

特許審査官が垣間見た深層学習 (Deep Learning) による自然言語処理の変革期 ～ Japio 特許情報研究所での業務を通じて～

審査第四部情報処理 石川 雄太郎

抄録

本稿では、深層学習を用いた機械翻訳や機械読解等の自然言語処理技術について解説し、その発展の背後に存在するオープンイノベーションの環境についても説明する。さらに、AIを用いた特許検索システム等についても紹介したい。

1. はじめに～深層学習 (Deep Learning) が与えたインパクト～

無料で利用できる機械翻訳を介して、世界中の情報を非常に流暢な和訳で読むことができる時代となった。Google 翻訳は、数年前までたどたどしい翻訳を生成していたが、今やよく読まなければ、人間による翻訳と見分けがつかないレベルの翻訳を作り出せるようになってきている。さらに2020年には、ドイツ発の高性能機械翻訳であるDeepL翻訳に日本語翻訳が追加され、より流暢な和訳を読むことができるようになった¹⁾。

機械翻訳の流暢性や正確性が飛躍的に上がったことから、出願人が明細書を翻訳する際にも変化の兆しが見えており、この数年で、一から人間が翻訳するのではなく、機械翻訳で生成された文を人間が編集することで (MTPE: Machine Translation Post-Editing)、コストを削減するケースが増えてきている²⁾。特許審査においては、従前より、特に大量の先行技術文献の中から関係のある文献のあたりをつける目的で、機械翻訳による外国語の特許文書の和訳が欠かせない存在であったが、より正確な翻訳が求められる出願書類についても、人間による翻訳の

下訳として、機械翻訳の活用が様々な形で検討され始めている。

機械翻訳の精度が劇的に向上した鍵は、深層学習の適用である。2014年に深層学習を適用した機械翻訳であるニューラル機械翻訳 (NMT: Neural Machine Translation) が発表されて以来³⁾、様々な改良が重ねられ、従来の統計的機械翻訳 (SMT: Statistical Machine Translation) を置き換えてしまいつつある⁴⁾。また、ニューラル機械翻訳のモデルも日々革新が進んでいる⁵⁾。このように、深層学習は、自然言語処理 (日常的に使われている自然発生的に生まれた言語のコンピュータ処理³⁾⁶⁾ の分野において、多大なインパクトを与えており、我々特許審査官が日常携わる審査業務においても、この大きな変革の波が確実に影響を与えていくと考えられる。

筆者は、2018年7月～2020年6月まで、(一財)日本特許情報機構 (Japio) の特許情報研究所に出向する機会をいただいた。Japioでは、2018年2月に知財AI研究センター (Japio 頂) が設立され、特許・商標情報分野における人工知能 (AI: Artificial Intelligence) 技術の実用化に向けた研究・開発を加速することとなり、筆者は、特許情報処理に応用できる可能性の高い人工知能の情報を広く収集すると

1) <https://www.deepl.com/ja/blog/20200319.html>

2) 阪本章子, 山田優 (2019) 「巻頭特集: AI時代の産業翻訳」, 『JTFジャーナル』2019年11/12月号

3) 黒橋禎夫 (2019) 『改訂版 自然言語処理』一般財団法人放送大学教育振興会

4) <https://www.nict.go.jp/press/2017/06/28-1.html>

5) Vaswani, A., et al. (2017) .Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008) .

6) 吉村賢治 (2000) 『自然言語処理の基礎』サイエンス社

共に、それらを使ったシステムの研究開発業務に従事した。特に、近年深層学習が人工知能の主流になっていることや、また、機械翻訳に関する研究が古くから Japio のコアテーマの1つであることから、深層学習による機械翻訳については、力を入れて情報収集と研究開発を行うこととなった。そこで見てきたのは、Google の開発した Transformer モデル⁶⁾が、機械翻訳のみならず、機械読解にも大きな影響を与えて、その応用範囲が広がってきており⁷⁾、自然言語処理が大きな変革期を迎えているということであった。そして、その背後では、最前線の企業や研究機関が、論文、OSS (Open Source Software)、データセットを共有して、世界中の知恵を結集することで、革新への大きな推進力となっている点も垣間見えた。

本稿では、特許情報処理において、現在注目されている深層学習を用いた技術と、将来発展することが予想され、今後利用が拡大されることが期待される技術を紹介する。特に、出向中の経験を生かして、深層学習を用いた機械翻訳や機械読解等の自然言語処理技術について解説し、その発展の背後に存在するオープンイノベーションの環境についても経験に基づいて説明する。その他、出向中に触れた AI を用いた特許検索システム等についても述べたい。なお、本稿で示した見解はすべて筆者個人の見解であ

り、筆者の所属する組織、または、以前所属していた組織としての見解を示すものではない。

2. 深層学習を適用した機械翻訳～ニューラル機械翻訳～について

図1に機械翻訳システムの現在までの変遷を示す。機械翻訳システムの主流は、後述するルールベース機械翻訳 (RBMT: Rule Based Machine Translation)、統計的機械翻訳 (SMT)、ニューラル機械翻訳 (NMT) と移り変わっており⁴⁾、ニューラル機械翻訳の中でも、モデルが進歩して、後述する RNN ベースの sequence to sequence モデルから Transformer モデルに移り変わっている⁵⁾。機械翻訳は新たな方式を採用する度に性能が上がっていき、特に、ニューラル機械翻訳の登場によって、流暢性や正確性が劇的に向上している。

2006年に、ジェフリー・ヒントン博士等の論文“A fast learning algorithm for deep belief nets.”によって、深い層のニューラルネットワークが効率的に学習できることが示され、深層学習という言葉の利用が普及してから、各分野で深層学習を用いた驚異的な成果が発表されてきたが⁸⁾、機械翻訳においても、2014年に深層学習を適用した機械翻訳であるニューラル機械翻訳が発表され、様々な改良を経た後、

RBMTからSMT/NMTへ

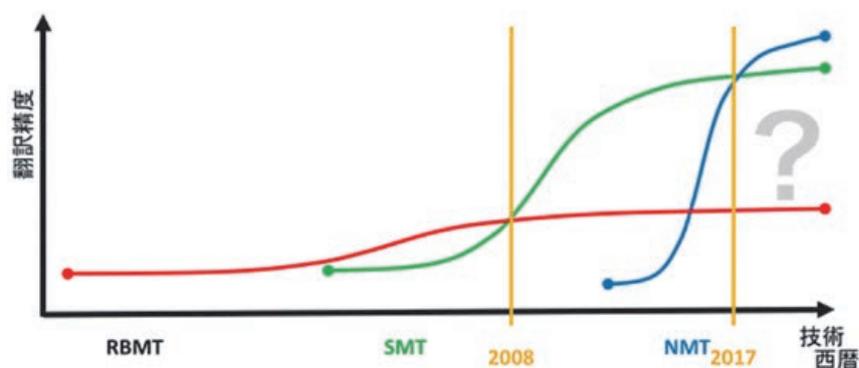


図1 機械翻訳システムの変遷 (NICTのプレスリリースより⁴⁾)

