自然言語処理 80－8
（1990．11．22）

変換主導型機械翻訳の実現手法

古瀬 蔵　隅田 英一郎　飯田 仁
ＡＴＲ自動翻訳電話研究所

本稿では、翻訳全体を変換部で制御する変換主導型機械翻訳について述べる。従来の解析主導型機械翻訳と異なり、この方式は、変換機構と対訳変換知識を中心に翻訳を行ない、高速、高品質で拡張性の高い翻訳システムの実現を目指すものである。まず、入力文をいかに訳すか決めるために、翻訳に必要な変換知識が検索される。これらの変換知識により必要最小限な処理のみが駆動され、効率的に翻訳結果を得る。筆者らは、国際会議の参加登録に関するドメインについて日英翻訳の基礎実験を行ない、変換主導型翻訳の基本原理の実現性を確認した。

A method for realizing Transfer-Driven Machine Translation

Osamu Furuse,　Eiichiro Sumita,　and　Hitoshi Iida

ATR Interpreting Telephony Research Laboratories

Sanpeidani Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

This paper describes the Transfer-Driven Machine Translation method, which controls the entire translation by transfer. Unlike conventional Analysis-Driven Machine Translation, the transfer engine and transfer knowledge play central roles in this translation method, which aims at translation with high speed, high quality, and high expandability. Necessary transfer knowledge data is retrieved to determine how to translate an input sentence. This transfer knowledge data drives only necessary processes and the translation is efficiently performed. The authors carried out a simple experiment of Japanese to English translation, using the corpus of registering for an international conference. The feasibility of this basic principle has thus been confirmed.

1. はじめに

機械翻訳の研究は近年盛んに行なわれ、現在多くの機械翻訳システムが実用化されている。しかし、構文構造の複雑さ、表層形と意味の対応の複雑さ、原言語と目的言語の表現方法の隔たりなどが原因で、正しい翻訳結果が得られないことがある。従来のシステムは、高品質の翻訳の実現を解析処理に依存させることが多く、多くの言語学的知見を取り入れたり、深い意味処理を試みてきた。このような翻訳方式を以下、解析主導型翻訳と呼ぶ。解析主導型翻訳では解析処理部が複雑で負荷の大きなモジュールになりがちであったため、次のような問題点が生じた。

- **規則による言語現象の記述の難しさ**

すべての現象を解析規則によって説明するのは、個々の言語現象の性質の多様性、規則数の多さ、説明の複雑さ、例外の多さなどによる難しさがあった。

- **処理の複雑さによる処理時間の増大**

従来の機械翻訳方式では、すべての文に対し、ほぼ均一な枠組みで処理を行なう場合、ダイレクトに目的言語にマッピングすれば翻訳できるような簡単な文に対して、深い意味処理などを行ない、翻訳に時間がかかり、慣用表現など特別な場合分けに従った。

- **解析処理の複雑さによる拡張の困難さ**

言語現象の多様性、複雑さのため、システムの処理が複雑なものとなり、拡張しようとすると知識表現の仕様変更、既存モジュールへの副作用などのため、首尾一貫した処理の枠組みを失いがちであった。

ところで、従来の解析主導型ではない新しい翻訳方式がいくつか考えらている。例えば、変換と解析の融合[1],[2]、アナロジー[4],[5]、実例[6]、用例[7]、翻訳対[9]などに基づく翻訳方式は、変換プロセスに重点を置いたものである。

翻訳は原言語から目的言語を導出するプロセスであり、機械翻訳全体の流れにおいて変換処理が翻訳の意味を最も反映したプロセスと言える。本稿では翻訳処理を全体を変換によって制御する変換主導型翻訳を提案する。変換主導型翻訳は、以下のことを目的している。

- **変換知識を利用した高品質の翻訳**

質の良い翻訳結果を得るのに、単語の意味を構成的に組み立てるやりかたでは難しいことがある。そこで、翻訳の傾向を捉えた変換知識を最大限に利用して質の良い翻訳結果を出そうとする。

