翻訳文評価方法の一考察

山内 佐敏
(株) リコー 研究開発本部
yamauchi@int.rdc.ricoh.co.jp

ＡＬＰＡＣレポートで組織的な機械翻訳システムの訳文評価がされて以降、幾つかの評価方法が提案されてきた。しかし、いずれの方法でも評価基準に対する解釈の違い（揺れ）や、その時々の時代的背景に影響を受ける可能性を持っていた。

そこで筆者は対象原言語の解析能力と共に、変換生成部分に付いても一義的な評価法の一つとして人間の翻訳を含め複数の翻訳システム間の相対的な翻訳の質の違いを数値化し、評価する方法を提案する。それは一対比較法の一種であり、原言語が正しく解析されている文の訳文を複数の評価者(訳文ネイティブ)に評価してもらい、その順位の平均を取って変換生成能力とするものである。また、本提案で２～３のシステムを評価した結果を考察し、客観的な翻訳能力を示す尺度であることが確認できた。

１．はじめに

従来から科学技術の進歩は対象を観測する、測定する、あるいは評価する技術の発展と共にあった。機械翻訳システムはある限定された分野での実用化がなされたものの更に一段の性能向上が望まれている。それは翻訳アルゴリズムのブレークスルーを待たなければならないが、その性能を評価する方法もより効果的で分かり易い方法の確立が望まれる。

機械翻訳に関する評価法としてはアメリカにおいて１９６６年に発表されたALPAC レポート[1]に初めてかなり厳密な評価法（ Intelligibility と Informativeness の二つの尺度）が登場した。その後、日本で行われた機械翻訳プロジェクト（通称：Mu プロジェクト）[2]では、それを簡易化し「忠実度」「理解容易性」の二つの尺度で表現する方法を用いたことにより、これに類した手法が普及した。また、成田[3]は基本的な文法を含む短文を比較的網羅的に組み込んだ解析力評価文セットを開発し、アルゴリズム開発者に提供し、品質の向上に貢献した。その後、電子協において開発者側からの評価と、ユーザ側からの評価に分け、それぞれに品質評価用テストセット（評価文）を準備し公開し、やはり多くのアルゴリズム開発者の利用に供している[4) 5) 6)]。

しかしながら、これらは定性的な部分が多く評価者集団の構成によるゆれや、評価する者の立場、評価する時点での社会的な背景が影響する余地が残っていたり、あるいはシステム開発者のツールとして用意されているため、実際に翻訳の現場で用いられている文とは若干異なる場合もあるので、ユーザ側からはどのシステムが良いのか分かり難い。

そこで筆者はシステム開発者のため（アルゴリズムの性能向上のため）ではあるが、示された数値はユーザ側からでも分かり易く、自ら測定しうる客観的で測定環境に影響されにくい翻訳システム評価方法を提案する。基本的には原文の解析能力と、訳文への変換生成能力の二つの尺度である。前者は特に目新しいものではないが、後者は対象原言語の解析能力を除いた部分の評価法の一つとして人間の翻訳文を含め複数の翻訳システム間の相対的な訳文の質の違いを数値化し、評価する方法である。また、本提案で２～３のシステムを評価した結果を示す。

２．評価方法

翻訳システム全体の評価を一つの尺度で行うのはやはり無理がある。一旦はMu プロジェクトが用いた尺度の様に２つに分ける。すなわち原文の解析能力を示す尺度と、変換生成能力（ピボットあるいは中間言語方式と称するシステムでは訳文生成能力）である。

２．１ 原文解析能力

一文単位で原文文法に則した訳文になっているかどうか（合否）を判定する。その時、意味的にも文脈に合致した意味を取りうる構文構造を得ているかをも考慮に入れる（但し、変換生成部分で訳語の選択の誤りによる意味の違いを排除するものではなく、あくまでも構文的に係り関係の合否をみる)。

評価者には ① 訳文側言語のネイティブであり、② 原文の文法をよく理解し、③ 記述内容の分野に関して知識を十分に持っていること（但し、後述の様に代案はある）、という条件が要求される。

