ISSN 1883-1818
No.73
December 2020

AAMT Journal

Asia-Pacific Association for Machine Translation

機 械 翻 訳

AAMT Journal
Asia-Pacific Association for Machine Translation 機 械 翻 訳
No.73

目　次

C O N T E N T S

巻頭言

翻訳システムの活用：COVID-19 世界情報集約サイト

黒橋 禎夫
京都大学大学院情報学研究科 教授

これまで永らく機械翻訳の研究開発に関わってきたが、自分自身が本当に機械翻訳を利用するということはほとんどなかった。研究者として最近の翻訳システムの向上は認識しつつ、英作文は自分の頭で行い、英語論文などの読解もそのまま英語で行ってきた。多くの研究者はそうであろう。このたび機会があって翻訳システムを実際に自分で利用することになった。その感想である。

少し長くなるが背景を説明したい。コロナ禍である。社会全体も、そして大学教育も大きな影響を受けている。未曾有の出来事であり、疾患自体についても対策と効果についても未知のことが多すぎる。しかし、世界に目を向けると、各国は様々な感染状況にあり、様々な対策がとられている。世界中のコロナに関する情報を集約・整理すれば、まだまだ出口の見えないコロナ禍において有効ではないかと考えた。コロナ禍において自分の専門性を活かして何か役に立てることはないかと考える専門家は少なくないだろう。情報系、自然言語処理の研究者と議論し、そのようなサイトをオープンコラボレーションで構築することとした。

2020 年 3 月頃から議論を始め、国内の約 10 研究室・研究機関が参加して開発を行い、世界のコロナに関連する情報を 9 地域（日本、中国、アメリカ、ヨーロッパ、アジア、南アメリカ、オセアニア、アフリカ、その他）、6 トピック（感染状況、予防・防疫・緩和、医療情報、経済・福祉政策、教育関係、その他）に整理して俯瞰できる「COVID-19 世界情報集約サイト」を 2020 年 6 月に公開した。具体的には、世界各国の有益なウェブ情報源から 1 日約 1 万ページをクローリングし、コロナ関係のフィルタリングを行って 1 日約 1 千ページを日本語へ翻訳し、さらに、トピック分類、スニペット（要約）抽出などを行う。この前処理のための翻訳には NICT の「みんなの自動翻訳」の API を提供頂き、ユーザがこのサイトから各ページを閲覧する際は Google 翻訳で動的に翻訳してもらう設定とした。深層学習の進化による翻訳精度の向上はもちろん、トピック分類の高度化や、その学習データ構築のためのクラウドソーシングの利用（約 2 万ページを人手で分類）、さらに、各地域の有益なウェブサイトを知るための現地でのクラウドソーシング、サイト情報の更新

https://lotus.kuee.kyoto-u.ac.jp/NLPforCOVID-19/

Use of Machine Translation: A System for Worldwide COVID-19 Information Aggregation
Sadao Kurohashi
Professor, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

を踏まえた正確なクローリングなどがこのサイトを支える技術である。また、開発の打合せには zoom や slack などのオンラインコミュニケーションツールを大いに活用している（もちろんのこと対面のミーティングはない)。まさに、現代の情報処理技術を総動員した「今だからできる」サイトである。

このサイトは現在も日々内容を更新し（すなわち、日々､クローリング､翻訳､分類などを動かしている)、その質の向上やさらなる機能追加についても議論を継続している。そして私自身も世界のコロナの状況把握にこのサイトを使っている。機械翻訳の質については（ここではユーザとしてページ閲覧する際に目にする Google 翻訳について)、専門用語の訳が、たとえばフランスのサイトの翻訳で「逆クラス」(「反転学習」が適当)、中国のサイトで「新しい冠状肺炎」(「新型コロナウイルス肺炎」などが適当）などこなれてないことも少なくない。しかし、全体として情報の概要を理解するには十分なレベルにあると言ってよい。英語はもちろん、フランス語、ドイツ語、中国語、韓国語などのページを翻訳を通して直接的に閲覧できる。日本のメディアが切り取った情報ではなく､各国･地域のウェブ情報を直接的に閲覧すると、各ページに貼り付けられている広告も含めて、全体としての空気感が伝わってくる。機械翻訳の原点である「世界のコミュニケーションの促進と、世界平和への貢献」に一歩近づいた気がする。一方、あらためて感じるのは、ある種の評価尺度の値に一喜一憂するのではなく、ユーザの立場でこのような実際の翻訳結果をより深く分析し、健全に研究を進めていくことの重要性である。もちろん、翻訳だけでなく、ページのコロナとの関連性判断、トピック分類、スニペット抽出など、自然言語処理技術全般としてまだまだ改善点がある。しかし、大きな一歩を踏み出していると感じる次第である。

第１５回 AAMT 長尾賞受賞記念

コニカミノルタが目指す，現場で使われるソリューションとは

川﨑 健
コニカミノルタ株式会社

1. はじめに

このたびは，AAMT 長尾賞という大変栄誉ある賞をいただき，誠に光栄に存じます．医療機関向けハイブリッド通訳システム「MELON」をご評価いただいた選考委員の方々をはじめ協会関係者の皆様には，この場をお借りして，厚く御礼申し上げます． このたびの受賞に至りました MELON につきまして，その特徴と開発経緯，医療現場に導入された理由などを簡単に紹介させていただきます．

2. MELON の特徴

MELON には大きく２つの特徴がございます．１つ目は，人間の通訳と機械の通訳が両方使えるハイブリッドな通訳機能を搭載している点です． ２つ目に，医療現場に特化して開発された通訳システムであることです．この２点の特徴について，開発経緯とともに説明いたします．

医療機関と外国人患者を聴診器とメロンの果肉で表した MELON ロゴマーク（https://www.konicaminolta.jp/melon/）

3. MELON の開発経緯

MELON は，コニカミノルタの新規事業研究開発部門である BIC Japan で開発され，2016 年 11 月にリリースいたしました．BIC Japan では，あえてコニカミノルタがこれまで手掛けたことのないビジネス領域を狙い，顧客起点を持った新規事業を生み出すことをミッションとしております．つまり，コニカミノルタのコア技術やこれまでの強み・経験に頼らず，顧客の目線に立って新規事業を作り出すことを重要視しているのです．それゆえ，MELON も例外なく「ふとした疑問」から新規事業の企画がスタートいたしました．我が国で外国人が増えてきているニュースを耳にし，外国人が日本の医療機関にかかる際に多くの課題があるのではないかと想像し，現場にお役に立ていただけるソリューションの開発が進められることとなりました．

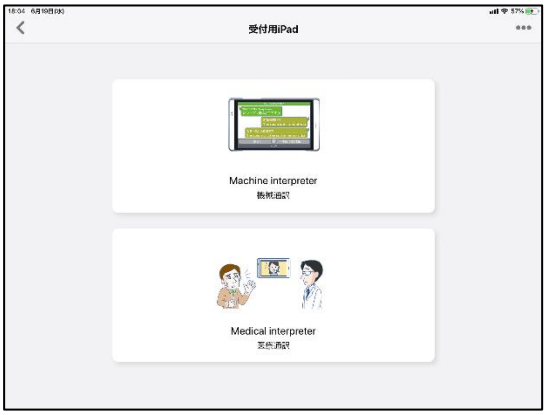
Business Innovation Center (BIC) ロゴマーク
（https://bic.konicaminolta.jp/）

外国人患者を受け入れる医療機関の負担を減らしたい想いで完成した初代 MELON は，受付から問診・診察・会計までの医療機関の業務フローに沿って機械とヒトの通訳が利用できるアプリケーションとしてリリースされました．この時の通訳機能の位置づけは，主に診察時の利用を想定して作りこまれましたが，ユーザーへのヒアリングを繰り返していくと，診察以外のプロセスでも通訳機能を広く利用したいことがわかってきました．同時に，日本人患者への対応と比べ

What are the Konica Minolta solutions that are sincerely needed at customers' sites?
Ken Kawasaki
KONICA MINOLTA,INC.

て外国人患者への対応にはイレギュラーなフローが発生することも見えてきました．そこで，日本人患者への対応フローを前提に作りこんだ業務フローベースの構成を見直し，2 代目 MELON ではフローに特化しないフレキシブルな通訳機能を中心に据えるアプリケーション構成へと作り変えました．

業務フローの流れで画面遷移していく初代 MELON



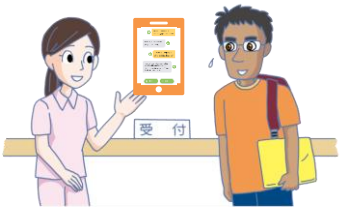
フローから通訳機能を独立させた 2 代目 MELON

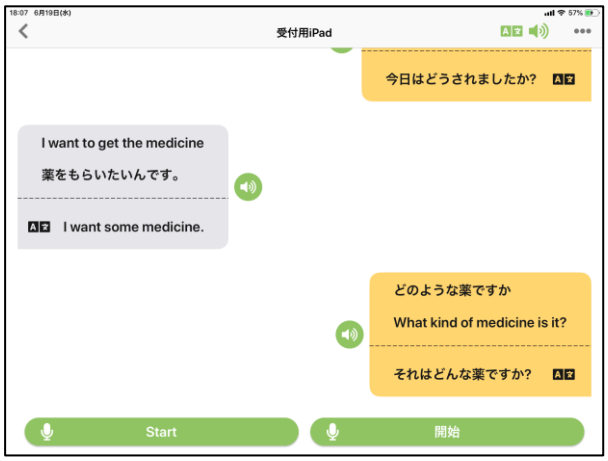
また，通訳者が遠隔で応じる「ヒトの通訳」に対し，機械通訳は回線が埋まって「ヒトの通訳」が利用できない際のつなぎ役として設けていたのですが，医療現場の皆様がアプリに使い慣れ始めると，機械通訳をメインとして利用しながら「ヒトの通訳」をサブ機能として運用されるケースが増えてきたのです．特に，方言・重い症状の説明・患者とのトラブル時のコミュニケーションの際にヒトの通訳が活躍することがわかり，機械通訳は簡単な日常会話レベルの通訳やスキンシップで気軽に使えることがわかりました．つまり，双方がシーンに応じた利用を求められていることが判明しました．そこで，3 代目 MELON では，Android 版と iOS 版に対応して機械通訳のみでもアプリ配信できるようにハード依存を脱却することとしました．これにより，利用者の好みや利用シーンで機械通訳とヒトの通訳を自由自在に選べるようになり，医療機関が眠らせていた保有タブレット端末を再利用して MELON を使えるようになりました．

4. 医療現場に導入できた理由

MELON は，これまで医療現場と何度も向き合いながら改修・開発を繰り返してきました．様々な ICT や医療機器で囲まれる現場で，患者の命と向き合う医療スタッフは常に多くのタスクに追われています．だからこそ，はじめて MELON を手にとっても直観的に操作できるユーザーインターフェースにこだわり，説明書を見ずとも利用できるアプリケーションの確立を目指して改修を行ってきました．また，2019 年 4 月の入管法改正に伴う働く外国人の増加，盛り上がったラグビーワールドカップの開催と予定していたオリンピック対応に向けて，医療機関はこれまでにない数の外国人患者を受け入れられる体制を求められ，微力ながら弊社も MELON を用いた受入体制のご支援を一緒に努めてまいりました．今年はコロナで多くの方々，医療機関の関係者様が想定外の事態に見舞われ，検疫所などでもコロナの問診で MELON が利用されることになりました．環境変化に応じて，顧客がいま何を求めているかを感じながら，現場に向き合って開発をすすめてきたつもりです．

外国人患者へのコミュニケーション支援ツールとして生まれた MELON ですが，外国人への対応は社会課題につながり，医療機関の皆様とともに未曽有の対応に取り組んだことが評価されて導入をいただけたものと考えております．もちろん，NICT の継続的な翻訳エンジン高度化の技術研究や AAMT の機械通訳発展に向けた啓蒙活動に支えていただきながら導入実績を築き上げて来られたことは言うまでもありません．





現在の MELON の機械通訳画面

5.　おわりに

MELON は，総務省が示す，「機械の通訳・遠隔でヒトが対応する通訳・現地でヒトが対応する通訳」で織りなすグラデーションの実現を理想として，来るオリンピックと万博に備えた医療機関の外国人受け入れ体制のご支援に尽力したいと考えております．また，医療業界での実績ノウハウをベースに，自治体・行政向けの多言語音声翻訳アプリ KOTOBAL も同時並行で展開していきます．社会インフラである医療と官公庁自治体を中心に，BIC Japan のモットーとする顧客起点に徹しながら，社会課題の解決に向けてソリューションを提供できるよう，この受賞を励みにより一層の努力を重ねてゆく所存です．

第１５回 AAMT 長尾賞受賞記念

特許庁におけるニューラル機械翻訳の実用化

西本 俊之
東芝デジタルソリューションズ株式会社

1. はじめに

東芝デジタルソリューションズ株式会社では、特許庁　機械翻訳システム（日英翻訳機能：2019 年 5 月リリース／中日・韓日翻訳機能：2020 年 4 月リリース）の開発・運用を行っています。この度、本システムが第 15 回 AAMT 長尾賞を受賞するという栄誉に与ったことを心より御礼申し上げます。

本システムでは、主な機械翻訳エンジンとして、国立研究開発法人情報通信研究機構（NICT）のニューラル機械翻訳エンジンを採用しています。このことは徐々に周知されてきたと感じますが、同時に「ところで東芝デジタルソリューションズは何か貢献しているのか？」とよく聞かれます。（社内からもよく聞かれます。本当に酷い（笑）。）この貴重な機会に、受賞理由とされた本システムの特徴を中心に、採用に至った経緯や実装における当社の取組を具体的に紹介します。

2. NMT エンジンを核に特徴の異なる複数のエンジンを活用

特許庁が「機械翻訳システム」の調達を検討していた 2017 年の初頭は、Google 翻訳がニューラル機械翻訳（NMT）に替わるなど、従来のルールベース機械翻訳（RBMT）、統計的機械翻訳（SMT）から急速に NMT へと時代が移り変わるタイミングでした。また、特許庁の要件（翻訳対象）は特許公報だけでなく、意匠公報、商標公報、審決公報、審査中間書類（拒絶理由通知書や意見書、手続補正書など）といった、用語や表現が互いに大きく異なる多様な知財文書が混在していました。当社の提案チームの事前検証において、NMT は学習データの豊富な特許文献（明細書、請求項、要約書等）には抜群の翻訳結果を出すのに対し、学習データに含まれない意匠公報や商標公報に対しては明らかに見劣りする状況でした。また、審査中間書類の中で最も重要な拒絶理由通知書は独特の表現が多く、特許庁の求める高い翻訳品質に達することはどうしてもできませんでした。

度重なる事前検討を経て、翻訳対象となる全種類の知財文書の特徴を分析し、知財文書の特徴に応じて複数モデルの NMT、RBMT、SMT の翻訳エンジンを切り替える方式を考案・提案に至りました。その結果、求められた高い翻訳品質を実現することができました。

本システムでは、翻訳要求された知財文書の識別コード（特実・意匠・商標・審判や書類の種類を示すコード）だけでなく、文書中の節目となるキーワードを目印に段落を分割し、最適な翻訳エンジンへの振り分けを行っています [1] 。

3. 当社の自然言語処理技術の活用

さらに特許庁　機械翻訳システムでは、レイアウト解析、化合物表現処理、翻訳メモリ処理、翻訳不要句処理、繰り返し表現回避処理等、多くの翻訳前後処理を実装しています [2] 。

ここには東芝グループが長年、ルールベース翻訳エンジン「The 翻訳™」の研究開発で培った自然言語処理技術を引き継いでいます。これらの機能を単に実装するだけでなく、各処理に用いる辞書や翻訳メモリの整備を大規模に行っています。機械翻訳をする上で固

Implementation of neural machine translation at the JPO
Toshiyuki Nishimoto
Toshiba Digital Solutions Corporation

有名詞が未知語になってしまうことはある程度、やむを得ないことと思いますが、特許情報の利用者にとっては権利者である社名、出願人名の誤訳は見過ごせない事項です。そこで当社は、本システムの構築期間中に 7000 社の国内外の企業名、25000 人の人名辞書を作成しました。辞書整備には終わりがありませんが、現在も定期的にログを解析し、参照頻度の高い固有名詞を中心に整備を続けています。翻訳メモリも 10000 レコード以上を新たに登録しています。翻訳メモリは翻訳品質を高めるだけでなく、翻訳サーバの計算リソースの有効利用、翻訳レスポンスの向上と一石三鳥の効果を生んでくれますので、日々統計をとり翻訳要求の多い順に翻訳メモリ化する運用も行っています。

4. 最新のクラウド技術

特許庁　機械翻訳システムでは、翻訳要求ごとに翻訳処理をするリアルタイム翻訳が採用されました。ニューラル機械翻訳で高速にリアルタイム翻訳するためには、複数台の GPU サーバに分散処理するよりほかはありません。また本システムではいつピークの翻訳要求がくるか分かりませんので、常にピークの処理件数を処理できる台数の翻訳サーバを用意しておく必要があります。これを低コストにかつ安定的に実現するために採用したのがパブリッククラウドです。当社のプラットフォーム構築チームにより、恐らく中央省庁のシステムの中でも最大級の GPU サーバ基盤をパブリッククラウド上に構築しています。そしてサーバ基盤の正常動作を常時監視し、不測時には適切にリカバリするための機能も備えています。

5. AI 機械翻訳導入を検討している方へ

このように、特許庁　機械翻訳システムが導入した独自の方式はどれも特許庁特有の課題への対策コレクションと言っても過言ではありません。ニューラル機械翻訳（AI 翻訳）については、その登場に期待した多くのユーザが無料サイトで試したものの何らかの課題に直面し、「まだ時期尚早だ」といって導入が取りやめられてしまうケースがあるように思いますが、たいへん勿体ないことです。ひょっとしたら当社のようなソリューションベンダが解決できる課題かもしれません。

例えば入力ファイルからテキストデータを綺麗に抽出できなかったり、特殊な構文・表現が入っていたり、多くの固有名詞が使われていたりする場合には、前処理アプリケーションで補正や退避を施すことで訳質が改善する可能性があります。

2020 年現在、NICT の機械翻訳研究者の皆さんの研究により育てられたニューラル機械翻訳の翻訳精度は、本当に素晴らしい高みに到達しています。この素晴らしい技術はこれからますます実用化されていくべきですし、その社会実装に東芝デジタルソリューションズもアプリケーション開発の力で貢献していきたいと願っています。

参考文献

[1] TOSHIBA CLIP あなたの知らない機械翻訳の世界 日本の知財戦略を支える翻訳システム
<https://www.toshiba-clip.com/detail/7768>

[2]園尾　聡，「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，Japio YEAR BOOK 2019(2019) 296-300.

第 7 回 AAMT 長尾賞学生奨励賞受賞記念

大語彙フレーズ翻訳機能付き SMT・NMT ハイブリッド翻訳方式

龍 梓
深圳技術大学

1. はじめに

機械翻訳の研究分野においては，近年，従来の統計的機械翻訳（Statistical Machine Translation; SMT）モデルに代わってニューラル機械翻訳（Neural Machine Translation; NMT）モデルによる機械翻訳方式の研究が行われている．NMT においては，原言語文の意味的要素を抽出し，固定長ベクトルへ写像してから目的言語文を生成する．そのため，NMT は意味的要素の翻訳に非常に優れており，SMT を上回る翻訳精度を達成し，非常に流暢かつ自然な訳文を生成できるようになっている[1]~[5]．しかしながら，NMT の弱点の一つとして，扱える語彙に限りがある点が知られている．具体的には，扱う語彙のサイズの増加に伴い，NMT モデルの訓練及び翻訳の計算量が増加し，結果的に要する時間が増す点が課題となっている．NMT においては，予めに設定した語彙辞書に含まれていない単語は未知語トークンとして扱われるため，誤訳の原因の一つとなる．日本語特許文を NMT によって中国語に翻訳した場合の誤り例を図 1 に示す．図 1 においては，ベースライン NMT によって生成された日本語入力文の中国語翻訳文と参照用中国語文を比較している．図 1 に示すように，略称「cmac」は原言語・目的言語とも語彙辞書に含まれておらず，出力文を生成する際に未知語として扱われ，未知語トークン< *unk* >に置き換えられて，誤訳となった．また，日本語単語「ブリッジ」（原言語の語彙辞書には含まれる）の訳語である中国語単語“桥架”も目的言語の語彙辞書に含まれていないため，出力文を生成する際に未知語として扱われ，未知語トークン< *unk* >

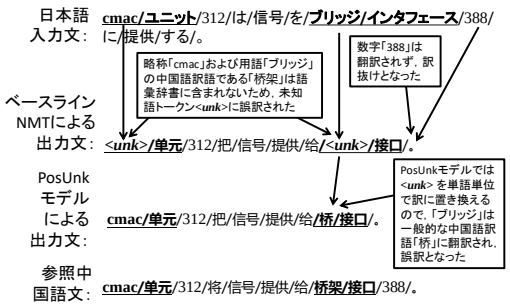


図 1　NMT によって日本語特許文を中国語に翻訳した場合の誤り例

に置き換えられ，誤訳となった．更に，入力文に含まれる数字「388」は原言語・目的言語とも語彙辞書に含まれていないため，未知語として扱われ，結果的にその訳は出力文に含まれず，訳抜けとなった．

この問題に対して，これまでにも，NMT が扱える語彙の規模を拡大する方式についての研究が幾つか行われてきた．[6]~[9] においては，訓練コーパス中の単語をサブワードもしく文字単位に細かく分割し，未知語となる語彙を減らす手法を提案している．[2]においては，大規模語彙を複数の部分集合に分割し，各語彙の確率を近似的に求めることによって，NMT が扱える語彙数を拡大する方式を提案している．[10]においては，未知語を同義となる既知語に置き換えた後，NMT モデルを訓練する方式を提案している．また，[4] においては，二言語間の単語対応関係を記録した未知語トークンを導入し，出力文中のトークンを訳語に置き換える方式によって，語彙辞書に含まれない未知語をより正確的に翻訳する NMT モデルを提案している．ここで，これらの先行研究は，いずれも，未知語となる単語をより細かい単位に分割することによっ

SMT/NMT Hybrid Machine Translation with Large Vocabulary Phrase Translation
Long Zi
Shenzhen Technology University

