JSTにおける英語・中国語科学技術文献の機械翻訳

Machine translation of English and Chinese science and technology literature in Japan Science and Technology Agency

国立研究開発法人科学技術振興機構　（JST）情報企画部主任調査員
岩城　修

1980 年東京工業大学大学院総合理工学研究科電子システム専攻修士課程修了。日本電信電話公社（現 日本電信電話株式会社）、株式会社 NTT データ等を経て、2015 年 7 月より現職。機械翻訳システム等の開発･運用業務に従事。博士（工学）。

国立研究開発法人科学技術振興機構　（JST）情報企画部主任調査員
松永　務

1988 年電気通信大学大学院通信工学専攻修士課程修了。株式会社 NTT データ技術開発本部にてデータマイニング、機械翻訳に関する技術開発に従事。2017 年 10 月より出向し、現職。博士（工学）。

国立研究開発法人科学技術振興機構　（JST）情報企画部調査役
堀内　美穂

1988 年北海道大学薬学部製薬化学科卒業。同年、日本科学技術情報センター（現、科学技術新興機構）に入職。文献データベースの作成、開発、研究者データベースの運営に従事し、2018 年より現職。

1 はじめに

JST では、我が国の研究開発活動の効率的実施を促し、科学技術の振興を図ることを目的に、国内外から収集した科学技術や医学・薬学関係の文献に抄録・索引を付与した文献情報データベースを整備し、J-GLOBAL[1] や JDream Ⅲ[2] などのインターネットアクセス可能な検索サービスを提供している。これらは科学技術に関する国内誌を網羅的に収集している唯一のデータベースであり、外国誌（英語・中国語）については出版者から許諾を得て日本語訳を作成することにより、全て日本語での検索を可能としている。ここに、その日本語への翻訳は永く人手で実施していたが、収録する外国誌の規模拡大と情報提供の迅速性を確保するため、2014 年から機械翻訳の導入を図っている。

本稿では、文献情報データベースに蓄積された標題・抄録の日本語訳に基づく対訳コーパスをはじめ、JST において構築整備している科学技術用語辞書[3] を活用して導入した科学技術文献の機械翻訳について述べる。

2 JST における機械翻訳の導入実績

JST における年間記事収録件数の過去 5 年間の推移と外国誌への機械翻訳導入比率を図 1 に示す。2014 年度からまず中国語文献に機械翻訳を導入し、英語文献には 2016 年度から導入したことで、外国誌の年間記事収録件数の大幅な増を図から読み取ることができる。2018 年度からは外国誌に対して機械翻訳を全面導入し、英語文献で約 100 万件、中国語文献で約 50 万件の実績で、2019 年度はさらに拡大する計画である。

