英日機械翻訳文の読解に関する評価実験

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、機械翻訳を読解支援のための道具としてとらえた場合、どのような言語現象が読解に影響を与えるかを調べるための実験を行なった。今回の実験では、200名規模の被験者に対して長文読解形式の問題を解かせることによってデータを収集した。

2 背景と目的

機械翻訳が一般個人に普及するにつれて、読むための道具として使われる機会が増えてきている。読むための翻訳では、個々の言語現象の正確さよりも、訳文全体として意味が通じるかという観点での評価が重要になってくる。

一方、機械翻訳結果を後編集して利用する際にも、訳文の質が意味の伝わるレベルまで到達していれば、後はネイティブによるチェックのみの作業となり効率はアップすると考えられる。

このようにして、機械翻訳システムを開発する上で、訳文全体として意味が通じるかどうかが重要な評価基準となってきている。本研究の目的は、現状の機械翻訳システムの開発において、どの言語現象を改善すれば意味の通じるような品質に到達することができるかを見積もることにある。

3 従来の研究

機械翻訳の品質評価の研究は古くから行なわれており、方法も多岐に渡る。古くは1960年代のALPACレポートの評価から、最近では電子協の品質評価用テストセット[6]まで数多くの研究成果が発表されている。

従来の評価では、

- 文中の個々の要素が適切に訳されているかの評価
- 文書全体に対する評価

の2種類の評価方法のうちでは、前者の方が中心であった。前者は、システム開発者が評価者となる場合が多く、実際にシステムを使うユーザの立場に立った評価は少なかった。

後者に関しても研究は行なわれているが[1]、機械翻訳システムの開発における言語現象毎の問題点の切りわけにはなかなか結びついていない。

4 実験方法

今回の実験は、実験方法や使用例文を試行錯誤によって決定し、最終的に以下に述べるような形態に落ち着いた。

4.1 実験の必要条件

各出力文について言語現象毎の評価をするには、システムからの出力を期待される訳文と比較すれば良い。また、システムの内部状態を観察することもできる。しかし、文書全体に対する読解の場合は、読者が読解の結果を明示的に表明（出力）することはないので、評価実験としては、読者にある課題を課してこれに対する回答（出力）をさせる必要がある。つまり、一種の心理学実験を行なうことになる。

心理学実験では、個人の能力によるばらつきが出るため、ある程度の人数の被験者に対して同じ条件で実験を行ない、得られたデータを統計処理する必要が出てくる。また、大人数に対して課題を与える際には次のような項目が実験の遂行のための条件となってくる。

- **課題がわかり易く、回答方法が簡単**
  多くの被験者に徹底させる必要があるため。
- **実験の所要時間が短い**
  多くの被験者を集める必要があるため。具体的には、１０分前後が現実的。
- **実験結果が機械的に処理できる**
  実験データが大量になるため。記述方式だと採点に揺れが生じ手間もかかるので、○×式や選択式などの方が好ましい。
- **一斉試験が可能**
  均一な実験環境を保証するため。微妙な実験環境の違いが読解正解率や読解所要時間に影響を与える。

4.2 実験方法の選択

以上述べた条件を満たすような実験方法はいろいろ考えられるが、今回は、実施が比較的簡単であると思われる長文読解問題を使った実験を行なった。

具体的には、機械翻訳された長文に対して被験者に長文読解問題を解かせて正解率を計算する。そして、この翻訳文に対して言語現象毎に順次人手で修正（後編集）を行ない、読解正解率がどのように変化するかを観察した。ここで、正解率の向上に大きく寄与した言語現象が、長文読解における重要な要素であると考えることができる。

なお実際の実験では、実験回数を減らすため、順次修正を行なっていくのではなく、様々な修正レベルの訳文を事前に用意しておき一斉に読解試験を行なうという方法をとった。

4.3 実験手順

今回の実験では、次の手順を踏んだ。

1. **長文読解用例文および質問文の収集**
   英語読解文章およびこれに対する英語の質問文一式を用意する。読解文章と質問は、市販の英検準一級のテキスト[3]に載っているもので、本来は英語文章の内容理解を試すものである。この英語の質問文を人手で日本語に翻訳する。
2. **読解文章の和訳および各種処理の適用**
   用意した英語読解文章に対して機械翻訳で和訳を行ない、さらにこの機械翻訳結果に対して言語現象毎に順次修正を行なう。修正の内容については次節を参照のこと。
3. **読解標準時間の設定**
   読解問題を解くための制限時間を設定するために予備実験を行なう。予備実験の方法としては、予備被験者数名に、日本語として最も適切である訳文（人手翻訳文）と和訳した質問を与え、なるべく短時間で問題を解いてもらう。全問正解した人の所要時間の平均を、標準読解時間として設定した。
4. **読解試験の実施**
   4回にわけて計約２００名の被験者に対して試験を行なった。
5. **集計**
   被験者の回答を読解文章毎に集計し平均正解率を計算した。また結果の統計的有意性を検証した。

