音声認識のための定型表現を用いた言語モデルの検討

西崎 博光　中川 聖一
豊橋技術科学大学 情報工学系
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

1 はじめに

近年の著しい計算機速度の向上、及び、音声処理技術/自然言語処理技術の向上により、ディクテーションシステムなど、音声認識技術が実用的なアプリケーションとして社会に受け入れられる可能性がでてきた [1]。一方、従来のテレビ放送のような情報伝達形態では、映像と音声の両媒体を組み合わせて、意味のある情報伝達/受容が可能となる場合が非常に多い。この場合、聴覚に障害を持つ者にとっては不完全な情報受容を強制されることとなる。特にニュース番組では、録画放送といった形態はその性質上避けるべきであり、また、放送の進行と同時に字幕を人手で入力するのは不可能に近い。かつ、放送原稿は番組の直前に作成されるなどの事情を考えると、アナウンサの発話をそのまま音声認識し、字幕へと変換することが望ましい。そこで、最近になってテレビニュース音声の認識が研究されている [2]。

このような背景を踏まえ、本稿ではニュース音声に対し、音声認識用の精度の良い言語モデルの構築を実験的に検討した。音声認識のための N-gram 言語モデルでは、N=3 で十分であると考えられる [3]。しかし、N=3 ではパラメータの数が多くなり、音声認識時の負荷が大きい。そこで、第 1 パス目は N=2 の bigram モデルで複数候補の認識結果を出力し、N=3 の trigram で後処理を行なう方法が一般的である。本稿では、第 1 パス目の bigram 言語モデルの改善を目指した。

音声認識において、認識単位が短い場合認識誤りを生じやすく、付属語においてその影響は大きいと考えられ、小林らは、付属語列を新たな認識単位とした場合の効果の検証をしている [4]。また高木らは、高頻度の付属語連鎖、関連率の高い複合名詞などを新しい認識単位とし、これらを語彙に加えることによる言語モデルの性能に与える影響を検討している [5]。なお、連続する単語クラスを連結して一つの単語クラスとする方法や句を一つの単位とする方法は以前から試みているが、いずれも適用されたデータベースの規模が小さい [6, 7]。

ニュース原稿には、通常の音声対話と比較して、使用頻度の高い (特殊) 表現や、固定的な言い回しなどの表現 (以下、定型表現と呼ぶ) が非常に多い。定型表現は、音声認識用の言語モデルや音声認識結果の誤り訂正のための後処理に適用できる。そこでまず、定型表現を抽出した。次に、これらの (複数形態素から成る) 定型表現を 1 形態素として捉えた上で、N-gram 言語モデルを構築する方法を提案する。評価実験の結果、長さ 2 および 3 以下である定型表現を 1 形態素化して bigram, trigram 言語モデルを作成することで、bigram に関しては、エントロピーが小さくなり、言語モデルとして有効であることを示す。

2 言語モデルの評価基準

2.1 エントロピーとパープレキシティ

言語モデルの評価基準として、エントロピーとパープレキシティを用いる。エントロピーとパープレキシティは共に、対象とする文集合の複雑さを定量的に示す指標で、その文集合が複雑なほど、それぞれの値は大きくなる。

単語列を生成する情報源をモデル化したものを言語モデルと呼ぶ。いま言語 L において、文 (単語列)$W_i = w_1 \cdots w_{L_i}$ の出現確率を $P(W_i)$ とすれば、文集合 $W_1, W_2, \cdots, W_N$ のエントロピーは次式で求められる。

$$H(L) = -\sum_{i=1}^{N} P(W_i) \log P(W_i)$$

テキスト文の連接を $W = W_1 W_2 \cdots W_N = w_1 w_2 \cdots w_T$ とすれば、テストセットのエントロピーは

$$H(L) = -\log P(W)$$

で示される。トライグラムを用いた場合、P(W) は

$$\begin{aligned} P(W) &= P(W_1)P(W_2)\cdots P(W_N) \\ &= P(w_1|*\ \#)P(w_2|\#\ w_1) \\ &\quad P(w_3|w_1\ w_2)\cdots P(w_T|w_{T-2}\ w_{T-1}) \end{aligned}$$

となる。(注:#は文頭を、*は文末を示す)

この時、一単語当たりのエントロピーは

$$H_0(L) = -\frac{\sum_i \log P(W_i)}{\sum_i L_i}$$

また、言語の複雑さ・パープレキシティは

$$PP = 2^{H_0(L)}$$

と定義される。

パープレキシティは、情報理論的にある単語から後続可能な単語の種類数を表している。この値が大きくなるほど、単語を特定するのが難しくなり、言語として複雑であるといえる。また逆に、この値が小さくなるほど、音声認識での後続予測単語を特定するのがやさしくなるので、認識率が上がる傾向にある [8]。