て未知語の問題を回避した後，NMT によって翻訳する，若しくは， 単語単位で未知語を訳語に置き換えるというアプローチとなっている．このため，未知語の翻訳の問題を，単語単位の構成的な翻訳の問題に帰着できる場合の対策にとどまっており，複合語的なフレーズの中でも， 特に，その構成単語の訳語を構成的に組み合わせる方式に帰着できない非構成的な複合語フレーズ翻訳が取り扱えない点が弱点となっている．例えば，図 1 においては，ベースライン NMT において未知語である日本語単語「ブリッジ」の中国語訳語が未知語トークン$< unk >$に誤訳されるのに対して，単語単位で未知語を訳語に置き換える PosUnk モデル[4] による翻訳結果においては，訓練文から学習された単語対応によって「ブリッジ」を一般的な中国語訳語 “桥”に置き換えて翻訳した．しかし，この中国語訳語も，日本語複合語フレーズ「ブリッジインタフェース」の参照訳“桥架接口”における「ブリッジ」の中国語訳語“桥架”とは異なっており，誤訳となっている．このように，日本語複合語フレーズ「ブリッジインタフェース」の中国語訳においては，「ブリッジ」の訳語を構成的に他の訳語と組み合わせる手法は不適切であり，非構成的な複合語フレーズ翻訳が不可避である．

以上の背景のもとで，本稿においては，ニューラル翻訳において，大規模フレーズ語彙に対応する方式について提案する．本稿の提案手法においては，SMT モデルを用いて訓練用対訳文におけるフレーズの二言語間対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．翻訳時には，NMT モデルの語彙集合中の語彙部分に対しては，NMT モデルによる訳文生成がなされ， 逆に，その他のフレーズまたは単語語彙部分に対しては，SMT モデルによる翻訳がなされる．日中，中日，日英，英日の各方向の翻訳において評価を行い， 提案手法の有効性を検証した．結果として，提案手法は，ベースラインである SMT モデル，及び，提案手法が適用されていないベースライン NMT モデルとの比較において，0.7 ポイント以上の BLEU の向上を達成できた．更に， NMT の弱点である訳抜けの改善においては，提案手法が適用されていない NMT モデルによる訳抜けを約 30%減らすことができた．

2. フレーズの抽出

本章では，branching entropy という統計的指標を用いて，前節においてトークンへの置き換え対象となる単言語フレーズを抽出する手法について述べる．

従来より，branching entropy を用いた手法は，文中におけるフレーズ分割[11], 及び, キーワードフレーズの抽出[12]等においてよく用いられてきた．それらの先行研究をふまえて，本稿では，left branching entropy（フレーズ候補の左に隣接する位置における branching entropy），及び，right branching entropy（フレーズ候補の右に隣接する位置における branching entropy）を用いてフレーズの境界を判定し，条件を満たすフレーズを抽出する．

フレーズtに対して，フレーズ t の左に隣接する形態素・単語の集合を$V_l(t)$，フレーズ t の右に隣接する形態素・単語の集合を$V_r(t)$とすると，left branching entropy 及び right branching entropy は次式で定義される．

$$H_l(t) = \sum_{v \in V_l(t)} P_l(v|t) \log_2 P_l(v|t)$$

$$H_r(t) = \sum_{v \in V_r(t)} P_r(v|t) \log_2 P_r(v|t)$$

ただし，形態素・単語の列 x の訓練文中における頻度を $f(x)$として，条件付き確率 $P_l(v|t)$，及び，$P_r(v|t)$は，それぞれ次式で定義される

$$P_l(v|t) = \frac{f(v,t)}{f(t)} \quad P_r(v|t) = \frac{f(t,v)}{f(t)}$$

以上の定義をふまえると，フレーズ t が他のより長いフレーズの部分列の場合には，t の左若しくは右には特定の形態素・単語が隣接し，branching entropy の値が小さくなるという傾向を示す．一方，tが他のフレーズの部分列ではなく，それ自身がフレーズを形成する場合， t の左右に隣接する形態素・単語の種類・

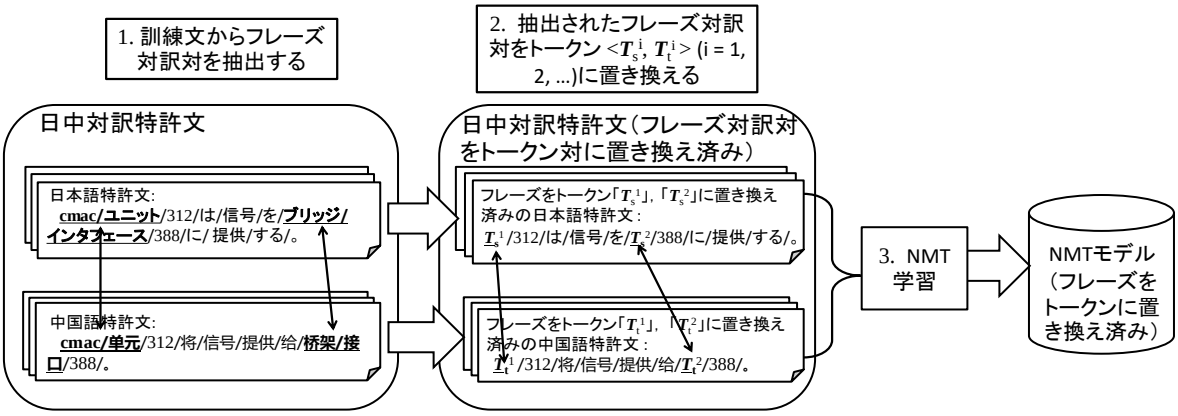


図 2　フレーズ対訳対をトークン対に置き換え済みの対訳文を用いた NMT モデルの訓練

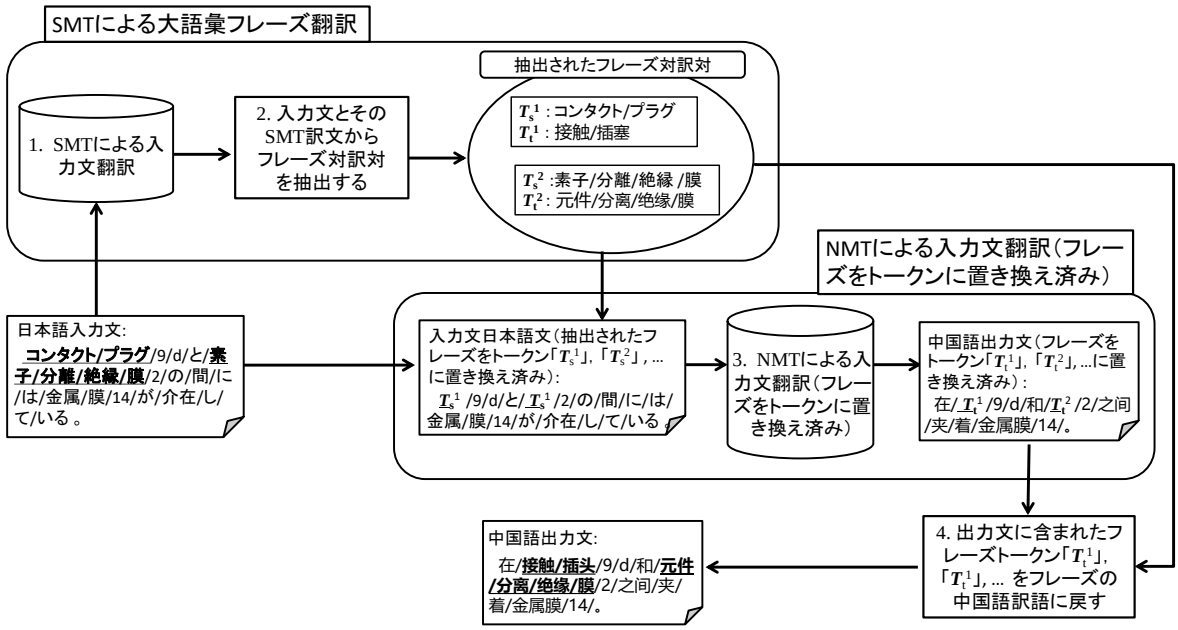


図 3　SMT による大語彙フレーズ翻訳とニューラルネットワークによる訳文生成

頻度が多様となるため，branching entropy の値が大きくなるという傾向を示す．本研究では，branching entropy の値に下限値を設定し，以下の条件を全て満たすフレーズ t を抽出する．

(i) $H_l(t)$ 及び $H_r(t)$ の値は下限値以上であり，t の任意の部分列を t_{sub} とすると，$H_l(t_{sub})$ 及び$H_r(t_{sub})$ の値は下限値未満である．

(ii) t には記号を含まない．

(iii) 訓練コーパスにおける頻度降順上位の 100 形態素・単語をストップワードとして，t は，ストップワード中の形態素・単語を含まない．

3. 大語彙フレーズに対応した NMT システム

3.1 大語彙フレーズに対応した NMT モデルの訓練

図 2 に沿って，対訳文から抽出されたフレーズの対訳対をトークン対$< T_s^i, T_t^i > (i = 1,2,\cdots)$に置き換え，その対訳特許文を用いて，大規模フレーズ語彙に対応した NMT モデルを訓練する流れを以下に示す．

ステップ 1　まず，図 2 の「1. 訓練文からフレーズ対訳対を抽出する」において，2. 節に述べた条件に従

表 1 自動評価の結果（BLEU）

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [18]	30.0	36.2	28.0	29.4
ベースライン NMT	34.2	40.8	43.1	41.8
Sentence Piece NMT	35.0	41.0	43.3	41.8
PosUnk モデル[4] による NMT	34.5	41.0	43.5	42.0
提案手法	**35.6**	**41.6**	**43.9**	**42.5**

（下線部 の BLEU スコア間には有意差なし（有意水準 5%））

い，対訳文$< S_s, S_t >$の原言語文 S_s からフレーズ t_s^i を抽出する．そして，抽出されたフレーズ t_s^i に対して，SMT モデルのフレーズ翻訳テーブル及び単語間対応を用いた訳語推定手法[13]によって目的言語文における訳語t_t^iを推定し，フレーズ対訳対 $< t_s^i, t_t^i > (i = 1,2,\cdots)$を抽出する．

ステップ 2 次に，図 2 の「2. 抽出されたフレーズ対訳対をトークン対に置き換える」において，ステップ 1 で抽出された対訳文$< S_s, S_t >$ 中のフレーズ対訳対 $< t_s^i, t_t^i > (i = 1,2,\cdots,k)$を，原言語文 S_sにおける原言語フレーズ$t_t^i(i = 1,2,\cdots)$の出現順に，それぞれ，トークン対 $< T_s^i, T_t^i > (i = 1,2,\cdots)$に置き換える．

ステップ 3 最後に，図 2 の「3. NMT 学習」において，フレーズ対訳対をトークン対に置き換え済みの対訳文を用いて，大規模フレーズ語彙に対応した NMT モデルを訓練する．

3.2 大語彙フレーズに対応した NMT モデルを用いた翻訳

図 3 に沿って，3.1 節 の手順によって訓練された，大規模フレーズ語彙に対応した NMT モデルを用いて，原言語文を目的言語訳文に翻訳する流れを以下に示す．

ステップ 1 まず，図 3 の上半分のうちの「1. SMT による入力文翻訳」において，原言語文 S_s に対して SMT モデルを適用し，SMT による目的言語訳文S_{SMT}を生成する．

ステップ 2 次に，図 3 の「2. 入力文とその SMT 訳文からフレーズ対訳対を抽出する」において，3.1 節のスッテプ 1 と同様に，2 節の手順によって入力文 S_s からフレーズ t_s^i を抽出する．そして，SMT モデルのフレーズ翻訳テーブル及び単語間対応を用いた訳語推定手法[13]によって，SMT による訳文S_{SMT} において，抽出されたフレーズ t_s^i の訳語 t_s^i を推定し，フレーズ対訳対$< t_s^i, t_{SMT}^i > (i = 1,2,\cdots,k)$を抽出する．

ステップ 3 ステップ 2 において抽出されたフレーズ対訳対 $< t_s^i, t_t^i > (i = 1,2,\cdots,k)$ に対して，原言語文 S_s 中の原言語フレーズ $t_s^i(i = 1,2,\cdots,k)$ をトークン $T_s^i(i = 1,2,\cdots,k)$に置き換え，フレーズをトークンに置き換え済みの原言語文を生成する．

ステップ 4 次に，図 3 の「3. NMT による入力文翻訳」において，フレーズ対訳対をトークン対に置き換えた後に訓練された NMT モデルを用いて，フレーズをトークンに置き換え済みの原言語文の訳文を生成する．翻訳結果の NMT 訳文においては，原言語文のトークン $T_s^i(i = 1,2,\cdots)$はトークン $T_t^i(i = 1,2,\cdots)$に翻訳されており，ステップ 2 で抽出された原言語フレーズの訳語を挿入する位置が，訳文における各トー

表 2 一対評価結果（ベースライン NMT との比較，スコアの範囲：−100 ~ 100）

手法	日中方向	中日方向	英日方向
PosUnk モデル[4] による NMT	13	12.5	14.5
Sentence Piece NMT	19.0	18	16
提案手法	**23.5**	**22.5**	**19**

表 3 JPO 基準に基づく絶対評価結果（スコアの範囲：1 ~ 5）

手法	日中方向	中日方向	英日方向
ベースライン SMT [18]	3.1	3.2	3.0
ベースライン NMT	3.6	3.6	3.7
PosUnk モデル[4] による NMT	3.8	3.9	3.9
Sentence Piece NMT	3.9	3.9	3.9
提案手法	**4.1**	**4.1**	**4.1**

表 4 入力文における訳抜けの形態素・単語の数

手法	日中方向	中日方向	日英方向	英日方向
ベースライン NMT	1,135	869	1,060	813
PosUnk モデル[4]によるNMT	1,112	846	1,031	794
Sentence Piece NMT	871	717	796	683
提案手法	**736**	**581**	**655**	**571**

表 5 名詞句翻訳性能の評価結果（対象：未知語を含む評価対象名詞句，異なり数に対しての再現率 / 延べ数に対しての再現率）

手法	日中方向	中日方向	日英方向	英日方向
ベースライン SMT [18]	28.1%(18/64) / 26.9%(18/67)	23.4%(15/64) / 23.9%(16/67)	35.4%(17/48) / 37.7%(20/53)	22.9%(11/48) / 26.4%(14/53)
ベースライン NMT	3.1%(2/64) / 3.0%(2/67)	6.3%(4/64) / 6.0%(4/67)	8.3%(4/48) / 7.6%(4/53)	6.3%(3/48) / 5.7%(3/53)
PosUnk モデル [4] による NMT	3.1%(2/64) / 3.0%(2/67)	6.3%(4/64) / 6.0%(4/67)	8.3%(4/48) / 7.6%(4/53)	6.3%(3/48) / 5.7%(3/53)
Sentence Piece NMT	7.8%(5/64) / 7.5%(5/67)	9.4%(6/64) / 9.0%(6/67)	12.5%(6/48) / 13.2%(7/53)	12.5%(6/48) / 15.1%(8/53)
提案手法	**18.8% (12/64) / 17.9% (12/67)**	**20.3% (13/64) / 19.4% (13/67)**	**18.8% (9/48) / 17.0% (9/53)**	**18.8% (9/48) / 17.0% (9/53)**

クンの位置によって特定されている．

ステップ 5 最後に，図 3 の「4．出力文に含まれたトークンをフレーズの訳語に戻す」において，ステップ 2 において抽出されたフレーズ対訳対$< t_s^i, t_{SMT}^i >$ $(i = 1,2,\cdots)$を用いて，NMT による訳文中のトークン$T_t^i(i = 1,2,\cdots)$を，それぞれ，SMT による訳文から抽出した訳語$t_{SMT}^i(i = 1,2,\cdots)$に置き換えて，最終的な訳文を得る．

4. 評価実験および考察

本稿では，WAT 2017 の特許翻訳タスク[1]において配布された 100 万文の日中対訳特許文，及び，100 万文の日英対訳特許文を用いて，提案手法の検証実験を行った．開発・評価用に，それぞれ 2,000 対訳特許文を用いた．

文単位の翻訳精度の自動評価尺度として BLEU スコア[14]を用いる．各手法に対する自動評価の結果を表 1 に示す．提案手法を用いた NMT モデルが最も高い BLEU スコアとなり，ベースライン SMT，ベースライン NMT，未知語をサブワードに分割することによって未知語の語彙を減らす Sentence Piece[2]を用いた場合の NMT，及び，単語単位で未知語を訳語に置き換える PosUnk モデル[4]の翻訳性能を上回った．

本稿の人手評価においては，[15]における人手評価尺度である「一対評価」及び「JPO 基準に基づく絶対評価」の二種類の人手評価尺度を用いる．評価用対訳特許文から無作為に抽出された 200 文を評価対象として，本稿の著者が評価を行う．PosUnk モデル[4]による NMT，Sentence Piece NMT，及び，提案手法をそれぞれ評価対象手法として，ベースライン NMT と比較した場合の「一対評価」結果を表 2 に示し、それぞれの「JPO 基準に基づく絶対評価」を表 3 に示す．これらの結果においては，提案手法が最も高い人手評価のスコアを得たことが示された．

翻訳文における訳抜けの問題の改善度合いを評価するために，評価用対訳特許文 2,000 文に対して，訳抜けした形態素・単語の数を人手で集計し，提案手法，ベースライン NMT，及び，比較対象の各手法を用いた NMT モデルとの間で比較を行った．集計結果を表 4 に示す．この結果においては，提案手法によって，ベースライン NMT における訳抜けのうちの約 30%が削減された．また，PosUnk モデルによる NMT，及び，Sentence Piece NMT と比較しても，提案手法による訳抜けの数の方が少なくなっている．

また，本稿の大語彙フレーズに対応した NMT において，大語彙フレーズの典型例である名詞句の翻訳性能がどの程度改善したのかの評価を行う．ここで，評価対象名詞句としては，構成形態素・単語として未知語を含むため， ベースライン NMT による翻訳が不可能である「未知語を含む評価対象名詞句」を用いた．人手評価において評価対象とした 200 文に対して，対象となる名詞句を人手で抽出し，各評価対象手法に対してそれらの訳語名詞句の再現率を測定した．評価

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html

[2] https://github.com/google/sentencepiece

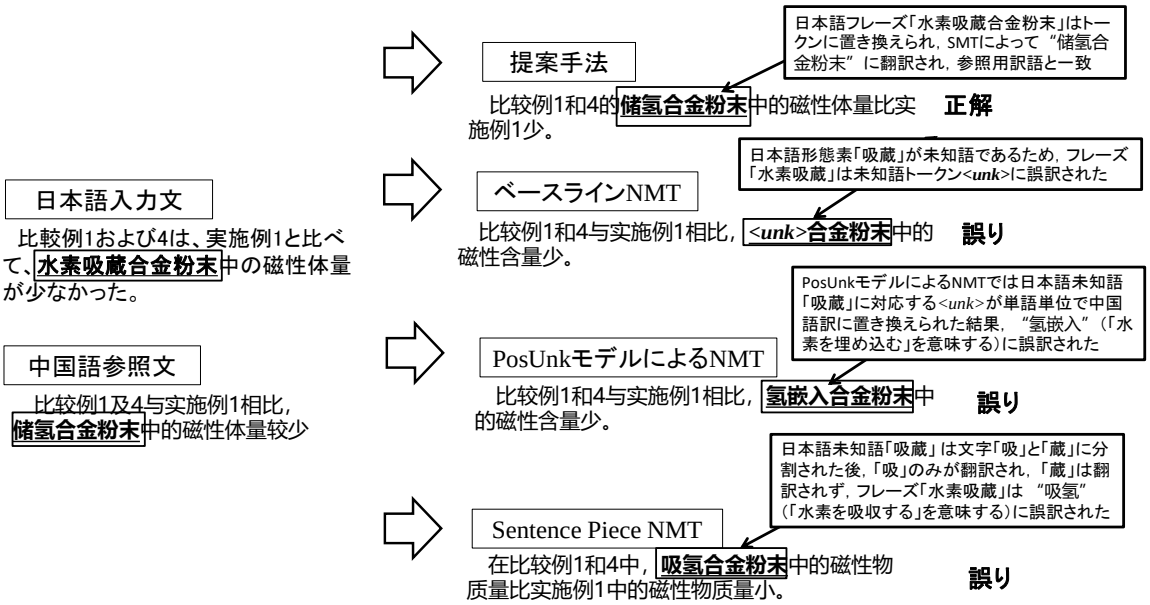


図 4　提案手法による改善例（日中翻訳）

結果を表 に示す．この結果においては， ベースライン SMT が最も高い再現率となり，提案手法はその次に高い再現率となった．この結果から分かるように，ベースライン SMT は，文単位の翻訳性能においては，自動評価・人手評価とも低い性能となっているが，翻訳結果における名詞句再現率においては，最も高い性能となった．また，ベースライン SMT 以外のベースライン手法と比べて，提案手法は相対的に最も高い性能を達成した．

提案手法による翻訳の具体例として，提案手法によって，三種類のベースライン手法であるベースライン NMT, PosUnk モデルによる NMT, 及び, Sentence Piece NMT の翻訳誤りを改善する例を, 図 4 に示す．

5. おわりに

本稿では，NMT モデルが大規模フレーズ語彙に対応できないという問題に対して一つの解決手法を提案した．提案手法では，まず，訓練用対訳文においてフレーズ間の二言語対応の情報を収集し，二言語間で対応済みのフレーズ対訳対を同一のトークンに置き換えた後，NMT モデルの訓練を行う．次に，大語彙フレーズ翻訳に対応した NMT モデルによる翻訳時には，NMT モデルの語彙集合に含まれる語彙部分に対しては，NMT モデルによる訳文生成がなされ，一方，その他のフレーズまたは単語語彙部分に対しては，SMT モデルによる翻訳がなされる．日中，中日，日英，英日の各方向の翻訳において評価を行い，提案手法の有効性を検証した．評価結果として，ベースライン SMT モデル，及び，ベースライン NMT モデルとの比較において，0.7 ポイント以上の BLEU の向上を達成できた．また，NMT の弱点の一つである訳抜けの改善においては, ベースライン NMT モデルによる訳抜けを約 30%削減することができた．今後の課題として，SMT による大語彙フレーズ翻訳と Sentence Piece に基づく NMT モデルを併用することにより翻訳性能を改善することが挙げられる．また，NMT モデルにおける訳抜けを改善するための既存の手法（例えば，[16]等）の枠組みを導入することにより，提案手法における訳抜けの改善を更に促進することが挙げられる．本稿の内容の詳細は文献[17]を参照されたい．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Proc. 3rd ICLR, pp.1–15, 2015.

[2] S. Jean, K. Cho, Y. Bengio, and R. Memisevic, “On using very large target vocabulary for neural machine translation,” Proc. 28th NIPS, pp.1–10, 2014.

[3] M. Luong, H. Pham, and C.D. Manning, “Effective approaches to attention-based neural machine translation,” Proc. EMNLP, pp.1412–1421, 2015.

