

パターン変換型機械翻訳システムにおける 日中翻訳文の生成方式について

謝軍 西愛珍 河原功幸 兵藤安昭 池田尚志

岐阜大学大学院工学研究科

1 はじめに

日本語は膠着語で、「～は・～を」などの助詞が特徴である。中国語は孤立語で、言葉の配列順序によって意味を表すのが特徴である。このように構造の違う日中両言語を機械翻訳するために、我々はパターン変換型の翻訳システムを構築中である。本報告では、目標言語パターンからの訳文の生成方法を中心に述べる。

2 パターン変換型機械翻訳システム

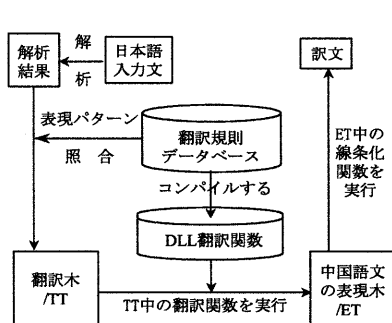


図1: 翻訳システムの流れ

図1に翻訳システムの概要を示す。

翻訳規則データベースは、係り受け関係を使って表現した日本語表現パターンと、それに対応する中国語の表現構造(翻訳規則)を記述しているデータベースである。日本語入力文の解析結果を翻訳規則データベースのパターンと照合し、対応する翻訳規則の木構造(TT: Translation Tree)を出力する。翻訳規則データベース中の翻訳規則はあらかじめコンパイルされてDLL中の関数になっており、TTの中にその関数の名前が取り込まれる。次に、TT中のこの関数を実行すると、中国語の意味構造/修飾構造であるET(Expression Tree)に変換される。最後にET中の線条化関数を実行すると翻訳文が得られる。

3 中国語訳文の生成方式

中国語の文の意味構造を表すのに、オブジェクト指向言語のクラスを使っている。つまり修飾される構造と線条化関数する規則が同一のものをクラスとしてまとめ、修飾する語句をクラスのメンバーとしている。クラスのオブジェクトを使って翻訳文の意味構造/修飾構造を表現したものがETである。各クラスは、その意味構造/修飾構造を線条化して1次元の言語表現を作り出すためのlinearize関数をメソッドとして持っている。ETに対してlinearize関数を実行すると翻訳文が得られるという仕組みである。

3.1 表現クラス

図2に、現段階で設定している名詞クラスのメンバーと派生クラスを示す。

CNoun	CNoun	*adnominal;	// ① AのB
	bool	adnoTEKI;	// ①'「AのB」に「的」を使うか
	CEmbed	*embedding;	// ② 連体修飾
	CString	designator;	// ③ 指示代名詞
	CxQNoun	*quantity;	// ④ 数量表現
-CnGeneral	CProposition	*predicate;	// ⑤ 被修飾名詞が述語パターンの引数でない連体修飾
-CnConstant			// 数字、記号などの直接変換

図2: 名詞クラスのメンバーと派生クラス

3.2 線条化関数

中国語の文構造は基本的に「SVO」の語順であるが、S,V,Oそれぞれの前後にあれこれの修飾要素が挿入できる。そこで動詞のlinearize関数では、各修飾語句は①主語前 ②主語 ③主語と述語の間 ④述語部 ⑤述語後 ⑥目的語の後のいずれの位置に置くかということを記述しておき、最終的にこれらと接続して1次元の文とする。

4 名詞句の翻訳

中国語では名詞を修飾する場合、一般に修飾語句(+的)+名詞という形をとるが、「的」を挿入するか否かについて様々の場合がある。表1に「的」を使うか使わないかの規則をまとめた。名詞の修飾語順序は、ふつう「的」を使う修飾語を「的」使わない修飾語の前に置く。ただし、指示代名詞③と、数量表現④は「的」を使わないが、文頭側に置く。つまり、③④⑤(②①)的(②①)的なしの順序がある。

表 1: 名詞修飾表現の翻訳規則

修飾語	「的」を使うか否か	パターン
名詞	1~4しか使わない	1. 人称代名詞+の+親族関係名詞
		2. 固有名詞+の+名詞
		3. 天気名詞・時間名詞+の+時間名詞
		4. 役職名+の+人名
形容詞	1と2では使わない	1. 中国語訳語が1文字の形容詞 2. 名詞+的(+な)+名詞
動詞・補助動詞	使う	動詞・補助動詞+名詞
数量詞	使わない	数量詞+の+名詞

・ 修飾語が名詞である場合

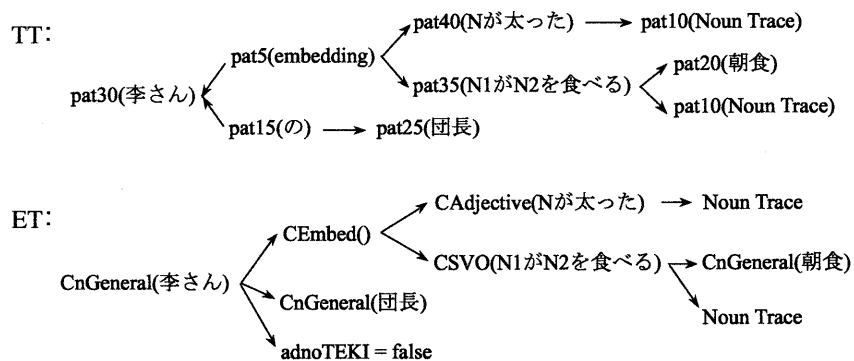
表1の「AのB」の規則は、次図の例のように翻訳規則データベース中の、係り受けの条件欄を使って取り込むことができる。

係りの条件	キーワード	係り先の条件	中国語翻訳規則
A-団長：役職名	の	B-李さん：人名	B-CnGeneral; A-CnGeneral, adnominal; #A > adnoTEKI=false; // 「的」を使わない

この翻訳規則は、AとBとも普通名詞クラス CnGeneral であること、AがBのメンバー adnominal であること、Bの線条化関数でメンバー adnominal を線条化する際、「的」を使わないことを表現している。

・ 修飾句が述語である場合

一般に連体修飾の TT は図3のように特別のノード embedding を用いて述語句を接続して表現する。被修飾名詞が述語パターンの引数である場合、述語句中のその引数は空位 (Noun Trace) となる。表1中の「中国語訳語が1文字の形容詞」という条件は、形容詞のクラスとその訳語を見て調べればよい。



中国語訳： 吃早饭的胖团长老李

図 3: 連体修飾の TT、ET と中国語訳の例 (朝食を食べる太った団長の李さん)

名詞の修飾語順序は CnGeneral の linearize 関数のプログラム中に表現される。図3の ET を線条化すると、中国語の訳が得られる。

5 おわりに

日中機械翻訳システムのパターン変換手法と名詞句の翻訳例について述べた。より多くの日中翻訳規則の実データでシステムを充実していきたい。