自然言語の語彙知識記述の評価のための解析システム

7H－5

中山拓也　松本裕治
奈良先端科学技術大学院大学
e-mail: {takuya-n,matsu}@is.aist-nara.ac.jp

1　はじめに

自然言語の文を解析する際には、さまざまな曖昧性を解消しなければならない[1][2]。曖昧性（単語区切り、品詞、係り受け、語義）の解消のために利用できる辞書的情報としては、シソーラスや格フレーム情報がある．これまでに、これらの情報を手作業で構築したりコーパスなどから自動的に獲得しようとする研究が幾つかなされてきているが、手作業・自動獲得いずれの場合も、構築された語彙情報が実際にどのくらい妥当なものかを評価する必要がある。特に自動獲得の場合は、評価尺度をあらかじめ用意しておき、その尺度による評価値が最も良くなるような学習獲得をしていくことが重要である。

語彙知識の評価の方法としては様々なものが考えられるが、曖昧性の解消に使用するという観点から見れば、実際に語彙知識情報を使って曖昧性が解消できるかどうかを試してみるという方法が考えられる。ここでは、そのような観点からの評価を容易に行うための解析システムを作成した。

2　曖昧性解消の観点からの評価

評価基準としては、実際に曖昧性解消を行った際に

1. 最終的な解釈に正解が含まれるかどうか
2. 構文解析の過程で生じる曖昧性の数(最終的な結果だけでなく解析途中の曖昧性の数も含む)を、いかに減らすことができるか

といったものが考えられる。特に2.の基準を使った評価を行うためには、語彙情報を用いた優先付けによる曖昧性解消を行いながら、実際に構文解析する必要がある。

3　優先付けの行える場面

曖昧性解消のための優先度が得られる場面としては、次のようなものが考えられる。

1. 単語や句の間に意味的関係が生じる際の優先度
2. 単語自体の優先度
3. 文法規則の優先度

1.は、動詞が後置詞句を補語としてとったり(補語構造)、形容詞が名詞を修飾したり(付加構造)する際に生じる優先度である。これには、例えば「猫」は「飼う」のヲ格は取りやすいが、ガ格は取りにくいといったものがある。また、「鯨にはヘソがある」のように「鯨」と「ヘソ」に部分・全体の関係があるという情報からの優先付けのように、複数の格を見ることで初めて分かるものも考えられる。

2.には、例えば「僕」は「ぼく」の意味の方が「しもべ」の意味で使うことが多いといったものがある。

また、1.や2.のように語彙情報によって得られる優先度の他にも、確率文法のように文法規則に優先度が付く場合(3.)も考えられる。

4　システム

4.1　概要

実際に様々な語彙情報を使って曖昧性を解消しながら構文解析するためには、

- **語彙情報を用いた優先付けが簡単に行える。**
- **様々な語彙情報を簡単に組み合わせて使える。**

ことが望まれる。本システムでは、これらの点について、なるべく容易に行えるような工夫をしている。

なお、実装については、現在のところ対象言語として日本語を考えており、形態素解析と構文解析にはJUMAN およびSAX を利用している。

4.2　文法

構文解析を行うためには、何らかの形の文法が必要となる。本システムで使う文法としては、3節で述べたような場面について、優先付けを行えるだけの記述が出来れば良い。

現在、HPSG[3](JPSG[4]) を参考にして、このような優先付けを行える試験的な文法を作成しており、解析可能な文をより多くするために改良を加えているところである。

A Disambiguation System for Lexical Knowledge Evaluation
Takuya Nakayama and Yuji Matsumoto
Nara Institute of Science and Technology

4.3 構成

システムは、図 1 に示すような構成になっている。統語解析を行う部分の他に、入力文の各単語について、文法で利用する為の素性構造を語彙情報から作る Make-FS と、必要に応じて語彙情報を用いた優先付けを行う Sem-Const がある。

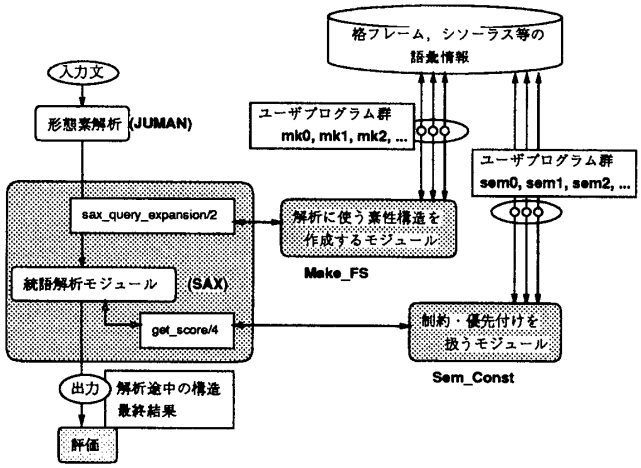

図 1: システムの概観

Make-FS

Make-FS では、登録されたユーザプログラム mk0 があれば、入力文の各単語について

mk0　品詞 単語 読み 基本形 活用型 活用形

のように引数を付けて呼び出し、mk0 が (標準出力に) 出力した結果を用いて素性構造を作成する。複数のユーザプログラムが登録されている場合は、それらから得る情報を統合して素性構造を作成する。

ここでは選言的な設定ができるようになっており、例えば、動詞「かける」には「掛ける」「欠ける」などの意味があるといった設定をする時に使う。なお、単語自体の優先付けは、ここで行うことも可能であるが、意味的優先付けを一括して扱うために Sem-Const で行う。

Sem-Const

Sem-Const では、登録されたユーザプログラム sem0 があれば、語彙情報による優先付けが行われる場面で

sem0　意味情報 1 関係 意味情報 2

のように引数を付けて呼び出し、sem0 が (標準出力に) 数値の形で出力する優先度 (スコア) を得る。例えば、[猫を]+[飼う] といった補語構造が現れた時は

sem0　[猫 : *] ヲ [飼う : *]

のように呼び出される。

ユーザプログラムは複数登録可能であり、それぞれのプログラムから得たスコアは、ユーザが設定した計算式によって 1 つにまとめたスコアにされる。システムは、この値をもとに曖昧性解消を行いながら解析を進める。曖昧性解消は、設定された閾値よりもスコアが低いものを取り除くという方法で行う。

評価したい語彙情報があったとき、ユーザとしては、その語彙情報のみを使って得られるスコアや素性構造の一部を出力するプログラム群を作ればよい。それらを登録しておくことで、それらを使った曖昧性解消を行いながら例文を解析できるので、結果的に、評価したい語彙情報の曖昧性解消への有効性を知ることができる。

5 おわりに

評価をする際、どの解釈が正解であるかを知る必要があるが、基本的には人手で判断するしかない。但し、構文解析 (bracket) 済コーパスを用いることで、係り受けの曖昧性解消についての評価を行うことは可能である。

今後は、解析済コーパスを用いることで、自動的にある程度の評価をする機構を作成し、実際にいくつかの語彙情報を評価してみたり、システムを用いた語彙知識獲得を試みる予定である。

参考文献

[1] 奥村 学. 自然言語の意味的曖昧性の解消法. 人工知能学会誌, 10(3):332–339, May 1995.

[2] 長尾 確. 自然言語理解のための意味・文脈処理に関する研究. PhD thesis, 東京工業大学 情報工学科, Mar. 1994.

[3] C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar.* The University of Chicago Press, 1994.

[4] T. Gunji. *Japanese Phrase Structure Grammar.* D.Reidel Publishing Company, 1987.