3.5.2　ユニバーサルコミュニケーション研究所　多言語翻訳研究室

室長（兼務）　隅田英一郎　ほか 3 名

多言語翻訳システムの実現に必要なアルゴリズムと対訳データの構築法の研究開発

【概　要】

ユニバーサル音声・言語コミュニケーション技術の研究開発、すなわち「コミュニケーションのグローバル化が進む中、言語・文化にかかわらず、また、システムの介在を意識することなく、いつでも、どこでも、だれもが必要な情報に容易にアクセスして、その内容を分析し、互いの円滑なコミュニケーションを可能とするため、音声・言語コミュニケーション技術の研究開発及び実証実験を行うとともに、研究開発成果のデモンストレーションを実施することにより、アジア諸国における成果の活用促進及び言語基盤の強化に貢献する」という中長期目標の下に、本研究室は、日本語と外国語の間の翻訳を主たる対象として研究を進めている。対訳データ（原文と訳文の対を集積したもの）に基づいて翻訳する手法を採用し、同手法の基盤になる大規模な対訳データを構築し、特定分野専用の高精度の自動翻訳システムを実現してきている。また、2020 年までに多言語音声翻訳の社会実装を目指す総務省の『グローバルコミュニケーション（GC）計画 *1』を推進している。

【平成 27 年度の成果】

●年度計画に対して、以下の通り目標を達成した。

【音声翻訳】観光向け音声翻訳「VoiceTra」を復活した *2。

【短文】の自動翻訳の多言語化・多分野化のため下記の研究を実施した。

■《多言語化》

翻訳の要素技術や言語資源の多言語化を進めた。

① 「対訳関係を利用して目的言語 A の文法知識を原言語 B の文法知識に変換する」提案手法（図 1）によって、原言語の文法解析が存在しない場合でも原言語の文法解析を推定し、構文解析を利用して事前語順変更（Pre-ordering）と訳語選択のためのモデルを対訳コーパスから学習する事前語順変更型構文利用統計翻訳と組み合わせて、多言語（ドイツ語、フランス語、スペイン語、ポルトガル語、韓国語、タイ語、インドネシア語、ベトナム語、アラビア語⇒英語）で自動翻訳を実現した。

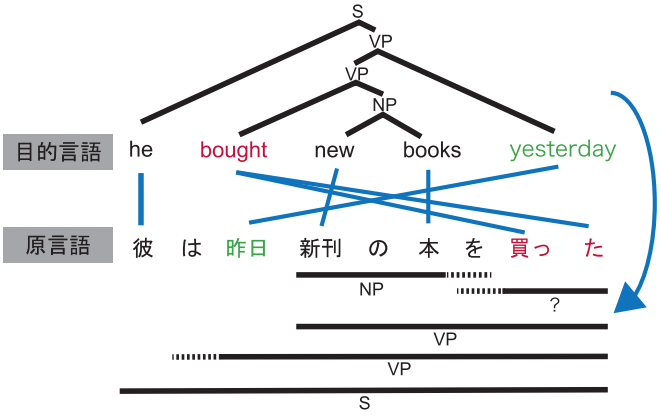

図 1　文法知識の変換

② 構文解析技術が未開発のミャンマー（ビルマ）語と高精度の構文解析技術のある英語と日本語について、20,000 文のツリーバンクを構築した（英語と日本語は、対照研究用に用意した）。次年度以降、構文解析技術が未開発の他の言語を追加する予定。これにより、①に比べて、より高精度な多言語自動翻訳の実現を目指す。

③ 英語を仲介とする手法で、自動翻訳を多言語で実現した（日本語⇔ドイツ語、フランス語、スペイン語、ポルトガル語、タイ語、インドネシア語、ベトナム語、アラビア語）

■《多分野化》

整備した対訳コーパス（医療、災害分野を含む生活分野　40 万～ 160 万文（前記 GC 計画の 10 言語））に基づいて医療分野、災害分野の翻訳システムの構築と評価実験をした（翻訳精度を測る BLEU スコアは、医療では日英 18.08、英日 23.54、日韓 51.49、韓日 56.67、災害では日英 18.35、英日 24.74、日韓 49.34、韓日 55.49 であった。英語との翻訳精度は改良が必要で、韓国語との翻訳は使える水準といえる）。

*1 http://www.soumu.go.jp/main_content/000285578.pdf

*2 http://www.nict.go.jp/press/2015/10/22-1.html

【長文】の自動翻訳の基礎技術の研究のため下記の研究を実施した。

① 対訳依存性のない高精度化のためコンパラブルコーパス（対訳でない同分野の２言語コーパス）を利用することを目指して、セクションの対応関係も考慮したバイリンガルのトピックモデルを提唱し、これを単語の対訳抽出に適用し、Wikipedia を使った実験で 80％弱の精度を達成した。

