ニューラル機械翻訳による特許機械翻訳システムの開発

Development of Neural Machine Translation System for Patent Translation.

東芝デジタルソリューションズ株式会社
園尾　聡

2009 年九州工業大学生命体工学研究科博士課程修了。博士（工学）。同年株式会社東芝入社（2017 年より現所属）。機械翻訳、自然言語処理の研究開発に従事。

satoshi.sonoo@toshiba.co.jp

1 はじめに

知財戦略のグローバル化に伴い、特許文献を翻訳する頻度・件数は急増している。一方で、人手による翻訳作業では、時間がかかる、コストが高い、大量の翻訳が難しいなどの課題があり、これらの課題を解決する手段として、機械翻訳の活用が進んでいる。

機械翻訳技術は、従来のルールベース機械翻訳（Rule-Based Machine Translation; RBMT）、統計的機械翻訳（Statistical Machine Translation; SMT）を経て、近年、ニューラルネットワークによるニューラル機械翻訳（Neural Machine Translation; NMT）が主流となってきた。NMT のモデルとして、注意機構（アテンション）を用いた Encoder-Decoder モデル [1] や Transformer モデル [2] が提案されており、どのモデルも大規模な対訳データからモデルを学習し、SMT を大きく上回る正確さと流暢さが実現可能であることが報告されている。

このような背景の中、特許庁では、特許情報プラットフォーム（J-PlatPat; JPP）等で利用される機械翻訳システムを刷新し、NMT をベースとした機械翻訳サービスの提供を 2019 年 5 月 [1] に開始した [3]。この新しい機械翻訳システムでは、国立研究開発法人情報通信研究機構（NICT）が開発した NMT を大規模クラウドプラットフォーム上に構築し、NMT 特有の課題に対して様々な前後処理を適用することで高精度な機械翻訳を実現している [4]。

本稿では、特許翻訳における NMT の実用化を目指して開発された、機械翻訳システムの概要および特徴について紹介する。

2 機械翻訳システム

2.1 システム概要

本機械翻訳システムは、以下の機械翻訳サービスを提供するためのシステムである。

1）特許情報プラットフォーム（JPP）審査・審判書類 [2] 情報翻訳
JPP からの翻訳要求に応じて、審査・審判書類の日英翻訳を行い、翻訳結果を返却する。

2）日本公報情報等翻訳
JPP からの翻訳要求に応じて、日本公報や出願人名情報等の日英翻訳を行い、翻訳結果を返却する。

3）ワン・ポータル・ドシエ（OPD）審査・審判書類情報翻訳
OPD からの翻訳要求に応じて、審査・審判書類の日英翻訳を行い、翻訳結果を返却する。

4）中韓文献翻訳文作成
特許庁内システムからの翻訳要求に応じて、中韓文献の中日翻訳・韓日翻訳を行い、翻訳結果を返却する。翻訳

1　2019 年 5 月、審査・審判書類・日本公報の日英翻訳機能をリリース。2020 年度、中国公報の中日翻訳機能、韓国公報の韓日翻訳機能をリリース予定。

2　出願手続において審査・審判官等と出願人・代理人等との間で交わされる文書。

結果は DB に蓄積され、中韓文献の日本語検索サービスに利用される。

翻訳対象となる特許文献は、特許出願に係る審査・審判書類、日本公報、中国公報、韓国公報と多岐にわたり、想定される利用ユーザーも、国内一般利用者、国内審査官、海外審査官と様々である。これに対して、本機械翻訳システムでは、各翻訳要求に応じたインターフェースを提供し、各インターフェースで処理される特許文献の特性を考慮し、NMT をはじめとした複数の翻訳エンジンと翻訳前後処理を組み合わせて、翻訳処理を行う。本機械翻訳システムの詳細について以降で紹介する。

