英日機械翻訳システムShalt2の日本語生成文法

6Q－2

荻野　紫穂　　武田　浩一
日本アイ・ビー・エム株式会社　東京基礎研究所

1. はじめに

現在開発中の英日機械翻訳システムShalt2は、日本語生成部に、疑似単一化文法による解析／生成双方向文法を用いている。本稿では、この文法の生成用文法としての面を取り上げ、並列句処理などにおけるセットの処理のための記述拡張、生成順序制御に使用する情報とその処理、及び、長距離依存の処理について述べる。

2. 文法の枠組と生成時の素性構造処理

まず、第3.章以下の説明に使用する、疑似単一化文法の等式と、生成の際の素性構造処理について、簡単に説明する。疑似単一化文法については、詳しくは[5]などを参照されたい。

疑似単一化文法はの句構造規則は、

$r1$: $X_0 \rightarrow X_1 X_2 \ldots X_m (X_i \in V_N)$

$r2$: $X_0 \rightarrow a_1 a_2 \ldots a_n (X_i \in V_N, a_i \in V_T)$

という句構造の記述と、単一化に関する0個以上の等式$eq_1, \ldots, eq_k$の集合からなる（V_Nは非終端記号の集合、V_Tは終端記号の集合）。ここで、各eqは、

$e1$: $\langle x_i L_1 \rangle = \langle x_j L_2 \rangle$

$e2$: $\langle x_i L_1 \rangle = a$

$e3$: $\langle x_i L_1 \rangle = c\ a$

$e4$: $\langle x_i L_1 \rangle = *defined*$

$e5$: $\langle x_i L_1 \rangle = *undefined*$

という形をしている（aは定数）。

$e1$は、$x_i L_1$と$x_j L_2$との単一化の結果で双方を置き換える。$e2$、$e3$は、素性が定数か、空でない素性構造の場合に成功するが、単一化の後、$e3$で参照された素性は定数となる。$e5$は、素性が空または存在しない場合に成功する。

生成においては、素性構造を句構造に分解し、単語導出に関係する各素性を、上の等式で参照し単一化することによって《外す》、つまり、空の素性にする作業を繰り返す。素性構造を句構造に分解する際には、原則として、各構文要素について、構文的・形態素的にそのヘッドから遠い、《外側の》句構成要素に対応する素性から外してゆき、最後にヘッドとなる素性構造が残るようにする。これは、いわゆる文節間の修飾関係を表す規則が適用される場合でも、文節内の形態素の接続関係（例えば付属語の接続）を表す規則であっても同じである。

全ての単一化が成功して終端記号、或いは、前終端記号に至った時は、基本的に、単語導出に関係する素性が全て参照され、レキシコンに書き込まれた基幹情報が定数として残ることになる。

3. セットの扱いと修飾句のリスト

名詞句や動詞句などの並列構造は、並列句を構成する要素のどれかをヘッドに定めるのではなく、セットが1つの構文要素として、その外側の構文要素のヘッドとなっているものとする[6]。セットは、以下のように記述される。

```
(*set*
 (セット共通の素性構造)
 (セット・メンバーの素性構造)
    ......
 (セット・メンバーの素性構造)  )
```

最初の*set*はこの構文要素がセットであることを示す。セット共通の素性構造には、セット自体が持つ素性が記述されるが、実際には、接続詞など、セットのメンバー同士の関係を示す素性が含まれており、品詞や活用型・接続型などの素性は、セット共通の素性構造には含まれない。セット共通の素性構造に対する等式は、

$e1'$: $\langle x_i L_1 \rangle = s \langle x_j L_2 \rangle$

$e2'$: $\langle x_i L_1 \rangle = s\ a$

という形をしている。これらは$e1$、$e2$に対応する等式で、x_iをセットと見做し、そのセット共通の素性構造のパスL_1と右辺の値を単一化する。

ヘッドの品詞や活用型・接続型などの素性の値を単一化する必要がある場合、セットの最後のメンバーの品詞や活用型・接続型が、セット全体のそれの代わりに参照される。日本語において、セットがヘッドである構文要素の活用型・接続型は、ほとんどの場合、セットの最後のメンバーのそれが受け継がれる。また、ある1つのセットは、ほぼ同じ構文的性質をもった複数のメンバーで構成されるため、レキシコンそのものの情報を参照する時以外は、最後の要素の素性をセットの素性として参照しても問題はない。

Grammar for Japanese Sentence Generation of E-to-J Machine Translation System Shalt2
Shiho OGINO, Koichi TAKEDA
IBM Research, Tokyo Research Laboratory

