3.6.2 先進的翻訳技術研究室

室長（兼務）　隅田 英一郎　ほか17名

自動翻訳技術の研究・開発と多言語・多分野での社会実装

■概要

自動翻訳の多言語化、多分野化技術を研究開発しつつ、並行して大規模な対訳データを収集し、多様な言語、多様な分野に対応した高精度の自動翻訳システムを構築する。特に、訪日外国人観光客の急増に対応するため、生活一般での利活用を目的として、10言語（日本語、英語、中国語、韓国語、タイ語、インドネシア語、ベトナム語、ミャンマー語、スペイン語、フランス語）に関して、旅行、医療、防災等の分野に対応した実用レベルの音声翻訳システムの社会実装を目指した研究開発を実施している。

一方、2020年以降の世界を見据えた研究開発として、翻訳処理の漸次化等同時通訳システムの基盤技術を確立するための基礎技術の研究開発を行う。また、自動翻訳システムの汎用化を妨げている対訳データ依存性を最小化するため、同一分野の対訳でない異言語データを利活用する技術と同義異形の表現を相互に変換する技術の研究開発を進めている。

■平成 30 年度の成果

1．自動翻訳技術（自ら研究）

旅行、医療、防災等の10言語の話し言葉の対訳コーパスを目標300万文を大きく上回る454万文（日本語、英語、中国語、韓国語各11万文、タイ語38万文、インドネシア語、ベトナム語、ミャンマー語各80万文、スペイン語、フランス語各66万文）拡張した。本コーパスは次年度で開発を完了する見込みで順調に整備が進んでいる（図１）。さらに、下記の２委託研究それぞれでの話し言葉で58万文、書き言葉で140万文及び翻訳バンクでの書き言葉の対訳データ収集を併せて1,000万文を越える対訳データ増強を実現した。対訳コーパスを多分野で大規模に構築したことは、話し言葉と書き言葉の双方の自動翻訳技術の高精度化を加速するために、要諦の成果である。

前年度のニューラル機械翻訳（NMT）化は日英双方向のみであったが、本年度は、上記対訳コーパスを用いて、多言語化を達成し、技術移転をした。実際に自動翻訳システムを多言語化・多分野化できたことは2020年に向けて社会実装を加速するうえで特筆に値する。

汎用の対訳データを収集する活動「翻訳バンク」を進め、製薬会社をはじめとする民間会社や中央官庁より汎用の対訳データを取得した。さらに、適応技術によって、高精度な製薬分野向け専門システムを構築し技術移転した（図２）。「翻訳バンク」はポジティブなスパイラルに成長していくと期待される類例をみない画期的な活動であり、次年度以降の発展が大いに期待される。

NMTについて、並列化に関する新技術を創案し、特許出願した。ニューラルネットの学習の高速化は焦眉の課題であり、効率よい方法を創出できたことの実装上の意義は極めて大きい。

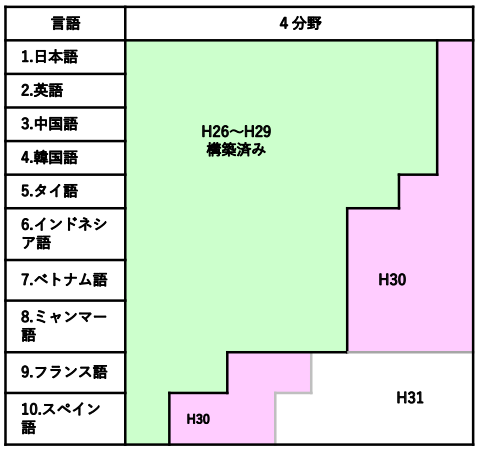
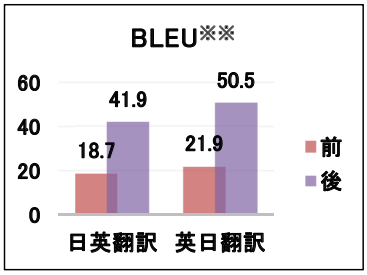


図1　世界最大規模の話し言葉の対訳データを平成31年度に完成予定で順調に整備



※※ BLEUとは、複数のプロ翻訳者が予め作成した訳文と自動翻訳の訳文との類似度であり、自動翻訳の品質評価のため広く利用される。類似度が大きいほど良い訳と判断される。

