自然言語処理 80－6
（1990．11．22）

格関係からの中国語生成

王啓祥* 王錫江* 黄英** 安原宏**
* 南京大学計算機科学系
** 沖電気工業(株)総合システム研究所

概要

日本語文は、述語と、その述語と何らかの関係を持つ格助詞をマーカーとする格によって文の意味が決まる。一方、中国語は語の形態変化がなく、語間関係と語の文中における役割は語順によって決まる。従って、中国語文を生成する時、語順の問題は極めて重要である。本稿で述べる中国語ジェネレータは、日本語の文を形態素／構文／意味解析して得られた格情報・意味コードなどを持ったノードを入力とし、日本語の深層格に中国語の文法成分を対応させ、日本語の動詞別に中国語の文法成分の位置を決める最適な最大表層構造チェーンをサーチし、正しい語順で中国の文を生成する。付録に本ジェネレータによって生成された中国語の例文を示す。

THE GENERATION OF CHINESE TEXT FROM THE CASE RELATIONS

QiXiang Wang, XiJiang Wang
Department of Computer Science, Nanjing University, CHINA.
Ying Huang, Hiroshi Yasuhara
Oki Electric Industry Co.,Ltd. Systems Laboratory
4-11-22 SHIBAURA, MINATO-KU, TOKYO 108, JAPAN.

Abstract

Deciding the order of words in a Chinese sentence is very important in the generation of Chinese text. In our Chinese generation(from Japanese) system , we generate the order of words in every generation process, such as the word generation process, the phrase generation process and the sentence generation process. In the sentence generation process, we make a correspondence between the Japanese deep case with the Chinese grammatical element, and define a logical chain which is composed of Chinese grammatical elements, called "The Longest Surface Syntactical Chain". In this paper, several translation examples will also be shown.

1. はじめに

近年、機械翻訳の研究が盛んになっており、既に日本でも多くのシステムが実用化されている[1]。しかし、その中のほとんどが日本語と英語の間の翻訳システムであり、中国語を扱ったものは少ない。

日本語を解析するには、格文法がよく使われる[2]。言語表現上、日本語は格を格助詞で表し、語順も比較的自由である。一方、中国語は日本語と違って格を語順で表現する。このように、言語によって格の表現の仕方が違う。翻訳の基礎となるものは深層格である。日本語を中国語に訳す時、正しい語順の訳文を得るために、深層格から表層格への変換のアプローチが必要となってくる。

例えば、深層格が、

・深層格（格の対応関係）：

	日本語	中国語
(命題情報)	述語 ：行く	去
	主題格：我々（は）	我們
	道具格：車 （で）	車 （乗）
	場所格：東京（に）	東京
(法情報)	テンス：Past（行った）	Past(去了)

のような日本語と中国語の文がある場合、その表層構造はそれぞれ、

・表層構造：

日本語：我々は車で東京に行った。

中国語：我們乘車去了東京。

のようになる。

本稿は、日本語を解析して得られた格情報を利用して、深層格から表層格へ変換し、正しい語順で中国語を生成する日中翻訳システムの一部である中国語ジェネレータについて述べる。

2. 中国語ジェネレータの設計思想

中国語は日本語と違って、形態変化が少ない。語順は文の意味を決める重要な要素である。語順以外にも、訳語の選択や、テンス及びアスペクトなどの問題にも注意を払う必要がある。以下は本ジェネレータが考慮に入れた幾つかの問題である。

2.1 訳語の選択

対訳辞書に語の構文情報（品詞）と意味情報を入れる。例えば、動詞の場合、動詞の格フレームを作り、中に動詞が支配する格の意味コードを入れる。名詞の場合、名詞が取る助数詞を入れる。このように、構文と意味情報を利用して多義語に対する適当な訳語を選択している。例えば、「引く」という日本語の多義語に対して、その訳語選択は以下の様になる。

見出し	格情報	語義コード	訳語例	訳文例
引く1	動作主格	OH（人間）		鉛筆で線を引く
	対象格	AF（形態）	划	用铅笔划线
	道具格	CE（手段）		
引く2	動作主格	OH		ロープを引く
	対象格	OM	拉	拉绳索
引く3	対象格	OM		7から4を引くと3になる
	ソース格	MN	减去	从7中减去4等于3

図1 訳語選択

2.2 格助詞の訳し方（訳文成分の添削）

中国語には日本語のような格助詞がないため、格助詞が表す意味を適切に表現するには何らかの単語を名詞や動詞の前や後ろに付け加えたり、訳さなくていいものを削ったりする必要がある。格助詞「の」と「に」を例にすれば、

格助詞	日本語	中国語
「の」	私の先生の本	我老師的書
「に」	実験室に学生がいる	实验室里有学生
	南京に住んでいる	住在南京
	技術に詳しい	熟悉技術(訳さない)
	王さんに本を貸す	借書給小王

