3.6.2 先進的翻訳技術研究室

室長（兼務）　隅田 英一郎　ほか15名

グローバルコミュニケーション計画に向けた翻訳技術の研究開発

■概要

当研究室では、東京2020オリンピック・パラリンピック競技大会に向けた自動翻訳技術として、１．多分野・10言語の話し言葉の対訳コーパスの拡張、国内に散在する対訳データを集積する「翻訳バンク」の創設、２．日英双方向翻訳についてニューラル翻訳（NMT）を実装し大幅に精度向上、３．NMT向けのオープンソースコードprimitivを公開。また、2020年以降の世界を見据えた自動翻訳技術として、４．同時通訳プロトタイプシステムを完全ニューラルネット化、５．「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」から対訳辞書を構築する手法の提案を行った。

■平成29年度の成果

１．東京2020オリンピック・パラリンピック競技大会に向けた自動翻訳技術

（1）医療分野をはじめとする多分野において10言語の話し言葉の対訳コーパスを、目標100万文を大きく上回って拡張（日本語、英語、中国語、韓国語各25万文、タイ語86万文、インドネシア語、ベトナム語、ミャンマー語各41万文、フランス語35万文）した。特に、実証実験等からニーズが急速に高まっているタイ語を重点的に拡張した（図１）。蓄積・整備された全対訳コーパスを利用して、翻訳ステムを構築し、全言語において順調な精度向上を確認した。このように、多言語翻訳システム開発の重要な基盤である対訳コーパスの構築を、計画を上回るペースで進めたことは、音声翻訳技術の社会実装を加速する重要な成果である。

また、総務省と連携して国内に散在する対訳データをNICTに集積する仕組「翻訳バンク」を立ち上げた。全国から多分野の対訳をWEBからのアップロードをはじめとする様々な手段で効率的に収集可能となり、汎用の自動翻訳の高精度化が期待される（図２）。「翻訳バンク」は寄付ベースの新たな収集法の創出と言える。さらに、提供データをNICTの知財のライセンス料の算定の際に勘案して負担を軽減することが可能とした。これらが奏功して提供が加速している。NICTを中心として大規模データが集積され、これに基づいて汎用の高精度の自動翻訳システムが現出することが期待される。

