様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２７ 日現在

機関番号：１６３０１
研究種目：若手研究
研究期間：2018～2019
課題番号：１８Ｋ１８１１０

研究課題名（和文）異構造の言語間翻訳の精度改善のための構文森に基づくニューラル機械翻訳の研究

研究課題名（英文）Forest-based neural machine translation for improving translation performance between distant language pairs

研究代表者

田村　晃裕（Tamura, Akihiro）

愛媛大学・理工学研究科（工学系）・助教

研究者番号：２０８０４１６５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、機械翻訳の手がかりとして翻訳元の文の構文森（構文解析結果の複数の解析結果を圧縮して表現したグラフ）を使うことで、ニューラルネットワークに基づく機械翻訳（NMT）の性能改善を実現した。実施期間中に、二つのNMTモデル（再帰型ニューラルネットワークに基づくNMT及びTransformerに基づくNMT）で構文森を活用する手法を開発した。そして、評価実験を通じて、構文森を活用することで日英翻訳の性能が改善できることを示すとともに、最先端の従来手法の性能を凌駕する日英翻訳性能を実現した。

研究成果の学術的意義や社会的意義
近年のグローバル化の進展とともに、外国語の利活用を支援する機械翻訳の需要が高まっている。しかし、機械翻訳では、構造が異なる言語間の翻訳は難しく、その翻訳性能の改善が大きな課題の一つとなっている。本研究では、その課題を解決するため、翻訳元の文の構文森の情報をNMTで活用する初めての試みに取り組んだ。そして、構文森を活用することにより、構造が異なる言語間の代表例である英語と日本語間の翻訳性能を改善できることを示し、今後の機械翻訳の研究開発において、構文森を活用する重要性を示唆した。

研究成果の概要（英文）：This research aims to improve the performance of neural network-based machine translation (NMT) by using a source-side syntax forest, which is a compact representation of many parse trees. The study has proposed a novel method for incorporating source-side syntax forests into recurrent neural network-based NMT and Transformer-based NMT. The evaluations show that the syntax forests improve the English-Japanese translation performance and the proposed model achieves a state-of-the-art performance.

研究分野：情報学－人間情報学－知能情報学

キーワード：ニューラル機械翻訳　構文森　ニューラルネットワーク　機械翻訳　Transformer　RNN

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

## １．研究開始当初の背景

近年、グローバル化の進展とともに外国語に接する機会が増加しており、外国人とのコミュニケーションや外国語による情報の利活用等を補助する高精度な機械翻訳の需要が高まっている。機械翻訳は、自然言語処理の研究分野において古くから研究されており、近年ではニューラルネットワークに基づく機械翻訳（NMT）が出現したことで、翻訳精度が大きく向上し、世の中への普及が進んでいる。

しかし、NMT を含めた機械翻訳では、語順などの構造が似ている言語間（例えば、英語とフランス語）の翻訳に比べて、構造が異なる言語間（例えば、英語と日本語）の翻訳は難しく、その翻訳精度の改善が大きな課題の一つとなっている。そこで、構造が異なる言語間の翻訳精度を向上させるため、原言語（翻訳元の言語）あるいは目的言語（翻訳先の言語）の文構造を活用する NMT がここ数年で提案されている。しかし、従来の文構造を活用する NMT は、まず、構文解析器を用いて原言語の文や目的言語の文の構文解析を行い、構文解析の結果、最適解となった文の構文木を NMT で活用するため、解析結果の構文木に誤りが含まれていると翻訳精度が低くなるという問題があり、改善の余地を大きく残している。

## ２．研究の目的

本研究では、構造が異なる言語間の翻訳精度を向上させることを目的として、原言語の文を構文解析した結果得られる最適解に加えて可能性のある複数の文構造を表現する構文森を文構造の情報として NMT で活用する方法を創出することで、構文解析誤りに頑健な文構造に基づく NMT の実現を目指す。