[4] M. Luong, I. Sutskever, O. Vinyals, Q.V. Le, and W. Zaremba, “Addressing the rare word problem in neural machine translation,” Proc. 53rd ACL, pp.11–19, 2015.

[5] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to sequence learning with neural machine translation,” Proc. 27th NIPS, pp.3104–3112, 2014.

[6] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” Proc. 54th ACL, pp.1715–1725, 2016.

[7] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” http://arxiv.org/abs/1609.08144, 2016. [Online; accessed 28-December-2017].

[8] M. Luong and C.D. Manning, “Achieving open vocabulary neural machine translation with hybrid word-character models,” Proc. 54th ACL, pp.1054–1063, 2016.

[9] M.R. Costa-Juss`a and J.A.R. Fonollosa, “Character-based neural machine translation,” Proc. 54th ACL, pp.357–361, 2016.

[10] X. Li, J. Zhang, and C. Zong, “Towards zero unknown word in neural machine translation,” Proc. 25th IJCAI, pp.2852–2858, 2016.

[11] Z. Jin and K. Tanaka-Ishii, “Unsupervised segmentation of Chinese text by use of branching entropy,” Proc. COLING/ACL 2006, pp.428–435, 2006.

[12] Y. Chen, Y. Huang, S. Kong, and L. Lee, “Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features,” Proc. 2010 IEEE SLT Workshop, pp.265–270, 2010.

[13] Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto, “Translation of patent sentences with a large vocabulary of technical terms using neural ma- chine translation,” Proc. 3rd WAT, pp.47–57, 2016.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” Proc. 40th ACL, pp.311–318, 2002.

[15] T. Nakazawa, H. Mino, I. Goto, G. Neubig, S. Kurohashi, and E. Sumita, “Overview of the 2nd workshop on Asian translation,” Proc. 2nd WAT, pp.1–28, 2015.

[16] I. Goto and H. Tanaka, “Detecting untranslated content for neural machine translation,” Proc. 1st NMT, pp.47–55, 2017

[17] 龍梓，木村龍一郎，飯田頌平，宇津呂武仁，三橋朋晴，山本幹雄．“フレーズ・トークン込み NMT モデル及び SMT による大語彙フレーズ翻訳によるハイブリッド翻訳方式”，電子情報通信学会論文誌,Vol. J102–D No. 3, 2019.

第 7 回 AAMT 長尾賞学生奨励賞受賞記念

潜在変数モデルに基づいた非自己回帰型ニューラル機械翻訳

朱 中元
東京大学

1. はじめに

2014年にSutskeverら [1]によってニューラルネットワークで実現したエンコーダ・デコーダモデルが提案されて以来，ほとんどのニューラル機械翻訳モデルがこの設計に従っている．エンコーダ・デコーダモデルでは，エンコーダによって入力文の隠れ状態が計算され，デコーダによって出力側の単語が順に予測される．ここで，デコーダは入力文の情報と既に出力した単語の情報を条件にした言語モデルと考えられる．このように定式化したモデルは自己回帰型モデルと呼ばれる．翻訳時には beam search による翻訳文の探索を行うのが一般的である．

Transformer モデル [2]の提案によってニューラル機械翻訳の翻訳品質は著しく向上した．しかし，条件付き言語モデルを用いる自己回帰型モデルは翻訳時間が出力単語数に対して線形に増加する欠点が存在する．一方で，直近の研究では従来の文単位ではなく文書単位の翻訳が注目されており，翻訳速度の高速化は重要な課題となっている．

出力単語数 N に対し，従来の自己回帰型モデルの翻訳時における計算量は O(N)である．理論計算量を向上させるためには，対数時間 O(log N) か定数時間 O(1)を目指す必要がある．自己回帰型モデルでは，各単語の予測確率がそれよりも前に出力した単語に依存するため，これらの理論計算量を達成できない．必然的に，機械翻訳の高速化を目指す研究は新たな方針を考案しなければならない．

2018 年の Gu らの研究 [3]により，GPU で定数時間 O(1)を目指す非自己回帰型翻訳モデル[1]が提案された．Gu らのモデルでは，まず各入力単語の繁殖率（各入力単語に対して対応づけられる出力文中の単語数）を予測し，出力文の文長を確定する．次に，予測した繁殖率を条件にし，各出力単語を独立に予測する．繁殖率の予測モデルは，fast_align で実装された IBM Model 2 によって作成した単語対応付けデータを用いて教師あり学習したものである．ここで，入力単語の繁殖率を教師データありの潜在変数Zとみなすことができる．この場合，入力文Xから出力文Yを予測する翻訳モデルは$p(Y|X,Z)=\prod_{i=1}^{N}p(y_i|X,Z)$と定式化できる．この式からわかるように，潜在変数が確定した後は各y_iの出力確率は互いに依存性を持たない．

直観的には，潜在変数は翻訳プランを表していると理解できる．例えば，「ジョンは傘を持っている」という文に対して，“John has an umbrella”と“John is holding an umbrella”の二つの翻訳候補が考えられる．しかし，「持っている」→“is holding”という翻訳プランを採用したならば，最終的な翻訳文はこの時点で後者のほうに確定する．このように，高品質の翻訳文を出力するためには，潜在変数が出力単語の選択における曖昧性をほとんど解消している必要がある．

Gu らのモデルでは入力単語の繁殖率を潜在変数として用いたが，繁殖率が翻訳文における曖昧性解消に最適な潜在変数である保証はない．本研究では，非自己回帰型翻訳モデルを構築するために，最適な潜在変数を自動的に獲得するモデル，および潜在変数の推論

[1] 定数時間で計算するために，GPU が全ての単語に対して，同時に計算できることが前提条件である．

Non-autoregressive Neural Machine Translation Based on Latent-Variable Models
Raphael Shu
The University of Tokyo

方法を提案する[2]. WMT'14 英独翻訳タスクにおいて，提案モデルは Transformer ベースラインモデルから BLEU スコアを 0.9% しか落とさず，約 6 倍の翻訳速度の向上を実現できた．本稿では定量評価の結果を報告し，翻訳速度及び変数の解釈性について考察を行う．

2. 連続潜在変数を持つ非自己回帰型ニューラル機械翻訳モデル

本研究では，潜在変数を自動的に学習するために，学習が比較的に安定的である変分オートエンコーダ (VAE) [4]を採用する．近年，GumbelSoftmax [5]と VQ-VAE [6]の提案によって，離散潜在変数の学習も可能になったが，学習時のバリアンスが大きいという欠点がある．そのため，本研究では扱う潜在変数を連続値とした．提案モデルでは，各入力単語x_iに対し，潜在変数z_iを付与する．すなわち，潜在変数の個数は入力文の単語数と一致する．ここで，z_iは低次元連続ベクトルである．潜在変数の系列 $\{z_1, \ldots, z_{|x|}\}$ をZと表記すると，翻訳モデルの対数尤度は下の式のように周辺化できる．

$$\log p\,(Y|X) = log \int p(Y|X,Z)\, p(Z|X)\, dZ$$

$$= log \int \prod_i p(y_i|X,Z)\, p(Z|X)\, dZ\,. \qquad (1)$$

モデルを学習するために，変分法を用いて対数尤度の変分下限 (ELBO) を目的関数にし，最適化を行う．

$$log\, p\,(Y|X) \geq \text{ELBO(X,Y)}$$

$$= \mathbb{E}_{z\sim q_\phi}\left[\log \prod_i p_\theta\,(y_i \mid X,Z)\right]$$

$$-\text{KL}\big[\, q_\phi(Z|X,Y) \,||\, p_\omega(Z|X)\big]. \qquad (2)$$

式（2）に示しているように，変分下限はデコーダ分布 $p_\theta(y_i|X,Z)$と近似事後分布 $q_\phi(Z|X,Y)$，事前分布 $p_\omega(Z|X)$の三つの分布によって構成される．本研究では，事前分布と近似事後分布は共に多変量正規分布とし、共分散行列は対角行列とした．簡便化のため，以降ではデコーダ分布を$\prod_i p_\theta\,(y_i \mid X,Z)$ に展開せず，$p_\theta(Y|X,Z)$で表す．

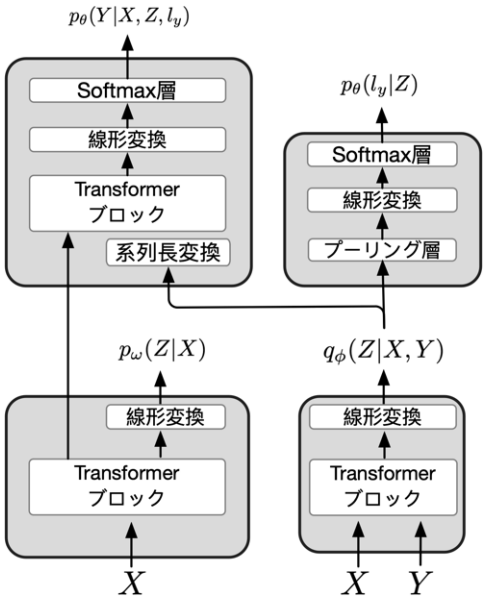


図 1: ニューラルネットワーク構造

2.1 ニューラルネットワーク構造

潜在変数 Zを含めて全ての分布の入出力はベクトルの系列のため，Transformer モデルのネットワーク構造を用いて計算できる．図 1 に具体的なネットワーク構造を示す. 事前分布と近似事後分布は多変量正規分布であるため，線形変換で分布の平均と分散ベクトルを予測する．また，デコーダ分布を $p_\theta(Y|X,Z) = p_\theta(Y|X,Z,l_Y)\, p_\theta(l_Y|Z)$ のように分解することによって，系列長予測モデルを独立させる．出力単語Yと出力系列長l_Yを予測する分布は共にカテゴリカル分布とする．

デコーダ（図 1 の左上）はl_Y個の単語を予測する必要があるが，事後分布$q_\phi(Z|X,Y)$からサンプリングした潜在変数のベクトル数は入力系列長l_Xであるため，サンプリングした潜在変数の系列長をl_Xからl_Yに調整する必要がある．本研究では，Local Attention [8]を用いて，ベクトル数の調整を行い，各出力単語y_iに対して $\frac{l_X}{l_Y} i$ を Attention の中心点とする．

[2] 本研究のソースコードは https://github.com/zomux/lanmt にて公開している.

2.2 潜在変数の推論

図１で示す翻訳モデルを学習し，潜在変数から出力単語のマッピング関数を得る．

$$F(X,Z) = \text{argmax}_Y\, p_\theta\left(Y \middle| X, Z, \widehat{l_Y}\right), \quad (3)$$

$$\widehat{l_Y} = \text{argmax}_{l_Y}\, p_\theta(l_Y|Z). \quad (3)$$

ここで，翻訳文Yの品質は潜在変数 Z の選択によって決まる．潜在変数の導入によって，良い出力文に導くZを推論することが重要な課題である．そこで，本研究では正確かつ高速に計算できる推論法の提案を目指す．

ナイーブ推論法　生成モデルの学習によって事前分布$p_\omega(Z|X)$が得られるので，最も簡単な推論方法は下記のように事前分布の期待値をとって出力文を生成することである．

$$\hat{Z} = \mathbb{E}_{p_\omega(Z|X)}[Z], \quad (3)$$

$$\hat{Y} = F\left(X, \hat{Z}\right). \quad (4)$$

多変量正規分布の場合，期待値は分布の平均ベクトルである．

デルタ分布推論法　ナイーブ推論法は事前分布の期待値をとるが，これによって得られた翻訳文$\hat{Y}$が変分下限 $\text{ELBO}\left(X, \hat{Y}\right)$ を最大にする保証はない．本研究では，変分下限を最大化するために，式（２）の近似事後分布$q_\phi(Z|X,Y)$を決定的デルタ分布 $r(Z)$ に置き換えることを提案する [9]．デルタ分布$r(Z)$は下記の性質を持つ．

$$r(Z = \mu) = 1, r(Z \neq \mu) = 0. \quad (5)$$

すなわち，この分布のすべての確率質量はある特異点に置かれる．$\text{KL}(\, r(Z) \,||\, q_\phi(Z|X,Y)\,)$を最小化することで，$\mu = E_{q_\phi(Z|X,Y)}[Z]$であることが分かる．

近似事後分布をデルタ分布に置き換えた変分下限は下記のようになる．

$$\text{ELBO}(X,Y) \approx \text{ELBO}(X,Y,\mu)$$

$$= \log p_\theta\,(Y|X, Z=\mu) - \log p_\omega\,(\mu|X)\,. \quad (6)$$

新たな変分下限$\text{ELBO}(X,Y,\mu)$を最大化すると，出力文の最適値は $\text{argmax}_Y \log p_\theta\,(Y|X,Z=\mu)$，すなわち$F(X,\mu)$であることが分かる．こうして，KL ダイバージェンスと変分下限の最大化を交互に行う更新アルゴリズムリズムを導出できる．

$$\mu = \mathbb{E}_{q_\phi\left(Z|X,\hat{Y}\right)}[Z], \quad (7)$$

$$\hat{Y} = F(X,\mu)\,. \quad (8)$$

ここで，出力$\hat{Y}$の初期値はナイーブ推論法の結果，すなわち$F\left(X, \mathbb{E}_{p_\omega(Z|X)}[Z]\right)$を用いる．翻訳時に，式（７）と式（８）を交互に実行し，出力文を更新する．出力$\hat{Y}$の変化が観測されない，あるいは最大更新回数に達すると推論を終了する．

エネルギー推論法　デルタ分布推論法は近似事後分布を利用して簡単に実装できるが，出力$\hat{Y}$を更新するために，式（８）のデコーディング関数を繰り返して計算する必要がある．デコーダの Softmax 層の計算を重ねることによりデルタ分布推論法の速度は大幅に低下する．さらなる速度の向上を目指すためには，潜在変数の推論を完全に連続ベクトル空間の中で完結する必要がある．

式（７）と式（８）による更新アルゴリズムは本質的に，ある推論ステップの潜在変数μ_tをμ_{t+1}に更新する過程である．変分下限の向上によりμ_{t+1}の品質はμ_tを上回ることが分かる．ここで，もし連続ベクトル空間上の全ての可能な潜在変数Zの品質を正確に測定できるエネルギーモデル$E_\psi(Z;X)$を学習できるならば，エネルギーモデルの勾配$\nabla_Z E_\psi(Z;X)$に従って，潜在変数を更新することができる．エネルギー関数は Softmax 層を持たないため，推論速度の向上が期待できる．

連続空間上に任意の潜在変数Zの品質を評価できる関数を$\tau(Z;X)$とする．本研究は異なる潜在変数ベクトルZとZ'に対して，それぞれのエネルギー値の差分で$\tau(Z';X) - \tau(Z;X)$を近似するように学習を行う．

$$\min_\psi || \left(-E_\psi(Z';X) + E_\psi(Z;X)\right) - \left(\tau(Z';X) - \tau(Z;X)\right) ||^2, \quad (9)$$

$$\approx \min_\psi || \nabla_Z E_\psi(Z;X) ||^2 + 2(\, \nabla_Z E_\psi(Z;X)^\top \nabla_Z \tau(Z;X)\,). \quad (10)$$

式(９)で示す損失関数をテイラー展開することで，式（１０）のようにエネルギー関数の勾配に関する損失関数に変換できる．このようにエネルギー値ではなく，勾配を学習の対象にするアプローチは Score Matching [10]と呼ばれる．

しかし，品質関数$\tau(Z;X)$の計算において，どうしても決定的デコーディング関数$F(X,Z)$を取り入れる必要があるため（例えば，$\tau(Z;X) = log\, p\,(F(X,Z)|X)$と定義する場合），通常$\tau(\cdot)$は微分不可能な関数となる．本研究では，この問題を避けるために，デルタ分布推論の結果を用いて品質関数の勾配を近似する [10]．潜在変数Zに対してデルタ分布推論を行った結果を$\mathrm{Delta}(Z;X)$で表す．本研究は下記の損失関数でエネルギーモデルの学習を行う．

$$\mathcal{L} = ||\, \nabla_Z E_\psi(Z;X)\, ||^2 + 2(\, \nabla_Z E_\psi(Z;X)^\top (Delta(Z;X) - \mathrm{Z})\,). \quad (11)$$

実験では，Song ら [12]の研究に従い，エネルギーの勾配$\nabla_Z E_\psi(Z;X)$の代わりにニューラルネットワーク$f_\psi(Z;X)$で近似する疑似勾配モデルに対しても評価を行う．

3. 実験結果と考察

実験では，潜在変数に基づいた非自己回帰型翻訳モデルに対して，主に翻訳文の品質及び翻訳速度の評価を行う．まず，WMT’14 英独翻訳データセット [13]を用いて定量評価を行う．そして，ASPEC 日英データセット [14]を用いて定性評価を行う．それぞれのデータセットは 450 万と 300 万の対訳文ペアから構成される．WMT’14 データセットの前処理は sentencepiece [15]を用いて最終語彙数が 3.2 万になるように処理した．

ネットワークの構成　本実験では，ニューラルネットワークの複雑度とハイパーパラメタを可能な限り Transformer Base [2]に一致させた．提案モデルでは，事前分布とデコーダ分布の計算に 6 層の Transformer ブロックを使用し，近似事後分布には 3 層の Transformer ブロックを使用した．エネルギー推論法で用いるエネルギーモデルの構成には，6 層の Transformer ブロックを使用した．エネルギー値を予測するために線形変換及び平均プーリング層を加える．疑似勾配モデルでは，エネルギー値が不要なため，Transformer ブロックの出力で直接エネルギー勾配を近似する．

表 1　WMT’14 英独翻訳における定量評価の結果

モデル	BLEU (%)	実時間 (ms)	速度向上
Transformer Base (ビーム幅=3)	28.3	251	1x
Transformer Base (ビーム幅=1)	27.5	291	1.1x
潜在変数モデル ナイーブ推論	25.7	19	15x
潜在変数モデル デルタ分布推論	26.1	46	6.3x
潜在変数モデル エネルギー推論	26.1	50	5.8x
潜在変数モデル エネルギー推論・疑似勾配	26.3	29	10x
潜在変数モデル エネルギー推論・疑似勾配 ・並列化探索	27.4	47	6.2x

知識蒸留法　非自己回帰型ニューラル機械翻訳の既存研究にならい，本実験も知識蒸留法によって学習データに対して事前翻訳を行う．具体的には，ベースライン Transformer モデルを用いて学習データを翻訳し，その翻訳結果を提案モデルの教師データにした．

潜在変数の並列化探索　デルタ分布推論法及びエネルギー推論法は事前分布の期待値を潜在変数の初期値としたが，事前分布から複数の潜在変数をサンプリングすることで同時に探索することもできる．これによって得られる複数の翻訳結果に対して Transformer ベースラインモデルを用いてリランキングを行い，スコアの最も高い翻訳結果を最終出力とする．本実験では，潜在変数を 5 個サンプリングし，それぞれに対してさらに出力文長の候補を 5 個探索した結果を報告する．

3.1 定量評価

表 1 に提案モデルとベースラインモデルの翻訳品質（BLEU 値）と翻訳速度を示す．実翻訳時間は，Nvidia V100 を 1GPU 使用し，WMT’14 英独翻訳のテストデータ 3,003 文を用いて計算した 1 文あたりの平均実翻訳時間である．

ナイーブ推論法の翻訳速度はベースラインの 15 倍であるが，BLEU 値は 2.6%下回っていることが分かる．この速度は提案モデルが達成できる翻訳速度の最

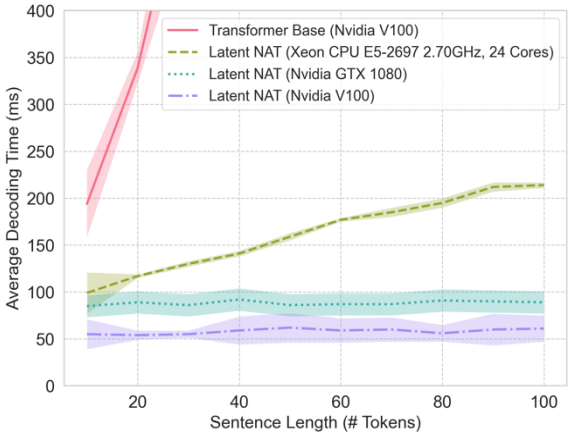


図 2: 提案モデルにおける入力文の単語数と翻訳時間の関係

速値と考えられる．デルタ分布推論法とエネルギー推論法では約 6 倍の速度向上が見られ，ナイーブ推論法の翻訳品質を上回ることができた．疑似勾配モデルを用いた場合，連続空間内で勾配逆伝搬なしに行われるため，ベースラインと比較して 10 倍の速度向上を達成できた．潜在変数の並列化探索を行う場合，翻訳速度を落とすことで，Transformer ベイスラインの BLEU スコアとの差分を 0.9 まで減少させることができた．

3.2 並列処理性能と翻訳速度

異なる単語数の入力文に対する提案モデルの実翻訳時間を図 2 で示した．本実験では，デルタ分布推論法を用いる．計算マシンの並列処理性能による影響を明らかにするために Xeon E5 CPU，Nvidia GTX1080 (1GPU), Nvidia V100 (1GPU) のそれぞれで翻訳速度を測定した．

Transformer ベースラインは自己回帰型モデルであるため, GPU を使用した場合でも翻訳時間が線形に増加してしまう．入力文の文長を 100 単語に設定した場合，平均翻訳時間は約 1,000ms に達する．一方，提案モデルを GPU で実行した場合，入力文の単語数によらず翻訳時間が一定であることが分かる. CPU で提案モデルを実行した場合，計算デバイスの並列化処理性能が低下しているため，翻訳時間が線形に増加した．

表 2　潜在変数の更新による翻訳文の変化

例 1　表記減衰量の確立を試みた．	
翻訳出力（推論前）	an attempt was made establish establish damping attenuation standard ...
翻訳出力（推論後）	an attempt was to establish the damping attenuation standard ...
正解翻訳文	the establishment of an optical fiber attenuation standard was attempted .
例 2　...「線膨張係数の取り扱い」について述べた．	
翻訳出力（推論前）	... “ handling of of linear expansion coefficient ” are described .
翻訳出力（推論後）	... “ handling of linear expansion coefficient ” are described .
正解翻訳文	... handling of linear expansion coefficient .
例 3　...マイクロマニピュレーションへと発展しており ...	
翻訳出力（推論前）	... micro micro manipulation and ...
翻訳出力（推論後）	... and micro manipulation , and it has been developed , and ...
正解翻訳文	... with wide application fields so that it has been developed ...