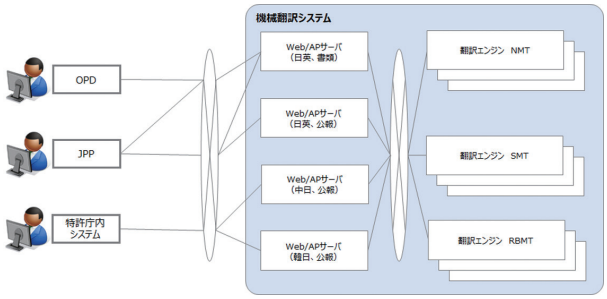

図 1　機械翻訳システムの概要

2.2　システム構成

2.2.1　翻訳エンジン

本機械翻訳システムで利用している翻訳エンジンとその主な翻訳対象を表 1 に示す。NMT は、NICT が世界最大規模の特許対訳データから構築した特許文書向けニューラル機械翻訳エンジンを採用し[5]、SMT も同様に NICT が開発した特許文書向け統計翻訳エンジンである。NMT は、流暢で高精度であるという特長を最大限に活かし、内容の理解度が求められる特許公報[3]（要約、請求項、実施形態等）が主な翻訳対象となる。さらに、国内審査官が審査した審査結果を海外審査官に正確に伝えるために、特許 NMT モデル（公報版）に対して審査結果の対訳データでカスタマイズ（ドメイン適応）した特許 NMT モデル（審査結果版）を審査結果文の翻訳に適用する。また、言語類似性から従来から高い翻訳精度を実現していた韓日方向および、原文が比較的短文かつ名詞句としての訳出が必要となる発明の名称許の翻訳については、特許 SMT モデルを利用する。

一方で、NMT は、語彙数の制約から出願人名や住所などの固有名詞を網羅的にカバーすることは難しく、また学習時の対訳データとドメインが異なる文書（特許公報以外の公報）については翻訳精度が劣化し、訳抜けや湧き出しといった NMT 特有の誤訳が生じる恐れがある。このため、書誌情報および意匠・商標・審決公報 の翻訳に関しては、大規模な固有名詞、専門用語辞書が利用可能で、特許翻訳において長年の実績のある RBMT（The 翻訳エンタープライズ TM[6]）を採用している。

2.2.2　翻訳フロー

本機械翻訳システムでは、翻訳対象となる特許文献に対して、公報フロー、審査結果フロー、固定表現フローといった複数の翻訳フローが用意されている。 翻訳フローとは、対象となる原文（段落）と、その翻訳を実行する翻訳エンジンおよび翻訳前後処理の組み合せを定義したものである。例えば、公報フローでは、特許 NMT（公報版）をベースとし、後述する翻訳前後処理を適用するとともに、NMT で誤訳を検出した場合は RBMT で訳出するという枠組みとなっている。

本機械翻訳システムの特徴として、1 つの特許文献（審査・審判書類）に対して複数の翻訳フローを用いて翻訳

3　本稿では、公開特許公報を含む公報全般を指す。

表 1　機械翻訳システムで利用する翻訳エンジンとその主な翻訳対象

エンジン名称	方式	翻訳方向	主な翻訳対象
特許 NMT（公報版）	NMT	日英	特許公報
特許 NMT（審査結果版）	NMT	日英	審査結果
特許 SMT	SMT	日英	特許公報（発明の名称）
RBMT	RBMT	日英	書誌、意匠・商標・審決公報
特許 NMT（中日）	NMT	中日	特許公報
特許 SMT（中日）	SMT	中日	特許公報（発明の名称）
特許 SMT（韓日）	SMT	韓日	特許公報

することが挙げられる。例えば、重要書類の１つである拒絶理由通知書は、図２に例示されるように出願番号や出願人名等、および定型文が記述される書類ヘッダー部、審査結果文が記述される審査結果部、および注意事項等の定型文と審査官名が記述される書類フッター部から構成されている。これに対し、書類ヘッダー部および書類フッター部は、RBMT を主エンジンとする固定表現フローへ入力し、審査結果部は、特許 NMT（審査結果版）をベースとしている審査結果フローへと入力する。各構成要素の切り分けは、キーワードによる分割ルールを適用することで実現している。これにより、原文の特性に適した翻訳エンジンが選択され、種々の情報を含む複雑な書類であっても高精度に翻訳することが可能となる。

