マルチコーパスを利用した多段用例翻訳方式

4L－9

池田修一　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

本稿では、「会話文」や「電子メール」など、論文やマニュアルに比べて、人間の感情や意志表現を多く含む文を対象にした日英用例翻訳システムの構築について論じる。用例翻訳では、一つの入力文に対し、一つのコーパスの中から、一つの類似用例を検索することを基本とするものが多い。それに対し、本システムでは、「辞」「述語をもつ詞」「述語をもたない詞」のそれぞれのレベルに固有のコーパスと類似評価法を用意し、一つの入力文に対し、複数の用例を適用する「多段用例翻訳方式」を提案する。これらのコーパスは構文解析済みの対訳コーパスで、状況によっても分類されており、状況によってどのコーパスから類似文検索するかを選択できる。類似文検索は、入力文の木とコーパスの木の間の「共通木」と「差分木」に重みをつける方法を用いるが、重みの付け方はどのレベルのコーパスを検索するかに依存する。また、用例翻訳では、二言語間の表現の対応づけが一つの問題となっているが、本稿では、一つの入力文に対し、日本語文の差分と英語文の差分の対応がとれるように二対の対訳用例を検索し、それらの目的言語側の二つの木の間で「共通木」と「差分木」をとる方法について述べる。

2　システムの概要

本システムの最大の特徴は、翻訳の際に、一つの入力文に対して複数の用例を適用するところである。コーパスは、構文解析済みの対訳コーパスで、いくつかの文法段階（レベル）に分かれ、状況によっても分類されている。これをマルチコーパスと呼ぶ。また、図1に示される複数のレベルモジュールが呼び出される順序は「辞」→「詞」→「句」のように一様に決まっているわけではなく、各モジュールがどのようなデータを次のモジュールに受け渡すかによって決まる。辞中心のコーパスの中に類似度の高い用例があれば一気に英文構造を抽出し、詞レベルのモジュールを飛ばし、句レベルのモジュールを呼び出すこともある。また逆に、類似度の高い用例がなければ、多くのモジュールを駆使して翻訳を進めることとなる。

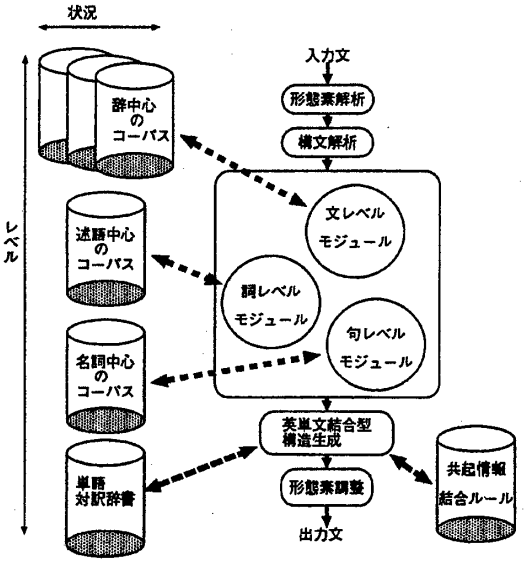

図1: システムの概要

3　モジュールの働きとデータ構造

入力文は形態素解析、構文解析を経て、木構造として翻訳部に受け渡される。木構造のトップノードによりどのモジュールが呼び出されるかが決まる。選ばれたモジュールは、外部記憶装置に収められているコーパスの中から、一次検索として類似文と対訳を n 文抽出し、それらを主記憶に収める。「二つの対訳用例による共通構造と差分構造の抽出法」により、「日本語側の差分構造」と「英語側の共通構造」を抽出し、そして英文生成の助けとなる「継承情報」を生成する。日本語側の差分構造が次のモジュールのデータとなる。英語側の共通構造は差分のところにリンクポインタを付与する。継承情報はリンクポインタをつ

Example-based Translation with several Level Module in available for Multi-Corpus