3.3 潜在変数の更新による翻訳文の変化

潜在変数Zを更新することによって翻訳モデルの出力$F(X,Z)$も変化すると予想される．表 2 に ASPEC 日英翻訳タスクにおける 3 つの更新例を示す．それぞれの例では，日本語側の入力，初期潜在変数による翻訳結果，更新後の潜在変数による翻訳結果および正解翻訳文を示す．潜在変数の更新による翻訳文の変更箇所を下線で示した．

例 1 では，潜在変数の推論により，出力文の単語数を変更させずに重複単語を修正している．例 2 では，重複単語を消去することで翻訳文を改良している．例 3 では，翻訳漏れに対して追加情報を翻訳文に差し込むことに成功している．

4. 潜在変数の解釈性に対する考察

提案モデルでは各入力単語x_iに対して潜在変数z_iを付与するため，各潜在変数は対応する単語の部分翻訳を捉えていると考えられる．学習後の潜在変数が捉えている情報を分析するために，潜在変数を部分的に変化させ，これによって翻訳文がどのように変化するのかを見ていく．

表３　一部の潜在変数を変化させたときの結果

入力文　DERS ソフトウエアを用いて「ふげん 発電所」の線量率を詳細に計算できる。	
翻訳出力 1	using the DERS software , the dose rate of “ Fugen plant ” can be calculated in detail .
翻訳出力 2	using DERS software , the dose rate rate “ Fugen power plant ” can be calculated in detail .
翻訳出力 3	using DEDERS software , the dose rate of “ nuclear plant ” can be calculated in detail .

表 3 に結果を示す．潜在変数の初期値を用いて翻訳した結果が「翻訳出力 1」である．続いて，「ふげん 発電所」に対応する潜在変数のみに対して新たにランダムサンプリングを行う．部分的に変化した潜在変数を用いて得られた翻訳文が「翻訳出力 2」または「翻訳出力 3」である。「ふげん 発電所」に対応している翻訳箇所を下線で示した．

表 3 で示されている翻訳結果の変化から分かるように，潜在変数の一部を更新した場合，該当箇所の翻訳文のみが変化する傾向がある．この例文では，“Fugen plant”から“Fugen power plant”または“nuclear plant”に変化している．該当箇所以外の潜在変数は更新されていないため，翻訳文の変化は少ない．この結果は各潜在変数が対応している単語の部分翻訳を捉えていることを示している．

次に，潜在変数を連続ベクトル空間内で移動させた場合の翻訳文の変化を見てみる．表 4 に示した日本語入力文に対して，まず異なる翻訳文を出力する 2 つの潜在変数を選定する．それぞれの翻訳結果を「潜在変数 1」と「潜在変数 2」に示す．この 2 つの潜在変数を用いて線形補間を行い，補間率を調整した際の翻訳文の変化を表 4 にまとめた．

補間率が 0.2 未満，あるいは 0.32 を超える場合，翻訳結果の変化が見られなかった. 補間率が 0.2~0.32 の場合，出力文の単語が局所的に置換される．他の例文においても，潜在変数を移動する際に翻訳文が大幅に変化することはなかった．

表 4　潜在変数を線形補間したときの結果

入力文　リサイクルに関する最近の話題を紹介した。	
潜在変数 1	**this paper introduces recent topics on recycling.**
補間率 0.20	this paper introduces recent topics on recycling.
補間率 0.24	recent paper introduces recent topics on recycling.
補間率 0.28	recent paper introduces recent recycling are recycling.
補間率 0.30	recent topics introduces the recycling are recycling.
補間率 0.32	recent topics on the recycling are introduced.
潜在変数 2	**recent topics on the recycling are introduced.**

5. おわりに

本研究では，ニューラル機械翻訳の翻訳速度の向上を目指す潜在変数モデルに基づいた非自己回帰型翻訳モデルとその推論法を提案した．また，潜在変数モデルを用いて定式化する場合，翻訳文の品質が潜在変数の選択によって決まるため，潜在変数の推論が極めて重要な課題となる．そこで，正確かつ高速に推論を行うために，デルタ分布推論法 [9]およびエネルギー推論法 [10]を提案した．現段階では，これらの推論法を適用し，潜在変数を並列探索することで，WMT’14 英独翻訳タスクにおいてはベースラインモデルと近い翻訳品質で約 6 倍の翻訳速度向上を実現することができた．本稿では，さら学習した潜在変数の解釈性について分析を行い，各潜在変数がそれぞれ対応している入力単語の部分翻訳を捉える傾向があることがわかった.

本研究の最も著しい貢献は，連続空間での潜在変数ベクトル更新による非自己回帰型ニューラル機械翻訳を実現するフレームワークを提案したことである．離散空間での単語精緻化モデル [15]と比べて連続空間での操作は Softmax 層の計算が不要なため，容易に高速化できるという利点がある．今後，連続空間におけるより正確な推論方法が提案されることで，非自己回帰型ニューラル機械翻訳の更なる翻訳品質の向上が期待できる．

参考文献

[1] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *NIPS*, 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[3] J. Gu, J. Bradbury, C. Xiong, V. O. K. Li and R. Socher, "Non-autoregressive neural machine translation," in *ICLR*, 2018.

[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *ICLR*, 2014.

[5] E. Jang, S. Gu and B. Poole, "Categorical Reparameterization with Gumbel-Softmax," in *ICLR*, 2016.

[6] A. Oord, O. Vinyals and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *NIPS*, 2017.

[7] M.-T. Luong, H. Pham and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *EMNLP*, 2015.

[8] R. Shu, J. Lee, H. Nakayama and K. Cho, "Latent-Variable Non-Autoregressive Neural Machine Translation with Deterministic Inference Using a Delta Posterior," in *AAAI*, 2020.

[9] A. Hyvärinen, "Estimation of Non-Normalized Statistical Models by Score Matching," in *Computer Science, Mathematics J. Mach. Learn. Res.*, 2005.

[10] J. Lee, R. Shu and K. Cho, "Iterative Refinement in the Continuous Space for Non-Autoregressive Neural Machine Translation," in *EMNLP*, 2020.

[11] Y. Son and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in *NeuIPS*, 2019.

[12] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut and L. Specia, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.

[13] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi and H. Isahara, "ASPEC: Asian Scientific Paper Excerpt Corpus," in *LREC*, 2016.

[14] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *EMNLP*, 2018.

[15] M. Ghazvininejad, O. Levy, Y. Liu and L. Zettlemoyer, "Mask-Predict: Parallel Decoding of Conditional Masked Language Models," in *EMNLP/IJCNLP*, 2019.

解説記事

Red Hat が目指す翻訳の未来

燃脇 綾子
レッドハット株式会社

1. はじめに

本稿では、米国に本社を置く Red Hat で、カスタマーエクスペリエンスの観点から筆者が取り組んできた翻訳状況の改善活動と、それに伴い弊社が採用してきた機械翻訳の活用方法を紹介する。本稿は、弊社非公式ブログ「赤帽エンジニアブログ」で公開した[1]「RHEL8 ドキュメント翻訳の裏側「GA 公開」プロジェクト」および[2]「機械翻訳を活用して RHEL ドキュメントの定期更新を実現するまでの裏話」、ならびに 2020 月 8 月 20 日に Women in Localization Japan で発表した[3]「Red Hat が目指す翻訳の未来 - 機械翻訳との共存」に加筆修正をしたものである。

2. Red Hat の紹介

Red Hat は、クラウドを中心にした技術サービスを提供し、Linux ディストリビューションの Red Hat Enterprise Linux (RHEL)を製品として開発/販売/サポートしている。また、オープンソースソフトウェアを利用したビジネスを展開し、ソフトウェアのアップデート/アップグレード/保守サポートなどを一体化したサブスクリプションを販売する事業モデルを採用している。

オープンソースソフトウェアは、オープンソースコミュニティーと呼ばれる団体のもとでソースコードを一般公開している。この団体は有志により組織され、その多くはオープンソースソフトウェアの利用者や開発者により構成され、非営利目的で運営されている。このように利用者の目的を問わずソースコードの使用/調査/再利用/修正/拡張/再配布を許可することでソフトウェアの開発が発展し、オープンソースソフトウェアの文化が作られていった。それは弊社の企業文化にも強く影響している。オープンソースで作られたカルチャーから生まれた理念やバリューを総称した Red Hat カルチャーという存在を、社員は日々意識して働いている。翻訳に関することで、このカルチャーを体現した出来事を 1 つあげると、2015 年に当時 Red Hat の CEO であった Jim Whitehurst が上記カルチャーを説明する書籍[4]「The Open Organization」を出版したのだが、その際、日本支社では、より多くの（特に新入）社員にそのカルチャーが浸透するように、数十名の有志により翻訳プロジェクトグループが結成された。このグループが目指したのはあくまで社員向けの日本語版の作成だったが、その後、外部の書籍翻訳者の力を借りて 2016 年に一般発売されるまでになった。筆者もこのプロジェクトに参加して翻訳/出版/社内セミナー開催を手伝ったが、この過程でオープンソースカルチャー（そして Red Hat カルチャー）の力を直に感じることができた。

3. Red Hat におけるローカライズ事情

Red Hat には、ローカライズを担当するチームが 2 つある。そのうち、技術用 Web サイトであるカスタマーポータルの翻訳を担当しているのが、筆者が所属するカスタマーエクスペリエンス＆エンゲージメント(CEE)のローカライゼーションチーム(l10n)である。ローカライゼーションチームは 2 つともオーストラリアを拠点とし、日本語や他言語の翻訳者を含めたほぼすべてのメンバーがオーストラリア支社に籍を置いている（おそら

Red Hat's Vision for the Future of Translation
Ayako Moewaki
Red Hat K,K.

く各言語のネイティブスピーカーを一か所に集めやすいからだろう)。日本には筆者他数名しかいない。筆者が所属する CEE l10n では製品マニュアルが主なサービスとなるが、その他にも、GUI、セキュリティー情報などを手掛けており、現在その多くは日本語である。翻訳およびポストエディット作業は主に社内翻訳者が行う。以前はより多くの言語を扱っていたが、近年は日本語を初めとした顧客の要望が高い言語に限定し、その言語のサービス改善に力を入れてきた。

筆者が弊社に入社した経緯は、他の翻訳者とは大きく異なる。筆者は、l10n チームではなく、技術サポートを提供するグローバルサポートサービスに、日本支社初の翻訳者として入社した。カスタマーポータルで利用できるナレッジベース（主にサポートエンジニアが顧客への技術サポート提供時に扱った事象を社員/顧客向けに公開/共有するページ）の翻訳を担当した。その頃も l10n チームが存在したにも関わらず、日本側は、自分達に必要なものが翻訳してもらえない、または翻訳してもらえたとしても時間がかかるというような不満を常に抱えており、その結果としてグローバルサポートサービスが独自に予算を取り、サポートエンジニアの代わりに翻訳者を 1 人採用することとなったためである。

筆者は、l10n チームに異動するまで、グローバルサポートサービス（その後、名称が CEE に変更となり、筆者が現在所属する l10n チームも同組織に編入した）で 5 年を過ごした。その名が示すとおり、本組織では、カスタマーエクスペリエンスを非常に重要なものと位置付けているが、筆者は、幸運にも日々顧客と接する技術サポートチームに長期間所属したおかげで、弊社が目指すカスタマーエクスペリエンスの形を十分すぎるほど叩き込まれた。これが CEE l10n への異動後の改善活動にも大きく役立っている。

4. これまでの改善活動

前項で示したとおり、入社当時、（少なくとも）日本支社における翻訳に対する期待値はそれほど高くなかった。日本語の資料が圧倒的に少なかったからだ。たとえば社員は日本語に翻訳した資料を顧客に求められるが、l10n チームに依頼してもその多くは翻訳してもらうことができない。したがって、たとえ語学が堪能でなくても、技術サポートや営業部の者たちは自分たちで翻訳せざるをえなかった。元の原稿（英語）を作成する担当者（主にアメリカ）も、翻訳をする部署が存在することを意識せず、独自に予算を取って外部に委託したり、機械翻訳で翻訳したものをそのまま使用したりして問題になったことも何度もあった。したがって、筆者が最初にしたことは、翻訳/翻訳者の重要さを各部署/各人に示すことであった。つまり、最初に自分で依頼を受け、その必要性が示されると l10n チームに依頼するように依頼者に促すことで自分の作業量を調整するとともに、l10n チームの存在が周知されるようにした。また、機会があれば、日本支社の社員や CEE の幹部に翻訳の重要性を訴え続けた。技術文書以外のものでも求められれば手を差し伸べ、その活動が別の l10n チーム（現在オーストラリアを拠点としている 2 つ目のローカライゼーションチーム）の結成にもつながった。現在は、当時に比べて環境がそろってきているのに加え、筆者自身も l10n チームに所属しているため、そういう草の根的な活動は必要なくなったが、代わりに、冒頭で紹介したように、弊社顧客向けにプレゼンテーションやブログ公開などを積極的に行っている。

また、CEE l10n に異動した後は、同チームの主力サービスである製品マニュアルの翻訳状況の改善に力を入れてきた。元々、製品の一般公開日(GA)に製品マニュアルの翻訳が終わっていることはなかった。GA 後に顧客がすぐに必要とするリリースノートなどの数冊は翻訳されることがあったが、多くは GA 後に翻訳されていた。また、各マニュアルの使用状況や、営業戦略により、翻訳されるマニュアルが限定されていた。さらに、英語の更新は頻繁に行われるが、それに合わせて日本語が更新されるとは限らなかった。弊社の主力製品で、他の製品と比べて利用者が桁違いである Red Hat Enterprise Linux (RHEL)でも状況はあまり変わらなかった。もちろん、コストなどの問題でそう

せざるを得ない状況ではあったし、その事情は日本も理解していた。しかし、それでは前述した不満はいつまでたっても解消されない。

したがって、まずは RHEL の翻訳プロジェクトコーディネーターになり、翻訳者という立場のまま、プログラムマネージャーに協力して RHEL の翻訳状況の改善に力を入れた。力を入れれば当然成果が出るため、それを社内外に広め、広めることで集まった声を海外にいる上司達に届け、その改善から得られる価値を会社に示し続けた。この過程で、もちろん避けて通れないのが機械翻訳である。こうして、機械翻訳の導入も積極的に行ってきた。

5. 機械翻訳の活用

機械翻訳はスピードには定評があるが、その導入にあたって最も大きな壁となるのは常に品質である。顧客は「問題のある」翻訳を許容できるのだろうか。また、許容できるとして、どこまで許容できるのだろうか。筆者が担当した RHEL を例にあげて詳しく説明しよう。

RHEL は、日本時間 2019 年 5 月 7 日にメジャーバージョンである RHEL 8 がリリースされたが、弊社は、その一般公開日(GA)に、日本語の製品マニュアルを 20 冊以上公開した。これまでのメジャーバージョンと異なり、原文である英語の製品マニュアル自体も「一から書き直し」が計画されていて、日本語だけでなく元となる英語も GA に間に合うか分からないと言われていた中での快挙であった。当時はそのようには意識していなかったが、GA 直前に英語の更新が続いたため時間が足りなくなり、最終的にはライトポストエディットで公開した(GA 後に品質をあげて再公開)。その 2 か月後、残りの製品マニュアルを機械翻訳で公開して、当初の目標だった全冊公開をはたした。この 2 つの実績は社内で周知され、日本支社の社員全員を対象としたプレゼンテーションを頼まれ、最終的には表彰もされた。そしてこの一連の出来事は、多くの日本社員にとってまだまだ遠い存在であった CEE l10n の実績が広く知られるようになるきっかけとなった。

それから遡ること 2018 年 8 月、本製品の翻訳プロジェクトコーディネーターに任命された時に、筆者は古巣である技術サポートチームに、RHEL 7 までの翻訳状況とそれに対する要望についてヒアリングを行った。そして直近の目標を「RHEL 8 の GA 公開」とし、前述した理由による全冊公開の難しさから、GA に公開するマニュアルと、その後機械翻訳で公開するマニュアルにわけ、まずは対象マニュアルの GA 公開を目指したのである。詳細は[1]のブログを参照してほしい。

このヒアリングの時に、技術サポートチームから出た一番シンプルな要望は「GA に全冊公開」であった。顧客が実際に読むかどうかでなく、日本語があってしかるべきなのだ。技術サポートチームで彼らの大変さを目の当たりにしていた筆者にとってもごく当たり前な要望であったが、先に述べたとおり、l10n チームの今までの戦略とは大きく異なっていた。その後、機械翻訳のみを使った公開の可能性についてもヒアリングを行った。この時期の機械翻訳の品質はまだまだ十分とは言えず、特にタグが文章を分断するためきちんとした「文章」になっていないことも多かった。たとえば「インストールする xxx パッケージをブート時に。」というような表現である。これは、翻訳を提供する側からすれば、許容できないレベルである。しかし、技術サポートチームの意見は分かれた。必要な単語がすべて訳されているため、文章として成り立っていなくても、意味が分かるから問題ない、英語しかないよりはましという人も少なくなかった。一方、彼らは、自分たちのビジネスに関する文言には敏感である。たとえば「xx is <tag>not</tag> supported」という文章で否定形がきちんと処理されずに「xx はサポートされる」と訳されることがあったが、この文章は顧客との会話で重要になってくるため、この間違いは許容できないと強く主張した。

以上をもとに、最終的に選んだのは「最低限の修正だけ行い（便宜上ウルトラライトポストエディットと名付ける）最短期間で公開」することで全冊公開を実現することだった。これにより人手翻訳では数か月かかる量を数日で公開することが可能となった。初めて

の試みで、しかも対象となるのは主力製品で待望のメジャーバージョンである。公開にあたっては多少の不安もあったが、想像していた以上に温かい声が、想像していた以上に様々な人から届いた。

次の目標となり得るのは、機械翻訳を他製品に導入して公開する範囲を広げることだったが、この時点では他製品に手を伸ばすのは総合的に困難であった。社内の意見は常にシンプルである。日本語がなくて困った経験のある人や、現在も日常的に日本語の不足を実感する人は、機械翻訳で公開することにほとんど躊躇いがない。一方で、日本語が充実している製品を担当している人、製品マニュアルを実際に読まない人、または読んだとしても日本語版を読まない人は多かれ少なかれ懸念を示す。それは CEE l10n 内でも同様である。さらに CEE l10n は、立場上コストと品質のバランスなど、その他の問題も考慮していかなければならない。明らかに、機は熟していなかった。したがって、機械翻訳を他製品に導入する代わりに、来る日のために RHEL の状況を改善し続けることにした。具体的な活動は[2]のブログを参照してほしいが、効率的に作業し、更新頻度をあげ、それを周知することで、これまで RHEL が抱えていた問題を解消していった。

6. 本格的なライトポストエディットの導入

2019 年 7 月に RHEL の製品マニュアルを機械翻訳で公開してから年度がかわり、これまでの経験をもとに CEE l10n チームはとうとう全製品の全冊公開を目指すことになった。その先駆けとして、2020 年 3 月に、RHEL の関連製品である Red Hat Insights を、ライトポストエディットのみで新たに公開した。RHEL 8 の（ウルトラ）ライトポストエディットのときと異なり、スクリプト等を多用して、翻訳者が作業しやすい環境を作ることで効率化を目指し、適切なライトポストエディット時間を確保した。これにより、通常よりも短い時間でも品質を飛躍的に向上させた。ほぼ同時期、RHEL 8 にもライトポストエディットが本格的に導入された。

この時、筆者に課せられた課題は、RHEL に影響を与えずに、筆者が対応する製品を増やすことだった。このため、2020 年 3 月に、弊社戦略製品である Red Hat Ansible のプロジェクトコーディネーターを兼任することになった。Ansible には、Ansible Engine と Ansible Tower の 2 つの製品がある。前者はオープンソースコミュニティー製品であるが、後者は前者をベースとして機能を拡張した弊社製品である。したがって弊社のサポート対象（および翻訳対象）となるのは Tower のみとなる。しかし、ユーザーは最初にコミュニティー製品の Engine を使い、その後 Tower を導入する。そのため、コミュニティー製品であるにも関わらず 1 年以上前から Engine の翻訳を求める声が何度もあがっていたが、コストの問題で中々取り掛かることができないでいた。これが、RHEL のコスト削減によりとうとう実現したのである。

この時期、筆者は 2 つのことを試した。1 つは、RHEL のライトポストエディットを、翻訳ベンダーに発注することである。先に述べたように、CEE l10n チームではこれまで、主に社内翻訳者が作業し、残りを翻訳ベンダーに依頼していたが、当時ベンダーに依頼していたのはフルポストエディット＋バイリンガルチェックだった。その作業内容をライトポストエディットに変えてベンダーへ依頼し、社内ではほとんど作業せずに公開するという試みである。これが成功すればより大きなコスト削減へつながる。RHEL は、他の製品と比べて学習データがそろってきていて機械翻訳の品質も向上していたが、当時まだ更新頻度も高く、その更新量も、ユーザー数も他の製品と比べて桁違いに多いため、顧客の反応を知るのに最適な製品だった。協力を依頼した翻訳ベンダーとは事前に打ち合わせを何度か重ねたが、実際の作業はスムーズに進み、CEE l10n にその可能性を示すことができた。

7. 人手翻訳や機械翻訳に対する顧客の期待

RHEL における上記の取り組みと同時に力を入れたのが Ansible チームとの協働である。成熟した

RHEL とは異なり、製品としてもまだ若い Ansible は弊社の製品の中でもオープンソースコミュニティー色が強く、それを支える弊社社員は、技術、営業などの部署間の垣根を超えてユーザー向けに様々な活動を行っている。製品理解のためもあり、筆者も翻訳を進めながらその活動に参加していたのだが、それがなぜか数日に渡る Ansible イベントの[5]パネルディスカッションへの登壇につながった（これこそが Red Hat カルチャーである）。

このときオーガナイザーに頼んだのが、翻訳に関するフィードバックを何らかの形で参加者から集めることだったが、今年はコロナの影響でオンラインイベントとなったため、その場で質問に答えてもらう形で簡単に実現した（図 1）。

図 1：質問「日本語ドキュメントを利用しているか」

非常に興味深いのは、日本語のドキュメントに対する顧客の反応は必ずしも良いとは言えないという点である。まず、先に述べたように、日本語への翻訳状況はまだまだ改善途中であり、製品マニュアルがすべて翻訳されておらず、翻訳されていても最新であるとは限らなかったこともある（図 2）ため、「古い」または「翻訳されていることを知らなかった」という声も多かった。

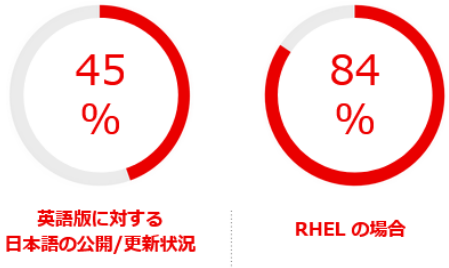


図 2：日本語製品マニュアルの提供状況
(2019/09 ~ 2020/04)