- **要求駆動による高速の翻訳**

人間が翻訳する場合、文によって反射的に訳文がでてくるものもあれば、文内容を熟考して訳文を作り出しているものもある。また、一見構造が複雑そうな文でも、訳す方略が決まっていてるため比較的容易に訳文を作り出せることもある。変換主導型翻訳では入力文に応じて処理方法を与え、できるだけ浅い処理によって翻訳を行ない、必要があればその都度深い処理を行なう。

- **拡張の容易さ**

各処理モジュールの処理が比較的独立していること、規則の代わりに変換知識が使われ拡張の際の副作用が起こらないのでシステムの拡張が容易である。

以下、2節で変換主導型翻訳の原理、3節で変換主導型翻訳のドメインとデータ、日英翻訳を例に4節で変換主導型翻訳のシステム構成、5節で基礎実験について述べる。

2. 変換主導型機械翻訳の原理

本稿で提案する変換主導型翻訳方式の主な基本構成を図2-1に示す。翻訳処理は変換機構を中心に行なわれるが、変換前には形態素解析、変換後には生成が行なわれる。以下、主な構成要素である変換知識、対訳決定部、変換知識検索部、変換機構部、解析処理部について説明する。

2.1. 変換知識

変換知識は翻訳処理を方向づけるデータとなる。変換主導型翻訳では、変換知識を階層化し、処理方法の深さのレベルを翻訳に必要な変換知識のレベルに対応づける。そして、できるだけ浅い階層の変換知識で翻訳しようと試みる。

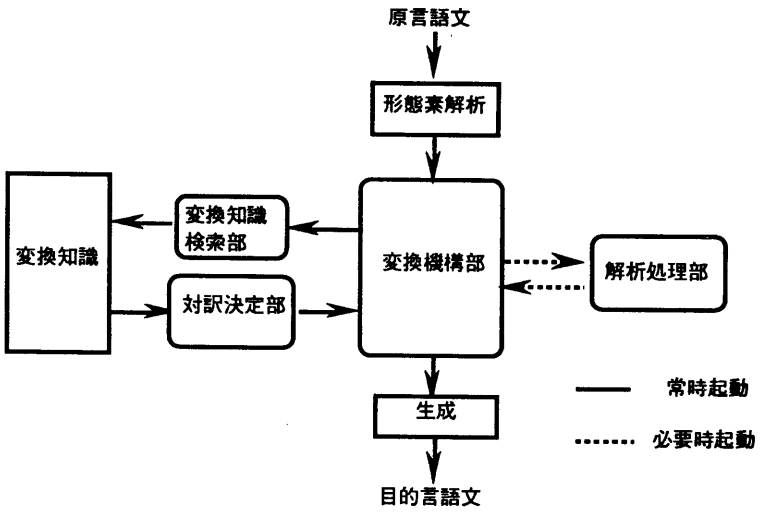

図2-1　変換主導型翻訳の基本構成

変換知識は原言語と目的言語の対応を記述する。原言語と目的言語の対応が一対一の時は無条件に変換知識に記述された目的言語表現に変換される。しかし、ある原言語表現に対応する目的言語表現は必ずしも一つとは限らない。そこで、変換知識は、次のようにその目的言語表現に変換されるための条件を記述する。目的言語表現が一つの場合は条件は記述されない。

原言語表現　＝＞　目的言語表現１（条件１），
　　：　　：
目的言語表現ｎ（条件ｎ）

2.2. 対訳決定部

対訳決定部は変換先の目的言語を決定する。変換知識において対訳が複数個ある場合、その表現の出現する環境と照合され、最も適合する条件を持つものが目的言語表現として選ばれる。文型であれば、文を構成する語句など文内文脈によって、文であれば、前文の内容、話者などの文外文脈に関する条件を記述する。