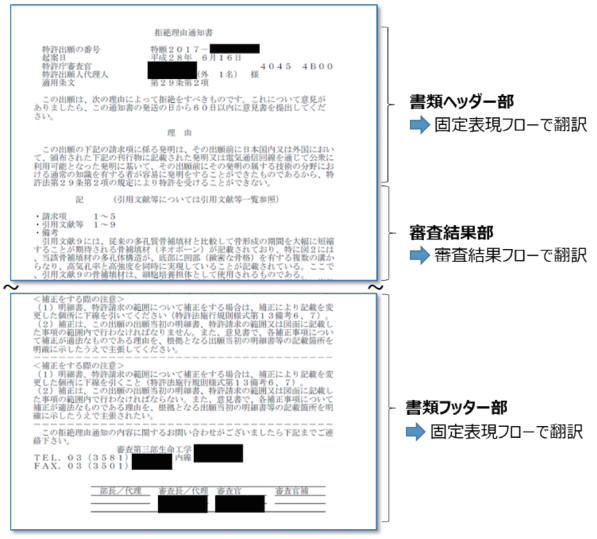

図２　審査・審判書類（拒絶理由通知書）の構成例

2.3　翻訳前後処理

大規模な特許対訳データから学習した NMT は、従来の翻訳エンジンに比べて非常に高い精度を実現しているが、実用化に向けて以下のような課題も挙げられる。

1）化合物や数式の正確な訳出

特許文書に出現する化合物表現、DNA 配列、数式等の記号列の正確な訳出を担保することが困難。

2）訳抜け・湧き出し

原文にある単語情報が抜け落ちてしまう訳抜け（Under-translation）や、原文にない単語情報が意図せず訳出されてしまう湧き出し（Over-translation）が発生する。

3）繰り返し出力

湧き出しの一部であり、意図しない単語（列）が繰り返されて訳出される現象。リカレントな構造を持つモデルに起因する誤訳であり、抜本的な解決が難しい。

特に、1）および 2）に関しては、訳文自体は非常に自然な文となっているので誤訳となっていることに気づきにくく、間違った情報伝達に繋がり、特許翻訳において大きな課題となる。これらの課題を解決するために本機械翻訳システムで開発した翻訳前後処理について以降で説明する。

2.3.1　固有表現退避

NMT では、図３に示すように、局所的に数字や記号が抜け落ちる（または沸き出す）誤訳が発生する。これを回避するために固有表現待避処理を行う。

固有表現（化合物表現）退避処理の概要を図４に示す。まず、原文中の化合物表現や数式をパターンマッチによって抽出し、抽出された固有表現を特殊タグへと置き換える。この際、化合物表現であれば、化学物質名は対象言語に逐次翻訳（リバビリン→ ribavirin、リボフラノシル→ ribofuranosyl、トリアゾール → triazole、カルボキサミド→ carboxamide）し、記号列であれば全角半角変換をすることで対象言語の文字列へと正規化する。置き換え後の原文を NMT によって翻訳したのち、訳文に現れる特殊タグを正規化された文字列で復元させる。このような処理により、化合物や数式を訳抜け・湧き出しを軽減し、正確な訳出を実現する。

[原文]リバビリン（1-β-D-リボフラノシル-1 H-1,2,4-トリアゾール-3-カルボキサミド）は、イノシン5'-一リン酸デヒドロゲナーゼ（IMPDH）の阻害剤であり、HCVの治療においてIFN-αの有効性を増強する。