また、その当時は機械翻訳がほとんど使用されていなかったのにもかかわらず、品質に対する評価も必ずしも高いとは言えなかった。同様の意見は、弊社の顧客向けの定期アンケートでも見られた。ただし、これは翻訳だけの問題ではない。そもそも英語のマニュアルに問題がある可能性もあるし、技術的に理解が難しいことを顧客が英語のマニュアルまたは翻訳の問題ととらえる可能性もあるからだ。もしくは単に弊社社員とのコミュニケーションになんらかの問題があって弊社自体に否定的な印象を持っているだけなのかもしれない。非常に参考になる意見ではあるが、顧客の不満の原因を特定するのはそれほど簡単ではないため、1 つ 1 つの意見に一喜一憂するのは得策ではない。しかし、全体的な印象としては、翻訳に携わる者が述べる「人間の翻訳は完璧で、機械翻訳になると誤訳の問題で企業の信用が落ちる」というような意見が必ずしも正しいとは言えないということが分かる。そして、なによりも、日本語が提供されているかどうかが重要となる。表 1 は、日本の顧客からあがってくるフィードバックを、その重要度に応じてまとめたものである。

	フィードバックの種類	重要度
(1)	**英語しかない!!!!** Red Hat Enterprise Linux (RHEL) などの一部の製品は翻訳されているが、多くの製品が翻訳されていない。	高
(2)	**日本語はあるけど、古くてほしい情報が見つからない!!** 「Release early, Release often」(Red Hat カルチャーの 1 つ) により、翻訳しても英語版がすぐに更新される。	
(3)	**日本語は最新だけど、誤訳等でほしい情報が得られない!** 日本語が間違っているため操作に失敗する等でサポートチーム等がお叱りを受けることもある。	
(4)	**ほしい情報は得られたけど、日本語に間違いがある or 読みにくい...。** 特に、パートナー企業や、カスタマー企業 (の業種) によっては、一次ソースである Red Hat 文書の修正が必要となる。	低

表 1：日本の顧客によくあるフィードバック

翻訳に携わる者にとって重要なのは主に下の 2 点であろう。実際、CEE l10n チームの戦略でもこれまでは下の 2 点が重視されてきた。しかし、顧客が一番求めているのは上の 2 点である。それは、これまで述べてきたことから考えても明らかである。

1 つ面白い例がある。筆者が機械翻訳に関する意見を様々な部署から集めていた頃の話である。CEE l10n と技術サポートチームのマネージャーがそれぞれ同じ例を出し、それに対して逆の解釈を示したのである（図 3）。

この解釈の違いは、顧客への責任がどこにあるかが両部署で異なっていることに関係してくるのかもしれないが、技術サポートチームのマネージャーが一番重視していたのは、製品マニュアルが日本語であるということだった。

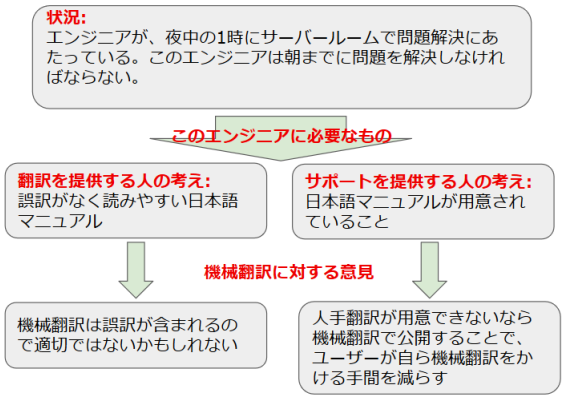


図 3：機械翻訳に対するニーズ

同マネージャーは「実際に時間に追われて作業したことがない人が機械翻訳に反対する」と熱く語っていた。確かに、翻訳に携わる者としては、CEE l10n マネージャーが話していたように、品質を軽視することはできない。しかし、顧客にとって重要なのはまず日本語が存在することなのだ。

このようなヒアリングを続けた結果、今夏 CEE l10n では機械翻訳とライトポストエディットを大々的に導入して、製品マニュアルの全製品全冊公開に踏み切ることとなった。それに先駆けて、筆者は日本支社に向けて再度プレゼンテーションを行った。人手翻訳（またはフルポストエディット）の代わりに（ウルトラ）ライトポストエディットを採用することにより、品質は下がるが、その代わりに機械翻訳を最大限活用して全製品の全冊公開を実現していくことに触れたが、その反応は非常に良いものであった。後日、全製品の全冊公開プロジェクトが始まったことを言及したときも、コンサルティング部と営業部のマネージャーがともに、「潜在顧客によっては、日本語になっていない製品は採用の候補にもならない」と現状を語り、このプロジェクトを支持する声をあげた。

8. Red Hat が選んだ機械翻訳との共存の道

CEE l10n チームは、現在、変革の真っ只中である。本稿を執筆しているのは 9 月なのだが、その約 2 か月前からライトポストエディットが本格的に導入され、これまで未翻訳だった製品が次々に翻訳公開され、それでもリソースが足りない製品はとりあえず機械翻訳のみでの公開となった。技術サポートチームに育ててもらった筆者が常々「当たり前」と思っていたことがこれでやっと実現する。

しかし、これで終わりではない。まず、これまでと比べて飛躍的に増えるメンテナンスコストに対応しなくてはいけない。

この課題に対応するため、CEE l10n では、図 4 を、これまで未翻訳だった製品に対する次のステップとした。

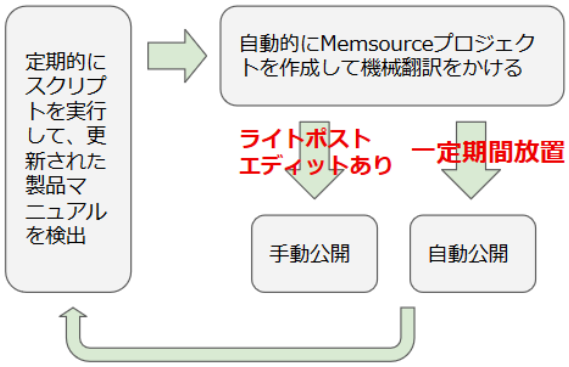


図 4：未翻訳製品の更新公開ワークフロー

これを実現するには技術的そして会社からのバックアップが必要であり、CEE l10n チームだけで実現することは難しく、現在は完全オートメーションではなくセミオートメーションでの対応となるが、近い将来、完全オートメーションに移行することが期待される。

では、少し先の未来として、CEE l10n の業務はどのように変わっていくと予想されるのか。今回の全冊公開と図 4 のワークフローにより、表 1 の(1)と(2)は解消される。したがって、表 1 の(3)や(4)がクローズアップされるかもしれない。しかも、そこで採用されているのは、人手翻訳ではなく、機械翻訳やライトポストエディットである。現時点で、図 4 はこれまで未

翻訳であった製品に対する取り組みであるため、営業戦略としても実際の使用状況からみても、誤訳や表現の拙さが実際に大きな影響を及ぼすことはあまり想定されていないが、本当にそうであろうか。

ここで、先に述べた品質に対する顧客のニーズに話を戻そう。先に述べたとおり、顧客は常に高品質のものを求めているわけではないし、人手翻訳が常に評価が高いわけでもない。そして評価が低い原因を特定するのは想像以上に骨の折れる作業である。

去年の 5 月に初めてライトポストエディットを採用した RHEL 8 の GA から、機械翻訳やライトポストエディットの本格導入を決めるまでの間、CEE l10n マネージャーはその品質について慎重に検討を重ねてきた。マネージャーは日本語話者でもなく、日本在住でもないが、品質を含めた全体のパフォーマンスに責任を持つ立場であるため、その慎重さは当然である。筆者は翻訳者として、日本支社に籍を置く者として、そして顧客と直接話をする日本社員たちの率直な意見を簡単に得られる立場にいる者として、マネージャーが求める意見や情報をできるだけ客観的に集める努力をしたつもりだ。そして長い時間をかけて集めたものを検討してきた結果、CEE l10n マネージャーが出した答えは「機械翻訳 vs.人手翻訳における品質の戦いではない」。筆者も同意見である。これについても技術サポートチームから興味深いフィードバックをもらったことがある。RHEL 8 を初めて機械翻訳のみで公開した 2019 年 7 月前後、「機械翻訳の問題は指摘があれば修正する」と説明したとき、あるエンジニアが「機械翻訳だと思って報告したら実は人手翻訳だったらどうなるのか」と質問してきたのだ。これまでも人手翻訳を利用してきた彼らの正直な判断基準は、「正確/読みやすい」か「間違い/読みにくい」でしかなく、それは人手翻訳にも機械翻訳にも当てはまるのだ（もちろんその割合は多少異なるかもしれないが）。つまり、彼らの求める品質、または印象として受ける品質について、翻訳を提供する側が考えるほど、機械翻訳と人手翻訳に大きな違いはないのである。この現実を冷静に受け止め、顧客の求めるものを都度判断して柔軟に対応する姿勢が、今後翻訳に携わる者に求められていくはずだ。

一般的に、弊社のように、これまで人手翻訳に長く取り組んできた組織に新たに機械翻訳を導入するのは、技術的な問題だけでなく、機械翻訳に対する（翻訳を提供する側だけでなく受ける側の）各人の受け止め方の違いなどで、その一歩を踏み出すまでが最も大変だと思われるかもしれない。しかし、我々の戦いはまだ始まってもいない。全冊公開が完了して、新たなワークフローが確立されて、やっとスタート地点に立てるのだ。ここからが、ローカライズの専門チームである CEE l10n の腕の見せ所である。当事者であるが、これから弊社がどう変革していくのか楽しみだ。

9. さいごに：目の前に迫る現実

筆者が CEE l10n で全製品の全冊公開に向けて様々な準備に取り組んでいたとき、社内でちょっとした動きがあった。先に述べたとおり、弊社はオープンソースを利用し、無料で使用できるソフトウェアに付加価値を付ける形でビジネスを展開している。したがって、ユーザーが手に入れることができる資料は弊社が提供するものに留まらないが、ここでも言葉の壁が大きな問題となっている。数年前から、日本社員は、この状況を改善すべく、日本語での情報発信に力を入れてきた。その 1 つが、弊社 Web サイトや無数にあるコミュニティーサイトで公開されているブログを日本語で、公式/非公式の形で提供することである。本稿で紹介したブログも、その非公式サイトで公開されている。その活動が、DeepL の出現によって拍車がかかった。無料の DeepL のサイトではなく、正規のサブスクリプションを購入して、未翻訳の資料をどんどん翻訳/公開していくという議論がなされ、筆者もいくつか質問をうけた。また、筆者が社内プレゼンテーションで機械翻訳を最大限利用して全製品の全冊を公開することに触れた翌週、それを聞いていたエンジニアの 1 人が DeepL を使って製品マニュアルを翻訳する API を作り、社内チャットで紹介した。オープンソースカルチャーが浸透している弊社で

は、これは当然予想される出来事で、筆者が最悪のケースとしてCEE l10n に常に訴えてきたことだった。このように、翻訳サービスを提供する側が、提供する側と受ける側の両方が納得できるような仕組みを構築していくことに躊躇すれば、他の誰かが代わりにそれを実現していく。しかし、翻訳に従事する者にとっても、そして顧客にとっても、それが最善な方法とは限らない。したがって、このような未来を回避するためにも、互いが望む形を認めつつ、様々な立場の人達が共通の目的のもとに団結する仕組みを作り上げていく責任が、提供する側にある。機械翻訳との共存への流れはもう止まらない。直視しようが目を背けようがそれは変わらない。たびたび、機械翻訳が生成した誤訳を企業や公的機関がそのまま使用して話題になることがある。これは、翻訳(者)の価値を理解しない使用者の過ちと切り捨てていいものなのだろうか。単にそこまで気軽に使えてしまうほど機械翻訳が身近になっただけなのではないだろうか。改善点を見つけ、時には失敗をしつつも少しずつ前に進む。これはオープンソースカルチャーが浸透している弊社が最も得意とする分野である。では、他の企業/組織ではどうなのか。決して楽観視することはできない。だからこそ、翻訳サービスを提供する側が、提供する側と受ける側の両方が納得できるような仕組みを構築していくことに力を注ぐべきで、それは、クライアント企業の翻訳部署、翻訳ベンダー、フリーランス翻訳者…どの立場にいても変わらないはずだ。めざすのは、最終的な読者が十分な価値を得られるようにすることだ。

筆者やCEE l10n は、すでに製品マニュアル以外のものを次の候補として視野に入れ始めた。現時点では具体的な計画はないが、全冊公開と図 4 が軌道にのる頃には、テクノロジーの発展もあり、現在予測しているよりも効率的な成果が得られることが期待される。次に求められるのは何なのか。より広範なサービスなのか、翻訳者の枠を超えた深い専門性なのか、それともその両方なのか。翻訳に対するニーズは今後もなくならないかもしれない。しかし私達はいつまでも同じ場所に留まることはできない。せめて、その変化を今後も楽しんでいきたい。

参照

[1] 赤帽エンジニアブログ「RHEL8 ドキュメント翻訳の裏側「GA 公開」プロジェクト」
https://rheb.hatenablog.com/entry/RHEL8-tranlation-project

[2] 赤帽エンジニアブログ「機械翻訳を活用して RHEL ドキュメントの定期更新を実現するまでの裏話」
https://rheb.hatenablog.com/entry/RHEL-document-with-MachineTranslation

[3] Women in Localization Japan オンラインイベント（2020 年 8 月 20 日）「Red Hat が目指す翻訳の未来 - 機械翻訳との共存」

[4] The Open Organization
https://www.amazon.co.jp/gp/product/B00O92Q6CQ/ref=dbs_a_def_rwt_hsch_vapi_taft_p1_i0

[5] Red Hat Ansible Automates Tokyo 2020 に伴い開催されたコミュニティーイベント Ansible Night
https://ansible-users.connpass.com/event/176436/

解説記事

QRpotato: Web から専門用語対訳対を収集する

阿辺川 武
国立情報学研究所コンテンツ科学研究系

1. はじめに

翻訳において専門分野の文書を翻訳するときは、その分野で普段用いられている専門用語を選択する必要があり、専門用語辞書の使用が欠かせない。最新の多言語専門用語のリファレンスリソースに対するニーズは非常に高いが、一般的に人手で編集された専門用語辞書では専門用語の増加のスピードに追いつけないことが認識されている。このギャップを埋めるために、パラレルコーパスやコンパラブルコーパスから対訳用語を自動的に抽出する手法の開発に多くの努力が費やされてきた。コンパラブルコーパスの利用は、パラレルコーパスよりも幅広い言語とテキストタイプのコーパスを利用できるため、特に重要であると広く認識されている[1]。

しかし、翻訳者は自動化手法を用いてコーパスから構築された用語集をあまり利用しておらず、別のアプローチを使う傾向がある。手元の用語集、あるいはオンラインの辞書サイトに関連する項目が見つからない場合、多くの翻訳者は Web 上で用語の対訳対が共起していると仮定して翻訳候補を生成し、Web 検索エンジンを使用して、これらの対訳対を検証する。

そこで我々は、このような翻訳者の行動を参考にして、専門用語対訳対を収集するシステム QRpotato を開発した。その仕組みは最初に専門用語からなるシード対訳対を利用して、Web 上でその出現ページを検索し、同一ページ内でシード対訳対と同じパターンで出現する文字列対を対訳候補として網羅的に収集する手法である。本稿では手法の説明と、QRpotato を使用して実際に収集した専門用語対訳対の例を紹介する。

2. 専門用語対訳対の存在するページ

まず、Web 上で専門用語対訳対がまとめて存在する可能性が高いページの例を紹介する。

- 単語集リスト

Web 上には特定の概念の単語対の集合を編集したページが数多く存在する。その中にはある領域の専門用語をまとめたページがある一方で「中学 3 年で習得する単語リスト」といった一般用語に関してリストを作成している場合もあり、これらと区別する必要がある。

- 専門文書

専門文書（technical document）では、この文のように文章中で専門用語（technical term）に対し、カッコ表記で対訳を併記することがある。また、学術論文では分野や出版社に依るがアブストラクト部分やキーワードリストに対訳対が記述されている。

- 目次、索引

目次、索引でも同様に見出し語に対し、英語が併記されていることがある。特に学術系のコンテンツに多く、大学のシラバスにも同様の傾向がある。

3. 収集方法

本節では、専門用語対訳対を収集する手法を説明する。以下に収集の簡単な流れを述べる。

1. シード対訳対集合を用意する。
2. シード対訳対をクエリーとして、検索エンジンから対訳対を含むページを収集する。
3. 収集されたページごとにシード対訳対の出現パターンを求める。

QRpotato: A system that exhaustively collects bilingual technical term pairs from the Web
Takeshi Abekawa
National Institute of Informatics

表 1. 得られた文字列パターンの例

	左側終端記号	左側用語	中間文字列	右側用語	右側終端記号
1	‘、'	シェーグレン症候群	‘('	Sjogren syndrome	‘)'
2	‘>'	ラポール	“	rapport	‘と'
3	‘['	代謝	‘]('	metabolism	‘)'
4	‘●'	アンダーカット	‘【'	undercut	‘】'
5	‘>'	アンタゴニスト	‘'	antagonist	‘<'
6	‘>'	イベント	‘（<i>'	event	‘<'
7	‘>'	光	‘</td><td><font>'	light	‘<'

4. 得られたパターンを用いて、同様のパターンで出現する用語対を収集する。

3.1 専門用語対訳対をクエリーとした Web 検索

Web 検索エンジン[1]を用いて、シードとなる専門用語対訳対を含むページを検索する。その際、対訳対がその順番で隣接して出現するように、2 つの文字列をそのまま結合したフレーズで検索を行なう。そして、検索結果として得られた URL 集合から HTML および PDF ファイルをクロールする。

3.2 対訳対の抽出

- 中間文字列の取得

既存の検索エンジンの多くは、HTML タグを含め記号全般を除去した検索インデックスを構築しており、フレーズ検索を行なっても、間に存在している記号類はすべて無視した文字列として検索される。そこで、シード対訳対の間に含まれる中間文字列を取得するため、得られた HTML ソースコードを対象に、プログラムを用いて再度シード対訳対による検索を行なう。ここでは、先程のフレーズとしてではなく、対訳対の間に記号文字（空白を含む）からなる文字列の挿入を許した正規表現検索をする。これにより表 1 にあるような中間文字列が取得できる。

- 終端記号の検索

次に、左側の用語、右側の用語双方に対して、終端文字列を決定する。中間文字列内に括弧記号や HTML タグがある場合には、それらに対応する終端記号を決定する。例えば、中間文字列に‘('があったとき、右側の用語の終端記号は‘)'となり、‘]'がある場合は、左側の用語も終端記号は‘['となる。中間文字列に括弧がないときは、用語から終端方向へ向けて記号文字を検索し、最初の記号文字を終端記号とする。これは中間文字列に HTML のタグがあるときなどが当てはまる（表 1 の 5, 6, 7）。終端記号方向へ記号文字を探しているとき、別の文字種が出現した場合（表 1 では 1 の左側、2 の右側）は、ここで探索を打ち切り、文字種を用いた正規表現としてパターンを作成する。1 では「日本語文字列[KanjiKana+]」、2 では「アルファベットと数字からなる文字列[AlphabetNum+]」となる。このため文章中に用語が出現した場合など、専門用語の区切りが判断できないことがあり、この正規表現を使ったパターンの信頼度が低くなる。

- 同一パターンの検索

上記で得られたパターンについて、中間文字列と終端記号から正規表現を作成し、そのページ内で同一のパターンで出現する文字列対を検索する。これは 1 つのページ内では、専門用語対訳対が同一のパターンで出現する傾向が高いというヒューリスティックスを利用し

[1] 現在は Google Custom Search API を使用している。

ている。獲得された文字列対については、どのようなパターンを用いて検索されたかについて保存しておく。前節で述べたように端が終端記号で抑えられるものと、正規表現の境界であるものを区別するためである。

3.3 本手法の評価

3.2 の手法を評価するために、科学技術用語集 210,328 用語対[2]を用いて対訳対の収集をおこなった。その結果、3,486,125 対の新しい対訳対の候補が得られた。サンプルとして 300 対を調べたところ 216 対（72%）が正しい対訳対で、22 対（7%）が部分的にマッチした対訳対であった。詳しい結果は紙面の都合もあり[2]を参照されたい。専門用語対訳対でない対を多く含むページには次のようなものが挙げられる。

- 読みだけのページ

カタカナからなる日本語用語に対応する英語用語が、たまたまローマ字綴りの表記と同じとき（例、グアノ | guano）に、日本語の読み方を紹介しているページを獲得してしまうことがある。特に日本の文化を海外に紹介するページに多い。形態素解析器を用いて読みを取得しローマ字に変換したものと、得られた英語用語を比較すれば、これらのページを除去できると思われる。

- 他の言語のページ

今回用いたシード用語対は日英のものであるが、たまたま英語の用語が他の言語と同じスペルであったために検索でヒットしてしまったことが原因である。

対応策として、その言語に対応した一般辞書が用意できれば、新たに獲得した用語について、その言語特有の単語を含むかどうかで判断できると思われる。

4. QRpotato

これまで説明してきた対訳対の収集方法を実装したシステムが QRpotato である。この名称は、シード対訳対から対訳対を収集し、その対訳対を新たなシードとして…というように芋づる式に対訳対を収集できることから命名された。接頭辞 QR は我々の開発している一連の翻訳支援ツールに共通する名前である。

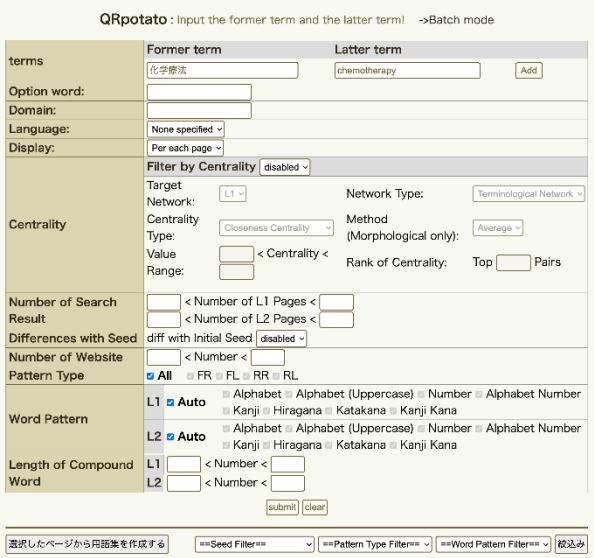


図 1. QRpotato のインターフェース

図 1 が現在開発している QRpotato の検索インターフェースである。ここではシードとなる対訳対の入力ボックス（terms）だけでなく、オプションワードを指定することができる。オプションワードという表記であるが、使い方として例えば、ある用語 t の対訳を知りたいとき、オプションワードに t を入力し、同一分野で既知の対訳対を term pair として入力する。そうすると対訳語を含み、かつ t を含むページが検索され、既知の対訳対と同様のパターンで t の対訳語が出現していることが期待できる。

図 2 はシード対に「化学療法｜chemotherapy」で検索した結果のうち正しく対訳対のリストが取得できた例である。対訳対パターンを見ると“;○(□)” となっており（○が日本語、□が英語を表す）、日本語の始め、英語の終わりが記号で区切られている。このようなパターンは正しく対訳対が取得できることが多い。QRpotato では、検索結果のうち、うまく取得できているページの対訳対リストを人手で選択し、それらのリストだけを CSV 形式としてダウンロードすることができる。

[2] JST 科学技術用語日英対訳辞書

学会雑誌抄録｜日本甲状腺学会

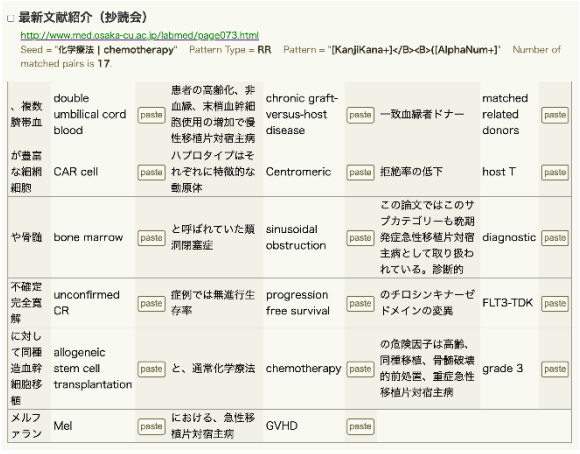
http://www.japanthyroid.jp/doctor/abstract/abstract08.html

Seed = "化学療法 | chemotherapy" Pattern Type = FL Pattern = "(_)" Number of matched pairs is 33.

