# 多言語話し言葉翻訳に関する変換主導翻訳システムの評価

古瀬 蔵　美馬 秀樹　山本 和英　Michael Paul　飯田 仁
E-mail:{furuse, mima, yamamoto, paul, iida}@itl.atr.co.jp
ATR 音声翻訳通信研究所

## 1 はじめに

多言語話し言葉翻訳では、多様ないい回しに対して、受け手が理解可能な翻訳結果をできるだけ速く出力する汎用的な枠組が必要である。筆者らは、多言語話し言葉翻訳の手法として変換主導翻訳 (Transfer-Driven Machine Translation, 以下、TDMT と呼ぶ)[2] を提案し、現在、旅行会話文を翻訳対象として、日英、日韓、日独、英日、韓日の話し言葉翻訳システムを構築している。

本稿では、TDMT システムの多言語話し言葉翻訳に対する性能と問題点を把握するために、1997年2月に行なった評価とその結果について報告する。各言語ペアの翻訳で、異なり1000文以上に対してオープンテストおよびクローズドテストを行ない、翻訳品質、原言語構造解析、翻訳時間について評価した。また、対訳用例学習量や翻訳言語ペアによる翻訳性能の違いなどに触れながら、今後取り組むべき多言語話し言葉翻訳の問題点について検討した。

## 2 TDMT システム

### 2.1 変換知識の利用

TDMT システムは、変換知識を用いて原言語構造および目的言語構造を導出する変換処理が翻訳処理の中心であり、多言語翻訳において原言語と目的言語のどのペアについても変換モジュールは共通の処理を行なう。形態素処理と生成処理については、各言語固有のモジュールを用意している。変換知識は原言語表現と目的言語表現の対応関係を意味的にまとまった単位でパタンによって記述し、原言語表現パタンごとに対訳用例を収集、編集することにより作られる。例えば、「ホテルの住所」→ "*the address of the hotel*" や「英語のパンフレット」→ "*the pamphlet in English*" などの対訳用例を収集し、日本語表現パタン「X の Y」についての日英の変換知識を作る。

TDMT は、最尤目的言語表現および最尤構造を決定するのに、用例に基づく手法を利用する。変換知識の中から入力に最も意味的に類似する対訳用例を意味距離計算により求め、その対訳用例を模倣することにより翻訳結果を得る [6]。例えば、翻訳の入力が「日本語のパンフレット」とする。「*X* の *Y*」に関する変換知識の中で「英語のパンフレット」が意味的に最も近ければ、*Y′ in X′* を使って "*the pamphlet in Japanese*" という翻訳結果を得る。対訳用例の学習量を増やすことは、訳し分け条件の精度向上につながり、高い構造解析成功率や翻訳率を達成するために欠かせない。

### 2.2 システムデータ

TDMT システムの翻訳対象は、音声翻訳を使用する場面を想定した旅行会話である。ATR では、通訳を介したバイリンガル模擬会話 [1] を収録したり、基本表現を網羅するための対訳つき基本表現集を作成して、言語データベースを構築している。言語データベースのトピックは、ホテルの予約、ホテルの紹介、ホテルでのサービス、乗物の切符購入、道案内、交通手段問い合わせ、観光ツアーの案内など旅行会話全般に渡っている。TDMT システムが広範囲に旅行会話文を翻訳できるように、この言語データベースを使って形態素辞書データや変換知識などの TDMT システムデータを構築している。

言語データベースはテキストデータや Tagged データなどから構成される。テキストデータからは、その文をシステムが翻訳ができるように対訳用例を追加し変換知識の更新を行なう (この作業を以下、翻訳訓練と呼ぶ)。翻訳訓練は、ホテルの予約に重点をおいて、さまざまなトピックの会話文や基本表現について行なっている。Tagged データからは、翻訳対象の語彙となる形態素辞書データなどを学習する。表 1は、1997年2月時点におけるシステム規模を示す。

## 3 評価方法

### 3.1 評価項目

今回の TDMT システムの評価項目は以下の通りである。

- 翻訳品質
  オープンテスト文、クローズドテスト文で評価
- 原言語構造解析
  オープンテスト文で評価
- 翻訳時間
  オープンテスト文で評価

翻訳品質に関してはさまざまな評価手法が提案されており、翻訳結果の品質をいくつかの尺度で採点する方法 [4] や、いろいろな言語現象を含む評価用例文に対する翻訳結果から評価項目をクリアしているかどうかを調べ、システム改良の参考データを求める方法 [3] などがある。ただし、これらはほとんど日英間の書き言葉翻訳

