英日機械翻訳システムShalt2における並列句の取り扱い

5P－2

浦本直彦・武田浩一・那須川哲哉・荻野紫穂・堤泰治郎
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

本論文では、現在開発中の英日機械翻訳システムShalt2[1]における並列句の処理、特に多義性の表現および解消について述べる。本論文で扱う並列句の多義性は、次の2種である。

1. 並列要素の同定に関する多義性
2. 修飾句の並列句への係り方に関する多義性

例えば"A of B and C on D"のような並列構造の時、1.は並列要素(接続詞によって等位関係にある単語)に関してAとC、BとCの複数の解釈がある場合をいう。2.は、Dの係り先がC、並列要素両方およびAの可能性がある場合である(これは、前置詞句の係り受けに関する構造的多義性とみなすこともできる)。また、並列要素の品詞は名詞に限らないものとする。

2. 多義性解消の統合的なアプローチ

並列句の処理に関して、我々が考慮したのは、次の2点である。

(2-1) **並列句に関する多義性は解消しなければならない。**

Shalt2は、中間(意味表現)言語を用いる機械翻訳システムであり、入力文は、フレームをベースにした意味表現に変換される。この意味表現は多義性を許さないので、入力文が意味表現に写像される時点で、文の解釈は一意に決っていなければならない。よって、意味表現の変換を行なって、こなれた訳文を生成するといった高度な処理を実現するには並列句に関する多義性を含め、全ての多義性は解消されなければならない。

(2-2) **他の種類の多義性の解消と同じ枠組で、並列句の多義性も処理されなければならない。**

並列句を扱う従来の研究では、並列句に対し、係り受けや語義の多義性等の他種の多義性解消処理とかけ離れた処理を行うことが多かった。しかし、これらの多義性は、全て統一的に扱えると非常に都合が良い。なぜなら、これらの多義性は、互いに関係を持っており、一つの多義性を解消することで、他の多義性も同時に解消できるといった利点があるからである。例えば、次の文において、

(S1)Use the Backspace key to move the cursor to the left and the spacebar to erase.

(S2)The scrolling indicator appears after a line of dashes and the word "More".

(S1)では、"to the left...", "to erase"の係り先に関して多義性があるが、正しく並列構造が認識されれば、これらの多義性は解消できる。また、(S2)では、lineとwordが並列要素であることが分かれば、lineの訳は"行"の方がもっともらしいといった解釈が可能である。我々は既に係り受けと語義の多義性を協調的に解消する手法[2]を提案しており、並列句も同じ枠組で取り扱う。つまり、

- 解析木は依存構造で表現し、複数の多義性は、その構造上でパックする。
- 多義性をパックした依存構造に対し、制約と、事例ベースから計算される優先度を用いることで、多義性を解消する。

3. 多義性のパック表現

自然言語文中には勿論、様々な種類の多義性が複数存在する。ここで、各々の解釈に一つの解析木(本論文では、依存構造を用いている)を作ると、計算量の問題はいうまでもなく、多義性がどこにあるのかといった情報も表現しにくいし、多義性解消のための様々な知識を効率良く適用することも困難である。また、多義性をパックした形で解析木上に保持しておくことで、多義性解消に必要な知識が揃う時まで、最終的な解釈を遅延できるといった利点もある。これらの理由から、一つの、あるいは少数の依存構造に多義性がパックされていることが必要である。本章では、並列句の多義性を含めた多義性がどのようにパックされているかについて述べる。

4. 多義性解消のメカニズム

1章であげたように"A of B and C on D'という構造では、並列要素に関して、"A of [B and C]"と"[A of B] and C"の2通りの解釈がある(図1の(a), (b))。また、前置詞句Dの係り先についても、複数の解釈があ

Handling of Coordinated Structure on E-to-J Machine Translation System Shalt2
Naohiko URAMOTO, Koichi TAKEDA, Tetsuya NASUKAWA, Shiho OGINO, and Taijiro TSUTSUMI
IBM Research, Tokyo Research Laboratory

