音声言語情報処理 35－8
（2001．2．3）

Webからの音声認識用言語モデル自動生成ツールの開発

西村 竜一＊　長友 健太郎＊　小松 久美子＊＊　黒田 由香＊＊＊
李 晃伸＊　猿渡 洋＊　鹿野 清宏＊

＊ 奈良先端科学技術大学院大学 情報科学研究科
＊＊ イメージ情報科学研究所　＊＊＊ TIS 株式会社

あらまし　本報告ではWebページからの音声認識用N-gram言語モデルの自動作成ツールの開発について述べる．言語モデルの作成は大量のテキストが必要で高いコストを要する．また，ユーザの使用する語彙は変化するため，常に新しい言語モデルを作成する必要がある．本ツールでは，大量に存在するWebページからキーワードによる検索を利用して，タスクに応じたテキストの収集を行ない，タスク適応した言語モデルを手軽に作成できる．また，Webページは更新されるので，新しい語彙を持つ言語モデルの更新ができる．さらに文字パープレキシィを評価基準とするテキスト整形手法の検討をした．実験では，本ツールを用いて「医療」をキーワードとして言語モデルを作成した．その結果，健康相談タスクに対して新聞記事による言語モデルに比べ約9%の認識率の向上が得られた．

キーワード　大語彙連続音声認識，N-gram言語モデル，Web

Automatic Language Model Building Tool using Web Texts

Ryuichi NISIMURA＊　Kentaro NAGATOMO＊
Kumiko KOMATSU＊＊　Yuka KURODA＊＊＊
Akinobu LEE＊　Hiroshi SARUWATARI＊　Kiyohiro SHIKANO＊

＊ Graduate School of Information Science, Nara Institute of Science and Technology
＊＊ Laboratories of Image Information Science and Technology　＊＊＊ TIS Corp.

Abstract　This paper describes a tool to automatically build N-gram language models for speech recognition using Web texts. It takes high costs to make language models, because huge texts are required. The performance of old corpus-based-model is not enough, because newer words are often treated as unknown words. By collecting large amounts of Web texts for a specific task via key-word based Web search and building task adapted language model, users can get new vocabularies and put them into the models. We also develop a text filtering algorithm based on character perplexity for full automatic building. On a dictation task of medical consulting, the word recognition rate of the language model made by the proposed tool is 9% higher than a newspaper model.

Key words　Large Vocabulary Continuous Speech Recognition, N-gram Language Model, Web

1　はじめに

ここ数年の大語彙連続音声認識の発展はめざましく，音声インターフェイスを組み込んだアプリケーションソフトウェアも登場しつつある．これは，認識アルゴリズムの進化や計算機の高性能化の他にもテキストコーパスや音声データなどの学習用データベースの整備がすすんだことの影響も大きい．しかし，データベースの作成は，非常に高いコストの作業が必要であり，ユーザがモデルを自由に作成するのは依然として困難である．

言語モデルの学習には，大量のテキストコーパスが必要である．これまでは大量のデータを整備し易い新聞記事からモデルを作成することが多かった．しかし，特定の目的で音声認識を利用する場合，新聞記事による言語モデルを利用するよりも，それぞれのタスクに適応した言語モデルを利用した方が良い．そのためにはタスクに特化したテキ

ストコーパスでモデルの学習をする必要がある．しかし，テキストを大量に収集，整備するのは非常に困難な作業である．

また，ユーザの使用する語彙は常に変化するものであり，数年前のテキストより作成した言語モデルでは，新しい語彙に未知語になるものがある．未知語の音声認識は困難なので，ユーザの使用語彙に対して，未知語が多いと実用的な認識性能を得ることは難しい．このため，テキストコーパスとモデルの更新が必要となる．

そこで，本報告では，膨大な量のテキストを持つインターネット上の Web ページから音声認識用 N-gram 言語モデルを手軽に構築するツールの開発について述べる．本ツールは，キーワードによる Web ページ検索を利用して，タスクに特化したテキストを収集することで，タスクに応じたモデルを自動作成できる．また，Web ページは，頻繁に更新されるので，新しい語彙を持つテキストを得てモデルの学習に利用できる．よって，最新の情報を元にした言語モデルの更新が行なえるなどの特徴を持つ．

