接続確率による日本語形態素解析の曖昧性解消法

5B－7

大川克利　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

日本語の形態素解析では、単語分割や同形語による曖昧性が大きな問題となる。従来、このような問題を解決するために文節数最小法や分割数最小法といったヒューリスティックスがよく用いられてきた。しかし、このようなヒューリスティックスだけでは限界があり、十分な解決策とはいえない。構文・意味情報や文脈情報を利用するのも一つの解決策だが、形態素解析の段階でこれらの情報を利用することは、処理速度の面からは望ましくない。

本稿では、形態素間の接続情報に着目し、隣接する形態素間に接続確率を設定し、接続確率を用いた評価を従来のヒューリスティックスに加えた日本語形態素解析の曖昧性解消法を提案し、その有効性を示す。

2　接続確率の設定

日本語形態素解析 [1] では、接続ルール [2] を参照して解析を行なっている。接続ルールには、隣接する２つの形態素間の接続情報が記述できるようになっているが、一般的な接続とはいえないものに接続確率という形で接続の強さを下げているだけで、ほとんどの接続が同じ強さになっている。しかし、形態素間の接続は、用言の語幹と語尾のように接続しやすいものや名詞と句点のように接続しにくいものといったように、それぞれ接続の強さが異なる。

本稿では、接続の強さを接続確率として扱い、

1. 接続ルール型
2. 品詞フィルタ型

の２つに分けて細かく設定する。

2.1　接続ルール型

[2] の接続ルールは、接続確率が記述可能になっている。ここに、接続確率を記述することによって隣接２項間の接続の強さを表すことができる。接続確率は、日本語文コーパスから、接続データを収集し、統計をとることによって獲得する。将来的には、学習機能によって接続確率の精度を上げていくことを考えている。

2.2　品詞フィルタ型

形態素の中には連語のように３項以上の組合せによって強固な接続をするものもある。しかし、接続ルールは基本的には２項関係を記述する形式になっているので３項以上の接続は記述しにくい。そこで、３項以上の形態素列の接続の強さを記述した品詞フィルタを接続ルールとは別に用意する。

code1（字面），code2（字面），…，codeN（字面）：p

code ： 品詞コード
p ： 接続確率
※字面は省略可

図 1: 品詞フィルタ型

3　接続確率による評価

形態素解析の結果は図2のように曖昧性をもったグラフ構造で出力される。このグラフ構造を展開し、展開された解析結果に対してそれぞれ以下のような評価を行なう。

Disambiguation of Japanese Morphological Analysis Using Connection Probability
Katsutoshi Ohkawa , Masahiro Miyazaki
Niigata University

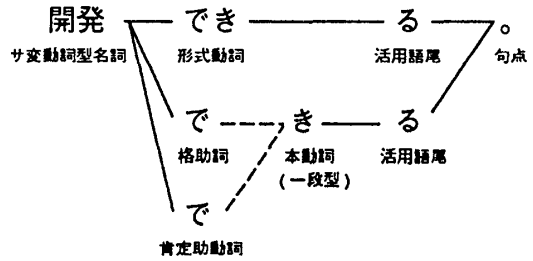

図 2: 形態素解析結果

評価は、まず文頭から接続ルールを参照していき文末までの接続確率を加算していく。さらに、品詞フィルタに当てはまる場合はその点数を加える。この点数を接続数、すなわち分割数で割った値を評価点とする。

$$P = \frac{1}{n}\left(\sum_{i=1}^{n} p_i + F\right)$$

n：接続数
p_i：接続確率 $(i = 1, 2, 3, \ldots, n)$
F：品詞フィルターによる点数

4 曖昧性の解消

4.1 同形語のパック化

同形語の曖昧性の解消のためには、構文・意味情報が必要であり、上記の評価では解消は困難である。同形語の曖昧性は、形態素解析の段階では保持しておき、複合語の構造解析や構文・意味解析で解消するのが得策である。そこで、あらかじめ統語的特徴が同じ同形語をパック化してから評価を行なう。

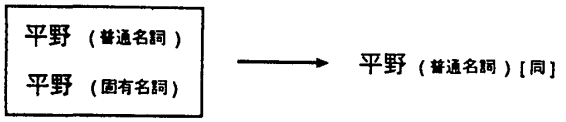

図 3: 同形語のパック化

4.2 評価

同形語のパック化をした後、図2のようなグラフ構造で出力された解析結果を展開し、各解析結果に対して評価を行なう。評価は、接続確率による評価、文節数による評価、分割数による評価に適当な重みをつけて行なう。

$$Z = P - S \times 0.3 - D \times 0.1$$

P：接続確率による評価の点数
S：文節数
D：分割数

4.3 絞り込み

評価結果から評価点が最も高い解析結果を正解として選ぶ。これによって、単語分割の曖昧性が解消される。これでもまだ同形語の曖昧性が残るわけだが、この曖昧性は構文解析以降に委ねることにする。しかし、形態素解析の段階で同形語の曖昧性を絞り込む必要がある場合は、同形語の使用頻度や局所的な共起関係などを基に絞り込むしかない。

5 おわりに

従来のヒューリスティックスに接続確率による評価を加えることによって、日本語形態素解析の曖昧性解消の精度の向上が期待される。今後、本手法の定量的評価を行ない、重みの再調整を行なう予定である。また、接続確率を形態素解析内で学習できる機構について検討する必要がある。

参考文献

[1] 高橋、佐野、宍倉、前川、宮崎：頑健性を目指した日本語形態素解析システムの試作、「自然言語処理における実動」シンポジウム論文集、pp.1-8(1993)

[2] 宮崎、高橋：三浦文法に基づく日本語形態素処理文法の構築、情報処理学会研究報告、92-NL-90、pp.1-8(1992)