5G-8

日本語辞書の開発と評価について

小林 賢一郎' 斎藤 裕美"
'東芝オーディオ・ビデオエンジニアリング株式会社 "株式会社東芝

1. はじめに

今日、仮名漢字変換をはじめとして文書校正支援システムや機械翻訳システムなど日本語処理を応用した製品が各社で開発されているが、語数の増強などに対応して効率良く品質の高い単語辞書の開発・管理がますます重要になってきた。

そのため辞書データの作成、修正からその評価までが短期間に行なえるようなシステムが必要となってくる。

しかも、これらの日本語解析に用いてる仮名漢字変換辞書と日本語形態素解析用辞書の単語辞書は形態素レベルまでの情報が共通であるにもかかわらず、必ずしもそのことを考慮にいれた管理方法が取られてはいなかった。

それらの点に着目して、仮名漢字変換辞書と日本語形態素解析用辞書を一つにまとめてマスター辞書として管理する単語辞書の開発環境を試作した。

この方法は、単語辞書の評価を日本語形態素解析を応用して行なうもので、単語辞書の評価のみならず、仮名漢字変換、日本語形態素解析のロジックの評価も同時に行なえるシステムとする。

本報告では大量の単語辞書を管理する際に日本語形態素解析を応用することにより単語辞書の評価を自動的に行えるシステムの設計およびその応用について報告する。

2. システムの構成

本システムは大きく分けて辞書変換部、漢字仮名変換部、仮名漢字変換部、形態素解析部、仮名漢字変換評価部、頻度チェック部から構成されている。各構成要素は次のような働きをする。

辞書変換部

マスター辞書から仮名漢字変換用辞書、日本語形態素解析用辞書のそれぞれのデータ形式に変換する。

漢字仮名変換部

日本語形態素解析を利用してテキスト原文を仮名漢字変換するための仮名文字列に変換する。対話的補助機能を持つ半自動式のものである。

仮名漢字変換部

仮名文字列を漢字仮名混じり文字列に変換するロジック。

評価部

日本語形態素解析を利用してテキスト原文と仮名漢字変換結果の双方の漢字仮名混じり文字列を文節解析することにより評価を行なう。

- 仮名漢字変換評価
  テキスト原文と変換文との差異を分析する。
- 頻度チェック
  仮名漢字変換を評価した際に同音語として選択された単語の使用頻度をチェックする。

Development and Evaluation of Japanese-Dictionary.
Ken'ichirou KOBAYASHI', Hiroyoshi SAITO'
'TOSHIBA AUDIO VIDEO ENGINEERING., Ltd., 'TOSHIBA Corp.

3．処理の流れ

処理の流れは図１に示すような形を取り、次のようになる。

（１）マスター辞書から仮名漢字変換用の辞書、日本語形態素解析用の辞書のそれぞれを辞書変換部により作成する。

（２）原文に対して漢字仮名変換を行うことにより原文の仮名テキストを求める。未登録語や解析の結果の曖昧なものに対しては対話形式の入力により行う。出来上がった仮名テキストは原文に対応付けて保存される。

（３）得られた仮名テキストを仮名漢字変換する。仮名漢字変換は実際のキーボードなどからの入力を変換するものを再現できるようにしたものである。

（４）仮名漢字変換の結果として得られた出力を漢字変換評価部に対して渡す。ここでは、形態素解析を利用して仮名漢字変換の出力結果を文節に切り分ける。切り分けられた各文節は原文を同様に文節に切り分けたものと比較される。

この文節単位での比較は仮名漢字変換の失敗による文節のずれを考慮して原文と変換結果の文節が最多一致する状態を求めて行う。

（５）このときに仮名漢字変換の結果が原文と一致しなかった部分を未登録語として出力すると同時に、第一候補として出力されて原文と一致しなかった候補と全ての候補の中で一致した候補をそれぞれ集計して各単語の出現頻度を調べる。

（６）出現頻度情報を基に各単語の優先順位の変更を行うと共に未登録語として出力された単語の追加、削除をマスター辞書に対して行う。

この新しいマスター辞書を基に新しい辞書を作成して一連の評価を再度繰り返して行う。

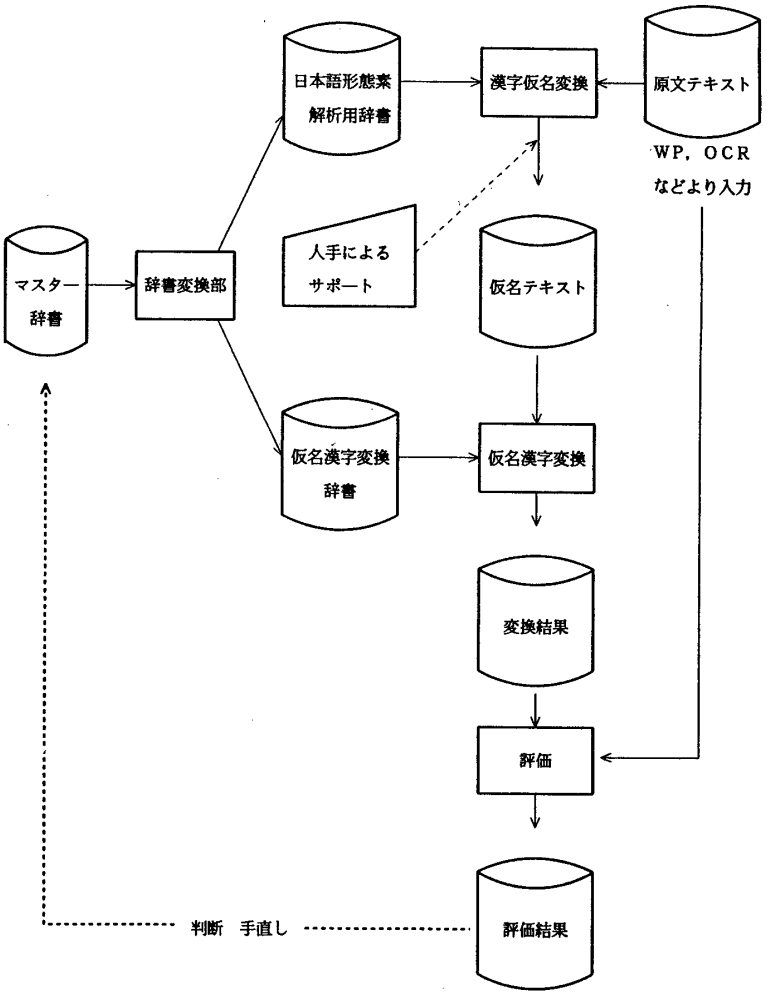

図１．処理の流れ

4．応用

この辞書開発環境の応用としては仮名漢字変換辞書を固定にして仮名漢字変換ロジックを変更することにより仮名漢字変換ロジックそのものの評価を定量的に行うことが考えられる。

5．おわりに

今後はこの辞書開発環境を用いて辞書の評価および作成を行っていきたいと思う。また、合成語と接辞の関係を種々の辞書に対して評価実験することにより調査していきたいと思う。