2 Web からの言語モデルの作成

本研究では，Web ページを用いて言語モデルを作成する手法を提案する．WWW（World Wide Web）はインターネットでもっとも良く利用されているサービスであり，今では Web ページの量は膨大なものになる．これらを収集してテキストコーパスを整備できれば，言語モデルの学習に利用できる．また，Web ページのテキストは，新聞記事に比べてやわらかい会話調のものが多く，話し言葉にもある程度対応できる．

以下では，Web ページからのテキストの収集方法の検討を行ない，その検討を元に言語モデル自動作成ツールの開発をすすめる．

3 テキスト収集方法の検討

Web ページからのテキストの収集の方法を 3 つ検討する．今回の実験でのモデルの作成には，前 2 つの方法を用いた．それぞれの方法で作成したモデルを以下では，モデル A，モデル B と呼ぶ．

3.1 掲示板からの手動収集（モデル A）

Web には，特定のトピックに基づいた発言がされている電子掲示板のページが存在する．商用サービスなどの有名な掲示板では，参加者が多く，多くのテキストが登録されている．これらのテキストを収集することで，トピックに対応したテキストコーパスの作成ができる．

この方法の利点は，比較的，整理されたテキストで書かれていることが多いことである．また，トピックに基づいて発言がなされているので，その話題の範囲が特定しやすい．よって，タスクに応じたテキストを効率良く収集できる．

欠点として，人手で適切な掲示板を探す必要があり，自動化には向かない．また，収集できるテキストの量に限界があり，大量のテキストの収集は困難である．

3.2 キーワード検索を用いた自動収集（モデル B）

インターネット上では数多くの Web ページの検索サービスが提供されている．これらは入力したキーワードに関連の高い Web ページへのリンクを提供するサービスである．このサービスを利用して，音声認識のタスクから連想されるキーワードを与えることで，タスクに関連性のあるテキストを大量に自動収集できる．

この方法の利点は，大量のテキストを自動的に収集できることにある．実際の検索サービスでは，検索結果の数に限りがあるので，結果の各 Web ページのリンク先をさらに取得することにより，関連する Web ページをさらに大量に取得できる．

しかし，検索結果やリンク先のページは，内容がタスクに関連する Web ページであるという保証はない．よって，想定したタスクと関連性の低いテキストも収集してしまう可能性が高いという問題点もある．また，この方法で取得した Web ページは，一般に多種多様のため，内容や記述方法が統一されていない．

3.3 アクセスログを用いた収集

個人やサイトごとの Web のアクセスログを利用して Web ページを収集することで，個人やネットワーク内のユーザを対象としたタスク向けのテキストコーパスを作成できる．

しかし，同じネットワーク内の複数ユーザのアクセスログを収集するには，サイト内に設置された Web キャッシュ（Proxy）サーバのアクセスログを利用できるが，個人ユーザのアクセスログを収集する方法は別途検討が必要である．

また，ユーザ個人や少数ユーザの環境では，モデル作成に十分な量の Web ページを取得するために必要なアクセスログを収集するのに長い期間が必要である．このため，短期間のアクセスログより取得した少数のテキストから，高性能な言語モデルを作成する手法が必要になる．

4 テキスト整形フィルタの検討

これまで新聞記事などによる言語モデルの作成では，ヘッダや記号などの文章のコンテクストに関係ない文字をテキスト整形フィルタを用いて，あらかじめ削除してきた．これには言語モデル作成

時の語彙数の制限による未知語率の上昇を防ぐなどの言語モデルの性能向上に効果がある．本節では，このテキスト整形フィルタについて検討する．

4.1 固定ルールフィルタの問題点

3.1節で述べたモデルA（手動収集）で使用するテキストについては，従来からの正規表現などで経験的にルールを定義するフィルタでテキストの整形を行なった．なお，以下では，この従来からのフィルタを固定ルールフィルタと呼ぶ．しかし，3.2節で述べたモデルBで使用するテキストでは，HTML（Hyper Text Markup Language）のタグのような明確にルール化された記号を除いては，固定ルールフィルタでの整形処理は難しい．