第2.章の5種類の等式に対応して、セットの最後のメンバーである素性構造に対する等式は、

$e1''$: $\langle x_i L_1 \rangle = l \langle x_j L_2 \rangle$

$e2''$: $\langle x_i L_1 \rangle = l\ a$

$e3''$: $\langle x_i L_1 \rangle = lc\ a$

$e4''$: $\langle x_i L_1 \rangle = l\ *defined*$

$e5''$: $\langle x_i L_1 \rangle = l\ *undefined*$

という形をしている。セットの最後のメンバーの素性を参照すること以外、働きは同じである。

これらに加え、

$e6$: $x_i = *set*$

$e7$: $x_i = *plain*$

という等式を便宜的に導入する。$e6$ は x_iがセットである時に成功し、$e7$ は x_iがセットでない時に成功する。

連体修飾などの修飾要素では、1つのヘッドについて同じ役割を持つ修飾句が複数存在することがある。この場合、複数の修飾要素は、以下のように1つにまとめられ、リストとして扱われる。

```
(*multiple*
 (修飾要素の素性構造)
    ......
 (修飾要素の素性構造)  )
```

リストはセットとは違い、共通の素性構造を持たない。また、このリストはそれぞれのメンバーが同じ構文的性質を持つとは限らない順序リストであるため、最後の要素の素性をリスト共通の素性の代用とすることはできない。つまり、リストのメンバーは、個々別々に処理する必要があるため、リストの素性構造に対する等式は、

$e1'''$: $\langle x_i L_1 \rangle > \langle x_j L_2 \rangle$

となる。この等式は、$x_i L_1$をリストの可能性がある素性構造と見做し、その最初の要素と $x_j L_2$とを単一化する。このリストは順序リストなので、単一化は、必ずリストの最初の要素と行なわれる。この等式が成功すると、$\langle x_i L_1 \rangle$ は、最初の要素が取り除かれたリストと置き換えられる。

4. レキシコン情報と文法による生成制御

"know"/"知る・知っている" のように、日本語において、英語側に特別な情報がなくても、日本語側でテイル形をとる動詞がある。こういった、日本語側の各語に依存した情報は、日本語側のレキシコン情報として記述する。このテイルは、動詞そのものの直後に密着して生成しなければならない。また、このテイルが生成される場合は、現在進行形のテイルが生成されてはいけない。

このテイルの生成に限らず、素性構造の中に、付属語に対応する素性が複数含まれている場合、付属語同士の間で、その生成順序や共起関係を制御しなければならない。第2.章でも述べた通り、生成の際には《外側の》形態素、つまり、ヘッドより遠い形態素から順に、素性構造から《外されて》いく。よって、ある形態素生成に対応する規則が適応される際には、素性構造中に、ヘッドからその形態素よりも遠い位置にあるべき形態素に対応する素性が残されていてはいけない。特に、いわゆる同一文節内での形態素生成制御については、文法中で、その規則から生成される形態素よりも《外側で》生成されるべき形態素、或いは、その形態素と共起してはいけない形態素に対応する素性のチェックを行なうことによって、形態素の生成順序や組み合わせを制御する。

5. 長距離依存等の扱い

副詞の中には、文頭に置かれると、"first"/"第一に" のように特別な役割（文副詞）を担ったり、"once"/"一旦 (〜すると・すれば)"(cf. "一回〜する") のように、呼応表現を導く働きをするものがある。これらの副詞は、位置的条件などを満たすかどうかによって、呼応表現を導くものと導かないものとにマッピングが分かれる。呼応表現を導くものに関しては、レキシコンに導かれる表現を記述しておく。文法では、そのレキシコンの情報を参照して、呼応表現を生成する。

"誰が〜するか"、"どこの〜も" のように、不定詞とカ・モなどの呼応については、カ・モが《外される》時点では、不定詞がそれ以降に解析される部分木のどこに現われるかが分からない。従って、カ・モが外されたヘッドの素性構造に特別な素性を与えておき、不定詞が部分木に現われた時点で、その素性を《外す》、つまり空にする手法をとる。ヘッドが終端記号まで至るまでの間に、その素性が空になっていれば、この呼応を含む規則における単一化が成功する。

6. まとめ

英日機械翻訳システム Shalt2 の日本語生成部における疑似単一化文法の枠組と、長距離依存などの生成制御について述べた。自然な語順の生成や読点の出力などが、今後の課題として残されている。

文献

[1] 尾上圭介 (1983) 不定語の語性と用法, 「副用語の研究」明治書院, pp. 404 - 431.

[2] 南不二男 (1986) 質問文の構造, 「朝倉日本語新講座 4 文法と意味 II」朝倉書店, pp. 39 - 74.

[3] 池田光生, 幡野浩司, 福島秀顕, 重永信一 (1988) 汎用日本語処理系 (LTB) 文生成部の生成方式, 「日本ソフトウェア科学会第 5 会大会論文集」, pp. 29 - 32.

[4] 徳永健伸, 乾健太郎 (1991) 1980 年代の自然言語生成, 「人工知能学会誌」, vol. 6, no. 3 - 5.

[5] 武田浩一 (1991) 疑似単一化文法の双方向性, 「情報処理学会第 42 会全国大会講演論文集」(3), pp. 116 - 117.

[6] 浦本直彦 (1993) 英日機械翻訳システム Shalt2 における並列句の取り扱い, 「情報処理学会第 47 会全国大会講演論文集」(3), pp. 177 - 178.