3.4.3　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

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】

本グループは、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、言語基盤グループと連携して、音声翻訳を研究する MASTAR（Multi-lingual 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 22 年度の成果】

対訳データに基づいて翻訳する手法

図 1 に示したように、対訳データから 2 言語間の対応関係をモデル化する翻訳モデル（直感的にいうと、確率付き対訳辞書である）と目的言語らしさをモデル化する言語モデル（例えば、英日翻訳の場合、日本語の単語の並びの自然さを表す確率付き日本語辞書である）を導出し、両者に基づく確率を最大化するように翻訳する（この技術を統計翻訳と呼ぶ）。

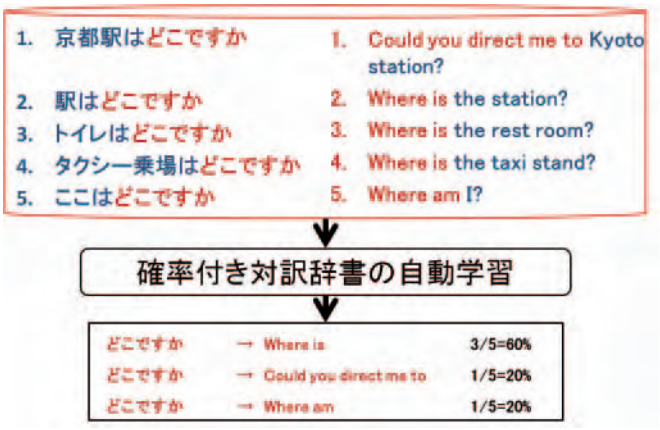

図 1　統計翻訳の概要

対訳データから翻訳システムが自動的に構築できるわけである。この手法のメリットの 1 つに多言語化の容易性がある。N 個の言語からなる多言語対訳データを用意すれば、その全ての組合せ、N（N − 1）個の翻訳システムが自動的に構築できる。我々は、旅行会話の分野で多言語対訳データ（N = 21）を構築し、420 通りの翻訳システムを実現し、実用レベルの翻訳品質（図 2）を確認した。さらに、音声認識と音声合成と組み合わせて、スマートフォン用の多言語音声翻訳アプリケーション VoiceTra として全世界に向けて公開した。

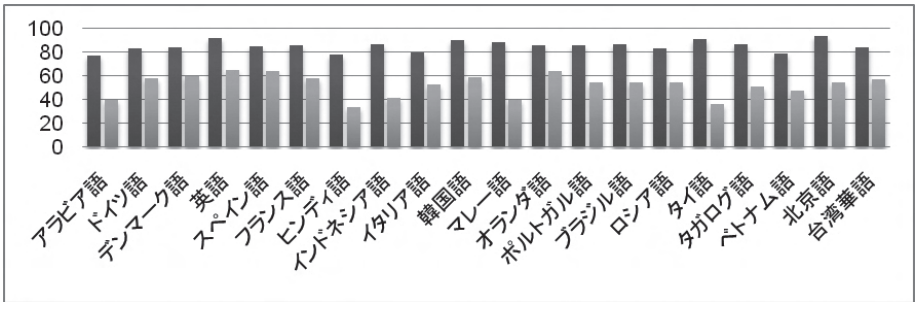

図 2　多言語翻訳での翻訳品質比較（縦軸が日本語への翻訳率、横軸が翻訳元の言語）

対訳データ構築

対訳データを効率的に収集するために、2 つの補完的なアプローチがある。(A) Web から対訳データをクローリングすることや文章レベルの対訳から自動的に文レベルで対応付けする技術などのコンピュータ中心のアプローチと (B) ボランティア翻訳のホスティング・サービスや外部機関との提携など、人や社会中心のアプローチである。NICT では、両方のアプローチを併用して精力的に対訳データを集め、第 2 期中期計画開始よりの集積で 2,800 万文を達成し、高度言語情報融合フォーラムを通じて公開を開始している。

(B) の 1 つの例として、「みんなの翻訳」(http://trans-aid.jp/) を紹介する。「みんなの翻訳」では、品質の良い

辞書や使いやすいエディターなどの翻訳のツール（図3）を公開している。利用者は原文と翻訳文を「みんなの翻訳」で蓄積・公開する。この蓄積されたデータは対訳データとして翻訳システムの構築に利用出来る。今年度、英語に加えて中国語・韓国語にも対応した。

このサイトの利用者は1,697名、登録文書数は7,294、公開文書数は2,862、対訳の英語単語数は864,547に成長した。また、NGOで採用されることが多く、Amnesty International Japan、Democracy Now! Japan、GlobalVoicesOnline Japanese teamなどのメンバーに活発に利用されている。

