社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－185 (6)
2008－SLP－ 71 (6)
2008／5／23

ポッドキャスト音声認識の性能向上手法：集合知によって更新されるWebキーワードを活用した言語モデリング

松原 勇介† 緒方 淳‡ 後藤 真孝‡

† 東京大学
‡ 産業技術総合研究所

†matubara[at]is.s.u-tokyo.ac.jp
‡{jun.ogata,m.goto}[at]aist.go.jp

あらまし 本稿では，ポッドキャスト音声認識の性能向上のための，言語モデリング手法について述べる．ポッドキャスト音声認識においては，あらゆるタスクが認識対象となること，常に最新の話題をカバーする必要があること，などから従来の言語モデルでは高精度な認識は望めない．そこで，本研究では，集合知によって日々更新されるWeb上の辞書サービス「Webキーワード」を活用した言語モデリングを行うことで，ポッドキャスト音声認識の性能向上をはかる．実際にポッドキャストを対象とした認識実験を行い，提案手法の評価を行ったところ，性能向上に有効であることを確認した．

キーワード 大語彙連続音声認識，ポッドキャスト，言語モデル，形態素解析，Webキーワード

Improvements of Podcast Transcription: Language Modeling Based on Web Keywords Maintained Through Wisdom of Crowds

Yusuke Matsubara† Jun Ogata‡ Masataka Goto‡

†University of Tokyo
‡National Institute of Advanced Industrial Science and Technology (AIST)

Abstract This paper describes language modeling techniques to improve automatic transcription of podcasts. Most previous language models had difficulties in transcribing podcasts because podcasts include various kinds of tasks and cover recent topics that tend to have many out-of-vocabulary words. To overcome such difficulties, we improve our speech recognizer by using language modeling that utilizes "Web keywords" updated on a daily basis through wisdom of crowds. From our experimental results for actual podcast speech data, the effectiveness of the proposed language modeling was confirmed.

Keyword LVCSR, podcasts, language models, morphological analysis, Web keywords

1 まえがき

音声情報検索は，音声認識技術のアプリケーションの1つとして重要視され，近年でも活発に研究が展開されている [1]．しかしながら，現状の音声認識技術では，あらゆる音声データから検索に必要な索引情報(テキストやキーワード等)を，精度よく抽出することが困難なこともあり，Google等の代表されるテキストの検索のように日常的に利用されるには至っていない．一方，最近では，音声版のブログ(Weblog)ともいえる「ポッドキャスト」が普及し，Web上の音声データとして多数公開されるようになったため，そうした音声データに対する検索の重要性がより一層増してきたといえる．

そこで我々は，Web上の日本語のポッドキャストを音声認識によって自動的にテキスト化することで，それらをユーザが全文検索できるだけではなく，詳細な閲覧，編集も可能なソーシャルアノテーションシステム「PodCastle[2]-[4]」の開発を行っている．PodCastleでは，検索したポッドキャストの全文をテキスト表示することで，音声再生環境がなければ内容を把握できないポッドキャストを「読む」ことも可能にする．従来こうしたシステムが実現困難だったのは，ポッドキャストの多様な音声に対して，高い音声認識率を達成することが難しかったからである．本研究では，これを解決するために，システムが持つすべての情報を積極

的にユーザに開示し，多数のユーザに認識誤りを訂正 (アノテーション) する協力をしてもらうことで，音声認識率をシステムの運用中に向上させる枠組みを採用している．こうすることで，検索サービスとしての質を向上させるだけでなく，音声認識技術の底上げをはかることも狙っている．

以上述べた Web サービスを運用するために，我々は実環境音声データであるポッドキャストの音声認識手法について検討を行っている [4]．本研究では，ポッドキャスト音声認識のための言語モデリングに着目する．ポッドキャストは，幅広い話題，タスクの音声データが日々増え続けるという特徴を持っているため，言語モデルをいかにして学習，構築するかが認識性能を左右する大きなポイントとなる．本稿では，言語モデリングにおいて「Web キーワード」を活用することで，ポッドキャスト音声認識の性能向上をはかる．Web キーワードは，Web において，不特定多数のユーザによって整備，更新されているキーワードリスト (辞書) であり，様々な分野，話題のキーワードをカバーしている．Web キーワードは本来，Web 上に大量に存在するブログや記事などの関連付けを行い，必要な情報の検索，アクセスを効率化することを目的として構築されたものであるが，後述するように，音声認識にとっても有用となる特徴，機能を持っている．提案手法では，Web キーワードの特徴，機能を，単語分割 (ここでは形態素解析) も含んだ一連の言語モデリングの過程において導入する．具体的には，Web キーワードを考慮した形態素解析による N-gram モデリング，単語の読み (発音) の自動獲得，各 Web キーワードに関連付けされたテキスト (ブログ) を用いた N-gram の学習，を行う．