2.3. 変換知識検索部

変換知識検索部は、入力の原言語を翻訳するのにどのような変換知識を必要とするかを調べる。

2.4. 変換機構部

変換機構部は変換知識を参照しながら、効率的な翻訳を行なう。変換主導型翻訳には、処理のレベルに応じた多重の対訳変換知識があり、変換機構部がどの変換知識をどう使うかを決めて翻訳の処理が制御される。すなわち、入力文に対し、変換知識を利用してどのように翻訳すべきか決定したうえで翻訳を行なう。従来の解析主導型翻訳は図2-2に示すように、すべての入力文に対して単一の変換知識により同じ深さの処理を行なう。それに対して変換主導型翻訳では、図2-3に示すように翻訳に必要最小限の処理だけを、文からその構成要素へというように、全体から部分へと再帰的に抽出していく。そして、それぞれの処理のレベルに応じた変換知識を用いて処理を実行する。

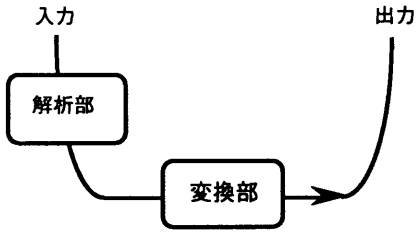

図2-2　従来の解析主導型翻訳での翻訳処理の流れ

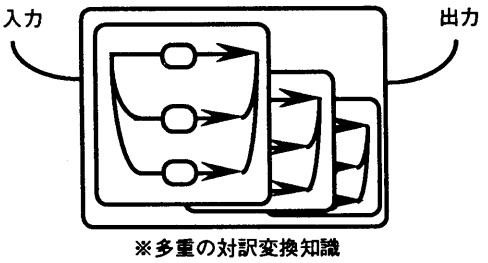

図2-3　変換主導型翻訳での翻訳処理の流れ

2.5. 解析処理部

変換知識を適用することができない部分について解析処理を行なう。この構成要素はできるだけ起動されないほうが望ましい。

3. ドメインと頻出表現

変換主導型翻訳では最大限、浅い変換知識を利用し、できるだけ解析処理をしないようにする。もし、ほとんどの文が浅い変換知識を利用できないならば、解析主導型翻訳とほぼ同じ動作をすることになるので、変換知識をいかに定めるかが問題となる。しかし、起こりうるすべての可能性の表現を変換知識として持つことは、無限の知識を持つことにつながり、実現は不可能である。従って、いかに効果的に変換知識を選定するかは重要な問題である。変換知識を選定する厳密な指標はない。しかし、少なくとも翻訳対象となるドメインをカバーするように高頻度の表現を中心に変換知識のデータベースを構築することが肝要である。

われわれは、国際会議の登録に関する電話会話とキーボード会話をドメインとしている。その会話コーパスについて変換知識のデータベースを構築するため、頻度調査や検索を行なっている。調査対象となる文の数は現在、14616文である。使用頻度の高い上位10文を表3-1に示す。

表3-1　文の使用頻度

順位	文	件数	頻度
1	はい	1882	(12.8%)
2	もしもし	274	(1.9%)
3	わかりました	261	(1.8%)
4	はい、わかりました	177	(1.2%)
5	そうですか	152	(1.0%)
6	どうも、ありがとうございました	92	(0.6%)
7	はい、そうです	75	(0.5%)
8	ありがとうございました	65	(0.4%)
9	さようなら	61	(0.4%)
10	あ、そうですか	56	(0.4%)

また、機能語を中心にデータベース検索を行なって頻度統計をとった上位10文型の結果を表3-2に示す。

表3-2　文型の使用頻度

順位	文型	件数	頻度
1	～ます	1441	(9.9%)
2	～です	912	(6.2%)
3	～ました	733	(5.0%)
4	～ですか	522	(3.6%)
5	～ております	511	(3.5%)
6	～でしょうか	505	(3.5%)
7	～致します	444	(3.0%)
8	～ございました	322	(2.2%)
9	～ですね	295	(2.0%)
10	～ございます	237	(1.6%)

上の結果を見てわかるように、国際会議の参加登録というドメインは、かなりの割合で定型表現によって占められており、変換知識によるドメインのカバー率の高さが期待できる。

しかし、変換知識として認めるためには高頻度な表現であるだけではだめで、変換先の目的言語表現がある程度方向づけられないといけない。すなわち、出現する毎に全く違う表現に訳されるものは変換知識として設定できない。ドメインによ

