機械翻訳システムの評価：客観と主観

東京大学大学院情報理工学系研究科教授
コンピュータ科学専攻
辻井　潤一

PROFILE マンチェスター大学，英国国立テキストマイニングセンター所長

tsujii@is.s.u-tokyo.ac.jp

1 はじめに

アジア太平洋機械翻訳協会(AAMT)では，Japioと協同して，特許文書の機械翻訳に関する研究会（AAMT/Japio特許翻訳研究会）を開催している．年に10回の研究会を開催しており，参加者には，日本における機械翻訳の代表的な研究者（大学，研究所から10名，機械翻訳メーカー4名）が参加している．また，この正規メンバーの以外にも，翻訳家の方や博士課程の学生がオブザーバーとして参加するなど，毎回，活発な議論があり，結構楽しい会になっている．また，本年度には，この研究会とは別であるが，国立情報学研究所の主催するNTCIR(NII Test Collection for IR System)がJapio提供の翻訳データ（公開特許公報とその英語版）を使って，特許の機械翻訳システムの評価型プロジェクトを企画するなど，特許の機械翻訳への関心が急速に高まってきている．

この特許翻訳への関心の高まりには，JAPIOが作成している和文特許の翻訳（PAJ - Patent Abstracts of Japan)が，NTCIRや我々の研究会に提供されるようになったことが，大きく寄与ししている．

近年盛んな統計的な機械翻訳では，PAJのような原文とその翻訳のデータが大量にあれば，それだけで機械翻訳システムが構築できる．このことから，JAPIOがPAJを研究用に使うことを許可してくれたことは，日英の機械翻訳を研究するグループの数を飛躍的に増やすことに貢献した．

我々，東京大学のグループでも，現在，日中の機械翻訳，特に，科学技術論文に特化した日中翻訳の研究プロジェクトを推進している．将来的には，PAJの中国語版，あるいは，中国語の特許の日本語翻訳版の作成に貢献できれば，と思っている．

本稿では，AAMT/Japio特許翻訳研究会と日中機械翻訳プロジェクトでの，翻訳結果の評価での経験を簡単にまとめておく．

2 翻訳の評価

NTCIRのような評価型プロジェクトでは，翻訳結果の評価基準が大問題となる．幾つものグループが開発したシステムに優劣をつけることになるので，公平さの担保が大きな問題となる．「評価の目的は，今後の研究に資することで，システム間に順位を付けることではない」とされるが，結局は，順位の高いシステムの翻訳方式がよいとされ，将来の研究方向を決め，アメリカの場合には，グループの研究資金に影響する．評価の方法とその公平さが，大きな関心事となる．

評価には，人間が翻訳文を読んで行う主観評価と，評価用プログラムが自動的に評価値を算出する客観評価がある．客観評価がよいように聞こえるが，大学の入試での客観評価が抱えるのと同じ問題がある．客観評価可能な問題だけでは，測れない学生の能力がある．機械翻訳の評価でも同様で，現在の客観評価法は，システムの能力を十全に評価できていないとの批判が，大きくなってきている．

ここ10年間，アメリカの研究グループは，Blueと呼

ばれる客観評価値の向上に大変な努力を注いできたので，この反省は，これまでの研究方向を見直すことにつながっている．

3 翻訳の客観評価：Blue値

科学技術，とくに，技術の研究では，効果的な性能評価法が研究を推進する．乗り物の高速化や燃料消費量の削減，半導体のサイズやスイッチング速度など，量的指標で性能が評価できる技術は，技術の進歩が明示化され，競合する技術間の優劣もつき，どの技術に重点を置くかの決定も，この評価結果に従って行われる．これに対して，「翻訳の質」という本来定性的なものを量的指標におきかえることができるのだろうか？

AAMT/Japio研究会の目的の一つも，翻訳システムの性能評価方法の確立である．実際，翻訳システムの導入やそのための辞書など，リソースの構築を行うJPAIOにおいて，投資を合理的に行うために，システムの性能評価は不可欠となる．

アメリカが中心の客観評価は，システムが出力する翻訳と人間の理想的な翻訳との近さを評価する．この近さを機械的に評価するプログラムは，いわゆる文の「意味」をつかうことはできない．与えられた文の「意味」を取り出す技術は，残念ながら，まだないから．結局，前述のBlue値と呼ばれるものは，表面上の近さを測る．たとえば，共通する単語が多いほど，2つの類似度は高い(uni-gram Similarity)とされる．これだけだと，「太郎が花子をぶった」と「花子が太郎をぶった」は，まったく同じとされる．もう少し精密に，連続する2単語の組(bi-gram)を考え，共通する2単語の組が多いほど，類似度が高いとすることもできる(Bi-gram Similarity)．Blue値とは，出力の翻訳文と人間翻訳の文との類似度を，単語のn-字組の重なり具合で計算し，それらを平均することでシステムの性能を評価するものである．

このやり方だと，「太郎が花子をぶった」と「太郎は花子に暴力をふるった」とは，類似度が低いとされ，一方の人間翻訳に対してもう一方を出力するシステムの性能は，低く評価されてしまう．なんか変である．

4 翻訳の主観評価

Blue値が，公平でない，偏向した評価であることは，多くの人が指摘してきた．ただ，最近，その欠陥が具体的に認識されるようになってきた．人間が明らかに優れていると判断する（すなわち，主観評価がよい）システムのBlue値での評価が，以外に低くなる．あるいは，もっと深刻なのは，統計翻訳方式という，特定の翻訳方式でのシステムのBlue値が相対的に高くなる．すなわち，同じBlue値を持つ他の方式（規則主導，例主導など）と比べると，主観評価が著しく低いことがわかってきたのである．

アメリカは，Blue値信仰が強く，Blue値が高いということで，統計翻訳の研究に集中的に資金を投入してきた．評価に基づく研究開発方針の決定という，これまでの技術開発に有効に機能した戦略を，機械翻訳にも適用してきた．これが，戦略上の大きな失敗につながったように見える．

翻訳の質という，「意味」に強く依存するものの評価を表面上の単語連続という，機械的に簡単に計算できる尺度に置き換えたこと，安易な効率の追求が大きな誤りにつながった．AAMT/Japio研究会では，この反省から主観評価と客観評価の相関，単語の並びではなく文の構造の評価，さらには意味を勘案した評価という正統的な方法で，翻訳システムの評価を考えている．

5 おわりに

評価の方法は，技術の方向を決める大きな要因である．アメリカの経験は，我々にとって良い反面教師になると考えている．