これは，Webページのテキストの記述の多様性が非常に高く，単純なルールでは大量のWebページの多様性をカバーしきれないためである．このため，固定ルールフィルタを利用するためには，Webページごとにルールを作成する必要がある．しかし，ルールの作成は，容易な作業ではなく，大量のWebページに固定ルールフィルタを利用するのは現実的ではない．

そこで，統計量を用いて判別処理をするフィルタの検討をした．以下では，これを統計ルールフィルタと呼ぶ．

4.2 統計ルールフィルタ

統計ルールフィルタは，新聞記事など整備されたテキストを学習した文字レベルや単語レベルなどの統計量で入力テキストの文章らしさを評価し，閾値によって選別することでフィルタリングを行なう．

統計量のモデルには表1に示す文字3-gramモデルを用いた．これを基準言語モデルと呼ぶ．

評価尺度には式1に表すパープレキシティを使用した．単語パープレキシティは，言語Lにおける各文字の後に続く可能性のある単語数の平均を表しており，言語Lを評価用のテキストとして，作成した言語モデルの性能の評価に使用されている．今回のようにフィルタとして利用する場合には，基準言語モデルに基づいた入力テキストの文章らしさの評価尺度と考えることができる．よって，基準言語モデルを日本語テキストで学習した場合，パープレキシティの値が小さいものを日本語文章らしいと判別できる．今回は形態素解析の影響を防ぐため文字パープレキシティを用いた．

なお，入力テキストの評価は，行単位に行ない，テキストの選別も行単位に行なった．

表1: 基準言語モデル

モデル構成	文字3-gram
学習テキスト	新聞記事1年分
学習テキスト容量	95.2MB
異なる文字数	4890文字

言語Lの文字あたりのエントロピー：

$$H(L) = -\frac{1}{n}\sum_{i=1}^{n} p(c_i|c_{i-1}c_{i-2}) \log p(c_i|c_{i-1}c_{i-2})$$

言語Lにおけるパープレキシティ：

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

式1: 文字パープレキシティ

5 言語モデルの作成と評価

3.1，3.2節で述べた方法でのモデル作成及び比較評価を行なった．今回，作成する言語モデルの想定するタスクは，健康相談である．

5.1 テキストの収集

モデルA インターネット上の商用プロバイダのWeb掲示板サービスから健康相談をトピックとするテキストを人手で収集した．HTMLのタグを削除した後，固定ルールフィルタによってヘッダやシグネチャ（署名）を中心にコンテクストに無関係な記号などを削除した．また，文字化けしたテキストも手動で削除した．表2に，その結果を示す．参考までに新聞記事1年分相当のテキストの値も合わせて掲載する．

表2: 手動収集したテキスト（モデルA）

テキスト容量	11.5MB
異なる語彙数	41973個
文章数	319498文
（参考値）新聞記事1年分	
テキスト容量	92.0MB
異なる語彙数	142202個
文章数	906106文

モデルB Webページのキーワードによる検索サービス[1]を用いてテキストを収集した．入力キーワードには，「医療」を使用した．さらに，検索結果からリンクをたどることによって2階層先までのWebページを収集した．はじめに固定ルールフィルタによりHTMLのタグの除去をした後に，

[1] http://www.google.com/

表 3: 自動収集したテキスト（モデル B）

閾値 P	P < 50	P < 100	P < 200	P < 400	P < 600	P < ∞
テキスト容量（MB）	97.6	148.0	179.6	195.6	200.9	212.0
異なる語彙数（個）	138287	209623	254077	277200	283986	272005
文章数	2093310	3402167	4161743	4681117	4874011	4473410

