ICOT Technical Memorandom: TM-1049

TM-1049

「実例に基づく翻訳」における類推

佐藤　理史（京都大学）

May, 1991

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

TM–

「実例に基づく翻訳」における類推

佐藤理史
京都大学工学部 電気工学第二教室
(ssato@kuee.kyoto-u.ac.jp)

1991 年 4 月 18 日

要 旨

機械翻訳の分野で研究されている「実例に基づく翻訳」は、自然言語処理を対象とした応用指向の類推研究と考えることができる。本稿では、「実例に基づく翻訳」で用いられている類推について、類推を用いる目的、類推のタイプ、類推の実現法、翻訳の特異性等について検討する。

1 はじめに

松原 [11] は、類推の説明として、三省堂新明解国語辞典第三版の説明を引用している。

> 既得の知識を応用して、同じ条件にある未知の事物について多分そうではないかと判断を下すこと。

このような言葉で言い表される類推は、人間の日常的な推論形態の 1 つであり、人工知能の分野では、古くから研究されてきている [1]。

一方、近年、機械翻訳の分野で現れてきた「実例に基づく翻訳」[17][3][13][7] は、以下のようなパラダイムである。

> 既得の翻訳例を模倣利用することによって、翻訳を行なう。

そもそも、その発端は、長尾による「アナロジーによる翻訳」[2] であり、また、Case-based Reasoning[9]、あるいは、Memory-based Reasoning[5] の機械翻訳への応用 [14] ともみなせることからも、「実例に基づく翻訳」は類推と深く関連していることがわかる。類推研究における「実例に基づく翻訳」の位置付けは、自然言語処理を対象とした応用指向の類推研究ということになるだろう。

本稿では、「実例に基づく翻訳」における類推について検討し、以下の疑問に答えることを目的とする。

1. 何故、類推を用いるのか。

2. どのような類推を用いるのか。

3. 類推のコンポーネントをどのように実現しているのか。

4. 「翻訳」というタスクの特異性は何か。類推ではそれをどのように利用しているか。

5. 類推研究へのフィードバックは何か。

以下、まず、2章では、1について検討する。次に、3章では、以降の議論の準備として、変換透過型類推について議論する。4 6章では、実例に基づく翻訳の3つのタイプのシステムに関して、2-4を中心に議論し、最後にそれらをまとめて、5への答を与える。

2 何故、類推を用いるのか

松原 [11] は、類推を使う理由を以下のように述べている。

> 何か未知の領域について知りたくなったとき、その領域に関する情報が少なすぎてまっとうな手段 (たとえば演繹) では推論ができそうもなければ、それに類似した既知の知識の領域の知識を用いて推論を行なう。それが類推であり、...

