# 1　これまでの修学内容

私は高専ではメカトロニクスを学修し、現在大学では情報工学を学修している。大学の卒業研究では自然言語処理に関する研究を行っている。以下では、現在取り組んでいる卒業研究のテーマである「日本語の膠着語的性質を考慮したマルチタスク学習によるニューラル機械翻訳」について述べる。

近年、ニューラル機械翻訳 (NMT) は非常に高いパフォーマンスを発揮している。しかし、Pan et al. (2020) は形態学的に複雑 (morphologically-rich) な膠着語の翻訳は NMT モデルにとって依然として困難なタスクであると指摘している。Pan et al. (2020) はこの問題を解決するために、トルコ語-英語およびウイグル語-中国語の双方向翻訳タスクにおいて、膠着語であるトルコ語とウイグル語のステミングタスクを補助タスクとしてマルチタスク学習を行うことで、翻訳性能が向上することを示した。トルコ語やウイグル語の単語は典型的には 1 つの語幹 (ステム) にいくつかの接辞が膠着している構造であり、ステミングタスクは単語の接辞を除去して語幹のみを抽出するタスクとして定義されている。

私は膠着語のなかでも特に日本語に着目し、Pan et al. (2020) の手法を日本語と英語の双方向翻訳に適用することで、翻訳性能の向上を試みる。日本語の文節は典型的には 1 つの自立語に 0 個以上の付属語が膠着することで形成される。このことから、日本語のステミングタスクは文節中の自立語の抽出および語の原形化と定義した。また、マルチタスクニューラルモデルとして Transformer (Vaswani et al., 2017) を用いる。タスク毎に固有のタグを各ソース文の先頭に加える方法 (Johnson et al., 2017) を用いることで単一のモデルでマルチタスク学習を行う。タスクを示す固有のタグとして <ST>(ステミングタスク) と <MT>(翻訳タスク) を用いる。図 1 にモデルの概念図、表 1 に各タスクの訓練データの例文を示す。

以上のモデルの翻訳性能を BLEU などによって測定し、ステミングタスクを用いない翻訳モデルと比較することで本手法の有効性を評価する。また、文節チャンキングタスクの付加やサブワード分割の適用、単言語資源の活用などによる拡張を検討している。

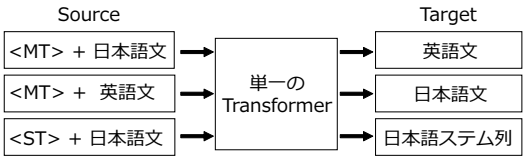


図 1: モデルの概念図

表 1: 各タスクの訓練データの例文 (上:ソース文、下:ターゲット文)

| タスク | 訓練データの例文 |
|---|---|
| 日英翻訳 | <MT> 私 は リンゴ を 食べ させ られ ました 。 |
| | I was forced to eat an apple. |
| 英日翻訳 | <MT> I was forced to eat an apple. |
| | 私 は リンゴ を 食べ させ られ ました 。 |
| 日本語ステミング | <ST> 私 は リンゴ を 食べ させ られ ました 。 |
| | 私 リンゴ 食べる |

# 2　NAIST で取り組みたい研究

## 2.1　はじめに

私は NAIST で言語的な構造や性質に関する情報を NMT へ応用する研究に取り組みたいと考えている。以下では、具体的なテーマとして「ツリーバンクを活用した Syntax-BERT-fused モデルによるニューラル機械翻訳」を挙げ、その内容について述べる。

## 2.2　背景・関連研究

現在主流の NMT モデルは、前章で述べたモデルも含めて、入力される文の文法構造を明示的には考慮しない (Nguyen et al., 2020)。しかし、自然言語の文中の単語は表層的な順序だけではなく、潜在的なネスト構造によっても体系づけられる。このネスト構造は多くの場合木構造 (構文木) で表現される。したがって、構文木を明示的に考慮できるようにすることは、構文のより正確な把握や構文的に正しい翻訳の出力を促し、NMT の性能向上に寄与することが期待できる。実際、 Eriguchi et al. (2017) は句構造解析と文生成の同時学習モデルである RNNG (Dyer et al., 2016) を用いて目的言語の係り受け構造を明示的に考慮する NMT モデルの RNNG+NMT を提案し、翻訳性能が向上することを示した。また、Nguyen et al. (2020) は句構造木を明示的に考慮する Attention 機構を組み込んだ Transformer である Tree Transformer を提案し、このモデルで NMT を行うことにより翻訳性能が向上することを示した。

