

日本語句構造文法を用いた
文生成アルゴリズム

7F-6

石崎 雅人

NTT 情報通信処理研究所

1. はじめに

文生成の問題は「何を言うか (what to say)」、 「どのように言うか (how to say)」、及び処理のモデルの問題に分類できる。第1番目は中間表現をいかにつくるかという問題、2番目は中間表現と表層表現をいかにマッピングするかという問題、3番目は内容と言い方の決定を逐次的に行なうか互いに影響させながら文を生成するかという問題である。ここでは句構造文法を用いた中間表現と表層表現とのマッピングの問題について考察する。議論をわかりやすくするために中間表現は何らかの方法により与えられるとし、処理は逐次的に進むものとする。

Shieber[1]では句構造文法を用いて文解析・文生成を同一のアーキテクチャで実現する方法が述べられている。文生成に注目してみると句構造文法からの単純なアルゴリズムの効率の悪さを改良するために意味の単調性 (semantic monotonicity) を用いている。しかしながらこのアルゴリズムには以下の問題がある。

- (1) 意味情報を十分に利用していない
- (2) 意図した文を出力するための生成過程の制御ができない

本論文では句構造文法を用いた生成において意味情報を用いた効率を上げるための手法と機能要素の導入による文を絞り込むための手法について述べる。

2章では文生成のための中間表現について、3章では効率改善のための技法、4章では生成する文の絞り込みの手法、5章ではアルゴリズムと例題について6章ではまとめと今後の課題について述べる。

2. 中間表現

中間表現を命題内容を表現する意味情報とそれ以外のものを表現する談話情報の両方で定義する。

意味情報を表現するための意味表現言語はHobbs[3]で提案されている言語を利用する。この言語の特徴は談話処理のために非常に単純な形式を仮定したことである。具体的には(1)構造は基本述語の接続で表現する(2)すべての変数は可能な限り最大のスコープで存在限量子化されていると考える(3)様相オペレータ、内包オペレータ、ネストされた限量子は利用しないという特徴を持つ。例1には少年xがボールyを蹴るという事象eを表現する中間表現を示す。

例1 少年がボールを蹴る $kick(e, x, y) \wedge boy(x) \wedge ball(y)$
久野[6]では談話情報として情報の新旧、視点が挙げられているが、何を談話情報とするかについては統一された見解があるわけではない。ここでは意味情報とは別に情報の新旧、視点、主題を談話情報として指定する。例2ではくによってboyが最も情報が古くkickが最も情報が新しいこと、視点がxにあること、主題は指定されていない

ことを示している。

例2 [boy < ball < kick; x; --]

談話情報の内容及び表現方法にはいくつか問題があるが、これについての問題点については4章で触れる。

3. 素性によって規則を駆動する文生成

3.1. 効率の改善

句構造文法を用いて文を生成する場合に最も単純なアルゴリズムは生成検査法を用いることである。これは句構造規則を非決定的に探索して文を生成し、そこで作られた意味情報を入力の意味情報と比較することで正しい文が生成されたかどうかを検査する方法である。生成検査法を用いた方法は初期の段階で文にならないとわかっていてもそれを排除することができないので効率が悪い。効率の改善のためShieber[1]では部分構造の意味情報は全体構造の意味構造に含まれるという意味の単調性の仮説を用いて無駄な句構造規則の適用を避けている(これによる最も分かりやすい効果は関係ない語彙の導入を避けることができることである)。

日本語句構造文法[4]のような単一化文法を用いて入力の部分意味情報に関連する素性の束を組み合わせていくことにより文を生成すると、部分構造の持つ素性の検査によって無駄な構造を作らないで済むようになる。例えば部分構造から別の構造を作る場合に意味情報を利用することを考えてみる。意味情報は部分意味情報を関係付けるものと関係付けられるものに分類され、意味規則の役割は前者と後者を関係付けることにある[7]。すなわち意味情報を考慮に入れることにより無駄な部分構造の組合せを排除することができる(部分構造の数をNとすると組合せの検査の回数は $(N!)^2/N$ から N^2 になる)。

