新しい機械翻訳自動評価基準NMGの提案

諏訪東京理科大学システム工学部電子システム工学科教授
江原　暉将

PROFILE 1967年早稲田大学理工学部電気通信学科卒。同年NHK入局、技術現業局を経て1970年より放送技術研究所に所属。情報検索、音声認識、機械翻訳の研究などに従事。2003年より現職。

eharate @ rs.suwa.tus.ac.jp

1 はじめに

機械翻訳は発展途上の技術であり、常に改良が加えられている。その改良内容は、辞書の充実であり、また構文規則の精緻化である。時には、全く新しい翻訳方式にパラダイムシフトされることもある。これらの改良が加えられると、その有効性を評価する必要がある。従来は、翻訳試験文を改良前と改良後の2つのシステムで翻訳させ、両者の翻訳結果を人手によって評価することで改良の有効性を確認してきた。しかし、このような人手による評価はコストが高く、時間もかかるという欠点がある。そこで正解翻訳文（基準翻訳文という）を人手で作成しておいて、機械翻訳結果と基準翻訳文を自動比較することで機械翻訳の精度を評価する新しい評価手法が考案された。本文では、これらの評価手法の概要を解説すると共に、筆者らが考案した新しい自動評価基準NMGについて述べる。

2 人手による評価の基準

人手による評価基準の一つとして、Muプロジェクトで用いた（1）理解容易性と（2）忠実度がある※1。前者は出力の訳文が翻訳先の言語表現としてどれだけ自然であるかを評価する指標であり、5段階で評価される。後者は出力文が入力文の持つ情報をどれだけ忠実に反映しているかを評価する指標であり、7段階で評価される。

最近では、同様の評価基準として、（1）Fluency（流暢さ、自然性）と（2）Adequacy（適切さ、妥当性）が用いられている。例えば、「私は少年である」という日本文を英語に翻訳した場合、"I am a girl"は（1）は高いが（2）は低い。一方、"Me is boy"は（1）は低いが（2）は高い。これらの評価基準を用いて人手によって機械翻訳結果を評価することができる。次節で述べる自動評価基準は、（1）と（2）の人手評価基準を置き換えうるものでなければならない。

3 自動評価の基準

人手による評価の高コスト性を克服するものとして機械翻訳結果の自動評価基準が提案された。その中でBLEU（Bilingual Evaluation Understudy）が最も良く利用されている。

BLEUは評価対象となる機械翻訳文と人手で作成した正解翻訳文（基準翻訳文）を比較して、その表現の一致度合いに基づいて評価値とするものである※2。翻訳には多数の正解がありうるので、正解文を複数用意して比較するのが普通である。このような自動評価手法は、高速かつ低コストで評価を行える利点があるが人手評価と比較しての欠点も指摘されている。例えば文献※3では、意味不明の文でありながら、BLEUの評価値が変わらない場合があることが述べられている。例えばアラビア語から英語への機械翻訳結果 "Appeared calm when he was taken to the American plane,　which will to

Miami, Florida." のBLEUの値と、意味不明な文 "Was being led to the calm as he was would take carry him seemed quite when taken" のBLEU値が一致することが指摘されている[注1]。

このようなBLEUの欠点は、次に示す（1）や（2）などの点から生じている。(1)単語の文字面のみの一致に基づいて評価しており意味の類似性を考慮していない。(2)構文的な自然性が評価に盛り込まれてない。（1）の欠点を克服する評価基準の一つとしてMETEORが提案されている[※4]。METEORでは、語の屈折などを標準化するstemming（例えば、boysをboyにする）やシソーラス（意味辞書）を用いて意味の類似性を考慮できるようにしている。また、（2）の欠点を克服するために、構文的な自然性を評価基準に含める手法の一つとしてBllip score（Brown Laboratory for Linguistic Information Processing score）がある[※5]。この評価基準は、評価対象文と基準翻訳文の双方を依存解析し、得られた依存構造の一致率に基づいてスコアを求めるものである[注2]。さらに（1）と（2）の両方の欠点を補う基準としてSemantic Roll Overlapが提案されている[※6]。

