旅行会話向け日中機械翻訳システムの開発

Development of Japanese-Chinese Machine Translation System for Travel Conversation

長田 誠也　　徐 金安　　山端 潔
日本電気株式会社 共通基盤ソフトウェア研究所

１．まえがき

我々は旅行会話向け自動通訳システムとして、日英・英日機械翻訳システムを開発してきた[1][2][3]。しかし英語だけでなく他の言語への翻訳の要望も高いために、その中で特に要望の高い中国語への機械翻訳システムの開発を行なった。この日中機械翻訳システムの構成と評価について報告する。

２．システム構成

今回開発した日中機械翻訳システムは、今までに開発してきた日英機械翻訳システムをベースとしたものとなっている。本システムは語彙化ツリーオートマトン文法[4]に基づいており、文法規則の適用を単語ごとにきめ細かに制御できることを特徴としている。

次にシステムの構成を説明する。本システムは図１に示すように、大きくわけて形態素解析部、構文解析部、構文生成部、形態素生成部からなっている。

日本語解析部の形態素解析部は日英機械翻訳システムで使用していたものをそのまま使用し、その後の構文解析部は、日英機械翻訳システムで使用している辞書やルールなどを参照しながら中国語用に新規に作成した。

中国語生成にあたる構文生成部は中国語向けに新規に作成し、形態素生成部では中国語は活用変形が不要な言語であるために構文生成部で使用した形態素をそのまま使用して文を生成している。

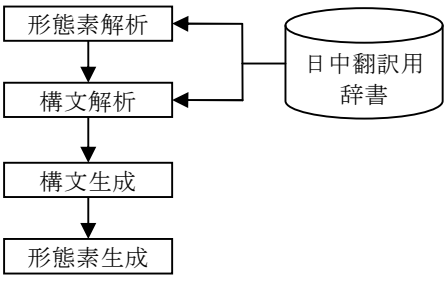

図１：日中翻訳システム構成図

３．本機械翻訳システムの動き

「私が夕食を食べる」という入力文を例に、本機械翻訳システムの動きを簡単に説明する。

3.1 解析部

形態素解析部で上記入力文を形態素解析して「私(名詞) / が(格助詞) / 夕食(名詞) / を(格助詞) / 食べる(動詞)」のように品詞つきの形態素単位に分割する。

構文解析部では最初に、分割された形態素に対して日中翻訳用の辞書から文法規則をロードする。

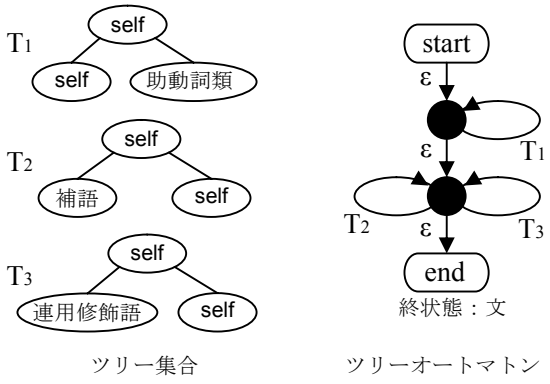

図２：「食べる」のツリーオートマトン

図２は「食べる(動詞)」がロードしたツリーオートマトンを示している。「食べる」が３つのツリー（T_1, T_2, T_3）を持ち、「食べる」自身(self)がヘッドワードとなり、それぞれ「助動詞類」、「補語」、「連用修飾語」を取り込んで成長することを示している。そして、このツリーによる遷移と、非決定性オートマトンにおけるイプシロン遷移の組み合わせで「食べる」のオートマトンが構成され、終状態として「文」になることを示している。

同様に、「私」と「夕食」は「格助詞」を取り込んで「補語」に成長する文法を持ち、それぞれ「私が」と「夕食を」という「補語」になる。

「食べる」は図２における T_2 のツリーにより、この「夕食を」と「私が」という「補語」を取り込み、「文」を生成する。図３にこの入力文で生成されたツリーを示す。

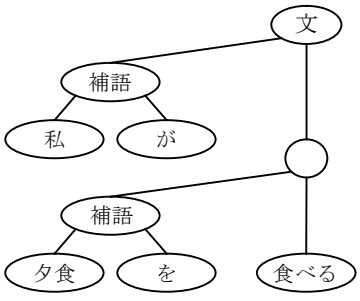

図３：オートマトン文法によって生成されたツリー

3.2 生成部

中国語の基本的な語順は英語と同様な SVO 型となっているが、把構文と呼ばれるような「把」を伴って述語の前に目的語を表示する構文などがあり、中国語の語順決定が課題になる[5]。