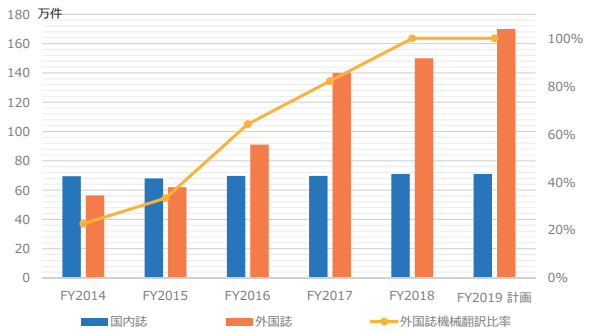

図１　記事収録件数の年次推移と外国誌の機械翻訳導入比率

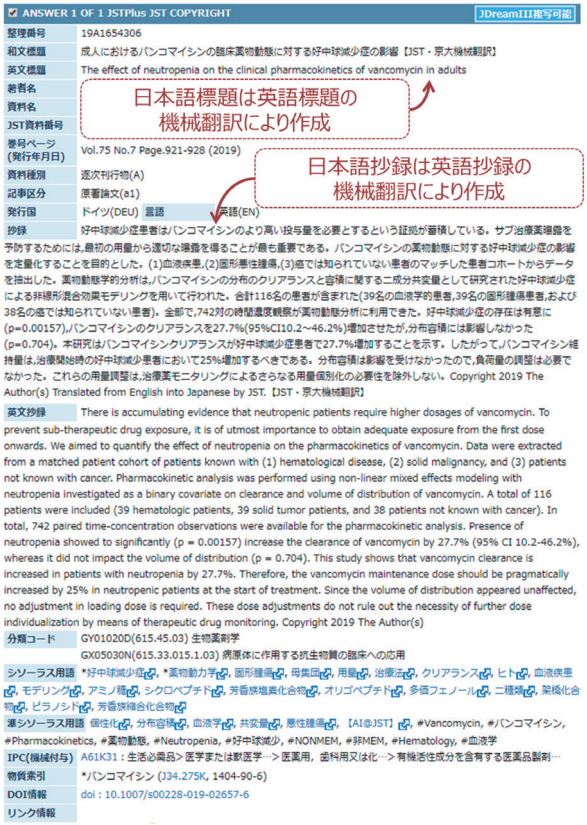

図２　JDream Ⅲでの検索結果の回答表示画面の一例
（「バンコマイシン」の検索語でヒットした記事候補から選択して得られた表示画面例）

図２はJDream Ⅲでの検索結果の回答表示画面の一例で、日本語標題および抄録がそれぞれ英語標題および抄録の機械翻訳により作成されている。機械翻訳の導入にあたり、必要に応じて原文（英文）を参照できるよう、図のように併せて表示することとしている。

3　ニューラル機械翻訳の導入

3.1　ニューラル機械翻訳エンジンの開発

JSTでは、科学技術文献に対する実用的な翻訳精度を達成することを目標に2013年から2017年まで日中・中日機械翻訳実用化プロジェクト[4]を推進しており、ここで開発されたニューラル機械翻訳エンジン（KyotoNMT）[5]を現在採用している。2016年12月に開催されたアジア言語を対象とした国際的な機械翻訳のワークショップ「WAT 2016」（Workshop on Asian Translation 2016）[6]での評価において、ニューラル機械翻訳方式を用いて科学技術情報の日中・中日機械翻訳タスクで１位の精度を達成している。

図３はJSTにおける機械翻訳の処理フローの概要である。収録される科学技術文献の記事原文に含まれる特殊文字列の置換処理を経た上で、対訳用語辞書に登録された科学技術用語に対しては、前処理でマスク処理して

