TR-IT-0189

日英間変換主導翻訳の中間時評価

Report on Mid-term performance evaluation of Transfer-Driven Machine Translation between Japanese and English

美馬 秀樹　　古瀬 蔵
Hideki MIMA　Osamu FURUSE

1996 年 11 月 14 日

概要

本稿では、変換主導翻訳機構（Transfer-Driven Machine Translation, 以下、TDMT と呼ぶ）に対し行った中間時評価試験の内容、及び日英、英日翻訳の評価結果について述べる。
著者らは、経験的な言語知識の利用という統一的な処理機構による多言語間の対話翻訳の実現を目的とし、変換主導翻訳機構について研究を進めている。TDMT では、多言語による話し言葉に対し、実時間でのコミュニケーションのための効率的な翻訳処理、文法から逸脱した表現等の多用な言い回しに対する頑健な翻訳処理が実現可能となる。
翻訳システムの評価においては、評価基準の設定や評価結果の解釈等、難しい問題が存在するが、話し言葉翻訳に対しては、その機能的な側面より、会話としての自然さや、理解の容易さが重要であると考え、本評価では、翻訳結果の理解可能性や、伝わる情報の正確さに重点を置いた評価を行った。

ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

目次

第 1 章

概要

本稿では、変換主導翻訳機構（Transfer-Driven Machine Translation, 以下、TDMT と呼ぶ）に対し行った中間時評価試験の内容、及び日英、英日翻訳の評価結果について述べる。

著者らは、経験的な言語知識の利用という統一的な処理機構による多言語間の対話翻訳の実現を目的とし、変換主導翻訳機構について研究を進めている [古瀬 94][Furuse94][古瀬 95a][Furuse96]。

TDMT では、多言語による話し言葉に対し、実時間でのコミュニケーションのための効率的な翻訳処理、文法から逸脱した表現等の多用な言い回しに対する頑健な翻訳処理が実現可能となる。

このような翻訳システムの評価方法において、"can" の訳し分け等の機能的な品質評価、原言語の意味内容がどの程度伝えられるかというような翻訳正当性の評価 [安藤 96] などが提案されているが、評価基準の設定や評価結果の解釈等、難しい問題が存在する。しかし、話し言葉翻訳においては、その機能的な側面より、会話としての自然さや、理解の容易さが重要であると考えられるので、本評価では、翻訳結果の理解可能性や、伝わる情報の正確さに重点を置くことにした。

以下、第 2章で，今回行った評価方法、第 3章で，評価結果，第 4章で，翻訳精度と形態素数との関連，原言語構造評価との関連等の評価結果における諸データ，第 5章で，翻訳速度について示し，第 6章で，まとめと今後の課題について述べる．尚、日韓、韓日翻訳についての評価は [山本 96b]、日独については [Paul96] で報告されている。

第 2 章

評価方法

2.1 評価項目

TDMT において、以下の項目に対する評価を行う。

(1) 翻訳精度

(2) 原言語構造

(3) 翻訳速度

2.2 評価文の選定

- 形態素解析の問題を翻訳評価から切り放すため、入力はストリングではなく、タギングデータとした。したがって、本評価では、翻訳における形態素解析誤りは出現しない。尚、TDMT 用形態素解析についての評価は [山本 96a] で報告されている。
- オープンテスト文は、音声言語データベース [浦谷 94]、及び言語データベース [古瀬 95b] における日英バイリンガル会話、及び日英韓基本表現集より無作為選定。日英、日韓、日独翻訳のオープンテストでは言語ペアの差を観測するため、同一のバイリンガル会話文を使用した。
- 選定されたバイリンガル会話評価文中の単語はデータベース全体の単語に対しのべで 92 % 程度カバーするなど、高頻度語、多会話に現れる語をほぼ網羅している。
- バイリンガル会話はネイティブ話者部分の文のみを評価対象とする。
- 翻訳処理時におけるパタンマッチングのビーム幅は 1 のみとする。

- 予備試験で有為な差が見られなかったため。

- 評価文には文脈を与える。
  - バイリンガル会話は、目的言語側のネイティブ話者の発話（日英翻訳評価では英語話者部分）を付与。
  - 基本表現集には、発話状況、及び話者情報を付与。

- 日英翻訳に関しては、言語翻訳研究の中心テーマであることから、日英バイリンガル会話、及び日英韓基本表現集での翻訳訓練済みのクローズド文のうちバイリンガル会話のネイティブ話者部分について翻訳精度の評価を行った。表2.1に、各翻訳処理に対し、選定の対象となった評価文とその文数を示す。

表 2.1: 選定の対象となった評価文

(a) オープンテスト

	バイリンガル会話	基本表現集
日英	日英会話より69会話（異なり1000文以上)	500文（20トピック25文ずつ）
日韓	日英と同会話	なし
日独	日英と同会話	なし
英日	日英会話より77会話（異なり1000文以上）	なし
韓日	日韓会話より92会話（異なり1000文以上）	なし

(b) クローズドテスト

	バイリンガル会話	基本表現集
日英	全訓練会話75会話（のべ1500文以上）	なし

原言語構文構造の評価は、表2.1(a)におけるバイリンガル会話に対し、すべての翻訳について行う。

2.3 評価基準

2.3.1 翻訳精度

翻訳結果は5段階評価とし、翻訳出力にA,B,C,Dのランクづけを行う。本翻訳評価試験においては、C以上の評価を翻訳成功と考える。

A : **No problem**、問題なし

B : **Fair**、小さい問題点はあるが、表現は整っており容易に内容が理解できる

C : **Acceptable**、何とか内容が理解できる

D : **Nonsense**, Wrong sense、内容が理解できない、入力と全く異なる内容

NIL : 出力なし

上記結果がb,c,dの場合、次のうちより問題点を指摘する（複数回答可）。

1. **Word Selection**、語の選択が不適切
   - おいしい店 → delicious store
   - 京都の寺 → a temple for Kyoto
   - 私は学校を行った、 私田中ます
2. **Conjugation**、出現語の形態、活用
   - 走るなさい
   - I goed yesterday
   - two book (bookes)
3. **Word Order**、語順の誤り、主格や目的格の欠落による構文誤り (with syntactic error)
   - I prefer a hotel cheap
   - I give her the book to Mary
   - I give to Mary
4. **Sentence Style**、文体 (without syntactic error)