我々は日英翻訳システムで開発した日英翻訳用の格フレーム情報をベースとして日中翻訳用の格フレームを新規に作成した。日中翻訳システムでの中国語用の文の構成要素を表１に示す。

表１：中国語の文の構成要素

主語	動作の主体や主題を表わし、述語の前に置かれる
述語	主語について述べる部分で、主語の後ろに置かれる
目的語	述語が表わす動作の及ぶ対象などを表わし、述語の後ろに置かれる 複数の目的語（間接目的語、直接目的語）を持つ語や、介詞を伴うもの（在+場所や到+時間）もある
状語	述語を修飾する要素で、主語の後ろ、述語の前に置かれる 介詞を伴うもの（把+名詞句など）もある
補語	述語の様態や結果、方向などを表わし、述語の後に置かれる
離合詞要素	複数の文字からなる動詞の中で、その文字の間に目的語や補語が入るものがあるので、この動詞の後ろ側を離合詞要素として記述して、目的語や補語との位置関係を示す（ここには日本語の格情報は入らない）
補句	句を文の要素に取るものがあり、その内容を記述する（日本語の「～と思う」の「～と」のような引用部）

本システムでは前述したとおり文法規則を単語ごとに制御できることを特徴としており、生成部でもこの特徴を用いて、単語ごとに記述された日中翻訳用の格フレーム情報に基づいて中国語の語順を決定している。

「食べる(動詞)」の辞書は、日中翻訳用の格フレームの情報と、訳語の情報を持つ。

```
<格フレーム>
   <主語>ガ格</主語>
   <述語>self</述語>
   <目的語>ヲ格</目的語＞
</格フレーム>
<訳語>吃</訳語>
```

図４：「食べる」が持つ辞書情報

構文解析時にこの日中翻訳用の格フレームの情報を参照しながら補語の解析を行い、構文生成部でこの中国語の語順で構文生成をする。この例の場合は「ガ格」を持つ補語である「私が」が「主語」に入り、自分自身の「食べる」が「述語」に入り、「ヲ格」を持つ補語である「夕食を」が「目的語」に入る。そしてこの語順で中国語の構文生成を行なう。

次に形態素生成部では、構文生成部で生成された語順で訳語を出力（中国語では英語や日本語のような活用変形がないために、訳語をそのまま出力）して、「我吃晚饭」と翻訳する。

４．文字コード

日中翻訳エンジンでは中国語を扱うために、日英翻訳エンジンで使用していた JIS コードでは対応できない。このため、エンジンを JIS コードから Unicode で動作するようにした。具体的には、辞書やプログラムの文字コードの変換、文字を扱う関数の修正、エンジンで使用している辞書のデータ構造の修正などを行なった。

５．日英翻訳と日中翻訳の違い

日本語はまわりの状況からわかることは極力言葉として表現しない言語であり、英語は SVO の情報は言葉として表現するといった言語となっている。これに対して中国語は日本語と英語の中間的な感じで、例えば英語では言葉として表現する主語に関して、中国語では表現しないことがある。また、日本語で言葉として表現されていないとしても中国語では表現しなくてはいけないことがある。

また日本語と英語では同じ概念を持つため直接翻訳できるものが、中国語ではできないことがある。例えば日本語英語では「はい」/「いいえ」と”Yes”/”No”はある程度同じ概念として対応しているけれども、中国語では同程度に対応するものがなく、似たものとして「对」「是」/「不」「不是」などになる。また、日本語での可能表現「れる/られる/～ことができる」が英語で”can/be able to”とある程度は対応するのに対し、中国語では助動詞「能/可以」などだけでは対応せず、可能補語を使った表現にしないといけないこともある[6]。

ここでは、特に旅行会話で頻出する現象において日英翻訳と日中翻訳で異なる現象について報告する。

5.1 量詞

旅行会話に特徴的なものの１つとして、数量表現を使った文や指示代名詞を使った文が頻出することがあげられる。

日本語における「本」や「冊」のように特定の名詞に対応する「助数詞」が、中国語にも「量詞」として存在する。以下は各中国語における量詞の例であり、括弧内の単語は参考のための日本語である。