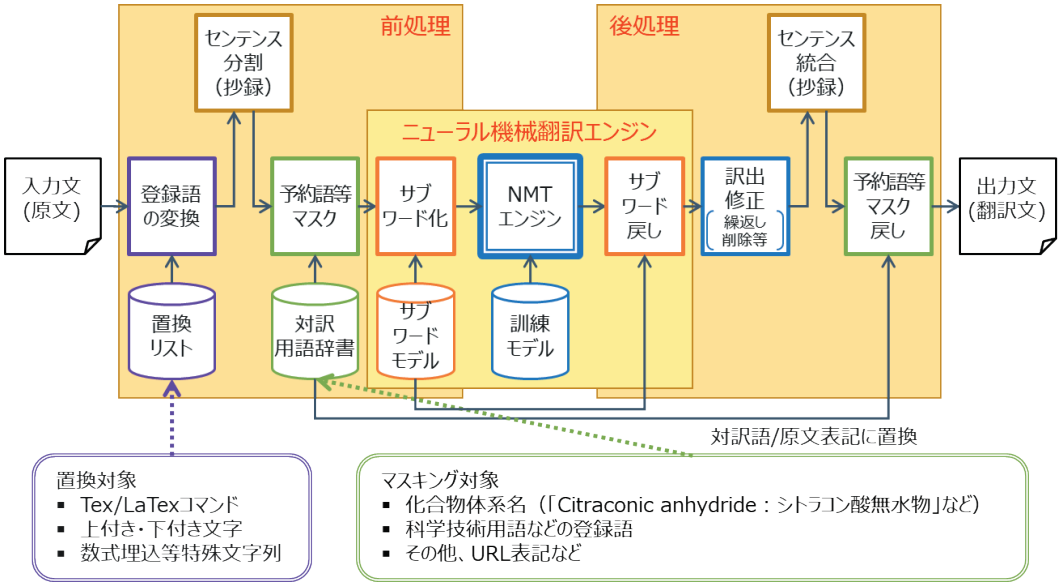

図３　ニューラル機械翻訳の処理フロー概要

後処理で対応する語に置換して翻訳文を得る予約語化処理を行っている。（4 章で事例を詳説）

3.2　対訳コーパスの構築と整備

JST では、医学・薬学分野を含めて広く科学技術文献を扱う中で、英語標題・抄録の人手による日本語訳作成から英日対訳を蓄積保有し、また記事への索引付けを目的に科学技術用語辞書を構築整備しており、これらの言語データ資産を活用して機械翻訳を開発している。

対訳コーパスの構築において、複数文からなる抄録に関しては文アライメント処理 [7] により文対形式にして機械翻訳に用いる対訳コーパスとしている。中国語・日本語の対訳コーパスについては日中・中日機械翻訳実用化プロジェクトにおいて中国科学技術信息研究所（ISTIC）と連携して整備した。現行の対訳コーパスの規模は、英日対訳の標題約 2,300 万件、抄録約 1,300 万件で、中日対訳は標題と抄録で約 500 万件である。

一般に、対訳コーパスの量が多いほど翻訳精度が向上するとされているが、上記文アライメント処理時に科学技術用語含有で高スコアのものから選択するなど、科学技術文献の機械翻訳に向け、対訳コーパスと用語辞書の整備を併せて行っている。

4　科学技術用語の翻訳

機械翻訳の導入にあたり、特に科学技術用語の正しい訳出に注力しており、保有する科学技術用語辞書 [3] を活用して、用語辞書に登録された語に対する予約語化処理（図 3 参照）を導入している。以下、代表的な科学技術用語として化合物名を取り上げ、その効果について述べる。

図 4a および b は化合物（有機低分子化合物）体系名を含む英日機械翻訳例であり、化合物名を予約語化対象に適用した場合としない場合で翻訳出力を示している。

原英文	Citraconic anhydride-modified ovalbumin (Cit-OVA), as model antigen, was incorporated into PHMs via electrostatic interaction, giving antigen-loaded micelles of around 150nm in size.
機械翻訳出力和文（予約語化処理有）	モデル抗原としてのシトラコン酸無水物修飾卵白アルブミン（ＣＩＴ‐ＯＶＡ）を静電相互作用によりＰＨＭに組み込み，サイズ約１５０ｎｍの抗原負荷ミセルを得た。
機械翻訳出力和文（予約語化処理無）	モデル抗原としてのクエン酸無水物修飾卵白アルブミン（ＣＩＴ‐ＯＶＡ）を静電相互作用によりＰＨＭに組み込み，サイズ約１５０ｎｍの抗原負荷ミセルを得た。

