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 13 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を克服するため、多言語翻訳の研究を実施している。

また、多言語翻訳技術は音声・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 1 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、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を目標としている。より、詳細には、①（言語基盤グループと共同で）対訳データ構築の自動化やコミュニティとの協業によって **1,000 万文の対訳データを構築する**こと、②この対訳データ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、③さらに、アジア言語の言語資源を開発し公開していくこと、を目指している。

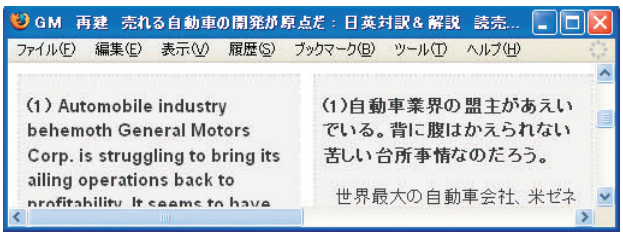

図 1　Web の同一ページ内の対訳データ

平成 21 年度の成果

①**【対訳データの構築】**対訳データ構築の自動化やコミュニティとの協業により 250 万文の対訳を構築し、別途特許に関して 1,800 万文の対訳を構築した。

- **Web 等に存在する大量の文書に対する機械学習の適用**（平成 21 年度に新たに提案・実装した、図 1 のような同一ページ内にある対訳を Web から発見し抽出するアルゴリズム及び自動文対応技術）、並びに**人手による作業**（開発した翻訳支援サイト「みんなの翻訳」（図 2）の 1,000 人を越える利用者（図 3）による翻訳作業）**の併用**により、対訳データの多分野化（旅行、論文、特許、一般）と規模拡大を実現した。これにより、**新たに 250 万文を越える規模の対訳データを構築し、平成 20 年度までの成果と合わせて合計 750 万を越える対訳データを構築した。**
- 翻訳支援サイトについて補足説明する。翻訳者支援サイト「みんなの翻訳」は東京大学と共同で開発し、平成 21 年度に公開した。　図 2 に「みんなの翻訳」の利用画面を示した。左側の窓が原文を表示し、右側の窓が訳文を表示している。原文にある下線は自動的に辞書引きされた語彙やイディオムを表す。利用者が下線にアクセスすると、辞書引き結果がポップアップし、COPY・PASTE で訳文画面に入力することが出来る。翻訳時間の

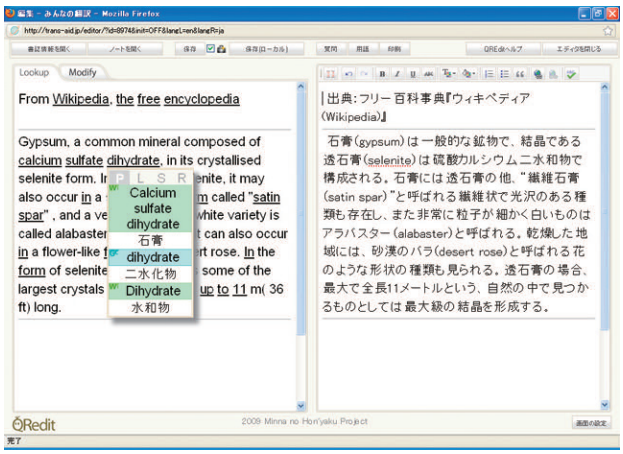

図 2　翻訳支援サイト「みんなの翻訳」

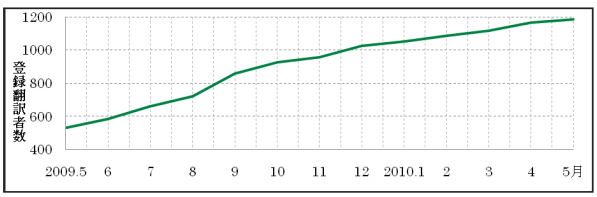

図 3　「みんなの翻訳」利用者数の増加

1/3 程度が辞書引き時間であることがわかっているので、この自動辞書引きの利用によって、翻訳作業が大幅に効率化出来る。図 3 にあるように、「みんなの翻訳」は広く一般から評価され、順調に利用者を増やしている。「みんなの翻訳」では、上述の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に加えて、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。