- ぎこちない、native は使わない表現

5. **Information**、情報が不適切

不足：昨日そこへ行った　→　I went there

過多：そこへ行った　→　I went there yesterday

- 情報の不適切性により、非文法的になる場合は、項目 3 （Word Order）へ

6. **Determiner**、冠詞や限定詞

- two milk、A man you saw is his father

7. **Others**、その他 （具体的内容を下欄に、日本語か英語で記入）

きれいな文でも文脈上不適切の時は項目 5、結果がめちゃくちゃの時は項目 7 とする。

評価上の注意事項

- 対象はあくまで話し言葉 (Spoken language) である
- 出力結果は、文字でなく音で聞いたと仮定する
- 大文字、小文字、ハイフン、句読点 (punctuation) の問題は無視する
- 同じ出力結果でも、文脈に応じて評価ランクが変わることもありうる
- mr-ms Suzuki （鈴木様）など文脈でも決められない場合は誤りとしない（他にも、sir-madam, he-she)

2.3.2　原言語構造評価

原言語構造に対しては 以下に示す基準において、○、×、NIL の 3 段階で評価する。

原言語構造評価基準

- 各言語の構造が構文として正しいか
- 係り先が意味的に正しい
- 用いたルールのカテゴリが正しいか

- ルールの変数に対する制限が正しいか

評価に対する注意事項

- 名詞連続や助詞抜けでマーカを用いたルール、助詞「で」や係助詞「は」及び副詞など、係り先が複数考えられるものについては意味的に適切な構造であるならば正解とする。
- 日英／日独／日韓それぞれにルール作成方針の違いが見られるが、ターゲットとなる言語を翻訳するために必要と思われる例外処理は評価の考慮に入れる。
- 原言語の表現自体に乱れ、ねじれがある場合は評価の基準を若干ゆるめる。

2.4 評価者

翻訳精度評価者には原言語と目的言語の両方に堪能であること、原言語構文構造評価者は、TDMT の翻訳データ作成に習熟していることをそれぞれ条件とし、表 2.2に示す人数により評価を行った。日英翻訳は、中心的研究テーマであることから、翻訳精度の評価者数を他の翻訳より多くしている。

表 2.2: 各翻訳処理別の評価者数

	翻訳精度	原言語構文構造
日英	3名	1名
日韓	2名	1名
日独	2名	1名
英日	2名	1名
韓日	2名	1名

第 3 章

評価結果

3.1 翻訳評価

3.1.1 日英オープン (バイリンガル会話)

3 評価者の平均

Rank	のべ (%)	異なり (%)
A	25.8	14.5
B	16.5	16.2
C	17.9	21.3
D	25.6	30.7
NIL	14.2	17.3
C 以上	60.2	52.0

・評価者別

評価者 1

Rank	のべ	(%)	異なり	(%)
A	312	25.020055	142	13.907935
B	253	20.288695	201	19.686585
C	146	11.70815	145	14.201765
D	359	28.78915	356	34.867775
NIL	177	14.194075	177	17.335955
sum	1247	-	1021	-
C 以上	711	57.016840	488	48.188051

評価者 2

Rank	のべ	(%)	異なり	(%)
A	348	27.906985	170	16.650345
B	101	8.0994395	69	6.758085
C	341	27.345635	333	32.615085
D	280	22.453895	272	26.640555
NIL	177	14.194075	177	17.335955
sum	1247	-	1021	-
C 以上	790	63.352045	572	56.023506

評価者 3

Rank	のべ	(%)	異なり	(%)
A	304	24.378515	133	13.026445
B	264	21.170815	225	22.037225
C	184	14.755415	173	16.944175
D	318	25.50125	313	30.656225
NIL	177	14.194075	177	17.335955
sum	1247	-	1021	-
C 以上	752	60.304731	531	52.007835

3.1.2 日英オープン (基本表現集)

3 評価者の平均

Rank	のべ (%)	異なり (%)
A	12.8	11.0
B	11.5	14.5
C	33.2	30.2
D	33.1	34.9
NIL	9.3	9.4
C 以上	57.5	55.7

・評価者別

評価者 1

Rank	のべ	(%)	異なり	(%)
A	36	7.1428585	36	7.2
B	103	20.436515	102	20.4
C	122	24.206355	121	24.2
D	196	38.888895	194	38.8
NIL	47	9.3253975	47	9.4
sum	504	-	500	-
C 以上	261	51.785714	259	51.8

評価者 2

Rank	のべ	(%)	異なり	(%)
A	74	14.682545	74	14.8
B	22	4.365085	22	4.4
C	234	46.428575	232	46.4
D	127	25.198415	125	25.0
NIL	47	9.3253975	47	9.4
sum	504	-	500	-
C 以上	330	65.476190	328	65.6

評価者 3

Rank	のべ	(%)	異なり	(%)
A	55	10.91275	55	11.0
B	94	18.65085	93	18.6
C	101	20.039685	100	20.0
D	207	41.071435	205	41.0
NIL	47	9.3253975	47	9.4
sum	504	-	500	-
C以上	250	49.603175	248	49.6

バイリンガル会話と基本表現の評価の差

- バイリンガル会話の特徴
  - 使用頻度の高い文型が多く出現し、また、”もしもし”等の exact-match も多い
  - より実際の会話状況に近い
- 基本表現集の特徴
  - 模擬的な表現
  - 文型の偏りが少なく、のべ文数と異なり文数の差が小さい
  - （文脈を設定しているが）基本的に文単位の翻訳評価

バイリンガル会話の評価における異なり文数での集計に相当すると思われる。

3.1.3 日英クローズド (バイリンガル会話)