図 4a　化合物名を含む英日機械翻訳例
（予約語化処理が無いと「シトラコン酸無水物」を「クエン酸無水物」に誤る）

原英文	In this study, gel-like liquid silk (LS) obtained from the middle division of the middle silk glands of fully grown larvae of the domesticated Bombyx mori silkworm was dissolved into 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP) and electrospinning was performed, and then the crystal modification of silk fibroin (SF) in the nanofibers by the water vapor or ethanol treatment at room temperature was investigated.
機械翻訳出力和文（予約語化処理有）	本研究では，栽培されたＢｏｍｂｙｘ　ｍｏｒｉカイコの完全に成長した幼虫の中絹糸腺の中央部から得られたゲル様液体絹（ＬＳ）を１，１，１，３，３，３‐ヘキサフルオロ‐２‐プロパノール（ＨＦＩＰ）に溶解し，電気紡糸を行い，次に，室温での水蒸気またはエタノール処理によるナノ繊維中の絹フィブロイン（ＳＦ）の結晶改質を研究した。
機械翻訳出力和文（予約語化処理無）	本研究では，栽培されたＢｏｍｂｙｘ　ｍｏｒｉカイコの完全に成長した幼虫の中絹糸腺の中央部から得られたゲル様液体絹（ＬＳ）を１，１，１，３，３，３，３，３‐ヘキサフルオロ‐２‐プロパノール（ＨＦＩＰ）に溶解し，次に，室温での水蒸気またはエタノール処理による絹フィブロイン（ＳＦ）の結晶改質を研究した。

図 4b　化合物名を含む英日機械翻訳例
（予約語化処理が無いと「3，3」の過剰な出力がみられる）

翻訳出力の比較から、予約語化対象に適用しない場合には、図 4a では「シトラコン酸無水物」を「クエン酸無水物」に誤る翻訳文、図 4b では「3，3」の過剰な出力がみられる。低頻度語や記号列に対して不正確となるニューラル機械翻訳の問題[8]との関わりが考えられるが、対応する化合物名を予約語化処理することで正しい訳出となることが図から確認できる。

なお化合物体系名については、日本化学物質辞書[9]を用いることにより IUPAC 命名規則に準拠して予約語化の処理を行っている。

新語への対応をはじめ、訳質維持改善に向けて対訳コーパスの拡充に取り組んでいるが、ここで取り上げた化合物名については、日本化学物質辞書において新規登録されたもの（2018 年で例えば「1,5 - dibutoxynaphthalene - 2,6 - dicarbaldehyde」や「(2E) -3- (4-chlorophenyl) -N- (4-oxocyclohexyl) prop-2-enamide」）を基に、それらを含むようなコーパスから効率的に収集するなど、用語辞書の更新と併せた対応に取り組んでいる。

5 おわりに

文献情報データベースは、これまでの書誌、抄録、索引情報から全文データ活用の方向性にある。また、科学技術戦略や研究テーマの策定、研究パートナーの探索、研究成果の評価など、科学技術文献情報の利用ニーズはこれまでの文献検索から付加価値利用に拡大している。

外国語文献を幅広く大規模に扱う重要性が増して機械翻訳の役割がますます高まる状況にある中で、構築整備している科学技術用語辞書の活用展開を基に、機械翻訳の効果的導入の取り組みを進めているところである。

参考文献

1） J-GLOBAL：http://jglobal.jst.go.jp/

2） JDream Ⅲ：http://jdream3.com/，提供は株式会社ジー・サーチ

3） 川村隆浩，渡邊勝太郎，松邑勝治，櫛田達矢，古崎晃司：JST 科学技術用語シソーラスの Linked Data 化：科学技術情報をリンクする知識インフラの構築に向けて，情報管理，59 巻，12 号（2017）pp.839-848

4） 日中・中日機械翻訳実用化プロジェクト：https://jipsti.jst.go.jp/jazh_zhja_mt/

5） KyotoNMT：https://github.com/fabiencro/knmt/

6） WAT 2016：http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2016/, Proceedings of the 3rd Workshop on Asian Translation（WAT 2016）, December 11-16, 2016, Osaka, Japan

7） Antoine Bourlon, Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi: Simultaneous Sentence Boundary Detection and Alignment with Pivot-based Machine Translation Generated Lexicons, Proceedings of the 10th Conference on International Language Resources and Evaluation（LREC2016）, Portoroz, Slovenia（2016.5）

8） 本間奨：機械翻訳の近未来 第 6 回 NMT の今後，日本翻訳ジャーナル，No.289（2017 年 5 月 /6 月号）

9） 木村考宏，櫛田達矢：日化辞 RDF データの公開と化合物情報の統合，情報管理，58 巻，3 号（2015）pp.204-212