話し言葉の日英翻訳システムの評価法

6B−4

浦谷則好[1]　菊井玄一郎[1]　田代敏久[1]　田窪行則[2]　定延利之[3]　成田一[4]
[1]ATR自動翻訳電話研究所　[2]九州大学文学部　[3]神戸大学教養部　[4]大阪大学言語文化部

1　はじめに

機械翻訳システムの評価方法は今までも各所でいろいろ考案されてきたが、定評のあるものがいまだに存在しない。機能的な面を評価する場合、できれば適当な試験文の翻訳結果を調べることで行えることが望ましい。この観点に立って日英機械翻訳システムの評価法を検討したので、これについて報告する。また、我々が開発した音声言語翻訳実験システムASURA[1]に対する評価結果についても述べる。

2　翻訳評価法の考え方と機能評価項目

我々は評価の観点として、以下の点を考慮したいと考えている。

I．入力日本語の難易度（解析能力の評価）
　表現の難易度を何らかの基準を用いて分け、どこまで解析できるかを評価する。
II．出力される英文の評価（変換／生成の評価）
　・翻訳の正確さ　（時制、相、照応関係等）
　・翻訳の的確さ　（構文、訳語の選択等）
　・英語の品質　（英語としての自然さ）
III．その他
　文脈や言語外の情報をどこまで利用できているかを評価する。

外国人に対する日本語教育を参考にし、日本語表現の洗い出しを行い「目的指向型」電話会話表現形（以下単に「表現形」と呼ぶ）と呼ぶものを設定した[2]。この中の各項目には我々のターゲットである「目的指向型」電話会話に即して頻度と基本的か否かの観点から5段階（A〜D，X）のランクを付与した。全部で467項目あり、A，B，C，D，Xにそれぞれ214，70，46，67，70項目がランク付けされている。参考のため、日本語教育での水準も付加しておくことにした（A：初級262項目、B：中級205項目）。表1に表現形の一部を示す。

この表現形の各項目に沿って例文を作成し、翻訳システムに掛けて解析できたか否かを調べれば、解析能力を評価することができる。

IIの評価法については3で詳述するが、IIIについては情報の体系的な分類ができていないので、今回の報告には含めない。

表1　「目的指向型」電話会話表現形（一部）

項　目	水準	(教育)
I．ア．文末に関する基本的表現		
3．判叙表現		
(2)態の表現		
[1]〜ている（継続・進行）	A	(A)
[2]〜ている（結果の存続・経験）	B	(A)
[3]〜ている（形容詞的）	B	(A)
[19]〜しつつある	D	(B)
(5)断定の表現		
[1]事物の一致・不一致		
①〜が（は）〜だ	A	(A)
③〜というの（こと）は〜だ	A	(A)
[14]適当・許容・不許可		
①（〜する／した）ほうがいい	A	(A)
⑥（〜し）てもかまわない	A	(B)
4．要求表現		
(6)消極的行為要求		
〜（したら）どうでしょうか	A	(A)
〜（し）てほしい	A	(B)
〜（し）ていただきたい	B	(B)
〜（し）ませんか	X	(A)

3　生成英文の評価法

2で述べた解析能力の評価に続いて、例文を変換／生成フェーズに送りこんで、生成された英文を調べれば変換／生成の機能評価項目毎の評価ができることになる。しかし、単に「調べる」と言っても解析フェーズのように解析結果（木や素性）が期待されるものか否かという1か0かの判断を行うことはできない。「正確さ」「的確さ」「自然さ」にはそれぞれ程度が考えられる。また、英文の自然さの評価はネイティブでないとほとんど不可能である。一方、日本文の意味を正確に理解できなくてはシステムが生成する英文（以下「生成英文」と呼ぶ）を評価することはできない。ネイティブで日本語を十分に理解できる評価者を確保するのは容易ではない。

An Evaluation Method of Japanese-English Translation System for Spoken Language
Noriyoshi URATANI[1], Gen-ichiro KIKUI[1], Toshihisa TASHIRO[1], Yukinori TAKUBO[2], Toshiyuki SADANOBU[3], Hajime NARITA[4]　[1]ATR Interpreting Research Lab.　[2]Kyushu Univ.　[3]Kobe Univ.　[4]Osaka Univ.