つまり、類推を用いる理山は、

- 問題領域における情報の不足
- 演繹不能

の2点に要約される。この2点は、多くの場面において、類推を用いる理由として適切であろう。

しかしながら、実例に基づく翻訳では、それより強い、以下のような主張がなされる。

> 演繹的翻訳より類推的翻訳の方が優れている。

つまり、これは、たとえ演繹的な方法が取れたとしても、類推的方法の方を優先して用いようというものである。この主張には、以下のような背景がある。

- 工学的見地から
  - システムの構築容易性 : 実例に基づく方法では、実例をそのまま知識源として利用するため、規則獲得のプロセスが不要である。

- 知識のポータビリティと安定性 ： 実例は、他のシステムに移植可能であるとともに、長期に渡って安定しており、風化することが少ない。

- 翻訳タスクの特異性

  - 翻訳においては、原理・原則より、個別性の方が優位である。言い換えれば、少数の一般的な規則で記述することが困難である。

これらを要約すると、

> 安定した一般理論がない (作れない) 分野に対して、類推による推論システムが有効である。

ということであり、実例に基づく翻訳は、それを実証しようとする試みと考えられる。

このように、実例に基づく翻訳では、一般の類推よりも、より積極的に類推を使おうとするものであり、その点は特筆に値する。

3 変換透過型類推

ここでは、4-6 章の準備として、変換透過型類推について議論する。

3.1 属性投射型類推と変換透過型類推

類推の基本原理から出発して、以下のような 2 つのタイプを導入しよう[1]。

まず、基本原理の 1 つめは、以下のようなものである。

> あるものに成り立つことは、それと似たものにも成り立つ (ことが多い)。

このような基本原理から、あるものに成り立つ性質 (属性) を、似たものにも成り立つとして投射するというモデル化がなされる。これを属性投射型類推と呼ぼう。有馬 [10] が扱っている類推は、このタイプの類推である[2]。

基本原理の 2 つめは、以下のようなものである。

> 類似性は、ある種の変換において保存される (ことが多い)。

ここで、ある種の変換とは、原因から結果を求めるとか、入力から出力を求める等であり、その変換において、入力が似ていれば、出力も似ているだろうということである。 Winston

[1] このタイプ分けは、網羅的ではない。説明のための便宜的なものと考えてほしい。

[2] 本稿では、このタイプの類推については、これ以上議論しない。このタイプの基本図式等は、有馬 [10] に述べられている。

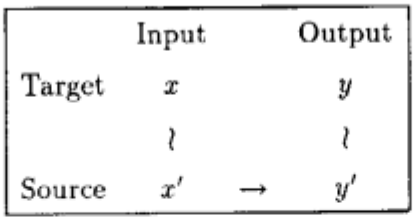

図 1: 変換透過型類推の基本図式

の類推原則 (類似性は因果関係を保存する。すなわち、似た状況は似た結果を生じやすい)[8] 等がこれに含まれると考えてよいだろう。これを変換透過型類推と呼ぼう。

実例に基づく翻訳は、

> 似たような文 (入力) は、似たような翻訳文 (出力) に翻訳されるだろう

という原則を用いていることから明らかなように、変換透過型類推に分類される。以下では、このタイプの基本図式と下位分類を導入する。

3.2 変換透過型類推の基本図式

変換透過型類推の基本図式は、図 1で与えられる。すなわち、入力 x に対して出力を得る場合、

1. まず、x に良く似た入力を持つ類推源 $x' \rightarrow y'$ を探し、
2. その出力 y' から、x に対する出力 y を推定する

いうことである。なお、この基本図式では、対象とする変換を、入出力変換とし、利用する類推源 (実例) を 1 つとした。

この類推は、上記の 1, 2 に対応したコンポーネントによって実現される。すなわち、

1. 類似性の検出、あるいは、類推源の選択 (決定)
2. adjustment[3]、あるいは、差異の解消、出力の生成

前者は、類推で使う類推源の決定と、入力側の類推目標と類推源の比較、差異の抽出までを担当する。すなわち、入力側の全ての処理を担当する。これに対して、後者は、出力側の処理を担当する。すなわち、入力側の差異を出力側で解消し、出力を生成することを担当する。

[3]適切な訳語が思い当たらないので、このまま用いる。具体的には、ターゲットとは部分的に異なっている類推源を、ターゲットに適用できるようにする処理のことを指す。

表 1: 入出力構造による分類

type	Input	Output
SY–SY	symbol	symbol
SY–ST	symbol	structure
ST–SY	structure	symbol
ST–ST	structure	structure

3.3 入出力構造による分類

上記の基本図式において、入力、出力のそれぞれが、1つのシンボル (SYmbol) か、それとも、シンボルを原子とするある種の構造 (STructure) かどうかによって、表1のタイプを導入する。

このうち、SY–SY, SY–ST は、研究する意味がほとんどない。なぜならば、シンボル間の類似度は、外部から定義するしかなく、定義された類似度だけで類似性の判定がなされ、類推結果が定まるからである。重要なのは、ST–SY と ST–ST である。ST–SY は、いわゆる選択問題である。すなわち、シンボルを要素とする出力集合 Y が与えられたとき、入力 x に対して、出力 $y \in Y$ を Y から選ぶという問題である。この問題は、前記の1のコンポーネントだけで実現できる。これに対して、ST–ST は、出力が部分構造を持ち、単純な選択問題にはならないような問題である。この場合は、2のコンポーネントが必要になる。