る対訳の傾向も変換主導型翻訳の重要な要素となる。例えば、前述の対話コーパスで「～致します」は第7位の頻度の272件、「～たいの（ん）ですが」は108件あり、それぞれ対訳が"we will ～"，"I would like to ～"にほぼ方向づけられるので、変換知識として適当であると言える。

4．翻訳実験システムの構成

日英翻訳を例に変換主導型機械翻訳のシステム構成について述べる。

4.1．変換知識の実現

翻訳処理に必要な変換知識として現在は次の3つのレベルを設定している。

（1）単語列に関する変換知識

「もしもし」、「はい」、「ありがとうございました」、「分からない点」など構成する単語がすべて確定している表現についての変換知識

（2）空所パタンに関する変換知識

「～たいのですが」のような文型、「～の～」など確定しない部分～を持った表現についての変換知識

（3）意味構造に関する変換知識

深い解析処理により得られた意味構造を変換するための知識

これらの知識はそれぞれのレベルの原言語表現と目的言語表現の対となっている。

（1）には「もしもし」という単語列の原言語表現と"hello"のような単語列の目的言語表現の対となっている。

（2）には原言語表現「～たいのですが」と目的言語表現"I would like to ～"のような文型の対として記述される。

（3）は意味表現の対により表現される。

変換知識は（1）、（2）、（3）の順にそれを使って行なう処理が複雑になっており、できるだけ（1）、（2）の範囲で翻訳しようと試みる。

（1）、（2）の知識には次のような観点によって選ぶ。

- 3節で示した文や文型のような高頻度な表現や言い回し
- 単語から構成的に目的言語表現を作るのが困難な表現

4.2．対訳決定部の実現

原言語表現の変換先の対訳候補が複数ある場合、変換知識の種類により変換の条件の記述内容が異なり、最適な目的言語表現を選ぶ対訳決定の方法も異なる。

- 前文参照による決定

単語列に関する知識の場合、次のように条件は前文や周囲の情報によって記述される。

```
はい  ＝＞  yes     （前文が一般疑問文），
           no      （前文が否定疑問文），
           hello   （前文なし）
```

「はい」を翻訳する場合、前文を参照する。前文が例えば「会議事務局ですか。」の場合、"yes"に、「はい」が会話の先頭ならば"hello"に変換される。

- 用例類似度計算による決定

空所パタンに関する知識の場合は空所～の用例類似度計算[7]によって、変換する目的言語表現を決定する。すなわち、空所パタンの場合、条件部には用例が記述され、空所部に照合した入力文の語句に対して、シソーラス上の類似度が最も高い用例を持つものを対訳として選ぶ。

```
～は～です  ＝＞
        ～is～      （（名前,鈴木）,...），
        payment should be made by～
                    （（会費,現金払い）,...）
```

例えば、「参加費は銀行振込です。」は、「～は～です」の空所部に相当する（参加費,銀行振込）と用例（会費,現金払い）との類似度が

会費≈参加費、現金払い≈銀行振込

であるので、「～は～です」の対訳として"payment should be made by ～"が選ばれる。この変換を個々の単語の変換結果から構成的な手法で実現するのは難しい。「～は～です」の空所部がそれぞれ、支払名目、支払方法に関する語句の時、このような翻訳を行なうという手続きを用例類似

度に基づいて行なっている。なお、用例類似度はシソーラス上の概念間の距離に基づき計算され、0の場合が最も類似度が高いことになる。

4.3.形態素解析の実現

日本語辞書を参照して形態素解析が行なわれる。各形態素には単語列および空所パタンの変換知識へのインデックスが与えられている。例えば、「たい」、「の」、「です」、「が」は「～たいのですが」に対するインデックスを持っている。インデックスにより示された変換知識のみが入力文の構造の候補となり、探索空間を小さくすることにより高速な翻訳処理の実現に貢献する[8]。

4.4.変換機構部と検索部の実現

変換主導型の翻訳システムは入力文の翻訳に必要な変換知識が何なのかを抽出する。

- 日本語の構造抽出

形態素からのインデックスにより抽出された表現が、入力文に適合するかどうか検証する。「です」が示す空所パタンについて、「こちらは会議事務局です。」という入力文に、「～は～です」は適合するが、「～たいのですが」は適合しない。