3 評価者の平均

Rank	のべ (%)	異なり (%)
A	40.2	26.9
B	24.0	36.1
C	28.3	28.0
D	7.5	9.0
C以上	92.5	91.0

・評価者別

評価者 1

Rank	のべ	(%)	異なり	(%)
A	540	46.03585	350	36.842115
B	163	13.895995	155	16.315795
C	435	37.08445	412	43.368425
D	35	2.9838025	33	3.4736845
sum	1173	-	950	-
C 以上	1138	97.016198	917	96.5263157

評価者 2

Rank	のべ	(%)	異なり	(%)
A	403	34.356355	224	23.578955
B	399	34.015355	364	38.315795
C	228	19.437345	225	23.684215
D	143	12.190965	137	14.421055
sum	1173	-	950	-
C 以上	1030	87.8090366	813	85.578947

評価者 3

Rank	のべ	(%)	異なり	(%)
A	349	29.752775	193	20.315795
B	576	49.104865	511	53.789475
C	160	13.640245	160	16.842115
D	88	7.5021315	86	9.0526315
sum	1173	-	950	-
C 以上	1085	92.4978687	864	90.9473684

3.1.4 英日オープン (バイリンガル会話)

2 評価者の平均

Rank	のべ (%)	異なり (%)
A	29.9	18.7
B	19.1	19.2
C	10.8	13.6
D	25.5	29.0
NIL	14.7	19.5
C 以上	59.8	51.5

・評価者別

評価者 1

Rank	のべ	(%)	異なり	(%)
A	375	28.344675	159	15.868265
B	364	27.513235	271	27.045915
C	125	9.4482235	119	11.876255
D	264	19.954655	258	25.74855
NIL	195	14.739235	195	19.461085
sum	1323	-	1002	-
C以上	864	65.306122	549	54.790419

評価者 2

Rank	のべ	(%)	異なり	(%)
A	414	31.292525	215	21.457095
B	142	10.733185	114	11.377245
C	162	12.24495	154	15.369265
D	410	30.990185	324	32.335335
NIL	195	14.739235	195	19.461085
sum	1323	-	1002	-
C以上	718	54.270597	483	48.203593

3.2 翻訳失敗の原因

3.2.1 オープンデータ評価

1. NIL（翻訳出力なし）の原因
   - 原言語の解析失敗。パターンが未登録（学習量の不足）
     - “十月二十八日 ですと 二万三千円となりますけれども”
       ’(X ですと Y)’ 等のパターンがあれば翻訳可能
2. 表現があいまい
   - ”all right” → ”結構です”
3. 表現が冗長
   - ”here at the hotel” → ”ホテルにここに”
4. 原言語の表現が話し言葉としても特殊
   - ”京都名産品 で お持ち帰りいただけるよう当ホテルでご用意してあるものです”
   - ”運行日は毎週金曜日土曜日日曜日あと祝日に 行っております”
5. 翻訳結果出力のあいまい性
   - 翻訳結果の候補として内部ではより良いものが出力されているものがあるが、最初に得られた翻訳結果のみを出力している
     - 翻訳結果が複数できるものが、全翻訳出力に対し 10% 程度あると思われる

3.2.2 日英クローズドでのD評価の原因

日英オープンでも以下のような傾向は見られるが、クローズドでの処理であるにもかかわらず翻訳不可の原因として見受けられたもの。

1. 原文に忠実に訳しているため、英語らしさが損なわれている

   - ”バスの時刻のほうはよろしいですか” → ”Is the time of the bus good?”
     ”Do you want a bus timetable?”
   - ”見て回る” → ”see around” ; ”see it all” が正解
   - ”サムニコラス様でいらっしゃいますね” → ”You are Mr. Sam Nicolas 〜”
     ”The name is 〜”が自然

2. 英語としては冗長な表現

   - “場所はどちらに伺えばよろしいんでしょうか”→ “as for the locaion, where should I come”
     ”as for 〜” は不必要
   - “OK”、”All right” のつながりに対し、” 結構です” を連発
     – 訳出不要な表現の扱いが重要

3. 状況、立場の認識

   - ”はいそちらで結構です” → ” Yes, that is fine”
     ”Thank you very much”ぐらいがよい
   - “sure” の状況に応じた訳し分け等。”Do you know 〜” 等の問いに対しては現状の” もちろんです” でも不自然ではないが、”Could you tell me 〜” 等の依頼表現に対しては、” 分かりました” などに訳し分けしなくてはいけない
   - 客に対し “お名前は何ですか”、” 何人いますか” 等の立場的に不適切な表現が多い

4. 主語の間違い

   - ”you”, ”he” ←→ ”I”

5. 訳語選択に問題

表 3.1: 各翻訳処理の学習量と NIL 数

	英日	日英	日韓	日独	韓日
パターン数	1274	991	680	623	311
用例数	7008	10241	3605	2935	1701
NIL数(オープン)のべ -異なり(%)	15% - 19%	14% - 17%	28% - 34%	30% - 36%	41% - 47%

同じ評価文を使用

3.3 登録パターンと NIL 数の関連

3.2で述べたように、翻訳結果が NIL となる原因のほとんどがパターンの未登録と考えられるため、概ね、パターン数が少ないものほど NIL 数が増える結果となっている。特に、表 2.1にある通り、今回の評価実験においては、日英、日韓、日独の翻訳には同じ入力を用いているため、表 3.1の日英、日韓、日独での比較より、登録パターン数が多いほど、翻訳処理で NIL （翻訳出力なし）となる割合が低くなると言える。

3.4 評価結果と意味距離との関係

TDMT においては、意味距離が小さいほど、より意味的に類似した対訳用例が翻訳処理に適用されたと考えられる。以下では、それぞれの評価結果と意味距離の関係を考察する。尚、以下の表中での morph、 distance の値は、それぞれの平均値である。また、比較用に、一形態素当りの意味距離 distance/morph の値を d/m に示す。