しかし、文字面以外の一致率に基づく評価には、構文（依存）解析器やシソーラスなど、何らかの高度なツールや言語資源を必要としているという欠点がある。そこで筆者らは、単純なテキストコーパス以外の言語資源を必要とせず、しかも表現の自然性を考慮した評価基準としてNMG （Normalized Mean Grams）を提案した。

4 新しい自動評価基準NMG

先行研究の調査結果に基づき、表現の自然性を評価する新しい評価基準を提案する。この方法は、評価対象文を目標言語の大規模コーパス（比較コーパスと呼ぶ）と比較し、n-gramの一致度を計測する。大規模コーパスを利用することで翻訳先言語のさまざまな表現と比較でき、表現の自然性を評価に盛り込むことができる。以下、NMGの計算法を具体的に説明する。

（1）評価対象文（機械翻訳文）Cを単語列W_1, …W_nに分割する。

(2) $i=1, \cdots, n$に対して、W_iから始まるgramで比較コーパス中に存在する最大のgram数をgrams（W_i）とする[注3]。そのとき評価値NMG（Normalized Mean Grams）を以下で定義する。

$$NMG(C) = \log_e \left(\sum_{i=1}^{n} grams(W_i) / n \right)$$

たとえば、比較コーパスが以下の4文から成るとして

- i am a boy
- you are a girl
- he is a man
- she is a woman

評価対象文（C）を

- she is a girl

とすると、n=4であり、評価対象文の4個の単語に対するgramsは

- grams(she)=3
- grams(is)=2
- grams(a)=2
- grams(girl)=1

となるので

$$NMG(C) = \log_e ((3+2+2+1)/4)$$
$$= \log_e (200) = 0.69$$

となる。NMGの値は使用する比較コーパスに依存する。比較コーパスとして大規模な翻訳先言語のコーパスをとることもできるし、基準翻訳文集合のみをとることもできる。

注1 この評価での基準翻訳文の一つは以下のとおり。
"Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida."

注2 依存解析とは日本語の係り受け解析に相当し、依存関係は係り受け関係に相当する。

注3 gramとは連続する単語のことである。gram数とは連続する単語数のことである。

5 NMGによる評価実験

NMGを用いた評価実験を行う。評価対象システム（system）は文献[※7]で用いたものであり、以下に示す。

・PAJでの翻訳結果（基準翻訳文）: ref
・特許専用規則方式機械翻訳による日英機械翻訳結果: rmt
・rmtに対して統計的後編集を施した結果: spe
・統計方式機械翻訳のみによる日英機械翻訳結果: smt

テストデータ（sample）として以下を用いた。これも文献[※7]で用いたものと同一である。

・closed test: sample1 188文
・open test: sample2 188文

比較コーパス（corpus）としては以下の2種類を利用した

・基準翻訳文（ref）1文のみ: REF
・米国特許アブストラクトデータ[注4]: ABS

systemが4通り、sampleが2通り、corpusが2通りあるので、実験の組み合わせは16通りとなるが、corpus=REF、system=refの組み合わせは意味がない。そこで、実際には14通りを実験した。

6 実験結果

評価実験の結果得られたNMGの平均値（μ）を図1に示す。

実験した各システム間についてt検定を行った。その結果、corpus=REF，sample=1の場合で、rmtとsmtとの間で有意水準1%で平均値間に有意差がないと判定された。それ以外のすべての組み合わせについては有意差があると判定された。