4 用例検索による翻訳支援

実例に基づく翻訳の第1のタイプは、用例検索による翻訳支援システムである。隅田らの ETOC[6]、中村のシステム [12] がこのタイプに属する。その基本思想は、以下で与えられる。

翻訳したい文に似た文とその翻訳例を提示することによって翻訳支援を行なう。

例えば、翻訳したい文、

君は水泳が大変うまい

を入力として、データベースを検索し、それに似た文とその対訳、

君は全く芝居がうまい → You're a great actor.

を出力しようということである。

このタイプのシステムの問題設定は以下のようになる。

入力　文(単語列)
タスク　入力に最も良く似た文(複数可)をデータベースから検索する。
出力　文(単語列)とその翻訳例

つまり、システムがすべきことは、データベースの中から、入力に最も良く似た文を選択することである。比較の対象となる文は、単語列、すなわち、シンボルを要素としたリストであり、出力は、データベース中のデータのIDであるから、この問題は、典型的なST SY型となっている。すなわち、すべきことは、文間の類似度を定義すること、すなわち、最適照合(best match)を見つける手続きを与えることである。

文(単語リスト)間に類似度を定義する方法として、ETOCと中村のシステムの両者とも、基本的には、まず、シンボル(単語)間に等価性を定義し、それに基づき構造(単語リスト)間に類似度を定義しているとみなせる。この点では、両者のシステムは共通であるが、実際にどのように類似度を定義するかについては、以下のように、かなり方針が異なっている。

- ETOCでは、主動詞と付属語、すなわち、構文的パターンを重視する。具体的には、入力文に対して、重要度の低い単語から順に変数化する一般化規則を適用し、その結果とデータベース中の文を照合する方法で、最適照合を実現している。つまり、構文的情報を重視した類似度を採用している。

- 中村のシステムでは、文中に含まれる自立語だけに注目し、それぞれの文から抽出された自立語の集合の積集合の大きさによって2つの文の類似度を定義している。すなわち、意味的情報を重視した類似度を採用している。

この差異の当然の帰結として、得られる最適照合の検索結果に差異を生じる。すなわち、ETOCは、翻訳したい文と文パターンが良く似た翻訳例が得られるのに対して、中村のシステムでは、自立語をどう訳したら良いかを示す翻訳例が得られる。

以上まとめると、このタイプのシステムでは、比較的良く研究されているST-SY型の類推を用いており、文(単語リスト)に対する類似度の定義法を与えた点に類推研究に対する貢献がある。特に、何が知りたいか(どのような文パターンに訳されるか、それとも、どのような訳語に訳されるか)によって、類似度の定義が異なるということを例示した点が注目に値する。

5 部分的な訳語選択

実例に基づく翻訳の第2のタイプは、文の部分的な訳語選択を実例に基づいて行なうシステムである。その基本思想は、以下で与えられる。

最も似ている翻訳例に基づいて訳語を決定する。

このタイプに属するシステムには、MBT1[17][16] と EBMT[7] があり、前者は、動詞とその必須格 (名詞) からなる動詞フレームの訳語選択問題を、後者は、日本語の「A の B」の翻訳パターンの選択を扱っている。以下では、MBT1 に沿って議論を進める[4]。

MBT1 は、動詞とその必須格 (名詞) からなる動詞フレーム、例えば、

(eat he vegetable)

を相手側、つまり、

(食べる 彼 野菜)

に翻訳する問題を扱う。すなわち、MBT1 の問題設定は、以下で与えられる。

入力	動詞フレーム (動詞 +N 引数)
タスク	入力を翻訳する。
出力	相手言語の動詞フレーム (動詞 +N 引数)

