3K-6

会話文翻訳システム
— 全体設計と英日翻訳文法の実装 —

山端 潔 安藤 真一 三村 清美
NEC C&C メディア研究所
e-mail: {yamabana, ando, mimura}@ccm.cl.nec.co.jp

1. はじめに

近年のメディア処理技術の進歩に伴い、言語処理の対象が広がりつつある。二者間の対話にあらわれる会話文は、音声言語処理技術の進歩とともに、処理対象としての重要性を急速に増しつつある。

我々は、このような会話文を対象とした機械翻訳システムを開発している。本稿では、そこで採用した文法形式を中心に、全体設計と英日翻訳文法・辞書の実装について述べる。

2. 会話文の処理と文法形式

会話文には、書き言葉の文法を逸脱した会話特有の表現や、特定の単語の組み合わせに基づく熟語的表現など、多様な表現が現れる。そのため、翻訳においては、汎用的な一般文法による処理に加えて、語彙に依存した個別処理を詳細に記述することが重要である。

このために、構成素境界を手がかりに表層レベルのパターン文法を記述する方法[4]や、語彙化の技法[7]を応用して、個別語彙に特有の規則をパターン文法として記述する方法[6]などが提案されている。ところが、精度向上のためには、パターン一つだけでなく、複数のパターンの組み合わせ方やその適用条件等、規則間の制御を個別に記述したい場合が少なくない。従来の語彙化文法の枠組みでは、規則の組み合わせ方をあらわすツリー演算はグローバルに定義されているため、このような目的には、ノードの素性構造を通じてその動作を間接的に制御するしかなかった。

次節では、汎用の一般文法と語彙やドメインに特有の個別処理を統一的に記述するのに適した新しい文法記述の枠組みとして、ツリーおよびツリー演算の一部を語彙化して、規則の精密な制御を可能にした語彙化ツリーオートマトン文法を導入する。

3. 語彙化ツリーオートマトン文法

3.1. 句構造文法と語彙化

句構造文法は、一般に、非終端記号と終端記号(単語)をノードに持つ有限のツリーの集合と、ツリーを組み合わせて別のツリーを構築するツリー演算の組として定義される。例えば、文脈自由文法は、高さ1のツリーの集合と、非終端記号の一致によるツリーの連結演算により定義される文法である。また、語彙化文法は、各ツリーに単語が関連付けられている文法として定義される。

3.2. 語彙化ツリーオートマトン文法の定義

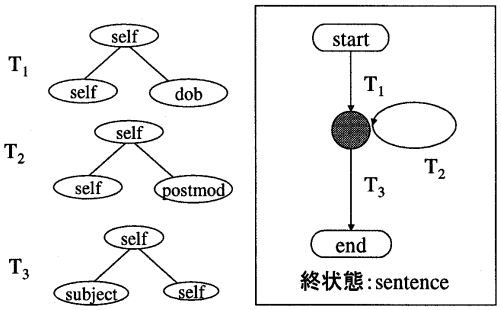
語彙化ツリーオートマトン(Lexicalized Tree Automata, LTA)文法とは、各単語に対し、その単語をヘッドワードとする要素ツリーの集合と、要素ツリーを組み合わせて構成されるツリーのうち、文法が許容するツリーのみを受理するツリーオートマトンの二つが付随する文法形式である。受理されたツリーの集合は、その単語をヘッドとしてどのようなツリーが成長可能かをあらわす。終状態に達すると、ルートに非終端記号が与えられる。一方、各単語から成長したツリー同士の演算は、非終端記号の一致による通常の連結演算とする。

すなわち、LTA 文法とは、ツリー演算を、ある単語をヘッドとするツリーを形成する部分(ローカル文法)と、それらのツリーを連結する部分(グローバル文法)に分解し、前者を単語に付随するツリーオートマトンとして表現した文法形式である。

3.3. 例

図 1 は、動詞 eat の持つツリー集合とツリーオートマトンの一例である。eat からは、自身をヘッドとして、図2に示すツリーが成長するものとする。eat に付随するツリー集合(図 1(a))は、直接目的語を取り込むツリーT_1、副詞等の自由修飾要素を取り込むツリーT_2、および主語を取り込むツリーT_3を有する。ルートおよびリーフの一箇所がself とマークされているのは、このノードを重ね合わせながらオートマトンによるツリーの受理が進むことをあらわす。

一般に、ある単語をヘッドとするツリーは、ルートからその単語に至る背骨(spine)に沿って並べることにより、要素ツリーをアルファベットとする文字列と同一視することができる。この同一視により、語彙化ツリーオートマトンは、要素ツリーをアルファベットとするストリングオートマト



(a) ツリー集合　　(b) ツリーオートマトン

図1:eatの辞書内容

A Dialogue Translation System: Design and Implementation of English to Japanese Translation Module.
Kiyoshi YAMABANA, Shinichi ANDO and Kiyomi MIMURA
C&C Media Research Laboratories, NEC Corporation

ンと同一視することができる。図 1(b)は eat に付随するツリーオートマトンを後者の形式で表現したものであり、図2のツリー、すなわち要素ツリーの列 $T_1 \cdot T_2^* \cdot T_3$ を受理するように構成されている。