芽細胞発癌	fetal cell carcinogenesis	paste	甲状腺幹細胞	thyroid stem cell	paste	甲状腺芽細胞	thyroblast	paste
リバースアプローチ	reverse approach	paste	幹細胞危機	stem cell crisis	paste	濾胞癌	follicular thyroid cancer	paste
分子診断	molecular diagnosis	paste	遺伝子発現プロファイル	gene expression profile	paste	遺伝子変異	gene mutation	paste
転座	translocation	paste	甲状腺分化癌	differentiated thyroid cancer	paste	予後因子	prognostic factor	paste
甲状腺未分化癌	anaplastic or undifferentiated thyroid cancer	paste	パクリタキセル	paclitaxel	paste	治療	treatment	paste
NCD	National Clinical Database	paste	甲状腺	thyroid	paste	副甲状腺	parathyroid	paste
副腎	adrenal gland	paste	内分泌	endocrine	paste	がん登録	cancer registry	paste
甲状腺癌	thyroid cancer	paste	一般社団法人 National Clinical Database	NCD	paste	甲状腺未分化癌	anaplastic thyroid cancer	paste
予後	prognosis	paste	臨床試験	clinical trial	paste	甲状腺分化癌	differential thyroid cancer	paste
アイソトープ内用療法	radionuclide targeting therapy	paste	ヨウ素	iodine	paste	Cowden 症候群	Cowden syndrome	paste
ポリポージス	polyposis	paste	甲状腺全摘	total thyroidectomy	paste	腺腫様甲状腺腫	adenomatous goiter	paste

図 2.「化学療法 | chemotherapy」の検索例 1

一方、図 3 は同じシード対でうまくいかなかったページ例である。実際のページ中での出現は“通常化学療法（chemotherapy、chemo）”であり、ここから対訳対パターンは“[KanjiKana+]([AlphaNum+]”が得られた。このパターンは用語境界が正規表現で定義され、正しい対訳対が得られない事が多い。また、このページではカッコの使用が対訳対だけでなく補足説明でも使われており、特定パターンの使用方法が統一されていないページではうまくいかないことが多い。



最新文献紹介（抄読会）

http://www.med.osaka-cu.ac.jp/labmed/page073.html

Seed = "化学療法 | chemotherapy" Pattern Type = RR Pattern = "[KanjiKana+]</B><B>([AlphaNum+]" Number of matched pairs is 17.

、複数臍帯血	double umbilical cord blood	paste	患者の高齢化、非血縁、末梢血幹細胞使用の増加で慢性移植片対宿主病	chronic graft-versus-host disease	paste	一致血縁者ドナー	matched related donors	paste
が豊富な細網細胞	CAR cell	paste	ハプロタイプはそれぞれに特徴的な働原体	Centromeric	paste	拒絶率の低下	host T	paste
や骨髄	bone marrow	paste	と呼ばれていた類洞閉塞症	sinusoidal obstruction	paste	この論文ではこのサブカテゴリーも移植発症急性移植片対宿主病として取り扱われている。診断的	diagnostic	paste
不確定完全寛解	unconfirmed CR	paste	症例では無進行生存率	progression free survival	paste	のチロシンキナーゼドメインの変異	FLT3-TDK	paste
に対して同種造血幹細胞移植	allogeneic stem cell transplantation	paste	と、通常化学療法	chemotherapy	paste	の危険因子は高齢、同種移植、骨髄破壊的前処置、重症急性移植片対宿主病	grade 3	paste
メルファラン	Mel	paste	における、急性移植片対宿主病	GVHD	paste			

図 3.「化学療法 | chemotherapy」の検索例 2

5. おわりに

本稿では、Web 上から専門用語対訳対を収集するために、シード対訳対を含むページを獲得し、ページ中でシード対訳対と同じパターンで出現する文字列対を新たな対訳対候補として収集する手法を紹介した。獲得した対訳対リストの精度は抽出に使用したパターンにより異なるため、パターンを限定する必要がある。

今後の分析として、シードとして使用した用語と実際に収集した用語の関係や、異なるシードを使用して収集した用語間の関係について検討する必要がある。また予測の観点からも、新しい用語対を収集するために有効なシードの特徴を明らかにできれば、非常に有用である。

最後に、QRpotato の仕組みは対訳対の収集に活用できるだけでなく、任意の文字列の組み合わせについても動作する。例えば“United Nation | UN”のような国際組織名とその略称、「坊っちゃん | 夏目漱石」で小説作品とその作家名、「栄枯盛衰 | えいこせいすい」として四字熟語を収集するなどである（これらは実際うまくいく）。また、「菅義偉 | 70」のように政治家、芸能人、スポーツ選手などの名前と歳を対とすれば、その人物がその年齢の年度に公開された記事のみがクロールされ、同時代に同分野で活躍する人々がリストとして作成できる。

参考文献

[1] Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (Eds.). (2013) Building and using comparable corpora. Berlin: Springer.

[2] Abekawa, T. and Kageura, K. (2011) Using seed terms for crawling bilingual terminology lists on the Web. Translating and the Computer 33.

解説記事

人間の翻訳と機械の翻訳（３）：翻訳プロセスは何から構成されるか？

影浦 峡
東京大学　大学院教育学研究科

1. これまでのまとめと今回の話題

本連載では、第１回の記事で、そもそも翻訳は何を対象としているのかという問いに「文やテキストではなく文書」という答えを与えました。文書を代表する「本」というのは不思議なもので、翻訳することもできれば、燃やすこともできます。それは単に「本」の概念が二重性を持つからだというのはその通りではあるのですがとはいえそれは言語やテキストを抽象体として捉えた上で本が担うそこに収まらないいわば物理的な領域を立てるという視点に立つからで、文書という社会的存在を翻訳の対象として同定したときに述べたことはまさに翻訳はそのような視点に基づく二重性を基本とした行為ではないかもしれないことを示唆しています。

それを踏まえて前回（第２回）は文書とはどのようなものかについて、位置付け・性格・属性・構成要素の観点から簡単に整理しました。それを踏まえて、いわゆる言語データと文書の間にあるギャップについても言及しました。

今回は、ではその文書を対象とするという翻訳のプロセスはどのような要素から構成されているのかを見ることにします。実際のところ翻訳においてどのようなプロセスが取られるかは個別的で多様であり、また、詳細な記述的明確化も進んでいません。その詳細な記述は私たちが進めている研究プロジェクト「翻訳規範とコンピテンスの可操作化を通した翻訳プロセス・モデルと統合環境の構築」（科学研究費基盤 S）[1]の中心的課題の一つで、現在、研究進行中ですので、ここで述べることは必然的に概略的なことになります。連載の第１回で、

> 例えば翻訳の先端で活動している高橋さきの氏のようなプロフェッショナルの眼には、言語処理が語っている翻訳は、「人間の脚や腕を、まるで一本の枝でも描くように」描いているようなものと映るのかもしれません。

と述べました。研究としては、きちんと描くことができることときちんと描くことがどのようなプロセスから構成されているかを体系的かつ明確・詳細に把握し言語化できることのギャップを埋めることが大きな課題となります。

2. 翻訳プロセス

ここでは、まず、翻訳プロセスとは何か、そしてそれはどのようなものか、を簡単に整理します。

最近、翻訳研究の世界ではいわゆる「翻訳プロセス研究」が活発に行われています。これは、アイトラッカーなどを用いて翻訳者が翻訳時に何をしているかを追うもので、実験的・実証的な方法により翻訳者の認知的なプロセスと対応する反応を明らかにすることに重点が置かれています[2]。多少単純化するとここでの「翻訳プロセス」は「翻訳者が翻訳しているときの認知プロセスおよびそれに対応した行動プロセス」です。

これとは全く異なる用法があります。翻訳産業界の用法で、クライアントとの交渉・契約、要項の作成、プロジェクトの構成と管理、調査、翻訳・修正・レビュー・校正、検証、納品からなる工程を指します。例えば、翻訳サービスの要件を規定した ISO 17100 には、次のように書かれています。

Of what does “translation process” consist?
Kyo Kageura
Graduate School of Education, The University of Tokyo

[ISO 17100] specifies requirements for all aspects of the translation process directly affecting the quality and delivery of translation services.[3]

ここでは「翻訳プロセス（translation process）」という言葉が翻訳者個々人の認知プロセスではなく、発注から納品までの流れを指すものとして使われています。

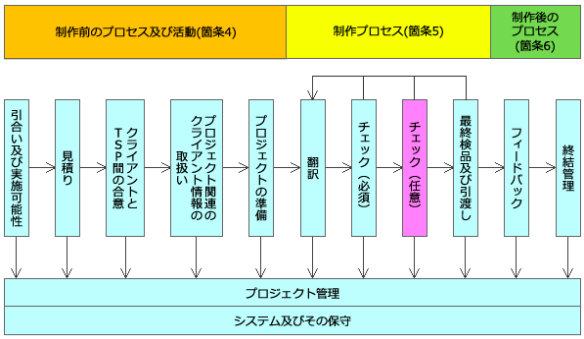
本稿（および本連載）では、「翻訳プロセス」を、後者の意味を基本とし、それを、いわゆる狭義の、翻訳者個々人が「翻訳」（修正・レビュー・校正等）をするときに行なっている行為にまで適用して考えます。

翻訳プロセスは、原則として、共有可能な外在的対象に対する外在的な操作として具体的に存在します。翻訳は知的営為であるからとりわけ翻訳者個々人の翻訳プロセスは本来的に内的・認知的なものであるという誤解もときに見られますが、翻訳時の認知プロセスは翻訳プロセスに対応した認知プロセスだから翻訳の認知プロセスなのであって、それは読字障害を有する人たちの脳の特定部位のパルスが揺れることが観察されたときにその揺れが読字障害なのではなく読字障害が既に存在し認識されているからこそそもそもその揺れが読字障害に対応していると認識することができそのようなものとして同定することができそう語ることができるのと同じであり、それゆえ翻訳時の認知プロセスに対する認識の解像度は翻訳プロセスそのものに対する認識の解像度に論理的に依存し、後者は前者に先行しまた前者を束縛します。例えば線形代数はそうと名指すことができない原初における生成においては個々人の認知的営為を第一義的なものと見なすことができるかもしれませんが（もちろん事実問題としてある一人にその生成を帰することはでませんが）そのプロセスは線形代数の生成プロセスとして捉えることはできずそう捉えることが可能であるためには対応する記号の体系を有し外在化された知識として線形代数が既に存在していることが求められます[4]。余談になりますが、これは本連載の第１回で紹介した、名古屋大学の佐藤理史先生による「AI の 1 つの本質というのは、今、擬人化してしか説明できないことを、擬人化せずに説明する方法なんじゃないかと思う」という発言にも繋がるポイントで[5]、end-to-end の学習によって言語操作のプロセス自体がブラックボックス化される前の言語処理研究が共有していた課題でもあると考えられます。

というわけで、本稿では、翻訳プロセスを、翻訳ニーズに始まり納品とそれに対するフィードバックで一応終わりになるプロセスであり、起点言語文書を目標言語文書に変換することに関与する様々な要素に対する具体的な操作から構成されるとします。

3. 巨視的な翻訳プロセス

巨視的な翻訳プロセス、すなわちクライアントのニーズに始まり翻訳会社がプロジェクトを組んで目標言語文書を生成し納品するまでのプロセスを大まかに見たときの標準的な構成として ISO (2015)が示しているものは下図のようになっています[6]。ISO 17100 はガイダンスではなく認証規格として、プロの翻訳サービスが満たすべき条件をまとめたものなので（すべての翻訳案件に適用されるわけではないにせよ）この図は翻訳プロセスにおいてなされなくてはならない標準的な要素を示していると考えることができます。



翻訳プロセス全体は、「制作前のプロセスおよび活動」、「制作プロセス」、「制作後のプロセス」の３つに分かれます。いわゆる「翻訳」は、11 ステップに分かれるサブプロセスの一つとして位置付けられていること（あるいは一つにすぎないこと）、全体の翻訳プロセスが制作前のプロセス及び活動の、「引合い及び実施可能性」というクライアントのニーズと条件を考慮したサブプロセスから始まっていること、制作後のプロセ

スとしてフィードバックがあり、その後にプロセスが終了することなどは、実際に産業翻訳のマネジメントを担当している人にとっても独立して産業翻訳の仕事を受注する個人翻訳者にとっても当たり前のことでしょうが、出版翻訳を副業とする個人翻訳者やMT研究開発従事者には見えにくい部分かもしれません[7]。

次に、それぞれのサブプロセスはどのようなものを対象にしたどのような行為からなっているのかが2節後半の主張との関係で問題になります。実のところ、ISO 17100の本文は、この図（及びそれを少し詳細化した図）に対応するサブプロセスが満たすべき要件を記述しているのですが、残念ながらそこでとられるべき行動が具体的に記述されているわけではありません（ここでの具体性のレベルとしては料理の本を想定するとよいと思います。それに従えば一応誰もがその料理を作ることができるような記述です[8]）。ここから先は、実務翻訳を意識した翻訳研究でも実用的な文書でも、体系的に整理されたものは見当たりません。

科学が対象を how の観点から捉えそれを具体的に共有可能な記号表現・操作の体系で表現するものであるとするならば、翻訳に対する私たちの理解はまだ17世紀以降の近代科学の意味での科学に到達していないと言うことができそうです[9]。なお、このプロセスの中で比較的具体化が進んでいるのは制作プロセス後半のチェックから制作後プロセスのフィードバックまでに関与する翻訳の品質管理で、これについては例えばMQM[10]が翻訳業界で標準的に使われているだけでなくMTを含む比較評価の試みでも用いられており[11]、外在的対象をめぐるそれなりに具体的な記述の有用性を示すものとなっています。一方、前半の契約等のプロセスにおいては、想定読者や目的等を含め、翻訳における重要な要件が検討されて定められ、それはプロジェクトの要項として制作プロセスをコントロールするのですが、この部分は事後的な品質管理と比べると記述の具体性が低く、一定程度の知識と専門性を前提とした人間の翻訳者間の理解に依存しています[12]。

4. 微視的な翻訳プロセス

ISOの図が示す通り、「翻訳」は、この翻訳プロセスの一サブプロセスと位置付けられます。一般に、ここがいわゆる「翻訳」の専門性が最も発揮されるところです。このサブプロセスがどのようなものかについて例に基づき考えてみます[13]。

> However, the set of all points on any straight line in R^2 not going through the origin (0,0) is not a vector subspace.[14]

当然のことですが、唐突にこの文だけが与えられて日本語に訳すよう求められた場合、プロの翻訳者は「訳せません」というか、訳すために必要な条件を明確にするやりとりを行うでしょう。内在的な条件としては、出典は何でどの範囲の翻訳が求められているか、どのような読者を想定するか、公表の媒体は何か[15]、などで、外在的な条件として期限や予算・単価などがあります。便宜的に図書全体の訳で想定読者は英語文書と同様、統計学や関連領域の学部上級から大学院初年次程度、教科書としての利用を想定します。

いわゆる「文意を取る」こととは別に直接の微視的文脈とは独立に専門概念の体系に対応した専門語彙の体系を想定して用いられている専門用語と記号について日本語で対応する体系を想定して対応する用語を確定する必要があります[16]。 名詞系では“set,” “points,” “(straight) line,” “origin,” “vector subspace” が、量化子として“all,” “any”が、記号として“R^2”がその対象となるでしょう。その際、標準的な数学辞典の表記を採用するか、あるいは予定されている出版社が学部向けに出している標準的な日本語教科書とのつなぎを考え、その教科書の用語を基本とすることもあるかもしれません。専門用語は表現の厳密性と唯一性を目指しはするものの、例えば “vector subspace”に対して基本的に「部分空間」「部分ベクトル空間」「ベクトル部分空間」あるいは「線形部分空間」のいずれをあてるかといった判断は必要で、いずれにせよこうした点をめぐる決断は具体的な文書が与えられ、翻訳目的等が絞り込まれた上でなされます。

この文における“vector subspace”に対応する日本語の選択は起点言語文書における異形がどのようになっているかによっても変わり得ます。この概念を指すときに単体の“subspace”が用いられることが決してないなら“vector subspace”であることはテキスト全体にわたる著者の意思的な選択であることが推測されるのでそれを考慮するかどうかの判断が必要になります。“vector subspace”と“subspace”が混在しているならば、まさにこの文で“subspace”ではなく“vector subspace”が用いられていることがどのような作用を持つかを把握し例えば少し重く感じられても「部分ベクトル空間」を当てる必要が出てくるかもしれません[17]。

以上の準備を踏まえ、文を日本語にしなくてはなりません。このとき、例えば量化子として検討対象となった“any”をどう扱うかを検討する際には、文書内での一貫性とは別にこの事例における判断が求められます。判断にあたって、翻訳者はほぼ確実に、“any”を“a”に変えても命題としては等価であると考えるでしょう。

> However, the set of all points on a straight line in R^2 not going through the origin (0,0) is not a vector subspace.

その上で、“any”を“a”のように変えて弱めても問題がないか、「任意の」という比較的分野特異性の高い表現を用いるか、あるいは「なんであれ」のようにできるかを検討し、実際に複数の訳文を作った上で、最終的な訳文を決定することになります。

以上は一つの例を起点とした、翻訳の中核プロセスの断片的な紹介にすぎませんが、こうした判断のほとんどが言語学的な意味での言語に関わるものでもTOEIC で測られるようなものでもないことは見て取れると思います。また、前提として必要な基本的理解は分野の専門的内容に関係しますが、“subspace”か“vector subspace”かの選択は分野の専門知識をめぐるものとは言えません。さらに、一部は読解一般の話と重なりますが、用語の選択などはやはり翻訳に固有の意思決定になります。

5. MT（+PE）はどこを扱っているか

MT はここで述べてきた翻訳プロセスのうち、巨視的なプロセスの中では「翻訳」の部分、そしてその「翻訳」の中では、テキストの言語的処理に、分野の用語体系をプラグインした範囲を扱っているというのが現在の状況だと考えられます[18]。扱っている範囲がそれに限られるのはもちろん技術的な問題でもありますが、翻訳とはどのようなプロセスか、そこにはどのような要因が関与しているかをめぐる理解の不十分さ（これは、明示的な知識に基づく理解という点ではここで見てきたようにMT だけの問題ではなく翻訳論の問題でもあります）も関係しています。まぐれあたりでもない限り、技術は問われていない問いへの答えを与えることはできません。

翻訳プロセスに照らしてみたときMT が扱っている範囲がこのように限られていることは、MT+PE の枠組みにおける PE の位置付けについても多少の示唆を与えます。現在のMT+PE が翻訳の完全な代替となりうると仮に考えるならば、翻訳プロセスのうちMT が扱っていないサブプロセスを PE が担うことになります。通常の翻訳では目的や想定利用者といったパラメータを事前に定めるのですが、MT は現在それを扱えないので（分野適応とは異なります）、PE は本来事前に定められそれに従って目標言語文書が作られるべきパラメータを考慮しないで作られた目標言語文書の文をそうした要因を事後的に考慮し直して修正する必要があります。これは自然な順序に逆らうかなり負荷の高いプロセスなので、翻訳の完全な代替となることを目指す場合、MT+PE が求める効率が犠牲になる可能性が高いことが推測されます。だとするとそのようなMT+PE は効率を求めるMT+PE と品質を求める翻訳との間で適切な位置付けを持たないことになるかもしれません。一方、MT+PE を翻訳と等価なものを実現するものとしてではなく、多少の品質を犠牲にしても効率を求めるものであるならば、PE の作業は MT の出力結果に対して、どちらかというと言語的な観点から微修正するものに止まることになりそうです。そ

の場合、現在の枠組みで MT がさらに進歩するならば、PE は不要になるかもしれません。

6. おわりに

本稿では、翻訳プロセスの概略を示しました。関与するサブプロセスと要素の一部を指示的に紹介するかあるいは事例ベースで断片を示すにとどまりましたが、それは紙幅の問題（だけ）ではなく、そもそも、翻訳プロセスの記述が、一定程度の専門性を有する翻訳者が理解することを前提になされていると同時にその専門性は明示的な知識として定式化されていないことからくる問題であり[19]、この点は今後の研究で明らかにされるべき課題として残されています[20]。

次回は、翻訳プロセスを記述する言語（メタ言語）の位置付けか、あるいは、機械翻訳と翻訳研究のメタ科学論的な対比か、いずれかを扱う予定です。いずれにしても、翻訳そのものというより翻訳を研究すること・語ることに関わるテーマになります。

謝辞

本記事の一部は、科研プロジェクトにおける論文／発表 “Metalanguage for the translation process” Translation in Transition: Human and Machine Intelligence, 15-17 October, 2020（オンライン会議）に依拠しています。筆頭著者山田優氏および共著者山本真佑花氏・大西菜奈美氏・藤田篤氏・宮田玲氏に感謝します。

注・参考文献

[1] https://tntc.p.u-tokyo.ac.jp/

[2] Jakobsen, A. (2017) “Translation process research,” In: Schwieter, J. and Ferreira, A. (eds.) *Handbook of Translation and Cognition*. Malden, MA: John Wiley & Sons. pp. 19-49.