評価文はシステムが狙っている分野の文、あるいはユーザが利用したい分野の文からランダムに選ぶ。

指標としては

$$M_A = \frac{C_{sent.}}{T_{sent.}} \times 100 \quad (\%) \qquad \cdots (1)$$

$T_{sent.}$ ：評価対象文数
$C_{sent.}$ ：正解文数

として単純に表現する。

2．2　変換生成能力（訳文生成能力）

訳出された文が如何に読みやすいかを求める尺度である。

手法としては一対比較法[7]の一つであるが、調査方法がより簡便で、得られた結果を解釈するにもより直感的に判り易くなるようにその指標が０～１０の範囲に収まるように変更した。

この場合、評価者には訳文を利用するユーザに相当する人を選択する。すなわち、訳文側のネイティブであり、対象文書分野の一般的な知識を持っていればよい。ここではパネラーと呼ぶ。

評価文は原文解析能力の測定時に正解と判定された文であり、評価対象のシステムが共通して正解を出した文のみを用いる。

実験に際して用意した翻訳文［実例は付録を参照］は

Ｈ：人間翻訳文：商品として用いられている翻訳文（あるいは商品として耐えうる翻訳文）で、意訳を含んでいて良い。システムの一つと見なす。

Ｉ：間接翻訳文：原文の文法の係り受け関係を維持または論理的な変換を加味して自然な目的文を作成した人間翻訳文（直訳調ではなく、現時点で機械翻訳システムが訳出しうる最高の訳文）。システムの一つと見なす。

Ｓ：評価対象システム訳出文：一般的には評価したい機械翻訳システムが訳出する文。

（但し、評価対象システムの解析能力評価で正解となった文のみを対象とする。）

評価手順は、それぞれの翻訳文をパネラーに一対ずつ提示し、優劣を決めてもらう。その後、評価文毎に、パネラー毎に、悪い順に番号を振る。

指標としては

$$M_G = 10 \cdot \frac{1}{l} \cdot \frac{1}{m} \cdot \frac{1}{(n-1)} \sum_{i=1}^{l} \sum_{j=1}^{m} (N_{ij} - 1) \quad \cdots (2)$$

l ：評価文の数
m ：パネラーの数
n ：評価対象システム数
N_{ij} : 注目システムの訳文 i 毎にパネラー j に
より与えられた n 段階の評価値

を用いる。

2．3　総合翻訳力指数

これは単純に解析能力と変換生成能力の掛算した結果を用いる。すなわち

$$M_T = M_A \times M_G \qquad \cdots (3)$$

である。

3．評価実験

以上の考え方に基づいて２つの異なった形で実験を試みた。一つは「現在得られるシステム（現在、実現している技術によるシステム）の訳文が本来欲しい（目標としている）翻訳文に対してどの程度かけ離れているか。」もう一つは「市場にでているシステムの中でどれが一番良いか（似たようなシステム同士でどちらが優れているか）。」である。

3．1　評価文

５分野（航空機整備、工作機械、ソフトウエアインストール、プリンター、ネットワーク）のマニュアルからそれぞれ３０文を選んで計１５０文を評価文とした。

注）S_A，S_B共に解析能力の評価では正解となった文を選んでいる。

3．2　パネラー

訳文を利用しうる人（機械翻訳の開発に従事していない人）で訳文生成能力を測定する際に訳文の良否を判定する人：５名

3．3　評価者

原文解析能力測定時は解析の合否を判定し、訳文生成能力測定時には、事前に設定した評価手順、および、評価基準に従いパネラーに評価の手順、基準を示し、評価に取り組む姿勢や環境の違いを最小限に押さえる（機械翻訳システム開発従事者）：４名