3.4.1 日英オープン（バイリンガル会話）

評価者 1

Rank	count	morph	distance	d/m
A	348	4.158	0.084	0.020
B	101	6.455	0.274	0.042
C	341	10.226	0.933	0.091
D	280	12.336	1.611	0.131
NIL	177	15.542	-	-

評価者 2

Rank	count	morph	distance	d/m
A	304	3.901	0.050	0.013
B	264	7.750	0.465	0.060
C	184	9.543	0.906	0.095
D	318	12.742	1.640	0.129
NIL	177	15.542	-	-

評価者 3

Rank	count	morph	distance	d/m
A	312	3.865	0.070	0.018
B	253	7.115	0.300	0.042
C	146	10.925	1.017	0.093
D	359	12.365	1.616	0.131
NIL	177	15.542	-	-

3.4.2 日英オープン（基本表現集）

評価者 1

Rank	count	morph	distance	d/m
A	74	5.743	0.389	0.068
B	22	6.773	0.611	0.090
C	234	6.940	0.914	0.132
D	127	7.945	1.298	0.163
NIL	47	10.085	-	-

評価者 2

Rank	count	morph	distance	d/m
A	55	5.764	0.390	0.068
B	94	6.021	0.597	0.099
C	101	7.050	0.872	0.124
D	207	7.787	1.234	0.158
NIL	47	10.085	-	-

評価者 3

Rank	count	morph	distance	d/m
A	36	5.722	0.267	0.047
B	103	6.184	0.662	0.107
C	122	6.762	0.848	0.125
D	196	7.852	1.223	0.156
NIL	47	10.085	-	-

3.4.3 英日オープン（バイリンガル会話）

評価者 1

Rank	count	morph	distance	d/m
A	375	2.744	0.057	0.021
B	364	6.162	0.328	0.053
C	125	8.280	0.985	0.119
D	264	10.174	1.423	0.140
NIL	195	10.733	-	-

評価者 2

Rank	count	morph	distance	d/m
A	414	3.727	0.162	0.043
B	142	6.965	0.409	0.059
C	162	8.136	0.715	0.088
D	410	7.666	0.972	0.127
NIL	195	10.733	-	-

形態素数が多いほど適用されるパターン数が増えると考えられ、また、パターン毎の意味距離の総和が文の意味距離として計算されるため、パターン数に比例して、意味距離も増加する。したがって、比較の基準となるパターン当りの意味距離を、単純に、形態素当りの意味距離（d/m）で近似できるものと考えても、意味距離が小さくなるほど、評価が良くなると言える。つまり、用例のバリエーションが多いほど、より良い翻訳が行えると考えられる。

3.5 評価結果に対する考察

1. 「登録パターンと NIL 数の関連」より登録パターン数が増えれば NIL 数が減少すると考えられる。

2. 「評価結果と意味距離の関連」より用例数が増えればより良い翻訳結果が出力できると考えられる。

したがって、今後、学習量が増えるに連れて、翻訳結果の正当率（評価 C 以上）も上がる可能性がある。しかし、今後の学習量の増加に伴い、登録用例数、登録パターン数の飽和点、つまり、どれくらいの量まで増やせるのかを（対象言語毎に）見極めておく必要があると思われる。

3.6 評価者の指摘した問題点

3.6.1 日英オープン（バイリンガル会話）

評価者 1

Rank	selection	conjugation	order	style	info	determiner
B	8	5	2	79	1	21
C	71	56	63	306	18	45
D	106	111	114	254	110	15
計	185	172	179	639	129	81

評価者 2

Rank	selection	conjugation	order	style	info	determiner
B	75	42	31	136	14	39
C	94	74	57	49	24	36
D	235	171	209	31	54	115
計	404	287	297	216	92	190

評価者 3

Rank	selection	conjugation	order	style	info	determiner
B	15	3	3	238	2	13
C	75	27	43	130	6	10
D	292	112	236	281	39	14
計	382	142	282	649	47	37

3.6.2 日英オープン（基本表現集）

評価者 1

Rank	selection	conjugation	order	style	info	determiner
B	4	0	0	7	1	12
C	48	23	17	175	13	42
D	45	18	27	77	43	2
計	97	41	44	259	57	56

評価者 2

Rank	selection	conjugation	order	style	info	determiner
B	43	24	5	30	12	35
C	73	52	30	9	27	36
D	167	130	94	23	87	80
計	283	206	129	62	126	151

評価者 3

Rank	selection	conjugation	order	style	info	determiner
B	38	16	16	59	11	23
C	86	41	32	52	38	27
D	151	96	79	27	111	38
計	275	153	127	138	160	88

3.6.3 日英クローズド（バイリンガル会話）

評価者 1

Rank	selection	conjugation	order	style	info	determiner
B	18	11	4	79	13	64
C	109	78	32	363	43	38
D	9	7	4	22	18	3
計	136	96	40	464	74	105

評価者 2

Rank	selection	conjugation	order	style	info	determiner
B	159	117	43	180	35	119
C	161	139	69	83	32	90
D	113	88	43	40	32	37
計	433	344	155	303	99	246

評価者 3

Rank	selection	conjugation	order	style	info	determiner
B	15	3	8	564	0	26
C	30	6	8	155	0	8
D	41	5	14	75	1	2
計	86	14	30	794	1	36

3.6.4 英日オープン (バイリンガル会話)

評価者 1

Rank	selection	conjugation	order	style	info	determiner
B	18	2	3	159	155	0
C	41	2	5	25	27	1
D	130	2	38	1	28	0
計	189	6	46	185	210	1

評価者 2

Rank	selection	conjugation	order	style	info	determiner
B	18	10	6	95	26	0
C	57	28	12	89	27	0
D	296	68	83	70	96	0
計	371	106	101	254	149	0

- 日英
  - 日英では、style(ニュアンスの違い、ネイティブはそうは言わない等) の問題が多い
  - オープンテストでは特に、Conjugation(活用)、Order(ワードオーダー) の指摘が多くなる
- 英日
  - 英日では、Selection(訳語選択) に対する指摘が多くなる
  - 英語をそのまま訳したための情報量の過多の指摘も多い