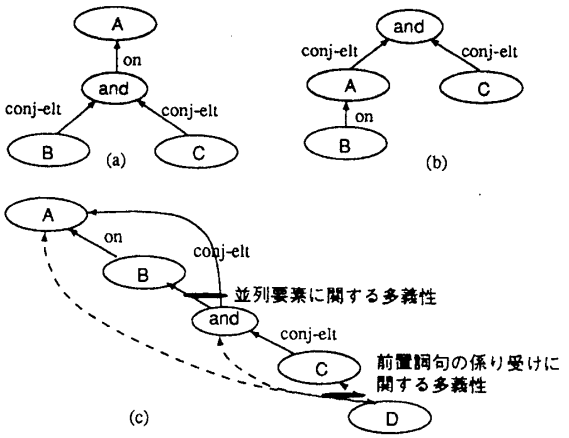

図 1: 多義性がパックされた依存構造

る (これは構造的多義性とみなすこともできる)。これら2種の多義性を表現したのが、図1の (c) である。一つのノードから複数のアークが出ている箇所が多義性を示し、各々のアークが一つの解釈を示している。語義の多義性も同じように表現でき [2]、種類の異なる多義性を一つの構造で表現することが可能である。ただし、他の多義性と異なり、並列要素に関する多義性に関しては、最終的に選択された解釈に応じて、(c) の依存構造を、(a) あるいは (b) の構造に変換しなければならない。この変換操作は、簡単にいうと、(a) の時は and ノードと B ノードを、(b) の時は、and ノードと C ノードの位置を入れ換えることで行なう。

Shalt2 における多義性解消は次のように行なわれる。まず、それぞれの多義性について、制約を適用することで解釈の候補を絞る。各多義性の間には制約マトリックスを使って依存関係が記述されており、一つの多義性の解消の結果を他の多義性の解消のために伝播させることができる。残った解釈については、事例ベースを用いて優先度を計算することにより、最終的に文全体の解釈を決定する。前章の例であれば、図1の (a) と (b) の2通りの解釈について優先度を計算することになる。計算された優先度は、各多義性毎に保持され、全体的な文の解釈は、全ての多義性における解釈が矛盾しないように制約マトリックスをチェックしながら行なわれる。

並列構造の推定や多義性解消に関する研究は、これまでにいくつかなされている [3, 4, 5]。その中で共通に使われている手法は、構造的および意味的バランスを優先するものである。構造的バランスとは、(4-1) それぞれの並列要素の部分木構造が類似している。例えば、並列要素が同じ種類 (あるいは数) の単語から修飾されている。(4-2) 並列要素の文法的属性 (単／復、加算／非加算、冠詞の有無 (以上名詞)、自動詞／他動詞等) が類似している、ことをいう。また、意味的バランスは、並列要素のシソーラス上での距離や意味マーカ・格フレームの一致度を用いて計算されることが多い。本論文でも、基本的には、この手法を用いて優先度を計算する。構造的バランスに関しては、本論文では (4-1) だけを用いている。(4-2) については、Shalt2 が現在ターゲットとしている計算機マニュアルでは、でこれらの属性が一致しない場合が多いからである。意味的バランスでは次章で説明するように、シソーラスによる類似度に加え、事例ベースを用いている。

5. 事例ベースを用いた多義性解消

筆者らが現在使用している事例ベースは、多義性が解消された依存構造で表現された８０００文 (一般辞書例文４０００文、計算機用語辞典定義文４０００文が含まれる) から構成されている。それに加えて、一般の類語辞典のデータや一般辞書から抽出した上位下位関係を用いている。根本らの指摘 [4] にもあるように、並列要素間の関係は多岐にわたっており並列要素を同定するのに、シソーラスを用いた意味的な類似性をだけを用いるだけでは不十分である。例えば、"document" という単語について考えると、シソーラス上で類似性の高い単語は"certificate", "instrument", "papar", "record", "report" 等であるが、事例ベースから検索された document と並列している単語は、"folder", "page", "note", "information", "comment", "mail", "string", "file","copy","code", "function" 等であり同義関係以外のものも多い。従来のシソーラスに加え、これらの事例を検索することで、並列要素間の関係をより正確にとらえることが可能である。

6. おわりに

Shalt2 における並列句の取り扱いについて述べた。現在、Shalt2 上で、実際の文を使って本手法の評価を行ない、優先度計算の詳細を検討中である。

参考文献

[1] K. Takeda, N. Uramoto, T. Nasukawa, and T. Tsutsumi. "Shalt2 - A Symmetric Machine Translation System with Conceptual Transfer". In *Proceedings of the COLING-92*, 1992.

[2] 浦本. "制約と事例による優先度を組み合わせた英文の多義性の解消". 自然言語処理研究報告 (NL-90-9), pp. 65–72, 1992.

[3] 黒橋, 長尾. "長い日本語文における並列構造の推定". 自然言語処理研究報告 (91-NL-86), 1991.

[4] 根本, 藤田. "名詞句並列表現の解析". 情報処理学会第 36 回全国大会, pp. 1155–1156, 1988.

[5] 田添, 相馬, 河合, 椎野. "英日機械翻訳における名詞句並列を含む長文解析について". 第７回人工知能学会全国大会, pp. 479–482, 1993.