構文意味情報を利用して辞書とルールを設計する時にどの場合にどの使い方をするかを制限したり、区別したりする必要がある。本ジェネレータでは格助詞だけの辞書を別に設けることにしている。

2.3 中国語の語順の決め方

中国語では、ある単語の文中での位置はその語の文における役割（どの文法成分であるか）とその語と他の語との関係（何を修飾するか）を決めている。例えば、「考試」という単語はそれが置かれる場所によって意味なども違ってくる。

他　明天　参加　考試。(彼は明日試験に参加する)
p(主語) adv(状語) v(謂語) n(直接目的語)

他　明天　考試。(彼は明日試験を受ける)
p(主語) adv(状語) v(謂語)

語順を正確に決めるために、本ジェネレータは以下に示すような**最大表層構造チェーン**を設計し、この最大表層構造チェーンを用いて語順を決めるようにしている。

最大表層構造チェーン：
従文＋前置状語＋主語＋伴随語＋状語＋謂語＋
後置状語＋間接目的語＋直接目的語

語順の決め方は、最大表層構造チェーンに基づいて日本語格パターン pp から生成した中国語に対し、2次元の番号を付ける。2次元の番号のうち、1つは文全体の各文法成分（即ち、主語、述語、目的語など）の番号であり、1つは従文や埋め込み文内の各文法成分の番号である。この2次元の番号を用いて一度に中国語の文法に合う文を生成することができる。

2.4 テンスとアスペクトの対応

日本語の解析から得たテンスとアスペクトの情報に応じて、対応した中国語の表現を用いる。しかし、中国語のテンスとアスペクトについての研究はまだ不十分なため、その対応関係はまだ一部分しかできていない。

3. 中国語ジェネレータの実現

3.1 ジェネレータの構造

本ジェネレータは3つの部分からなる：入力部、生成部及び出力部。生成部は文法部と実行部からなる。文法部は生成辞書と生成規則を含む。その構造図を図2で示す。

入力部	生成部	出力部
格情報を	【1】文法部	中国語文
含んだ	(1)生成辞書	
ノード	(2)生成規則	
	【2】実行部	

図2 ジェネレータの構造図

各部分の構造について 3.1.1 節以降で詳しく述べる。

3.1.1 入力部

入力は、日本語の文を形態素解析、構文解析、意味解析をして得られた格情報を含んだファイル形式のノードの順列。各ノードは次のような形をしている。

node(<level>, <cat>, <lex>, <type>, <orid>, <sem>)

ここで、<level> ::= ノードのレベル番号
<cat> ::= 品詞¦構文属性
<lex> ::= 単語¦句¦文
<type> ::= 活用形¦表層格¦表層態（ボイスやテンス、アスペクト）
<orid> ::= 単語原形¦深層格¦深層態（ボイスやテンス、アスペクト）
<sem> ::= 意味情報

例： node(2, "pp", "我々は", "は", "主題格", "NOH")

3.1.2 生成部

生成部は文法部と実行部から構成される。

3.1.2.1 文法部

(1) 生成辞書の内容

生成辞書は対訳辞書と格助詞辞書2つの辞書を含む。それぞれの辞書項目の構造を以下に示す。

(i) 対訳辞書

対訳辞書の辞書項目の構造は、

word(見出し, 品詞, 意味コード, 訳語見出し,
訳語品詞, 訳語意味コード)

例：
word(はな, n, plant, 花, n, plant)

(ii) 格助詞辞書

格助詞の辞書項目の構造は、

part-word(格助詞見出し, 格情報, 文節意味コード,
前置訳語, 後置訳語)

例：
part-word(に, 場所格, location, nil, 里)

格助詞辞書は文節 pp を処理する時に使われる。その役割は訳文を生成する時に添加すべき単語とその位置を決めることである。

上の例は、格助詞が「に」である時、格情報が「場所格」で、意味コードが「place」の場合、訳語の後ろに「里」を付けることを意味する。

(2) 生成規則の形式

生成規則の構造は、

rule(<Condition>, <Action>, <Backmessage>)

ここで、<Condition>: 条件部：動作をするための条件を記述する。
<Action>: 動作部：条件が満たされた時の動作を記述する。
<Backmessage>: 返答部：動作を実行した後に返すメッセージを記述する。

生成規則には 4 種類のものがあり、それぞれ名詞句生成規則、動詞句生成規則、文節生成規則及び文型生成規則である。詳しい構造を次に示す。

(i) 名詞句生成規則

名詞句生成規則の形式：

np_rule(<Condition>, <Action>, <Backmessage>)

