4T-07

日本語文でとりうる係り受け解のとりやすさの数値化

田中 蘭† 松澤 智史†
東京理科大学 理工学部 情報科学科†

1. はじめに

文章推敲支援を目的とする自然言語処理技術を用いた研究はさまざま行われている。

本研究では、文章推敲の際の着眼点の一つである係り受けの曖昧性に着目する。自然言語の文には複数の係り受け解をとりうるものが多く存在する。文に複数の係り受け解が存在する状態を回避しようと文を分割すれば、含まれる情報量が少ない文がたくさん生成されてしまう。

そこで、文に複数の係り受け解が存在する状態を回避する必要はないと仮定する。このとき、それらの係り受け解は「書き手の意図する係り受け解」と「そうでない係り受け解」に分類できる。読み手がそれらの解のどれを正しいと解釈するかといった情報があれば、書き手は文を書き直すかどうかを判断できる。例えば「書き手の意図する係り受け解」のとりやすさが「そうでない係り受け解」のとりやすさと大差ない場合や後者のほうが上回っていた場合は書き直したほうがよいと判断できる。

しかし、関連技術や関連研究では複数の係り受け解に対しそのとりやすさといった情報の提示は行われていない。

本研究では、推敲対象の文を入力した際その各係り受け解について推敲の判断材料にできると感じる数値を出力するシステムを構築する手法について提案する。

2. 基礎知識

2.1 形態素解析

形態素解析とは、単語辞書に記述された単語の品詞や活用形などの情報に基づき、与えられた文や文章を意味をもつ最小単位 (形態素) に分割することである。

2.2 係り受け解析

係り受け解析とは、構文構造の考え方の一つである係り受け構造を求めることである。日本語の場合、文中の各文節がどの文節と依存関係にあるかを明らかにすることがこれにあたる。

2.3 ニューラルネットワーク

ニューラルネットワークとは人間の脳の神経回路構造を模した計算モデルであり、複数のニューロンモデルを結合することで構成される。ニューロンモデルは生体ニューロンの動作を簡単な数式でモデル化したものである。

2.3.1 LSTM

LSTM とは、再帰的な構造により時系列データを扱えるニューラルネットワークの一種である。

3. 関連技術・関連研究

3.1 文章校正・推敲支援ツール Tomarigi

文章校正･推敲支援ツール Tomarigi は、入力文の解析結果に対し、文章を執筆する際の定型的な構成ルールを適用し誤り候補の指摘を行う。このツールでは Cabocha による係り受け解析結果を可視化することで、それが書き手の意図する係り受け解と一致するかどうかを確認することが容易である。Cabocha が出力した解以外にどのような係り受け解があるかどうかの提示は行わない。[1]

3.2 文章推敲支援における読み手に誤解される文の抽出

菅沼らは、係り先を複数持つ文節が含まれる文を「誤解される文」とし、文節のランク付けによる制約や共起情報による絞り込みを導入することで、誤解される文であるかどうかの判定と誤解の原因となる文節の指摘を行う手法を提案した。この研究では起こりうる誤解の度合いに関する情報の提示は行っていない。[2]

4. 提案手法

4.1 概要

本研究の提案手法では、推敲対象を想定した文を入力し、その文がとりうる係り受け解と各係り受け解のとりやすさを想定した数値を出力する。

入力を受け取ってから各係り受け解のとりやすさを出力するまで

1. とりうる係り受け解を列挙
2. 係り受け解とそのとりやすさを学習したと想定されるモデルへ、各係り受け解を入力し、そのとりやすさを予測

という過程を経る。

4.2 とりうる係り受け解の列挙

入力文を係り受け解析器 Cabocha で解析し、各単語の基本形や品詞情報、各文節においてどの形態素が主辞であるかという情報を取得する。これをもとに各文節に対し主辞の品詞を割り当てる。そして中学国語文法に基づいた品詞の制約 (例:連体形は体言を修飾する, 副詞は用言を修飾する等) や、

Quantifying the possibility of dependency solutions of Japanese text
†Ran Tanaka, †Tomofumi Matsuzawa
†Faculty of Science and Technology Dept.of Information Sciences, Tokyo University of Science

1. 後方修飾性
   各文節は後方の文節を修飾する
2. 非交差性
   係り受け関係は交差しない
3. 係り先の唯一性
   係り要素は受け要素を一つだけ持つ