2.2 補正パープレキシティ

本稿で使用した CMU SLM toolkit[9] では語彙に含まれないものは全て一つの未知語のカテゴリにまとめら

れ、語彙に含まれる形態素と等価に未知語のカテゴリは扱われる。そのため語彙サイズのセットが小さい程(カバー率が小さい程)、パープレキシティは小さくなるということになり好ましくない。そこで評価テキスト中に出現した未知語の種類 m と、未知語の出現回数 n_u を用いてパープレキシティを補正する。補正パープレキシティは

$$APP = (P(w_1...w_n)m^{-n_u})^{-\frac{1}{n}}$$

で与えられる。これは、複数の未知語はそれぞれ等確率に生じると仮定して、補正したものである。

3 定型表現

ニュース原稿には、通常の音声対話と比較して、非常に定型表現が多いことに着目し、これらの高頻出定型表現を1形態素として捉えた上で、言語モデルを構築すれば、より精度の良いモデルが出来ると考られる。

今回、使用頻度の高い定型表現を抽出するアルゴリズムとして、池原、白井らの提案した方法 [10] を用いた。この方法では、最長一致の文字列抽出 (ある文字列が抽出されたとき、その文字列に含まれる部分文字列は統計量を求める際にはこの部分文字列を定型表現とはカウントしない) を条件とし、任意の長さ以上、任意の使用頻度以上の表現を、もれなく自動的に抽出する。文献 [10] では文字列単位で抽出していたが、これを形態素単位で抽出するようにした。抽出例を表 1に示す。

表 1: 定型表現抽出例

連語数	定型表現 (頻度)
2	い/ます (131793) こと/です (18499) 日本/の (11485)
3	して/い/ます (15171) これ/まで/の (7039) これ/に/対して (5357)
5	こと/に/して/い/ます (2497) こと/を/明らかに/し/ました (621) たい/と/話して/い/ます (531)

4 言語モデルの構築

4.1 標準言語モデル

標準言語モデルは、表 2に示す学習用データから作成した。まず、JUMAN[11] を用いて形態素解析を行ない、出現頻度が上位 2 万番目までの形態素を語彙として辞書に登録した。言語モデルの構築には、CMU SLM Toolkit Ver.1[9] を用いた。バックオフ・スムージングには Good-Turing 推定を用いた。

表 2: 標準言語モデルの仕様

コーパス	NHK ニュース原稿
学習用データ	1992 年 7 月~1996 年 5 月分 総形態素数 : 21177320 異なり形態素数 : 117183
評価用データ	1996 年 6 月分 総形態素数 : 962963 異なり形態素数 : 27142
語彙サイズ	20000
言語モデル	bigram,trigram

4.2 定型表現を用いた言語モデル

言語モデル構築のための手順を以下に示す。

Step.1 連語数 2 または 3 の定型表現を抽出する。

Step.2 各形態素の頻度リストを求める。

Step.3 定型表現を 1 つの新しい形態素としてまとめる。

Step.4 高出現頻度順に 20000 形態素を求め、語彙サイズ 20000 の辞書を作成する。

Step.5 言語モデルを構築する。

4.2.1 定型表現抽出

JUMAN[11] を用いて形態素解析を行なったニュース原稿に対して、定型表現抽出プログラムを実行し、連結数 2 または 3 の定型表現を抽出する。

4.2.2 頻度の計算

定型表現を用いる前のトレーニングデータから、各形態素の頻度リストを求める。上位 15000 番目くらいの形態素の出現頻度が 50 回であるので、定型表現の出現頻度が 50 回以上のものを新しい形態素候補として用いることにする。

4.2.3 定型表現の連結

Step.2 の定型表現を用い、トレーニングデータ内の定型表現を図 1のように 1 つの単語にまとめる。

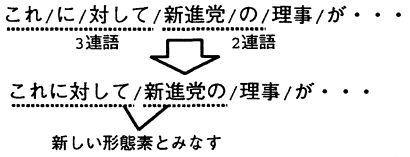

図 1: 形態素の連結例

4.2.4 語彙サイズ 20000 の辞書作成 (1 回目)

トレーニングデータから出現頻度の多い順に 20000 を求め、語彙サイズ 20000 の辞書を作成する。このとき、上位 20000 の辞書に登録された定型表現は 2 連結で 9514 個、3 連結で 8773 個 (2 連語:5497、3 連語:3276) である。登録されなかった定型表現が多数あるので、これは未知語の数を増やすだけなので図 2 のようにもとの形態素に分解しておく。