7) Kenton, J. D. M. W. C., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

8) Ian Goodfellow, 他 (2018)『深層学習』KADOKAWA

2015年には、WMT2015 (Workshop on Statistical Machine Translation : 世界的に有名な機械翻訳の評価型ワークショップ) にて、カナダ モントリオール大学のチームが、英独翻訳で、ニューラル機械翻訳の一つのモデルであるRNNベースのsequence to sequenceモデル(RNNsearch architecture)を用いて、当時主流であった統計的機械翻訳を上回る性能を発揮して、1位を獲得した⁹⁾。以来、2016年にGoogle翻訳が、統計的機械翻訳からニューラル機械翻訳に置き換わる¹⁰⁾等、ニューラル機械翻訳が主流になっている。さらに、2017年には、Googleから“Attention is all you need”という論文で、ニューラル機械翻訳の新たなモデルであるTransformerモデル⁵⁾が発表され、今日ではRNNベースのsequence to sequenceモデルに代わって主流になっている。なお、筆者が出向した2018年は、Transformerモデルが評価型ワークショップでも主流になり始めた年であり、聴講したWAT2018 (Workshop on Asian Translation : 日本語や中国語等のアジア言語を使った機械翻訳の評価型ワークショップ) でも、多くの参加チームがTransformerモデルを採用していた¹¹⁾。ニューラル機械翻訳の研究は今も衰えるところを知らず、プレプリント・サーバ(査読を経ないで論文を公開可能)のarXiv¹²⁾で、neural machine translationと検索すると、126件(2016年)、209件(2017年)、394件(2018年)、503件(2019年)と年々公開される論文数が増えている。

ここで、各機械翻訳方式の特徴について概説する。機械翻訳は、ルールベース機械翻訳と、コーパスベース機械翻訳(コーパス:ここでは、学習データとして用いるため、翻訳関係にある二言語の文書対を集めたもの³⁾)に分類される。ルールベース機械翻訳は、辞書や文法ルールを人手で作成し込む方式であり、複雑で手間がかかるデメリットがあることから、現在は大量のコーパスから辞書も文法ルールも自動的に学習させるコーパスベース機械翻訳が主流であり、その著名な方式として、統計的機械翻訳

とニューラル機械翻訳が存在する。統計的機械翻訳では、学習データとして対訳文データを用いて、原文の語と訳語の対応や、その言語らしい訳語の語順の並べ替えを統計的に学習し、それに基づいて翻訳を行う³⁾。具体的には、訳語の尤もらしさを表す翻訳モデルと、訳文の単語の尤もらしさを表す言語モデルを使って、2つのモデルの尤もらしさの積を最大化するように翻訳を行う。一方、ニューラル機械翻訳では、学習データとして対訳文データを用いて深層学習を行い、エンコーダ・デコーダモデルを使用して翻訳を行う³⁾。

深層学習によるエンコーダ・デコーダモデルは、機械翻訳の文脈において、原文をコンテキストベクトルに符号化するエンコーダ(符号化器)と、コンテキストベクトルから翻訳文を復号するデコーダ(復号化器)からなる。文は形態素解析(形態素:意味を担う最小単位で語よりも小さいもの。例えば、「お花」は単語で、「お」と「花」がそれぞれ形態素。単語分割のことを慣習的に形態素解析と呼ぶこともある³⁾¹³⁾)されて分かち書きされた後、各形態素が、辞書によって、ニューラルネットが扱える単一のIDに変換される。IDは単語埋め込みと呼ばれるベクトルに変換される。単語埋め込みにおいては、意味が近いもの同士がベクトル空間上で近くに配置されるようになっている。RNNベースのsequence to sequenceモデルとTransformerモデルでは、この後入力するエンコーダとデコーダが異なっている。(図2(RNNベースのsequence to sequenceモデル)、図3(Transformerモデル)参照。)RNNベースのsequence to sequenceモデルでは、可変長を取り扱うことができるRNNをエンコーダやデコーダとして使う。その際、エンコーダやデコーダとは別に、アテンション機構も用いて、デコーダが符号化された入力文のどこに注目すべきかをコントロールすることが一般的である。一方、Transformerモデルでは、アテンション機構を内蔵したTransformerエンコーダやTransformerデコーダを用いる。Transformerエンコーダの大きな特徴

9) Jean, S., et al. (2015). Montreal neural machine translation systems for WMT' 15. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 134-140).

10) <https://japan.googleblog.com/2016/11/google.html>

11) <http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2018/index.html>

12) <https://arxiv.org/>

13) 工藤拓 (2018) 『形態素解析の理論と実装 (実践・自然言語処理シリーズ)』近代科学社

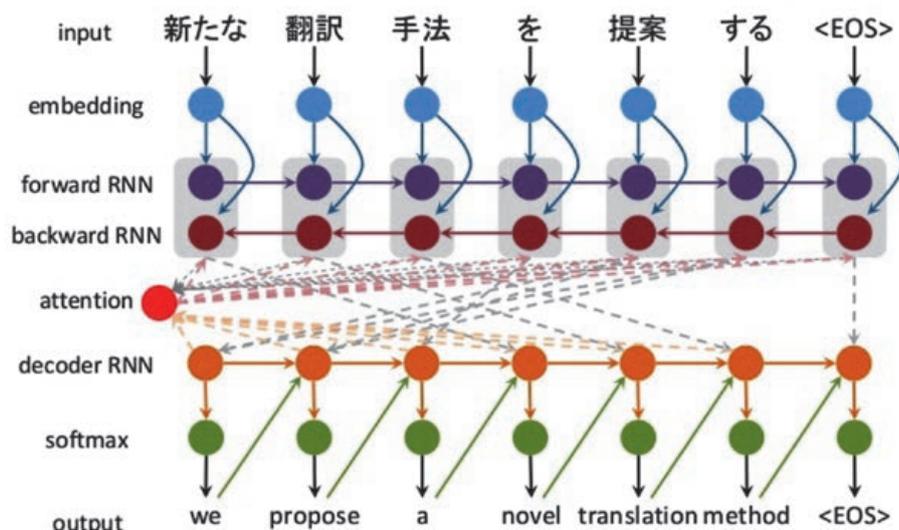


図2 RNNベースのsequence to sequenceモデル

(ゼロから始めるニューラルネットワーク機械翻訳 中澤 敏明 科学技術振興機構 NLP2017 チュートリアルより¹⁵⁾)

としては、位置エンコーディングを用いることである。RNNベースのsequence to sequenceモデルでは形態素をRNNに入力する順番により入力文全体の構造を保持していたが、Transformerモデルでは位置エンコーディングと呼ばれる、前から何番目の要素であるかを示す情報を、単語埋め込みに足し合わせることで、入力文全体の構造を保持している。Transformerモデルは、RNNベースのsequence to sequenceモデルのように再帰構造がないため、並列化が容易であり、学習時間を大幅に短縮することが可能である。なお、Transformerモデルの実装については、Harvard NLPのAnnotated Transformer¹⁴⁾に詳しく、実装も理解しやすい。

次に、統計的機械翻訳と、ニューラル機械翻訳を比較して、各々の特徴について解説する。統計的機械翻訳と比べて、ニューラル機械翻訳では、学習データは大量でクリーンである必要がある。まず、「大量」の意味としては、学習データが少なすぎる場合、統計的機械翻訳もBLEU値（機械翻訳において、翻訳品質を測るのに一般的に使われる指標。BLEU値が高いほど翻訳品質が高い。3. E. に詳述。）が低くなるが、ニューラル機械翻訳の方が、顕著にBLEU値が下がる¹⁶⁾。そして、「クリーン」の意味と

