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　　隅田英一郎　ほか 10 名

多言語翻訳システムの構築に必要な対訳データと翻訳アルゴリズムの研究開発

【概　要】

本研究室は、人と人との言葉の壁を克服するため、日本語と英語のような異なる言語間の翻訳の研究を実施している。特に、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、自動化やコミュニティとの協業など新たな手法によって対訳データの構築を効率化し、同手法の基盤になる大規模な対訳データを構築した。さらに、この対訳データを用いて旅行分野において高精度翻訳を実現した。

また、音声コミュニケーション研究室、情報分析研究室と連携して、音声翻訳を研究する MASTAR（Multilingual Advanced Speech and Text Research）プロジェクトを実施しており、これは同時に総合科学技術会議の社会還元加速プロジェクトの 1 つに選定されている。さらに、高度言語情報融合フォーラム（ALAGIN Forum: Advanced Language Information Forum）を通じて、研究成果の社会還元も行っている。

【平成 23 年度の成果】

基礎技術

固有名詞の総数は多いのでその全てを辞書に登録することはコストを考慮すると困難であり、かつ、固有名詞は日々自由に生産・利用されるので、辞書にない語の翻訳（翻字と呼ばれる）の高精度化が求められていた。このため、コンパクトなモデル化手法を提案するとともに、当該分野の世界のトップの学会である計算言語学会（ACL）主催の国際コンペ NEWS において 15 言語対中 12 対で 1 位という好成績を達成した。

長文翻訳に関して、特に文長が長くなる特許文を対象として、①文分割法：長文を表層の特徴によって分割し翻訳結果を統合する手法と、②名詞句カプセル化法：名詞句をカプセル化し、文を短縮して翻訳、名詞句の翻訳を埋め戻す手法を創出し、これらを併用して、大幅な性能改善を実現した。

構文利用のアラインメント手法を提案し、モデルをコンパクトにできることを実証し、同プログラムを公開した。また、構文利用の翻訳混合アルゴリズムを提案し、従来法と同等の性能を、少ない空間計算量で実現できることを示した。

音声翻訳技術

技術開発

対訳コーパスの自律成長的学習技術の高度化の一環として、音声翻訳実行の履歴に基づく音声翻訳ソフトウェアの精度改善の研究を進め、正解データを人手で作成するとコストがかさむのでこれを避けて、自動評価によって履歴データの取捨選択を行った上で作成したモデルを線形補間で適応する手法を提案し、実証実験の履歴データによって有効性を確認した。さらに、旅行分野で必要となる固有名詞や一般用語を増強することを中心に、日英、日中、日韓対訳辞書を約 20 万語に大語彙化し、NICT の 1 端末で利用する音声翻訳実証実験ソフトウェア VoiceTra に組み込んで公開した。

標準化と国際連携 U-STAR

NICT の提案の ITU－T の標準勧告に基づいたプロトコル MCML を実装し、多数端末で利用できる音声翻訳実証実験ソフトウェア ChaTra に組み込んで公開した。さらに、同標準勧告を普及し多言語音声翻訳を効率的に実現するために、アジア・ヨーロッパを中心とした約 20カ国の代表的な音声・言語処理の研究機関との協力体制を U-STAR として構築し、そのリーダーとして研究協力を推進、全世界規模での音声翻訳の実証実験を計画している。

図 1　NariTra のポスター

事業化

また、VoiceTra は累計 60 万ダウンロード（概算で、利用者は日本人の 200 人に 1 人、スマートフォン所有者の 33 人に 1 人に相当）を達成し、技術の見える化、ひいては事業化の引き合いに大きく貢献した。また、音声翻訳のプログラム・特許を 3 事業者に技術移転した。そのうち 1 社は、成田国際空港株式会社のサービス NariTra として事業化した（図 1）。