表 1: システム規模

| | 日英 | 日韓 | 日独 | 英日 | 韓日 |
|---|---|---|---|---|---|
| 辞書の語彙数 (概算) | 10000 | | | 6000 | 3000 |
| 翻訳訓練文数 (異なり) | 2602 | 1195 | 1553 | 2431 | 493 |
| 翻訳訓練文の平均語数 (異なり) | 10.1 | 9.0 | 9.3 | 8.4 | 7.5 |
| 変換知識のパタンの種類 | 887 | 624 | 787 | 1194 | 320 |

を対象としており、多言語話し言葉翻訳についての評価手法は提案されていない。

われわれは、翻訳対象である旅行会話での TDMT システムの性能を把握するために、ATR の言語データベースから評価例文を選定し、翻訳結果の品質を採点する方法を採った。また、システム改良の参考データとするために、翻訳結果に問題がある場合について、問題点や模範出力を評価者に記入させた。

原言語構造解析評価は、変換知識の原言語パタンを組合せた原言語構造を、システムが正しく導くことができるか評価した。入力全体の構造を正しく解析できていれば成功、部分的にでも誤った構造になっていれば失敗と判定する。

## 3.2 翻訳品質の評価値

それぞれの会話場面で翻訳結果が話者の発話内容を正しく相手に伝えられるか、という単一の尺度で翻訳品質を採点することにし、以下の 4 段階の評価値を設定した。

**評価値 A:**
問題なし。

**評価値 B:**
小さい問題はあるが、内容が容易に正しく理解できる。

**評価値 C:**
くずれた出力であるが、何とか内容が正しく理解できる。

**評価値 D:**
内容が理解できない。または、誤った内容を伝えている。

これらの評価値から、以下の二つのレベルの翻訳率を設定した。

(1) **評価値 A の割合**
翻訳結果として全く問題がない。

(2) **評価値 C 以上の割合**
最低、相手に自分の言ったことが伝わる。

翻訳システムとしては (1) の数値を高くすることも重要であるが、できるだけ汎用的に、異なる言語を使う話者の間の意思疎通を可能にするという観点から、われわれは TDMT システムの改良に関して (2) の数値を重視している。

## 3.3 評価者

翻訳品質を評価値で採点する方法は、評価結果に個人差が出ることが予想されるため、各翻訳での評価者を複数にした。評価者数は、日英で 3 名、日韓、日独、英日、韓日それぞれ 2 名である。また、翻訳品質の評価では、入力文の意味が出力に反映されているかを判断する必要があるので、原言語にも堪能な、目的言語のネイティブを評価者とした。

原言語構造解析評価では、TDMT の言語構造表現に習熟した 1 名を評価者とした。

## 3.4 評価例文

バイリンガル会話では、異なる言語を使う話者が通訳を介して対話を行なっているが、通訳者の発話は評価対象から除外し、オリジナルの発話部分のみを評価の対象とした。すなわち、日英翻訳の評価では日本語話者の発話のみを、英日翻訳の評価では英語話者の発話のみを評価対象とした。

評価例文は、話し言葉では状況に依存する表現が多いことが予想されるため、評価文が存在する文脈が分かるよう、相手話者の発話内容も文脈情報として含めて、会話単位で評価例文を評価者に示した。

翻訳品質のオープンテスト文は、できるだけ多くの基本語彙やさまざまな言語表現を含むように、異なりの文数が 1000 文以上となることを目安として、TDMT システムが翻訳訓練していないバイリンガル会話から無作為抽出した。日本語を入力とする日英、日韓、日独翻訳については、比較検討のため、同じ文を評価に使用した。表 2に、オープンテスト用に選定した会話と文の数を示す。

表 2: オープンテスト文

| | 日英、日韓、日独 | 英日 | 韓日 |
|---|---|---|---|
| 会話数 | 69 | 77 | 87 |
| のべ文数 | 1247(9.4 語 / 文) | 1323 (7.1) | 1169 (8.0) |
| 異なり文数 | 1021(11.0 語 / 文) | 1002 (8.8) | 997 (9.0) |

表 3: オープンテスト文に対する翻訳率 (平均値)

| | 日英 | 日韓 | 日独 | 英日 | 韓日 |
|---|---|---|---|---|---|
| 評価 A (のべ / 異なり、%) | 30.1 / 19.4 | 46.6 / 35.5 | 27.6 / 17.5 | 23.7 / 17.4 | 34.4 / 28.4 |
| 評価 C 以上 (のべ / 異なり、%) | 68.8 / 62.4 | 90.1 / 88.0 | 49.6 / 40.2 | 59.5 / 52.7 | 73.1 / 69.4 |

