JSTにおける英日機械翻訳システムの構築

Development of English-Japanese Machine Translation System in JST

国立研究開発法人科学技術振興機構　情報企画部
関根　基樹

文献データに関する機械処理システムの開発・運用に従事。本稿は 1、2、3 章を担当。

国立研究開発法人科学技術振興機構　情報企画部
菊井　玄一郎

NTT 研究所、岡山県立大学情報工学部、農研機構・農業情報研究センターをへて現職。自然言語処理技術の文献データへの適用検討に従事。本稿は 4 章を担当。

国立研究開発法人科学技術振興機構　情報企画部
水田　寿雄

ファンディング業務等を経て、現在情報資産の提供や企画等に従事。本稿は主に 5 章を担当。

1　はじめに

国立研究開発法人科学技術振興機構（以下「JST」）は、国内外の科学技術や医学薬学関係の文献情報を収集・整理・加工し、日本語で検索できる日本最大級の科学技術文献データファイルを作成し、J-GLOBAL[1] や JDream Ⅲ [2] などの高度な検索、分析･可視化サービスにデータ提供している。

海外文献の科学技術文献データファイルの作成では、2014 年より英日機械翻訳システムを導入し、2018 年にはニューラル機械翻訳エンジンに切り替え、翻訳精度の向上や作成コストの削減を実現し、速報性・提供数の拡充を果たすことにより、科学技術情報の流通に寄与してきた。また、JST の機械翻訳システムは、これまで JST の業務において蓄積した数千万件の英文、和文の文対データを対訳コーパスとして用いていることなど特徴・強みを有していた。しかしながら、近年は民間企業等において最新技術の導入などにより機械翻訳の性能向上が著しく、JST においては精度向上スパンが長周期化しつつあることが課題となり、強みを活かした新しい機械翻訳システムを確立したいと考えた。

このため、昨年度より、まず JST にて稼働中の機械翻訳システムについて現状認識からはじめ、問題点を洗い出し、問題を分類して課題を設定し、課題解決の方法を決定した。そして、他社サービス・エンジンを対象に、仕様等のオープン情報を調査して無料トライアルを行い、続けて有料フィジビリティスタディを実施して、最適と考えられるエンジンとして OpenNMT[3]（以下「ONMT」）を選定し、外注によりエンジン切り替え・システム開発を実施することとした。

本稿では、エンジン選定に至るまでの過程や、ONMT の試行等を紹介し、今後の展望について述べる。

2 課題解決のための検討

最初に、課題解決のための検討イメージを図 1 に示す。

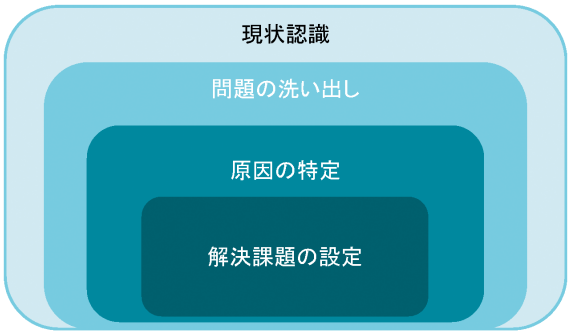

図 1　課題解決のための検討イメージ

現状認識から解決課題の設定までを検討して課題解決へ進むイメージである。

(1)JST における機械翻訳の現状認識

まず JST の機械翻訳について現状を整理し、その現状認識を対象にして、改善や不足事項等を洗い出し、問題事項として分類した。

現状認識の対象は、機械翻訳を導入してからの年度ごとの機械翻訳件数の推移、人手修正作業の件数や作業時間、システム改修遍歴、自動評価値（BLEU 値、RIBES 値）の把握、システムの機能構成、体制、予算などである。

(2) 問題の洗い出し

認識された現状に対して、システム運用・管理・利用ユーザー・技術等の様々な切り口から問題提起を行い、問題や解決の類似性などを踏まえて整理した。

一例として、次に示す「評価手法の確立」を挙げる。

問題提起：人による（ユーザー視点での）品質評価が十分ではない
・品質評価が無いと何が問題となるか
→（問題）翻訳を見るのは人であり、人手評価が無いと訳質の良し悪しを判断できない、人手の評価はブレがある、時間・コストが必要だが得られる情報は限定的、都度・任意に行うことが困難
→（代表的な問題）訳質向上の判断をどう行うべきか、求める訳質の目標をどうすべきか
→（分類）評価手法の確立