3.4. 構文解析アルゴリズム

LTA 文法に対し、ボトムアップチャート法をベースとした構文解析アルゴリズムが定義できる。CFG の場合、アクティブエッジは、ルール右辺の受理がどこまで進んだかを示すドットつきルールにより表現されるが、LTA 文法では、アクティブエッジは、ローカルツリーの受理がどこまで進んだかを示すツリーオートマトンの状態として表現することにより、CFG と同様の解析アルゴリズムが適用できる。エッジのパックの可否は、オートマトンの未適用部分および非終端記号の一致で判定する。

4. 会話文翻訳システム

以上の枠組みに基づき、PC の Windows 環境で動作する英語・日本語双方向の会話文翻訳システムを構築した。以下、エンジンの設計の概要と英日翻訳文法・辞書について述べる。

4.1. 翻訳エンジン

エンジンの実装にあたっては、単語辞書内に分散された文法を効率的に共有する仕組みが重要である。そのために、ルールテンプレート機構と共通ルール機構を設けた。ルールテンプレートは、ツリー集合とツリーオートマトンを単語間で共有する仕組みであり、初期チャート作成直後に実体がロードされる。共通ルールは、個々のツリーをツリー集合間で共有する仕組みである。共有されたツリーはポインタとして表現され、参照時に実体がロードされる。

言語変換には同期導出[8]をベースとする方式を採用した。ただし、変換後のツリーを、構文木自体ではなく、構文木を生成する手続きの呼び出し関係の表現とすることにより、生成過程の自在な制御を可能としている。生成についてはさらに[2, 5]も参照されたい。

4.2. 英日翻訳文法と辞書

英語文法は、標準的なXバー理論に準拠した句構造を採用している。LTA 文法では、異なる単語をヘッドに持つツリー間の連結には、従来と同様、品詞(素性構造)のマッチングを用いているが、この際に語彙的な統語的能力の違いを細かく表現する手段として、Link Grammar[9]と同様、品詞名ではなく文法関係名を中心に記述するアプローチをとった。英日辞書は見出し語数約 7 万である。そのうち、個別のツリーセットとツリーオートマトンの記述が必要となった単語は数千であった。

5. 考察

LTA 文法は語彙化文法の一形式であり、文法の実効的な大きさが、入力文中の単語に関連付けられた部分に縮小されるなどの語彙化文法の利点[7]を共有する。

文法の能力としては、ツリーオートマトンのクラスを(ツリー列に対するストリングオートマトンとしての表現で)正規オートマトンに限ると、その生成能力は文脈自由文法と等価である。プッシュダウンオートマトンを使うと、LTAG と弱同値な文法を構成することができる。ツリーオートマトンは、Extended Domain Of Locality (EDOL)を表現するローカル文法の有限な表現であり、LTAG 等と異なり、EDOL の非有界性の表現のために adjoining 等の特殊なツリー演算を用いる必要がないのも利点である。

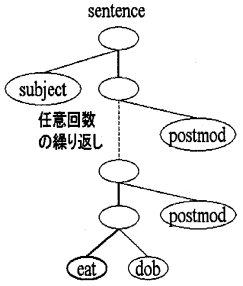


図2:eatをヘッドとするツリー

本稿の手法には、表層パタン[4]や語彙化 CFG[6]に基づく手法に比べて、抽象度の高い一般文法規則から用例に近いパタン規則までを統一的な枠組みで記述でき、さらに規則の適用を単語ごとに精密に制御できるという特徴がある。この特徴は、実際の文法記述においても訳質向上に大きく寄与している。文法としてオートマトンを記述するのは一見複雑だが、実質は単語をヘッドとするツリーを定義する従来の文法記述の作業と同等であり、かえって、ツリーがうまく成長するように品詞や素性構造を操作する必要がないぶん、見通しの良い文法記述が可能となっている。

本稿の手法と、オートマトンベースの解析の手法として従来提案されている手法[1,3]の相違は、従来の手法のオートマトンが、非終端記号列を受理する有限オートマトンであるのに対し、LTA 文法ではツリーを受理するツリーオートマトンである点にある。また、従来手法には、単語に関連付けられたツリーの作るローカル文法の概念がなく、例えば[3]では、オートマトン化は構文解析の高速化手法として位置付けられている点でも異なる。

6. まとめ

語彙化文法の一形式として、語彙化ツリーオートマトン文法による文法記述の枠組みを提案した。さらに、この枠組みを用いて会話文翻訳システムを構築した。

参考文献

[1] H.Alshawi: ``Head Automata and Bilingual Tiling.'' ACL'96, pp.167-176 (1996).

[2] 安藤 他:「会話文翻訳システム —生成処理と生成における事例の利用—」情処第 60 回全国大会 2K-4 (2000).

[3] R.Evans et al.: ``A structure-sharing parser for lexicalized grammars.'' ACL'98, pp.372-378 (1998).

[4] 古瀬 他:「構成素境界解析を用いた多言語話し言葉翻訳」自然言語処理, 6 (5), pp.63-91 (1999).

[5] 三村 他:「会話文翻訳システム —英日処理における省略を用いた対話文生成について—」情処第 60 回全国大会 3K-7 (2000).

[6] 長瀬 他:「ルールベース翻訳とパターンベース翻訳の融合」言語処理学会第 4 回年次大会, pp.496-499. (1998).

[7] Y.Schabes et al.: ``Parsing strategies with `lexicalized' grammars.'' COLING'88, pp.578-583 (1988).

[8] S.Shieber et al.: ``Synchronous Tree Adjoining Grammars.'' COLING'90, pp.253--258. (1990).

[9] D.Sleator. et al. ``Parsing English with a Link Grammar''. CMU TR CMU-CS-91-196 (1991).