翻訳訓練による翻訳品質向上の見込みや TDMT の処理機能の問題点を把握するため、日英についてはクローズドデータに対する翻訳品質の評価も行なった。日英翻訳で翻訳訓練を行なった日英間バイリンガル 97 会話の日本人話者の発話 (のべ 1410 文、異なり 1144 文) を評価対象とした。オープンテスト同様、文脈が分かるよう会話単位で評価者に翻訳結果を示した。

### 3.5 評価の前提条件

本評価における主な前提条件を示す。

1. 本評価では、形態素解析以降の構造解析、変換、生成などの問題点を明らかにすることを目的とし、翻訳評価の入力はすべて正解形態素解析列とした。すなわち、形態素解析の誤りは排除して評価を行なった。形態素解析の性能は別途、評価を行なっている [7][1]。

2. 話し言葉翻訳という前提を考慮して、音声として現れない句点やコンマなどは入力形態素に含めなかった。また、音声的に曖昧なアラビア数字は入力、出力いずれにも使用しなかった。例えば、12 は、日本語では「一二」または「十二」、英語では"one two" または"twelve" と発声された音声により形態素を区別する。また、出力文の翻訳品質は会話中に音声で聞いた話し言葉として評価した。例えば、大文字と小文字、ハイフンなど、話し言葉に関係ない問題は無視した。

3. 最尤原言語構造解析結果や最尤翻訳結果が複数ある場合、システムが最初に提示する 1 個の解のみを評価対象とした。これは、音声対話で複数の翻訳結果を提示する状況は実際的でなく、複数の最尤解をできるだけ出さない方がシステムとして望ましいからである。

## 4 評価結果

### 4.1 翻訳品質オープンテスト

表 3に、オープンテスト文についての翻訳率を、複数評価者の平均値で示す。

[1]現在、1-best で形態素解析の文正解率は、日本語約 93%、英語約 81%、韓国語約 74%、処理時間は 10 形態素入力の文であれば、SPARCstation10 を使ってほぼ 0.2 秒以内である。

類似した言語を扱う日韓、韓日翻訳は翻訳率が高く、原言語と目的言語が言語的に遠い日英、英日、日独翻訳はより多くの翻訳訓練を行なう必要があることが示された。ただし、平均翻訳率をシステム性能の指標とするには評価者のばらつきの問題がある。図 1は、評価者の翻訳率のばらつきを翻訳訓練量とともに示す。翻訳率はのべで計算した値である。

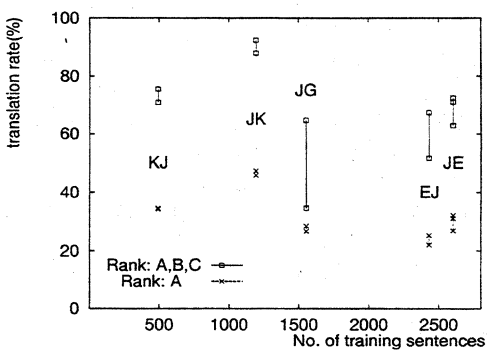


図 1: 評価者ごとの翻訳率

英日と日独で評価値 C 以上の翻訳率は評価者による差が大きかった。これは、言語的な遠さのため、原言語の表現に引きずられた目的言語表現が多く、これらの文に対する評価が分かれたのが主な原因である。英日と日独については、より客観的な翻訳率を得るために評価者数を増やす必要があると思われる。

### 4.2 翻訳品質クローズドテスト

表 4に、日英翻訳に関するクローズドテスト文の翻訳品質の評価結果を示す。

表 4: クローズドテスト文に対する日英の翻訳率

| | 評価者 | #1 | #2 | #3 | 平均 |
|---|---|---|---|---|---|
| 評価 A | のべ (%) | 54.4 | 34.8 | 40.4 | 43.2 |
| | 異なり (%) | 46.8 | 24.4 | 32.8 | 34.6 |
| 評価 C 以上 | のべ (%) | 96.2 | 89.3 | 91.7 | 92.4 |
| | 異なり (%) | 95.5 | 87.3 | 90.1 | 91.0 |

クローズドテスト文の翻訳品質が悪い原因として、出現した状況で表現が不適切、表現がこなれていない、などの指摘が多かった。この問題に対処するために生成の改良や文脈処理の導入を行なう必要がある。