[3] ISO (2015) *ISO 17100: 2015 Translation Services-Requirements for Translation Services*. Geneve: International Standard Organisation.

[4] 少しずれますが、これに関連して、公理的定義は対象の認識において特異な位置付けにありそうです。

[5] 池上高志・石黒浩・梅田聡・佐藤理史・中島秀之・開一夫 (2019) 「[座談会] 人工知能研究は何をめざすか（後編）」『科学』89(5), pp. 460-469.

[6] ISO (2015) 付属書 A。画像は日本規格協会ソリューションズ審査登録事業部サイト（https://shinsaweb.jsa.or.jp/MS/Service/ISO17100）から。

[7] 例えば、村上春樹・柴田元幸 (2000) 『翻訳夜話』文春新書（こちらは文芸翻訳談義ですが）とは、扱われている範囲が全く異なります。

[8] 料理に関しても対応する認知プロセスを考えることができますが、料理のプロセスと言ったときに認知的なプロセスを考える人は少ないようです。影浦峡 (2020) 「翻訳プロセス・モデルの構築を目指す研究」 *AAMT Journal*, 72, pp. 25-28. で関連する議論を行なっています。

[9] ここで科学にこだわる必要があるのかというと、やはりあって、原理的に同じものを具体的に共有することが可能な記号表現・操作の体系は暗黙知と徒弟的伝達により支えられる職人的な専門性と比べるとその性質において民主的で、公正な共有と規模拡大に資するためです。17 世紀には一部の特別な人たちしか理解できなかった微分積分のかなりを高校で学ぶことができるのは科学のその面がもたらした恩恵と考えることができます。ちなみに翻訳プロセスの記述においては、チョムスキーが理論言語学の科学的探求においてモデルとした物理学よりも、科学的分類に基づく生物学につながる普遍言語・普遍分類の構想が参考になると考えられます。後者については例えば Slaughter, M. M. (1982) *Universal Languages and Scientific Taxonomy in Seventeenth Century*. Cambridge: Cambridge University Press.が参考になります。

[10] Burchardt, B. and Lommel, A. (2014) *QT LaunchPad Supplement 1: Practical Guidelines for the Use of MQM in Scientific Research on*

Translation Quality. http://www.qt21. eu/downloads/MQM-usageguidelines.pdf.

[11] Specia, L. and Shah, K. (2014) "Predicting human translation quality," *Proceedings of the 1th AMTA*, pp. 288-300.

[12] 翻訳論の中に、要項で標準的に定められる項目の一つである目的を起点として翻訳戦略を定義するスコポス理論があり、実務翻訳や実務翻訳教育への有効性が認められていますが、記述をさらに具体化する余地は残されています。例えば Vermeer, H. J. (2000). "Skopos and commission in translation action," In Venuti, L. (ed.). *The Translation Studies Reader*. London: Routledge. pp. 221-232.

[13] 翻訳のポイントを事例ベースで指南したものには、例えば田辺希久子・光藤京子 (2008)『英日日英 プロが教える基礎からの翻訳スキル』三修社. など複数の良書があります。それらがいわば産物ベースであるのに対し、ここではプロセスを重視します。

[14] Searle, S. R. and Khuri, A. I. (2017) *Matrix Algebra Useful for Statistics*. Second Edition. Hoboken, NJ: John Wiley. p. 5.

[15] 公表媒体が内在的かどうかは議論の余地がありますが、翻訳が文書を対象とするという点を意識し、ここでは内在的属性に入れました。

[16] 本来は索引を中心に文書全体で用語の扱いを考える必要があります。この例では数学辞典や標準的な教科書が想定されますが、企業の技術文書等ではクライアントや翻訳サービス側が管理している場合もあり得ます。ここでは述べませんが翻訳メモリも類似の位置付けにあり、「単に近い」だけでは不十分で与えられた文書が適切にその中に位置付けられる文書集合の例を利用する必要があります。

[17] ちなみにこの本の索引に"vector subspace"はあるけれども"subspace"はなく、目次と前書きを除いて本文や章節見出しに 25 回現れる"subspace"のうち"vector"が付いていないのは 7 回で、規則性は確認できませんが、定義と索引で "vector subspace"を使っていることからかなり"vector subspace"という表現を重視していることが伺えます。

[18] 2020 年 9 月 24 日の時点で、4 節に挙げた例文の MT 結果は以下のようになります。

Google 翻訳：ただし、原点 (0,0) を通過しない R2 の直線上のすべての点のセットは、ベクトル部分空間ではありません。

みらい翻訳：ただし、R 2 の直線上で原点 (0, 0) を通過しないすべての点のセットは、ベクトルサブスペースではありません。

DeepL：しかし、原点(0,0)を通過しない R2 上の任意の直線上のすべての点の集合は、ベクトル部分空間ではありません。

みんなの自動翻訳 (特許)：しかしながら、原点 (0,0) を通らない R2 の任意の直線上の全ての点のセットは、ベクトル部分空間ではない。

[19] European Master's in Translation (2017) *European Master's in Translation Competence Framework*. は翻訳プロセスで発動することが求められるコンピテンスを具体化するための土台として作られていますが、それでも "Translate general and domain-specific material in one or several fields from one or several source languages into their target language(s), producing a 'fit for purpose' translation"のように中核において同義反復的な記述が残されています。

[20] この問題は翻訳に限るものではなく、例えば「読む」といったときにそもそも読むとは何をすることなのかもあまり具体的に共有されていません（これについては例えば、新井紀子 (2018)『AI vs. 教科書が読めない子どもたち』東洋経済新報社.）。

解説記事

書評: Neural Machine Translation by Philipp Koehn

渡辺 太郎
奈良先端科学技術大学院大学

1. はじめに

2016 年に Google がニューラル機械翻訳のサービスを始めて以来[1]、翻訳の精度が大幅に向上し、様々な分野で影響をもたらしたと思っています。今や 100 言語以上を対象としたサービスをしています[2]。私は初期のシステムの立ち上がりを横で眺めて、時々お手伝いをするだけでしたが、サービスとして立ち上げるため、論文にならないような様々な試行錯誤が行われた、という記憶があります。 また、ライバル企業による様々な研究開発が行われています[3, 4]。

2. Neural Machine Translation

本書 (Koehn, Neural Machine Translation, 2020) はニューラル機械翻訳の初めての入門書であり、同じ著者による統計的機械翻訳による初めての入門書 (Koehn, Statistical Machine Translation, 2012)の続きにあたります。序文にありますが、

> あなたが手にしている本は、この教科書の第二版の一つの章として計画されていましたが、新たな手法があまりにも発展してしまい、以前の手法がほぼ無関係になってしまい、結局一つの本になりました。

確かに、4 章の評価に関する内容以外、重なるものが全くと行ってなく、これまでの知識の積み重ねが全く役に立たない、といえます。ただ、Koehn 先生によりますと「この数年に始まったことなので、最先端の知識を勉強するのに、遅れはないですよ」と慰めてくれます。日々新しい論文が量産される世の中になりましたので、これも怪しいものですが。

本書は、三部構成になっていまして、第一部は翻訳とはどういうことをするのか、また、機械翻訳の種々のアプリケーション及び評価法についてまとめています。第二部は、ニューラル機械翻訳の仕組みについて、深層学習に始まり、言語モデルから翻訳モデルまで幅広く扱っています。第三部ではより深く、どのような研究課題があるのかを網羅しています。機械翻訳は知っている、という人は、第二部から始めてもいいですし、すでに深層学習に触れていて、機械翻訳に関する研究について知りたいようでしたら、第三部から読み始めてもいいでしょう。

2.1 機械翻訳について

第一部ではニューラル機械翻訳について解説する前に、機械翻訳がどのようにして発展してきたのかを解説しています。

1 章では、翻訳において、単語や句、構文、意味など様々な曖昧性をいかにして取り扱うのか、また、言語学的な知見および機械学習などデータに基づく手法からの知見について議論しています。翻訳はどうあるべきか、という難しい課題については「読み手に翻訳したことがばれない」と結論づけています。個人的には、例えば辞書目的で翻訳する場合なるべく直訳調の文体が良いなど、「使用目的に依存している」と思っていますので、異論等あるかもしれません。

[1] https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/
[2] https://blog.google/products/translate/new-look-google-translate-web/
[3] https://www.deepl.com/
[4] https://ai.facebook.com/blog/facebook-leads-wmt-translation-competition/

Book Review: Neural Machine Translation by Philipp Koehn
Taro Watanabe
Nara Institute of Science and Technology

2 章では、機械翻訳のアプリケーションを紹介しており「情報アクセス」「翻訳者の手助け」「コミュニケーション」と定義しています。機械翻訳の一般的な教科書ではよく「情報理解(assimilation)」「情報拡散(dissemination)」「コミュニケーション(communication)」と定義していますので、微妙に違うのでお気をつけください。

3 章では、研究開発の歴史についてまとめています。印象的なのは、ニューラルネットワークの説明から始まっていることでして、間違った教科書を読んだのかと数回読み返してしまいました。ニューラルネットワークを活用した深層学習の進展により難しいとされていた様々な AI の問題が解決されましたし、機械翻訳もその影響を大きく受けています。一番印象的なのは冒頭部分にあり、機械翻訳は長年の研究開発を経ているため、時には過大な期待を受け、また、失望も味わっている、という点でしょうか。今「人間と同等の能力」と宣伝されていますが、まだ解けていない問題も数多くあり、警鐘を鳴らしています。

4 章では、機械翻訳の評価手法について紹介しています。長年 BLEU (Papineni, Roukos, Ward, & Zhu, 2002)が指標として使われ、様々な批判について議論をしていますが、なかなかこれを超えるものがない、というのが残念とも言えます。この章はほぼ更新されていないため、人間と同等の能力を得た機械翻訳システムの微妙な差分をどのように評価したら良いのか、という課題について議論が欲しかったです (Läubli, et al., 2020)。

2.2 深層学習による機械翻訳

第二部は、この本の主題かもしれません。ニューラルネットワークは何なのかを説明したあとで、まずは言語モデルを解説し、現在の機械翻訳で用いられているエンコーダ・デコーダモデルについて説明しています。

5 章では、ニューラルネットワークについて解説しています。初期のパーセプトロンに始まり、その構造を深くして活性化関数を導入する、ということで非常にわかりやすく説明しています。さらにモデルの学習がどのように行われるのか、具体例を示しながら勾配を計算していますので、自分で計算してみることで深層学習がどのように行われるのかを実感できると思います。

6 章では、深層学習の発展に貢献した「計算グラフ」という概念について解説しています。計算グラフは、ニューラルネットワークを構築する時に、その構造を抽象化して表現するために用いられ、TensorFlow[5]や PyTorch[6]は Python により記述できるようにしています。一番大きいのは計算グラフを用いることで自動的に勾配を計算できることでして、研究者が微分するアルゴリズムを実装する必要がなくなりました。本章では、どのようにして計算グラフが構築され、さらに、微分を計算する時にどのように計算グラフが利用されているのかわかりやすく説明しています。

7 章では、ニューラルネットワークによる言語モデルについて解説しています。ただし、いきなり「N-gram」や「条件付き確率で履歴部分を省略することで近似する」など確率についての知識や N-gram 言語モデルを知っていることを前提としているので注意してください。前の著書 (Koehn, Statistical Machine Translation, 2012)の抱き合わせ販売を目的としているのか、とちょっと疑ってしまいました。N-gram 言語モデルについて知りたい、また、ついでに深層学習における自然言語処理が何をしているのかを知りたいのでしたら、「Speech and Language Processing」(Jurafsky & Martin, 2008)をおすすめします[7]。

本章では、Feed-Forward ネットワークによる言語モデルに始まり、そのために必要な単語ベクトルについて議論し、さらに、長い系列を表現可能な回帰型ネットワークについて簡潔に説明をしています。ただし、LSTM や GRU の解説については、Wikipedia や TensorFlow、PyTorch の API ドキュメントを読んだほうがわかりやすく、また、LSTM の解説では小さいバグがあるので気をつけてください。

[5] https://tensorflow.org/

[6] https://pytorch.org/

[7] 最新版のドラフトがあります。https://web.stanford.edu/~jurafsky/slp3/

8 章のニューラル機械翻訳では、ニューラル言語モデルを拡張したエンコーダ・デコーダモデルを解説しており、更に、アライメントおよび注意機構や、層を積み重ねた深いネットワークを説明しています。この章はたった 16 ページですが、既存の注意機構のモデルとは微妙に異なる、独自のデコーダのネットワークを解説していまして、各ステップで「注意機構で得られたベクトル表現を入力」「RNN と前のステップの目的言語の単語ベクトル表現、注意機構の出力から今のステップの目的言語を予測」「予測された目的言語の単語のベクトル表現を出力」ということで、理解するのに非常に時間がかかりました。TensorFlow や PyTorch のチュートリアルの方がわかりやすいかもしれません。

9 章は実際に翻訳を生成するビーム探索について説明しています。入力を与えられた時に、モデルから出力を得る問題ですので、機械学習では「推論(inference)」と呼ばれていますが、機械翻訳の分野では従来「雑音のある通信路モデル」に基づいていましたので「復号化(decoding)」とも呼ばれます。基本的なビーム探索に加えて、複数のモデルによる合意を取るアンサンブルについても議論しています。

2.3 ニューラル機械翻訳の発展

第三部ではニューラル機械翻訳の最近の研究成果についてまとめています。深層学習の様々な課題からデータの活用法および可視化まで非常に多岐に渡っています。

10 章では、モデルを構築し、学習する時の様々なコツをまとめています。過学習を防ぐための手法や学習時のパラメータの更新時に勾配が消滅あるいは発散する問題を解決する手法、さらには最適化まで広範囲の話題を扱っています。特に機械翻訳に特化した内容ではなく、また、すべての手法についてなぜうまくいくのかをすべて説明しているわけではありませんが、簡潔にまとめられています。

11 章では、Transformer (Vaswani, et al., 2017)など最新のモデルについて解説しています。8 章および 11 章では、回帰モデルによるエンコーダ・デコーダについて解説しており、予測された目的言語のベクトル表現を出力としていましたが、この章では、改められていて、逆にわかりやすいと思いました。

12 章では、自然言語を扱う時に最小の単位となる「単語」をどのように表現するのか、また、どのようにして単位を定義するのか、という課題について言及しています。テキストデータから単語ベクトルを学習する手法について、CBOW (Mikolov, Chen, Corrado, & Dean, 2013) や GloVe (Pennington, Socher, & Manning, 2014)などの手法や文脈に依存した BERT、多言語化についても解説しています。ニューラル機械翻訳ではモデルが扱える語彙数が多くなるとメモリーが増え、さらに、学習が難しくなります。語彙数を少なくするために、単語をさらに小さい単位へと分割する、Byte Pair Encoding (BPE) (Sennrich, Haddow, & Birch, 2016)や unigram に基づく手法 (Kudo, 2018) について解説しています。

13 章では、ドメイン適用について説明しています。モデルを学習する時に、実際に翻訳したい対象となるドメインのデータが大量にあるとは限りません。そのため、現存する学習データに対して学習して、目的とするドメインのデータに対して最適化をする、という手法が用いられていまして、様々な手法について議論しています。

14 章では、対訳データ以外の、単言語データを活用する手法について言及しています。ニューラル機械翻訳では、原言語の入力に対し、目的言語の出力が与えられる、教師あり学習によりモデルを学習しています。対訳データの量が多くない場合、単言語のデータを既存の機械翻訳システムにて翻訳することで対訳データの量を増やすことが可能であり、その場合でも翻訳精度の向上が見られます。他にも、二言語から多言語を活用する手法についても議論しています。せっかくなので、教師なし機械翻訳につてもより深く解説していたら良かったと思いました。

15 章では、単語の対応付や構文など言語学的な知識を活用する手法について言及しています。あまり深く解説していないのが残念ですが、逆に、あまり成功し

ていない、という現実をそのまま反映しているような気がします。

16 章では、現在課題となっている「ドメイン」「データ量」「低頻度語」「ノイズ」「探索」「アライメント」について言及しています。ついでに「偏り」の問題についても解説していただけたら、と思いました。

17 章では、モデルの分析方法について解説しています。エラーの分析手法に始まり、可視化により研究者がどのようにしてモデルの挙動を解析できるのか、その研究成果についてまとめています。この分野は非常に難しく、たとえ可視化したとしてもどのようにして修正したら良いのかわからない、という問題もあります。

3. おわりに

第一部は機械翻訳とはどのようなことをしているのか、また、どのように応用されているのかを知るためによくまとまっています。第二部に関しては実を言うと TensorFlow や PyTorch などのチュートリアルのほうがわかりやすい、という印象があります。機械翻訳はエンコーダ・デコーダに基づいており、この手法は要約や対話生成など幅広い分野で応用されています。このため、ツールキットは幅広い分野でユーザを増やそうと努力をしていますので、ドキュメンテーションが充実していますし、使い方まで勉強できてお得な感じがしました。第三部は機械翻訳の性能を向上するための様々な手法を解説していますので、これから研究を始めたい、という人には最初に読んでみると論文を検索するきっかけとなり、新たな研究の方向性が定まるかもしれません。ただし、この本を読んだあとで、機械翻訳に特有な技術は何なのでしょうか、という疑問が湧いてきました。その問いかけを探るには、やはり、過去の研究を見直さなければいけないのでは、結局、抱き合わせ商法と思いながらも前の教科書(Koehn, Statistical Machine Translation, 2012)も手にしないといけないのでは、と思いました。

数カ月間、学生とともに教科書を読む機会を得まして、自然言語処理や機械学習、深層学習の基礎から勉強する時には何から始めたらいいのか、ということを考えるよいきっかけになりました。このような観点から、入門書として非常によい教科書ですし、この本を、きっかけに様々な論文へのハブとなり、機械翻訳の研究がさらに発展することを期待したいと思っています。

参考文献

Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing. Prentice Hall.

Koehn, P. (2020). *Neural Machine Translation.* Cambridge University Press.

Koehn, P. (2012). *Statistical Machine Translation.* Cambridge University Press.

Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (pp. 66–75).

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. Journal of Artificial Intelligence Research.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (pp. 311-318).

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), (pp. 1532–1543).

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, (pp. 1715–1725).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems 30*, (pp. 5998-6008).

イベント報告

2020年度第1回JTF関西セミナー

『製薬業界におけるAI翻訳の現状と将来性』参加報告

東山翔平
国立研究開発法人情報通信研究機構

1. はじめに

本稿では，2020年6月23日に開催された，本年度第1回となるJTF関西セミナー『製薬業界におけるAI 翻訳の現状と将来性』について報告する．本セミナーは，新型コロナウイルス感染症（COVID-19）流行への対応措置としてオンライン形式での開催となり，JTF 会員・非会員の申込者に対して無料公開された．約 400 名の聴講者が参加し，各講演ならびにパネルディスカッションを通じて活発な質疑応答が行われた．

本セミナーの報告として，4 名の講演者による各講演と，同講演者をパネリストに迎えて行われたパネルディスカッションの内容を報告する．

2. 内山将夫氏：アダプテーションによる製薬専門NMTの研究開発

情報通信研究機構（以下，NICT） 内山将夫氏による講演について報告する．内山氏は，NMTの概要の解説と，製薬分野へのアダプテーション事例の紹介を行った．

コーパスに基づく機械翻訳（MT）のアルゴリズムは，1980年代の用例ベース機械翻訳（EBMT）以降，統計的機械翻訳（SMT），ニューラル機械翻訳（NMT）と発展してきた．内山氏は，NMT をさらに第 1 世代（Sutskever らの sequence-to-sequence）と第 2 世代（Vaswani らの Transofrmer）に分けて紹介した．第2世代モデルによる精度向上は，人が見ても翻訳文の質的な違いが感じられるレベルだという（Vasiwani らの論文では英独翻訳でBLEU 4ポイントの向上）．NICTの自動翻訳サービス『みんなの自動翻訳＠TexTra』でも，「汎用 NT」として第2世代モデルがすでに公開されている．アルゴリズムの発展により精度向上が実現される一方で，翻訳モデルの学習のための大規模な対訳データや，人間による評価の必要性は変わらないと述べた．

分野に特化した翻訳モデルを構築する方法としては，標準的な対訳データで訓練済みのモデルを，特定分野の対訳データで再訓練するアダプテーションがある．NICTでの製薬分野における取り組みとして，『翻訳バンク』の枠組みにより R&D Head Club のメンバー 8 社から320万文以上の日英対訳データが提供され，これをアダプテーションに用いて製薬専門の NMT モデルを構築したことが紹介された．NICT の内部評価では，第1世代から第2世代で BLEU 10ポイント前後の向上に加え，汎用から製薬専門で 20 ポイント近い向上が見られ，第 1 世代汎用モデルと比べると 30 ポイント近くに達する劇的な向上が達成されたという．事業会社により，この製薬専門 NMT を用いたサービス提供がすでに開始されているとのことである．

3. 木下潔氏：製薬会社での「翻訳作業」とAI翻訳への期待

MSD 株式会社（以下，MSD） 木下潔氏による講演について報告する．「これからの製薬会社の翻訳のウチアケ話」という主旨の講演であった．

Report on the JTF Kansai Seminar: the Present Status and Prospect of AI Translation in Pharmaceutical Industry
Shohei Higashiyama
National Institute of Information and Communications Technology

まず，木下氏は，製薬会社において翻訳がなくなることがないことを強調した．新薬開発では，非臨床・臨床試験後に承認申請の過程があり，国内・国外の当局とのやり取りで様々な文書を扱うためである．翻訳対象は，多量（～数百ページ）だが時間的余裕のある「重厚長大型」と，少量（～数十ページ）だが短期間で作成しなければならない「軽薄短小型」があり，状況に応じて一般社員，派遣社員が翻訳したり，翻訳会社に委託するという．