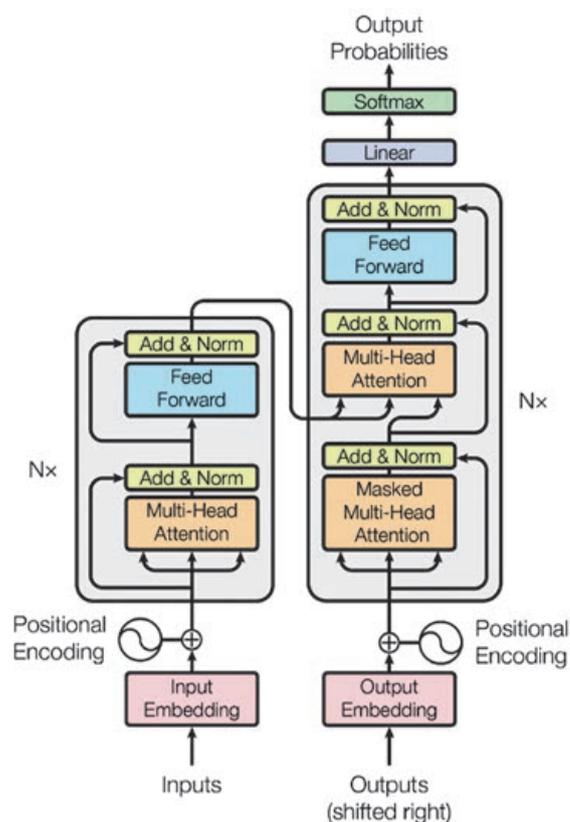


図3 Transformerモデル

(Vaswani, A. et al. (2017) Attention Is All You Needより⁵⁾)

14) <https://nlp.seas.harvard.edu/2018/04/03/attention.html>

15) <https://www.slideshare.net/ToshiakiNakazawa/nlp2017-nmt-tutorial>

16) Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation (pp. 28-39).

しては、統計的機械翻訳では、学習データは質より量、つまり、誤ったデータが多少混入していたとしても、学習データの量が多い方がBLEU値が高くなるが、ニューラル機械翻訳においては、誤ったデータが混入していると、学習データの量を増やしてもBLEU値がかえって低くなることが知られている¹⁷⁾。また、両者の翻訳の仕組みが全く異なることに起因して、エラーの種類も異なる。このことから、ある機械翻訳による和訳が、(フレーズベース)統計的機械翻訳か、ニューラル機械翻訳かは、エラーの含まれ方である程度見分けることができる。統計的機械翻訳では、翻訳モデルで単語(フレーズ)の置き換え候補を列挙し、言語モデルで単語(フレーズ)の置き換え候補を並び替えて尤もらしい翻訳を出力する。その際、並び替えについては、数語程度の狭い範囲の尤もらしさを見ているため、長文になると、例えば、係り受け等の文法がおかしくなることがある。また、単語(フレーズ)の置き換えについても、意味としては正解なのだが、あまり使われないものが選択され、理解することが難しくなる場合がある。一方、ニューラル機械翻訳では、不自然な言い回しはほぼ発生せず、係り受け等の文法的な正しさも格段に向上し、訳文は流暢で読みやすい。その代わりに、大量の文がごっそり抜けたり、特定の文節や、列挙している例の1つについて、訳さずに無視しまったり、特定の語をよく似た別の語に翻訳してしまうことがある。言い回しが自然である分、このようなエラーが起こっても、人間が気づかずに素通りしてしまう危険性があるため、審査業務においてニューラル機械翻訳を利用する際は、特に引用文献として用いる場合は翻訳文だけでなく可能な限り原文を確認する等、十分注意する必要がある。

3. 特許文書和訳におけるニューラル機械翻訳の課題や注目技術について

ニューラル機械翻訳は、様々なジャンルの翻訳に

適用可能であり、対象言語ペア×ジャンルに応じて、代名詞を含んだ文の適切な翻訳や、性差別(gender bias)のない翻訳等、様々な取り組みが存在する。例えば、医師等の社会的地位の高い役職を主語にすると、動詞の性が自動的に男性になってしまうのは性差別的だとして、Google翻訳でも問題視されていたが¹⁸⁾、技術文書の翻訳では、これはあまり起こりえない問題である。ここでは、ニューラル機械翻訳の現状の課題や注目技術のうち、特許文書和訳に関するものをピックアップして紹介する。特許文書和訳と一口に言っても、検索、粗読(先行技術のあたりをつける)、精読(先行技術の内容を正確に把握する)等の先行技術調査等のためにある程度のレベルで翻訳されていれば良いのか、あるいは、出願書類作成のために正確な翻訳が必要なのかによって、要求される翻訳品質は異なるが、いずれの場合においても、主な現状の課題や注目技術と考えられている点について説明する。

A. 不適切な訳語、一文献内での訳語のバラつき

統計的機械翻訳も含めたコーパスベース機械翻訳の問題点ではあるが、訳語が確率的に決まってしまうため、意味的には正解であっても、希望の訳語が選択されない場合がある。このため、一文ごとに翻訳されることで、原文側の文献で同一であった単語が、異なる訳語としてバラバラに翻訳され得るので、文献を精読する際に発明の正確な把握が難しくなるケースがある。このような事象への対応策として、語彙の指定をして翻訳する研究が進んでいる。語彙指定を行う方法については、原文側の単語を『<PLS1>』等の特殊な記号(プレースホルダ)に置き換えて翻訳した後、訳文側に出力されたプレースホルダを指定単語で置き換える方法¹⁹⁾、翻訳時に指定単語を含むようにビームサーチを行うことで訳文側に指定単語を出力させる方法²⁰⁾、外国語が混入された文を翻訳する場合のコードスイッチという枠組みを用いて、原文側の単語を指定語に置き換えた

17) Khayrallah, H., & Koehn, P. (2018) . On the Impact of Various Types of Noise on Neural Machine Translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation (pp. 74-83) .

18) <https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html>

19) Post, M., et al. (2019) . An Exploration of Placeholder in Neural Machine Translation. In Proceedings of Machine Translation Summit XVII Volume 1: Research Track (pp. 182-192) .

20) Hokamp, C., & Qun, L. (2017) . Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) .

り、原文側の単語の横に指定語を並べたりして翻訳する方法²¹⁾等があるが、どの方法にも一長一短がある。筆者自身も業務の一環としてこの問題に取り組み、マルチソーストランスフォーマ(Multi-Source Transformer) という複数の入力を受け付けるTransformerモデルの亜種を利用して、翻訳時に指定語を出力させる方法を提案した²²⁾。語彙指定ができると、一文献の中で訳語がバラバラに翻訳されることがなくなり、易読性が向上する。また、出願書類を作成するために機械翻訳を利用する翻訳者の観点からも、クライアントから提供された用語集が機械翻訳に反映できるようになるため、翻訳作業が容易になると考えられる。

B. 長文

特許請求項は、1つの請求項だけで半ページを超える長文になる場合があるが、ニューラル機械翻訳では、長文を翻訳すると精度が悪化することが知られている。特に、Transformerモデルでは、RNNベースのsequence to sequenceモデルよりも、単語

同士の依存関係を捉えられる距離が短いことが知られており、長文のまま翻訳することで、翻訳の精度が下がってしまう。Transformerモデルの改良型であるTransformer-XLというモデルでは、長距離の依存関係を捉えることができるが、論文においては、エンコーダ部分のみの発表であった²³⁾。その他、マルチホップ注意機構を用いる研究²⁴⁾や、長文自体を適切な長さに分けてから翻訳を行う研究²⁵⁾も行われている。

C. 低頻度語、未知語

特許文書は、その性質上、新しい技術を扱うため、必然的に学習データに含まれていない新たな語を含む可能性が高い。また、特許法施行規則第二十四条(様式第29)によると、明細書には、「平仮名(外来語は片仮名)、常用漢字及びアラビア数字を用いる」とあり、アルファベット等で記載された外国語をそのまま文中に残しておくことができないため、一般的な訳語が確定していない語についても、原語の発音に基づいてカタカナ語で表記

A computer-implemented method, comprising: ▲ generating, at a computer system having one or more processors, a first machine translation of one or more words from **a source language to a target language** using a first machine translation system; generating, at the computer system, a second machine translation of the one or more words from the source language to the target language using a second machine translation system; ▲ and generating, at the computer system, a consensus machine translation of the one or more words from **the source language to the target language** based on a combination of the first and second machine translations when a degree of correlation between the first and second machine translation systems is less than a threshold indicative of a minimum acceptable gain in translation quality achievable by the first and second machine translation systems.