4.4 用意した読解文章

本実験の目的が機械翻訳文の品質を上げることにあるので、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。以下で、訳文１と訳文２は機械翻訳の出力であり、訳文３以降は人手による修正を加えた訳文である。

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。今回用いた文章の内容は、医学に関する新聞の解説記事であり、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、基本構成のままの機械翻訳をかけて得られた日本語文。
- **訳文２**
  原文に対して、辞書登録および学習機能を施した機械翻訳をかけて得られた日本語文。
- **訳文３**
  訳文２に対して、主に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**
  長文読解問題の回答に説明用の人手による意訳文。当然、機械翻訳の結果には制約されていない。

＜英語＞

原文	英語の原文

＜日本語＞

訳文１	原文 ＋ 基本辞書のみの機械翻訳
訳文２	原文 ＋ チューニング付き機械翻訳
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

表１．読解対象例文

4.5 用意した質問文

長文読解問題に用いられている質問文を用いた。元々の質問は英語であるが、これを手翻訳で日本語に翻訳して実験に用いた。問題数は5問であり、各問は4つの選択肢から正しいと思われるもの1つを選ぶという、四者択一式である。

質問は、読解文章の大意の理解度を試すような質問となっている。つまり、「何が」、「いつ」、「どこで」のような質問ではなく、「なぜ」や「どのように」を問うような質問で構成されている。

4.6 読解試験の実施環境

- 全被験者を実験会場に集合させ一斉に実験を行なった。実験に先だって、回答方法の説明を行なった。なお、実験の目的や内容に関する事前の説明は一切行なわなかった。
- 制限時間内に全質問に答えるよう指示した。
- 実験会場に時計を用意し、制限時間内に回答するよう指示した。

4.7 被験者

大学および短期大学の学生を中心におよそ２００名に実験の被験者として協力をお願いした。学生の専攻は、理系と文系がほぼ同数である。

なお、実験計画法としては、被験者内計画ではなく**被験者間計画** [5] を採用した。つまり、各訳文を複数の被験者に振り分けることにより、各被験者は必ず1種類の訳文のみを読むようにする。これは、一旦いずれかの訳文を読むと何らかの理解が生じ、別の訳文の読解に影響を与えてしまうからである。

5 実験結果

実験は、予備実験と本実験に大きく分かれる。

5.1 予備実験

被験者5名程度に、訳文7および質問文を渡し、質問に答えてもらった。この際、なるべく速く回答するよう指示した。また、実験方法そのものに関する感想なども聞いた。

全問正解した被験者は、4～5分程度回答に時間がかかった。今回の実験が、普通の読解速度における測定を目的としているため、4分を本実験の**標準回答時間**として設定した。また、じっくり読んだ時の読解も測定するため、標準時間の倍である8分を、**熟読回答時間**として設定した。

なお、予備実験で被験者より指摘のあった、わかりにくい質問や表現などは修正してから本実験に取り掛かった。

5.2 本実験

予備実験によって実験の問題点を一通り洗い出した後、200名規模の本実験に移った。

図1. は標準回答時間における読解正解率の平均値を表したものである。正解率は、5問全問正解で100％に、5問全問不正解で0％となる。各文章毎に正解率を計算してグラフ化してある。なお、比較のために、原文の英文に対する正解率を破線で同じグラフ上に記した。