ここで、
<Condition> ::= con(品詞情報, 意味情報)
<Action> ::= inout ¦ act(De) ¦ add(liang_word) ¦ change
inout : 翻訳不要な単語の処理
act(De) : 「的」の処理
add(liang_word) : 数助詞の処理
change : 名詞句内の語順の調整
<Backmessage> ::= 意味情報

(ii) 動詞句生成規則

動詞句を生成する時に起動される規則である。その形式は、

vp_rule(<Condition>, <Action>, <Backmessage>)

ここで、
<Condition> ::= con(品詞情報, 意味情報)
<Action> ::= change
change: 動詞句内の語順の調整
<Backmessage> ::= 文法成分（つまり述語(謂語)）

(iii) 文節生成規則

日本語の文節に適用される規則である。その形式は、

pp_rule(<Condition>, <Action>, <Backmessage>)

ここで、
<Condition> ::= con(品詞情報, 意味情報, 格助詞, 格情報)
<Action> ::= inout ¦ act1(x) ¦ act2(格助詞)
inout : 直接ソース言語の単語を訳語として出力する。
act1(x) : x の値を付け加えて訳語を出力する。
act2(格助詞) : 格助詞により、格助詞辞

書を調べ、適当な単語を添加し、訳語を出力する。

<Backmessage> ::= pp に対応した中国語の文法成分。

(iv) 文型生成規則

日本語の格情報、格助詞及び文末の用言 vp（動詞、形容詞、形容動詞）を利用して、文型を選択し、最大表層構造チェーンを得る規則である。その形は、

cvp_rule(<Condition>, <Action>, <Backmessage>)

ここで、

<Condition> ::= con(格情報, 格助詞, 用言の原形, 用言の意味情報)

格情報： vp が支配する各格情報

格助詞： 原文中の各格助詞

用言の原形： v（動詞）， a（形容詞）,av（形容動詞） の原形

用言の意味情報： v, a, av の意味情報

<Action> ::= search(i)

search(i): 日本語文型番号 i によって、その文型に対応した最大表層構造チェーンを得る。

<Backmessage> ::= サーチした中国語の文型の番号

3.1.2.2 実行部

実行部は３つの処理に分けられる。それぞれは、訳語生成処理、句生成処理と語順調整処理である。

(1)訳語生成処理：訳語を生成するモジュールからなる。

(2)句生成処理　：３つのモジュールからなる。

・名詞句(np)を生成するモジュール

・日本語の文節(pp)に対応する部分を生成するモジュール

・動詞句(vp)を生成するモジュール

(3)語順調整処理：語順を調整するモジュールからなる。

この３つの処理について 3.2 節で詳しく述べることにする。

3.2 文生成の各処理

[1] 訳語生成処理

訳語生成する時、日本語の単語について以下の分類をする。

(1)そのまま訳語として出力する単語

記号など。例えば、「CPU」。

(2)翻訳不要な単語

格助詞。訳語を nil とする。

(3)単義語

例えば、私(日) →我（中）

(4)多義語

多義語の品詞や語義コードで訳語を選ぶ。

(5)未登録語

ユーザに単語を登録してもらうか、文生成処理を終わらせる。

[2] 句生成処理

(1) np の生成

日本語の np の定義は、

np ::= n ¦ p ¦ np

np ::= np + 「の」 + np

である。名詞句生成規則に基づいて訳語を生成する時に、中国語らしい表現にするには「的」を名詞句に入れたり削ったりする必要がある。「的」の使い方はとても複雑であり、今のところまだ完全に明らかにされていないが、本ジェネレータは今わかっている幾つかのルールを使うことにしている。例えば、

「p(人間) + n(人間)」で構成される np :

「的」を入れない。

「p(人間) + n(物)」で構成される np :

「的」を入れる。

としている。

例：「私の妹」は「我妹妹」になり、「彼のノート」は「他的本子」になる。

(2) pp の生成

日本語の pp ノードの語義コードと格成分の情報を利用して、文節生成規則に基づいて対応する中国語の文法成分を生成する。そして、格助詞辞書を調べ、添加すべき単語を訳語に付け、それを pp の訳語として出力する。

(3) vp の生成

動詞句生成規則と以下のような中間情報チェーン L に基づいて中国語の vp を生成する。

vp の中間情報チェーン L:
v(a,av)の訳語 + テンス（アスペクト） + ボイス + vp に対応する中国語文法成分.