これまで構文森を NMT で活用する試みは行われていない。そして、NMT で構文森を有効に活用する方法は自明ではない。そこで本研究では、NMT における構文森を活用する方法として、「系列化した構文森を活用する方法」と「構文森を直接エンコードする方法」の二つの方法を実現して比較することで、NMT における構文森の有効な活用方法の検討も行う。

構造が異なる言語間の翻訳実験として英日翻訳実験を行い、構文森を活用することで構造が異なる言語間の翻訳精度が向上するかどうかを明らかにする。また、系列化した構文森を活用する方法と構文森を直接エンコードする方法のどちらが NMT において有効であるかを実験的に確認する。

## ３．研究の方法

### (1) 構文森に基づく NMT の開発

本研究では、提案手法で活用する構文森は既存の構文解析手法[1]を用いて獲得する。また、提案手法のベースライン手法として、研究開始当初に標準的であった再帰型ニューラルネットワーク（RNN）に基づく NMT[2]と、近年多くの翻訳設定で最高精度を達成している Transformer に基づく NMT[3]を採用し、これらの手法を拡張することで提案手法を開発する。

### (2) 英日翻訳実験による評価

英日翻訳実験では、開発した提案手法の翻訳性能評価とともに従来手法の翻訳性能も評価し、それらを比較することにより、構文森の活用が、構造が異なる言語間の翻訳精度の改善をもたらすかを検証する。提案手法は、構文木（従来活用されている文構造）に基づく NMT と文構造を利用しない単語列系列に基づく NMT の両者と比較する。英日翻訳実験は、評価型ワークショップ Workshop on Asian Translation 2017 の Scientific paper Subtasks の英日翻訳の設定で行う。

## ４．研究成果

(1) 系列化した構文森を活用する NMT を創出した。具体的には、提案手法では、原言語の文を構文解析して獲得した構文森を系列データに変換し、変換した構文森の系列データを入力として RNN に基づく NMT で翻訳を行う。本手法では、構文森を系列データに変換する方法が必要となるが、その方法はこれまで研究されていない。そこで本研究では、単語間の文における前後関係と構文森における親子関係を反映する系列化手法を開発した。また、構文森の情報を効果的に活用するために、構文森に付与されている解析の確信度を表すスコアを、RNN に基づく NMT の単語埋め込み層や注意機構内で反映して翻訳を行う方式も開発した。これら提案手法の具体的なアルゴリズムや方法は ACL2018 にて発表した。

(2) 研究成果(1)記載の提案手法の英日翻訳性能を評価し、構文木を用いる場合と単語系列を用いる場合の性能と比較した。実験結果を表 1 に示す。なお、表 1 の全ての手法は RNN に基づく NMT である。表 1 より、提案手法の BLEU（翻訳性能の評価指標）は、構文木を用いた場合より 2.75 ポイント、文構造を利用せずに単語系列を用いた場合より 5.07 ポイント高

表 1: 系列化した構文森を活用する NMT の有効性

| 手法 | 文字単位 BLEU |
|---|---|
| 単語系列に基づく NMT | 37.10 |
| 構文木に基づく NMT | 39.42 |
| 系列化した構文森に基づく NMT | 42.17 |

いことが分かる。この実験結果により、系列化した構文森を NMT で活用することで英日翻訳性能が改善することを実験的に示した。

また、構文森スコアの反映方式を使用しない提案手法の翻訳性能も評価した。実験結果を表 2 に示す。この実験結果により、構文森を NMT で有効活用するためにはスコアを反映する必要があることを示した。また、RNN に基づく NMT における構文森スコアの活用方法は、単語埋め込み層で反映するよりも注意機構内で反映する方が効果的であることを示した。

表 2: 構文森スコアの反映方式の有効性

| 構文森スコアの反映方式 | 文字単位 BLEU |
|---|---|
| 構文森スコア不使用 | 37.92 |
| 単語埋め込み層で反映 | 41.35 |
| 注意機構で反映 | 42.17 |