（2）半年という短期間に日英双方向翻訳についてニューラル翻訳（NMT）を実装し、各分野で20％前後の精度向上（図３）を確認し、技術移転した。今後、

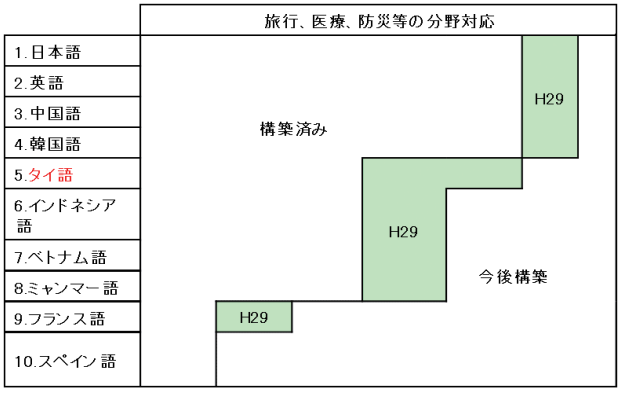

図1　対訳コーパスの拡張

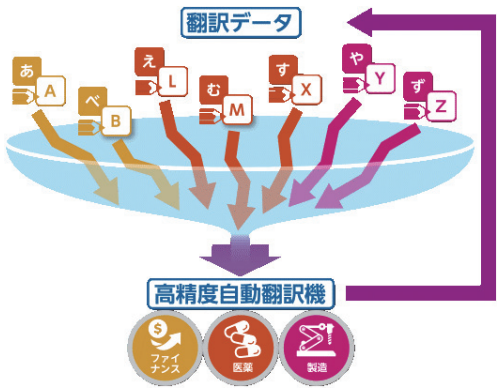

図2　『翻訳バンク』のコンセプト

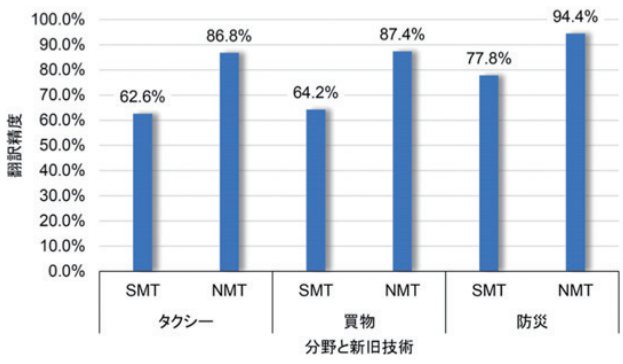

図3　NMTによる精度向上（従来技術のSMTから約20%改善）

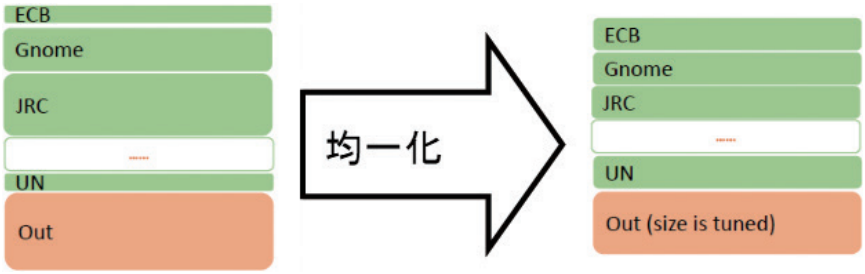

図4　多分野同時適応による高精度化

順次多言語化を進めていく。NMTの従来の適応は分野ごとにする必要があり多数のNMTと学習し稼働させなくてはいけないという課題があったが、単一NMTで多分野同時高精度化する手法を発明し特許出願した（図４）。NMTについて極めて短い期間でDEPLOY（展開）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。また、適応方法に関する前記の発明は多分野のシステムを短時間に、しかも省メモリで稼働させることができ、社会実装を加速するうえで特筆に値する。

（3）NMTの研究開発はオープンソースを活用することが多く、このことがその改良を加速している。NICTが中心となって開発した新しいNMT向けのオープンソースコードprimitivを公開した。この基盤をNICT内に構築したことの意義は今後の研究開発を加速すると期待でき意義が大きい。

2．2020年以降の世界を見据えた自動翻訳技術

（1）平成28年度に構築した同時通訳プロトタイプシステムを完全にニューラルネット化した。平成32年以降の実用化を目指しているところの同時通訳プロトタイプシステムを全面的にニューラルネットに変換したことは、今後の研究開発の基盤を確立できたことになり意義が大きい。

（2）「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」から対訳辞書を構築する手法を提唱した。「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」からの対訳知識を抽出技術は、対訳が十分揃わない状況は頻繁に起こるので、その基礎的な解決策を提示できたことの意義は大変大きい。

3．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

図5　窓口業務用のユーザーインターフェース

の社会実装をめざす委託研究である。

当年度は以下の研究開発を行った。

- 子育て／年金コーパス（日英15万文、日越23万文）、住民登録・国保コーパス（日越23万文、日中５万文）を構築した。
- 自治体用語（日本語）4,443語を収集し、英語、ブラジルポルトガル語に翻訳し、この対訳辞書に発音を付した。
- ロールプレイによる窓口対話実験を行い、これに基づいて要件定義を行い、設計・開発を経て、窓口業務用のユーザーインターフェース（図５）を実装した音声翻訳アプリケーションを開発し、板橋区、前橋市、綾瀬市において実証実験を行った。
- ビジネス化に向けた可能性検討の一環として、外国人比率２％超の自治体約250団体を対象に多言語化取組状況のアンケート調査（回収率45％）を実施し分析した。さらに、自治体への普及啓発活動として、報道発表４件、展示会出展７件を実施した。

実用に向けた研究開発を着実に推進すると同時に市場調査及び想定顧客である自治体との連携関係構築を進めており、研究開発からビジネスへのスムースな移行が期待される。