コンピュータ実装方法であって、以下の工程からなる方法。▲1つ以上のプロセッサを有するコンピュータシステムにおいて、第1の機械翻訳システムを使用して、**原語から目的言語**への1つ以上の単語の第1の機械翻訳を生成するステップと、コンピュータシステムにおいて、第2の機械翻訳システムを使用して、原語から目的言語への1つ以上の単語の第2の機械翻訳を生成するステップと、を含む。▲第1および第2の機械翻訳システム間の相関度が、第1および第2の機械翻訳システムによって達成可能な翻訳品質の最小許容ゲインを示す閾値未満である場合に、コンピュータシステムにおいて、第1および第2の機械翻訳の組み合わせに基づいて、**ソース言語からターゲット言語**への1つまたは複数の単語のコンセンサス機械翻訳を生成すること。

図4 無料のウェブサービスで米国特許(US8855995B1)の請求項を翻訳した様子(赤字と▲は筆者による。)

本文中のsource languageとtarget languageがそれぞれ異なる訳語に翻訳されている。(A. 不適切な訳語、一文献内での訳語のバラつき) また、▲は長文翻訳を避けるためにシステムが前処理として文分割したと思われる点を示す。(B. 長文)

- 21) Dinu, G., et al. (2019). Training Neural Machine Translation to Apply Terminology Constraints. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
- 22) 石川雄太郎, 江原暉将 (2020)「マルチソーストランスフォーマと専門用語辞書を用いた訳語の制御方法」言語処理学会 第26回年次大会
- 23) Dai, Z., et al. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2978-2988).
- 24) Iida, S., et al. (2019). Attention over heads: A multi-hop attention for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 217-222).
- 25) Kuang, S. & Xiong, D. (2016). Automatic long sentence segmentation for neural machine translation. In Natural Language Understanding and Intelligent Applications (pp. 162-174). Springer, Cham.

(トランスリタレーション)する等して対処しなければならない。このため、同じ原語であっても、翻訳者によって、別々のカタカナ語として翻訳されてしまい、アルファベット等で記載された原語自体は低頻度語(学習データに出現する頻度が低い語)や未知語(学習データにない語)でなくても、翻訳されたカタカナ語は翻訳の揺れにより低頻度語や未知語になってしまうケースも出てくる。これらの事実から、特許文書は、低頻度語や未知語が含まれやすい性質を持つといえる。

一方、ニューラル機械翻訳においては、語をIDに変換する辞書のエントリー数に制限があるため、低頻度語がうまく翻訳できない問題が存在し、低頻度語に関しては何らかの工夫をして翻訳する必要がある。また、そもそも未知語については、語をIDに変換する辞書に存在しないため、一律未知語を示すIDに変換されてしまい、対応する部分の訳語も出力されない。これらの現象への対応策として、SentencePieceやBPEを用いたサブワード化という手法が挙げられる²⁶⁾。サブワード化を用いると、低頻度語や未知語は、最短1文字単位にまで細かく分解され(例えば、「スマホ」であれば、「ス マホ」、「ス マホ」、「ス マホ」と分解される可能性がある。)、このように分解された語として辞書に出現するため、未知語を示すIDに変換されてしまう可能性は少ない。しかしながら、後述する訳抜けというエラーと組み合わせられてしまうと、カタカナ語の途中一文字だけが抜けて翻訳される等の問題が生じてしまう。また、英語に関しては、綴りと発音が必ずしも規則的でないため、何を指しているか分からないカタカナに翻訳される可能性もある。

D. 訳抜け、湧き出し

原文に存在する内容が訳文から欠落するエラーを訳抜けといい、逆に原文に存在しない内容が訳文に出現するエラーを湧き出しという。ニューラル機械

翻訳に特徴的な誤りとして、一見流暢な文章ではあるが、湧き出しや大規模な訳抜けをしていることがあり、これらは、出願書類作成の際は勿論のこと、検索時の漏れや、粗読、精読時の誤解等にもつながるため、特に重大なエラーである。ニューラル機械翻訳モデルにおける訳抜け検出としては、アテンションの累積確率と逆翻訳の確率を組み合わせる手法が研究されている²⁷⁾。

E. 評価手法

機械翻訳の評価は、自動評価と人手評価に分けられる。まず、自動評価としては、BLEU値が広く一般的に使われており、他に、RIBES値等の有名な指標も存在する。BLEU値は、翻訳システムの翻訳文中の単語列が参照訳に含まれる割合を1-gram、2-gram、3-gram、4-gram (n-gram : 分かち書きを終えた後のトークン(後述)がn個連続すること。)についてそれぞれ計算して相乗平均を取り、翻訳文が参照訳に比べて短い場合のペナルティを乗じることで、文としての近さを示す指標である³⁾。一方で、BLEU値は、表面的に連続する語のまとまりを見ているため、言い換え文には不当に低いスコアを与えるデメリットがある²⁸⁾。BLEU値のデメリットを克服する新しい指標として、2019年、BERT(4.に詳述)を使って、機械翻訳と正訳の類似度を測ることで、BLEU値より人手評価と相関関係が高い評価ができることが示された²⁹⁾。さらに2020年、GoogleからBLEURT²⁸⁾という、同じくBERTを使って機械翻訳を評価するシステムが、OSSとして提供された。しかし、これらのシステムは、学習方法や学習データによってスコアが変わってしまう点が懸念される。

次に、人手評価については、例えば、特許庁が機械翻訳に関連する事業の調達や管理を実施するにあたって、機械翻訳結果の品質を適切に評価するために、特許文書機械翻訳の品質評価手法に関する調査

26) Sennrich, R., et al. (2016) . Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) (pp. 1715-1725) .

27) Goto, I., & Tanaka, H. (2017, August) . Detecting untranslated content for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation (pp. 47-55) .

28) Sellam, T., et al. (2020) . BLEURT: Learning Robust Metrics for Text Generation. arXiv preprint arXiv:2004.04696.

29) Shimanaka, H., et al. (2019) . Machine translation evaluation with BERT regressor. arXiv preprint arXiv:1907.12679.

