文書認識における言語情報の活用 (1)
－概要－

2J－8

小川 知也　斉藤 孝広　松井 くにお
富士通研究所

1 はじめに

既存印刷文書を電子ファイリングして活用するために, OCR は有力な手段である. しかし日本語文書を対象とする場合, 日本語は文字種が多いことなどから認識誤りを完全に無くすことは困難である. [1]

文字認識結果中の候補文字列から最適な単語のパスを探索することによる誤り訂正・指摘方式を開発したので, その概要について述べる. また, 文字認識における切り出し誤りへの対応も考慮した拡張形態素解析について論じる.

2 概要

2.1 誤りの分類と対策

プロトタイプシステムによる予備実験の結果, 文字認識における誤りを次のように分けて考えることにした.

1. 一文字認識における誤り
   (a) 候補文字に正解が含まれているもの
   (b) 候補文字に正解が含まれていないもの
2. 文字切り出しにおける誤り

候補文字に正解が含まれているものに対しては, 候補文字を考慮した形態素解析の精度を上げることで対処する. 日本語文章としての自然さを的確に反映するように言語辞書や文法知識のチューニングを行う.

候補文字に正解が含まれていないものは, 画面の不鮮明さ (かすれ, つぶれなど) やゴミ, その他が原因と思われる. これらのうち, 文字認識プログラムに特徴的な誤りに対しては, 類似文字 DB による候補文字補完で対応する.

文字切り出しにおける誤りに対処するには, 正しい切り出しを含む複数切り出しを扱う形態素解析が有効と考える.

Application of linguistic information in document recognition
Tomoya OGAWA, Takahiro SAITO, Kunio MATSUI
Fujitsu Laboratories Ltd.

2.2 構成

以上の考察をふまえ, 誤り訂正・指摘方式の全体構成を図 1に示すものとした.

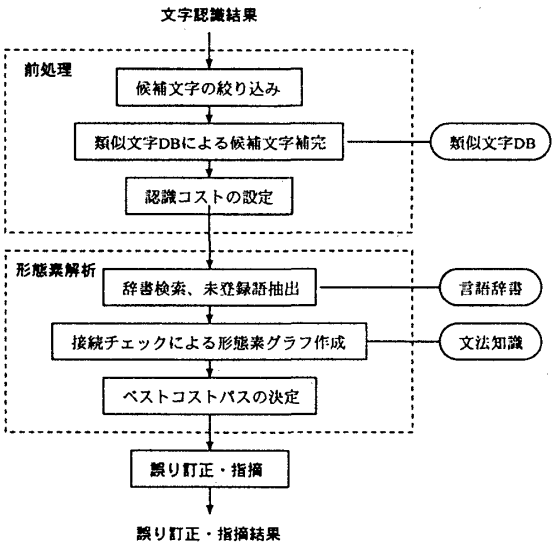

図 1: 誤り訂正・指摘方式の全体構成

本方式における形態素解析は,

1. 辞書検索と, 未登録語の抽出
2. 形態素間の接続チェックによる, 形態素グラフの作成
3. コストに基づくパスの決定

というステップから成る. 解析においては, 文節数最小法などの一般化として知られるコスト付き形態素解析を行う. 解析コストとして扱うものは,

接続コスト

形態素コスト
自立語 / 付属語, 品詞, 単語長, 未登録語などに応じたコスト

連接コスト
前後の形態素の連接情報に応じたコスト

認識コスト
　認識距離値に応じたコスト

である．実際にはそれぞれに重みをかけその和を解析コストとすることで，バランスの調整を容易にしている．

累積認識率と訂正性能，処理速度を考え合わせ，デフォルトでは文字認識結果の4位候補までを解析の対象としている．文字認識結果には，イメージとしての確からしさをある程度反映する認識距離値が含まれる．この値を利用して，さらに候補文字の絞り込みを行う．これにより，訂正性能をほとんど低下させることなく，処理速度を向上させ，改悪も減らすことが可能となる．

類似文字データベースによる候補文字補完では，文字認識プログラムの傾向を反映させた補完を行う．補完する文字に与える認識コストは，改悪が起こさずに効果が発揮されるような数値を与える．

こうして前処理された文字ラティスを拡張形態素解析エンジンに与え，日本語として最も自然と思われるパスを選択する．

拡張形態素解析の結果得られるパス情報を元に，文字認識結果の誤り訂正を行う．また，日本語として不自然な部分 (解析コストの高い箇所など) に関する情報などを元に誤り指摘を行う．

3　拡張形態素解析

文字認識結果に含まれる切り出し誤りに対処するために，複数切り出しを扱う形態素解析が有効と考える．

これはCYK法を用いる文字列形態素解析 [2] に次のような拡張を行うことで実現される．

3.1　文字ラティス

複数切り出しを含む文字ラティスの例を図 2 に示す．

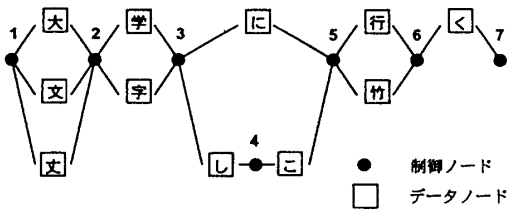

図 2: 文字ラティス

形態素解析処理において，各形態素は制御ノードを基準に管理される．

3.2　辞書検索

言語辞書としては，処理効率などを考慮してトライ構造の辞書を用いる．辞書検索は，文字ラティスを先頭ノードをルートとするツリーと考え，トライ構造辞書の各インデックスと対応付けることにより，キー主導型の検索を行う．

3.3　未登録語抽出

未登録語抽出では，2位候補以下の文字も対象とした抽出を行う．これは数詞抽出を未登録語抽出で行っていることなどのためである．そのため，同じ開始位置 (制御ノード) から複数のパターンが抽出されることがあり得るので，抽出処理中は状態の集合を管理する必要がある．

抽出パターンを柔軟に定義でき，効率的な処理ができるよう，図 3 のような状態遷移図に基づく未登録語抽出を行う．

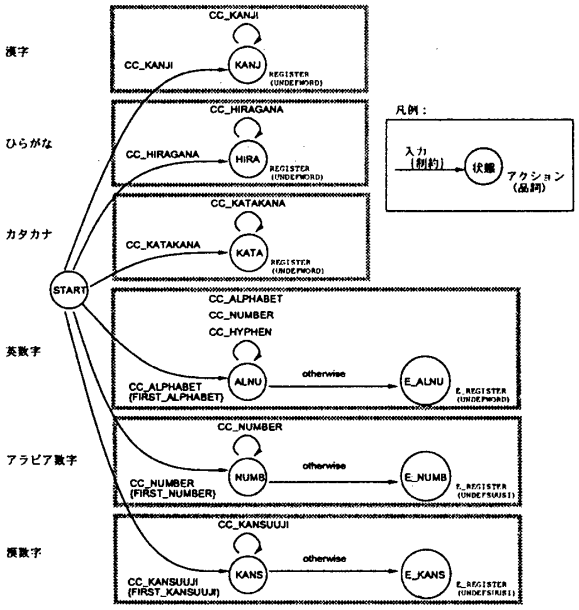

図 3: 未登録語抽出

4　まとめ

言語情報を活用する文字認識誤り訂正・指摘方式について，その概要を説明した．今後は切り出し誤りに対する有効性の検証を含め，さらに高度な誤り訂正・指摘方式の開発を目指したい．

参考文献

[1] 西野文人：文字認識における自然言語処理，情報処理，Vol. 34, No. 10, pp. 1274–1280 (1993).

[2] 田中穂積：自然言語解析の基礎，産業図書 (1989).