同様に、図2. は熟読回答時間における読解正解率を表したものである。

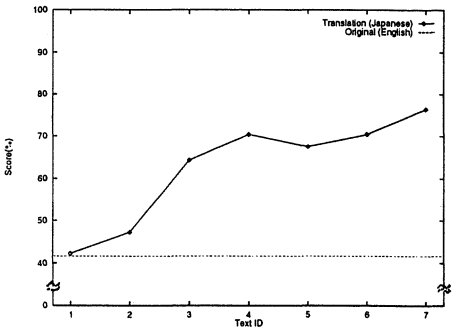

図1. 標準回答時間における文章毎の平均正解率

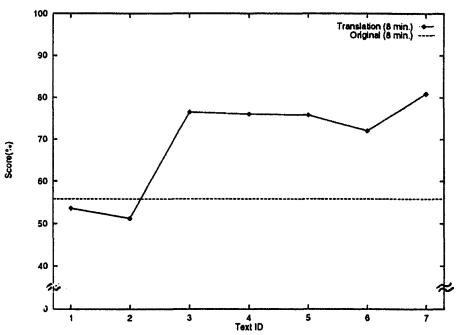

図2. 熟読回答時間における文章毎の平均正解率

グラフでは手翻訳の場合の正解率が100％になっていないが、これは質問として「ひっかけ」的なものがあるなど、微妙な内容となっているからだと思われる。また、問題は四者択一式なので「あてずっぽう」で回答した場合には25％の正解率が得られてしまうことに留意したい。さらに、本来は修正を加えるにつれて正解率が上がるはずであるが、実験の誤差によりそのようになっていない部分がある。

5.3 分析

本実験の最終目標は、どの言語現象が読解正解率に影響を与えるかを調べることである。これは、結果データの中で、どの文章からどの文章に移る際に計画的変化分（訳文の質による変化分）が現れるを分析することに等しい。

しかし、前項の実験結果のグラフは平均正解率をそのまま表したものであり、平均正解率は無計画的変化分（偶然のユレ）を含むものとなっている。このため、無計画的変化分を特定し計画的変化分から統計的に分離するために分散分析の手法を用いた。

5.3.1 分散分析

本実験の実験計画は、「２要因７水準の被験者間計画。被験者：２００名」（ABSタイプ）である。分散分析表を作成し、F比を求めた結果、分散分析結果が有意であることがわかった。

5.3.2 多重比較

分散分析の結果が有意であったため、多重比較を行なった。全文章の平均正解率から、２平均ずつの対を順次取りだし多重比較を行なった。

文章#	2	3	4	5	6	7
1	=	<	<	<	<	<
2		<	<	<	<	<
3			=	=	=	=
4				=	=	=
5					=	=
6						=

不等号：$p < 0.05$、等号：有意差なし

表 2. 文章対の多重比較結果

以上の多重比較結果から、訳文１と訳文２の平均正解率の間には有意差がなく、また訳文３から訳文７までの間の平均正解率の間にも有意差がないことがわかった。そして、前者の訳文群の平均正解率と後者の訳文群の平均正解率の間には有意さがあることがわかった。

本実験では、訳文１から始まって訳文２、訳文３... と訳文を順次修正しているが、訳文２から訳文３への修正の際に統計的有意差が認められたことになる。

6 結論

以上の実験結果から次のような結論を導き出した。

1. 長文読解に関しては、訳文２と訳文３の間に有意な差が認められた。つまり、対象とした機械翻訳システムでは複合語や句表現などが現在は不十分であるが、これらが登録されると読解性が確実に向上することがわかる。
2. 長文読解に関しては、係受け解析や文脈処理などの精度を上げても、読解正解率の有意な向上には結び付かない。

また、統計的な根拠は十分ではないが、以下のようなことも言えそうである。

1. 長文読解では、読解に時間をかけた方がより正確な理解が得られるという傾向がある。
2. 現状では、チューニングを行なっていない機械翻訳の和訳から得られる理解は、原文の英語から得られる理解とさほど変わらない。

7 おわりに

読むための翻訳の評価を、長文読解という観点から行なった。この結果、機械翻訳システムの翻訳精度向上のための一つの指標を得ることができた。

これからは、訳文が、長文読解のみならず他の用途のために使われた場合の評価を行ない、より総合的な結論を出したい。

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の海保博之教授ならびに同研究科の平山祐一郎氏、そして実験にご協力いただいた米沢女子短期大学の高尾哲康助教授に深く感謝いたします。

参考文献

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

[2] Eric Visser, Masaru Fuji, and Makoto Shiotsu. "the world on your shoulders: The development of atlas". In *MT Summit V*, July 1995.

[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] 成田一. 「翻訳ソフト こんなに使える！」. In *Professional English*, February 1995.

[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.