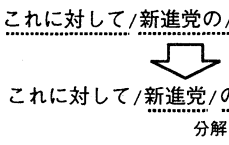

図 2: 形態素の分解例

4.2.5 語彙サイズ 20000 の辞書作成 (2 回目)

分解後のトレーニングデータから、もう一度語彙サイズ 20000 の辞書を作成する。これは、4.2.4 節で定型表現を分解したことによって形態素の出現頻度が変わってしまうためである。当然ここでも登録されない定型表現がでてくる。登録された定型表現は 2 連結で 9075 個、3 連結で 8702 個 (2 連語:5479、3 連語:3223) になった。ここでも、登録されなかった定型表現はもとの形態素に分解する。

厳密に行なうなら、辞書作成と定型表現の分解といった作業を繰り返し行ない、辞書に登録される定型表現の数が収束するまで行わないといけないが、今回は 1 回だけしか行なっていない。

4.2.6 言語モデルの構築

CMU SLM Toolkit[9] を用いてトレーニングデータから、語彙サイズ 20000 の辞書を作成し、bigram,trigram 言語モデルを構築する。

4.2.7 評価

評価を行なう時、注意しなければならないことは、いずれの比較対象に対しても同じ定義の 1 形態素あたりのパープレキシティを求めないといけないということである。通常パープレキシティを求める式は **bigram** の場合で、

$$PP_0 = \sqrt[M]{\prod_{i=1}^{M} P(w_i|w_{i-1})^{-1}} \qquad (1)$$

であるが、これは 1 連結形態素 (定型表現として形態素を連結したもの) あたりのパープレキシティを求めている。形態素を連結する前の従来の 1 形態素あたりのパープレキシティを求めるには、

$$PP_1 = \sqrt[N]{\prod_{i=1}^{M} P(w_i|w_{i-1})^{-1}} \qquad (2)$$

表 3: 語彙サイズの増加による定型表現の評価結果

(a) 語彙サイズ 22000、定型表現を含まない

データセット		トレーニングデータ	テストデータ
bigram	PP	71.90	90.23
	APP	89.20	108.47

(b) 語彙サイズ 22000、定型表現 2000 個を含む

データセット		トレーニングデータ		テストデータ	
定型表現の連結数		2 連結	3 連結	2 連結	3 連結
bigram	PP	50.49	49.79	71.93	74.34
	APP	64.59	63.69	88.33	91.98

(c) 語彙サイズ 25000、定型表現を含まない

データセット		トレーニングデータ	テストデータ
bigram	PP	72.34	86.90
	APP	91.71	107.74

(d) 語彙サイズ 25000、定型表現 5000 個を含む

データセット		トレーニングデータ		テストデータ	
定型表現の連結数		2 連結	3 連結	2 連結	3 連結
bigram	PP	43.31	40.89	67.57	69.24
	APP	55.52	52.45	82.99	84.93

を用いなければならない。ここで

M:定型表現を 1 つの形態素としたときの連結形態素と従来の形態素の総数

N:定型表現を使わなかったときの従来の形態素の総数

5 評価実験

標準言語モデルと、前節に述べた方法で定型表現を用いた言語モデルを構築し、その評価を行なった。トレーニングデータには、標準言語モデル作成の場合と同じ、表 2の学習用データを用いている。テストデータには、標準言語モデル、定型表現を用いた言語モデルとし、ともに表 2の評価用データを使用した。

実験結果を図 3に示す。まず、bigram モデルでは、トレーニングデータに関しては約半分、テストデータに関しては約 3 割、パープレキシティが減少しているのがわかる。しかし、trigram モデルではトレーニングデータでは効果があったが、テストデータに対しては効果が得られなかった。これは、トレーニングデータのパープレキシティとテストデータのパープレキシティの差が大きいことからも、トレーニングデータの不足であると考えられる。

次に、標準言語モデルを作成した時の語彙サイズ 20000 の辞書に、2 および 3 連結の定型表現をそれぞれ高出現頻度順で上位 2000 個,5000 個分を追加した場合の辞書で言語モデルを構築した。その評価結果を表 3に示す。この言語モデルの場合でも、パープレキシティの改善が見られた。定型表現 5000 個の方が、定型表現 2000 個のものに対して、語彙サイズが大きいのにも関わらず、パープレキシティが減少している。図 3と表 3から、定型表現の数が多い方が良い、つまり出来るだけ多くの定型表現を辞書に登録すれば良いということが言える。

