3.5.3　知識創成コミュニケーション研究センター　言語翻訳グループ

グループリーダー　　隅田英一郎　ほか10名

言語翻訳の研究開発

概　要

本研究センターでは、言語・文化・能力などの壁を越えて自由にコミュニケーションが行える環境を実現するためのユニバーサルコミュニケーション技術の研究開発を行っている。本グループは、特に、人と人との言葉の壁、人とコンピュータの言葉の壁を越えるスーパーコミュニケーション技術の中核である翻訳技術の研究を実施。また、翻訳技術は音声・言語技術のMASTAR(Multi-lingual Advanced Speech and Text Research)プロジェクト、総合科学技術会議の社会還元加速プロジェクトの一つに選定されたネットワーク音声翻訳プロジェクト、「高度言語情報融合フォーラム（ALAGIN Forum: Advanced Language Information Forum)」の必須要素でもある。

第２期中期計画では、次の目標をもって推進している。すなわち、日本と世界の間にある言語の壁の克服に向けて、多言語・多分野の高精度翻訳システムを社会へ提供すること。より、詳細には、①(NICTの言語基盤グループと共同で)対訳コーパス構築の自動化やコミュニティとの協業によって1,000万文の対訳コーパスを構築すること、②この対訳コーパスを用いて融合型翻訳技術によって高精度翻訳を実現することを目指している。

上記目標の下、次の概要の研究開発を実施。多言語を対象とした高性能のコーパスベース翻訳技術の確立・普及・展開を行う。①大規模対訳コーパスを効率的に構築するための自動的手法とWeb2.0的手法を確立し、②話題・分野などへの適応法や複数翻訳融合法をはじめとする翻訳のためのアルゴリズムを研究し、さらに、アジア言語の言語資源を開発し公開する。

①の例として、翻訳支援サイトの構築を説明する。東京大学、㈱三省堂の協力の下、多言語情報流通を促進する総合的翻訳者支援サイト・翻訳情報発信サイト「みんなの翻訳」を共同で開発した（図１)。「みんなの翻訳」では、高品質辞書とウェブ上の多様な情報源をシームレスに活用できる翻訳支援エディタQReditと、翻訳コミュニティ支援と翻訳情報発信基盤、翻訳メモリ共有といった翻訳者支援及び翻訳情報共有の基盤メカニズムを組み合わせて、高度な翻訳支援機能により翻訳者を支援する翻訳情報発信サイトとして実現した。「みんなの翻訳」は、クリエイティブ・コモンズ・ライセンスの考えに基づき、翻訳情報を共有することで、近年爆発的に活発になっているオンライン個人翻訳者の翻訳、NPO/NGOによる翻訳の効率改善と発展を促す。また、共有された翻訳情報を利用して、機械翻訳の品質を改良できる。

図１　翻訳支援サイト「みんなの翻訳」

②の例として、「Dynamic Model Interpolation for Statistical Machine Translation (ACL08-SMT, pp. 208-215, 2008)」について紹介する。

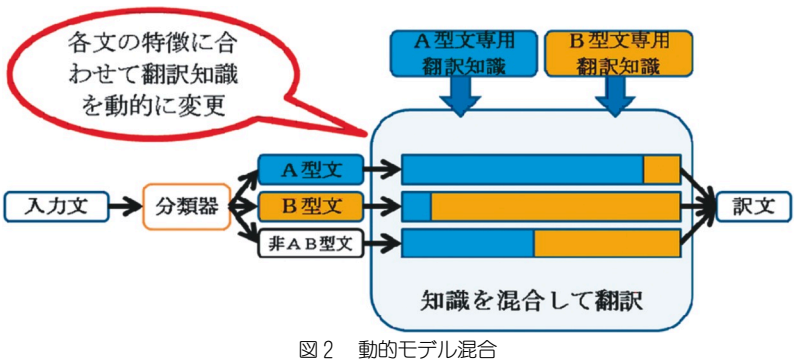

図２　動的モデル混合

NICTでは、多言語を対象として実用的な翻訳システムを効率的に開発するために、多言語処理を考慮した手法、日英翻訳など文法が異なる言語対の翻訳を扱う手法、対訳コーパスの増量ではカバーできない固有名詞の処理など様々な改良を行った。まず、音声翻訳に必要な統計的なモデルをあらかじめクラスタリングした分野毎のデータに基づいて構築し、推定される分野に適応してモデルを選択して翻訳する方式を実現した。さらに、本論文で、クラスに依存した動的モデル混合によるコーパスベース翻訳の新しい実行方式を提案した。本方式では、入力の特性

に合わせて動的に制御可能なモデルの確率的混合を実現する（図２）。異なるクラスのモデルを独立に構築し、分類器によって入力のクラスを判定し、動的に重みを変更する。アジア言語、ヨーロッパ言語を含むさまざまなバリエーションの言語対を翻訳する実験によって、広く有効性を確認した。これらによって、多言語の旅行会話の音声翻訳の精度を大きく改善できた。