3.2. 問題点

日本語の場合、入力の部分意味情報からの文生成では意味情報に寄与するかどうかによって語彙を実質的意味のあるもの(実質要素)と実質的意味のないもの(機能要素)に区別して扱う必要がある。実質要素とは通常の動詞のように語彙レベルから文レベルまで意味情報が保存されるようなものであり、これは意味の単調性の仮説を満足する。機能要素とは格助詞のように意味情報には直接は表現されないものであり、実質要素のように意味の単調性の仮説によっては意味情報に関係のない無駄な語彙の導入は排除することはできない。3.1で触れた文生成処理ではこの機能要素のように意味情報には影響を与えないものに関しては導入が任意となり、効率と多様な文の出力を保証するという観点からは望ましくない。以下では機能要素の導入に関する制約について考える。

機能要素は助詞のように表層に現れる要素と空所のように現れない要素に分類することができる。表層に現れる要素は統語情報、意味情報や談話情報から導入する必要性を判断できる。例えば用言とその格要素の関係を表現するために用いられる格助詞は用言の持つ下位範疇化情報から助詞句(名詞+助詞)を要求するので、用言と名詞との関係からどの格助詞を選択すれば良いかを判断することができる。また視点を変更する機能を持つ直接受身の助動詞は談話情報から導入すべきかどうかを判断

することができる。表層に現れない要素は現れる要素と違って導入の必然性を判断することができない。従ってまず空所を導入しないで生成を行ない、もし文を生成することができなかつた場合には空所を導入して生成し直すという方法を取る必要がある。

4. 多様な文の生成について

同じことを言うのにもいろいろな言い方が考えられる。すなわち意味情報は同一であるが、それに対応する表層表現が違ふことがある(これを多様な文と呼ぶ)。一般にこれをすべて数え上げるのは難しいので、ここでは井上[5]を参考にする。それによると多様な文は(1)要素の位置に関するもの(2)機能要素の導入に関するもの2種類に分類できる。

久野[5]は要素の位置に関する違いを情報の新旧によって説明している。ここでは2章で仮に情報の単位を意味情報における接続された1つの基本述語とし、その順序によって情報の新旧を表現した。これにより要素の位置に関する制御は新情報に相当する意味要素から構造をつくりあげていくことにより実現することができる。しかしながらこの方法は情報の単位と意味単位が対応するかどうかはわからないと言う問題と語彙と意味単位が1対1に対応している場合には直接に語順を決定していることに等しくなってしまうと言う問題がある。

機能要素については3.2で述べたように談話情報の指定が直接その導入を決定するものと間接的に導入されるものがある。前者の例は視点と直接受動化の関係である。視点が意味要素中の第1引数(主語に相当する)ではない場合に直接受動化のための語彙である「れ・られ」が導入される。後者の例は主題と疑似分裂文化の関係である。主題で指定された部分は最後に作り上げるという制約があるので、主題以外の要素から構造を作ろうとする。ところが主題には関係要素である用言が取り込まれているので主題以外の要素だけでは句構造規則を実行できない。そこで機能要素である「だ」を導入する。また主題の方を作り上げると主題以外の部分から主題に対して助詞句でなければならないという制約を課されるので形式名詞の「の」を導入する。