- これらに加えて、NICT の**自動文対応技術を用いて**、特許という特殊な分野について対訳データ 1,800 万文を構築した。これは現在他機関より公開されている対訳データの倍以上の量であり、**世界最大の規模を誇る**（図 4）。

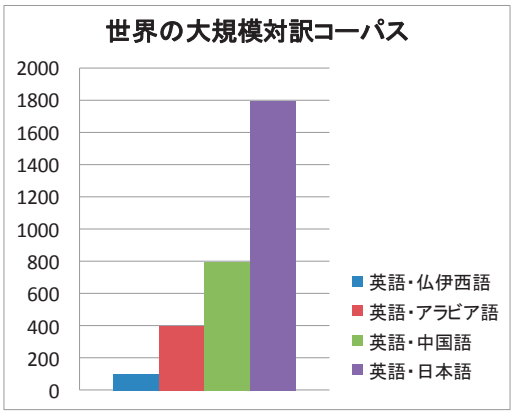

図 4　大規模対訳コーパスの比較

②**【高精度機械翻訳技術】**対訳データを用いて高精度翻訳を実現するため、平成 21 年度は、双方向翻訳技術、形態素解析の翻訳向け最適化技術（図 5）などを提案し、翻訳アルゴリズムを高度化した。

- **双方向翻訳技術**　統計翻訳技術においては、通常、文頭の単語から順方向の翻訳処理を行うが、ここに文末からの逆方向の翻訳処理を取り入れ、双方向翻訳とすることにより、翻訳品質を改善する手法を提案した。272（=17 言語× 16 言語）通りの翻訳実験で、順方向翻訳との性能比較を行うと、99%（=269/272）の割合で、双方向翻訳の性能が優るという強力な結果を得た。
- **形態素解析の翻訳向け最適化技術**　翻訳技術に必須である各言語の形態素解析プログラムは、母国語話者による研究が遅れていたり、また、よいプログラムが存在しても、種々の制約から入手困難な場合もある。また、既存のプログラムが翻訳に最適とは限らないという問題もある。そこで文字を分割の初期値とし、翻訳スコアが上昇するように単位を大きくする方向で学習する手法を提案し、図 5 にあるような様々な言語で、高精度翻訳ができる形態素解析技術を確立した（数字は翻訳スコアであり、値が大きいほど品質が高いことを表す）。

言語	サンプル	文字	学習
アラビア	نعم ، انه كذلك.	58.60	63.70
タイ	ใช่มันเป็นเช่นนั้น	44.41	55.00
ベトナム	Vâng, đúng rồi.	49.91	60.56

図 5　翻訳向け形態素解析

- **その他の技術開発**　旅行会話分野において省資源技術を用いて、メモリや処理能力が制限されるモバイル機器での実装を実現した。このシステムの翻訳品質は大手翻訳サイトのそれを大きく上回った。同様に、高性能機械翻訳技術に関して、平成 21 年度補正予算における全国 5 地域での音声翻訳の実証実験のための翻訳エンジンを開発提供した。また、音声翻訳の国際会議である IWSLT を開催し、研究フィールド全体の着実な技術の進歩に寄与した。さらに、次年度の開催に当たっては、対話に加えてスピーチも対象とすることで音声翻訳の新技術の研究開発を主導することとした。

③**【アジア言語】**「アジア言語の言語資源を開発・公開」する目的で、タイ自然言語ラボラトリーで活動した。

- 知識構築支援ツール KUI を使ってワードネット（意味辞書）の多言語化を推進し、これまでに、タイ語 80,098 語、インドネシア語 21,584 語、ラオス語 72,672 語、ベトナム語 17,767 語、韓国語 65,483 語、ミャンマー語 26,033 語、を構築。
- 知識構築支援ツール KUI やワードネットを始めとする自然言語処理に関する教育コース ADD の開催も 5 回目を迎え、ベトナム、カンボジア、ブータン、モンゴル、ラオス、ミャンマー、インドネシア、インド、パキスタン、ネパール、スリランカなどからの参加者に技術教育を実施した。