を行っており、成果物として、「特許文献機械翻訳の品質評価手順 Ver.1.0」を公開している³⁰⁾。この中で定められた「内容の伝達レベルの評価」とは、機械翻訳結果が、原文の内容をどの程度正確に伝達しているかを5段階評価にて主観的に評価するものであるが、この評価手法はWAT (Workshop on Asian Translation: 日本語や中国語等のアジア言語を使った機械翻訳の評価型ワークショップ) においても、別の人手評価手法と共に使われている。人手評価は手間がかかることから、評価型ワークショップや日常業務においては、自動評価も使われているものの、その正確性は人手評価に及ばないと考えられている。

F. モデル圧縮技術

特許文書に限らないが、基盤技術として注目されているのは、モデル圧縮技術(蒸留、枝刈り、量子化技術)である。モデル圧縮技術を用いると、モデルを小さくすることができるだけでなく、翻訳速度が上がることで、GPU等の特殊なハードウェアがなくても、一般的なCPUを使って実用的な時間で推論を行うことができる。クラウド化の流れが進んではいるものの、機密性の高い書類を扱うPCをインターネットに接続できない場合等、機械翻訳のみならず、深層学習を使った様々なタスクをCPUのみで可能とする技術は重宝されると思われる。

4. 深層学習を適用した機械読解～ニューラル機械読解～について

これまで、機械翻訳に関する話題について解説したが、この章では、機械読解について説明する。ところで、ウェブページ内で特定の用語を検索するために、キーワードのハイライト機能を用いた経験がある方は多いのではないだろうか。また、特許審査官であれば、先行技術調査の際、スクリーニング画面で本願関連用語等を入力し、それぞれの用語に対応する色を指定すると、先行技術文献内に同じ用語

があれば指定した色でハイライトされる機能を用いたことがないという方はいないだろう。このように、通常は入力したキーワードそのものがハイライトされるが、TensorFlow Blogで発表されたウェブページ内検索機能³¹⁾を用いると、例えば、ウェブページで、「天体観測に最も良い日は？」等の質問を入力し、答えに該当する「新月と次の満月の約3日前の間の夜」等の部分がハイライトされる。これは、機械読解を用いたQ&Aタスク(質問に対する答えを示す箇所を抽出する)の精度が向上したために、新たに実用化された技術であり、今後精度がさらに向上していくと、検索の在り方を大きく変える可能性があるものになると考えられる。この背後にある技術がBERTである。

2018年10月に、GoogleからTransformerモデルのエンコーダ(以下、Transformerエンコーダ)を利用したBERTが発表され、2020年7月現在、上述のQ&Aタスクを始めとして、自然言語処理の様々な分野で驚異的な性能を発揮している⁷⁾。BERTの大きな特徴の1つは、事前学習モデルであるということである。深層学習では、学習データは大量でクリーンである必要があるが、Q&Aタスク等の特定のタスクに特化した大量のクリーンな学習データを集めることは、コスト等の問題からかなりハードルが高い。一方、BERTにおいては、事前学習モデルと呼ばれる、大規模で個別タスクに依存しないデータを使った学習を行ったモデルを用意し、その後、Fine-Tuningと呼ばれるタスクに特化した学習を行う。この際、大規模で個別タスクに依存しないデータは、手に入りやすいテキストデータ(Wikipedia等)で済み、一方、Fine-Tuningに用いられる、Q&Aタスク等の特定のタスクに特化したデータは、一から学習させる場合に比べて、少量を用意すれば事足りるというメリットがある。このような有用な特徴により、BERTは学習データを大量に用意できない様々なタスクに適用可能であり、現在の主な用途として、Q&Aタスク、2文の同値判定、構文解析³²⁾、検索時の

30) https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/tokkyohonyaku_hyouka/01.pdf

31) <https://blog.tensorflow.org/2020/03/exploring-helpful-uses-for-bert-in-your-browser-tensorflow-js.html>

32) 柴田知秀, 河原大輔, 黒橋禎夫 (2019) 『BERTによる日本語構文解析の精度向上』言語処理学会 第25回年次大会

Re-ranking³³⁾、要約時の重要文抽出³⁴⁾、上述の従来人手で行っていた機械翻訳の自動評価等が挙げられる。

また、Fine-Tuning後のそれぞれの個別のタスク(downstream task)において、既存手法に比して、性能が高くなるのも大きな特徴である。論文発表時には、1つの事前学習モデルを、代表的な8つのデータセットにFine-Tuningし、その全てで世界最高性能(SoTA: State-of-the-art)を記録して、注目を浴びた。特に、SQuADという文書読解データセットでは、BERTを使うことで、人間の文書読解の成績を上回る性能を発揮するという驚くべき成果を挙げた。一方、加算等の計算問題や数字の比較(DROP dataset)は他の事前学習モデルに比べて苦手だという研究もある³⁵⁾。

次に、BERTの基本的な仕組みについて解説する。BERTは、上述のTransformerエンコーダを、所定の数、スタック(同じ層を縦に繰り返し連結する)している。その他、単語埋め込み、位置埋め込みの他に、セグメント埋め込み(2文のうち、前の文はA、後の文はBとして異なる文であることを明示するもの)を用いている。BERTの特徴的な部分は、事前学習に関する部分であり、学習データのうち15%のトークンを、[MASK]という記号に置き換えたり、ランダムに別のトークンに置き換えたりしたものを入力文とし、学習データに元々あった正しい文に復元して出力するタスクを学習する。また、2文が隣接関係にあるかどうかを判定するタスクの学習も行う。これにより、BERTは言語そのものに関する知識を大規模なコーパスから獲得できるとされている。またFine-Tuningにおいては、最後の1層

のTransformerエンコーダを追加して、タスクに特化した学習データで追加的に学習を行う。これにより、他の層に蓄えられた言語そのものに関する知識を利用して、高い性能を獲得することができる。

BERTは近年益々研究が盛んになっており、2019年言語処理学会全国大会では、予稿集の中でBERTに言及している発表は21本だったが、2020年言語処理学会全国大会では118本まで急増しており、全発表の約4分の1が何らかの形でBERTに言及するほど自然言語処理分野に大きなインパクトを与えている。また、BERTの弱点を補うような様々な後継システムも現れてきており^{36) 37) 38)}、上述のQ&Aタスクの正答率を始めとする性能の向上は未だ留まるところを知らず、日々進化し続けている。なお、BERTに関する解説は、特技懇誌のJapio前原義明課長代理の寄稿に詳しい³⁹⁾。

日本語におけるBERTの事前学習モデルについては、2020年7月現在では、京都大学、東北大学、NICT(情報通信研究機構)等から提供されているが、最初は無償で提供された事前学習モデルは、Ubic, Incの菊田遥平氏のものである⁴⁰⁾。筆者は、この菊田氏の前学習モデルの公開後に、BERTに関する複数の実験を経て、BERTを用いることで、一の請求項と引用文献の対比を行わせることができるのではないかと考えた。以下、筆者の研究について少し説明したい。

特許審査実務においては、請求項に係る発明と引用発明との対比を行い、一致点及び相違点を認定してなされるが、その際、請求項に係る発明の下位概念と引用発明とを対比することも可能である。すなわち対比とは、請求項の記述が、引用文献の記述と

33) Nogueira, R., et al. (2019) . PASSAGE RE-RANKING WITH BERT. arXiv preprint arXiv:1901.04085.

34) Liu, Y. (2019) . Fine-tune BERT for extractive summarization. arXiv preprint arXiv:1903.10318.

35) Wallace, E., et al. (2019) . Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5310-5318) .

36) Yang, Z., et al. (2019) . Xlnet : Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5753-5763) .

37) Lan, Z., et al. (2019) . ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.

38) Clark, K., et al. (2019) . ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In International Conference on Learning Representations.