5. アルゴリズムと例題

5.1. アルゴリズム

ここでは3, 4章の議論をまとめ、文生成のアルゴリズムを提示する。

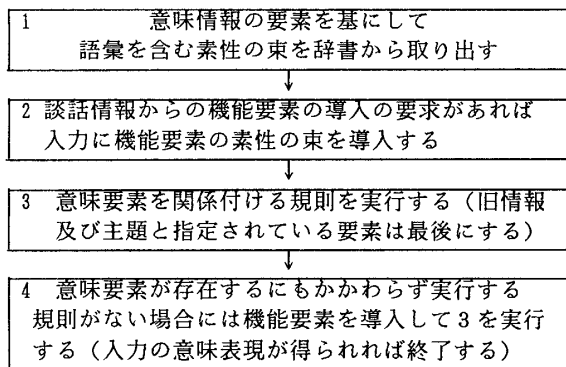
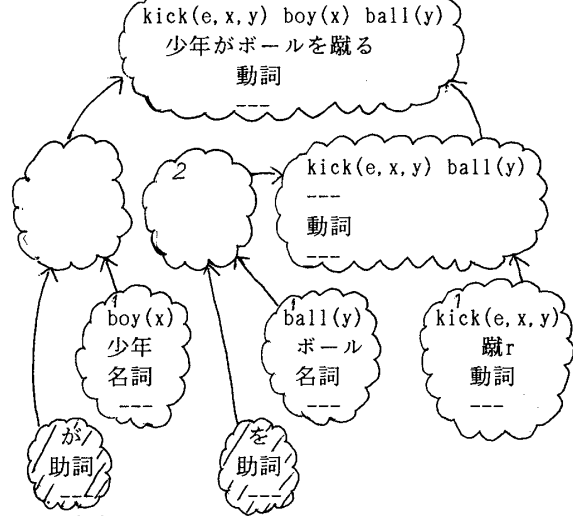


図1 素性駆動方式の文生成アルゴリズム

5.2. 例題

中間表現として例1, 2に示した意味情報と談話情報が与えられたとして図1に示した文生成アルゴリズムの処理の流れを図2でみてみる。☁は素性の束を表し、☉は

機能要素の素性の束を表す。まず辞書から意味要素に関連する素性の束が得る(1と番号付けされている)。談話情報からは機能要素の導入の要求はないので1で得られた素性の束から規則を駆動する。意味要素を関係付けることができる可能性があるのはkick(e, x, y), boy(x)及びkick(e, x, y), ball(y)だけである。情報の新しい要素から規則を実行するのでまず後者の方について規則を実行しようとするが、実行できない。そこでkick(e, x, y)の下位範疇化情報から機能要素を導入して規則を実行する(その結果得られた素性の束を2で示す)。さらに2と残りの素性の束について同じように規則を実行して最終的に「少年がボールを蹴る」という表現を得ることができる。



5. おわりに

本論文では意味情報, 談話情報を利用して日本語句構造文法から素性を利用してボトムアップ的に文を生成するアルゴリズムを提案した。意味情報に注目することにより情報の検査回数が(N!)²/NからN²(意味要素数をNとする)と大幅に削減できることを示した。また機能要素について談話情報及び適用規則がない場合に導入する方法について提案した。この機能要素を談話情報に応じて導入することにより文生成の過程を制御でき、生成する文を絞り込むことができる可能性を示した。

中間表現に関してはAppelt[2]で述べられている論理形式の同一性の問題(the problem of logical-form equivalence)を始めとして多くの問題がある。今後はその決定方法とともにそれらの問題を検討していく必要がある。

参考文献

[1]Stuart M. Shieber: A Uniform Architecture for Parsing and Generation, COLING-88, 1988
 [2]Douglas E. Appelt: Bidirectional Grammar and the design of natural language generation system, TINLAP-3, 1987
 [3]Jerry R. Hobbs: Ontological Promiscuity, Proceedings of the 23rd Annual Meeting, Association for Computational Linguistics, 1985
 [4]Takao Gunji: Japanese Phrase Structure Grammar, D. Reidel Publishing Company, 1987
 [5]井上和子: 日本語の文法規則, 大修館書店, 1978
 [6]久野すすむ: 談話の文法, 大修館書店, 1978
 [7]James Allen: Natural Language Understanding, The Benjamin/Cummings Publishing Company Inc, 1987