- 牛(牛)　→ 头(頭)
- 衣服(服)　→ 件
- 鱼(魚)　→ 条
- 书(本)　→ 本

また日本語の「個」と同じように、ある程度汎用的に使用できる「个」がある。

量詞に関して、日本語や英語と異なる点は、「この本」のような「連体詞＋名詞」でも量詞を入れる必要があることである。以下の例の下線部が量詞である。

- この本　→ 这本书

このように日中翻訳では指示代名詞を使ったときにも量詞を使用するため、量詞の使用頻度が高くなる。また、

日本語の助数詞と中国語の量詞が１対１に対応しているわけではないので、この量詞を名詞１つずつに付与して対応した。

なお日本語が指示連体詞ではなく指示代名詞のときも、基本的に量詞を補う必要がある。

- これを着る　→穿这件

ただしこの現象に対応するためには動詞などにも量詞を付与する必要があるが、現時点ではこのように動詞から対応する量詞を出さずに、「个」を使って「これ」→「这个」として翻訳している。

5.2 ～をお願いします

旅行会話での特徴的な表現として「～をお願いします」や「～してください」のような依頼表現が頻出することがあげられる。

日本語では「名詞をお願いします。」、英語では“*noun*, please.” というような「名詞(*noun*) に対して便宜を図ってください。」といった言い回しがあるが、中国語ではこのような言い回しがなく明示的に動詞を補う必要がある。我々の翻訳システムでは語彙に文法規則を簡単に記述できるようになっており、「お願いします」の辞書に名詞ごとに決められた動詞を出力するような文法を記述して対応している。以下の下線部は括弧内の日本語に対応する中国語である。

- 毛布をお願いします。
  →要毛毯。（～が欲しい。）
- 田中さんをお願いします。
  →请叫田中。（～を呼んでください。）
- コレクトコールをお願いします。
  →请接通对方付费电话。（～でつないでください。）
- ハイオクをお願いします。
  →请加高级汽油。（～を入れてください。）
- パスポートをお願いします。
  →请给我看一下护照。（～を私に見せてください。）
- お名前をお願いします。
  →请告诉我名字。（～を私に教えてください。）

5.3 その他の動詞の補完

「お願いします」以外にも動詞を補わなくてはいけない例をあげる。これも動詞の辞書に名詞との組み合わせで動詞を補完して訳出している。

- 映画に行く。(I go to the movies.)
  →去看电影。（映画を見に行く。）

5.4 象鼻構文

日本語には「象は鼻が長い」といった構造の文があるが、英語には対応する構造の文がなく構造変換をして翻訳する必要があった。しかし中国語で同様な構造の文があり、その構造を保持したまま翻訳できる文が存在する。

- 象は鼻が長い
  →象鼻子很长。

６．評価

本システムを WindowsMobile が動作している市販のＰＤＡ端末上に実装（図５）して、評価を行なった。

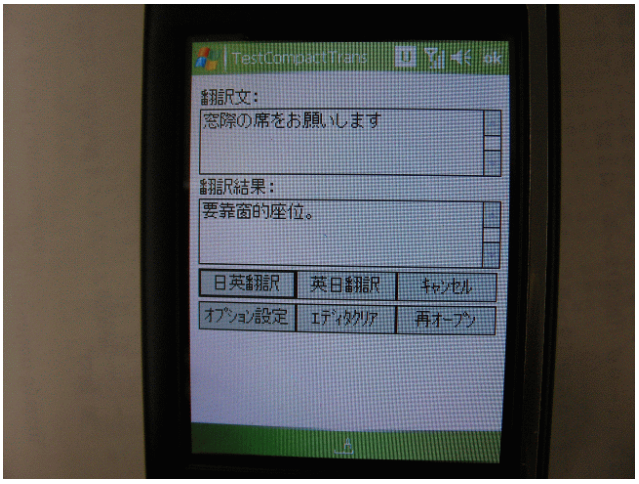

図５：ＰＤＡ端末上でのテストアプリ動作

翻訳システムの開発時は、１つの文に対する翻訳結果が頻繁に変化していく。しかし、中国語の訳質を確認できる人が限られており、翻訳結果が良くなったのか悪くなったのかを確認できないと開発がスムーズに進まない問題があった。

訳質評価作業の効率化とその評価結果の情報共有の効率化を図るため、開発用コーパス全文に対する毎日の翻訳結果を保存しておき、そのすべての翻訳結果に対して訳質の評価をした。

開発用コーパスとして、独自に収集した２つの旅行会話コーパス（コーパスＡ：約 2000 文で平均文字長 13.1 文字、コーパスＢ：約 12000 文で平均文字長 14.0 文字）を使用して開発をすすめ、訳質評価として 4 段階(a: natural, b: good, c: understandable, d: bad)の主観評価を行なった。また訳質評価で未評価のまま残ってしまった文を便宜的に'n'として、開発当初からの訳質評価の遷移を図６、図７に示す。左から右に向けた時間経過での訳質評価の遷移となっており、分布は下から a,b,c,d,n となっている。