6 まとめ

本稿では、ニュース原稿から抽出した定型表現を用い、N-gram 言語モデルを構築する方法を検討した。実験の

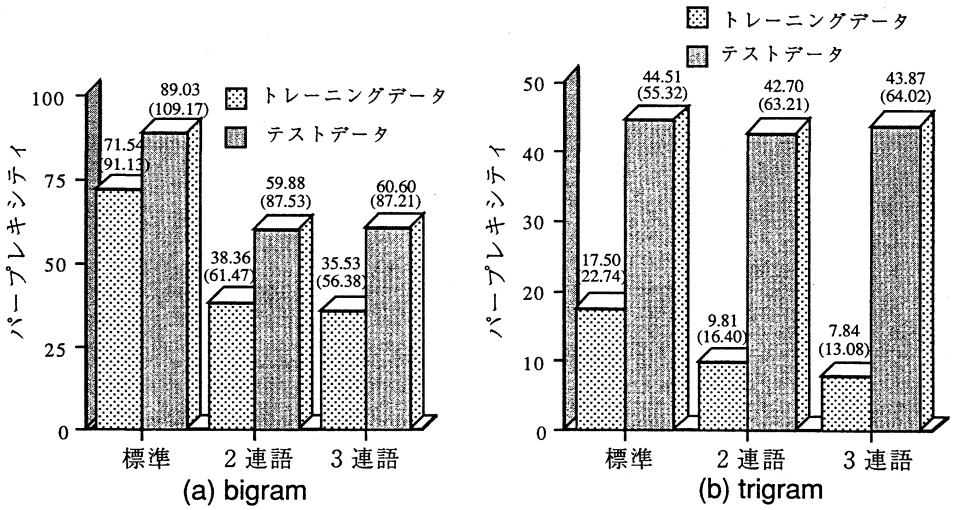

図 3: 定型表現の評価結果 [注:() 内の数値は補正パープレキシティを示す]

結果、定型表現を用いた言語モデルを作成することで、bigram モデルに関しては、テストデータに対し約 3 割程度パープレキシティを低く押えるとこができ、言語モデルの有効性を示すことができた。しかし、trigram ではトレーニングデータの量が不十分だったため、トレーニングデータでは効果があったがテストデータに対しては効果が得られなかった。そこで、トレーニングデータの量をもっと増やし、本方法の有効性を調べたいと考えている。

本稿では、定型表現の登録はヒューリスティックに行なったが、パープレキシティ(エントロピー) 最小基準で登録する方法 [6, 7, 12] を試みたい。

謝辞

この研究は、NHK 放送技術研究所との共同研究の一部として行なわれた。ニュース原稿データベースを提供された NHK 放送技術研究所の関係諸氏に深く感謝する。ニュース原稿の形態素解析には JUMAN、言語モデルの構築には CMU SLM Toolkit を利用させていただいた。

参考文献

[1] 西村雅史, 伊藤伸泰, 山崎一孝, 萩野紫穂:単語を単位とした日本語大語彙連続認識, 日本音響学会講演論文集,3-1-5(1997.9)

[2] 小林彰夫, 今井亨, 安藤彰男, 宮坂栄一, 赤松裕隆, 中川聖一, 小黒玲, 尾関和彦, 古井貞照, 鈴木順子, 白井克彦:ニュース音声認識システムの検討, 日本音響学会講演論文集,3-1-9(1997.9)

[3] 大附克年, 松岡達雄, 吉田航太郎, 古井貞照:高次 n-gram を用いた大語彙連続音声認識の検討, 日本音響学会講演論文集,2-6-2(1997.3)

[4] 小林紀彦, 中野裕一朗, 肥田木康明, 小林哲則:統計的言語モデルにおける付属語の扱いに関する一考察, 日本音響学会講演論文集, 2-1-6,pp.59-60(1997.9)

[5] 高木一幸, 小黒玲, 橋本顕示, 尾関和彦:ニュース音声認識における言語モデルの検討, 情報処理学会, 音声言語情報処理,19-12(1997.12)

[6] E.P.Giachin:Phrase bigrams for continuous speech recognition,Proc.ICASSP,pp.225-228(1995)

[7] 政瀧浩和, 松永昭一, 匂坂芳典:連続音声認識のための可変長連鎖統計言語モデル, 電子情報通信学会, 音声技報,SP95-73(1995.11)

[8] 中川聖一:情報理論の基礎と応用, 近代科学社 (1992)

[9] R.Rosenfeld:The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation, Proc. ARPA Spoken Language Systems Technology Workshop,pp.47-50(1995)

[10] 池原悟, 白井諭, 河岡司:大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌, Val.36, No.11, pp.2584-2596(1995)

[11] 松本裕治, 黒橋禎夫, 山地治, 妙木裕, 長尾真:日本語形態素解析システム JUMAN Ver3.1 使用説明書, 京都大学工学部長尾研究室 (1996)

[12] 森信介, 山地治, 長尾真:予測単位の変更による n-gram モデルの改善, 情報処理学会, 音声言語情報処理,19-14(1997.12)