39) 前原義明 (2020) 「寄稿5 トランスフォーマーを用いた特許審査支援の探求 -Detailed Description Is All We Need-」, 『特技懇』 296号

40) <https://github.com/yoheikuta/bert-japanese>

比較して上位概念化されていてもよいという条件の下で、1つの文である請求項と、1つの文書である引用文献の間に、意味的にどの程度一致する部分が存在するか判定することであると捉えることもできるのではないかと思います。

統計的機械翻訳もそうであるが、従来の統計方式のモデルにおいては、入力から出力まで、いわゆるEnd-to-Endで学習するわけではなく、条件付き確率に基づいて、サブタスクの種類を考え、サブタスクに特化した機構を用意し、人間が特徴量を設計する。特許審査実務における対比のような特殊なタスクの場合は、必要な各サブタスクを全て列挙すること自体、困難性が高く、さらにそれらに即した特徴量設計とデータ収集も、大変ハードルが高かった。一方で、深層学習を用いたモデルは、各サブタスクの列挙や特徴量設計、各サブタスクに特化した学習データの準備を必要とせず、入力から出力まで、いわゆるEnd-to-Endで学習して、精度の高い言語理解モデルを構築することが可能であるため、入力と出力の学習データさえ揃っていれば、特許審査実務における対比のような複雑なタスクも解くことができる。また、このような専門的タスクに適用する学習データは、クリーンなものを大量に集めることが難しいため、Fine-Tuningに比較的少量のデータしか必要がないBERTは大変マッチしていた。

実験では、入力としては請求項と引用文献をタグで繋げたものを用い、出力としては請求項内のトークン（分かち書きした各々の部分：例えば、「犬が好き。」を形態素解析すると、「犬 が 好き。」と分かち書きされ、「犬」、「が」等がトークンとなる。）が、トークンごとに一致点に含まれるかを判定することにした。このとき、一般に公開されているBERTの実装（GitHubにアップロードされているTensorFlowでの実装）には、文間の意味の類似性の判定（STS）のように、2つの文を繋げて入力し、全体として多値判定（例えば、「類似」、「非類似」）を行うパターンと、固有表現抽出（CoNLL）のように、1つの文内のトークンごとに多値判定（例えば、「人名」、「場所」）を行うパターンしかなかったため、プログラムを書き換えて、本願請求項と引用文献の記載の2つの文を繋げて入力し、本願請求項について、トークンごとに2値判定（例えば、「一致点」、「相違点」）を行って、一致点に含まれるか否かを判定するように改造した。

肝心の学習データとしては、審決公報を用いて、特許審決文における一致点の認定部分を利用することにした。審決文における一致点の認定部分は、審判官が、本願請求項の文言のうち、引用文献に記載されている部分であると認定した部分であるから、本願請求項と一致点の最長一致を取ることで、引用文献に書かれている部分、書かれていない部分が明確になると考えた。しかし、この手法で難しかったのは、BERTには入力トークン数に制限があるため、極めて短いデータしか学習に利用できない点であった。BERTは最長で512トークンしか入力できないが、本願請求項と引用文献を繋げて入力すると、512トークンを簡単に超えてしまう。したがって、学習時には、本願請求項と引用文献を繋げて入力しても512トークンに収まるデータという極めて限られたデータしか学習に使えなかった。BERTに代えて、長い入力にも対応しているXLNet³⁶⁾等のモデルを使うとうまくいくのかもしれない。

特許審決文における一致点の認定部分は、ルールベースで抽出した。多くの審決文を読み、審決文の特徴から、一致点を抜き出すことができるスクリプトを、正規表現を試行錯誤して作成したが、審決文ごとの記載が想像以上にバラついており、正確に抽出しようとする、多くの審決文では一致点を抜き出すことができなかった。なお、特許審決文から抽出した一致点の認定部分と引用文献の摘記箇所を使って、Fine-Tuningしたモデルをテストデータで試すと、人手で認定した一致点の認定と比較して、ほぼ同レベルの対比が可能ケースもあった。このように、少量のデータでもうまくいく例を確認することができたが、やはりFine-Tuningであっても、学習データの量は重要である。学習データ量と方式自体の問題を切り分けて研究を進めるべく、今後機会があれば、十分な学習データ量を確保した上で、本方式でできること、できないことを詳しく調べていきたい。

なお、大量の論文の査読や関連研究の引用のために、学界にも関連文献を引用するニーズが存在しており、実際に研究も進んでいるところ、BERTを始めとする機械読解技術は、このニーズに確実にマッチしている。研究が進んだ結果、少しひねりを加えることで新規性・進歩性を否定する文献を的確に探せる外部システムが生み出される可能性も見据

え、特許庁内でも新技術を幅広く継続的に注視し続け、特許情報処理にどのように応用できるか、自由闊達な議論を続けていく必要があるのではないかと考えている。

5. AI研究をする際の環境について～オープンイノベーションを実感する～

以上、深層学習を用いた機械翻訳と機械読解の概況について説明した。今日、最先端の論文が査読前に公開され、最新のOSSを無償で試してみることができるので、GAFa等以外の研究開発者であっても、最先端の成果に基づいて研究開発ができるだけでなく、個人的にAI研究を行ってみたいと思っている方にも、これらの成果が利用可能である。この章では、特に個人でAI研究を進める際に利用可能な研究環境等の情報について、筆者自身の経験も交えながら紹介したい。ご興味のある方は以下参考にしていただければ幸いである。

A. 論文等の情報収集について

オープン化された情報としては、arXiv⁴²⁾ というコーネル大学図書館が運営する論文のプレプリント・サーバが有名である。arXivでは、査読前の論文が無料で入手できるが、後に有名カンファレンスで発表される論文も多く掲載されており、旬の情報を素早く入手できる点で大変有用である。小技ではあるが、arXivを読む際は、arXiv Vanity⁴¹⁾ というサイトを使うと、掲載されたPDFがHTML化されるので、これをGoogle翻訳でサイトごと翻訳してから読むと日本語で読めて便利である。有名カンファレンスについては、機械翻訳を含む自然言語処理分野では、ACL、EMNLP、NAACL、CoNLL等があり、機械学習や人工知能分野では、NeurIPS、ICML、ICLR、AAAI、IJCAI等がある。その他、MTSummit⁴²⁾のように、機械翻訳の研究開発者だけ

でなく、翻訳者やクライアント等のユーザを含めた関係者が一堂に会し、機械翻訳の研究開発に関する発表や、翻訳業務で実際に活用する方法等の発表が行われるユニークなカンファレンスもあり、筆者も参加した際、ユーザ側からの目新しい課題を知ることがもできた。なお、国内学会としては、人工知能学会、言語処理学会、情報処理学会の自然言語処理研究会(NL)等がある。

論文以外の情報収集の手段としては、機械翻訳分野の著名な先生がTwitterに投稿している場合があるので、Tweetを追うことで、どのような技術が話題になっているか分かり、技術の流れをフォローするのに大変効果的だった。

B. OSS (Open Source Software) について

ソフトウェア開発のプラットフォームであるGitHub⁴³⁾には、OSSとして多くのプログラムがアップロードされている。これらは、無料でダウンロードして実行することができ、商用利用可能なライセンスで提供されているOSSも多い。有名カンファレンスに投稿される多くの論文は、GitHubに実際のプログラムコードがアップロードされていることも多く、先端の研究について手元の環境で実験ができることも特徴である。