したがって、訳語選択と、文生成系の改善が当面の課題と思われる。

3.7 評価方法に関する考察

評価結果、及び評価方法に関して、以下のような点で今後考察の余地があると思われた。

1. 評価者の主観
   - "All right." → " 結構です " に対して訳語選択に問題があるため D 判定
2. 評価のばらつき

- A 判定、D 判定は比較的収束
  - 3 人の評価者の判定が BCD、DDC のように別れた場合の判定をどう扱うか

3. 評価方法

- 評価基準があいまい (特に B,C 評価)
- 対話に対する翻訳の評価方法として十分とは言えない（対話相手の立場、対話の状況等を考慮に入れ、どのくらい自然な対話が行えるか等)
- 紙面で評価を行うため、書き言葉翻訳としての評価を行いがちになってしまう

第 4 章

評価結果諸データ

4.1 翻訳精度と形態素数との関連

4.1.1 日英オープン (バイリンガル会話)

評価対象文一文あたりの平均形態素数: 9.5

評価者 1

Rank	count	ave.	max.	min.
A	348	4.158	19	1
B	101	6.455	20	1
C	341	10.226	29	1
D	280	12.336	33	1
NIL	177	15.542	39	2

評価者 2

Rank	count	ave.	max.	min.
A	304	3.901	18	1
B	264	7.750	23	1
C	184	9.543	25	1
D	318	12.742	33	1
NIL	177	15.542	39	2

評価者 3

Rank	count	ave.	max.	min.
A	312	3.865	17	1
B	253	7.115	23	1
C	146	10.925	25	1
D	359	12.365	33	1
NIL	177	15.542	39	2

4.1.2 日英オープン（基本表現集）

評価対象文一文あたりの平均形態素数: 7.3

評価者 1

Rank	count	ave.	max.	min.
A	74	5.743	14	1
B	22	6.773	11	4
C	234	6.940	18	1
D	127	7.945	24	2
NIL	47	10.085	22	4

評価者 2

Rank	count	ave.	max.	min.
A	55	5.764	12	1
B	94	6.021	12	2
C	101	7.050	16	2
D	207	7.787	24	1
NIL	47	10.085	22	4

評価者 3

Rank	count	ave.	max.	min.
A	36	5.722	12	1
B	103	6.184	16	2
C	122	6.762	18	2
D	196	7.852	24	1
NIL	47	10.085	22	4

4.1.3 日英クローズド（バイリンガル会話）

評価対象文一文あたりの平均形態素数: 9.2

評価者 1

Rank	count	ave.	max.	min.
A	540	6.670	27	1
B	163	11.313	28	1
C	435	11.329	31	1
D	35	11.143	25	3

評価者 2

Rank	count	ave.	max.	min.
A	403	5.439	27	1
B	399	10.604	28	1
C	228	11.750	28	1
D	143	11.622	31	1

評価者 3

Rank	count	ave.	max.	min.
A	349	5.903	28	1
B	576	9.826	28	1
C	160	12.250	28	1
D	88	12.318	31	1

4.1.4 英日オープン（バイリンガル会話）

評価対象文一文あたりの平均形態素数: 6.9

評価者1

Rank	count	ave.	max.	min.
A	375	2.744	15	1
B	364	6.162	21	1
C	125	8.280	23	1
D	264	10.174	27	1
NIL	195	10.733	26	2

評価者2

Rank	count	ave.	max.	min.
A	414	3.727	19	1
B	142	6.965	22	1
C	162	8.136	22	1
D	410	7.666	27	1
NIL	195	10.733	26	2

評価は，概ね長い文になるほど悪くなるといえる。

4.2 翻訳精度と原言語構造評価との関連

以下の表に、翻訳精度と原言語構造評価との関係を示す。尚、表中の割合は、それぞれの構造評価に対する、翻訳評価結果の割合である。例えば、日英オープン（バイリンガル会話）での評価者1において、構造評価が○にもかかわらず翻訳評価ではDとなったものが18.8%あることを示す。

4.2.1 日英オープン（バイリンガル会話）

評価者1

Rank	○	(%)	×	(%)	NIL	(%)
A	345	38.3	3	1.8	0	0.0
B	99	11.0	2	1.2	0	0.0
C	286	31.8	55	32.2	0	0.0
D	169	18.8	111	64.9	0	0.0
NIL	1	0.1	0	0.0	176	100.0
合計	900	-	171	-	176	-

評価者2

Rank	○	(%)	×	(%)	NIL	(%)
A	301	33.4	3	1.8	0	0.0
B	254	28.2	10	5.8	0	0.0
C	156	17.3	28	16.4	0	0.0
D	188	20.9	130	76.0	0	0.0
NIL	1	0.1	0	0.0	176	100.0
合計	900	-	171	-	176	-

評価者 3

Rank	○	(%)	×	(%)	NIL	(%)
A	310	34.4	2	1.2	0	0.0
B	249	27.7	4	2.3	0	0.0
C	119	13.2	27	15.8	0	0.0
D	221	24.6	138	80.7	0	0.0
NIL	1	0.1	0	0.0	176	100.0
合計	900	-	171	-	176	-

4.2.2 英日オープン (バイリンガル会話)

評価者 1

Rank	○	(%)	×	(%)	NIL	(%)
A	374	40.7	1	0.5	0	0.0
B	339	36.8	25	11.8	0	0.0
C	96	10.4	29	13.7	0	0.0
D	108	11.7	156	73.9	0	0.0
NIL	3	0.3	0	0.0	192	100.0
合計	920	-	211	-	192	-

評価者 2

Rank	○	(%)	×	(%)	NIL	(%)
A	402	43.7	12	5.7	0	0.0
B	130	14.1	12	5.7	0	0.0
C	130	14.1	32	15.2	0	0.0
D	255	27.7	155	73.5	0	0.0
NIL	3	0.3	0	0.0	192	100.0
合計	920	-	211	-	192	-

