

4E-1

木元 逞

1.はじめに 自動翻訳電話研究用の基礎的言語データ[1][3][4]の一部として、話し言葉を主たる対象として形態素情報を中心とした日本語の単語に関する情報(以下形態素情報と呼ぶ)を収集している。本稿では、形態素情報収集の過程を支援し、効率的に形態素情報データを構築するためのシステムについて述べる。

形態素情報としては、会話(文章)に現れている単語そのものと、その文脈でのその単語の読みと、単語の標準表現(活用語の場合は終止形)、品詞、および活用のある単語の場合はその単語がどのように活用するかを示す活用型、その文脈での活用形、音便(表1では省略)を入れている。

単語	読み	正規表現	品詞	活用型	活用形
会議	かいぎ	会議	普通名詞		
に			格助詞		
参加	さんか	参加	サ変名詞	サ変	連用
した	した	する	補助動詞		連体
いの	いの	いの	助動詞		
です	です	です	準体助詞		
の			助動詞		終止
です			助動詞		
が	が	が	接続助詞		

形態素情報の収集を支援するシステムの概要を図1に示す。システムは日本語テキストを形態素解析する形態素解析システムと、その結果を人手によって効率よく修正する作業を支援する形態素修正システムと、形態素情報収集を間接的に支援する形態素情報修正履歴検索システム、文章比較システム、形態素情報検索システム、形態素情報表示システムからなる。

日本語テキストから形態素解析システムを使って自動的に形態素情報を収集する。現在の話し言葉を解析する技術では、完全に正しい解析結果は望めない。そこで、単語の接続条件レベルの解析で形態素情報を集め、形態素情報修正システムを使って形態素解析で間違えた部分を人手で修正するという方式を取る。実際のデータに対して、人手の修正作業を行うことにより、単語の区切り単位や品詞などに関する問題点を明確にできる。どのような修正が行われたかは修正履歴として蓄積することができる。修正履歴は修正履歴検索システムを使って分析

言語データは、高品質で均質で首尾一貫したものが要求される。そこで、原文のテキストと形態素情報修正結果を比較して、食い違いがあればその箇所を指摘する文章比較システムや、形態素情報の検索システムや表示システムを利用して高品質で均質で首尾一貫性を持ったデータを収集する。また、データ修正作業によるデータのゆれを最小限に押さえるため、作業マニュアルを用意してデータの首尾一貫性と均質性を計っている。

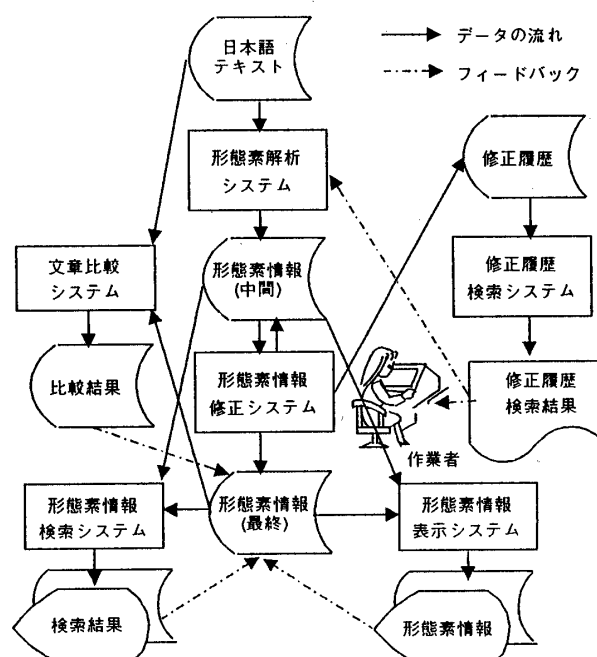


図1. 形態素情報収集支援システムの概要

形態素解析システム^{[5][6]}は日本語のテキストに対して形態素情報を付けるものである。図2に形態素解析システムとその周辺ツールの関係を示す。形態素解析システムの主な項目を以下に示す。

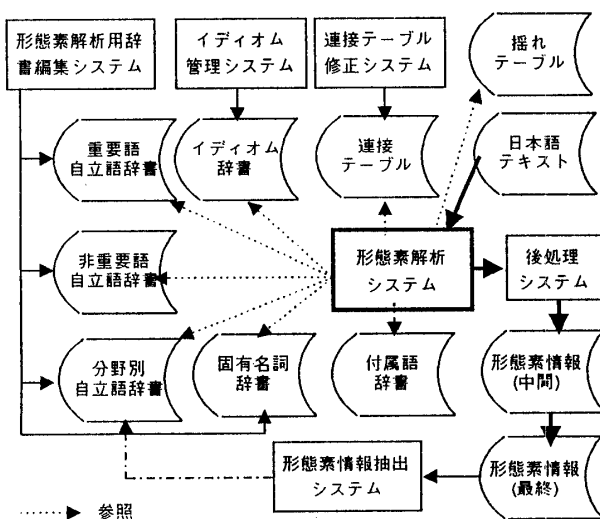


図2. 形態素解析システムとその周辺ツール

- ①入力: べた書きの仮名漢字混じり文
(表記のゆれも許す)
- ②出力: 単語、読み、標準表現、品詞、
活用型、活用形、音便情報、コメント
- ③解析手法: 最長一致法、最尤度法
- ④単語の接続条件: 接続テーブル
- ⑤辞書: 優先度付きの複数の辞書、イディオム
- ⑥未知語処理: 字種情報の利用