といった日本語の性質を用いて、とりうる係り受け解を列挙する。

また、本研究では図 1 のように文に対して文節番号を与えたうえで各文節に係り先文節番号を与えることで係り受け解を表現する。

(0)背泳ぎで (1)泳いでいる (2)少年を (3)追いかけた

解1: 背泳ぎで(3) 泳いでいる(2) 少年を(3) 追いかけた
解2: 背泳ぎで(1) 泳いでいる(2) 少年を(3) 追いかけた

図 1: 係り受け解の表現

4.3 LSTM を用いた機械学習

本研究では学習のためのテキストデータとして産経ニュースの記事を用いる。

1. 形態素ごとに数値へ変換したテキストデータ
2. 係り先文節番号を想定した数値列

のそれぞれを入力とし、図 2 のモデルで学習する。

また、「新聞の文は係り受けの曖昧な文が無い」「係り受け解析器 Cabocha は新聞の学習をもとに作られたため新聞の文に対してはおおむね正しい係り受け解析結果が得られる」と仮定する。これによって、各係り受け解へのラベルの与え方について Cabocha の出力した係り受け解に正解ラベル 1 をつけ、それ以外の係り受け解に不正解ラベル 0 をつけている。

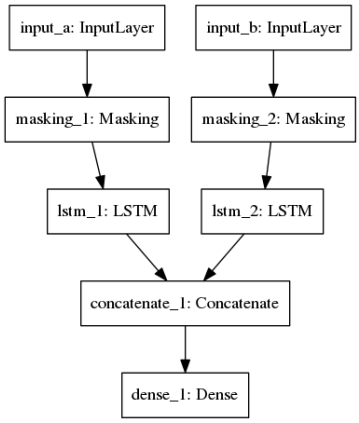

図 2: 学習モデル

4.4 各係り受け解のとりやすさの提示

システムへの入力文に対し、4.3 の要領で各係り受け解をそれぞれ学習済みネットワークへ入力する。それらの出力に対して (単一の係り受け解のとりやすさ)=(単一の係り受け解の出力)÷(各係り受け解の出力の総和) とすることで単一の係り受け解のとりやすさを列挙したものを本システムの最終的な出力とする

5. 評価方法

複数の係り受け解が存在する文を実験に用いる。それらの文についてどの解が正しいと感じるか、またそれらの文をシステムに入力した際の出力結果について妥当であるまたは推敲の判断材料にし得ると感じるかどうかといった項目でアンケートをとる。

6. 結果

図 1 に用いた文を入力した際の結果が図 3 である。

解1: 背泳ぎで(3) 泳いでいる(2) 少年を(3) 追いかけた … 42.13 %
解2: 背泳ぎで(1) 泳いでいる(2) 少年を(3) 追いかけた … 57.87 %

図 3: 出力結果の一例

この結果を見る限りではより距離が近い文節に係りやすいと学習できているようである。

この文についてどちらの解が正しいと感じるかアンケートを実施したところ、解 1 を正しいと感じると回答した人の割合は 76.92%、解 2 を正しいと感じると回答した人の割合は 23.08%であり、システムの出力とは乖離がみられる。この乖離の一因として学習データ作成の自動化のために係り受け解析器の出力した係り受け解に対して一律に正解ラベルを与えたことが考えられる。

また、その他の文についても同様にアンケートをとった結果を提示したうえで、このシステムの出力結果を妥当であると感じるかまたは推敲の判断材料にし得るかどうかについて再度アンケートを実施する予定である。

7. まとめ

本研究では、日本語文でとりうる係り受け解のとりやすさの数値化について深層学習を用いた手法を検討した。結果として推敲の判断材料にし得る数値化をできたとは言い難い。しかしより厳密なデータの整備による精度の向上は期待できるため、今後はさらにそちらにも注力する予定である。

参考文献

[1] 大野博之・稲積宏誠 (2011)「日本語文章校正・推敲支援ツール:Tomarigi:添削作業の支援と省略化に向けた取組概要」,『日本教育工学会研究報告集』No.1,pp.325-332, 日本教育工学会

[2] 菅沼明・小野貴博 (2007)「文章推敲における読み手に誤解される文の抽出」,『情報処理学会研究報告』No.50,pp.31-38, 情報処理学会.