原言語構造評価単体では、日英で、全評価文 1247 に対し、○が 900(72%)、英日で、全評価文 1323 に対し、○が 920(70%) と概ね 70% 程度が成功となり、良好な結果であると言える。しかし、翻訳評価との関連では、構造評価が○であるにもかかわらず、翻訳評価が D となるものが日英では 20 ～ 25%、英日では 10 ～ 28% 程度あり、翻訳における変換、または生成処理に少なからず問題があると考えられる。 また、構造解析が成功しているにもかかわらず、翻訳結果が NIL となるもの (日英で 1 文、英日で 3 文) が存在するが、これらは、変換、及び生成処理において処理失敗しているものと思われる。これらを含めて、今後、変換、及び生成処理の改善を重点的に進める必要がある。

第 5 章

翻訳速度

翻訳速度の評価は、翻訳出力を得ることが保証されているクローズドデータを対象として行われた。尚、先に述べたように、本評価では評価文としてタグ付けされたデータを用いているため、翻訳処理時間には、形態素解析に要する時間は含まれていない。

5.1 日英クローズド

図 5.1に形態素数と翻訳時間との関係をグラフで示す。

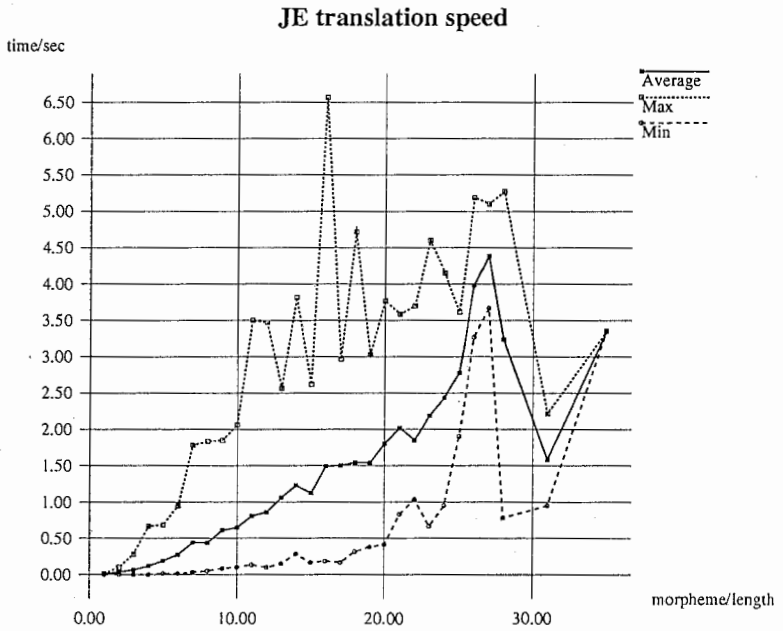

図 5.1: 形態素数と翻訳時間との関係（日英）

- 評価対象文一文あたりの平均形態素数: 9.2

形態素数	平均 (sec.)	最長 (sec.)	最短 (sec.)	文数
1	0.0096	0.0170	0.0000	21
2	0.0384	0.1000	0.0000	26
3	0.0650	0.2830	0.0000	59
4	0.1169	0.6670	0.0000	85
5	0.1875	0.6840	0.0160	76
6	0.2742	0.9500	0.0160	90
7	0.4435	1.7830	0.0340	92
8	0.4380	1.8330	0.0500	100
9	0.6120	1.8500	0.0840	94
10	0.6461	2.0670	0.1000	87
11	0.8068	3.5000	0.1330	113
12	0.8620	3.4660	0.1000	80
13	1.0607	2.5670	0.1500	50
14	1.2295	3.8160	0.2830	60
15	1.1230	2.6170	0.1660	45
16	1.4931	6.5670	0.1840	46
17	1.5038	2.9660	0.1670	39
18	1.5449	4.7170	0.3170	29
19	1.5413	3.0330	0.3830	27
20	1.8036	3.7660	0.4170	19
21	2.0188	3.5830	0.8330	16
22	1.8500	3.7000	1.0330	15
23	2.1872	4.5990	0.6660	13
24	2.4348	4.1500	0.9500	11
25	2.7720	3.6160	1.9000	3
26	3.9837	5.1840	3.2660	3
27	4.3835	5.1000	3.6670	2
28	3.2362	5.2670	0.7830	6
31	1.5835	2.2170	0.9500	2
35	3.3500	3.3500	3.3500	1