他方、近年の自然言語処理分野では事前学習モデルの研究が盛んである。代表的なものとして、Masked Language Modeling (MLM) と Next Sentence Prediction (NSP) によって事前に学習を行う Transformer ベースの事前学習モデル BERT (Devlin et al., 2018) が挙げられる。事前学習モデルに構文情報を組み込む研究も行われており、Jiangang et al. (2021) は学習済み BERT に句構造木と依存構造木の情報を考慮する fine-tuning を施した Syntax-BERT を提案し、含意関係認識や自然言語推論などの下流タスクにおける有効性を示した。Syntax-BERT などの構文情報を考慮する言語モデルの学習には、各文にその構文情報が付与されたコーパスであるツリーバンクを用いることができる。日本語のツリーバンクは Kaede treebank[1]や NPCMJ[2] などが利用可能である。ツリーバンクはしばしば構文解析の golden-standard として扱われ、構文解析器によって得た構文情報より質が高いことが期待される。

事前学習モデルの隆盛を受けて、NMT に事前学習モデルを利用する手法も研究されている。その手法のひとつに単言語資源で学習した事前学習モデルを NMT モデルに組み込むというものがある。Zhu et al. (2020) はソース文を入力した学習済み BERT の最終層の出力を NMT モデルに入力する BERT-fused モデルを提案し、翻訳性能が向上することを示した。翻訳性能が向上した理由は、この手法によって BERT に蓄えられた知識を NMT にうまく活用できたためであると考えることができる。

[1] https://github.com/mynlp/kaede
[2] http://npcmj.ninjal.ac.jp/

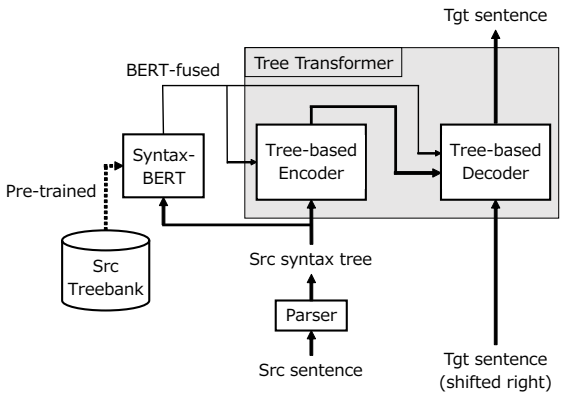


図 2: Syntax-BERT-fused モデルの概略図

### 2.3　提案手法 : Syntax-BERT-fused モデルによる NMT

構文情報を明示的に考慮する NMT では、訓練データとして { 原言語文, 目的言語文, 構文情報ラベル[3]} の 3 つ組が含まれるコーパスを使用することが理想的である。しかし、実際にはそのような 3 つ組のコーパスが利用可能であることはほとんどない。そのため、前節で述べた RNNG+NMT や Tree Transformer を用いた NMT では、既存の対訳コーパスを構文解析器によって構文解析することで疑似的に 3 つ組コーパスを作成して、そのコーパスをモデルの訓練データとして用いている。この疑似 3 つ組コーパスでは原言語文と目的言語文の対の品質は高いことが期待されるが、文と構文情報の対の品質については必ずしも高くないという欠点がある。

そこで私は、品質の高い文と構文情報の対であるツリーバンクによって学習した構文情報を考慮する事前学習モデルを用いて前述の欠点の補償を試みる。具体的には、事前学習[4]した Syntax-BERT を BERT-fused モデルの手法によって Tree Transformer に結合させた Syntax-BERT-fused モデルを提案する。このモデルでは Syntax-BERT に蓄積されたツリーバンクの知識を構文情報を考慮する NMT に活用でき、欠点の補償につながると考える。提案モデルの概略図を図 2 に示す。提案モデルでは以下の手順によって翻訳を行う。