企業等での製品開発においては、単体のプログラムコードというよりも、幾つかの機能をまとめて提供しているフレームワークを使っているケースもあるとのことだった。深層学習のフレームワークとしては、TensorFlowやPyTorchが有名だが、機械翻訳のフレームワークとしては、OpenNMT⁴⁴⁾、Marian NMT⁴⁵⁾、Sockeye⁴⁶⁾等が有名である。これらのフレームワークは、基本的にボランティアベースで開発されているため、どんな機能が導入されているかは、開発者の興味による部分が多い。そのため、複数のフレームワークにどのような機能が搭載されているのかを網羅的に調査し、各々のメリッ

41) <https://www.arxiv-vanity.com/>

42) <https://www.mtsummit2019.com/>

43) <https://github.com/>

44) <https://opennmt.net/>

45) <https://marian-nmt.github.io/>

46) <https://awslabs.github.io/sockeye/>

トとデメリットを比較して、どのフレームワークを導入するか検討することも、研究所における重要な作業だった。

また、GitHubには、深層学習に関するものだけでなく、その他のOSSも充実している。例えば、JavaScriptのOSSのライブラリも数多く存在し、これらを用いると、検索結果のグラフによる可視化や、テーブル作成等を簡単に行うことができる。そのため、筆者は、製品に必要な機能がOSSで実現可能かについての情報収集と、実装方法の検討も行った。

OSSで作業していると、理解不能なエラーが発生することも多々ある。この場合は、GitHubのissuesを見るか、エラーメッセージをGoogle検索して、プログラマのための技術情報共有サービスであるQiita⁴⁷⁾等に辿り着いて対処した。マニアックなエラーについては、英語のサイトを参照するのは勿論だが、英語のサイトに書かれていないことが、CSDN (Chinese Software Developer Network)⁴⁸⁾に掲載されていることが多々あり、中国のAI分野での研究開発者の裾野の広さを垣間見た気がして印象的であった。

OSSは大変有用であり、まず手を動かしてみる際のハードルを低くできることは確かである。しかしながら、筆者が実験を進める中で、大量のデータを処理して実験を行う際には、特に、処理速度が重要となってくるため、OSSに頼るだけでなく、アルゴリズム等の情報学的知識も必須であることは、以前と変わらないと感じた。例えば、筆者が前述の論文「マルチソーストランスフォーマと専門用語辞書を用いた訳語の制御方法」の執筆に向けて実験を行っていた際、最初期は、アルゴリズムを意識せず実装してしまい、実験を終えるのに数か月単位かかる試算になってしまったが、Trie木とマルチプロセスを用いてアルゴリズムを改良したおかげで、実行時間を数百分の一に抑えることができた。他にも簡単な工夫として、重複排除の際は、処理の最後に一気にするのではなく、データを追加するたびに小まめにしたり、一度読み込んだデータはpickleという

高速で読み込み可能な形式で保存し直したり、マルチプロセスで誤ったデータが作られた際、特定のプロセスだけをもう一度やり直す形に整えたりと、OSSを何も考えずにコピーするのではなく、主要な部分はOSSを使いながらも、時間のかかる前処理等を情報学的知識によってさらに改善することで全体の処理速度を向上させることができたケースが多々あった。

C. 計算機環境について

深層学習では、基本的にGPUや、深層学習を高速化すべくGoogleが開発したTPUを搭載した計算機によって学習時間を大幅に短縮することが可能になるため、Japioの研究所では、オンプレミス(自社運用)のGPUマシンを導入していた。GPUとは、Graphics Processing Unitの略で、その名の通り、リアルタイム画像処理に特化した演算装置であるが、深層学習に用いることもできる。代表的なGPUとしては、NVIDIA社のTesla V100が挙げられるが、1枚当たり100万円以上する。そして、機械翻訳では、4枚または8枚をまとめて使用して効率化を図ることも多い。さらに、GPUを搭載するサーバ自体の価格に加え、空調やセキュリティ等を含むサーバ管理費を考えると、とても個人では手が出せないのが実情である。一方、Amazon Web Services (AWS)のGPUインスタンスを用いると、Tesla V100等のGPUを搭載したクラウドにある仮想サーバ(インスタンス)をすぐに使うことができ、初期費用や管理費がかからないことがメリットである。しかしながら、Tesla V100が搭載されたインスタンスの料金は、4.194ドル/時(p3.2xlarge)⁴⁹⁾であり、学習に数日かかることを考えると、結局は、数千~数万円単位の出費が必要になり、個人にとっては大きな金額となる。時間当たりの単価が最も安いのは、産業総合研究所のAI橋渡しクラウド(ABCI: AI Bridging Cloud Infrastructure)⁵⁰⁾だが、こちらは基本的には日本国内の法人を対象にしており、個人の資格で契約することはできない。個人的に深層

47) <https://qiita.com/>

48) <https://www.csdn.net/>

49) <https://aws.amazon.com/jp/ec2/pricing/on-demand/>

50) <https://abci.ai/ja/>

学習を試す際は、Google Colaboratoryの無料GPU環境⁵¹⁾が代表的だが、時間制限が存在する点が難点ではある。

D. データセットについて

Japioの研究所では、特許情報に関するデータとしては、日本国内のものでは、①公開公報⁵²⁾、②審決公報⁵²⁾、③案件の情報が登録されている特許情報標準データ⁵³⁾等を取り扱った。出向するまで特許庁から公開されているデータのことはよく知らなかったが、特許に関する情報を使って研究しようとすると、これらの公報等を避けて通ることはできず、タグやデータ形式等が想像以上に専門的でそれらに関する深い知識が必要であった。例えば、公開公報に関しては、古い公報はXMLではなくSGMLになっている等、発行時期によってデータ形式自体が変わってしまうため、特に数十年単位で公開公報を処理する際には、これらのデータに関する正しい知識をもって適切に処理することが必要であることを知った。また、データ形式がXMLである期間内であっても、単にXMLタグに従って一括で処理することができないケースもあり、このようなデータをどのように扱うかについては、Japio研究所での試行錯誤の末のノウハウが詰まっていることを実感した。また、世界各国の公開公報がWIPOの標準を使っているため出現するタグの種類等が一緒であったのも興味深かった。なお、海外の特許情報データについては、Stephen Adams氏の Information sources in patent⁵⁴⁾に詳しい。

その他、自然言語処理や画像処理等の多くの分野で、学習データ、開発データ、テストデータからなるデータセットが無料で公開されており、データセットを学習させたモデルの性能が競われている。データセットには、Leaderboard (投稿したモデルの性能を公表する順位表) も公開されている場合が

あり、有名データセットで成績が1位になると、対外的なアピールになるため、大学や企業等の研究者が盛んにチャレンジしている。機械翻訳については、Leaderboardというより、WMTやWAT等の評価型ワークショップで性能が競われている印象があるが、評価型ワークショップで使われるデータセット自体は無料で公開されており、Transformerモデルの発表論文“Attention is all you need”⁵⁵⁾においても、WMTのデータセットでBLEU値の評価を行っていた。なお、WMT2020までは、日本語のタスクがなかったため、必然的に日本語を使った研究成果の発表がしにくかった可能性もある。特許文書の機械翻訳については、WATの特許翻訳タスクで使うデータセットを契約することによって、無償で利用できるが、その他に、NTCIR-10 PatentMT (特許機械翻訳テストコレクション)⁵⁵⁾ や、WIPOのCOPPA V2.0 (Corpus Of Parallel Patent Applications)⁵⁶⁾ も、研究目的であれば無償で用いることができる。