また、この主観評価でそれぞれの値を持つ文を表２に例として載せる。

表２：評価の例

ほとんどの抗生物質にはアレルギーなんです	a
对几乎所有的抗生素过敏。	
ここで食べますか持ち帰りますか	b
在这儿吃吗？带回去吗？	
ここは私が払います	c
我付这儿。	
このボールはサラダ用ですか	d
这个球是色拉用吗？	

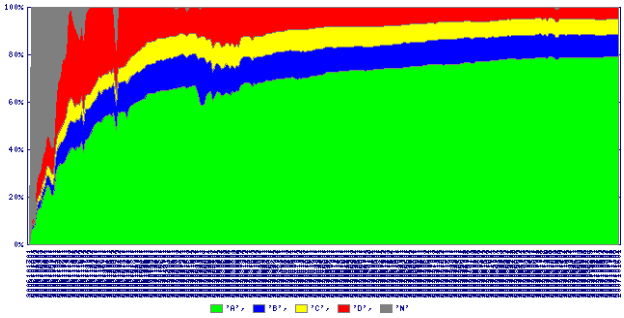

図６：コーパスA(約 2000 文)に対する訳質評価の遷移

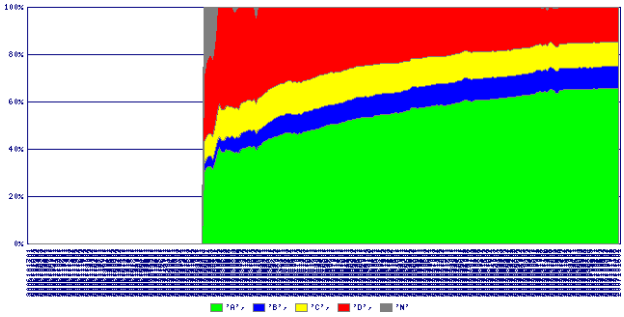

図７：コーパスB(約 12000 文)に対する訳質評価の遷移

２つのコーパスとも基本的に右肩上がり（時間が経つと訳質が上がっていく）の結果となっており、コーパスAでは初期の段階で understandable 以上の評価が 80%程度になったことがわかる。

コーパスBでの開発はコーパスAでの開発より後から開始したので、訳質評価のグラフが途中からとなっている。またコーパスBの開発開始とともにコーパスBに出現する単語を大量に追加したため、その時点でコーパスAの評価が悪化したこともわかる。しかしコーパスBを使った開発の開始直後で understandable 以上の評価が 60%程度あることから、それまでに開発した辞書や文法が汎用的なものとなっていたことが示されている。

最終時(グラフの一番右側)の評価結果の具体的な数字を表３に示す。

表３：最終時の評価結果

	a	b	c	d
コーパスA	79.1%	9.4%	6.5%	5.0%
コーパスB	67.3%	9.7%	9.8%	13.1%

７．おわりに

旅行会話向けの日中機械翻訳システムを開発し、ＰＤＡ端末上に実装して、その評価を行なった。開発コーパスの１つとなる約 12000 文のコーパスBに対して、understandable 以上で 87%, good 以上で 77%の評価結果となり、旅行会話向け日英機械翻訳システムとほぼ同等の結果[2]が得られた。ただし、今回の評価は完全にクローズな評価となっているので、今後はオープンなコーパスに対しての評価をしていくとともに更に翻訳精度を高めていく予定である。また、音声認識・合成と組み合わせた日中自動通訳システムとして統合していく予定である。

参考文献：

[1] Takao Watanabe, Akitoshi Okumura, Shinsuke Sakai, Kiyoshi Yamabana, Shinichi Doi and Ken Hanazawa, ”An Automatic Interpretation System for Travel Conversation”, Proc. *ICSLP-2000*, Vol. IV, pp.444-447, Oct. 2000.

[2] 山端潔 他, “PDA で動作する旅行会話向け日英双方向音声翻訳システム”, 情処研報, 2002-NL-150-9, 2002 年 7 月

[3] 長田誠也 他, “自動通訳システムのユーザインタフェイスレベルでの統合”, FIT2006, E-031, 2006 年 9 月

[4] 山端潔 他, “語彙化されたツリーオートマトンに基づく会話文翻訳システム”, 言語処理学会第 6 回年次大会論文集, pp.264-267, 2000 年 3 月

[5] 謝軍 他, “日中機械翻訳における中国語語順の決定法について”, FIT2002, E-45, 2002 年 9 月

[6] 徐金安 他, “日中機械翻訳における日本語可能表現の翻訳方法について”, 言語処理学会第 14 回年次大会論文集, PA3-5, 2008 年 3 月