② 漸次音声翻訳システムの試作として、音声入力を長さ制限なしで受け付け、適宜分割して翻訳の入力とする英日プロトタイプを実装し、課題抽出を行った。

●次に掲げるように年度計画の目標を大幅に上回る成果をあげた。

■総務省委託『グローバルコミュニケーション計画の推進 ―多言語音声翻訳技術の研究開発及び社会実証―Ⅰ.多言語音声翻訳技術の研究開発』を受託し推進した。

■特許、マニュアル等の TEXT 翻訳システム実用化のため、世界最大の日本語との対訳コーパスを構築し、これを用いて高精度の自動翻訳システムを実装（特許用システムは特許庁審査官より極めて高い評価を受けた）。

■パターン利用翻訳と事前語順変更型構文利用統計翻訳を統合した手法を提案し、（長年の課題であった）特許請求項を対象とした高精度翻訳システム（図 2）を世界で初めて構築し、WEB サービスの形で実証実験として公開した（https://mt-auto-minhon-mlt.ucri.jgn-x.jp/）。

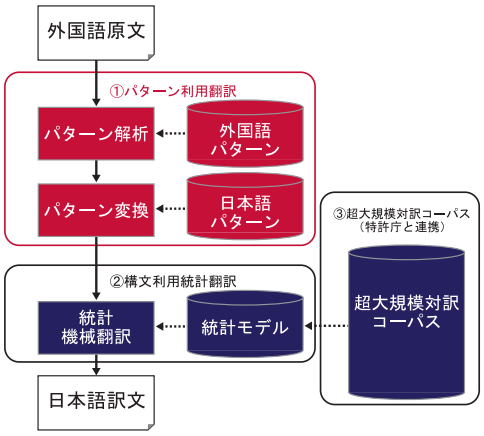

図 2　特許請求項を対象とした高精度翻訳システム

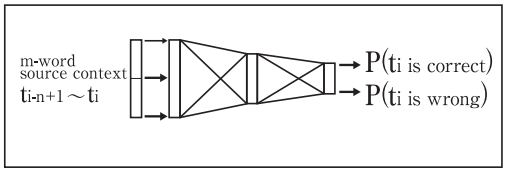

図 3　BNNJM

■自然言語処理や翻訳の要素技術のニューラルネット化　DNN（深層ニューラルネット）の構文解析や自動翻訳の全要素技術への適用を進め、STATE-OF-THE-ART を上回る成果をあげた。例えば、MS（マイクロソフト社）の SKYPE 翻訳で採用されている NNJM を高速化する BNNJM[*3] を提唱した（図３）。

●社会還元のまとめ

- 2014 年 6 月より公開している自動翻訳活用サイト「みんなの自動翻訳@ TexTra」（高精度の自動翻訳システムと同カスタマイズツール等を提供）は、利用者数が 950 人、日英対訳文数が 506,406 文に達した。
- 翻訳支援、自動翻訳技術のライセンス供与：商用ライセンスを ATR-Trek、凸版印刷、日本特許情報機構（JAPIO）、科学技術振興機構（JST）などに提供。複数社から、毎年ライセンス料の納付がある。
- 新たにニューラルネットに基づく自動翻訳のプログラム agtarbidir: agreement on target-bidirectional LSTMs for sequence-to-sequence learning を公開し、自動翻訳関係のオープンソースは下記と合わせて、6 件になった。
  - ◇ cicada: a hypergraph-based machine translation toolkit which supports {string, tree}-to-{string, tree} model
  - ◇ expgram: yet-another ngram toolkit with succinct storage
  - ◇ pialign: phrasal ITG aligner for phrase table induction
  - ◇ lader: latent derivation reorder for pre-reordering of MT input
  - ◇ trance: a transition-based neural network constituent parser
- 特許庁と共同で開発した世界最大の超大規模対訳コーパス（英日 347,950,000 文、韓日 83,460,000 文、中日 132,850,000 文）を ALAGIN で研究向けに公開した。

● JST、京都大学と連携して創設したアジア言語の自動翻訳にかかわるコンペ型国際会議 Workshop on Asian Translation（WAT）を成功裏に開催し、第３回を国際会議 COLING のワークショップとして提案した。

● 自動翻訳にニューラルネットを活用する手法の研究を開始し、ACL（Annual Meeting of the Association for Computational Linguistics）や EMNLP（Empirical Methods in Natural Language Processing）などの最難関国際会議での採録数を 12 件まで伸ばした。さらに、言語処理学会、情報処理学会、電子情報通信学会、「マルチメディア、分散、協調とモバイル（DICOMO2015）シンポジウム」で、合計 4 件の論文賞を受賞した。

[*3] Zhang, Utiyama, Sumita, Neubig, Nakamura: A Binarized Neural Network Joint Model for Machine Translation. EMNLP 2015.