5.2 英日クローズド

図 5.2に形態素数と翻訳時間との関係をグラフで示す。

- 評価対象文一文あたりの平均形態素数: 7.7

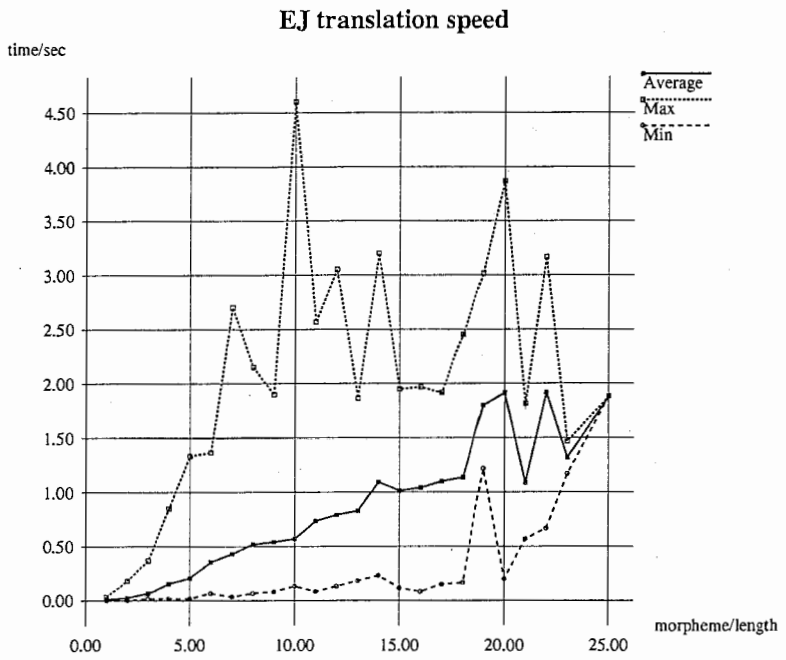

図 5.2: 形態素数と翻訳時間との関係（英日）

形態素数	平均 (sec.)	最長 (sec.)	最短 (sec.)	文数
1	0.0096	0.0340	0.0000	66
2	0.0259	0.1830	0.0000	77
3	0.0649	0.3670	0.0160	68
4	0.1538	0.8500	0.0170	108
5	0.2088	1.3330	0.0170	95
6	0.3543	1.3670	0.0670	113
7	0.4308	2.7000	0.0340	111
8	0.5175	2.1500	0.0670	84
9	0.5401	1.8990	0.0830	101
10	0.5689	4.6000	0.1340	83
11	0.7345	2.5670	0.0840	43
12	0.7911	3.0500	0.1330	60
13	0.8292	1.8670	0.1830	32
14	1.0925	3.2000	0.2330	38
15	1.0125	1.9500	0.1170	20
16	1.0397	1.9670	0.0840	26
17	1.0976	1.9170	0.1500	14
18	1.1351	2.4500	0.1660	9
19	1.8000	3.0160	1.2160	7
20	1.9129	3.8670	0.2000	9
21	1.0868	1.8160	0.5670	5
22	1.9160	3.1660	0.6670	7
23	1.3170	1.4670	1.1670	2
25	1.8830	1.8830	1.8830	1

グラフに見られるように、概ね形態素数に比例して翻訳処理時間が増える傾向にあるが、平均しても、日英翻訳で 0.7 秒程度、英日翻訳で 0.5 秒程度で翻訳処理が終了する。尚、形態素数が大きい部分での翻訳時間のばらつきは、評価データのサンプル数が比較的少ないことによる。形態素解析に要する時間が平均 0.2 秒程度 [山本 96a] であることより、概ね一秒以内で翻訳処理が終了すると考えられる。平均形態素数での翻訳処理時間を見ても、それぞれ、0.6 秒、0.4 秒程であり、話し言葉に対する翻訳処理時間としては良好な値であると言える。今後、インクリメンタル翻訳処理等の実現により、より長い文に対しても高速な翻訳が可能になると思われる。

第 6 章

まとめ

TDMT 研究における中間時評価として、日英間の翻訳処理に対し、翻訳結果の理解度を基準とした評価を行った。結果として、実際に近い会話において 60%、文単位での翻訳では 50% 程度に対し理解可能な文への翻訳処理が行えることが確認できた。学習量と翻訳成功率との関連により、今後学習量が増えるにしたがって、翻訳精度もより向上する可能性があると言える。今後も、適宜、翻訳評価を行うことにより、学習量と翻訳精度の関連を調査する予定である。また、翻訳文に対し評価者の指摘した問題点、原言語構造評価と翻訳評価の関連より、文生成系の改良が今後の重要な課題であると考えられる。さらに、評価結果（翻訳失敗）に対する考察と個別問題点への対応による翻訳処理の改善も進めていく必要があると思われる。

参考文献

[Paul96] Michael Paul, 伝康晴, 古瀬蔵. 日独変換主導翻訳の中間時評価. *ATR Technical Report*, No. TR-IT-0191, 1996.

[安藤 96] 安藤真一, 隅田英一郎. JEIDA 機械翻訳システムの評価基準を用いた TDMT の評価. *ATR Technical Report*, No. TR-IT-0188, 1996.

[浦谷 94] 浦谷則好, 竹沢寿幸, 松尾秀彦, 森田千帆. 音声言語データベースの構成. *ATR Technical Report*, No. TR-IT-0056, 1994.

[Furuse94] Osamu Furuse and Hitoshi Iida. Constituent Boundary Parsing for Example-Based Machine Translation. In *Proceedings of Coling '94*, Vol. 1, pp. 105–111, 1994.

[Furuse96] Osamu Furuse and Hitoshi Iida. Incremental Translation Utilizing Constituent Boundary Patterns. In *Proceedings of Coling '96*, Vol. 1, pp. 412–417, 1996.

[古瀬 94] 古瀬蔵, 隅田英一郎, 飯田仁. 経験的知識を活用する変換主導型機械翻訳. 情処学論, Vol. 35, No. 3, pp. 414–425, 1994.

[古瀬 95a] 古瀬蔵, 赤峯享, 河井淳, 金徳奉, 飯田仁. 経験的な言語知識を利用する対話翻訳機構 - 日英・日韓の対話翻訳システム -. 言語処理学会第 1 回年次大会, pp. 277–280, 1995.

[古瀬 95b] 古瀬蔵, 増井淳子. 言語データベースの概要. *ATR Technical Report*, No. TR-IT-0136, 1995.

[山本 96a] 山本和英, 隅田英一郎. TDMT 用形態素解析プログラム（日本語 / 英語 / 韓国語) - 性能評価報告 -. *ATR Technical Report*, No. TR-IT-0192, 1996.

[山本 96b] 山本和英, 古瀬蔵. 日韓間変換主導翻訳の中間時評価. *ATR Technical Report*, No. TR-IT-0190, 1996.

付録 A

TDMT 評価データの所在

それぞれの評価結果は、
/usr/local/TDMT/eval/
以下に格納されている。

各ディレクトリ、ファイルの説明

eval/

JOB-MANAGE　評価作業の管理に関することが記述されている。(評価データ管理者のメモ)

kiso-data/　評価データをプログラムで集計するたのデータ
(詳細については、kiso-data/README を参照のこと)

JE /　日英に関するコングレで評価してもらった生のデータ及び不具合を修正したデータ

translation /
open-bilingual-960614-template
(翻訳結果, バイリンガル会話,1996 年 6 月 14 日時点の TDMT)

translation / open-bilingual-960614-MH
(翻訳結果, バイリンガル会話,1996 年 6 月 14 日時点の TDMT,
評価者 (MH) による評価結果ファイル)

:

960801 - template (1996 年 8 月 1 日時点の...)

open-glossary-??????-template (基本表現集, ...)