重厚長大型の場面では翻訳会社に委託することになるが，外注サイトを作成して翻訳委託プロセスを標準化することで，このタイプにおける問題はおおよそ解決されたそうだ．

一方，問題となるのは軽薄短小型である．たとえば，翌日期限といった短納期の照会事項が米国当局から突発的に発生する場面があり，米国本社のメンバーとやり取りしながら寝る間を惜しんでの作業になるという．このような時間的制約の厳しい場面では，回答の技術的内容の作成に注力するため，翻訳作業に割く労力は極力削減できることが望ましい．そのため，下訳のレベルであっても自動翻訳文が有用であり，2018 年以降，社内での AI 翻訳の利用場面は増えているという．

さらに，木下氏は，NICT の翻訳バンクの枠組みを知り，コーパス提供によって製薬業界に特化した AI 翻訳エンジンを実現することに注目したことを述べた．その後，MSD による MSD マニュアルの提供を経て，R&D Head Club による対訳データ提供に至り，内山氏の講演で紹介された製薬専門 NMT の開発に至った．同エンジンの印象については，製薬分野の言い回しを学習して訳文の質が向上しており，将来性を感じるものであったという．木下氏は，特化型翻訳エンジンは業界関係者全員にとっての資産になり得，継続的に整備し発展させていくことがすべての人の利益につながると展望を述べた．

4. 津山逸氏：翻訳者による MT 後編集のさじ加減

フリーランス翻訳者 津山逸氏の講演について報告する．津山氏は，MT 出力のポストエディット（後編集，PE）について，製薬・医療系文書の英日・日英翻訳における実例十数文を紹介した．

国際標準規格 ISO18587:2017 では，「概ね意味が通じるレベルへの修正」を light PE，「文体・スタイルまで含めて人間に近い品質への修正」を full PE の要求レベルと定義している．内山氏が述べた製薬専門 NMT による翻訳文を基に津山氏が PE を行ったところ，誤訳や訳抜けは皆無に近く，light PE では少数の細かい修正を除いてほぼ許容できるレベルであったという．“sample” を「サンプル」から「検体」に修正する正確な専門用語の使用や，新型コロナウイルス感染症に関する文章において “who had visited Wuhan, China” を「中国の武漢を訪れた」から「中国の武漢市への渡航歴のある」に修正する，より文脈に即した訳への変更などの例が紹介された．MT は語順などについて原文に忠実な翻訳を行う傾向があるため，英日・日英翻訳の full PE では，人間訳と同等の自然な訳文に仕上げるために，構造的な書き換えも必要になることが多いと述べた．

PE が翻訳の時間短縮になるかという聴講者からの質問に対しては，津山氏は，「必ずなる」と答えた．PE の容易さは原文の質に依存し，「よい原文」の場合は PE により時間短縮され，主語が不明瞭，係り受けが複雑といった原文の場合は，MT 出力をそのまま採用せず一から翻訳し直すためだという．

5. 早川威士氏：AI 翻訳を取り入れた翻訳ワークフローの設計

株式会社アスカコーポレーション 早川威士氏の講演について報告する．翻訳会社の立場から，翻訳におけるワークフローの設計についての講演であった．

人手翻訳のワークフローは，原文書に対して，前処理，翻訳，品質管理（QC），品質保証（QA），フォーマッティング，ユーザレビューの各段階を経て，翻訳文書が出来上がる．これが MT と PE を用いるワークフローとなった際に，「翻訳」が「MT+PE」に置き換わるかというと，議論の余地がある．MT は，低コストで超高速という長所と，品質管理・品質保証ができず，用語の制御や個別資料への適合が難しいという短所がある．つまり，MT が「低価格化」，「短納期化」に貢献できる可能性は高いが，「高品質化」には結び付きにくい．早川氏は，このような MT の性質を理解した上で，MT を使う目的を明確化し，そのためのワークフローを設計することが必要だと述べた．

具体的には，「とにかく納期を短縮したい」状況では，目標日数を決めた上で日数がかかるタスクを特定し，人手作業負荷を軽減する自動化等の方法を検討することになる（たとえば，PE の負荷を減らすための MT のチューニングなどである）．また，「とにかく費用を下げたい」状況では，許容できる品質レベルを考慮しながら，不要な作業・工数を削減することになる．

また，長期的な視点では，適切なワークフローの持続には，品質水準維持のための蓄積データの管理・検証や，納期短縮・コスト削減のためのシステム改善なども必要となるという．

このように，MT+PE 時代の翻訳会社の役割には，顧客ニーズの特定，ワークフローの設計と最適化，ワークフロー実現のためのシステム開発が求められる．また，翻訳者とは，目標やニーズを共有し，コミュニケーションを行いながら一緒に効率化を目指したいと早川氏は括った．

6. パネルディスカッション「三方良しの AI 翻訳実用化～製薬翻訳を題材に～」

前述の講演者 4 名（内山氏，木下氏，津山氏，早川氏）をパネリストに迎え，製薬翻訳を題材としたパネルディスカッションが行われた．モデレータは，アジア太平洋機械翻訳協会（AAMT）／情報通信研究機構（NICT）隅田英一郎氏が務めた．隅田氏が提起した議題と聴講者からの質問を題材としてディスカッションが進められた．以下，中心的な話題を取り上げて要点を報告する．

「PE の呼称を変えるべきか」　NMT 以前の MT は翻訳文の質が低く，これを基に PE を行うのが大変な作業であったことから，PE が酷評され悪いイメージが生じてしまった過去がある．PE の呼称を変えるべきかという提言に対し，ポストエディットや PE という呼称が定着しているため，このままでよいのではないか，言葉を変えなくても言葉のイメージは変わり得るという意見にまとまった．なお，プレエディット（前編集）も略すと PE になり得るが，これは略称が定着するまで浸透していないようである．

「翻訳者と PE の役割は異なるか」　翻訳者にとっては，翻訳メモリと同様に MT 自体も一つのツールであり，本来，翻訳と PE を区別する必要はない．ただし，現実に PE だけの仕事は発生しており，PE 専門の養成もあり得る．たとえば，MT 特有の誤りとして，原文の一部がそのまま残される未訳や，同一文字列が反復される現象などを知っておくべきで，また日本語原文の曖昧さや省略が MT になじみにくいという性質も理解しておく必要がある．しかし，full PE の場合でも対応できる技量は持っているべきで，これは翻訳者の技量と同じとなるはずである．また，これから翻訳者になる人が PE を想定してどう訓練していくかについては検討の道半ばで，翻訳者と翻訳会社だけでなく，翻訳者養成学校を有する企業との連携も必要になる．PE への報酬については，必要となる労力や成果物の翻訳の品質の高さに応じて十分な対価が支払われるべきという点で一致した．

「翻訳者や PE の仕事はなくならないか」　MT は内部パラメータに基づいて入力文を出力文に変換する処理を行っているにすぎず，その正しさをモデルが保証することはできない．翻訳の正しさを評価・保証する役割として人間の翻訳者が必要である．また，MT を利用しつつ PE・翻訳を行う翻訳者，チェックを行う翻訳会社，校正を行うネイティブスピーカーと，それぞ

れに異なるスキル・経験が必要で，人の仕事がなくなることは当分ないと考えられる．特に日英翻訳は人間が書く日本語原文の曖昧さが MT にとって厄介な問題であり，もし人間が必要なくなるなら日本語原文を機械が書くようになったときではないかという意見があった．

「AI 翻訳ソフトウェアの選び方」 セキュリティ性，カスタマイズ性，現時点での精度や将来的な期待精度などいくつかの観点が挙げられた．すべてのユーザに自分で選ぶことを求めるのは難しく，業界でコンサルしていくことも必要になる．

「どのようにコーパスを集めるべきか」 翻訳者や翻訳会社にとって，文書と翻訳メモリをセットで提供すると得をする仕組みがあれば協力を望めるのではないか．機械処理しやすい文書形式であることも重要で，人類の共有資産として言語資源を蓄積し発展させていく必要がある．

7. おわりに

JTF 関西セミナー『製薬業界における AI 翻訳の現状と将来性』について報告した．最後に，報告者の主観によるまとめを述べる．本セミナーでは，翻訳を必要とする製薬会社，翻訳会社，翻訳者，機械翻訳の研究開発者という異なる立場から翻訳に関わる実務者が登壇し，MT を取り入れた翻訳業務の現状と展望が語られた．MT の性質を理解して適切な使い方に注意を払う必要はあるとした上で，その有用性や将来性については肯定的な意見が占めた．実用レベルの MT を誰もが簡単に利用できるようになっている現状や，低コストで高速な翻訳へのニーズがあるという背景の下，MT の利用を前提としてどのように最適な業務活動を行っていくかという点が各者共通の中心的な課題意識となっているようである．産業翻訳業界の持続的な発展のためには，MT の実態・利用法についての認知を広めるとともに，関係者・協力者の裾野を広げていくことが今後も必要と考えられる．

法人会員 PR

ニューラル機械翻訳を活用した特許検索サービス Japio-GPG/FX

木下　聡
一般財団法人日本特許情報機構

1.　はじめに

（一財）日本特許情報機構（以下 Japio）は、㈱発明通信社と共同で、世界の主要国・地域・機関の特許公報全文を日本語と英語で横断的に検索できる Japio 世界特許情報全文検索サービス(Japio-GPG/FX)を提供しています。本サービスでは、英中韓独仏語で記述された約 8,000 万件の海外文献を、原文に加え、日本語に機械翻訳して検索サーバに登録しています。それにより、元々の日本語公報だけでなく、言語が異なる全ての文献を日本語のキーワードだけで横断的に検索することができ、また検索された文献を瞬時に日本語で読むことが可能となっています。

Japio 世界特許情報全文検索サービス(Japio-GPG/FX)
https://japio.or.jp/service/service05.html

サービス開始時は、ルールベース翻訳(RBMT)を用いていましたが、その後統計翻訳(SMT)に移行しました。現在は、SMT に加えてニューラル機械翻訳(NMT)も利用しています。

2.　特許翻訳用フレームワーク X-STEP®

一般的な技術文献と比べ特許文献は長文が多いことが特徴の一つです。このような長文をそのまま翻訳しようと思っても、翻訳エンジンの文字数制限などにより適切に翻訳できない場合も少なくありません。この問題を解決するため、Japio では X-STEP®(**X**ML Translation Framework with **S**tate-of-the-art **T**ranslation **E**ngines and Automatic Claim **P**re-editor)と呼ぶ機械翻訳用フレームワークを独自開発しました（図 1）。翻訳に際しては、前処理として、請求項の原文を翻訳しやすい形に自動で書き換えてから翻訳を行うことで、一度に翻訳できないような長い請求項に関しても、なるべく読みやすく理解しやすい訳文が得られるようにしています。

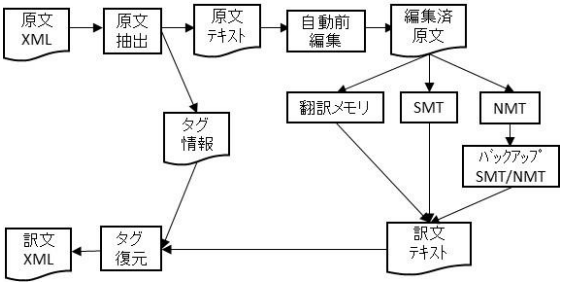


図 1　X-STEP®の処理フロー

3.　NMT の導入

Japio では、SMT の学習用に独自に収集した大規模な特許対訳コーパスを利用して、NMT についても独自に学習を行っています。2019 年 11 月には前述の Japio-GPG/FX に、特許文献をリアルタイムで翻訳して表示する機能を導入し、RBMT や SMT と比べて自然で読みやすい訳文を提供できるようにしました（リリース当初はβ版として全ユーザに開放していましたが、現在はオプション契約者のみが利用可能）。このほか、ヨーロッパ特許庁(EPO)が提供している世界各国の特許の要約についても、NMT で翻訳をしています。また、検索用のサーバに蓄積済みの文献データに加え、ユーザが入力したテキストを翻訳することも可能です。

Neural Machine Translation in Japio's Crosslingual Patent Search Service
Satoshi Kinoshita
Japan Patent Information Organization

今回 NMT を導入した言語は、英語と中国語、ドイツ語ですが、これらの中では中日翻訳での品質向上が顕著でした。SMT では原文での preordering（事前並び替え）を行っていたため、この段階での間違いがその後の処理に大きく影響していましたが、NMT では preordering をしなくとも高精度な構文的変換ができる学習がなされるため、大きな精度向上につながったと考えられます。

4. 訳抜けと湧き出しについて

上述の通り、NMT は一般的に、自然で読みやすい訳文を生成することができますが、SMT 以上に訳抜けや湧き出しが問題となっています。SMT の場合、それらは、原文中の一部の語が訳出されなかったり、原文には出現しない語が訳出される単語レベルの現象でしたが、NMT の場合には、句や節のレベルで欠落したり、同一の語句が繰り返し訳出される異常な訳文が出力される場合があります。これらの現象は、学習に用いる対訳コーパスを増やすことにより徐々に減少するものの、完全になくすことはできません。そのため、X-STEP では、NMT が出力した訳文を評価し、著しい訳抜けや湧き出しなどの不具合があった場合には、バックアップとして用意した SMT で翻訳しなおすハイブリッド方式を採用し、これらの課題に対応しています。

5. 訳揺れについて

RBMT が主流の時代には、MT を利用するメリットの一つは、専門用語辞書やユーザ辞書を使うことで、文献内に複数回出現する技術用語の訳を統一できるということでした。しかし SMT や NMT では、学習に使用する対訳コーパスの中に同一の見出しに対して異なる訳語が使われている事例があると、そのような事例から異なる訳し分けの仕方が学習されてしまいます。その結果、翻訳の際には、一つの特許文献でありながら、出現箇所によって異なる訳語が使われてしまうことがあります。

このような訳揺れは出版目的の翻訳では当然許されませんし、特許を調査したり審査したりする目的で翻訳結果を読む場合でも、理解の妨げになるため、統一された訳語を使用することが望ましいことは言うまでもありません。実際、Japio-GPG/FX の利用者の方々からいただいているフィードバックにおいても、訳語統一に対する要望は少なくありません。そのような声に応えるべく、現在 Japio においても精力的に研究を進めています。

6. 最後に

Japio が提供する特許検索サービス Japio-GPG/FX での NMT の活用について紹介しました。Japio では他言語の機械翻訳についても早期にサービスを提供できるよう鋭意努力しております。また、特許文献の検索や読解目的での利用に加え、外国出願のための明細書作成支援ツールなどにも NMT の活用の場を広げていきたいと考えています。

ところで、Japio では、新たに「ゆるキャラ」を作りました（図２）。「にゃぴお」が手にしている盾（A）と剣（I）に表れるとおり、Japio はこれからも AI に力を入れてまいります。

最後に、Japio-GPG/FX では試用 ID によるトライアルも可能ですので、興味のある方は Japio のホームページ(https://japio.or.jp/service/service05_05.html)にてお問合せください。

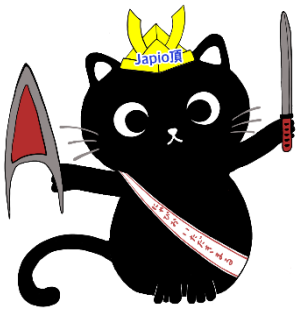


図 2 Japio 新キャラクター
「にゃぴお　いただきまる」
（商標登録出願中）

法人会員 PR

Flitto の翻訳ソリューションのご紹介

冨山 亮太
フリットジャパン株式会社 代表取締役

1. Flitto の翻訳サービス

フリットジャパン株式会社は、韓国で 2012 年にスタートアップ企業として創業した Flitto Inc.の日本法人です。機械（AI)翻訳とクラウドソーシング翻訳を組み合わせたプラットホーム「Flitto」で、ユーザー同士が自由に翻訳を依頼し合えるサービスを提供しています。2020 年現在、25 ヶ国語対応、世界 170 カ国でユーザー数世界 1,030 万人以上、登録翻訳家数 380 万人以上を抱えるグローバルプラットフォームへと成長しました。

また、サービスを提供するだけでなく、プラットフォームを通じて取得した言語データ（音声、テキスト、イメージ、翻訳）を収集・加工して提供することにも活発に取り組んでいます。

2. Flitto の AI 翻訳エンジン

プラットフォームを通じて独自に確保した膨大な言語データベース（DB、Database）と Flitto の技術によって商用化した AI 翻訳エンジンを提供しています。Flitto の翻訳エンジンは API でオーダーメイド型エンジンを提供することができ、データ学習を通じて産業、企業特化用の AI 翻訳エンジンとして利用できます。翻訳結果が不正確である場合はプラットフォームを通じてリアルタイムで全世界の翻訳家に翻訳をリクエストし、正確な翻訳データに更新できるのが Flitto エンジンの最大の特徴です。クラウド翻訳をリクエストした件数に比例して一定の手数料を受け取るのが Flitto の AI 翻訳エンジンの収益モデルでもあります。韓国では AI 翻訳エンジンを独自開発して商用化した企業が Naver、カカオなど大手企業に続き珍しいケースであるため、業界で注目されています。

3. Flitto の AI 翻訳エンジンの強み

翻訳エンジンの性能はデータ量に比例して、性能を向上していきます。「Flitto」は翻訳プラットフォームを運営しながら質の高い言語データの収集と分析を行ってきた為、言語データに対する知見が豊富です。もちろん、収集した言語データについては著作権に準じ、問題のないデータだけを扱っています。

また、最大の強みはリアルタイム学習が可能ということです。翻訳リクエストで更新されたデータは、再学習によってエンジンの精度を高め、各業界、企業に特化した翻訳性能を身につけることが可能になり、同じ言葉を入力しても、医学や科学用語へと翻訳する企業専用の機械（AI）翻訳エンジンを構築することができます。「Flitto」は翻訳サービス利用の費用をお客様にお支払いしていただくため、データの著作権はお客様であり、データ所有権も明確です。

4. Flitto の AI 翻訳エンジンの活用分野

翻訳が必要なグローバルサービスであれば、どの分野でも活用可能です。グローバルユーザーがレビューを共有する e コマースや、グローバルユーザーが多く、リアルタイムコミュニケーションをする観光・旅行・宿泊サービスに適しています。

また、新造語がよく使われるグローバルコミュニティサービス、多くのクライアントを持ち翻訳の正確さを重要とするメディア企業など、全ての分野におけるリアルタイムカスタマーサービスのための人工知能チャットボットに活用可能です。

Introduction of Flitto Translation Solution
Ryota Tomiyama
Flitto Japan Inc.

編集後記

石川 弘美
AAMT 編集委員会

今年は激動の年となりました。本来ならオリンピックイヤーでお祭りムード一色になるはずでしたが、新型コロナウイルスが私たちの生活を大きく変え、グローバリゼーションが進んだ世の中で、海外渡航が制限されるという未曽有の事態となりました。テクノロジーが進んだ現代でも、ウイルスひとつで世界中の生活様式を変えさせるほどの威力があることを見せつけられ、人類の無力さを痛感します。

しかし、この変化も悪いことだけではないようで、特に、セミナーやイベントはオンラインで開催されるものがほとんどとなり、これまでは場所の制約により参加できなかった方々にも参加していただくことができるようになりました。京都大学の黒橋先生に巻頭言でご紹介いただいた「COVIT-19 世界情報集約サイト」のように、制約がある中でテクノロジーを駆使することにより新たな進歩が生まれています。

さて、この事態の中ですが、設立 29 年目の今年、AAMT は任意団体から一般社団法人となりました。MT の技術が日進月歩で向上し、注目を集める技術になった今、AAMT が果たす役割はこれまで以上に大きなものとなっていくと思います。来年は AAMT 設立 30 周年を記念する年となります。皆様と一緒にお祝いできるような社会状況になっていることを祈ります。

今回寄稿していただいた AAMT 長尾賞および AAMT 長尾賞学生奨励賞受賞者の皆様には、本来なら 6 月に行われる総会において長尾真先生からの賞状授与と記念講演が予定されていましたが、代わりにメールで受賞をご連絡し、記念講演は 12 月 2 日の年次大会に持ち越されることとなりました。今年は、長尾賞 2 件、長尾賞学生奨励賞 2 件と、過去最多の受賞者数を記録しました。このことからも MT 技術への注目度の高さが窺われます。受賞理由は AAMT の Web サイトに掲載しておりますので、是非ご覧ください。長尾先生からは、お祝いとして「真意深（しんいはふかし）」という揮毫を頂きました。

レッドハット株式会社の燃脇綾子様には、製品リリースと同時に MT を活用して関連ドキュメントの翻訳版を公開するまでの過程を、現場の視点で解説した記事を頂きました。ユーザーにとっては、訳文の品質よりも、そこに書かれている情報の提供スピードの方が重要であることが分かります。今後の“翻訳”の在り方について考えさせられました。

国立情報学研究所コンテンツ科学研究系の阿辺川武様には専門用語の対訳対を収集するシステム「QRpotato」についての解説を頂きました。MT の品質を向上させる鍵となる専門用語集を機械的に作成できれば、世間を騒がすような MT 独特の誤訳を減らすことができるかもしれません。

東京大学の影浦峡先生には、翻訳プロセスについての連載第 3 回目を、奈良先先端技術大学院大学の渡辺太郎様には NMT の入門書である『Neural Machine Translation』の書評を、また情報通信研究機構の東山翔平様には、COVID-19 の影響で延期になっていた JTF 関西セミナー『製薬業界における AI 翻訳の現状と将来性』の参加報告を頂きました。

法人会員の PR 記事として、（一財）日本特許情報機構様とフリットジャパン株式会社様から寄稿を頂きました。引き続き法人会員の PR 記事を募集しておりますので、ご希望の会員様は、AAMT からの募集メールに記載されている連絡先までご連絡ください。

今後も、掲載をご希望のトピックや、記事に関するご意見などがありましたら是非お寄せください。

Editor’s note
Hiromi Ishikawa
AAMT Editorial Board

AAMTジャーナル「機械翻訳」No. 73

【　発　行　】アジア太平洋機械翻訳協会（AAMT）
ホームページ：https://aamt.info/
【　住　所　】〒619-0289
京都府相楽郡精華町光台3-5
国立研究開発法人情報通信研究機構　先進的翻訳技術研究室内
【編集委員会】内山将夫　後藤功雄　中澤敏明　新田順也　園尾聡　森口功造　隅田英一郎
仲山裕子　小谷克則　宇津呂武仁　山田優　石川弘美
【表紙デザイン】泉谷東十郎
【　題　字　】長尾真
【　事務局　】石川弘美
【　印刷所　】株式会社 プリントパック

Asia-Pacific Association for Machine Translation (AAMT)
c/o National Institute of Information and Communications Technology
3-5 Hikaridai Seika-cho Soraku-gun Kyoto 619-0289, Japan


AAMT
Asia-Pacific Association for Machine Translation
アジア太平洋機械翻訳協会