最寄りの内科を教えてください
お腹が痛いんですけど
風邪気味なので病院を教えてください

図 1: タスク内（健康相談）テキストの例

何かおいしいものを教えてください
煮物料理について教えてください
ケーキを買いたいんですけど

図 2: タスク外（グルメ・レシピ）テキストの例

4.2 節で述べた統計ルールフィルタを利用して，文字化けテキストや記号などの削除を試みた．文字パープレキシティの閾値 P の値には，50，100，200，400，600 を使用し，パープレキシティがこの値以上の行を除外した．また，取得した全てのテキストを利用したものも作成した（閾値: ∞）．収集結果を表 3 に示す[2]．

5.2 モデルの作成

収集テキストから言語モデルの作成には，CMU–Cam SLM Toolkit[1] を用いた．また，日本語形態素解析には，ChaSen 2.2[2] を利用した．さらに「日本語ディクテーション基本ソフトウェア（99 年度版）」[3] に含まれる読み変化プログラム ChaWan 及び数字読み付与プログラムを使用して読み仮名の付与を行ない，語彙辞書の作成をした．作成した言語モデルは，2-gram 及び逆向き 3-gram 言語モデルである．使用語彙は，収集テキスト中の出現頻度上位 2 万語である．

5.3 モデルの評価

単語パープレキシティ及び未知語率を用いて作成した言語モデルの評価を行なった．タスクは健康相談であり，図 1 の例のような評価用テキスト

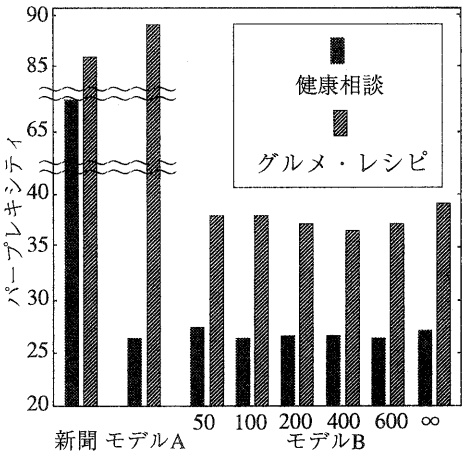

図 3: 3-gram 単語パープレキシティ

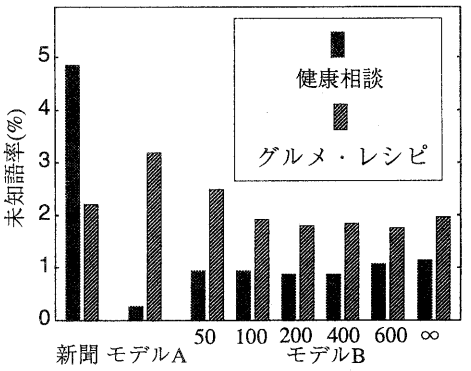

図 4: 未知語率

（150 文，総語彙数 1189 語）を使用した[3]．また，タスク外での傾向も調べるために図 2 の例のようなグルメ・レシピタスク（200 文，総語彙数 2046 個）での評価も行なった．

5.4 評価結果

図 3，図 4 に評価の結果を示す[4]．比較用に新聞記事 1 年分のテキストで作成した言語モデルの結

[2]閾値 P= ∞ の時に，異なる語彙数や文章数で他の閾値のものより小さい値になるのは，整形してないテキストの場合，プログラムが正しく処理を完了できないことがあるためである．