??????- template (....)

```
                    closed-bilingual -   (クローズドデータ)
                                         ?????? - template  (...)
                                         ?????? - template  (...)
      structure   /   (原言語構造)
      speed       /   (翻訳速度)
JK/               日韓   :
JG/               日独   :
EJ/               英日   :
KJ/               韓日   :
```

付録 B

評価実験データシートの例

B.1 バイリンガル会話

"評価実験データシート(target)"

"1996 6/24 14:25:41"

```
==========================================================================
入力ファイル: AT120011

モード: *translate-mode* = :E-J
        *disable-sem-code* = NIL
        *all-targets* = NIL
        *beamwidth* = 1
==========================================================================

==========================================================================
(1)
"こんにちは、ペニンシュラニューヨークです。" (2.0e-5)
("Good afternoon this is the Peninsula New York")
--------------------------------------------------------------------------

    Rank: (a :No problem,  b :Fair,  c :Acceptable,  d :Nonsense,  NIL :No outut)
           >>
    --------------------------------------------------------------------
```

```
     Problem: (1 :Selection, 2 :Conjugation, 3 :Order, 4 :Style, 5 :Info, 6 :Deerminer)
              >>
              ------------------------------------------------------
            Others
              >>

=========================================================================
(2)
"どのような御用件でしょうか。" (0)
("May I help you")
-------------------------------------------------------------------------

     Rank: (a :No problem,  b :Fair,  c :Acceptable,  d :Nonsense,  NIL :No outut)
              >>
     ------------------------------------------------------------
     Problem: (1 :Selection, 2 :Conjugation, 3 :Order, 4 :Style, 5 :Info, 6 :Deerminer)
              >>
              ------------------------------------------------------
            Others
              >>

=========================================================================
(3)
"そちらでやっている週末の宿泊割引を利用したいんですが"
=========================================================================
(4)
"二人でディナー込みで一泊三百ドルでお願いしたいんです けれども"
=========================================================================
```

B.2 基本表現集

"評価実験データシート(kihon-target)"

"1996 7/30 12:29:12"

```
==========================================================================
入力ファイル: "/data/as32/nisimura/tdmt-hyouka/data-sheet-96.7.30/je-kihon/MOTO-DATA/kiho
```

```
モード: *translate-mode* = :J-E
        *disable-sem-code* = NIL
        *all-targets* = NIL
        *beamwidth* = 1
==========================================================================
```

```
==========================================================================
会話名: "/DB/LDB/JEK/JEKTEXT/S9503/A4JP.JEKTEXT-1"
会話場面: "ホテルフロントにて、チェックイン時"
会話形態: "対面"

==========================================================================
(1)(フロント)
"Yes, we have a reservation for you.|Please fill out this registration form."
==========================================================================
(2)(フロント)
"You couldn't write the company's name especially" (3.3666868)
("会社名は別に書かなくてもよろしいですよ")
--------------------------------------------------------------------------

    Rank: (a :No problem,  b :Fair,  c :Acceptable,  d :Nonsense,  NIL :No outut)
            >>
```

```
------------------------------------------------------------
Problem: (1 :Selection, 2 :Conjugation, 3 :Order, 4 :Style, 5 :Info, 6 :Deermi
        >>
        ------------------------------------------------------
      Others
        >>


==========================================================================
会話名: "/DB/LDB/JEK/JEKTEXT/S9503/A4JP.JEKTEXT-2"
会話場面: "ホテルフロントにて、チェックイン時"
会話形態: "対面"

==========================================================================
(1)(フロント)
"We do not have a reservation under that name."
==========================================================================
(2)(フロント)
"Do you have a confirmation slip?"
==========================================================================
(3)(客)
"No I'm afraid I left it in Japan"
==========================================================================
(4)(フロント)
"Do you know the travel agency ?" (0.375005)
("旅行会社はわかりますか")
--------------------------------------------------------------------------
```

B.3 原言語構造

"評価実験データシート(source-structure)"

"1996 7/30 13:9:25"

```
==========================================================================
入力ファイル: "AT120072"

モード: *translate-mode* = :J-E
        *disable-sem-code* = NIL
        *all-targets* = NIL
        *beamwidth* = 1
==========================================================================

==========================================================================
(3)
"そちらにフィットネス宿泊パックというのが有ると聞いた んですけれども"
"
TOP [(?X けれども) --- SE]
 |--?X [(?X たのです) --- SM]
     |--?X [(?X と ?Y) --- SP]
         |--?X [(?X に ?Y) --- NP]
         |   |--?X [(そちら)]
         |   |
         |   |--?Y [(?X が ?Y) --- NP]
         |       |--?X [(?X というの) --- S+N]
         |       |   |--?X [(?X <SN-SN> ?Y) --- N+N]
         |       |       |--?X [(?X <CN-SN> ?Y) --- N+N]
         |       |       |   |--?X [(フィットネス)]
         |       |       |   |
         |       |       |   |--?Y [(宿泊)]
```

```
          |        |        |
          |        |        |--?Y [(パック)]
          |        |
          |        |--?Y [(有る)]
          |
          |--?Y [(聞い)]"
```

評価: (1:good 2:bad NIL:NIL)
>>

==========

(4)
"一泊は確か二百ドル程度でしたよね"
"

```
TOP [(?X よね) --- SM]
 |--?X [(?X でし た) --- SM]
     |--?X [(?X は ?Y) --- NP]
         |--?X [(1 泊)]
         |
         |--?Y [(?X <ADJNOUN-CN> ?Y) --- S+N]
             |--?X [(確か)]
             |
             |--?Y [(?X 程度) --- ND]
                 |--?X [(200 ドル)]"
```

評価: (1:good 2:bad NIL:NIL)
>>