用例翻訳の為の対訳例からの自動的翻訳パターン抽出の一手法*

6P−5

渡辺 日出雄†
日本アイ・ビー・エム株式会社 東京基礎研究所‡

1 はじめに

近年、Example-Based Approach (EBA)[3] を用いた翻訳手法が盛んに研究されてきている。EBA をトランスファー処理に用いるシステムとして [4, 6, 7] などがあるが、これらの手法では翻訳パターンとして解析木のペアを用いるため、翻訳パターンの収集が容易とは言えないのが現状である。また、従来の手法による翻訳システムにおいてもやはり翻訳規則を収集するのは人手に頼ることになり、同様の問題を抱えていると言える。一般に翻訳パターンの収集方法には、(1) 多くの翻訳例を収集しその中から発見的に翻訳パターンを見つける方法と (2) 誤った翻訳結果と正しい翻訳結果とを比べてその差から翻訳パターンを見つける方法の二つがある。本論文では、後者の手法を機械的に行なう方法について提案する。[8]

翻訳システム MT があり、入力文 S_s が与えられた時 MT の出力を S_t、正しい翻訳結果を S_c とする。また、S_s, S_t の依存構造 D_s, D_t 及び D_s と D_t の間の対応関係は MT により得られているものとする。この時、差分から翻訳パターンを得るプロセスは以下の3つの部分から構成されることになる。

(a) S_c の解析木 D_c を得る。
(b) D_s と D_c の間に対応関係を見つける。
(c) D_s と D_t の対応関係と D_s と D_c の対応関係とを見比べて必要な部分を翻訳パターンとして取り出す。

(a)(b) に関しては [2, 5, 1] などの研究がある。本論文では、(a) の解析を除いた (b)(c) に関する手法について述べる。

2 依存構造間の対応関係

D_s と D_c の間の対応関係を求めるために、まず日英辞書を用いて1対1の対応関係を付ける。[1] (これを語彙的対応関係 (lexical mapping) と呼ぶ。) しかし、以下の例でも分かるように、この語彙的対応関係だけでは不十分なことが多い。

彼女はいつもとても髪がきれいだ。
She always has very beautiful hair.

この例では、「きれい」↔"beautiful" という対応関係だけでは、「きれい」を修飾している「とても」の訳"very" が"beautiful" を修飾するが、「きれい」を修飾するもう一つの「いつも」の訳"always" が"has" を修飾することを説明できない。そこで、我々の EBA Transfer System (SimTran) [6, 7] では、より大きな句レベルの対応関係 (これを構造的対応関係 (structural mapping) と呼ぶ) を表す upward mapping($M\uparrow$) と downward mapping($M\downarrow$) を用いている。それを図2に示す。これによれば、「きれい」は $M\uparrow$によって"has" と、$M\downarrow$によって "beautiful" と対応し、「いつも」の様な動詞を修飾する副詞の訳語は $M\uparrow$ で関連づけられた単語を修飾し、「とても」の様に形容詞を修飾する副詞の訳語は $M\downarrow$で関連づけられた単語を修飾するというように表現可能である。

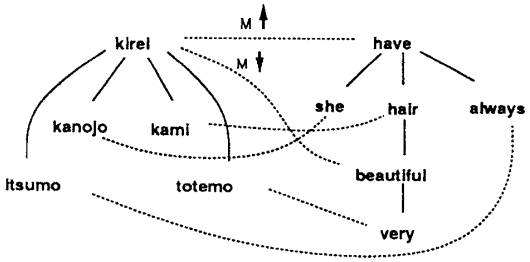

図 1: 対応関係付けの例

この構造的対応関係は以下のようにして求める。初期状態として語彙的対応関係を $M\uparrow$,$M\downarrow$とする。D_s の任意のノード x に関してその子孫ノード yの内で、x から $M\downarrow$で関連付けられた D_cのノードを x'、yのを y' とした時に、x'が y'の子孫ノードとなっているような y を見つける。次に、y'の祖先ノードを遡って、ルートであるか、もしくはその親ノードが対応付けられているノードを zとし、upward mapping から (x, x') を取り除き (x, z) を加える。

3 翻訳パターンの発見

ここでは、$\langle D_s, M_t, D_t\rangle$、$\langle D_s, M_c, D_c\rangle$ の差分を見つける手法について述べる。(ここで、M_t, M_cはそれぞれ D_s, D_t及び D_s, D_cの間の構造的対応関係である。) 翻訳システム MTによって、D_tを生成するために用いられた翻訳パターンは容易に得ることが出来る。翻訳パターン tp は $\langle P_s, M'_t, P_t\rangle$ の3つ組から構成される。(ここで、P_sは原言語側の依存構造、P_tは相手言語側の依存構造、M'_tはそれらの間の構造的対応関係である。) 次に、それぞれの翻訳パターンについて、$\langle D_s, M_c, D_c\rangle$ での対応する翻訳パターン (projected translation pattern, or

*A Method for Extracting Translation Patterns Automatically from Translation Examples for Example-Based Translation
†Hideo Watanabe (watanabe@trl.vnet.ibm.com)
‡IBM Research, Tokyo Research Laboratory
[1] これを厳密に行なうには [5, 1] の手法を用いれば良い。

