英日機械翻訳の訳質に関する評価実験

富士 秀
(株) 富士通研究所
fuji@flab.fujitsu.co.jp

1 はじめに

本研究では、これまで訳質と情報伝達度（長文読解）の関係を調べる評価実験[7][4]を行なってきたが、今回はこれに加えて分かり易さや自然さといった読み手の感覚に与える影響についても評価も行なって、情報伝達度との比較を行なった。これらの実験を通じて、「読むための道具」としての機械翻訳にとってどのような評価方法が有効であるか検討した。

2 背景と目的

近年、インターネット上の外国語文書を母国語に翻訳して読むためのブラウジングツールが脚光を浴びているが、本研究ではこのブラウジングツールに焦点を絞った評価実験を行なっている。

ブラウジングツールが広く使われるようになると、読み手（利用者）の観点に立った評価が重要になってくる。この傾向を象徴するように、最近では、各種翻訳システムに関する比較記事が新聞や雑誌などで取り上げられる機会が多いが、評価手法などに問題があり不適当な内容になっているものも多い。

そこで本研究では、読み手の立場に立った評価手法を策定することを目標とした。

3 従来の研究

欧米の研究などでは従来より翻訳評価に関する研究[1]が行なわれており、一連の研究を総括したある論文[2]では、翻訳評価の観点として「忠実度」、「自然さ」、「情報伝達度」などを上げている。ただし、これらの観点が存在するという認識はあっても、これに基づいた定量的な評価はあまり行なわれていない。また、開発の指針となるようなものも見当たらない。

一方、国内の代表的な研究では、電子協の評価用テストセット[6]などが知られているが、これは開発者のための評価方式である。この評価セットを用いれば、開発対象となっている特定の言語現象に関してはシステム性能が上がる。しかし、このテストセットだけでは、システムの利用者にとって訳文の質がどのくらい向上したかはわからないという問題点がある。更に、テストセットによって問題点がわかったとしても、効率的な改善を行なうには一体どの問題点から手をつけて良いかの指針が示されない。

本研究でも、以前「読むための道具」という観点で、開発の指針となるような研究[7][4]を行なってきたが、評価方法が長文読解に限定されており総合的な評価とはなっていなかった。

4 評価実験

「読むための道具」に特化した評価手法として「情報伝達度」、「分かり易さ」、「自然さ」といった評価手法を考案し、各々の評価手法としての有効性を比較する実験を行なった。更に、これらの手法がシステム開発の方向付けにつながるように、実際の機械翻訳文に対して評価手法を適用した。また、機械翻訳文に段階的に人手修正を加え、それぞれの評価手法による結果の推移を分析した。なお、本評価実験はどのような言語対に対しても実施することができるが、今回は英日翻訳の品質を評価する実験を行なった。

4.1 評価手順

1. **評価対象文章の収集**
   全ての評価手法に共通して使えるような英語文章を数種類用意する。「情報伝達度」では文章の内容に関する選択式の読解質問（質問は英語）が必要になるので、質問とセットになったものを英検準一級のテキスト[3]から選んで用いる。

2. **評価対象文章の翻訳**
   用意した英語文章を機械翻訳で日本語に翻訳する。

3. **機械翻訳文章の人手修正**
機械翻訳した文章を段階的に人手修正する。（詳細は4.2を参照）

4. **評価手法毎の実験用紙の作成**
人手修正を施した文章を、それぞれの評価手法に適するように加工する。（詳細は4.3を参照）

5. **評価実験の実施**
大量の被験者を用いて評価実験を行なう。

6. **実験結果の分析**
評価実験の結果を、統計心理学的手法を用いて分析する。

4.2 用意した訳文

本実験の目的が翻訳システムの改善の指針となることにあるので、読解文章としては機械翻訳文を出発点とし、手翻訳文を到達目標としておいた。

修正の方針としては、単語などの局所的な修正から始め、順次、複合語や句表現など少しずつ全般的な修正に移り、最後に文脈処理などの広域的な修正を行なった。以下に、各例文について説明する。

- **原文**
英検準一級の長文読解問題の英語読解文そのもの。主に訳文との比較の用途のために用いた。今回は、医学、政治などに関する新聞の解説記事数種類であり、長めの文が多く含まれたものである。

- **訳文1**
原文に対して、チューニングなしの機械翻訳をかけて得られた日本語文。

- **訳文2**
訳文1に対して、単語や名詞の複合語などを人手で修正して得られた日本語文。

- **訳文3**
訳文2に対して、主に名詞以外の複合語や句表現の誤りを人手で修正して得られた日本語文。

- **訳文4**
訳文3に対して、係受けレベルの誤りを人手で修正して得られた日本語文。なお、この修正には、解析が失敗して単語のみが羅列して出力されてしまっているような文に対する構造上の修正も含まれる。