例えば、L が、
「翻訳」 + past + 受け身 + 述語（謂語）
のような形の場合、L に基づいて、
「被翻訳了」
という中国語の vp が生成される。

[3] 語順調整処理

語順の調整は、まず pp の並ぶ順序と、pp ノード中の格情報と格助詞とによって、文型生成規則に基づいて対応する中国語の文型を選び出し、その文型に合った最大表層構造チェーンをみつける。そして、そのチェーンによって各 pp 中の中国語の文法成分に番号を付け、語順を調整し、訳文を生成する。

以上をまとめると、本ジェネレータの処理のフローチャートは図３の様になる。

4. 終わりに

本ジェネレータは日本語の文を形態素、構文、意味解析して得られたノードを入力とし、訳語の生成、句の生成と語順の調整の３のプロセスを経て中国語を生成することを試みた。

本ジェネレータの特徴は以下のようなものである。

（１）対訳辞書の他に格助詞辞書を用いて、訳語に適当な単語を付け加え、中国語らしい表現ができるように工夫をした。
（２）日本語の格情報に対応する中国語の文法成分を文型によって決め、文型に合った最大表層構造チェーンによって文法成分の順序を調整し、文を生成した。
（３）文法部（辞書と生成規則）と実行部はそれぞれ独立しており、単語やルールの追加や書き換えなどに便利である。
（４）格文法を用いる、よりよい日本語解析ツールとの結合が実現できれば、より良い翻訳システムができる。

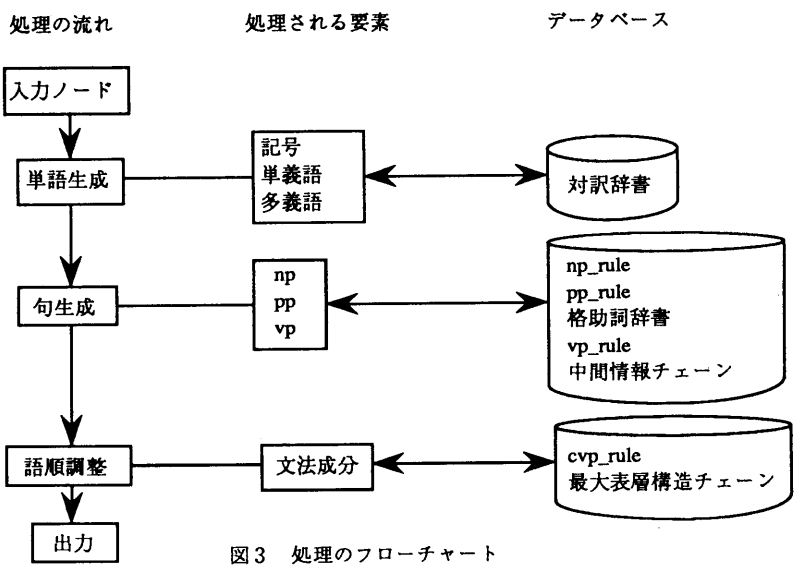

図３　処理のフローチャート

本ジェネレータはまだ実験段階にあり、日本語と中国語のより複雑な言語表現についての研究はまだ不十分である。今後は辞書やルールの情報の補足や、単語数、ルール数の増加などの作業をしていくと同時に、訳語の選択や、テンスとアスペクトを含んだ日本語と中国語の複雑な表現などについて更に研究、整備して行くつもりである。

参考文献

[1]野村 浩郷,田中 穂積: 機械翻訳, 共立出版,bit別冊 (1988.9)

[2]野村浩郷: 自然言語処理の基礎技術, 電子情報通信学会(1988)

[3]Fillmore.C: The case for case, In: E.Bach and R.Harms(Eds.), *Universals in Linguistic Theory.* (1968)

[4]田中 穂積,辻井 潤一: 自然言語理解, オーム社 (1988)

[5]辻井潤一: 文解析方式, 情報処理,Vol.27,No.8(1986)

[6]辻井潤一: 機械翻訳における文章の生成, 人工知能学会誌, Vol.4, No.6(1989)

付録　翻訳例

本ジェネレータを用いて生成した例文を以下に示す。

(1)我々は車で東京に行った。
　我们乘车去了东京.

(2)私は川で泳がない。
　我在河里不游泳.

(3)私はその試合に勝ちたい。
　我那场比赛想赢.

(4)彼は医者に行く。
　他去看医生.

(5)生徒は先生に答える。
　学生回答老师.

(6)日本経済は海外諸国の発展と密接に結び付いている。
　日本经济和世界各国的发展紧密联系.

(7)国際社会での日本の地位が向上し、日本語の文章を大量に翻訳する必要が生まれてきた。
　在国际社会中日本的地位提高，大量翻译日语文章的必要产生了.

(8)雨が降っている。
　下着雨.

(9)計算機が情報を記憶する容量は非常に大きい。
　计算机存储信息的容量非常大.

(10)この本とあの本と同じくらい面白い。
　这本书和那本书差不多有趣.

(11)象は鼻が長い。
　象的鼻子很长.

(12)私は万年筆がほしい。
　我想要钢笔.