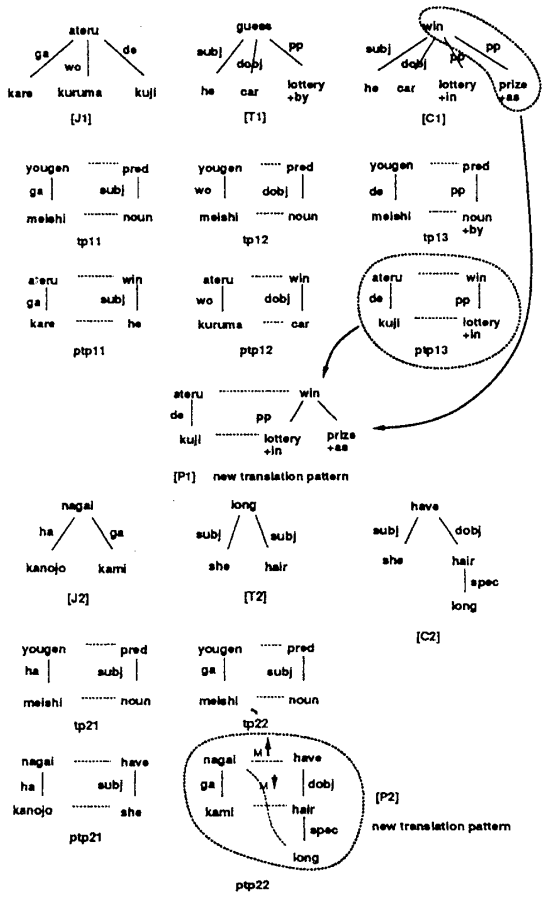

図 2: 翻訳パターン抽出の例

ptp) を求める。$tp = \langle P_s, M, P_t \rangle$ に関してその ptp は P_s、M_cの内 source が P_sであるものからなるサブセット (M'_t)、P_sから M'_tにより対応付けられている D_cのノードを含むような最小の連結サブグラフ (P_c) の3つ組 $(ptp = \langle P_s, M'_c, P_t \rangle)$ により表すことができる。

ある tp とその対応する ptp が得られた時、tp が ptp を包摂 (subsume) するとき tp と ptp は同型であるとする。よって、同型でない ptp を集めて結合することにより、その翻訳に不可欠な翻訳パターンが得られることになる。ただし、全ての ptp をもってしても D_cを全てカバーできていないことも有り得るので、その場合カバーされていない部分も新たな翻訳パターンの一部として結合する。この結合の結果は必ずしも一つの翻訳パターンとなる訳ではなく、場合によっては複数の翻訳パターンとなることがある。

4 例

実験は我々が開発した EBA Transfer System SimTran を用いて行なった。ただし、誤訳が出やすいように文法的なものを中心に 200 程度の翻訳パターンだけの状態にして行なった。以下に、日本文 [J]、SimTran での英語への翻訳結果 [T]、正しい翻訳結果 [C]、得られた翻訳パターン [P] を示す。また、図4にそれらと tp と ptp の依存構造を示す。

[J1] 彼はクジで車を当てた。
[T1]He guessed a car by lottery.
[C1]He won a car as a prize in lottery.
[P1] クジで当てる↔win as a prize in lottery

[J2] 彼女は髪が長い。
[T2]She hair is long.
[C2]She has long hair.
[P2] 髪が長い↔have long hair

[J1] に関しては、ptp13 が包摂されないので、これと [C1] でカバーされていない部分を結合して [P1] が得られている。[J2] に関しては、ptp22 が包摂されないのでそれがそのまま [P2] となっている。ptp22 の"nagai" は"have" と"long" にそれぞれ $M\uparrow$,$M\downarrow$で関連付けられている。

5 おわりに

本論文では、機械翻訳の結果と正しい翻訳結果を見比べることにより差分を見つけ、現在の翻訳システムにとって有効な翻訳パターンを抽出する手法を提案した。これは、用例ベースのトランスファーシステムだけに有効なわけではなく、広く一般のトランスファー手法においても、翻訳規則を見つける手法として有効である。ただし、問題点としては、誤りの度合が大きいものや正解が意訳されているなどの時に差分情報が大き過ぎてほとんど全文が対象になってしまうということがある。今後は、ユーザーインターフェースを提供してより簡単に翻訳パターンを取り出すことが出来るシステムを構築することを目指している。

参考文献

[1] 石本、宇津呂、松本、長尾、"日英対訳文間の構造照合"、自然言語処理研究会資料 95-11, 1993
[2] Kaji,H., Kida,Y., and Morimoto, Y., "Learning Translation Templates from Bilingual Text," Proc. of 14th Coling, 1992
[3] Nagao,M., "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," Elithorn, A. and Banerji, R. (eds.): *Artificial and Human Intelligence*, NATO 1984
[4] Sato,S. and Nagao, M., "Toward Memory-Based Translation"、Proc. of 13th Coling, 1990
[5] Utsuro, T., Matsumoto, Y., and Nagao, M., "Lexical Knowledge Acquisition from Bilingual Corpora," Proc. of 14th Coling, 1992
[6] Watanabe,H., "A Model of a Transfer Process Using Combinations of Translation Rules," Proc. of 1st PRICAI, 1990
[7] Watanabe,H., "A Similarity-Driven Transfer System," Proc. of 14th Coling, 1992
[8] Watanabe, H., "A Method for Extracting Translation Patterns from Translation Examples," Proc. of 5th TMI, 1993