対話データベースからの日英変換規則の抽出

4S−4

長谷川 敏郎
ATR 自動翻訳電話研究所

1 はじめに

機械翻訳システムにおいて、言語間で単語の対応づける変換規則は対応関係が複雑であり、その構築は容易ではない。ATRでは対話データベース[1]の構築が進められており、収録された対話に種々の言語的な分析がほどこされている。**単語に関する情報**としては、対話中に現れた単語、単語の読み、標準表現、品詞、活用型、活用形、音便が、**単語間の関係情報**としては、格・係り受け関係(係り受けの意味的関係も付与されている)が、さらに、文、文節、単語に対して**日英の言語間の対訳対応情報**が、既に付与されている。

ここでは、対話データベースから変換規則の自動抽出の第一段階として、単語情報(品詞、標準表現)、係り受け関係、単語における日英対訳対応情報を用いて、動詞の訳語選択規則の抽出を試みた。

2 言語データベースから訳語選択規則の抽出

動詞 v と v に係っている単語との共起関係から v の訳語を決定する規則を抽出すること試みた。一般には動詞の訳語は、動詞に係る複数の格関係によって決定されると考えられる。したがって、抽出すべき規則は格パタンを持ったものであるべきである。しかし、実際の対話では、動詞は任意格を伴ったり、必須格の省略が起こるので、係り受けのパタンは一様ではない。したがって、訳語選択規則の抽出を行なうにあたって、パタンが不定である係り受け構造は、そのままでは、訳語選択規則の抽出処理の操作対象に適していない。そこで、ここでは、動詞と動詞に係っている単語(格要素)、および、動詞と格要素との意味的関係(格役割)(以下、この三組を修飾子と呼ぶ)を規則抽出処理の操作対象とし、抽出される規則の形式も修飾子とした。

2.1 規則の抽出方法

ここでは、訳語選択規則の抽出アルゴリズムについて述べる。対話データベースから、動詞を中心とした係り受け構造と、その動詞の対訳を抽出する。ここでは、異る動詞間での操作は行なわないので、まず対話データベースから抽出した係り受け構造と動詞の対訳の対の集合を、動詞の標準表現で集合 $S(v_i)$ に分割する。さらに、集合 $S(v_i)$ を、動詞に与えられた対訳によって部分集合 $S(v_i)(t_j)$ に分割する。したがって、部分集合 $S(v_i)(t_j)$ の構成要素は、同じ対訳と同じ動詞の標準表現を持った係り受け構造である。

ある係り受け構造の集合 $S(v_i)$ において、下位の部分集合 $S(v_i)(t_j)$ 内で共通であり、$S(v_i)(t_j)$ 間で互いに相異なっている属性を抽出し、それらを組み立てて訳語選択規則を構成することができる。ここでは、訳語を決定するための属性抽出の操作対象は修飾子であり、部分集合間で互いに相異なる格要素を持つ修飾子を抽出し、それを訳語選択規則とした。

2.2 規則の一般化

抽出された規則(修飾子)に対して、似通った規則を一つの規則にまとめ上げることにより規則数を減らし、かつ、規則の適用範囲を拡大するために、規則の一般化を行なう。規則の一般化は、部分集合 $S(v_i)(t_j)$ 内の抽出された修飾子に対して、同じ格役割を持つ修飾子の格要素をシソーラスを用いて抽象化し、複数の格要素を一つの抽象的な概念で表現することで行なう。格要素をどの程度抽象化すべきか(シソーラス中のどの階層に位置付けるか)という問題がある。この問題に対して、(1)格要素(概念)の抽象度を上げた時、部分集合間で格要素が互いに相異なっていればその抽象化を認める、(2)部分集合間で互いに相異なっている格要素集合の Least Upper Bound を一般化の上限とする、という二つの方法が考えられる。ここでは、訳語選択規則を抽出する上で、対話データベースに含まれている語彙数が量的にまだ不十分であると考えられるため、過剰に一般化される恐れがあるが(1)の方法を採った。

2.3 シソーラスへの同定

規則の一般化は、格要素が同定されたシソーラス中の概念(シソーラス中のエントリ)を上位の方向に辿ることによって行なわれる。したがって、修飾子の格要素が複数の概念を内包している時には、どの意味で用いられているかによって、特定の概念に決定する必要がある。ここでは、動詞を特定した時、ある格役割に入り得る単語は似通っているという仮説に基づき、以下のヒューリスティックスで格要素をシソーラス中の特定の概念に同定した。

1. 単語が内包している各概念に対して、それらの概念の上位概念の頻度が高い概念を選択する
2. 単語が内包している各概念に対して、頻度の高い概念を選択する

二つのヒューリスティックスを用いても、概念を特定できない場合には、同レベルの複数の概念をパッキングし、一つの新たな概念として扱った。

A method to extract Japanese-English transfer rules from Dialogue Database
Toshiro HASEGAWA
ATR Interpreting Telephony Research Laboratory

