

漢字N-gramを用いた読み付与システム

5H-5

日本アイ・ビー・エム株式会社

東京基礎研究所

鳥原 信一

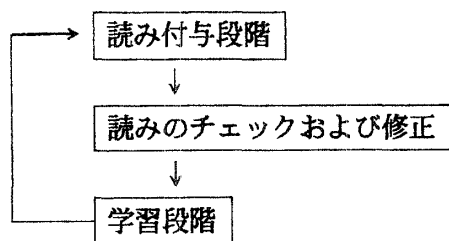
1. はじめに

日本語テキストに、より正確な読みを付与することは、テキスト音声合成および自動点訳といったアプリケーションにとって重要である。先行論文[1]においては、漢字の前後に環境文字をおくことにより、読みが絞れることを検証した。本論文においては、0から3文字の環境文字を前後におくことにより、実際に読み付与を行う実験システムを作成したので、その概要を記述する。また、「行っ(イッ、オコナッ)」および「通っ(トオッ、カヨッ)」を含むテキストに読みを付与して、その読み分け精度について実験を行ったので報告する。

2. 環境文字による読み付与システムの概要

2.1 読み付与段階と学習段階

この読み付与システムは、つぎの2つの段階と、その中間に人間が介在することにより、読み付与の正確さを増していくことができると考える。



2.2 読み付与段階の構成要素

3種類のデータ・ファイルおよび読み付与プログラムにより構成される。

System of Yomi-setting by Kanji N-gram

Shinichi Torihara

IBM Research, Tokyo Research Laboratory,

IBM Japan

ファイル(1)

漢字・読みファイル。

見出し漢字に対してその読みを列挙したテーブル。読みが単一のものは一意にその読みが採用される。

ファイル(2)

漢字数>読み数、読み・漢字逆順ファイル。

このシステムはある漢字には必ず読みが対応していると考えているが、漢字の文字数の方が読みの文字数より多い場合、漢文調の漢字などが例外なのでこのファイルに収めている。

例) 香具師 ヤシ

不忍池 シノバズノイケ

ファイル群(3)

前後0から3文字の環境文字を持つ読み付き漢字N-gramファイル。

15個を用意して、環境文字の多い方を先に検索する。

```

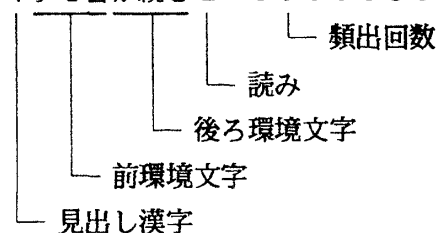
for (left_width=3; left_width>=0; left_width--) {
    for (right_width=3; right_width>=0; right_width--) {
        if (left_width==0 && right_width==0) break;
    }
}
  
```

left_width : 前環境文字数。

right_width : 後ろ環境文字数。

つぎに環境文字3+3のN-gramの例を示す。

例) 々する日が続きビ 00000001



2.3 学習段階の構成要素

読み付与段階の出力ファイル(漢字列とその読み列)を人間によってチェック・修正した後で、前述のファイル(1)を用いて単漢分解プログラムによって、漢字と読みの対応付けを行い、ファイル群(3)を作成する。また、必要に応じて、ファイル(2)、ファイル(1)を補充する。単漢分解では、後方より行い対応付けを行う漢字を含む3漢字に対して読みとの対応をとることにより正確さを向上させている。

3. 読み分け

「行っ(イツ、オコナッ)」および「通っ(トオッ、カヨッ)」について、本読み付与システムで、読み分けができるか実験を行った。実験に先立ちEDRの読み付きコーパス[2] および、かな漢字変換用辞書を用いて学習を行った。実験データは新聞記事[3]で1月から12月までの該当する漢字が現れたテキストを用いた。読み付与→学習を毎月行い、学習効果を観察することにした。表1および表2を参照されたい。

	正解率(%)	正解数	誤り数
Jan	90.12	374	41
Feb	92.98	358	27
Mar	90.52	430	45
Apr	92.09	408	35
May	92.87	378	29
Jun	94.85	406	22
Jul	91.23	406	39
Aug	94.88	371	20
Sep	90.50	343	36
Oct	93.42	426	30
Nov	94.96	396	21
Dec	93.75	390	26

表1. 「行っ」の読み分け

	正解率(%)	正解数	誤り数
Jan	74.75	77	26
Feb	75.64	59	19
Mar	78.82	67	18
Apr	78.82	67	18
May	88.23	75	10
Jun	85.88	73	12
Jul	89.41	76	9
Aug	84.28	59	11
Sep	90.76	59	6
Oct	87.09	81	12
Nov	86.20	75	12
Dec	88.67	94	12

表2. 「通っ」の読み分け

4. おわりに

環境文字(読み付き漢字N-gram)により、読み付与する実験システムの概要、読み付与段階・学習段階について述べた。「行っ」は92.68%の正解率で、「通っ」は84.04%の正解率で読み付与できた。しかしながら、学習効果はあまり働いていないようである。今後の課題として、つぎの2点を上げておく。

- (1) より正確な読み付与をするために、漢字ごとの環境文字の可変長化。
- (2) 人間の介在をなるべく少なくする学習方法の確立。

参考文献

- [1] 鳥原信一：漢字N-gramによる日本語テキストの読み付与
情報処理学会 第53回全国大会 3L-2
- [2] 日本電子化辞書研究所：EDR電子辞書，日本語コーパス JCO-V015 (1995)
- [3] 日本経済新聞社：新聞記事データベース