[NMT誤訳例]Ribavirin (1 - β - D-ribofuranosyl-2H-1 - 1 - triazole-2-carboxamide) is an inhibitor of Inosine 5 ' - monophosphate dehydrogenase (IMPDH) and enhances the efficacy of IFN - α in the treatment of HCV. 3, 1,2,4,

図３　化合物表現に対する誤訳例

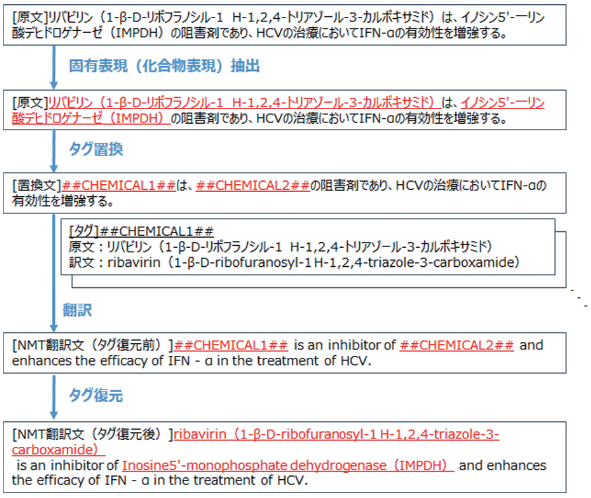

図 4　固有表現待避処理

2.3.2　繰り返し出力判定

NMT では、図 5 に示すように、訳文中に意図しない単語（この例では "（s）"）が繰り返されて訳出される場合がある（繰り返し出力と呼び）。これに対して、内部状態を示す分散表現を用いて繰り返し出力を軽減する手法 [7] も提案されているが、根本的な解決には至っていない。本機械翻訳システムでは、訳文中の繰り返し出力を検出し、検出された場合は原文を修正して再翻訳を行う。再翻訳後も繰り返し出力が改善されない場合は、別エンジン（RBMT）で再翻訳を行う。

[原文]本発明は更に、タンパク分子の解析のための対応する方法であって、ここに規定されるキットオブパーツを用いることによって存在するタンパク分子の少なくとも 1 のサブセットを標識化し、その後、標識に含まれる親和性タグを介して親和性精製によって標識化された分子を分離することを含む方法に関する。

[NMT誤訳例]The invention further relates to a corresponding method for the analysis of protein molecules, comprising labeling at least one subset of the protein molecules present by using a kit-of-parts as defined herein, and subsequently separating the labeled molecules by affinity purification via an affinity tag contained in the label (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) (s) 1.

図 5　繰り返し出力の例

2.3.3.　訳抜け・湧き出しの検出

訳抜け・湧き出しは NMT の大きな課題である。これに対し、アテンション確率および逆翻訳確率を用いて訳抜けを検出する手法 [8] 等が提案されている。しかしながら、仮に検出できたとしても、NMT において訳抜け・湧き出しの無い翻訳結果を得ることは難しい。したがって、本機械翻訳システムでは、繰り返し出力と同様に、訳抜け・湧き出しの可能性があれば、別エンジンで再翻訳を行う。すなわち、NMT の翻訳結果において、原文および訳文の単語数をベースに機械学習による識別器を用いて訳抜け・湧き出しの判定を行い、訳抜け・湧き出しと判定された場合は、別エンジン（RBMT）で再翻訳を行う。また、原文が特に短文である場合は、Encoder における文脈情報が十分に利用できず、訳抜け・湧き出しが発生する確率が高くなり、また、短文であれば RBMT でも比較的高精度な翻訳結果が期待できるため、特定の閾値以下の短文は RBMT で翻訳する。