Syuichi Ikeda , Masahiro Miyazaki
Niigata University

なぐための情報や文の上位構造から継承される情報をリストの形で保持する。

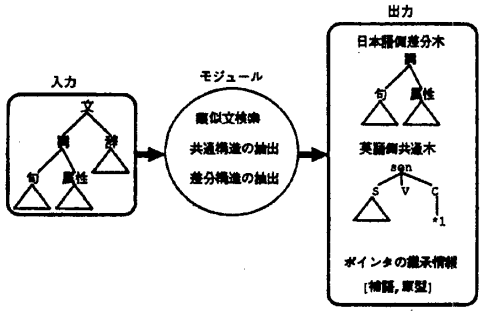

図 2: モジュールの働き

日英用例対の差分構造の対応づけとして、二つの対訳用例を利用する方法を提案する。検索された n 対の対訳用例をの中で、任意に二つの目的言語（英語）の木構造の組合せを選び共通構造と差分 Y を抽出する。ここで、日本語側の差分 X と英語側の差分 Y が対応すると考えられる。

抽出された共通構造が大きく、かつ、差分構造 Y が小さい二つの目的言語の組合せがより望ましい対訳用例対となる。この処理には ${}_nC_2$ 回の共通構造抽出が必要なため定数 n の決定は、計算機が実用時間上問題がない程度にしなくてはならない。

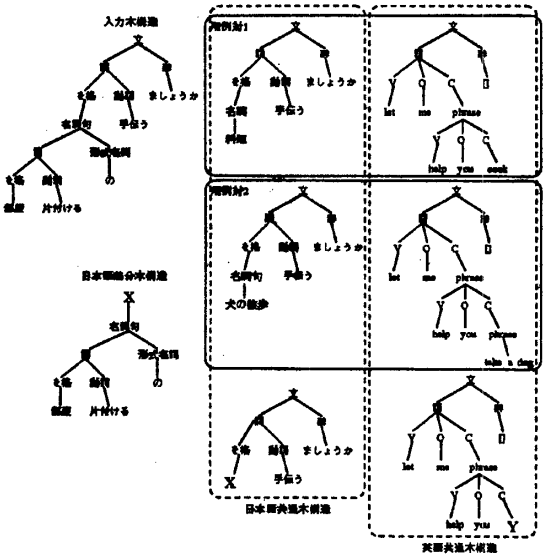

図 3: 二つの対訳用例による共通構造と差分構造

4 単文結合型英文生成

目的言語の生成は、各モジュールによって抽出された複数のポインタ付き共通木構造を継承情報を利用して、ポインタとポインタを結合し一つの木構造を生成する。

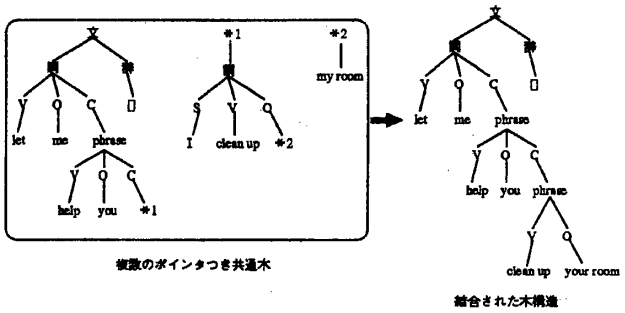

図 4: 複数のポインタ付き共通木構造

```
*1=[complement,infinitive,regular_varb]
*2=[noun,agent(you)]
```

ポインタとポインタをつなぐ時、木構造を変化させる必要がある。例えば、図4の2つ目の木構造は、ポインタルールに従い、原型不定詞にする必要があるし、また、主語も上位構造と一致しているため省略しなければならない。この変換はルールを用いる。

5 おわりに

本稿では、マルチコーパスを利用した多段用例翻訳システムの構築について論じた。コーパスを文法的レベルに分け、話題により分類することを提案した。各レベルのモジュールはシソーラスや類語動詞弁別ネットワークといった他の言語資源により、より精密で頑健性の高い類似用例の検索が可能になるだろう。今後は、このシステムのインプリメントを行ない、翻訳精度を評価する必要がある。しかし、このシステムの翻訳精度は、適切な用例があるかどうかにかなり依存してしまうためどのような方法でマルチコーパスを整備するかが問題として挙げられる。

参考文献

[1] 池田、宮崎 ： 用例翻訳における類似木構造生成法とその有効性、情報処理学第５２回全国大会,No.3-93(1996)