形態素解析をした後、入力文全体に対して、単語列、空所パタンの順に抽出を行なう。空所パタンでも抽出できなければ解析を行ない、意味表現を導出する。単語列パタンとマッチすればその部分の抽出は終了する。空所パタンとマッチすれば、空所部について同様に単語列パタンから抽出を行なう。入力の全区間にこれらのいずれかの処理がなされるまで同様の処理を行なう。

- 変換知識の検索

日本語の構造を抽出した後、変換知識を検索して各部分構造の表現ごとに英語部分構造表現へのマッピングを行なう。マッピング先の表現が複数個ある場合は、4.2節で示したように前文参照、用例類似度計算などによりマッピング先を一意に決定する。

4.5.英文生成の実現

変換されて得られた英語の構造表現から英文を生成する。変換された英語の部分構造表現を組み立てて英文生成を行なう。複数の構造が結果としてある場合は用例類似度計算の総和によって最適なものを選ぶ。

また、英語辞書により屈折、法などの処理を行ない、精密な英文を生成する。

4.6.動作例

例えば、「案内書に記載されている内容についてお聞きしたいのですが。」と言う入力文に対しては、まず、入力文全体に次のような空所パタンの変換知識を適用することができる。

～たいのですが　＝＞　I would like to～

次に、空所部に相当する「案内書に記載されている内容についてお聞き」には

～について～　＝＞　～about～

という空所パタンの変換知識を適用する。ここで二つの空所部「案内書に記載されている内容」、「お聞き」に対し、それぞれ変換知識の抽出を行なう。「案内書に記載されている内容」には、単語列、空所パタンいずれの変換知識も適用できず、この語句に対して解析処理を行なって意味構造を抽出し、意味構造の変換を行なう。

「案内書に記載されている内容」
の意味構造　＝＞
"the content which is written in the announcement"
の意味構造

「お聞き」に対しては次のような単語列変換知識を適用する。変換のマッピングが一意でないので文内文脈処理を行なう。

お聞き　＝＞　ask us　（命令文）
ask you　（非命令文）

「～たいのですが」という願望を表す文で使用されているので"ask you"を目的言語表現として選択する。

各目的言語表現を合成することにより対訳文を生成する。

"I would like to ask you about the content which is written in the announcement"

]

5．基礎実験

Symbolics Lisp Machineを用いて変換主導型機械翻訳のパイロット･システムを作成した。前節のシステム構成に準拠した設計となっている。ただし、基礎実験では動作原理を確認することを目的とし、単語列と空所パタンで国際会議の参加登録に関する電話会話の代表的文の翻訳をどれだけカバーできるかの検討に重点を置いた。そのため、解析処理、前文参照などのモジュールは今回組み込まず、システム構成を簡略化している。変換知識もコーパスの検索や頻度に基づくものでなく、暫定的な小規模なものを与えている。

パイロット･システムでは、図5-1のように、途中、変換知識に基づいて日本語、英語、それぞれの構造表現が作られ、表示される。

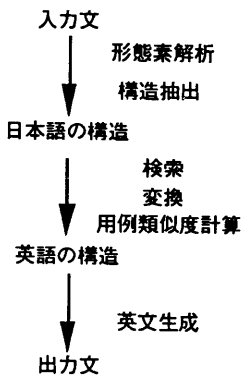

図5-1　パイロット･システムのデータの流れ

パイロット･システムで翻訳実験したいくつかの例を、この処理の流れに従って示す。

（a）単語列変換知識のみの適用による場合

入力文「もしもし」

↓

日本語構造　もしもし

入力文全体が次の単語列変換知識だけで覆われたので構造の抽出は直ちに終了する。

単語列変換知識　　もしもし　=>　*hello*

↓

英語構造　hello

↓

出力　hello

（b）空所パタン適用して翻訳する場合

入力文「会議に参加したいのですが。」

↓

日本語構造

([（会議）に（参加し）]たいのですが)

この構造に対して次のような変換知識を適用する。

空所パタン変換知識

～たいのですが=>　*I would like to* ～

$\sim_x$に$\sim_y$　　=>　$\sim_y$, $\sim_x$.