以上の様な取り組みを進めた結果として、主要な問題を次の 5 種類に分類した。

①評価手法の確立
②翻訳の改善
③記事翻訳件数
④サーバや運用
⑤コーパス

(3) 原因の特定と解決課題の設定

次に、分類した問題毎に、何故その問題が生じているのかという原因と、その解決策は何かという解決課題の設定を行った。例えば、以下のようなものである。

JST では翻訳の目的を「原文を忠実に日本語として翻訳される。自然な翻訳 (てにをは等) になっている」としているが、訳質として何を評価して、向上すべき目標をどう設定するか明らかではない	
解決の方策	翻訳目的を達成するために人手評価方法を確立する
解決課題の設定	正確性、流暢性、技術用語精度を指標として評価する

訳質の良し悪しを簡便に判断する指標および定量的な評価方法がない	
解決の方策	人手評価方法をより簡易使用できるよう自動評価との相関を向上させる
解決課題の設定	同じ評価用データに対して自動評価と人手評価を行って、自動評価値と人手評価値の相関関係を確認し、自動評価のパラメータを調整する

こうした問題点や解決策の中で特に共通する重大な解決課題が「高精度の翻訳エンジンの利用」であったため、

課題解決がなされる新エンジンの特定や特性設定を行うとともに、新しい翻訳サービス・エンジンの調査・導入を重点的に進めることとした。

具体的には、翻訳サービス・エンジンを対象に仕様等のオープン情報を調査して無料トライアルを行い、続けて有料フィジビリティスタディを実施して、最適と考えられるエンジンを選定した。

3 無料トライアル・有料フィジビリティスタディ

十数件の翻訳サービス・エンジンを対象に、Web サイトに掲載されているオープン情報をもとに、仕様、料金、データ利用、規約などを調査･比較した。このうち、サービス・エンジン提供ベンダ側でデータの二次利用が行われるもの、ユーザー側でのデータの営利利用が禁止されるもの、サービスの準拠法・裁判管轄が日本国内でないものは除外した。

そして、5 件を対象に無料トライアルを実施し、評価用データを翻訳した結果について、自動評価、繰り返し文字の検出を行ったほか、用語登録の反映が正しくなされているかなど確認した。同時に、JST 内製で ONMT を用いたエンジンを調査して試行版システムを構築し、比較検証することにした。JST 内製の ONMT 試行版システムの詳細は第 4 章で述べる。

フィジビリティスタディは対象を他社サービス・エンジン 3 件および JST 内製の ONMT 試行版システムの合計 4 件に絞って実施した。他社サービス・エンジンの調査にあたり、それぞれのベンダには、コーパス学習、用語登録とチューニング、JST から翻訳するための API 開発を発注し、さらに翻訳時間の計測、サーバ仕様または運用料金体系の提示を納品物として求めた。出力された翻訳結果に対して JST にて次の項目を評価した。

・人手評価（相対評価、絶対評価）
・自動評価（BLEU 値、RIBES 値）
・文末が体言止めと用言止めで使い分けられているか
・繰り返し文字が出力されていないか
・用語登録された語について変換エラーが生じていないか

このうち、人手評価については、4 サービス・エンジンを対象に相対評価と絶対評価を行うこととし、相対評価は評価用データから 8 分野、絶対評価は評価用データから 1 分野を選定した。相対評価と絶対評価ともに

表 1　人手評価による比較結果

	JST内製	参考：A社(注)	B社	C社
相対評価(※1)	**572**	**534**	**521**	**465**
うち正確性	216	184	210	165
うち流暢性	188	244	159	179
うち技術用語	168	106	152	121
絶対評価(※2)	**632**	**581**	**636**	**604**
うち正確性	211	186	213	200
うち流暢性	210	203	210	198
うち技術用語	211	192	213	206

(※1)2エンジン間を正確性、流暢性、技術用語精度のそれぞれで評価し、勝っている方に1点を加算して合計した

(※2)各エンジンの正確性、流暢性、技術用語精度をそれぞれ1～5段階（点）で採点して合計した

(注)有料フィジビリティスタディの同等費用内では十分な学習が出来ていない

表２　自動評価・費用評価による比較結果

		JST内製	参考：A社(注)	B社	C社
自動評価(※1)	抄録：BLEU	＋	－－	－	＋
	抄録：RIBES	＋	－	＋	＋
	標題：BLEU	＋	－－	＋＋	＋
	標題：RIBES	＋	－	＋	＋
費用評価(※2)	初年度	＋＋	＋＋	＋＋	＋＋
	２年間	＋＋	＋	＋＋	＋＋
	定常状態	＋＋	－－	＋	－