表 1: 対話データベースからの抽出データ

会話数	292	係り受け構造数	5482
修飾子数	12312	動詞異なり語数	389
部分集合数	1915		

(注) メディアはキーボードおよび電話、話題対象は国際会議および旅行に関する問い合わせ

表 2: 抽出された規則数

修飾子数	level1	level2	level3	level4	level5
1 以上	2278	815	724	640	441
2 以上	211	77	81	72	48
3 以上	74	24	30	24	19

2.4 頻度情報の利用

規則抽出において頻度情報を利用することが考えられる。規則の一般化の際に、ただ一つの反例によって、規則の一般化を無効にするのではなく、確率付きで規則を一般化する。集合 $S(v_i)$ の下位の部分集合 $S(v_i)(t_j)$ 間で、格要素が互いに相異なっているか否かを無視して抽象化し、同じ格役割を持つ修飾子間において、抽象化された概念の部分集合間での出現確率がある値以上のものを規則として抽出する。その際、抽象化のレベルに応じて概念の出現確率が異ってくる。ここでは、抽象化された概念の出現確率が 1 の時には、その概念のみを規則化し、その概念の下位概念は規則化しない。一方、抽象化された概念の出現確率が 1 より小さい時には、その概念の出現確率がある値以上であれば規則化し、さらに、その概念の下位概念に対しても同様の方法で規則化を試みる。この方法では、シソーラスの各階層に応じて複数の規則が抽出される。そこで、規則適用時に、抽象度の低い規則から適用を試み、適用が成功すればそれより抽象度の高い規則の適用は試みないという方法を取った。

3 実験

対話データベースから規則抽出のために抽出したデータの概要を表 1に示す。シソーラスには、角川類語辞典 [3] を用いた。対話データベースから抽出できた規則数を表 2に示す。表 2において、修飾子数とは規則の一般化の対象となった修飾子数である。また、レベルとは規則の一般化の度合を示している。シソーラスは 4 階層からなり、level1 は最上位にまで一般化された規則数であり、level5 は一般化されなかった規則数である。

3.1 抜き取り試験

対話データベースで出現頻度の高い上位 20 の動詞について、抽出した規則でどの程度正しく訳し分けができるか実験した。実験方法は、各動詞の集合 $S(v_i)$ から一つの係り受け構造を抜き取り、残りの係り受け構造から規則を抽出し、抜き取った係り受け構造に規則を適用した結果がその係り受け構造の対訳と一致するか調べた。

訳語の決定方法は、係り受け構造を構成する各修飾子に規則を適用し、適用が成功した総ての規則が同一の対訳を持つ時には、その対訳を解とした。適用が成功した複数の規則が異なった対訳をもつ場合、同じ対訳を持つ規則の出現確率の総和が高いものを解とした。出現確率は、100% および 75%、50% より大きなものについて、訳語選択規則を抽出し、各々について実験を行なった。実験結果の一部を表 3に示す。

表 3: 抜き取り試験の結果

単語	修飾子数	偏り度	正解率 (出現確率)		
			(1.0)	(0.75)	(0.5)
送る	415	0.753	0.546	0.657	0.713
有る	275	0.634	0.349	0.429	0.483
思う	260	0.5	0.126	0.203	0.25
申す	180	0.978	0.944	0.955	0.955
成る	141	0.78	0.595	0.659	0.695
する	132	0.399	0.318	0.348	0.378
聞く	91	0.383	0.109	0.109	0.109
行く	88	0.438	0.295	0.295	0.375
教える	85	0.493	0.247	0.247	0.247
支払う	65	0.923	0.646	0.676	0.738

4 考察

実験の結果、動詞によって正解率にバラツキがある。動詞の対訳に偏りが大きいものは正解率もよく、正解率と対訳の偏りとは相関があることが分かる。各動詞について、対訳の偏りの度合は $\sqrt{\sum_{i=1}^{n}(\frac{t_i}{N})^2}$ (ここで t_i は同一の対訳を持つ係り受け構造数、N は係り受け構造総数) で計算した。確率を用いた規則は正解率が上がっており、訳語選択規則への頻度情報の利用は有効であると考えられる。知覚動詞の正解率が悪い。また、全体的に正解率の悪い原因の一つに対話データベースの量的な問題がある。表 2に示すように、一般化された規則のもとの修飾子数が 2 以上の規則数は、1 以上の規則数に比べ著しく少なく、対話データベースがまだスパースであることが分かる。さらに精密な規則抽出のために、品詞、表層格、複数の格の共起関係など、ここで利用しなかった情報の利用が考えられる。

5 おわりに

ここでは、単語を処理の単位として扱ったが、さらに、対話データベースで日英間の対応率が高い文節単位での実験や、他の品詞の訳語選択規則の抽出実験も試みる。また、規則抽出結果は、訳語選択規則として変換過程で直接利用するだけでなく、人手による訳語選択規則の作成のための重要な参考データとして活用できる。

謝辞: 貴重な助言を下さった言語処理研究室およびデータ処理研究室の諸氏に感謝する。

参考文献

[1] 江原暉将, 小倉健太郎, 森元逞: “電話対話データベースの構築”, 情全大 40 回, 1990

[2] 隅田英一郎、飯田仁、幸山秀雄: “用例に基づいた翻訳”, 情全大 40 回, 1990

[3] 大野晋, 浜西正人: “類語新辞典”, 角川, 1984