1L－4

正例と負例のコーパスを用いた日本語形態素解析の確率論的曖昧性解消機構

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。そこで、近年では確率論的な手法がいくつか提案されている [1][2]。

本稿では、隣接2形態素間の接続強度による確率論的手法をベースに、連語コーパス（正例コーパス）や誤った解析結果を集めた解析誤りコーパス（負例コーパス）を用い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　日本語形態素解析システム

図1のように拡張CYK法による解析をベースに固有名詞解析、数詞解析、未知語解析、複合名詞解析から構成される日本語形態素解析システムが試作されている [3]。本システムでは、入力文に対して形態素辞書を検索し、CYK表に格納する。その際、固有名詞解析や数詞解析も随時行なう。次に、接続辞書 [4] を参照し、CYK法による解析を行なう。接続失敗などにより、文末まで解析ができなかった場合は、固有名詞解析、未知語解析を行ない、文末まで解析を成功させる。最後に、構造化ルールを用いた複合名詞解析を行ない、解析結果を出力する。この解析結果は、同形語や単語分割の曖昧性を含んだグラフ構造で出力される。

現在のところ、文当たり97～98%正解を含んだ結果を得ることができる。2～3%の失敗は、主に未知語解析における解析誤りや品詞誤りが原因である。

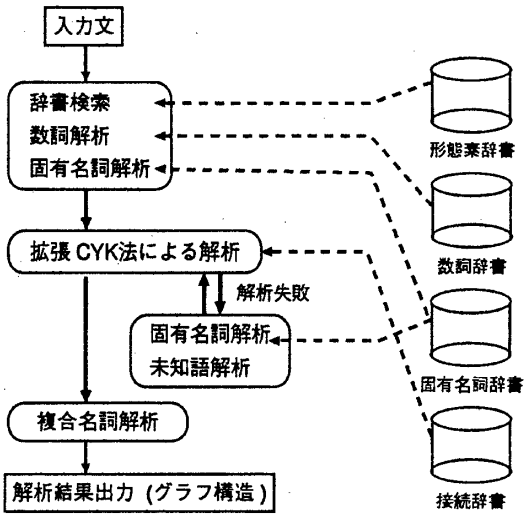

図 1: 試作された日本語形態素解析システム

3　曖昧性解消機構

このような曖昧性を含んだ解析結果を絞り込むために、下記に示すようなコスト最小法による評価をベースとした曖昧性解消を行なう。

3.1　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る。特に、名詞や動詞、助詞の同形語は形態素解析レベルで解消するのは困難である。そこで、統語的性質が同一な同形語の曖昧性は、パック化することによって保持しておき、複合語の構造解析や構文・意味解析において解消することにする。

3.2　コスト最小法による評価

同形語のパック化をした後、コスト最小法を用いて曖昧性解消を行なう。ここで、コストとしては2形態素間の接続コストを用いる。この接続コストは、接

Disambiguation of Japanese Morphological Analysis Using Positive and Negative Example Corpus
Katsutoshi Ohkawa , Masahiro Miyazaki
Niigata University

続辞書に記述されている接続確率と、連語コーパス（正例コーパス）、誤った解析結果を集めた解析誤りコーパス（負例コーパス）から得られる。

4 接続コスト

4.1 接続辞書

CYK 法による解析の際に参照する接続辞書には、接続確率が記述されている。この接続確率をコスト最小法に適応できるように修正する。つまり、接続確率が高いものを小さいコストにする。

4.2 正例コーパス

接続辞書の接続確率は、隣接 2 形態素間の接続しやすさを表しているが、形態素の中には、連語のように 3 つ以上形態素の組合せによって強固な接続をするものもある。そこで、連語のコーパスを用意し、連語に相当する形態素列に負（−）のコストを与え、評価を高くする。

4.3 負例コーパス

また、形態素の中には 3 つ以上の形態素の組合せによって接続を弱めるものもある。これに対しては、負例コーパスを用いる。負例コーパスは、形態素解析の解析結果の誤った部分をデータベース化したものであり、図 4.3のように品詞レベルと字面レベルから構成される。まず、品詞レベルのコーパスを参照し、マッチしたら形態素列に正（+）のコストを加え評価を低くする。さらに、字面レベルのコーパスを参照し、マッチしたら正のコストを加える。

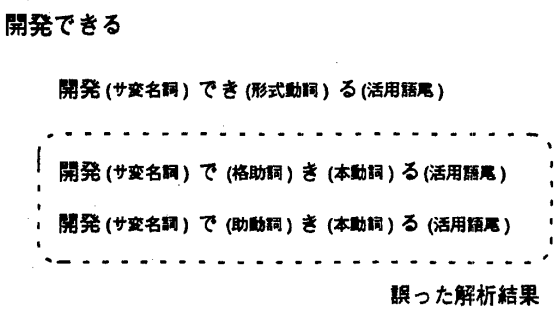

図 2: 誤った解析結果

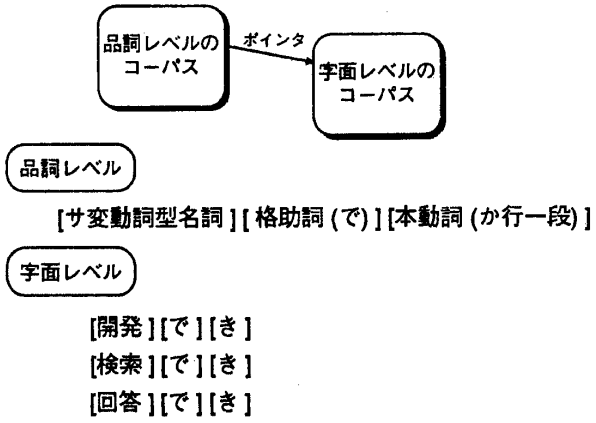

図 3: 負例コーパス

5 試験的評価

日本語文コーパス 300 文に対して試験的評価を行なった。分割数や文節数の最小法では、文当たり 88%の正解を得た。本手法においては、未知語がある文を含めても 95%以上の正解が期待される。

6 おわりに

接続確率、正例コーパス、負例コーパスの 3 種の接続コストを用いたコスト最小法による評価によって、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、これらの接続コスト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] 永田昌明：自然言語の確率モデルと統計的学習、「自然言語処理における学習」シンポジウム論文集、pp.17-24(1994)

[2] 長尾真：岩波講座ソフトウェア科学 15、自然言語処理、岩波書店、pp126-129(1996)

[3] 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)

[4] 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)