2.4　翻訳処理速度

本機械翻訳システムは、翻訳要求を受け付けてリアルタイムに翻訳結果を返すオンデマンドサービスを提供する。特許文書には長文が含まれることが多く、最新の GPU を用いた NMT であっても実用的な応答時間で結果を返すことが難しい。これに対して、本機械翻訳システムでは、翻訳要求を適切な長さに分割し、複数の GPU サーバー（GPU コア）で分散処理を行っている。翻訳要求の分割は、改行コードや句点での分割に加え、句点を含まない長文に対しては、規則ベースの分割処理を適用している。これにより、審査・審判書類の翻訳要求に対しては 1 件 [4] あたり 6.0 秒以内（80 パーセントタイル）、日本公報の翻訳要求に対しては 1 件 [5] あたり 5.0 秒以内（80 パーセントタイル）の翻訳処理速度を実現している。

3　翻訳精度評価

3.1　自動評価

対象項目（審査結果（日英のみ）、請求項、明細書、要約、発明の名称）のテストセットに対して、自動評価を実施した。自動評価手法として BLEU[9] および RIBES[10] を用い、英語の単語分割には Moses tokenizer[6]、日本語の単語分割には Mecab[7] をそれぞれ用いた。自動評価の結果、ほぼ全ての対象項目において、BLEU40 以上、RIBES80 以上を達成し、最新の特許 NMT をベースとした高精度な機械翻訳システムであることを確認し

4　審査・審判書類は 1 文献に対して 1 件の翻訳要求が行われる。

5　日本公報は 1 段落に対して 1 件の翻訳要求が行われる。

6　https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

7　https://taku910.github.io/mecab/

た。

3.2　人手評価

本機械翻訳システムの運用において、自動評価に加えて、人手評価を定期的に実施する。人手評価の対象は、実際の翻訳要求からサンプリングによって選定し、特許文献機械翻訳の品質評価基準[11]に基づいた以下の観点で評価を行う。

1）情報の過不足がない適切な内容伝達
2）技術用語・法律用語の最適な翻訳
3）翻訳文の流暢さ
4）文法・構文の適切な翻訳
5）その他（発明の名称や請求項特有の表現、化合物や数式の表現、箇条書き等）

人手評価結果をフィードバックし、リリース後も継続して翻訳精度の維持・向上を実施していく予定である。

4　おわりに

本稿では、特許翻訳における NMT の実用化を目指して開発した、機械翻訳システムについて紹介した。本機械翻訳システムは、特許翻訳に特化した NMT をベースとし、複数の翻訳エンジンおよび前後処理を組み合わせることで、特許文献の高精度な機械翻訳システムを実現した。

2019 年 5 月より本機械翻訳システムの日英翻訳機能がリリースされ、これまで[8]に 8 千万件以上の特許文献が翻訳されている。2020 年度には、中日翻訳機能、韓日翻訳機能もリリース予定である。

今後も、利便性の高い機械翻訳サービスが提供できるよう、翻訳精度の維持・向上の改善プロセスを継続していく予定である。

8　2019 年 8 月末現在

参考文献

[1] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[3] 古田敦浩，特許情報普及活用施策に関する最近の取組と今後の展開，Japio YEAR BOOK 2019（2019）: 66-71.

[4] 特許庁から受注した機械翻訳システムの稼動開始について，https://www.toshiba-sol.co.jp/news/detail/20190531.htm

[5] 特許庁“次期機械翻訳サービス”の中核技術として NICT の技術が採用，https://www.nict.go.jp/press/2018/07/10-1.html

[6] https://www.toshiba-sol.co.jp/pro/hon_yaku/seihin/server/index_j.htm

[7] Mi, Haitao, et al. "Coverage embedding models for neural machine translation." arXiv preprint arXiv:1605.03148 (2016).

[8] 後藤功雄，and 田中英輝．"ニューラル機械翻訳での訳抜けした内容の検出．" 自然言語処理 25.5（2018）: 577-597.

[9] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[10] 平尾努，et al. "語順の相関に基づく機械翻訳の自動評価法．" 自然言語処理 21.3（2014）: 421-444.

[11] 特許文献機械翻訳の品質評価手順について，https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html


123456789
Dddddd
Ccccc