4．評価結果と考察

4．1　解析能力

評価文の文長毎の分布は次のようである。

総文数	1～10	～20	～30	～40	～50語
150文	35文	80文	26文	7文	2文

機械翻訳システムが訳出した文に対し、人間翻訳文（H）を絶対的な正解とし、同一の係り関係となっている文のみを評価者が合格と判定する。

市販の機械翻訳システム2台（A，B）ではその結果を（1）式で計算すると

$$M_A(S_A) = 55.3 \quad (\%)$$

$$M_A(S_B) = 52.0 \quad (\%)$$

のようになった。

勿論、人間翻訳（H）と、間接翻訳（I）は

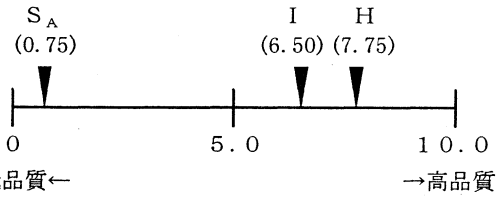
$$M_A(H) = 100 \quad (\%), \qquad M_A(I) = 100 \quad (\%)$$

である。

4.2 変換生成能力

解析能力評価で用いた文のうち1分野はシステムA，B共に共通する文が少なかったので割愛し、4分野各5文の計20文により評価を行った。

手順に従い、得られたデータを（2）式で計算した結果は次のようになった。



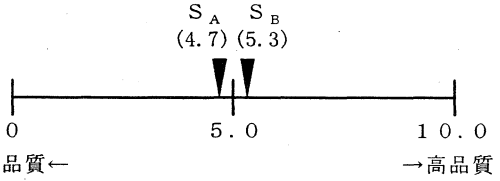
すなわち、

$$M_G(S_A) = 0.75$$

である。

このようにMGが1以下であることは、機械翻訳システムに期待している訳文の質には程遠いことを示している。

また、参考のためにA，Bの機械翻訳システム同士の比較を行ったが、次のようになった。



このように、2つのシステム間での優劣を決めるだけならば各評価文に対してパネラーにどちらが良いかを判定させ、良い方に1悪いほうは0として平均を取り 10 倍すれば同じ数値を得ることができる。また、今回は時間の関係から機械翻訳システムは2種類しか評価しなかったが、数システムを一度に評価し、同一軸上において見ることは可能で、全体の順位を知ることが出来る。

さらに、直接には測定しなかったが、前述の2つのデータから、MG(SB)を概算できる。

すなわち、$M_G(S_B) \approx 0.78$

である。

4.3 総合翻訳能力指数

前述の様にシステムA，Bとも、変換生成能力がほとんど意味がないほどの数値なので、この指数を計算してもあまり意味がないが、一通り計算してみると次のようになる。

SA：機械翻訳システムAの場合は

$$M_T(S_A) = 55.3 \times 0.75 = 41.5$$

SB：機械翻訳システムBの場合は

$$M_T(S_B) = 52.0 \times 0.78 = 40.6$$

となり、両者はほとんど同じ性能といえる。

4.4 評価式に対する考察

一般の一対比較法はパネラーに対して「AとBのどちらが良いか」を言わせるだけではなく、5段階評価させている。これはどちらとも言えない場合と、僅差で優劣がある場合、明確に優劣がある場合のように分けて点数を付けさせている。言い換えると、一つの事象を見た場合、見方にも何らかの分布が生じているので、それをも集約して結果に反映させようとしているわけである。この考え方は香り、味、色合い、音色のような人の五感に関係する嗜好の類を対象とした場合には重要である。しかしながら、評価対象が工業製品や論理的な物の場合は白か黒かはっきりさせたほうが良い。中間の分布を無視することによる歪みはサンプル数とパネラーの量で吸収できると見る。

また、ここで変換生成力は平均値だけでなく、標準偏差も考慮すべきだとの議論もあるが、かえって分かり難くしてしまうので、「その差が1以内の場合には個人によっては評価が逆転する場合もある。」というような言い方に止めるほうが一般ユーザには分

かりが良い。参考に機械翻訳システム同士の評価では評価者間の標準偏差がσ=0.68で、評価文間のそれはσ=3.81であった。

更には、人間翻訳と間接翻訳の両方を用意する必要があるかどうかも議論の余地は残るが、今後の機械翻訳アルゴリズムの大いなる発展を期待すれば、両者を同時に用いる方法が良いと思う。