2 ポッドキャスト音声認識における課題

ポッドキャストは，その発話内容や録音環境などが多種多様であり，従来のタスクを限定した場合の音声認識に比べて多くの問題を含んでいる．特に言語面においては，ニュース，講演，コラム，雑談などあらゆるタスクが認識対象となること，日々新しいエピソード (音声データ，mp3 ファイル) が追加されるため常に最新の話題をカバーする必要があること，などが音声認識を困難にする要因となっている．

そのような問題に対して，我々はこれまでに，Web ニューステキストを利用した言語モデリング手法を提案し，ポッドキャスト音声認識において有効性を示した [4]．しかし，本手法では，日々更新される Web ニューステキストを学習に用いることで，ニュース性のある新しい話題に対して有効であったものの，N-gram 学習の事前処理として行われる単語分割 (形態素解析) において，未知語 (新出語) に対する分割誤りに対処できない問題があった．これは，単語分割 (形態素解析) 自体が持つ問題であり，単語分割器 (形態素解析器) において学習されていない (辞書にない) 単語は，正しい切れ目で分割することができない (例：“ケータイ” ⇒ “ケー”, “タイ”)．特に，新しい話題のテキストを学習する場合では，多くの未知語が出現する可能性が高いため，このような分割誤りは音声認識性能を劣化させる原因となる．

また，ポッドキャストのように話し言葉を対象とした音声認識においては，個々の単語の認識難易度を表す要因として，言語モデル学習データにおける出現頻度，単語の長さ (単語を構成する音素数) などが大きく影響することが報告されている [5]．すなわち，出現頻度の低い単語，音素数の少ない単語は認識誤りとなりやすい傾向がある．前述したような，新しい話題において出現した未知語に関しては，基本的に言語モデルの学習データにおいて出現頻度はそれほど高くないと考えられる．したがって，新たに獲得した未知語については，他の既知語に比べて認識誤りとなる可能性が高く，いかにして未知語に関連するテキストを収集するかが重要なポイントとなる．一方，単語分割として音声認識において一般的に用いられている形態素解析では，日本語の解析的な最小単位 (形態素) として分割を行うため，N-gram としてモデル化される単語単位としては，人間が感覚的に捉えている「単語」よりも短いものとなる [6]．そのため，各単語を構成する音素数は全体的に少なくなり，認識時に認識誤りを引き起こしやすくなるといった問題もある．

3 Web キーワードを利用した言語モデリング

前節で述べた問題点を解決するために，本研究では「Web キーワード」を活用した，ポッドキャスト音声認識のための言語モデリング手法を提案する．以下，まず Web キーワードについて述べ，提案する言語モデリング手法について説明を行う．

3.1 Web キーワード

本稿では，ブログや Wiki で共有され，ユーザーによって追加，編集されている辞書のことを，「Web キーワード」と呼ぶ．Web キーワードは，単語を手掛かりにして，関連したブログとブログ，記事と記事の間で

のアクセスを促進するために用いられている．このようなWebキーワードをWebサービスとして運営しているサイトはいくつか存在するが[7][8]，本研究ではその中でも最も大規模なサービスとなっている「はてなダイアリーキーワード」[7]を利用する．

Webキーワード(ここでは，はてなダイアリーキーワード)には，次のような特徴がある．

1. **日々，様々なジャンルのキーワードが追加，編集されている**
   不特定多数のユーザの協力，すなわち「集合知」によって，様々なジャンルのキーワードとそれらに対する説明文が整備されている．さらに，キーワードとともにそれに対する読み(ふりがな)までも定型のフォーマットにて記述されており，必要なキーワードに対する読みを容易に取得できる．
2. **キーワードを含むテキストがハイパーリンクによって関連付けられている**
   Webキーワードを提供するサービスにおいて，ユーザーはキーワードを手がかりにして，他のユーザーのブログ等の記事にアクセスできるようになっている．たとえば，ブログサービス「はてなダイアリー」ではユーザーが書いた日記のテキスト中のキーワードが，そのキーワードを含むブログ記事の一覧へのハイパーリンクとなる．
3. **コミュニティによって質と量が維持されている**
   Webキーワードの質と量は，ユーザー同士のアクセスの利便性を高める目的で維持されている．たとえば，記事同士をつなぐ効果が薄い単語（単語として不完全な文字列や機能語など）は登録されない傾向にある．一方，新しい商品の名前，新しい映画のタイトルなど，ユーザーの共通の関心事項となりやすい単語は，迅速にWebキーワードに追加されていく．

