漢字Ｎ－Ｇｒａｍによる日本語テキストの読み付与

3Ｌ－2
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

日本語テキスト音声合成および墨字・点字変換プログラムにおいてより正確な読みを付与することはきわめて重要である．一般的には形態素解析の過程で読み付与を行うが，漢字のシソーラスを用いての読み付与[1]，読み付き漢字 Tri-gramにより，漢字未知語の読みの推定[2]などの研究もなされている．漢字の読みは，一意に決定できるもの，左側文字によって決定できるもの，右側文字によって決定できるもの，両側文字によって決定できるものがある．さらに，環境文字を増やすことにより読みを絞り込めるものがある．本稿では漢字 N-gram の読み付与・読み分けに対する有効性について検討をする．

２．読み付きコーパスの単漢分解

ＥＤＲ[3]の読み付きコーパス，1,236,390文を用いて，漢字と読みの対応をとるために 7,058漢字の読み付きテーブルを参照しながら，単漢分解を行った．これにより，漢字と読みの対応がとれたのは，598,341文48.4%であった．単漢分解できないものには，当て字，中国語の発音など固有名詞が大半である．

(例)早稲田(ワセダ),小百合(サユリ),広東(カントン)

このような単漢分解できないものに関する取扱いはきわめて大きな課題である．

３．単一読み漢字

7,058漢字の読み付きテーブルの中で，読みが単一のものがひらがな，カタカナも含め 406漢字ある．

(例)β(ベータ),燐(リン),俣(マタ)

これらの漢字は問題なく読みを付与できることになる．

４．Bi-gram

左側または右側に環境文字１文字を置くことにより，読みが付与できるものがある．接頭語または複合語の一部が左にあると右側が決定される．複合語の一部，接尾語または付属語がくると左側が決定される．

(左側文字で読みが決定される漢字の例)
亜，唖，阿，挨，始

(右側文字で読みが決定される漢字の例)
哀，握，扱，宛，綾

５．Tri-gram と環境文字

両側１文字，左側２文字・右側１文字，両側２文字の Tri-gram を作成した．

(両側文字1文字で読みが決定される漢字の例)
愛，悪，囲，易，為

(左側2文字右側1文字で読みが決定される漢字の例)
圧，暗，位，異，違

(両側文字2文字で読みが決定される漢字の例)
安，羽，雨，越，夏

ここまでの N-gram の結果を表１に示す．「複数読みの漢字」とは，与えられた環境が同一で読みが特定できなく複数ある漢字を表している．

(例)大-家-に(ヤ),大-家-に(カ)

環境文字を増やすことにより複数読み漢字数が減少していくことがわかる．

Japanese Yomi-setting by Kanji N-gram
Shinichi Torihara
IBM Research,Tokyo Research Laboratory,IBM Japan

N-gramの種類	1	1+	+1	1+1	2+1	2+2
ｷｰ漢字数	7,058	2,992	3,059	3,072	2,989	2,975
ﾚｺｰﾄﾞ数	－	164,664	141,891	506,079	1,140,261	1,565,077
複数読みの漢字数	6,652	1,326	1,123	539	252	153
複数読みのﾚｺｰﾄﾞ数	－	38,438	23,385	7,720	3,089	1,439

表１．漢字 N-gram と複数読み

６．順次的 N-gram

表１のレコード数が，環境文字を増やすにつれて増加していくことがわかる．本研究では，前の N-gramで読みが単一に絞れなかった「複数読み漢字」のみを次のN-gramの対象漢字とした．つまり，1 -> 1+ -> +1 ->1+1 -> 2+1 -> 2+2 の順序で行う．これにより，レコードの膨張を防ぎ，早い段階で読みが特定できると考えられる．表２にこれらを示す．

N-gramの種類	1	1+	+1	1+1	2+1	2+2
対象漢字数	7,058	6,652	1,326	965	527	250
ｷｰ漢字数	6,652	2,992	1,321	965	527	250
ﾚｺｰﾄﾞ数	－	164,664	113,298	376,448	619,902	573,930
複数読みの漢字数	6,652	1,326	965	527	250	152
複数読みのﾚｺｰﾄﾞ数	－	38,438	22,521	7,680	3,083	1,437

表２．順次的 N-gram

７．読み分けと読みの揺れ

表２の 2+2の N-gram の複数読み漢字のデータを観察すると，いわゆる「行った（イッタ・オコナッタ）」といった読み分けしたい漢字がある．これらの漢字は，手がかり単語が隣接している場合も離れている場合もあり格関係，共起・シソーラス情報など別の手法と組み合わせることが必要である．

また，「見え隠れ（ミエカクレ・ミエガクレ）」といった読みの揺れも多く含まれている．これらのデータは，一意に決定できるルールを発見したいと考えている．

８．おわりに

読みが単一の漢字があり，環境文字を増やしてN-gramを作成すると複数読みを持つ漢字数が減少する．また，前のN-gramの複数読み漢字のみを対象とする順次的N-gramによると早い段階で漢字の読みを絞り込めることがわかり，漢字N-gramは，読み付与に対し有効であると言える．すなわち，どの漢字がどのN-gramで読みが決定できるかわかる．本方式により3,398漢字に対し95,5%の精度で読み付与ができた．しかしながら，つぎの問題があることが明らかになった．

(1) 漢字と読みとを対応づける単漢分解に失敗する漢字．
(2) 読みの揺れを持つ漢字の読み統一．
(3) 本方式での読み分けの限界．

これらの問題は，日本語テキストへの読み付与において基本的なものであり，他の手法と組み合わせて解決したいと考えている．

参考文献

[1]鈴木,中挟知,近藤,佐藤,島田:漢字シソーラスの構築と語句解析への応用
第52回情報処理学会全国大会4B-6(1996)
[2]鈴木,鳥原,斉藤:日本語テキスト音声合成のための言語処理の検討　情報処理学会　音声言語情報処理研究会11,pp.1-5(1996)
[3]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)