2S-03

深層学習を用いた係り先の文字列予測

吉田　富雅[†, a)]　大野　誠寛[†, b)]　松原　茂樹[‡]

東京電機大学未来科学部[†]　名古屋大学情報連携推進本部[‡]

1. はじめに

同時通訳や字幕生成，音声対話などの音声言語システムでは，入力に追従して処理を行う必要がある．このようなシステムに構文的情報を提供するための漸進的係り受け解析の出力方式として，未入力の文節に係る既入力の文節が複数ある場合は，それらの係り先が同じか否かを明示した係り受け構造を出力する方式が提案されている[1，2]．この出力構造を拡張し，未だ入力されていない係り先文節の内容を予測し提供できれば，音声言語システムにおける処理の同時性は更に向上すると考えられる．

そこで本稿では，この漸進的係り受け解析の出力方式を前提として，未入力の係り先文節の内容を予測することを目標に，その端緒として2つの係り元文節の情報から，それらが共に係る文節を深層学習により予測する手法を提案する．

2. 漸進的係り受け解析の出力構造

大野ら[1]と相津ら[2]の漸進的係り受け解析では，1文の文節列$b_1 \dots b_n$を解析する際，文節b_x $(1 \leq x \leq n-1)$が入力されるたびに，図1のような係り受け構造を出力する．図1の係り受け構造は，文節b_5までに入力された$(x=5)$時の漸進的係り受け解析[1，2]の出力構造を示しており，未入力文節に係る既入力文節(b_2, b_4, b_5)が複数ある場合において，それらの係り先が同一か否か（b_4とb_5の係り先は同一で，b_2の係り先とは異なること）を明示している．この漸進的係り受け解析において，係り先である未入力文節（図1では，未入力文節AとB）の内容を予測することまでできれば，入力文のより詳細な情報を後段の音声言語システムに提供できると考えられる．

本研究では，その端緒として，人手により付与された正しい係り受け構造から，係り先が同一である2つの文節を取り出し，それらの情報から，その係り先文節の主辞[1]を推定することを試みる．これは，図1の例において，文節b_4「上に」とb_5「旗が」の情報から，未入力文節Aの主辞を推定することを模した問題設定である．

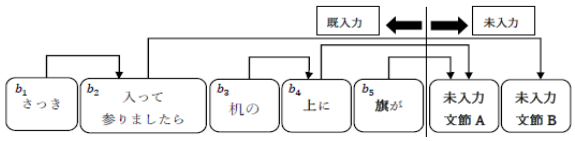


図1　相津らの手法が同定する係り受け構造

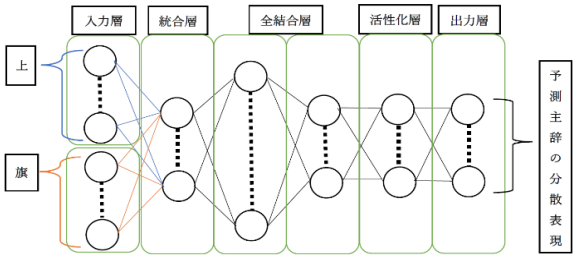


図2　深層学習のネットワーク図

3. 深層学習を用いた係り先の文字列予測

本手法では，同一の係り先を持つ2つの文節を入力として，その係り先の主辞を出力する．

本手法の流れは以下の通りである．

① 同一の係り先を持つ2つの文節の各々から主辞（となる単語）を抜き出す．

②2つの主辞をWord2Vecにより分散表現に変換し，それらを図2に示す深層学習の入力とする．

③ 深層学習の出力（予測した係り先文節主辞の分散表現）と，全単語集合中の各単語の分散表現との間でcos類似度を計測し，類似度が高い上位3つの単語を予測した主辞として出力する．

図2に本手法が用いる深層学習のネットワーク構成を示す．図2では，図1の文節b_4「上に」とb_5「旗が」の2つの文節から，手順①で主辞「上」と「旗」がそれぞれ抜き出され，手順②で変換された分散表現の2つが入力となっている．これら2つの入力層の次元数は主辞（単語）の分散表現の次元数となる．次に，これらの2つの入力を統合層によって統合する．統合層には，加算を採用した．その後，全結合層2層及び活性化層（tanh）を通して，予測した係り先文節の主辞を表す分散表現を出力する．出力層の次元数は，手順③でcos類似度をとるため，単語の分散表現の次元数と同一である．

String Prediction of Modifyee Words using Deep Learning
Fuga Yoshida[†,a)], Tomohiro Ohno[†,b)], Shigeki Matsubara[‡]
† School of Science and Technology for Future Life, Tokyo Denki University.
‡ Information and Communications, Nagoya University.
a) 17fi110@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