平成20年度の成果

本年度のトピックスを下記に列挙する。

○まず、グループ単独の成果としては以下がある。

(1)　新しい分野、マニュアル（Linux Japanese FAQ）、標準化文書の研究を企画し、対訳コーパス収集を進め翻訳システムの構築を実施。多分野翻訳の実現可能性を実証。

(2)　「みんなの翻訳」をはじめとするWeb2.0的対訳コーパス収集手法を提案・実装し、翻訳者コミュニティとの連携という新しいパラダイムの実現可能性を実証。

(3)　旅行会話の分野において18言語の全ての組合せである306通りの翻訳システムを、コーパスベース方式で直接構築し、実用に耐える高品質を確認し、多言語高品質翻訳の実現可能性を実証（図３）。

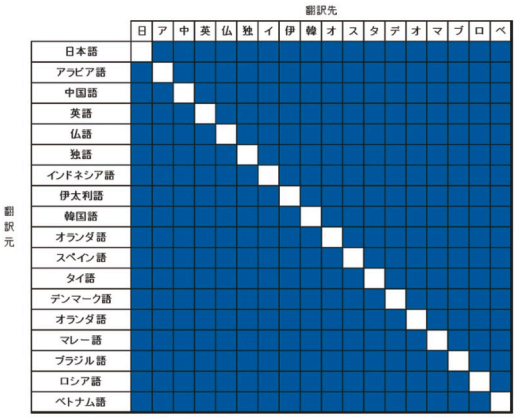

図３　18言語の全ての組合せの翻訳

○さらに、他のグループと共同の成果として以下がある。

(4)　NICT内の言語基盤グループと共同で、合計150万文対の用例ベースを構築した。これにより平成19年度までの成果と合わせて日本語に関しては前例を見ない合計584万文対の用例ベースを構築。

(5)　科学技術振興調整費により、言語基盤グループと共同で推進する日中論文翻訳の中間報告において、「(1)総合評価（所期の計画と同等の取組が行われている）、(2)今後の進め方（計画をさらに発展させるべきである）」と高い評価を得た。

(6)　音声コミュニケーショングループと共同で、音声翻訳に関する国際会議IWSLTを開催。多言語対訳コーパスBTECを提供して音声翻訳技術の比較を目的として運営している。参加・参照が年々増加しており、標準的な会議として認知されている。例えば、科学技術関係の世界最大の出版社Elsevierが運営するhttp://www.scirus.comでのキーワードのHIT件数で見ると、国際的認知度が年々上昇している（図４）。

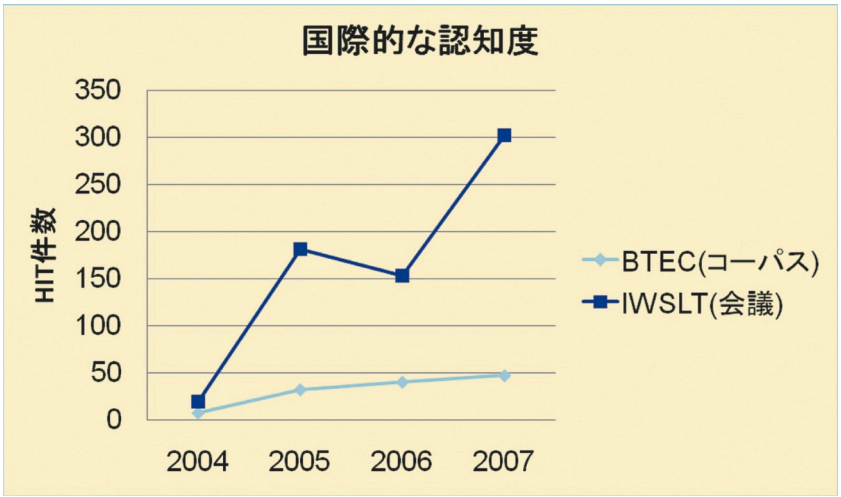

図４　音声翻訳国際会議IWSLTの世界的広がり

(7)　MASTARプロジェクト：キックオフシンポジウムを開催。

社会還元加速プロジェクト：ネットワーク音声翻訳が認定。

北京五輪音声翻訳モニター実験：北京五輪旅行時に、データ収集、サービス満足度アンケートなどのモニター実験を行った。

ユビキタス特区：総務省「ユビキタス特区」事業に採択。京都太秦地区で音声翻訳の実証実験を実施。

アジア音声翻訳標準化：アジア太平洋電気通信標準化機関（ASTAP）及びアジア音声翻訳コンソーシアム（A-STAR）においてネットワークを利用した分散型音声翻訳の接続標準化活動を継続して行った。

高度言語情報融合フォーラム総会：平成21年３月25日大手町サンケイプラザ（東京）で開催。