また、「みんなの翻訳」はアジア太平洋機械翻訳協会（AAMT）第5回長尾賞を受賞した。いわゆる学会の学術賞ではなく、たとえば、高性能の機械翻訳システムを商品化した、機械翻訳システムを使った新しいサービスを開始した、といった貢献を対象とした賞であり、「みんなの翻訳」が、社会に資するものと認められたといえる。

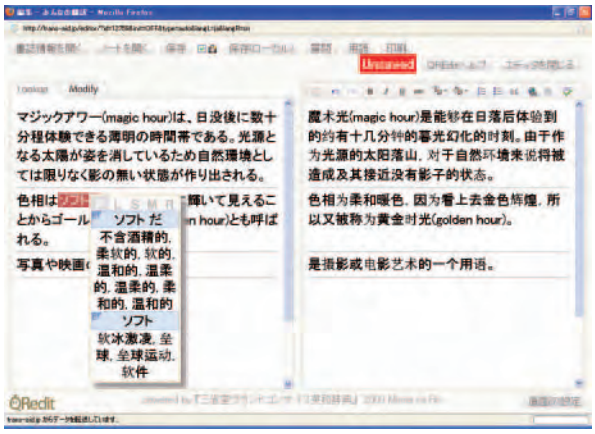

図3　翻訳支援サイト「みんなの翻訳」(日中版)

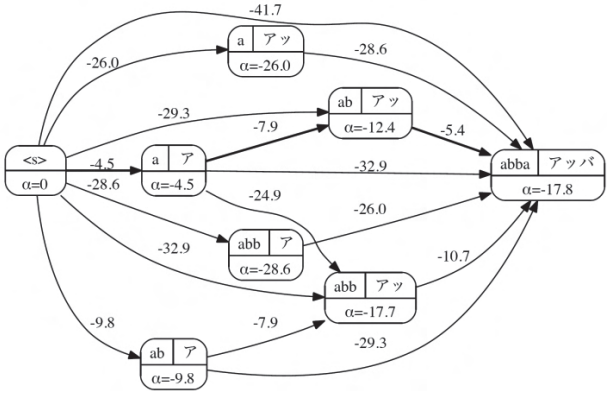

図4　日英間の翻字モデルの例

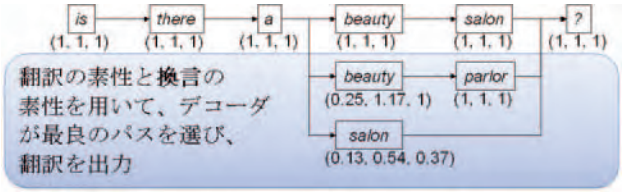

図5　換言による翻訳

翻訳技術

次の2つの新技術を創出した。

① 辞書や対訳データに現れない未知語は翻字（発音をなるべく変えずに2言語間で文字を翻訳すること）で処理することが出来る。当グループでは、ディリクレ過程を用いた新しいモデルを開発した（図4）。本手法の利点はモデルがコンパクトになることと過学習しない点である。本技術は、翻字に関する国際コンペACL/NEWS2010で8つの言語対のうち5つの言語対で1位の世最高性能を達成した。

② 入力と翻訳モデルは必ずしも一致しないことから、同じ意味でも表現が違うと翻訳できないことがある。例えば、入力の *beauty salon* に対応する翻訳モデルがなくても、入力を翻訳モデルが存在する *beauty parlor* や *salon* に置き換えたラティスを生成すれば翻訳できる（図5）。音声翻訳で成功したラティスデコーディング（音声認識の途中結果のラティスを探索する手法）を援用し、入力を同義表現で言い換えて翻訳適用範囲の拡大する手法を提案し翻訳品質を改善した。

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り、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。

② 国際会議NTCIRの一部として特許翻訳に関するPatentMTを主催。NTCIR7/2007～2008、NTCIR8/2009～2010は日英対訳データを提供して特許翻訳技術を比較。NTCIR9/2010～2011はThe Hong Kong Institute of Education（香港教育學院）と共同で日英・日中対訳データを提供して特許翻訳技術を比較、多数の機関の参加を得て、相互比較により、新たな知見を明らかにしてきた。

③ 音声研究に関するトップレベルの国際会議INTERSPEECHで渡辺太郎主任研究員が統計翻訳に関する招待講演 Foundations of Statistical Machine Translation: Past, Present and Future を行った。