(3) 構文森をエンコーダ内で直接エンコードする NMT を創出した。具体的には、Transformer に基づく NMT をベースとし、その特徴である位置エンコーディング及び自己注意機構の枠組みの中で構文森をエンコードする方法を開発した。また、研究成果(2)により、構文森を活用するためには構文森のスコアを反映することが重要であることが分かったため、構文森をエンコードする自己注意機構の中で構文森スコアを反映する方式を開発し、本手法に組み込んだ。これら提案手法の具体的な方法は会誌「自然言語処理」で発表した。

(4) 研究成果(3)記載の提案手法の英日翻訳性能を評価し、構文木を用いる場合と単語系列を用いる場合の性能と比較した。実験結果を表 3 に示す。なお、表 3 の全ての手法は Transformer に基づく NMT である。表 3 より、提案手法の BLEU は、構文木を用いた場合より 0.6 ポイント、文構造を利用せずに単語系列を用いた場合より 1.99 ポイント高いことが分かる。この実験結果により、Transformer に基づく NMT のエンコーダ内で構文森を直接エンコードすることで英日翻訳性能が改善することを実験的に示した。

表 3: 構文森を直接エンコードする NMT の有効性

| 手法 | 文字単位 BLEU |
|---|---|
| 単語系列に基づく NMT | 34.42 |
| 構文木に基づく NMT | 35.81 |
| 構文森を直接エンコードする NMT | 36.41 |

(5) 表 1 と表 3 より、研究成果(1)記載の提案手法の性能改善幅の方が、研究成果(3)記載の提案手法の性能改善幅より大きいことが分かる。また、翻訳精度自体も研究成果(1)記載の提案手法の方が高い。これらの実験結果により、実施した英日翻訳実験の設定においては、構文森をエンコーダ内で直接エンコードするよりも系列化した構文森を活用する方が有効であると考えられる。ただし、本比較においては、ベースとなる NMT が RNN に基づく NMT と Transformer に基づく NMT とで異なること、また、実験した言語対や実験設定（実験データやデータサイズの多様性）が少ない点は注意されたい。今後は、ベースラインを統一し、様々な実験設定において「系列化した構文森を活用する方法」と「構文森を直接エンコードする方法」を比較した上で、有効な構文森の活用方法を模索する必要があると考える。

(5) 提案手法と従来の構文情報に基づく NMT 及び単語系列に基づく NMT の性能比較を行った。結果を表 4 に示す。表 4 に示されている通り、本研究により、最先端の従来 NMT の日英翻訳性能を凌駕する NMT を実現した。

表 4: 従来の NMT との性能比較

| 活用する文構造 | 手法 | 文字単位 BLEU |
|---|---|---|
| なし | Bahdanau et al., 2015 | 37.10 |
| なし | Vaswani et al., 2017 | 34.09 |
| なし | Shaw et al., 2018 | 34.42 |
| 構文木 | Eriguchi et al., 2016 | 37.52 |
| 構文木 | Chen et al., 2017 | 36.94 |
| 構文木 | Li et al., 2017 | 36.21 |
| 構文森 | 系列化した構文森に基づく NMT | 42.17 |

＜引用文献＞

[1] Huang, L. Forest Reranking: Discriminative Parsing with Non-Local Features. In the Proc. of ACL-08: HLT, pp. 586-694, 2008.

[2] Bahdanau, D., Cho, K, Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015, 2015.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., Attention is All You Need, In the Proc. of NIPS 2017, pp. 5998-6008, 2017.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　3件／うちオープンアクセス　5件）

| | |
|---|---|
| １．著者名<br>Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Eiichiro Sumita, Tiejun Zhao | ４．巻<br>27(2) |
| ２．論文標題<br>Syntax-based Transformer for Neural Machine Translation | ５．発行年<br>2020年 |
| ３．雑誌名<br>Journal of Natural Language Processing | ６．最初と最後の頁<br>- |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>該当する |

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　5件）

| |
|---|
| １．発表者名<br>Chunpeng Ma, Akihiro Tamura, Masao Utiyama, Tiejun Zhao, Eiichiro Sumita |
| ２．発表標題<br>Forest-Based Neural Machine Translation |
| ３．学会等名<br>56th Annual Meeting of the Association for Computational Linguistics（国際学会） |
| ４．発表年<br>2018年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|