前章のシステムからのステップ・アップは、単に、入力 (eat he vegetable) に良く似た翻訳例、例えば、

(eat she potato) → (食べる 彼女 じゃがいも)

を検索するだけでなく、he や vegetable に対する訳語を決めて、完全な出力を作り出す点である。すなわち、出力の生成 (adjustment) を含んだ ST-ST 型の類推問題となる。

MBT1 では、この問題を以下のように扱った。

1. 入力を翻訳することを、入力の各構成要素に対する適切な訳語を選択することとする。可能な訳語の候補は前もって与えておく。

2. 要素構成原理[5]を仮定し、入力に対して出力の候補集合を与える手続きを導入する。これによって、問題を、その候補集合からの選択問題に帰着する。

[4]以下の議論は、MBT1 に対してのみ有効である。EBMT は、単純な ST-SY 型の類推であり、MBT1 で直面したような問題を扱っていない

[5]全体の翻訳は、部分要素の翻訳結果から合成される。

例えば、'eat(2引数)'は、「食べる(2引数)」か「侵す(2引数)」に訳される、'he'は「彼」に訳される、'vegetable'は、「野菜」か「植物人間」に訳されるというという訳語の候補を前もって与えておく。要素構成原理を仮定することにより、(eat he vegetable)の訳は、それらの可能な組合せの中の1つとなる。なすべきことは、それらの組合せの候補集合の中から、最も適切なものを1つ選ぶことである、ということである。つまり、これは、ST-ST型の類推問題を、ST-SY型の問題に帰着して解くということである。

残された問題は、構造(ここでは、入出力対となる)間に類似度を定義することである。MBT1では、以下のような方法を取っている。

1. 比較対象の限定 : 動詞対と引数の数が一致するもののみを模倣の対象とする。

2. シンボル(単語対)間の類似度をシソーラスによって定義する。

3. それに基づき、構造(動詞フレーム対)間に類似性を定義する。

最適照合を探す範囲を限定している点と、シンボル間に等価性ではなく、類似度を(外から)定義している点に前章のシステムとの相違が見られる。

以上まとめると、MBT1においては、ST-ST型の類推問題を、ST-SY型に帰着して解くことができることを示した点に注目したい。これが非常に簡単に行なえた理由は、入力が定型(レコード形式)であることと、要素構成原理を仮定できたことによる。このことによって、出力(候補)の生成(adjustment)の問題を回避できたわけである。しかしながら、一方、以下のような疑問が生まれてきたのも事実である。

類推では、結局、選択問題しか解けないのか?

6 一文全体の翻訳

実例に基づく翻訳の第3のタイプは、一文全体の翻訳を実例に基づいて行なうシステムである。これを完全に満たすシステムはまだ存在しない[6]。ここでは、それに最も近い位置にあるMBT2[13][4]に沿って議論を進める。

MBT2は、単語依存構造で表わされた文、例えば

```
[[buy,v],
 [[he,pron]],
 [[book,n],
  [[a,det]],
```

[6]現在、筆者は、これを完全に満たすシステムの開発を計画している[15]。

```
[[on,p],
 [[politics,n],
  [[international,adj]]]]]]
```

を、相手側の単語依存構造、

```
[[買う,動詞],
 [[は,助詞],
  [[彼,代名詞]]],
 [[を,助詞],
  [[本,名詞],
   [[た,助動詞],
    [[れる,助動詞],
     [[書く,動詞],
      [[について,助詞],
       [[国際政治,名詞]]]]]]]]]
```

に翻訳する問題を扱う。すなわち、MBT2 の問題設定は、以下で与えられる。

入力　文 (単語依存構造 = 木構造)

タスク　入力を翻訳する。

出力　相手言語の単語依存構造

ここでの最大の問題は、単一の実例を模倣するだけでは、出力が得られないという点である。ここでは、入出力が定型ではないため、MBT1 で採用したような方法をそのまま適用することはできない。つまり、出力の生成 (adjustment) の問題を真面目に解決しなければならないわけである。

