用例に基づく形態素解析の検討

Example Based Morphological Analysis

村上仁一
Jin'ichi Murakami
NTT 情報通信処理研究所
NTT Information and Communication Laboratories

1 はじめに

形態素解析は、従来から対話、翻訳、校正などの目的のために、自然言語処理研究の一つの分野として研究が続けられている。形態素解析は、漢字かな文を単語に分けて品詞ラベルを付与することであるが、通常、大量の候補が出力されるため、言語情報を持ちいてこれらの曖昧さを削除している。この言語情報として、単語を構文的意味的なカテゴリに分類してカテゴリー間の接続ルールや係受けルールなどが利用されている [1]。しかし、実際の日本語では単語の境界が明確でないことや単語の多品詞性や曖昧な係受けなどの問題があるため、精密なルールの作成は容易でない。そこで、本稿では、これらの辞書を利用する代わりに、既に形態素解析された結果を利用する、用例に基づく形態素解析を提案する。

2 用例に基づく形態素解析

本稿で提案する用例に基づく形態素解析は以下のような流れになる。

1. 形態素解析済みのデータの準備

   形態素解析が既にされたデータを大量に準備する。一般に、このデータは人間によるクリーニングがされているため、形態素解析の誤りは少ない。このデータを従来の形態素解析における単語辞書として扱う。

2. 形態素解析

   1. で用意された形態素解析済みのデータを利用して、新しい日本語の形態素解析をおこなう。このプログラムは、従来の形態素解析と同様、文節数最小法もしくは最長一致法を利用する。

以下に日本語の名詞連続複合語の形態素解析の例をあげる。辞書には、表 1 のような形態素解析済みのデータを辞書として登録する。

表 1: 形態素解析済み辞書

1	日本 { 一般名詞 } 電信 { 一般名詞 } 電話 { 一般名詞 } 株式 { 一般名詞 } 会社 { 一般名詞 }
2	東京 { 一般名詞 } 海上 { 一般名詞 } 火災 { 一般名詞 }
3	安田 { 一般名詞 } 海上 { 一般名詞 } 火災 { 一般名詞 }
4	総務 { 一般名詞 } 部 { 接尾語 }
5	秘書 { 一般名詞 } 部 { 接尾語 }

次に、新しい名詞連続複合語
「日本電信電話株式会社総務部」
を形態素解析することを考える。

最長一致法を用いて、表 1 を辞書として、表 2 の形態素解析結果が得られる。

表 2: 形態素解析結果

1	日本 { 一般名詞 } 電信 { 一般名詞 } 電話 { 一般名詞 } 株式 { 一般名詞 } 会社 { 一般名詞 }
2	総務 { 一般名詞 } 課 { 接尾語 }

ここでは、名詞連続複合語の形態素解析の場合について述べたが、一般的な日本語の形態素解析にも使用できる。

3 考察

この用例ベースの形態素解析には、以下の長所があると思われる。

1. 辞書作成の容易性

   単語辞書の代わりに、形態素解析済みのデータを使用するため、辞書の作成が容易である。

2. 形態素解析の精度

   人間によってクリーニングされた形態素解析の結果を使用するため、形態素解析の精度が高いことが予想される。

3. 安定度

   日本語は単語の境界が明確でないことや単語の多品詞性などの問題があるため、形態素解析結果に曖昧さがでる。しかし、用例ベースの形態素解析では、人間によってクリーニングされた形態素解析の結果を利用するため、この曖昧さが、少ないと思われる。

4 まとめ

本論文では、既に形態素解析された[illegible]について形態素解析をおこなう、用例ベースの形態素解析方法について述べた。今後、この手法による形態素解析の精度を求める実験を行なう予定である。

参考文献

[1] 長尾 真, “日本語情報処理,” 社団法人電子通信学会, pp.63-64 (1984).