さらに、予備評価実験の結果、普通のネイティブには文法性の評価はできないことが判明した。そこで、我々は以下のような方法を採ることにした。

[1] 日本人の通訳者が生成英文を文法性、入力日本語との整合性の観点で5段階で評価する。同時に、文法性や整合性の評点に影響を与えた項目をチェックシート（表2と表3）にマークする。チェック項目の指摘では不足な場合はコメントを付加する。さらに、入力日本語に対して翻訳の難易度を3段階（難しい、普通、易しい）で評価し、適切な翻訳英文（以下「修正英文」と呼ぶ）を作成する。

[2] ネイティブ（米国人）が修正英文と生成英文を比較して、整合性と自然さを5段階で評価する。整合性は通訳者と同様、チェックシートの該当項目にもマークを付ける。整合性と自然さを加味して総合点も付ける。さらに、生成英文を最小限いじった（多少非文法的なものも許して）意味の通る文と自然な英文を作成する。

[1],[2]とも高度な知識と判断を必要とするため、評価者は大学卒以上の教養を有するものに限定する。

表2　文法性のチェック項目

- □ 時制が正しくない
- □ 語順が正しくない
- □ 法助動詞の選択が不適切
- □ 冠詞（限定詞）が適切でない
- □ 前置詞の選択が不適切
- □ 代名詞の選択が不適切
- □ 受動態・能動態・使役の選択や変換が不適切
- □ 共起制限が守られていない
- □ その他（　　）

表3　整合性のチェック項目

- □ スタイルの選択が妥当でない
  - □ too formal
  - □ too informal
  - □ formalな表現とinformalな表現が混在している
  - □ 主語の選択が不適切
  - □ 法（仮定法、命令文、感嘆文）の選択が不適切
  - □ 語順が不適切
- □ 語彙の選択が妥当でない
- □ 日本文の省略要素が正しく特定できていない

4　ＡＳＵＲＡの評価結果

2，3で述べた評価法にしたがって我々のシステムＡＳＵＲＡの翻訳評価を実施した。表現形の項目に沿って約６００の例文を作成し、評価に掛けた。解析ではランクＡの２１４のうち１９６項目、ランクＢの７０のうち５５項目が正しく解析できることを確認した。つまり、「目的指向型」電話会話における日本語の基本表現の約９０％をカバーしていることが明らかになった。

3による生成英文の評価の中間結果の一部を述べると４０５文の整合性の分布は表4のようになっている。これを見ると当然のことながら通訳者とネイティブの評価には強い相関があることがわかる。同時にネイティブの方が少し評価が高めであることも見てとれる。問題となるのは一方が低いのに他方が高い場合である。通訳者＞ネイティブの場合は通訳者の制約知識の不足かと推察される。逆の場合は通訳者の英語直感が問題である。１例をあげると、

（通訳者２：ネイティブ５）

入力日本語：会議への参加を申し込みたいのですが。
生成英文：I'd like to apply for attendance to the conferenc
修正英文：I'd like to register for the conference.
意味の通る文：I'd like to apply to attend the conference.
自然な英文：I'd like to register for the conference.

つまり、通訳者は"apply for attendance"は英語として許容できないほどおかしいと判断しているが、ネイティブは十分意味が通じると評価しているわけである。現在、評価作業は継続中であり、同時に問題点の洗い出しを行っているところである。

表4　整合性の分布

通訳者＼ネイティブ	5	4	3	2	1	合計
5	64	34	11	2	0	111
4	54	52	29	4	1	140
3	23	27	26	6	2	84
2	6	11	20	10	1	48
1	2	3	7	7	3	22
合　計	149	127	93	29	7	405

5　おわりに

日英翻訳システムの評価のために「目的指向型」電話会話表現形を作成した。また、生成英文の評価法を確立した。我々の翻訳システムＡＳＵＲＡをこの方法で評価したところ日本語の基本的な表現の約９０％を解析できることを確認した。評価作業の最終結果と評価法の問題点については別の機会に報告したいと考えている。

参考文献

[1]竹沢寿幸ほか：“ＡＴＲ音声言語実験システムＡＳＵＲＡ”，本大会６Ｂ－５ (1993)

[2]浦谷則好ほか：“目的指向型会話文解析システムの機能評価法”，電通技報ＮＬＣ９２－１０ (1992)