これらの公開されている無償のデータセットについては、①少なくとも英語で説明文が書かれており、中身についても英語で書かれたデータが極端に多いこと、②機械的な抽出や、不特定多数のワーカーに少額で簡単なタスクを解いてもらうクラウドソーシングを利用して作られているものが多く、例えば、特許のような特定分野の専門家が人手で作ったようなデータをあまり見かけないことも大きな特徴である。一方、人文学オープンデータ共同利用センターのくずし字データセットのように、専門的ではあるが、データセットを公開しているものも一部存在している。元々同センターの研究者が、くずし字データセットを用いて、KuroNet⁵⁷⁾というくずし字認識システムを開発していたものの、Kaggleという世界中のデータサイエンティスト等が開発したモデルの性能を競い合うプラットフォームでコンペ

51) <https://colab.research.google.com/notebooks/welcome.ipynb?hl=ja>

52) <https://www.publication.jpo.go.jp/>

53) <https://www.jpo.go.jp/system/laws/sesaku/data/keikajoho/index.html>

54) Stephen, A. (2020). Information Sources in Patents (Guides to Information Sources) K G Saur Verlag GmbH & Co

55) <http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-ja-PatentMT.html>

56) <https://www.wipo.int/export/sites/www/patentscope/en/data/pdf/wipo-coppa-technicalDocumentation.pdf>

57) Lamb, A., et al. (2020). KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition. SN Computer Science, 1 (3), 1-15.

ティションにしたところ、KuroNetを超える精度のシステムが開発されると共に、公開された「ディスカッション」で実際の解法を含む重要な知見が得られたとのことであった⁵⁸⁾。このように、解くべき問題さえ明確であるならば、具体的な処理やモデルに関する検討を世界中のデータサイエンティスト等に委ねる環境も揃っているため、特許情報処理に関しても、必要なデータセットを作って、コンペティションを積極的に利用できること、AI化が一層進んでいく可能性がある。

6. その他、将来発展が予想される技術

A. AIを用いた特許検索

Japioでは、毎年、特許情報および知的財産関連の我が国最大の専門見本市である特許・情報フェア&コンファレンスを主催している。昨年度は、AIを用いた特許検索システムについてのブースが一昨年よりも増加していたことが印象的であった。AIを用いた特許検索システムの中で、筆者が特に注目したのは、Amplified AI⁵⁹⁾、Deskbee（検索結果のノイズ落としに用いる）⁶⁰⁾、AI Samurai⁶¹⁾、PatentSQUAREのAI検索⁶²⁾、XLPAT⁶³⁾等である。これらのうち、筆者が実際に試用した経験があるAmplified AIとXLPATについて紹介する。Amplified AIは、AIを使った特許検索システムであり、公開日と、特許番号または発明を要約した文を入力することで、最も関連性が高いと思われる特許を最初に20件表示し、表示された結果に対して、関連特許を指定したり、関連特許を削除したりすることで、より適切な関連特許を表示していくシステムである。また、特許番号の入力に加え、キーワードでの検索を重ねることもできる。XLPATは、インド発のAIを使用する特許調査・解析システムであり、先行技術調査については、NOVELTY CHECKERという機能を用いて、発明

を要約した文を入力し、単語に重みづけを行い、類義語を選ぶことで、最も関連性が高いと思われる特許を表示する。その他、特許番号を入力することで、INVALIDATORという機能で無効資料を出力することもできる。特筆すべき特徴としては、非特許文書が充実している点であり、驚くべきことにYouTubeの動画等も検索対象となっている。

このように市場には、AIを用いた商用の特許検索システムが出てきているが、以前から指摘されているように、AIを用いた特許検索システムの優劣を公平で的確に比較できるような透明性の高い指標が存在していないという問題がある。指標としては、例えば、NTCIRのPATENTタスクで用いた適合率等もあるが、検索式を変えることを前提にしていないため、検索結果を対話型インタフェースでインタラクティブかつ容易に更新可能なことを売りにしているタイプのシステムにはそぐわない。そのため、目的の文献を見つけるまでにかかる総検索時間を何かの形で表せる指標が作られると、システム本来の検索の実力が反映される指標になり得ると考えられる。

B. その他、特許審査に 응용が可能であると見込まれる技術

深層学習を用いたOCR技術が進んでおり、Google Patentでは、USPTOの古い公報について、深層学習を用いたOCRで解析し、全文検索を可能にしている⁶⁴⁾。特許審査においては、分野によっては技術が簡単に陳腐化せず、古い公報も十分引用文献になり得ることから、OCR化を進めることは、検索をより容易かつ網羅的に行う上で有意義であると思われる。また、深層学習を用いた誤記や文法チェッカーの研究開発も進んでおり、起案の品質を向上する上で非常に有用な技術となり得ると考えられる。

58) <http://codh.rois.ac.jp/kuzushiji-challenge/>

59) <https://www.amplified.ai/ja/home>

60) <http://www.ipfine.com/deskbee/>

61) <https://aisamurai.co.jp/service-ais/>

62) <https://www.panasonic.com/jp/business/its/patentsquare.html>

63) <https://www.raytec.co.jp/system-tools/xlpat>

64) <https://www.google.com/googlebooks/uspto-patents-grants-ocr.html>

7. まとめ

Japioへの出向期間は2年間と、研究開発のスパンからすると決して長いものではなかったが、その間に、GoogleのTransformerモデルが機械翻訳の主流モデルになり、またTransformerエンコーダを用いたBERTが機械読解に大きな影響を与える等、深層学習によって自然言語処理が大きな変革期を迎えている様子を見ることができて、大変幸運であった。

ニューラル機械翻訳は様々な課題がありつつも、研究の勢いは衰えるところを知らず、今後も益々高性能化が進んでいくことが期待される。また、機械読解分野でも、BERTとその後継システムが着々と成果を上げており、難易度の高いQ&Aデータセットでも高い成績を出せるようになってきている。

その背後では、最前線の企業や研究機関が、論文、OSS (Open Source Software)、データセットを世界中で共有して、データセットの公開や、コンペティション・評価型ワークショップの開催によって、解くべき1つのゴールを規定し、世界中の研究開発者の知恵を結集しながら、問題解決に向かって邁進する体制となっていることが垣間見えた。特許情報処理においても、AI化していきたい部分について、データセットの公開や、コンペティション・評価型ワークショップの開催によって、よりAI化を進めていくことが可能であると考えられる。

その他、将来発展が予想されるAIを用いた特許検索システムについても紹介した。今後、AIを用いた検索システムの優劣を公平かつ的確に比較でき

るような透明性の高い指標が設けられることで、システムをより身近に使えるようになると思われる。

本稿によって、特許情報処理の分野におけるAI活用について、特技懇会員の皆様のお役に立つ情報を提供できたならば幸いである。

末筆ながら、特技懇会員各位の今後のご活躍をお祈り申し上げますと共に、編集長はじめ、編集員の皆様に寄稿の機会をいただき、厚く御礼申し上げます。

謝辞

出向期間中は、(一財)日本特許情報機構の皆様、特に、小林明所長をはじめとする特許情報研究所の皆様にご大変お世話になりました。厚く御礼を申し上げます、感謝する次第です。

Profile

石川 雄太郎 (いしかわ ゆうたろう)

2011年4月	特許庁入庁 (特許審査第四部伝送システム)
2014年4月	審査官昇任
2016年10月	審査第一部 調整課 審査システム企画班 審査システム企画第二係長
2017年10月	審査第四部伝送システム
2018年7月	(一財)日本特許情報機構 特許情報研究所 調査研究部 研究企画課 課長代理
2020年7月	審査第四部情報処理