形態素解析の正解率を向上させ、人手による修正作業を減らすために、入力文の揺れへの対処や、最尤度法による解析、接続条件の改良、優先度付きの複数の辞書の利用、未知語処理、イディオム処理などを行っている。

“表す”、“表わす”のような表現の揺れやひらがな書きによる揺れについては、揺れテーブルで対処している。揺れテーブルでは、例えば、“表す”、“表わす”に対して同一の辞書エントリ(標準表現)“表す”を持たせている。最尤度法の尤度としては、単語の長さ、単語の頻度、品詞の優先度、品詞レベルの二項間頻度を用いている。単語の接続条件はテーブル形式でデータとして持っている。その接続テーブルは接続テーブル修正システムを使って、条件の改良を容易に行うことができる。

形態素情報から形態素情報抽出システムを使って、分野依存の辞書を構築できる。また、複数の形態素解析辞書を優先度を付けて利用することができる。形態素解析辞書は、形態素解析用辞書編集システムを使って容易に修正・改良できる。このことにより、対象とするテキストの分野により、分野別の辞書を取り替えることによって、解析システムの能力を向上できる。

“～のでしょうか”のような話し言葉の慣用的表現は、うまく形態素解析されない場合が多い。このようなものを慣用句登録して優先的に処理することにより、人手による修正作業をかなり軽減できる。慣用句はイディオム管理システムを使って、イディオム辞書に登録できる。全く慣用句登録を行わない場合と慣用句登録を行った場合を比べると、慣用句

登録を行った場合は形態素情報の1回目の修正率が27.4%から19.5%に下がった。慣用句情報の利用の有効性を確認できた。

形態素解析システムの後処理システムも人手による修正作業の負荷を軽減するためのもので、形態素解析システムで生じるパターン化した誤りを、パターン変換ルールを使って自動的に修正する。現在は、形態素解析システムが利用している新明解国語辞典の品詞や語の区切りの体系と、我々が定めたもの^[7]との違いなどをこれを用いて埋めている。また、人手による修正が終了した後でも、品詞や語の区切りの体系の修正に伴いデータを修正しなければならない時、有効に働く。

形態素解析システムの能力としては、4355語に対する形態素情報修正が、1回目が628件、2回目が65件、3回目が19件となっており、合計712件なので、正解率は84%以上である。(後処理を行っていない場合の数値である。)

5. 形態素情報修正システム

形態素情報修正システムの機能を以下に示す。

- ①挿入、削除、変更モード
- ②後戻り機能
- ③検索機能
- ④修正履歴保存機能
- ⑤慣用句の登録機能

人手による修正の効率を上げるために、修正したいパターンを見つけながらQuery-Replace的な修正を行う検索機能を用意している。これにより、パターン化した修正を容易に行うことができる。

6. おわりに 形態素解析システムを中心とした種々の収集支援システムにより、良好な支援環境を用意して効率良く形態素情報を集めている。形態素情報の修正は3回行っており、形態素情報修正作業の平均所要時間は、1ファイル(約300語)当たり、1回目が1時間20分、2回目が24分、3回目が10分計1時間54分程度である。ここで集めた形態素情報は、当研究所で開発した言語データベース統合管理システム^[2]を使って、当研究所で集めている係り受け情報や対訳対応情報等と組み合わせて、言語処理研究に有効に活用できる。

謝辞 日頃御指導いただく樽松明社長に感謝致します。
<参考文献>

- [1] 小倉・篠崎・森元、言語データベース収集支援システム、情処学会第36回全国大会4U-4、1988
- [2] 小倉・橋本・森元、言語データベース統合管理システム、情報処理学会第37回全国大会5B-6、1988
- [3] 篠崎・小倉・森元、言語データベースの品質管理、情処学会第36回全国大会4U-3、1988
- [4] 森元・小倉・飯田、自動翻訳電話研究用言語データベースの収集について、情処学会第36回全国大会4U-5、1988
- [5] 吉村・武内・津田・首藤、コスト最小法を用いた日本語文の形態素解析、情処学会NL研60-1、1987
- [6] 吉村・日高・吉田、日本語文の形態素解析における最長一致法と文節数最小法について、情処学会NL研30-7、1982
- [7] 吉本、日本語品詞の分類、ATRテクニカル・レポートTR-1-0008、1987