MBT2 では以下の方法で、この問題を解決した。

> 入力を翻訳することを、複数の翻訳例の部分を組合せ利用することによって実現する。

まず、翻訳例の部分的に対応関係が付く部分を翻訳ユニットとして定義する。次に、入力を翻訳ユニットの組合せとして表現する。これを表わす表現として、照合表現とよぶ表現を導入する。入力が一旦照合表現に変換されると、後は単純である。すなわち、照合表現に含まれる翻訳ユニットを各々の部分対応関係に従って相手側言語の照合表現に変換し、得られた照合表現を解いて、出力を得るわけである (図 2参照)。このような方法で、与えられた入力

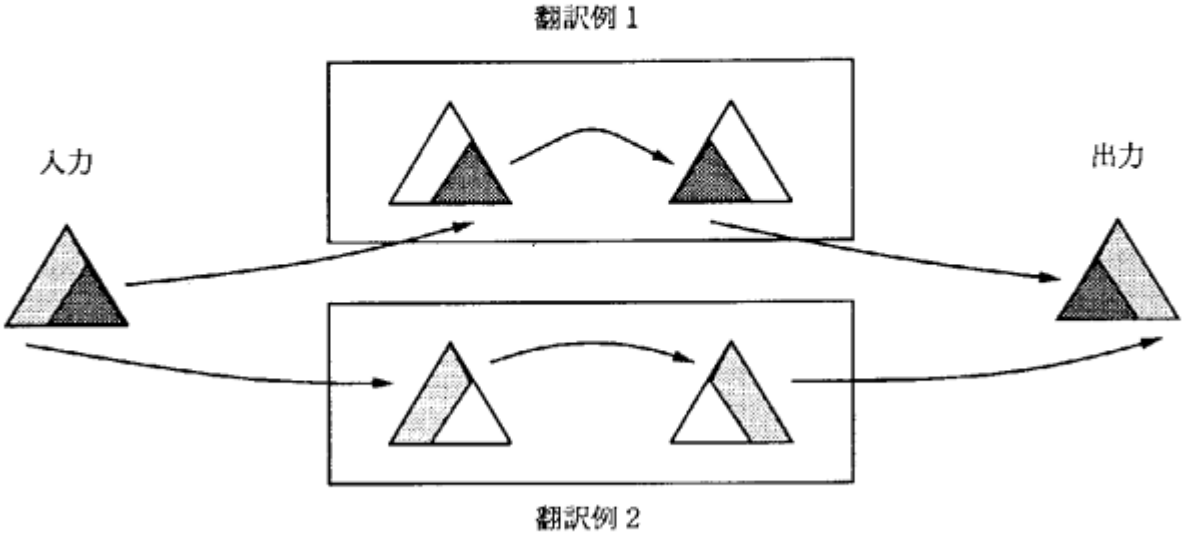

図 2: 複数の翻訳例の部分の利用

$$\begin{array}{ccc} x & & y \\ \wr & & \wr \\ x' & \rightarrow & y' \\ \hline x - x' & & y - y' \\ \wr & & \wr \\ x'' & \rightarrow & y'' \\ \hline & \vdots & \end{array}$$

図 3: 再帰的に類推を適用

に対して、出力の候補(複数)を求めることができる。これによって、MBT1と同様に、選択問題に帰着させることができる。

このことは、別な言葉で表現すると、複数の類推源を利用するということである。すなわち、MBT2では、図3のように、再帰的に類推を適用する。

このような方法が生まれた背景には、翻訳の特異性がある。すなわち、翻訳では、似たようなものを持ってくるだけでは不十分であり、完全照合しない限り、その部分の訳語がわからないという特徴がある。例えば、'a book on international politics'を訳す場合、

a book on economics → 経済学について書かれた本

という対訳を利用することができるが、この例からは、international politicsをどう訳せばよいかということは、決して得られない。

上記の、複数類推源の利用は、類推の確からしさを測る指標の変更を要請する。単一類推源を利用した類推では、その類推源との類似度がそのまま類推の確からしさを測る指標として利用できた。しかし、複数の類推源を利用する場合は、それでは足りなくなってくる。利用する複数の類推源とどのくらいうまく照合しているかということをもって、類推の確からしさを測る指標としなければならない。