[3]評価用テキストは，実際の対話例を参考にして作成した比較的丁寧な話し言葉である．

[4]付録 A に詳細な結果を添付する．

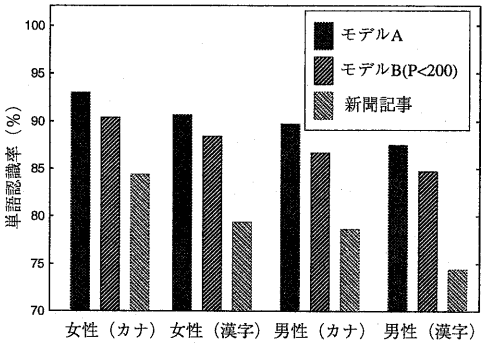

図 5: 単語認識率（健康相談タスク）

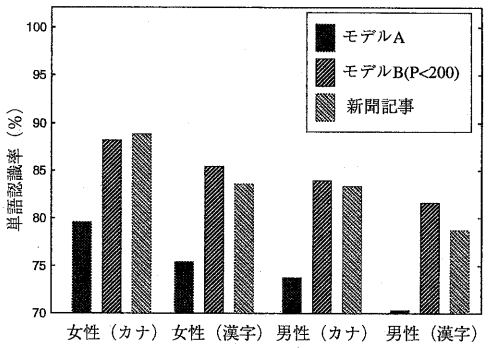

図 6: 単語認識率（タスク外 グルメ・レシピ）

果もあわせて掲載する．

結果より，モデル A，B ともに新聞記事モデルに比べて高い性能を示すことを確認できた．特にモデル A は，健康相談タスクに対して，未知語率が 0.27%になるなど，性能が非常に高い．モデル B に関しては，パープレキシティの値は小さく，性能は高いが，未知語率は 1.0%前後と大きく，モデル A に比べると性能が低いことがわかる．これは，3.2 節で述べたようにタスク外のテキストを学習テキストに含んでいることが原因であると考えられる．また，モデル B は，異なる語彙数が大きい．このため使用語彙が出現頻度上位 2 万語では十分ではない可能性がある．

タスク外であるグルメ・レシピタスクにおいては，モデル A，モデル B ともにタスク内評価よりもパープレキシティ，未知語率ともに性能が下がっており，本手法による言語モデルは，タスク適応をしていることがわかる．ただし，モデル B はモデル A と比較するとタスク適応度が低い．

モデル B に関して，統計ルールフィルタを適応することにより適応しないもの（$P<\infty$）と比較して未知語率を若干だが下げることができた．また，閾値 P の値が小さい時（P<50），学習テキスト数の量が十分でないため，パープレキシティが悪化する．この結果，閾値 P の値を 100 以上の適当な値を与えると良いことがわかる．しかし，今回の結果では，フィルタの影響は若干であり，今後も統計ルールフィルタの改良が必要である．

6 大語彙連続音声認識実験

モデル A，B の音声認識による評価を行なった．

6.1 実験方法

認識エンジンには，大語彙連続音声認識エンジン Julius[4] を用いた．また，音響モデルには，高齢者向け音響モデル（PTM[5]，2000 状態，64 混合，性別依存）[6] を用いた．ビームサーチ幅などのパラメータは，高精度認識用の設定に調整した．

評価用音声には，音響モデル同様に高齢者音声を用いた．話者は 60~90 歳の高齢者女性 50 人，男性 51 人で，それぞれが前述の評価用テキストの中から健康相談タスク 30 文，グルメ・レシピタスク 40 文を発話したものを使用した[5]．

6.2 結果

単語認識率の結果を図 5，図 6 に示す[6]．ただし，図 5，図 6 のモデル B に関しては，最も良い結果の閾値 P のものだけを掲載する．なお，カナ単語認識率は，認識結果をカナ（読み仮名）に変換して，単語認識率を求めたものであり，漢字仮名混じりの認識で起こる表記の違いを吸収したものである．

図 5 から，タスク内認識の結果において，モデル A，B ともに高い認識率を示していることがわかる．女性話者での認識で，新聞記事モデルと比較して漢字仮名混じり単語認識率でモデル A は約 11%，モデル B は約 9%の認識率の向上が得られた．よってモデル A，モデル B ともに Web ページからの言語モデルの作成が有効であることを確認できた．

図 6 のタスク外での結果では，モデル A の認識率が低く，モデル A のタスクへの適応度が非常に高いことがわかる．モデル B は，学習テキストの量が多いため，新聞記事モデルと同等の認識率であった．

7 おわりに

今回の実験では，モデル A に比べてモデル B は，若干の性能低下が見られた．しかし，モデル B でも性能は高く，その有効性を確認できた．また，3.1 節で述べたように，モデル A はテキストを収集するのに必要な作業量が多く，ツールとし