3.2 提案手法

提案するWebキーワードの特徴を活用した言語モデリングの概要を図1に示す．以下，各ポイントの詳細を順に説明する．

3.2.1 Webキーワードを利用した形態素解析辞書の拡張

まず，はてなダイアリーキーワードを通じて得られるWebキーワードを利用して，形態素解析辞書の拡張を行う．ここでの目的は，学習テキストに対して単語分割(形態素解析)を行う際に，Webキーワード中に含まれる様々なジャンルのキーワードに対する，分割誤りの発生を抑制することである．すなわち，Webキーワードが，最終的なN-gramの単語単位として登録されるように，学習テキストを正しく分割する．

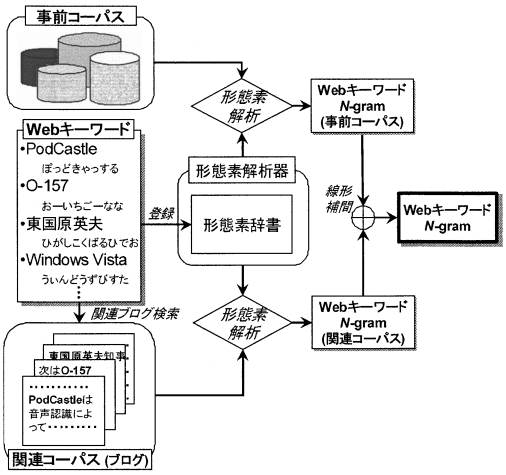

図 1: Webキーワードを利用した言語モデリング

本研究では，形態素解析器としてMeCab[9]を利用する．MeCabを含む多くの形態素解析システムの辞書に必要な情報は，表記，読み，品詞，コストである．表記と読みはすでにWebキーワードに含まれているため，品詞とコストの決め方が問題となる．Webキーワードのほとんどは，先に述べた性質により，名詞か固有名詞とみなすことができるので，本研究ではすべてのキーワードの品詞を「名詞, 一般」とする．コストの推定方法として，代表的なのは学習用の形態素解析済みコーパスを用いる方法である．しかしWebキーワードを含む形態素解析済みコーパスの入手は困難であるため，本研究ではすべてのキーワードのコストを一定の値4000とする．最後に，MeCabの辞書のフォーマットに合うように形式を変換し，既存の辞書と統合することで，Webキーワードベースの形態素解析辞書を作成する．以上の形態素辞書を用いて形態素解析を行うことにより，学習テキスト中にWebキーワードに相当する表記が出現すれば，1つの単語として正しく切り出すことができる．

WebキーワードをN-gramの単語単位として登録することのメリットとしては，Webキーワードに付随する読み情報により，正しい発音(音素列)を獲得できる点が挙げられる．Webキーワードは様々なジャンルの最新の単語を多く含み，そのような単語に対して正しい発音が付与できることは，日々更新されるポッドキャストを認識する際には特に有効に働くと考えられる．また，Webキーワードは，通常の形態素に比べ，比較的長い単位で構成されるものが多い．したがって

実際に N-gram を構築した際には，音素数の比較的多い単語が語彙として含まれることになるが，音声認識を考えた場合には，各単語間の音響的類似度が小さくなることで，通常の形態素 N-gram に比べて誤認識を削減できる可能性がある．

3.2.2 Web キーワードの関連コーパスの獲得

言語モデルの学習テキスト中で出現頻度の低い単語が，音声認識誤りを引き起こす傾向があること (2 章参照) の理由としては，その単語が現れる周辺文脈のテキストが十分に得られないために，推定される言語的確率値が頑健でないことが挙げられる．特に Web キーワードは最新の話題に関するキーワードを多く含むため，このような影響が比較的大きいと考えられる．そこで，3.1 節の 2. で述べた Web キーワードの特徴により，Web キーワードに関連付けられている記事 (ブログ) を収集，利用することを考える．Web キーワードに関連付けられた各ブログから得られるテキストは，キーワードが現れる文脈を保持しているため，各キーワードに関する N-gram の確率推定を有効に行うことができる．

以上のようにして得られたテキストデータを「関連コーパス」と呼ぶ．

3.2.3 Web キーワード N-gram の生成

N-gram の学習データとして，ここでは 2 種類のコーパスを用いる．1 つは，言語モデル学習において一般的に利用される，新聞記事コーパスや日本語話し言葉コーパス (CSJ) を合わせたもので，本研究ではこれを「事前コーパス」と呼ぶ．もう 1 つは，前節で述べた Web キーワードの「関連コーパス」である．これら 2 つのコーパスのテキストデータに対して，Web キーワードベースの形態素解析を行うことで，Web キーワードが語彙として学習された 2 種類の N-gram を生成する．そして，それぞれのコーパスから学習された N-gram を線形補間することで，最終的な「Web キーワード N-gram」を生成する．