- **訳文5**
訳文4に対して、時制、照応、省略などの文脈的な誤りを人手で修正したもの。

- **訳文6**
訳文5に対して、訳文の基本的な構造を活かしながらも、自然な文になるように人手で修正したもの。

- **訳文7**
長文読解問題の解説書に説明用に添付された、人手による意訳文。機械翻訳の結果には制約されていない。

表1. 評価対象例文

訳文1	原文 ＋ 基本辞書のみの機械翻訳
訳文2	訳文1 ＋ 単語、複合語、を人手修正
訳文3	訳文2 ＋ 訳語、句表現を人手修正
訳文4	訳文3 ＋ 係受けを人手修正
訳文5	訳文4 ＋ 時制、照応、省略を人手修正
訳文6	訳文5 ＋ その他の文脈を人手修正
訳文7	手翻訳の和文

4.3 評価手法毎の評価用紙の準備

- **情報伝達度の評価**
英日翻訳の情報伝達度を評価するために、英語の読解試験を用いた。英語の読解試験の読解用文章を機械翻訳で和訳し、4.2の訳文群を作成して読解文として用いた。読解試験の質問文は人手で完全な日本語に翻訳した。

- **分かり易さの評価**
英語の文章を機械翻訳で和訳し、4.2の訳文群を作成して対象文書として用いた。情報伝達度の評価との比較のため、同じ英語文章を用意した。

回答用紙としては、機械翻訳文をまず紙に印刷し、各文の右側に「分かり易さ」を4段階評価するための欄を設けた。4段階とは、「非常に分かりにくい」、「やや分かりにくい」、「やや分かり易い」、「非常に分かり易い」である。回答が真中の評価に集中することを防ぐために、選択肢は偶数個となるようにした。なお、文章の最後には文章全体としての分かり易さを記入するための欄を設けた。

回答方法としては、先ず最初に文章全体を最後まで一通り読むように指示することによって、全員に文脈を意識させるようにした。また、「分かり易さ」の判断は、なるべく時間をかけずに「直観的」につけるように指示した。

- **自然さの評価**

  上述の「分かり易さの評価」とほぼ同じ手順で実験を行ない、「分かり易さ」のかわりに「自然さ」について答えてもらった。英語文章は、上記2種類の評価と同一のものを用いた。

  回答用紙としては、機械翻訳文をまず紙に印刷し、各文の右側に「自然さ」を4段階評価するための欄を設けた。4段階とは、「非常に不自然」、「やや不自然」、「やや自然」、「非常に自然」である。なお、文章の最後には、文章全体としての「自然さ」を記入するための欄を設けた。

  回答方法は、分かり易さの場合と同様である。

4.4 被験者

およそ500名の大学生を被験者として評価実験を行なった。全ての評価試験は記名式で行なった。盲目的に全てに同じ答をつけるなどの無効だと判断される結果に関しては、取り除いてから統計処理を行なった。

4.5 統計分析

統計分析[5]は、基本的に分散分析手法を用い、各訳文間での多重比較を行なった。有意差の判断基準としては、統計心理学では一般的とされる95％検定を用いた。（ただし、自然言語処理の分野では多少厳密過ぎる可能性もある。）

5 実験結果

5.1 情報伝達度の評価

統計処理の結果、大きな訳質の改善が、訳文2から訳文3に移る部分に集中していることがわかった。つまり、訳文3までのチューニングを行なえばかなり最高値に近い性能が得られ、それ以降の修正を行なってもあまり情報伝達度に貢献しないことがわかった。

図1.はある文章について情報伝達度を評価したものであり、評価結果を表す一例である。他の文章についても似たような結果が得られた。

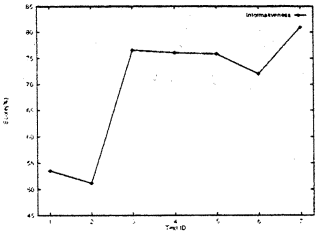

図1. 訳文による情報伝達度の推移の例

5.2 分かり易さの評価

「非常に分かり易い」が5点で「非常に分かりにくい」が1点という設定にしたが、評価結果は機械翻訳文（訳文1）がおよそ2点であって、手翻訳文（訳文7）がおよそ3点となっている。

訳文1から訳文7に移行するにしたがって分かり易さの値は上がったが、情報伝達度に見られたような局所的な変化は見られなかった。また、情報伝達度と比較して、文脈関連の修正の効果が大きく出ている。ただし、統計的に分析しても文脈関連の修正部分に評価値の改善が集中しているとは言えないことがわかった。