単語列変換知識

会議　=>　*the conference*

参加し　=>　*attend*

↓

英語構造

(I would like to [(attend)(the conference)])

↓

出力文

I would like to attend the conference

（c）類似度計算により対訳を決定する場合

入力文「参加費は銀行振込です。」

↓

日本語構造

（（参加費）は（銀行振込）です）

4.2節で述べたように、「～は～です」というパタンは少なくとも次の二つの対訳を持つので用例類似度計算を行なう。

↓

英語構造

① ((attendance fee) is(bank-transfer))

用例類似度　2

② (payment should be made by (bank-transfer))

用例類似度　1.5

用例類似度の良い（数値が低い）②を選択する。

↓

出力文

payment should be made by bank-transfer

（d）日本語の構造に複数の可能性がある場合

入力文「彼は会議に参加したいのですが。」

入力文の構造として適合する次の二つの変換知識があるとする。

空所パタン変換知識

～たいのですが　=>　I would like to～

～は～たいのですが　=>　～would like to～

それぞれの知識へのインデックスを基に複数の構造が得られる。

↓

日本語構造

（[（彼）は（（会議）に（参加し））]

たいのですが）

([彼]は

[（会議）に（参加し）]たいのですが）

↓

英語構造

③(I would like to [(he)((attend)(the conference))])

累計用例類似度　1.5

④([he] would like to [(attend)(the conference)])

累計用例類似度　1

「～は～たいのですが」の方が「～たいのですが」よりも累計用例類似度の値が低く、類似性が良い。

↓

出力文

類似性の良い④から得られる英文を解とする。

he would like to attend the conference

6．おわりに

高速、高品質の機械翻訳方式の実現を目指した変換主導型機械翻訳について述べた。変換主導型翻訳は変換知識を参照しながら必要最小限の処理だけを駆動し、効率的に翻訳を行なう。解析処理部起動については従来より研究を進めてきた語彙主導型解析手法[3]を適用することを試みる。基礎実験により変換主導型翻訳の動作原理の実現性が確認された。今後は、ドメイン上での検索、統計調査に基づく変換知識を作成し、翻訳実験を行う。高速、高品質の翻訳をドメイン上でどれだけ達成できるか、処理と知識の独立性という前提が適切かどうかなど、問題点の分析を通し、変換主導のより柔軟な機構を追及していく。

謝辞

本研究について有益な議論をしてくださったATR自動翻訳電話研究所の樽松社長、言語処理研究室、データ処理研究室の各位に感謝致します。

参考文献

(1) 池原、宮崎、白井、林：言語における話者の認識と多段翻訳方式，情報処理学会論文誌，Vol.28, No.12, (1987)

(2) 井佐原、田中：融合方式による機械翻訳，「自然言語処理技術」シンポジウム，(1983)

(3) Kogure, K. Iida, H. Hasegawa, T. and Ogura, K. : NADINE An Experimental Dialogue Translation System from Japanese to English, InfoJapan '90, (1990), Part 2: 57-64.

(4) Nagao, M. : A framework of a mechanical translation between Japanese and English by analogy principle, in Artificial and Human Intelligence, ed. Elithorn, A. and Banerji, R., North-Holland , (1984),

(5) Sadler, V. : Working with Analogical Semantics, Foris Publications, (1989).

(6) Sato, S. and Nagao M. : Toward Memory-Based Translation, Proc. of Coling '90, (1990).

(7) Sumita, E. Iida, H. and Kohyama, H. : Example-based Approach in Machine Translation, InfoJapan '90, (1990), Part 2: 65-72.

(8) Sumita, E. and Tsutsumi, Y. : A translation aid system using flexible text retrieval based on syntax matching, Proc. of 2nd MT Conference, (Pittsburgh,1988).

(9) Tsujii, J. and Fujita K. : Lexical Transfer based on Bi-Lingual Signs -Towards Interaction during Transfer-, Issues in Dialogue Machine Translation, CCL/UMIST Report number 90/5, (1990).