用例翻訳における類似木構造生成法とその有効性

6B－4

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、用例翻訳における基盤処理となる類似木構造生成法を提案し、その有効性について論じる。

用例翻訳では、コーパスの中からどのように入力文と最も類似する文を検索するかが大きな問題となっている。これについては、従来、数多くの研究がなされてきたが、単文を対象としてきたものが多かった。しかし、実際の翻訳では複文や重文などに対応しなくてはならない。しかし、複文や重文をコーパスの中にすべて用意しようとすると、必要とする用例が爆発的に増加してしまう。

このような問題を解決するために、本稿では、コーパスの中のすべての文に構文解析を施し、木構造の形で用意しておき、入力文が複文や重文であるとき、単文レベルの木構造を組み合わせることによって、入力文に最も近い構造を得る方法を提案し、その有効性を示す。

2　類似度評価法

構文解析は、三浦文法を元にDCG形式で記述されたSGLRパーザ[1]を利用する。三浦文法を用いる利点は、従来、類似度を評価する場合、格パターンを重視していたのに対し、助詞、助動詞、感動詞、接続詞、陳述副詞などの主体表現の入れ子構造を重視した類似度評価を可能とするところにある。

三浦文法によるSGLRパーザで「花が咲いた。」を構文解析した結果を図1に示す。この構文解析をコーパスの中のすべての文に施す。

類似度評価は、以下の手順で行なう。

1. 入力文を構文解析する。
2. 入力木構造とコーパス中のある一つの木構造の共通木構造をとる。
3. 共通木構造に評価点を与える。
4. 2と3の処理をコーパス中の全文に対し行ない、評価点の高いものを取り出す。

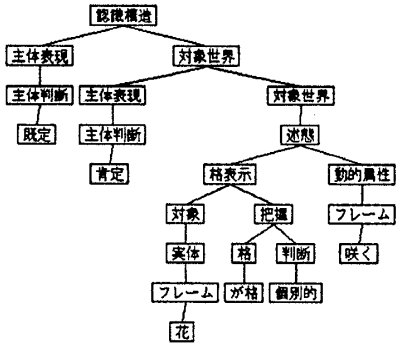

図 1:「花が咲いた。」の木構造

共通木構造はトップダウン横型探索で同一なノードを取り出す。類似度評価法は、木構造のすべてのノードを一律1点とし、その総和を評価点とし、入力木構造に対する共通木構造の割合を類似度とする。

例えば、「花が咲いた。」と「鳥が飛んだ。」（木構造は図2 a）とは、意味的には大きな違いがあるが、文の構造上は類似と言える。図1と図2 aの共通木構造を図2 bに示す。図2 bの共通木構造の評価点は、21点となる。また図1は評価点が23点であるため $21/23 \times 100$ で、類似度は91.3％となる。

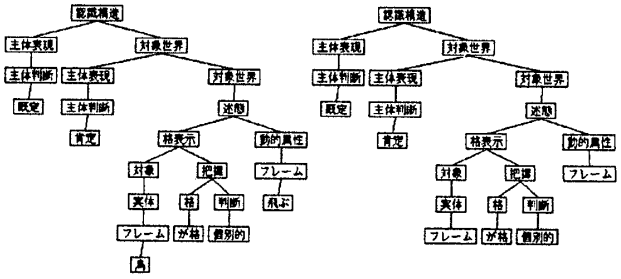

図 2: a.「鳥が飛ぶ。」の木構造　　b. 共通木構造

Generation of Similar Tree Structure in Example-based Translation and its Efficiency

Syuichi Ikeda , Masahiro Miyazaki

Niigata University

3 複文や重文に対する拡張

入力木構造が複文や重文で、類似度が高い木構造がコーパス中に存在しない場合、既存の木構造を組合せ入力木構造に近いもの、つまり類似度が高いものを生成する必要がある。そのための手順を以下に示す。

1. 共通木構造をとり、類似評価を行ないコーパス内から類似木構造を検索する。
2. 入力木構造にトップノード以外に「認識構造」のノードがある場合、入力木構造から「認識構造」以下の木構造を取り出し1へ戻る。
3. 検索した類似度が高い木構造の集合を組み合わせる。

