers, Journal of Artificial Intelligence Research, Vol. 26, pp. 101–126 (2006)

[Daumé III 07] Daumé III, H.: Frustratingly Easy Domain Adaptation, in Proc. of ACL, pp. 256–263, Association for Computational Linguistics (2007)

[Tokui 15] Tokui, S., Oono, K., Hido, S., and Clayton, J.: Chainer: a next-generation open source framework for deep learning, in Proceedings of workshop on machine learning systems (LearningSys) in NIPS 29, Vol. 5, pp. 1–6 (2015)

[黒橋 97] 黒橋禎夫, 長尾眞：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)

[松野 18] 松野智紀, 能地宏, 松本裕治：ニューラルグラフ型係り受け解析器のための文節ベクトル表現, 言語処理学会第 24 回年次大会発表論文集 (2018)

[内元 99] 内元清貴, 関根聡, 井佐原均 他：最大エントロピー法に基づくモデルを用いた日本語係り受け解析, 情報処理学会論文誌, Vol. 40, No. 9, pp. 3397–3407 (1999)