図2　翻訳バンクへの対訳拠出による製薬分野での精度向上・工期短縮の効果が証明され、通常業務に導入（アストラゼネカ）

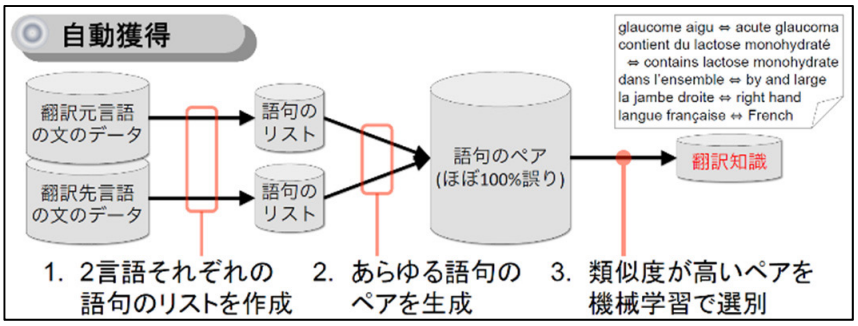


図3　コンパラブルコーパスからの対訳語抽出法

対訳データ依存性を最小化するため、同一分野の対訳でない異言語データを利活用する技術のひとつとして、「対訳でない２つの単言語コーパスと小規模の対訳データ」から構築した対訳辞書を種として用いることによって文単位の翻訳を実現する技術を開発した（図３）。分野や言語が多様である場合に対訳が十分そろわない状況は頻繁に起こるので、その解法を拡張できたことの意義は大変大きい。

2020年以降の世界を見据えて同時通訳システムの基盤技術を確立するため、昨年度ニューラルネット化した同時通訳プロトタイプシステムに関して、本年度は、五月雨式音声単語入力列から翻訳単位としての文章を切り出すモジュールを高精度化し、デモのブラッシュアップを行った。

２．委託研究No.180「自治体向け音声翻訳システムに関する研究開発」

外国人対応の多い自治体窓口のニーズを検討し、自治体で必要とされる対訳コーパスや音声データを収集し実証実験を行いながら、自治体窓口向け音声翻訳システムの社会実装を目指す委託研究である。訪日外国人観光客など一時滞在の外国人の急増に加えて、長期滞在や定住の外国人も今後急増する見込みである。

外国人材の受入れ・共生のための総合的対応策（外国人材の受入れ・共生の関する関係閣僚会議決定、平成30年12月25日）の中でも『多言語自動音声翻訳の利用促進の観点も踏まえ、「多文化共生総合相談ワンストップセンター（仮）」をはじめ、外国人と接する機会の多い行政機関の相談窓口においては、自動翻訳アプリ等を活用しながら、外国人の相談ニーズに適切に対応できる多言語対応を進める。』と明記されており、本委託研究の重要性は増すばかりである。

当年度は以下の研究開発を行った。

・子育て・年金コーパス（中国語15万文）、住民登録・国保コーパス（中国語３万文、ブラジル・韓国・タイ・インドネシア・ミャンマー語各８万文）を構築した（計58万文）。
・新たに自治体用語669語を収集し、合計5,112語の日本・英・中国・ベトナム・ブラジル語の発音付対訳辞書を作成した。
・全国12の市役所と連携して本システムを実際の業務で利用してもらうことにより、本システムの効用と、課題、要望を調査した結果にもとづいて、社会実装に向けた改善を行った。
・海外市場への応用展開の可能性を調査するため、ベトナム語対応の実証実験用のアプリNhaTra開発・無料公開し、訪日外国人の送り出し機関を対象に当該アプリの利活用状況に関する調査を実施した。
・音声翻訳システムを自治体以外の公的機関（教育委員会、消防署等）の窓口業務に展開する可能性について調査を行った。

実用研究を着実に推進すると同時に想定顧客である自治体との連携関係構築を進めており、次年度にはビジネス化が十分期待される。

３．委託研究No.197「深層学習によるマルチモーダル文脈理解と機械翻訳の高度化」

文脈やマルチモーダルの研究において、NICTにとっては自ら研究における基礎研究を補完して、外部にとっては実用研究に向けて加速できるという点で、シナジー効果を生み進歩を加速すると期待される。

初年度である本年度において次年度以降への準備として、以下の研究開発を行い、論文発表11件を実施した。

・対話対訳コーパス（4.5万文）、共参照アノテーション付き翻訳コーパス（2.8万文）、会議対話翻訳コーパス（３万文）、新聞記事対訳コーパス（20万文）を構築した。新聞記事対訳データ140万文対を収集した。
・記事からの見出しを生成する新しいアルゴリズムを提案した。
・文脈を考慮した機械翻訳の評価実験を行った。