### 4.3 原言語構造解析

表 5に、オープンテスト文に対する原言語構造解析の成功率を示す。

表 5: 原言語構造解析成功率

| | 日英 | 日韓 | 日独 | 英日 | 韓日 |
|---|---|---|---|---|---|
| のべ (%) | 76.2 | 67.4 | 66.1 | 72.6 | 51.2 |
| 異なり (%) | 70.9 | 60.2 | 58.7 | 63.9 | 43.3 |

成功率は、翻訳訓練量が最大の日英が最も高く、最小の韓日が最も低い。いずれの言語ペアの翻訳も、並列句などを含む長文は構造解析が困難になっている。

日英、英日、日独については、原言語構造解析で失敗すると、翻訳品質で良い結果を得ることは難しい。これらの翻訳は、訓練量をさらに増やして、原言語構造解析の精度と翻訳率を向上させる必要がある。一方、言語的に類似した日韓間の翻訳は、語順がほぼ同じで、同様の省略表現が出現することなどから、誤った構造解析結果でも理解可能な翻訳結果を得られることが多い。しかし、日韓での「かかる（対訳は、걸리다、들다等）」、韓日での「쓰다（対訳は、書く、使う等）」、日韓双方向での助詞などでは、正しい訳し分けを行なうために正しい依存関係を構造解析により求める必要がある。

### 4.4 翻訳時間

翻訳品質評価で使ったオープンテスト文について翻訳時間を計測した。翻訳時間の計測には256MB メモリを持つSPARCstation10を使用した。システムはCommon Lispで記述されている。図 2は、各翻訳における入力形態素数と翻訳に要したCPU timeの関係を示す。CPU timeは、異なりのクローズド文に対して、各形態素数ごとに計算した平均値である。ただし、形態素解析の時間は含めていない。

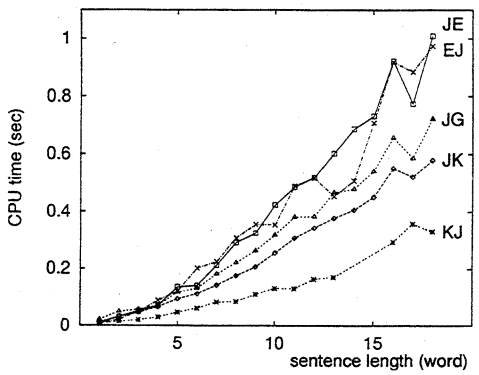


図 2: 入力形態素数と翻訳時間

いずれの言語ペアの翻訳でも高速な翻訳処理を実現している。翻訳訓練量が多い日英、英日は他の翻訳に比べて処理時間が長い。また、原言語構造解析を失敗すると、成功した場合に比べてやや処理時間が長くなる傾向が見られた。

## 5 おわりに

多言語話し言葉翻訳を行なうTDMTシステムの評価とその結果について述べた。評価結果により、現在、TDMTシステムは多くの旅行会話文を、相手の意図が理解可能な文に短時間で翻訳していることを確認した。また、原言語と目的言語が言語的に遠い日英、英日、日独翻訳はより多くの翻訳訓練を行なう必要があること、生成モジュールの改善や文脈処理の導入が今後の翻訳品質向上には必要であること、などの問題点も示された。

今後、評価用例文[5]による言語現象の網羅性の調査なども含めて、適宜、翻訳評価を行ないながら、システムの性能、問題点を明らかにしていく予定である。評価結果に基づいて翻訳処理機能の改良、および効率的な翻訳訓練を行なうことによりTDMTシステムの性能を向上させ、より高性能な多言語話し言葉翻訳の実現を目指す。

## 参考文献

[1] Furuse, et al.: "Bilingual Corpus for Speech Translation", *Proc. of AAAI'94 Workshop 'Integration of Natural Language and Speech Processing* (1994).

[2] Furuse, et al.: "Incremental Translation Utilizing Constituent Boundary Patterns", *Proc. of Coling '96* (1996).

[3] 井佐原 他: "開発者の視点からの機械翻訳システムの技術的評価 – テストセットを用いた品質評価法 –" 自然言語処理, Vol.3, No.3 (1996).

[4] 長尾 他: "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 NL47-11 (1985).

[5] 日本電子工業振興協会: "機械翻訳評価基準 – 品質評価用テストセット –" 日本電子工業振興協会 95-計-17 (1995).

[6] Sumita, et al. "Example-based transfer of Japanese adnominal particles into English" IEICE Transactions on Information and Systems, E75-D, No.4, (1992).

[7] 山本 他: "単語と品詞の混合 n-gram を用いた形態素解析" 情報処理学会第 54 回全国大会講演論文集 (1997).