MBT2では、これを照合表現の得点として定義し、類推の確からしさを測る指標、すなわち、選択問題を解く指標として利用する。照合表現の得点は、よい翻訳を得るためのヒューリスティクスに基づき、以下のように定義した。

$$\text{翻訳ユニットの得点} = \text{大きさ} \times (\text{内的類似度} + \text{外的類似度})$$

$$\text{照合表現の得点} = \frac{\sum \text{翻訳ユニットの得点}}{\text{大きさ}}$$

ここで、外的類似度がはじめて姿を表わしたことに注目してほしい。MBT2では、類推源(翻訳例)を部分的に利用する。そこで、その部分が、どのくらいうまく全体にフィットするかを測る必要が生じ、それを外的類似度によって測ろうというわけである。

以上まとめると、MBT2では、出力の生成(adjustment)の問題に真面目に取り組んだ結果、複数の類推源の部分を利用するという方法に至る結果となった。結果的には、ST-ST型の問題をST-SY型の問題に変換して解くという点では、MBT1と同じである。しかし、複数の類推源の利用は、単一類推源の類推では常識であった類似度をそのまま類推の確からしさの指標とするという点に変更が必要になった。この点が注目に値する。

7 議論

本章では、前3章の議論をまとめ、筆者の考えについて述べる。

7.1 類推の図式

実例に基づく翻訳は、翻訳という変換に対する変換透過型類推を用いている。用例検索による翻訳支援システムは、ST-SY 型であるが、他の 2 つは、ST-ST 型である。しかし、これを解く場合、入力から出力を求める何らかの方法を与えて出力の候補集合を得、そのなかから最も適当なものを選ぶという選択問題 (ST-SY 型) に帰着することを行なっている。

類推は選択問題しか解けないのか。この問題は、未回答の問題である。しかし、利用できる類推源は有限であると考えるならば、結局、そのどれを (あるいは、どのような組合せを) 選ぶかということに問題が帰着されてしまうのは、ある意味では当然であると考える。

7.2 類推のコンポーネント

類推は、類推源の決定と出力の生成 (adjustment) の 2 つのコンポーネントから構成される。このうち、後者は、用例に基づく翻訳支援にはなく、MBT1 ではないに等しい。唯一、MBT2 だけが、このコンポーネントを持っている。その実現法は、複数の類推源の部分を組み合わせるという方法である。複数の類推源の利用、それ自体は、一般的に利用できる方法ではあるが、出力の生成機構として、それをそのまま利用できるのは、翻訳のように、入出力の部分間に部分対応関係が存在するような問題に限られるかもしれない。

7.3 類似性、類似度

実例に基づく翻訳では、構造間に類似性を定義することを行なっている。扱っている対象は、シンボルを原子とする構造であり、シンボル間の類似度を定義し、それに基づいて構造間の類似度を定義するというように、ボトムアップに積み上げるという方法を取っている。シンボル間の類似度は、{0,1} の等価性として与える場合と、[0,1] の類似度を外から与える方法がある。複数の類推源の部分的利用を行なう場合には、内的類似度のほかに、外的類似度を持ち込んでいる。なお、上記の議論では、明示的には述べなかったが、類似度を測る対象が定型である場合と、不定型である場合において、類似度の定義の難易度 (あるいは類似度を計算する計算量) がかなり違う。定型の場合はやさしいが、不定型の場合は、難しい。

用例検索による翻訳支援システムでは、得たい結果が異なる場合、類似度の定義が異なることが示された。このことは、重要である。いずれにしても、類似度の定義は、領域依存・タスク依存であり、各領域・タスク毎に、適切な類似度の定義を見つける必要がある。