入力文が「新潟で作られた酒が好きだ。」のような埋め込み文を含む場合、木構造は図3のようになる。これをトップダウンに探索し、例えば「犬が好きだ。」程度の単文の類似木構造しか得られなかったとしても、2の処理を行ない、例えば図4のような「この人形は日本で作られた。」を得られたとすれば、図5の様な共通木構造を生成することができる。この共通木構造の類似度は95.7％となる。

共通木構造を生成する上でもうひとつ重要な処理として、入力木構造に対しコーパス内から検索された木構造のほうが大きい、つまり枝の数が多い場合、枝刈りという処理を行なう。図4の「この人形は」の部分の構造が枝刈りされ、図5のようになる。

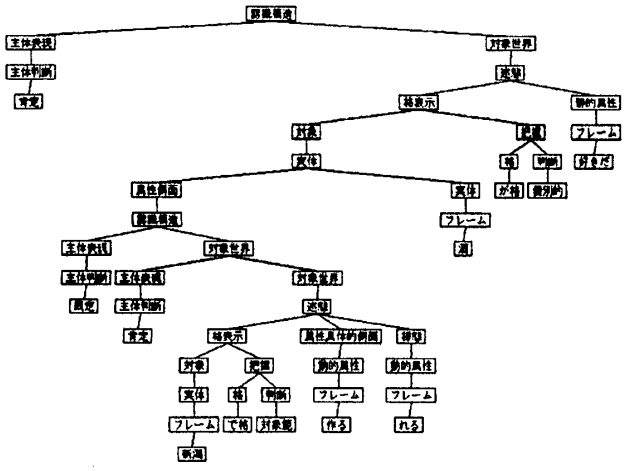

図 3: 「新潟で作られた酒が好きだ。」の木構造

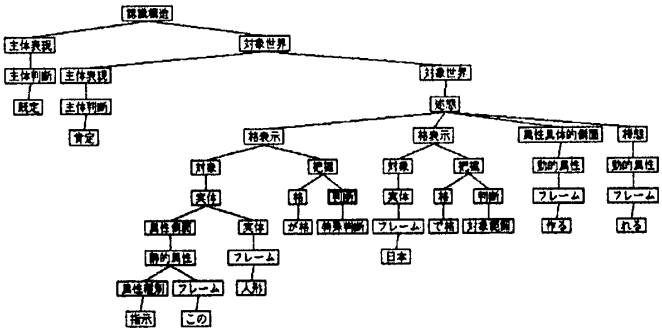

図 4: 「この人形は日本で作られた。」の木構造

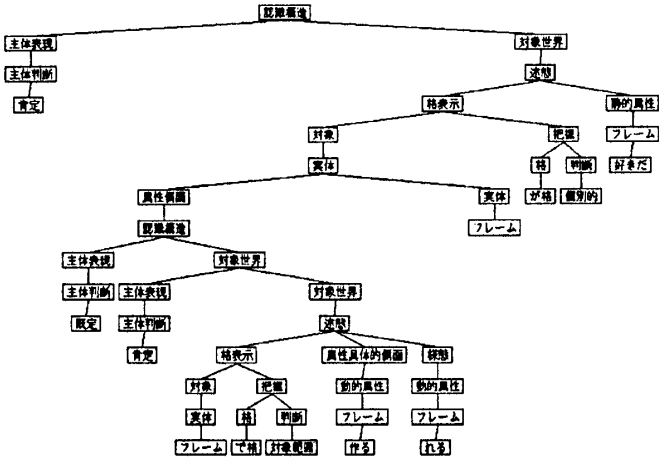

図 5: 生成された共通木構造

4 おわりに

本稿では、用例翻訳における基盤処理となる類似木構造生成法を提案し、その有効性について論じた。

今後、用例探索の高速化や文の構造だけでなくシソーラスなどを使った単語の意味の類似性を加えた類似評価を検討し、日英対訳コーパスを用いた用例翻訳システムに適用していく予定である。

参考文献

[1] 藤石、宮崎：三浦文法による日本語パーザの試作、情報処理学第51回全国大会,No.4H-9(1995)