[5]収録の関係で発話文が少ない話者もいる．
[6]付録 B に詳細な結果を添付する．

付録 A: 言語モデルの評価

	新聞	モデル A（手動収集）	モデル B（Web, 自動収集）					
			$P < 50$	$P < 100$	$P < 200$	$P < 400$	$P < 600$	$P < \infty$
評価文: 健康相談タスク								
PP（2-gram）	65.50	32.93	37.16	37.05	37.23	37.50	37.02	37.15
PP（3-gram）	69.39	26.38	27.42	26.38	26.63	26.65	26.38	27.14
未知語率（%）	4.84	0.27	0.94	0.94	0.87	0.87	1.07	1.14
評価文: タスク外（グルメ・レシピ）								
PP（2-gram）	100.71	93.31	57.38	57.37	57.52	57.37	58.14	59.56
PP（3-gram）	85.90	89.73	38.01	37.99	37.20	36.56	37.20	39.12
未知語率（%）	2.21	3.19	2.49	1.92	1.80	1.84	1.76	1.96

付録 B: 単語認識率（単位: %）

	新聞	モデル A（手動収集）	モデル B（Web, 自動収集）					
			$P < 50$	$P < 100$	$P < 200$	$P < 400$	$P < 600$	$P < \infty$
単語認識率	発話文: 健康相談タスク							
女性（カナ）	84.39	92.98	90.22	90.06	90.42	90.17	90.16	89.76
女性（漢字）	79.32	90.65	88.14	88.14	88.40	88.14	88.15	87.84
男性（カナ）	78.62	89.72	86.91	86.55	86.69	86.53	86.53	85.90
男性（漢字）	74.43	87.50	84.93	84.65	84.72	84.54	84.49	84.09
単語認識率	発話文: タスク外（グルメ・レシピ）							
女性（カナ）	88.84	79.60	87.04	87.85	88.23	87.94	87.96	87.59
女性（漢字）	83.58	75.42	84.81	85.22	85.48	85.34	85.34	84.66
男性（カナ）	83.34	73.74	82.86	83.39	83.98	83.95	83.96	83.26
男性（漢字）	78.74	70.29	81.01	81.07	81.63	81.72	81.66	80.71

て利用するには適さない．そこで，モデル B のアルゴリズムを用いてツールの開発を続けることにする．

本ツールの開発をすすめる上で，今後，以下のような予定を考えている．

- 統計ルールフィルタの改良
- テキスト収集量の調整による性能評価
- 他のタスクでの実験及び評価
- アクセスログを用いたテキスト収集法の検討
- ツールとしての実装

また，6.2 節で述べたように，Web ページから収集したテキストには表記の揺れが多く含まれており，これが異なる語彙数の増大に影響を与えていると考えられる．これは，未知語率の増加などの原因にもなるため，表記の揺れを統一するための手法の検討も行ないたいと考えている．

最後に，本ツールはフリーソフトウェアとして公開を予定している．

謝辞 本研究は，NEDO（新エネルギー・産業技術総合開発機構）の援助を受けて行われた．高齢者音声の収集・整備は本プロジェクトの一環として TIS 株式会社（株式会社東洋情報システム）によって行なわれた．ご協力いただいた関係各位に感謝いたします．

参考文献

[1] P.R. Clarkson, R. Rosenfeld: "The CMU-Cambridge Statistical Language Modeling Toolkit v2," http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[2] 松本, 北内, 山下, 平野, 松田, 高岡, 浅原: "日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書," 2000-12 http://chasen.aist-nara.ac.jp/

[3] 河原, 李, 小林, 武田, 峯松, 嵯峨山, 伊藤, 伊藤, 山本, 山田, 宇津呂, 鹿野: "日本語ディクテーション基本ソフトウェア（99 年度版）の性能評価," 情報処理学会研究報告,99-SLP-31-2,2000

[4] 李, 河原, 堂下: "単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識," 電子情報通信学会論文誌,J82-D-II No.1,pp.1–9,1999

[5] 李, 河原, 武田, 鹿野: "Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識," 電子情報通信学会論文誌,J83-D-II No.12,pp.2517–2525,2000

[6] 馬場, 芳澤, 山田, 李, 鹿野: "高齢者向け音響モデルによる大語彙連続音声認識の評価," 情報処理学会研究報告,2001-SLP-35-3,2001