5．結論

変換生成能力（訳文生成能力）を原文解析能力と同じように数値で表現する一つの手法を得ることができた。

また、これによってシステム開発者、ユーザ共に共通の尺度で意見交換が可能な評価方法、評価尺度が提案できた。

7．おわりに

今回の実験によって、提案した評価法から得られる指標は非常に客観的な指標だとの確信を得た。何時もどのようなジャンルの、どの程度の量の、評価文の選定を如何にするかという問題が残るが、外国語としての英語能力を測定するＴＯＥＩＣの試験問題のように、ある一定の基準を作る事は可能である。そうすることによって、例えば自動車の燃料消費率の表示の一つである 10･15 モード燃費のような使い方ができ、翻訳システムのカタログ仕様の一項目に記載義務化も可能となる。すなわち「原文解析力」「変換生成力」の２項目による表示である。

しかしながら、このような議論をしてみても、原文解析力が５０(%) 程度、変換生成力が１以下では話にならない。せめてそれぞれが７０(%) 以上、３以上とならなければ（総合翻訳能力指数としては２００を越えなければ）商品として胸を張って表示できる様にはならないだろう。

また、本提案での間接文(Ｉ)と機械翻訳システムの２者間比較で変換生成力 (M_G) が５以上となったら本当の意味でブレークスルーがなされたと言っていいだろう。そうなる時が早く来る事を期待する。

勿論、考察で触れたように本評価法も議論の余地が残っている。更に的確な評価が出来るように改善していきたい。

最後にこの評価実験を推進して頂いた成田真澄さんをはじめ機械翻訳システムの研究開発に携わった皆さんに感謝します。また、そのような環境を提供して頂いた研究開発本部のスタッフの方々に謝意を表します。

[参考文献]

1）ALPAC.(1966).“An Experiment in Evaluating the Quality of Translations.” LANGUAGE AND MACHINES —COMPUTERS IN TRANSLATION AND LINGUISTICS — Appendix 10, pp.67-75.

2）電子技術総合研究所,京都大学.(1986).「日英学技術文献の速報システムに関する研究——言語処理システムの開発に関する報告書」.pp.483-539.

3）成田一(1987).「機械翻訳における構造処理能力の評価」情報処理学会報告. NL_69-1. pp.1-9.

4）中岩浩巳,森本康嗣, 松平正樹, 成田真澄, 野村浩郷. (1993).「JEIDA 機械翻訳システム評価基準（開発者編）」.情報処理学会報告. NL_96-10. pp.73-80.

5）井佐原均,新納浩幸,山端潔,森口稔,野村浩郷. (1993).「JEIDA 機械翻訳システム評価基準（品質評価編）」.情報処理学会報告. NL_96-10. pp.81-88.

6）(社)日本電子工業振興協会. (1995).「機械翻訳システム評価基準—品質評価用テストセット—」. 95-計-17

7）西里静彦.(1982).「質的データの数量化—双対尺度法とその応用—」.朝倉書店. pp.162-171.

[付録] 実験に際して用意した翻訳文の例

原文：Data Input : The machine is kept inoperative unless the necessary operating data such as index memory, wafer size, cutting depth and cut mode, are correctly input.

H：データ設定：本機はインデックスメモリー、ウエハサイズ、切込み深さ、カットモードなどの作動に必要なデータが正しく入力されなければ作動しない。

I：データ入力：必要となるオペレーティングデータ（インデックスメモリー、ウエハーサイズ、切り込み深さ、切断モード等）を正しく入力しなければ、この機械は作動不能になる。

S_A : Data Input : 機械は、インデックスメモリー、ウエハーサイズ、切込み深さ、及び、切断モードのような必要な作動しているデータが、正確に入力されなければ、無効な状態にされる。

S_B : Data Input : インデックスメモリー、ウエハーサイズ、切込み深さ、及び、切断モードのような必要なオペレーティングデータが正しくインプットされない限り、機械は、動作不能状態に保たれる。