(※1)現行システムの自動評価値に対して、下記の区分により評価
＋＋　現行システムより大幅に向上（5ポイント以上）
＋　現行システムより向上
－　現行システムより低下
－－　現行システムより大幅に低下（5ポイント以上）

(※2)現行システムの運用経費に対して、下記の区分により評価
＋＋　現行システムより大幅に向上（20ポイント以上）
＋　現行システムより向上
－　現行システムより低下
－－　現行システムより大幅に低下（20ポイント以上）

(注)有料フィジビリティスタディの同等費用内では十分な学習が出来ていない

評価指標は正確性、流暢性、技術用語精度の３項目とし、相対評価では４サービス・エンジンを２種類ずつペアにして、どちらの翻訳結果が良いか３指標それぞれで比較して、勝っている方のサービス・エンジンに１点を加算する採点、絶対評価では３指標をそれぞれ１～５段階（点）で採点を行った。その結果を表１に示す。

また、表２の通り、それぞれのサービス・エンジンについて現行システムの自動評価値に対する比較、および現行システムの運用経費に対する比較を行った。

いずれもONMTを採用したJSTまたはベンダのサービス・エンジンが最良の結果を示した。ONMTはGPUサーバではなくCPUサーバのみで翻訳処理を実行でき、中長期に設備的・人的費用が大きく軽減できるといった利点もある。以上の結果からONMTを選定した。

4 ONMT

4.1 ONMTの概要

ONMTはオープンソースのニューラル機械翻訳のソフトウエアである。ONMTはpytorch版およびtensorflow版の２つの翻訳ソフトウエア群を中心にtokenizer、および、軽量な翻訳実行系であるCTranslate2の大きく４つのソフトウエア群から構成されている。pytorch版（以下「pt版」）およびtensorflow版（以下「tf版」）の翻訳ソフトウエア群はいずれもtransformerモデルを具現化した学習系と翻訳実行系を持つ。両者は一方が他方の単なる移植というわけではなく細かい設計が異なっているため、翻訳の結果の同一性は保証されない（パラメータセットすら異なる）。Tokenizerはsentence-piece(SP)およびbyte-pair-encoding(BPE)の両方が利用可能であり、それぞれのモデルを学習する機能も含まれる。CTranslate2（以下「CT2」）はCPUで動作する翻訳実行系であり、モデルはpt版あるいはtf版で学習させたものを付属の変換プログラムにより変換して利用する。

4.2 検討の概要

ONMT は類似機能を持つ複数のプログラムを含み、それぞれは多数のオプションを持つ。主な選択肢は以下の通りである。

1）サブワード分割：方式（SP/BPE）、語彙サイズ

2）翻訳学習系：実装（pt 版 /tf 版）、各種ハイパーパラメータ設定

3）翻訳実行）：モデル学習と同じ実装（pt 版 /tf 版）、あるいは、CT2

これらの組み合わせは膨大になるため、次のような手順で翻訳学習・翻訳を試み、評価セットに対する自動評価値を指標としてスコアの高いものを選んだ。

まず、抄録文 500 万文対を学習データとして、翻訳ソフトウエアを tf 版に固定し、tokenize 方式 (SP/BPE1)、語彙サイズ (32k/50k/100k, 日英同サイズ）を変えて実験したしたところ、BPE,SP とも語彙サイズ 50k がベストであり、BPE が SP より若干優れているという結果となった。次に語彙サイズを 50k に固定して pt 版で実験したところ、BPE,SP とも tf 版 BPE,50k とほぼ同等の性能となった。SP は形態素解析（pre-tokenize）を必要としないため、実運用を考えると有利である。ここまでで pt 版 -SP-50k, tf 版 -BPE-50k の２セットが候補となった。