1. 原言語 (Src) のツリーバンクで Syntax-BERT を学習する
2. 構文解析器によって原言語文を解析し構文木を得る
3. 原言語文の構文木を Syntax-BERT に入力する
4. Syntax-BERT の最終層の出力と原言語の構文木を Tree Transformer のエンコーダ (Tree-based Encoder) に入力する
5. Syntax-BERT の最終層の出力と Tree-based Encoder の出力と右シフトした目的言語 (Tgt) の出力文を Tree Transformer のデコーダ (Tree-based Decoder) に入力する
6. Tree-based Decoder が目的言語の文を出力する

Syntax-BERT-fused モデルによる NMT の翻訳性能を BLEU などによって測定し、Tree Transformer による NMT の翻訳性能と比較することで、提案モデルの有効性を評価する。

[3]原言語文と目的言語文のどちらの構文情報のラベルが必要であるかは手法によって異なり、両方の構文情報ラベルが必要な場合も考えられる

[4]学習済みの原言語 BERT にツリーバンクを用いた MLM と NSP の 2 つのタスクで fine-tuning を施す

### 2.4　検討事項

前述の提案手法以外に疑似 3 つ組コーパスの欠点を補償する方法として、適当な NMT モデルで原言語ツリーバンクの各文を翻訳し、それによって得た { 原言語文, 目的言語文, 構文情報ラベル } の組を別の疑似 3 つ組コーパスとして追加で用いるという方法を検討している。この別の疑似 3 つ組コーパスでは原言語文と目的言語文の対の品質は必ずしも高くないが、原言語文と構文情報の対の品質については高いことが期待される。そのため、対訳コーパスを構文解析することによって得た疑似 3 つ組コーパスと合わせて使用することによりそれぞれが相補的な役割を果たすと考える。なお、この手法に比べて、前節の提案手法ではツリーバンクの知識の活用のために事前学習モデルを用いるため、NMT モデルのアーキテクチャの変更や対訳コーパスの変更の際に以前学習した事前学習モデルを再利用することができるという利点がある。

構文木は大別して句構造木と依存構造木があり、そのなかでも様々な表現形式が存在する。そのため、どのような表現形式が提案手法に有効であるかを検討する必要がある。

モデルの翻訳性能は BLEU による評価以外にも、出力文の流暢性や構文的正しさの評価、ケーススタディによる定性的な評価など、多元的な評価を行うことを検討している。

以上のように、提案した手法あるいは評価方法には様々なバリエーションが考えられるため、適宜検討しながら研究を進めていきたいと考えている。

### 2.5　おわりに

前節までに、言語的な構造や性質に関する情報を NMT に効果的に活用するための研究テーマについて述べてきた。私は、多様なバックグラウンドを持つ学生の集う NAIST で様々な議論を通じて研究を進め、ひいては自然言語処理および言語学の分野に貢献していきたいと考えている。

## 参考文献

Pan Yirong, Li Xiao, Yang Yating and Dong Rui. 2020. Multi-Task Neural Model for Agglutinative Language Translation. In *Proceedings of ACL: Student Research Workshop*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Proceedings of ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Akiko Eriguchi, Yoshimasa Tsuruoka and Kyunghyun Cho. 2017. Learning to Parse and Translate Improves Neural Machine Translation. In *Proceedings of ACL (Volume 2: Short Papers)*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of NAACL*.

Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi and Richard Socher. 2020. Tree-Structured Attention with Hierarchical Accumulation. *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Bai Jiangang, Wang Yujing, Chen Yiren, Yang Yaming, Bai Jing, Yu Jing and Tong Yunhai. 2021. Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees. In *Proceedings of EACL*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li and Tieyan Liu. 2020. Incorporating BERT into Neural Machine Translation. *ICLR*.