応般

［シニアコラム］

IT好き放題

［No.75］

翻訳バンク

最近のコラムの中で「好き勝手に書いてよい欄だ」という一文に遭遇し，シニア冥利と思い書きはじめました．

2020年オリンピック・パラリンピックの東京開催が決定し日本中盛り上がっています．訪日外国人も順調に増え，政府目標も大幅に上方修正されました．日本語という言葉の壁と，この壁を越える技術がクローズアップされて，グローバルコミュニケーション計画☆1という多言語の音声翻訳システムの社会実装を目指す国家プロジェクトが着々と成果を出しています．

そんな中，過日，第一級の翻訳者の方から，英語から日本語への自動翻訳システムは，かつては箸にも棒にもかからなかったけれど，最近目に見えて良くなったと伺いました．この目覚ましい品質向上の背景に技術革新があります．翻訳データ（原文と翻訳の対を大量に集積したもの）に基づく手法が発明され熾烈な研究競争が行われアルゴリズムが日進月歩し続けています．構文解析を使う事前並べ替え型の統計翻訳やLSTM☆2を使うニューラルネット翻訳で大きく翻訳品質が改善しました．

翻訳データに基づく手法には2つの柱があります．①翻訳アルゴリズムと②翻訳データです．データ量が同じならアルゴリズムが翻訳品質を決定し，逆に，アルゴリズムが同じなら翻訳データ量が品質を決定します．研究では，共通のデータを使ったベンチマークをしますので，翻訳データは与えられた条件と捉え，研究者は翻訳データの収集を軽視しがちです．一方，翻訳データの収集は自動翻訳システムの実用化には不可欠です．昨今の人工知能が深層学習のアルゴリズムとビッグデータの対で成立しているのと同じ構図です．データがないと人工知能と呼ぶところの機能は実現できません．

この翻訳品質への貢献が大きい翻訳データを作成する方法はいろいろあります．Webからクロールする，クラウドソースで翻訳する，既存データの権利者に許諾をいただく等，あらゆる方法で集められています．ここで，日本で1年間に生産されている翻訳データの全体量を考えてみましょう．翻訳産業の年間総売り上げは2,000億円と言われています．1文字10円とすると200億文字になります．1単語2文字として100億語．1文20単語とすると5億文です．これはきわめて膨大な量です．これを1カ所（『翻訳バンク』と呼びます）に集めて，『翻訳バンク』の全データで自動翻訳システムを構築すれば相当の高品質が実現できます．さらに，構築された高品質の自動翻訳システムで下訳を作成し校正することにすれば，翻訳作業全体の効率化を実現できますから，翻訳データの集積が加速されます．翻訳データを再利用するこのエコシステムが完成すれば，想像を絶する高品質な自動翻訳システムが出現するでしょう．最近，私はこの『翻訳バンク』の実装・拡張に東奔西走しているところです．すでに，中央官庁，自治体，観光や飲食にかかわる民間企業，各種メディア企業などのご協力を得ることができており，NICTに大量の翻訳データが集まり始めています．

翻訳のイノベーションは本当に起こるのでしょうか？　イノベーションのジレンマに囚われて結局何も起こらないのでしょうか？　私には，イノベーションの可能性に気付いた人々が発しているシグナルがたくさん見えています．オセロゲームのように一気に局面は変わり，この夢が2020年以前に必ず実現するでしょう．私はこのイノベーションを楽しむ予定です．皆さんもいかがですか？

（2017年1月25日受付）

隅田英一郎 Eiichiro SUMITA

国立研究開発法人 情報通信研究機構

［正会員］ eiichiro.sumita@nict.go.jp

NICTフェロー．（株）日本アイ・ビー・エム東京基礎研究所，（株）国際電気通信基礎技術研究所を経て，NICT入所．京都大学大学院博士（工学）．機械翻訳，eラーニングの研究開発に従事．

☆1 http://www.soumu.go.jp/main_content/000285578.pdf
http://www.soumu.go.jp/main_content/000395359.pdf
http://gcp.nict.go.jp/about/index.html

☆2 https://en.wikipedia.org/wiki/Long_short-term_memory#/media/File:Long_Short_Term_Memory.png