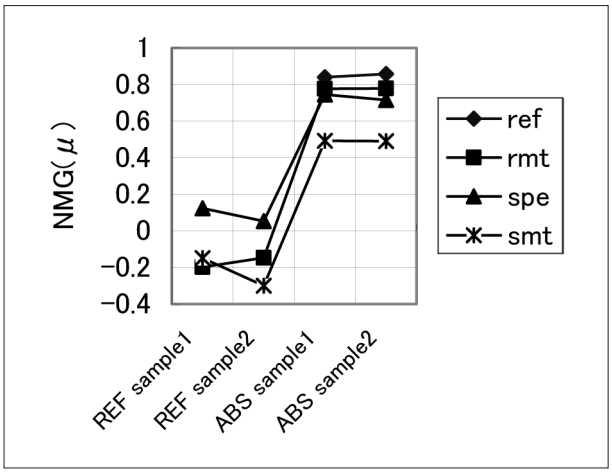

図1　実験結果

7 実験結果の考察

図1を見ると興味深い事実がいくつか分かる。

(1) 比較コーパスをREFとした場合、規則方式機械翻訳結果（rmt）よりも統計的後編集結果（spe）の方が評価値が高い。これは、文献[※7]に示したBLEUでの評価結果と一致している。しかしながら、比較コーパスをABSとするとspeよりrmtの方がNMGの評価値が高くなっている。これは、英語としての構文的自然性がspeよりrmtの方が高いことを示しており、rmtとspeの翻訳結果を目視によって比較した直感的考察結果と一致している。このことからNMGによる評価が英語としての構文的自然性を評価するという所望の性質を持っているといえる。

(2) 先に述べたようにcorpus=REF，sample=1においてsystem=rmtとsystem=smtとではNMGに有意な差はないが、corpus=ABSとした場合には、大きな差が出ており、少なくとも、今回用いた統計的機械翻訳（smt）の結果は、規則方式機械翻訳（rmt）の結果と比較して、英語としての構文的自然性を有していないといえる[注6]。

注4　米国において2000年に特許明細書として公開されたデータ（157,596件）から要約部分のみを抽出したもの。本研究が、「課題」部分を翻訳対象にしているので要約部分のみを用いるのが適切であると考えた。ABSの総文数819,123、総単語数は21,194,662である。

(3) 最後にcorpus=ABSとした場合、PAJの英文自体（ref）の評価値が最も高いのは当然であろうが、定量的に、例えばcorpus=ABS, sample=2, system=refのNMG（μ）=0.8575であり、これは、平均グラム数e^{NMG}が$e^{0.8575}$=2.35、つまりref中の各単語から始まるABSに存在するgram数が平均2.35 gramであるということになる。これは、refに存在する長い単語列に対してはABSに存在しない可能性が高いということを意味しており、ABSの規模が小さすぎたためと考えられる。

8 おわりに

機械翻訳の新しい自動評価基準としてNMG（Normalized Mean Grams）を提案し、PAJの基準翻訳文および米国特許アブストラクトデータを用いて、各種機械翻訳結果の精度や自然性を評価した。その結果、目視による直感的な評価結果に一致しており、本評価基準が妥当なものであることが分かった。今後の課題として以下が考えられる。

・一層大規模な英語コーパスを用いて自然性の評価をする
・意味や構文の類似性も考慮できるように評価基準を改良する。
・人手での評価結果とNMGの評価結果の関係を明らかにする

注6 ここで用いたSMTでは、92，855文対を用いてGIZA++によって翻訳モデルの学習を行った。

参考文献

※1 長尾真、辻井潤一：Muプロジェクトにおける日英翻訳結果の評価、情報処理学会研究報告、自然言語処理47-11、pp.79-88、1985。
※2 Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu : BLEU: a Method for Automatic Evaluation of Machine Translation, ACL2002, 2002.
※3 Chris Callison-Burch, Miles Osborne, Philipp Koehn: Re-evaluating the Role of BLEU in Machine Translation Research, 11th Conference of the European Chapter of the Association for Computational Linguistics, pp.249-256, 2006.
※4 Satanjeev Banerjee, Alon Lavie : METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65-72, Ann Arbor, June 2005.
※5 Michael Pozar, Eugene Charniak : Bllip: An Improved Evaluation Metric for Machine Translation, Brown University Master Theses, 2006.
※6 Jesú s Gimé nez, Lluí s Má rquez : Linguistic Features for Automatic Evaluation of Heterogenous MT Systems, Proceedings of the Second Workshop on Statistical Machine Translation, pages 256-264, Prague, June 2007.
※7 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳、Japio 2006 Year Book、pp.184-187、2006。