日中機械翻訳プロジェクトの技術的展望

東京大学情報理工学系研究科教授
国際言語処理学会（ACL）前会長（2006年）
アジア太平洋機械翻訳協会前会長
アジア言語処理学会連合（AFNLP）会長
辻井　潤一

PROFILE 機械翻訳、言語処理の研究に従事、現在、東京大学と英国・マンチェスター大学教授、同テキストマイニングセンター・センター長を兼任する

tsujii@is.s.u-tokyo.ac.jp　03-5841-4120

1 はじめに

アジア諸国の国際的な地位が急速に向上している。経済活動に続いて、科学技術に関する研究活動でもアジア諸国が占める位置は急速に大きくなり、北米・ヨーロッパと並んで、アジアが3極の一つになりつつある。それにともない、アジア言語からの、あるいは、アジア言語への翻訳の需要が急速にたかまっている。

実際、アメリカの機械翻訳に関する研究は、中国語とアラビア語を中核にしている。昨年度から開始された、科学技術振興調整費による日中機械翻訳システムの開発は、そういう観点からは、遅きに失しているということもできる。ただ、日本の研究プロジェクトは、アメリカ型の短期決戦的なものとことなり、中長期的な視野での技術開発を行うことから、機械翻訳のように一筋縄ではいかない技術のプロジェクトでは、むしろ良いかも知れない、とも思っている。特に、すこし無謀な枠組みでの研究だけに集中して投資し、最初は威勢が良かったのが、最近はすこし息切れ気味のアメリカの様子を見ていると、その経験から学ぶこともおおく、遅れ気味のスタートが幸いするかもしれない。

上の科学振興調整費による日中の機械翻訳プロジェクトには、情報通信研究機構（NICT）、科学技術振興機構（JST）、京都大学、静岡大学のグループとともに、東京大学の私の研究グループも関与している。この稿では、このプロジェクトの背景と技術的な方向を書いておこうと思う。

2 2つの機械翻訳の枠組み

日本では、機械翻訳技術の研究は、1980年代に一つのピークを迎える。これは、現在、国会図書館館長で前京都大学総長がはじめられたプロジェクト（MUプロジェクト）がきっかけになったもので、私も、そのプロジェクトの推進に関与した。この日本の機械翻訳がピークであったときには、米国と日本の現在の状況がちょうど逆の形で存在していた。すなわち、日本が規則主導・文法中心という枠組みの研究開発に集中投資をし、機械翻訳の研究に冷淡であった米国は、その欠点を見極めてから、10年以上遅れた形で、機械翻訳に投資を始めることになる。

すなわち、アメリカは、日本でのブームが去った90年代の後半に機械翻訳に膨大な資金と人材を投入し始める。実際、日本の規則主導型のシステムは、現実の多様な言語現象をカバーするための巨大な規則の集合をどのように作り、管理するかの方法論がなく、90年代に入る頃にはその限界が見えて来ていた。規則主導のシステムは、2つの言語間の複雑な構造の差を調整する能力は、可能的には非常に高い。ただ、その可能な能力を活用するには、2つの言語の構造の差を調整するための規則を膨大に作らなければならない、それができない、というわけである。

アメリカがはじめた枠組み、統計による機械翻訳というのは、この壁を超えようとするものであった。2つの言語のテキストさえ大量に収集すれば、人手をかりることなく、2つの言語の差を吸収する統計モデルが作れる、というわけである。

この枠組みは、計算機の能力が急速に高まったこともあり、その当初は非常にうまくいっているように見えた。大量のテキストを集めて来て、あとは、計算機をぶん回していれば、自動的に2つの言語の対応をとる統計モデルができる、というわけである。

ただ、ここ数年、その欠陥も明らかになってきた。作られる統計モデルが荒く、2つの言語に見られる構造的に複雑な対応をとることができない。作られる統計モデルがBlack Boxで、人手で改良することができない。改良するには、もっと大きなデータを与えるしかないが、あるレベルの性能に達すると、さらに性能を改良するには、指数関数的に大きなデータが必要となることも、わかってきた。

そのうえ、指数関数的に大きなデータを与えたとしても、2言語の複雑な構造対応が取れるかどうかの保証がない。

実のところ、評価の仕方にもよるが、規則主導のシステムと比べて、性能的に劣ることが認識されるようになってきた。翻訳の評価は非常に難しい問題を抱えていて、よい評価の手法を作ることが、我々のプロジェクトの研究項目にもなっている。

アメリカの研究者にとっては皮肉なことに、人間の評価にできるだけ合致するように、評価の仕方を工夫すればするほど、統計モデルによるシステムが、規則主導のシステムに劣ることがはっきりしてきている。初期のころ威勢が良かったのは、評価の仕方が荒くて、データを集めて計算機をぶん回すだけの手法と規則主導のシステムの差がうまく測れず、ごく短期間の間に、同じ性能にまでに達することができた、と錯覚したのである。

実際、評価を精密にすると、そうではないこと、また、短期間で性能が向上したのは良かったが、そこで性能は頭打ちになっていて、それ以上に良くしようとすると、大きな壁があることも分かってきた、というわけである。

3 第3の枠組み

では、この2つの枠組みの欠点を解消する第3の方法というのはあるのか、というのが次の問題になる。「ある」というのが、われわれの考えで、それを日本語・中国語で実証するのが、技術・研究面からのプロジェクトの目的になっている。

まず、第一には、規則主導のシステムの失敗原因は、規則や辞書という翻訳に必要な知識源を、すべて人手でつくろうとしたこと、であった。アメリカの統計機械翻訳の研究で判ったことは、データを大量に集めると、統計モデルの形でこの知識源の原初的なものをつくることができる、言い換えると、2つの言語間で、どのような単語や構造の対応があるかが確率付きでわかる、ということであった。

統計翻訳システムは、この対応の統計モデルをそのままBlack Boxとして、機械翻訳のエンジンに使った。これが、データから帰納的に構築される統計モデルの唯一の使い道というわけではない。構築された統計モデルを、人間が参照することで、規則主導の知識源を効率よく作ることができる。

また、集積されたテキストをそのまま統計モデルを作るための入力と考えることもない。集積されたテキストを規則主導型でまず処理して、その処理結果を統計モデルの入力とすることもできる。

あるいは、規則主導型で処理した結果を例文のデータベースと考えて、統計モデルで計算されるような対応や確率値を例文データベースからの例文検索の手がかりに使うこともできる。このような検索で、入力文と「よく似た」例文が検索できれば、例文の翻訳文をお手本として、入力文に対応する出力を合成することもできる（例文主導の翻訳システム）。

このような新たなアイデアを試してみること、それでアメリカの研究が遭遇する壁を乗り越えてみせること、これがプロジェクトの目的となっている。

4 終りに

実は、今回のプロジェクトには、前回のプロジェクトにプロジェクト要員として参加していた研究者も加わっている。20年の歳月を経て、次の挑戦ができることで、皆、わくわくしている。