[1] 内元らの研究[3]に基づいて設定したもの．

4. 評価実験

本手法の有効性を確認するために，日本語講演データを用いて評価実験を行った．

4.1. 実験概要

実験データは，同時通訳データベース[4]中の日本語講演音声書き起こしテキスト全16講演分から作成した．具体的には，同一の係り先を持つ2つの係り元文節（入力）と，その係り先文節（出力の正解）とから成る組を1セットとし，上述の対象データから14,872セット抽出し実験データとした．このうち，各講演から5セットずつランダムに選んだ80セットをテストデータとし，残りの14,792セットを学習データとした．なお，当該講演データには，形態素情報節境界情報，係り受け情報が人手で付与されている．また，当該講演データにおいて，係り先が同じ文節が3つ以上存在する場合は，その中から2つの係り元文節を選ぶ組合せを考え，各組合せと係り先から成る全セットを実験データに加えた．

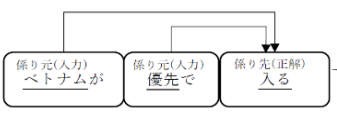
評価には，上位n位までの単語のいずれかと，正解単語とが一致すれば正解とみなした場合の正解率を$ACCn$とし，$ACC1$～$ACC3$を測定した．

Word2vecはGensim[2]を用いて実装した．学習データにはWikipediaコーパス5,923,213文に含まれる8,643,956,674単語を使用した．単語の分散表現の次元数を500，学習に使う前後の単語数の幅を5，イテレーション数とエポック数を15，単語を破棄する最低出現数を0に設定して学習を行った．

深層学習はKeras[3]を用いて実装した．損失関数は平均二乗誤差，最適化にはADAMを採用した．パラメータの更新は，ミニバッチ学習（学習率0.01）により行い，更新時にユニットを0.2の確率でドロップアウトさせた．また，バッチサイズは2,048，エポック数は200とした．深層学習のネットワークの各次元数は，入力層，統合層活性化層，出力層を500次元，全結合層を2層とし，その次元数をそれぞれ800次元，500次元とした．

4.2. 実験結果

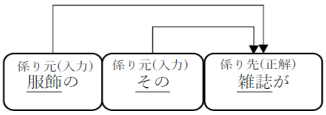
本手法における実験を20回実行した結果，$ACC1$，$ACC2$，$ACC3$の平均値はそれぞれ30.2%，38.7%，40.3%という結果となった．また，正解単語が名詞の場合には名詞が多く出力され，動詞の場合には動詞または助動詞が多く出力されていることから，文の構造を捉えた学習ができていることをうかがえた．以上より，本手法の実現可能性を確認した．



本手法の出力結果

	出力単語	cos 類似度
1位	入る(一致)	0.7195
2位	ない	0.5515
3位	ます	0.5015

図3　本手法の正解例



本手法の出力結果

	出力単語	cos 類似度
1位	問題	0.4222
2位	もの	0.4030
3位	課題	0.3696

図4　本手法の不正解例

正解した結果を分析すると，一般に頻出する「です」や「ます」などの助動詞は，数多く正解できていた．また，図3に示すように，1位の出力単語のみ（「入る」）が，他の出力単語である（「ない」，「ます」）と比べて，一段高いcos類似度となる例が，正解した結果には多く見られた．

一方，学習コーパス中で1，2回程度しか出現しない単語については著しく正解率が低下した．また，図4に示すように，不正解例では，出力単語のcos類似度がいずれも低い傾向にあった．以上より，深層学習の学習データ量が十分でないものと考えられる．

5. おわりに

本稿では，深層学習を用いて2つの係り元文節から係り先文節の主辞を予測する手法を提案した．実験の結果，本手法の実現可能性を確認した．今後は，学習データの拡張．文節を構成する主辞以外の単語を含めた学習などにより更なる性能向上を図りたい．

謝辞　本研究は，一部，科学研究費補助金基盤研究（C）No.19K12127 により実施した

参考文献

[1] 相津ら，“漸進的係り受け解析における未入力文節との構文的関係の同定，” 情報処理学会第82回全国大会講演論文集，pp. 441-442，2020.

[2] 大野，松原，“文節間の依存・非依存を同定する漸進的係り受け解析，”信学論，Vol. J98-D，No. 4，pp. 709-718，2015.

[3] 内元ら，“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，” 情処学論，Vol. 40，No. 9，pp. 3397-3407，1999.

[4] Matsubara et al.，“Bilingual Spoken Language Corpus for Simultaneous Machine Interpretation Research，”Proc. LREC2002，Vol. I，pp. 153-159，2002.

[2] https://radimrehurek.com/gensim/

[3] https://keras.io/ja/