Ｎ－ｇｒａｍによる同形語の読み分け

２Ｊ－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

テキスト音声合成において、同形異音語（ homography (例)「wind /wind/, /waind/」、「行っ /イッ/, /オコナッ/」）を読み分けることは、読み付与の精度向上の観点からきわめて重要である。

先行研究において、N - gram による読み付与の可能性についての基礎的調査がなされている [1] [2]。また、我々は実際に実験システムを作成して評価を行っている [3]。 同形異音語の読み付与（読み分け）も同一の枠組みで行っていたが、読みを決定する手がかり語（文字列）が、挿入句および長い付属語により、参照できないことがあった。 そこで、隣接の N - gram に加え、非隣接の N - gram を作成して手がかり語に参照するようにした。 評価実験の結果、最大０．５９％の精度向上が得られたので報告する。

２．同形異音語 (homography)

現代言語学辞典 [4] の「homograph(同綴異義語)」の項に次のような定義が見られる。

1. 綴りと発音が同じで意味が異なる語。
   bear/bɛər/(支える)、bear/bɛər/(熊)
2. 綴りが同じで発音と意味が異なる語。
   slough/slau/(泥沼)、slough/slʌf/(ぬけがら)

日本語は表意文字なので、1. の例は存在しないと思われる。 ただし、ひらがな表記をすると1. に該当するものがある。

(例) はし(橋、端、箸)

本研究では、読み分けの観点から2. を問題とする。 本論文では、これを同形異音語または単に、同形語と呼ぶことにする。

３．日本語における同形異音語

単語は、表出によって物事を指示したり、意味したりするという定義に基づいて、日本語処理用辞書の一般語から複数読みのあるものを抽出し、視察により同形異音語の絞り込みを行った。 すなわち、指示物が異なったり、意味するところが異なる単語のみを残した。 その結果、つぎのようになった。

721(同形異音語) / 4443(一般語の複数読み単語)

実際には、固有名詞は固有の指示物を指示するので、固有名詞も含めた分類が必要である。

(例) 一(いち、はじめ)

これらの同形異音語に対して、方法の改良、学習などにより、正しい読み付与を行う必要がある。

Reading homography by N-gram

Shinichi Torihara, IBM Research Laboratory, IBM Japan

4．隣接・非隣接環境文字による読み分け

従来は、０から４の左右環境文字列によって読みを付与していたが、前方には挿入句、後方には長い付属語がくる可能性があり、４文字の環境文字列では手がかり語（文字列）に届かず、ローカルに決定されていた。そこで、非隣接の、０から４の左右環境文字列の N - gram を作成し、これにも参照するようにした。図１に、非隣接な手がかり語への参照について示す。

それは一方通行を誤って通ってしまったことを笑っているようでした

←==→ xカヨ

←―― ==→ トオ

僕が通ったのは中学

←==――→ xトオ

←==― ―→ カヨ

図１．非隣接の手がかり語への参照

5．実験結果

表１．に実験結果を示す。

	隣接N-gram	隣接・非隣接N-gram
行っ	96.20%	96.20%
通っ	92.76%	93.24%
降り	94.42%	94.42%
金	94.81%	95.40%

表１．隣接 N-gram と隣接・非隣接 N-gram の実験結果

6．おわりに

同形語の読み分けがテキスト音声合成において重要であることを述べた。　そして、隣接に加え、非隣接の左右の環境文字に参照することにより、最大０．５９％の精度が向上した。　今後は、つぎに挙げる項目の検討・研究を進めるつもりである。

・テストケースによっては、非隣接文字列を参照することにより精度が下がる場合もあるので、さらに制御が必要である。

・数字・数字の前後文字列およびアルファベットの読み付与。

・漢字列と N - gram によって得られた読み列を入力とする読み誘導形態素解析の研究。

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(1996)

[2] 伊藤,他：N-gram を用いた言語コーパスへの読みの付与：日本音響学会春季研究発表会(1997)

[3] 鳥原信一：漢字 N-gram を用いた読み付与システム：情報処理学会第54回全国大会(1997)

[4] 田中春美編：現代言語学辞典，成美堂(1988)