次に翻訳実行系について検討した。tf 版、pt 版および CT2 を全て CPU モードで動かして翻訳時間を調べた。CT2 についてはモデルの変換が伴うため、変換前のモデルを当該モデルの学習で用いた実装（pt または ft 版）の実行系による翻訳結果と変換後のモデルによる CT2 の翻訳結果でデグレードがないかチェックした。なお翻訳実行時のパラメータセット（ビーム幅など）は極力揃えた。翻訳速度は CT2 が pt 版 , tf 版それぞれに対しておよそ 1/3 程度となり、高速に処理できることが分かった。翻訳結果についてはほぼ同じ（約 2600 文のうち 3 文に何らかの差が存在する以外は完全一致）であり、自動評価値（BLEU/RIBES）もほぼ一致という結果となった。

4.3 対訳辞書への対応

科学技術文献の翻訳においては入力文の専門用語を特定の用語に訳出することが必要となる。外部から対訳辞書を与えて、これを制約として翻訳を行う手法については現在研究段階であるため、今回の開発では単純な「ダミー語置き換え方式」により実現を試みた。ダミー語置き換え方式とは、１）原言語側で対訳辞書にヒットする単語を「ダミー語（例 :<x_01>)」に置き換え、２）翻訳処理ではこのダミー語をそのままの文字列で訳出し、３）出力に存在するダミー語を対訳辞書の目的言語側単語に置き換える、という処理である。ここで問題となるのが、原文側のダミー語が目的言語の適切な位置に訳出されるか、ということである。ニューラル翻訳においてこれを保証することは難しい（制約付き翻訳そのものの問題である)。我々は、学習用の対訳データの一部について、対訳関係にあると思われる単語のペアをダミー語に置き換えた「追加学習用コーパス」を作成し、ダミー語を含まないコーパスで訓練したベースモデルに対して追加学習を行うことにより、この課題への対応を試みた。追加学習用コーパスは翻訳学習用のコーパスに対して既存の手法により単語アラインメントを求め、１：１対応する単語対を同じダミー語に置き換えることにより自動作成した。なお、ダミー語は tokenizer によるサブワード分割の対象外とした。追加学習用コーパスは通常の翻訳にとっては「ノイズ」であるからその量や追加学習ステップ数を適切にコントロールすることが必要となる。そこで、ダミー語への置き換えを行わない検証用セットで自動評価値の変動とダミー語の湧き出し誤りをチェックし、対訳辞書にヒットした用語をダミー語に置き換えた検証用セットを用いて、ダミー語が過不足なく出力側に出現しているかをチェックした。その結果 tf 版と pt 版で若干数値が違うが予約語の消失は 0.3-0.8% 程度、湧き出しは 0.001% 程度（いずれも文単位）発生したが比率は小さく、また、これらの誤りは機械的に検出可能であるため実用可能と判断した。また、1200 文程度の評価文について、対訳辞書を反映するように訳文を書き換えて参照訳を作り、上述のダミー語置き換え方式により、対訳辞書を反映させた自動翻訳結果を作成して自動評価したところ、自動評価値の低下は見られなかったことから、この方式を採用することとした。

5 システム構築と今後の展望

３章、４章の通り、エンジンやその開発の主要事項が定まったことをうけ、既存エンジンの切り替え・システ

ム開発を実施することとした。約５年ぶりの機械翻訳の刷新の開始である。

刷新は、2023 年度の第 1 四半期の調達から始まり、第３四半期には本格的な開発着手、システムテストを予定している。試験稼働後は既存システムとの並行稼働を行い、新システムの運用を成熟させつつ、新たな翻訳システムに係る業務への影響把握・対策を講じる予定である。このようなシステム切り換え・開発・運用試行を経て、年度内には本格稼働の見通しを得る予定である。

これまで紹介したように、JST における新たな英日機械翻訳システムの構築は、現状認識から問題を抽出し、問題から課題を設定し、その課題解決方法を決定し、課題解決がなされる新エンジンの特定を行うことで進んでいる。数年ぶりのシステム刷新ということもあり、多くの手探りを伴う作業の結果でもあり、本出稿の機会を捉え、検討の経緯等を記録し共有することとした。今後のシステム更新において参考となれば幸いである。

また、今般の取り組みを通じて、過去から現在までの現状理解と課題、新システムの構築の見通しを得た。今後、これらを基盤として、更なる将来像・次期システムの検討を開始し、具体的な導入ロードマップを検討していきたいと考えている。

参考文献

[1]J-GLOBAL
https://jglobal.jst.go.jp/
[2]JDreamIII
https://jdream3.com/
[3]OpenNMT
https://opennmt.net/