図2はある文章に対する評価結果の一例だが、他の文章に対しても同様の結果が得られた。

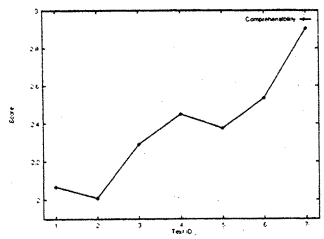

図2. 訳文による分かり易さの推移の例

5.3 自然さの評価

「非常に自然」が5点で「非常に不自然」が1点という設定にしたが、評価結果は機械翻訳文（訳文1）がおよそ2点であって、手翻訳文（訳文7）がおよそ2.7点となっている。

訳文2から訳文3に移るときに、情報伝達度と同様に、比較的大きな改善が見られた。しかしながら、この部分の改善は統計的には有意差が認められるほどのものではなかった。また、文脈的要素の修正においても局所的ではないにせよかなりの評価値の改善が見られた。全体としては、分かり易さと同様、全ての修正を行なわないと統計的に有意な差は生まれないことがわかった。

図2はある文章に対する評価結果の一例だが、他の文章に対しても同様の結果が得られた。他の文章の場合も、平均評価値は分かり易さより低めとなっている。

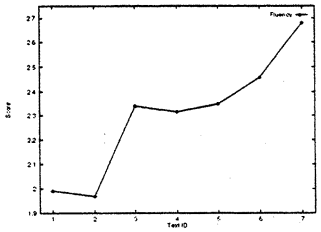

図3. 訳文による自然さの推移の例

5.4 被験者の感想

被験者の何人かから実験に関する感想を聴取した。その結果、「分かり易さ」と「自然さ」は分けて考えることが難しいことがわかった。つまり、両者は密接な関係にあり、例えば「分かりにくい文は不自然にも感じられる」というようなことが起こってくることがわかった。

6 結論

- **評価値の改善箇所**
  情報伝達度では評価値の改善箇所が一箇所に集中しているが、分かり易さと自然さではそのポイントを見つけることができなかった。つまり、情報伝達度では特定の修正を行なうことによってかなりの訳質改善が見込めるが、分かり易さや自然さでは全ての修正を行なわないと最終的な改善に結び付かない。

  違う見方をすると、現在の訳質ではいきなり分かり易さや自然さの向上を目指すよりも、まず情報伝達度の向上を目指す方が実現可能性が高いと言える。

- **2種類の分かり易さ**
  人間が分かり易いと感じる分かり易さと、実際に人間が内容について分かる（情報が伝達される）かどうかは、大幅に異なる場合がある。

- **分かり易さと自然さ**
  読者が感じる分かり易さと自然さは、統計データからも感覚的にも、密接な関係にあり、両者を切り離すことは難しい。

- **不自然でも情報伝達度の高い文章の存在**
  読み手としては不自然だと感じても情報伝達度の高い文章は存在する。つまり、品質に不満はあるが使う気が起きるレベルの訳文を目指すことができる。

- **情報伝達度評価の長所**
  情報伝達度評価では、英語の原文も同様にして扱うことができるので、これらの比較実験を行なうことができる。これに対し、分かり易さや自然さの評価では、原文との比較は難しい。

7 おわりに

機械翻訳の評価に関して従来研究を調査したり実験を行なってきて、評価に関する研究があまり行なわれていないことを感じた。研究レベルの向上のためには、評価自体に関する検討が必要であることを実感した。

8 謝辞

本研究の心理学実験を進めるにあたって貴重なアドバイスをいただいた筑波大学心理学系教授の海保博之先生ならびに同研究科の平山氏、荷方氏に感謝いたします。また、実験にご協力いただいた東京成徳大学の佐藤至英先生、静岡大学の比留間太先生、作新学院の西谷健次先生、静岡大学の村越真先生、米沢女子短期大学の高尾哲康先生に感謝いたします（順不同）。

参考文献

[1] AMTA, IAMT. *MT Evaluation: Basis for Future Directions*, November 1992.

[2] John S. White, editor. *"The Primacy of Core Technology MT Evaluation"*, October 1996.

[3] （財）日本英語教育協会（編）. 英検セミナー「準一級英文読解」. 旺文社, 1991.

[4] 富士秀.「英日機械翻訳文の読解に関する評価実験」. 言語処理学会第2回年次大会発表論文集, pp. 21–24, March 1996.

[5] 田中敏, 山際勇一郎. ユーザーのための「教育・心理統計と実験計画法」. 教育出版株式会社, 第2版, September 1992.

[6] 日本電子工業振興協会.「機械翻訳システム評価基準」, June 1995.

[7] 富士秀, Eric Visser.「機械翻訳文の理解容易度に関する評価実験」. 情報処理学会第51回全国大会予稿集, September 1995.