

4 B - 5

マルチコーパスを用いた用例翻訳における 単文結合型英文生成方式

富樫亮介 宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

用例翻訳において、一つの入力文に対して、一つの類似用例をコーパスから検索し、翻訳する方式では重文や複文といった、複数の動詞が含まれるような入力文を処理するには膨大なコーパスが必要である。本稿ではある程度の量のコーパスで重文や複文に対応するために、一つの入力文に対して、複数の対訳コーパスを参照し、この問題に対処する方法を提案する。対訳コーパスとしては文、単文、句、複合語など種々のレベルの用例をデータベース化したマルチ対訳コーパスを準備する。これらのマルチ対訳コーパスから入力文の日本語文に類似する様々な日本語類似表現素片を検索し、そこから類似表現素片とその差分である差分表現素片、その各々に対応する英語類似表現素片と差分表現素片を抽出し、それを組み合わせて入力文に対応する英語文を合成する方法を提案し、その可能性を示す。

2 マルチコーパスを用いた用例翻訳^[1]

本翻訳システムにおいては、文だけにとどまらず単文、句、複合語といった種々のレベルの対訳コーパスから複数の用例を適用する。コーパスは構文解析済みの対訳コーパスである。このような複数の対訳コーパスを有効に利用するため、それぞれのコーパスに対してそれぞれ固有の類似評価法をもつ複数のモジュールが存

在する。

入力文はまず形態素解析、構文解析を行ない木構造として、翻訳部に受け渡される。この木構造から各モジュールを呼び出し、日本語の類似表現素片と差分表現素片、および英語の類似表現素片を抽出する。そしてさらに日本語の差分表現素片をモジュールに受け渡す...という処理を繰り返し、最終的にいくつかの日本語の類似表現素片と差分表現素片、およびそれに対応する英語の類似表現素片と差分表現素片を得る。

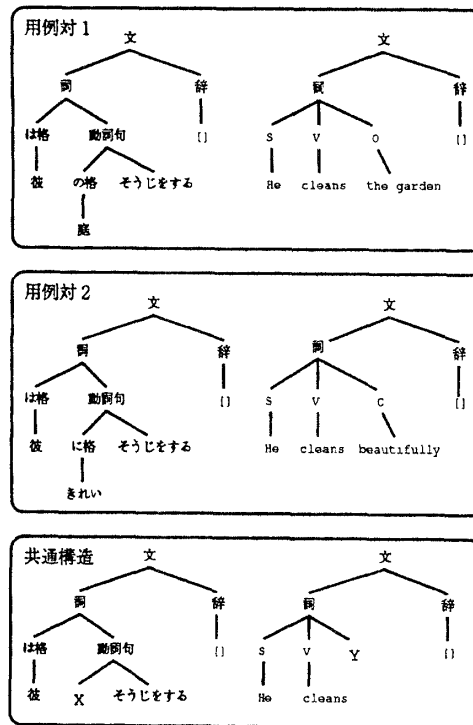


図 1: 二つの対訳用例の共通構造

ここで、英語の類似表現素片は二つの対訳用例を利用して抽出される。図1のように、二つの対訳用例の日英双方の木構造の共通構造と差分をとると、共通構造から日英それぞれの類

似表現素片が得られる。さらに図2のように日本語の共通構造と入力文の木構造の差分から日本語の差分表現素片がそれぞれ得られる。日本語、英語共に、類似表現素片と差分表現素片はポイントによってつながり、共通構造をつくるさいに、このポイントを求めることができる。

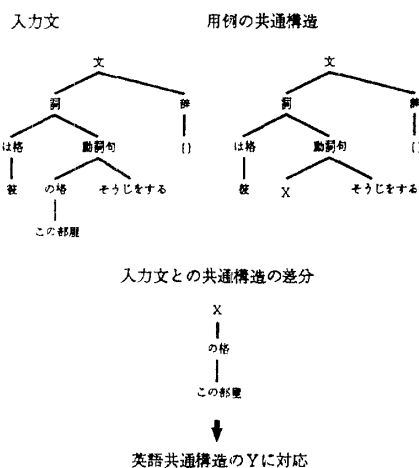


図2: 共通構造と入力文の差分と英文の対応

3 単文結合型英文生成

最終的に複数の、日本語と英語のそれぞれの類似表現素片と差分表現素片を得るわけであるが、ここから英文生成のための継承情報を生成する。

こうして得た、「継承情報」を用いて複数の英語類似表現素片の木構造のポイントとポイントを結合して、一つの木を生成する。

ポイントとポイントをつなぐ時、単純に結合すればいいというわけではなく、木構造を変化させる必要がある。この木構造の処理をする際に「継承情報」を利用する。この処理にはルールを用いる。

例えば、図3ではXのポイントに下位の構造に形式名詞「の」が存在し、上位の構造に動詞句に「が格」でつながっている、といった「継承情報」をのせ、ポイントYではただ単に上位が「を格」であるという情報をのせ、ポイントX'では動詞 like の目的語であるという情報をのせる。ここから、「英文を合成するさいに、動詞 like は目的語に to 不定詞をとり、ポイントX'以下を to 不定詞化する」といった処

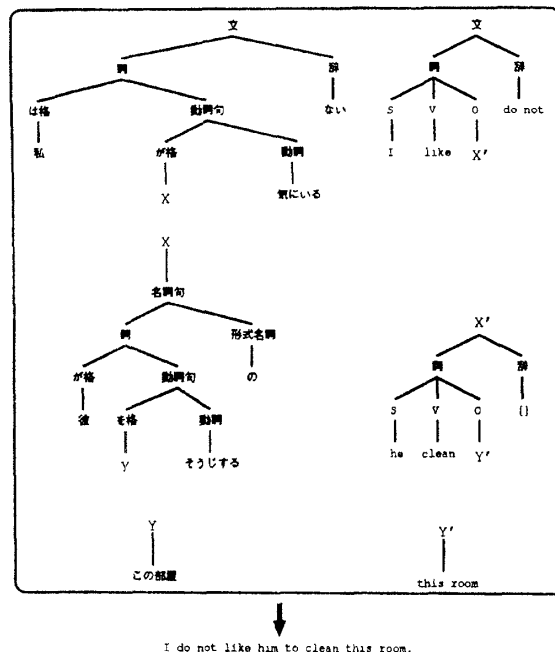


図3: 単文の結合

理をおこない、木構造を合成して英文を生成する。このように素片を結合するさいに特定の単語に固有の構造をつくることが多い。

4 おわりに

本稿では、ある程度の分量のコーパスで重文や複文に対応するためマルチ対訳コーパスから得られた、日本語、英語それぞれの類似表現素片と差分表現素片を組み合わせて英文を生成する単文結合型英文生成方式を提案した。この方式では翻訳精度が入力文と類似表現素片の類似度にかかなり依存してしまう。類似度が高ければ、構文解析の結果から比較的簡単な処理で翻訳精度のよい英文を生成できる。しかし、意味解析を行わない場合、接続詞をとまなうような英文等を生成しようとする精度が落ちる。今後は、こうした接続詞をとまなうような英文等に対する処理が課題としてあげられる。

参考文献

[1] 池田、宮崎：マルチコーパスを利用した多段階用例翻訳方式、情報処理学会第53回全国大会、4L-9(1996)