社会還元加速プロジェクト「言語の壁を乗り越える音声コミュニケーション技術の実現」を 5 年計画（平成 24 年度末終了）で実施していたところ、同研究計画を上回る成果を出したため、1 年前倒しで平成 23 年度末に成功裏に終了できた。これは、社会還元加速プロジェクト 6 件のうち唯一のことであり、非常に高く評価されたと言える。以下に、関連する科学技術政策担当大臣等政務三役と総合科学技術会議有識者議員との会合資料の 3. 節を引用する。「これらのことを総合的に考慮すると、普通の旅行者が、日本、英語、中国語圏でほとんど支障なく海外旅行を楽しめる環境の実現を加速するというプロジェクトの終了時の目標を、概ね達成したと考えられる。したがって、本プロジェクトは、当初のプロジェクト終了時期である平成 24 年度末を 1 年前倒しして、平成 23 年度末で終了することが適当と考える。」

TEXT 翻訳技術

技術開発

専門分野向け高精度自動翻訳システムを多分野で実現できる技術は、波及性も高く、社会経済的に我が国にとって不可欠である。例えば、特許庁の資料『国際知財戦略 （Global IP Initiative） ～国際的な知的財産のインフラ整備に向けた具体的方策～ 2011 年 7 月』が参考になるので次に引用する。「中韓文献が増大している現状を踏まえ、中・韓→日への翻訳機能を備えた外国特許文献検索システムの整備を行う必要がある。」この背景の下、長文翻訳の基礎技術の研究に注力した。

さらに、高精度翻訳システムが求められる分野として、電子通販（e コマースとも呼ぶ）を選択した。電子通販は成長産業であり、かつ海外進出が課題となっており、膨大な商品の量、商品回転の速さから自動化が必須であるにもかかわらず高品質システムが存在しなかったことが選択の理由である。

専門分野向け高精度自動翻訳システム実現のために、①翻訳支援技術による対訳の効率的構築、②対訳辞書自動構築技術による専門用語辞書の効率的構築、③構文に基づく統計翻訳技術を研究開発した。

同時に、汎用翻訳システムの構築をしつつ、効率的に多分野化を実現するための適応技術の検討を開始した。

事業化

電子通販向け高精度翻訳システムを実現して事業者に技術移転し、国内最大級アパレル電子通販のグローバルサイトで活用されている。

アカデミアでの主導性

共通の対訳データに基づくコンペ型の国際会議を主催したほか、統計翻訳に関するチュートリアル講演を行うなど、翻訳研究に関するコミュニティで主導的役割を果たした。具体的には次の 3 点を挙げることが出来る。

① 米国 CMU と欧州 BFK と協力して、音声翻訳に関する国際会議 IWSLT を主催。2004 年から毎年開催し、世界の研究機関が参加、標準的な会議として認知され、参加・参照が年々増加している。2011 年は、TED（Technology, Entertainment, Design）Talks をデータとして講演の音声翻訳技術を対象として取り上げ実施した。

② 国際会議 NTCIR の一部として特許翻訳に関する PatentMT を主催。NTCIR9 は、NTCIR7・8 の日英対訳データに加え、香港教育學院と共同で日中対訳データを追加して特許翻訳技術を比較。米国 IBM と BBN、ドイツ RWTH、フランス LIUM、に加え、中国 7 チーム、日本 7 チーム、台湾 2 チーム、韓国 1 チームで計 21 チームが参加し、2011 年末時点の最高の技術を比較した。本会議は、日本国特許庁（JPO）、欧州特許庁（EPO）、中国国家知識産権局（SIPO）などの各国政府機関からも注目されている。日英翻訳で、統計翻訳が規則翻訳より高品質を実現できたこと、中英の統計翻訳で、当該テストデータにおいて翻訳率 8 割（意味が通じる翻訳が 8 割）を実現できたことなど新たな知見を得た。

③ 渡辺太郎主任研究員が言語処理学会のチュートリアル「統計的機械翻訳の最先端」や人工知能学会論文誌の解説論文「統計的機械翻訳の現場」で、統計翻訳の普及に貢献している。