表 2: 実例に基づく翻訳における類推：まとめ

何故類推か	演繹より類推の方が優れている
類推の図式	変換透過型類推 ST-ST → ST-SY (選択問題に帰着)
類推源の検索	制約による絞り込み
類似性	シンボル間の類似性を定義 構造間へ積み上げる
妥当性	(選択問題に帰着)
その他	現実指向

7.4 翻訳の特異性

翻訳問題は、部分問題に分割可能である。あるいは、別な言葉で言うと、入出力間に、部分的な対応関係が存在する。この性質を、実例に基づく翻訳ではかなり利用している。また、もうひとつの翻訳問題の特異性として、完全照合しなければ、訳語がわからないという問題がある。つまり、似たような実例を探してくるだけでは、その差異を翻訳できない。このことから、再帰的に類推を適用して翻訳するという考え方が素直に生まれてきた。

8 おわりに

本稿では、「実例に基づく翻訳」における類推について議論した。以下にまとめを表2として示す。

最後に、翻訳というタスクは、類推の応用対象として、非常におもしろいタスクであるという筆者の感想を付記する。

謝辞

本稿をまとめるにあたって、90 年度 ICOT ANR-WG における議論が有益であった。ANR-WG の委員、オブザーバー各位に感謝する。

参考文献

[1] Hall, R.P., Computational Approaches to Analogical Reasoning: A Comparative Analysis, Artificial Intelligence, Vol.39, pp39–120, 1989.

[2] Nagao, M., A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in ARTIFICIAL AND HUMAN INTELLIGENC (Elithorn & Banerji, Eds.), Elsevier Science Publishers, pp173-180, 1984.

[3] Sadler, V., Working with Analogical Semantics, Foris Publications, 1989.

[4] Sato, S. and Nagao, M, Toward Memory-based Translation, Proc. of COLING90, Vol.3, pp247–252, 1990.

[5] Stanfill, C. and Waltz, D., Toward Memory-based Reasoning, Comm. of ACM, Vol.29, No.12, pp1213–1228, 1986.

[6] Sumita, E. and Tsutsumi, Y., A Translation Aid System Using Flexible Text Retreval Based on Syntax-Matching, TRL Research Report, TR87–1019, IBM, 1988.

[7] Sumita, E., Iida, H., and Kohyama, H., Example-based Approach in Machine Translation, IPSJ, Proc. of InfoJapan'90, Part:2, pp65-72, 1990.

[8] Winston, P.H., Learning and Reasoning by Analogy, Comm. of ACM, Vol.23, No.12, pp689–703, 1980.

[9] Case-based Reasoning from DARPA: Machine Learning Program Plan, Proc. of Case-based Reasoning Workshop 89, Morgan Kaufmann Publisher, pp1–13, 1989.

[10] 有馬淳, 類推の正当化問題に関する論理的分析と一弱正当化法, 情報処理学会研究報告, AI–75–13, 1991.

[11] 松原仁, 類推による学習, ICOT ANRWG 配布資料, 1991.

[12] 中村直人, 用例検索翻訳支援システム, KSA フォーラム / 自然言語処理分科会資料, 1989.

[13] 佐藤理史, 実例に基づく翻訳 II, 情報処理学会研究報告, AI-70-3, 1990.

[14] 佐藤理史, Memory-based Reasoning の挑戦 — もう、ルールなんていらない?—, 1990 年度 日本認知科学会シンポジウム資料集, p1–10, 1990.

[15] 佐藤理史, 実例に基づく翻訳の超並列化に向けて, Workshop on Learning '91, 1991.

[16] 佐藤理史, MBT1: 実例に基づく訳語選択, 人工知能学会誌, Vol.6, No.4, 1991.

[17] 佐藤理史, 長尾真, 実例に基づいた翻訳, 情報処理学会研究報告, NL-70-9, 1989.