4 実験と考察

提案する言語モデリングを，実際のポッドキャストを用いた音声認識実験により評価を行った．

4.1 学習データ

Web キーワードとしては，「はてなダイアリーキーワード」の 2007 年 8 月 1 日の時点でのキーワードリストを利用した．これらに対して種々のフィルターを施すことで，最終的に 188036 のキーワードを得た．

言語モデルの学習用データのうち，「事前コーパス」としては，毎日新聞記事 10 年分 (1991 年～2001 年) のテキストデータと日本語話し言葉コーパスの 2670 講演分の書き起こし，さらに本研究で収集した Web ニュースサイト (Google ニュース，Yahoo!ニュース) のテキストデータ (2006 年 9 月～2007 年 12 月) を用いた．一方，「関連コーパス」としては，上記の Web キーワード 1 つあたり平均 4.5 記事，合計 848816 記事の「はてなダイアリー」のブログを利用した．

表 1: 評価用音声データ

カテゴリ	エピソード数	時間 (sec.)
ニュース	4	2087.39
コラム	7	1434.30
対談	3	1012.65

4.2 評価用音声データ

本実験では，「ニュース」，「コラム」，「対談」の 3 つのカテゴリのポッドキャストを評価用音声データとして用いた (表 1)．「ニュース」はその日の様々な話題が網羅されており，背景には音楽が流れている．したがって，読み上げ音声であるものの，認識が比較的困難なデータとなっている．「コラム」は，経済に関する内容の独話であり，講演音声に比較的近い．「対談」は，全エピソードとも話者 2 人での対話であり，完全な自由発話音声となっている．また背景雑音も多い．

4.3 音声認識システム

音響モデルには，単語間の調音結合も考慮した状態共有型 triphone モデルを用いた．雑音対処として，雑音環境下音声認識において幅広く利用されている ETSI Advanced Front-End[10] を音響分析に適用した．学習データには，日本語話し言葉コーパス (CSJ) を用いた．また，評価用音声データの各エピソードごとに，教師なし MLLR 適応 [11] を行った．このときの教師ラベルは，音素認識を行うことで生成した．

本音声認識システムのデコーディングは段階的探索に基づいている．まず，bigram を用いた N-best デコーディングにより単語グラフを生成する [12]．次に，trigram を用いて，生成された単語グラフを，trigram 制約の単語グラフに拡張する．最後に，trigram 制約の単語グラフに対して，consensus デコーディング [13] を行い，confusion network 中の最尤候補を最終の認識結果とした．

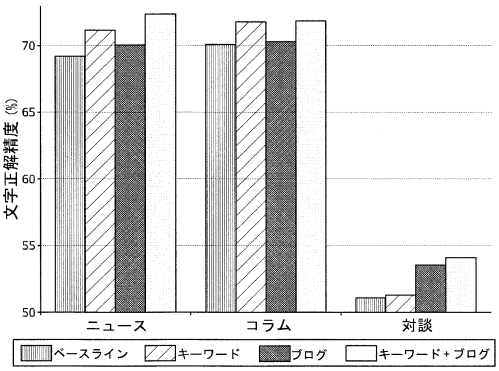

図 2: 各言語モデルにおける認識性能 (文字正解精度)

4.4 実験結果

ここでは，Web キーワードを N-gram の単語単位とすることの効果，Web キーワードの関連コーパスを学習に用いることの効果について考察を行うために，以下の 4 種類の N-gram(本実験では trigram) を作成した．

- **ベースライン**
  Web キーワードを利用しない通常の形態素解析により作成した N-gram で，学習には事前コーパスのみを使用．
- **キーワード**
  Web キーワードベースの形態素解析により作成した N-gram で，学習には事前コーパスのみを使用．
- **ブログ**
  Web キーワードを利用しない通常の形態素解析で作成した N-gram で，学習には事前コーパスと関連コーパス (ブログ) を使用．
- **キーワード+ブログ**
  Web キーワードベースの形態素解析により作成した N-gram で，学習には事前コーパスと関連コーパス (ブログ) を使用．図 1 中の最終的に生成される Web キーワード N-gram に相当．

N-gram の語彙選択は各単語の学習データ中の頻度により行い，本実験では，頻度が 100 以下の単語をカットオフした．その結果，上記 4 種類の N-gram の語彙サイズはそれぞれ，61799(ベースライン)，75063(キーワード)，76539(ブログ)，90444(キーワード+ブログ) となった．

各言語モデルにおける認識性能を図 2 に示す．上記言語モデルは単語の単位自体が異なるため，通常の単語正解精度では数値での厳密な比較はできない．したがって，ここでの評価には文字正解精度を用いた．結果より，ニュース，コラムにおいては Web キーワードを用いることで，認識精度の改善がみられた．これはニュースやコラムは，最新の話題や，Web キーワードに含まれるキーワードが話されることが比較的多いためである．一方，対談においては，Web キーワードを活用することの効果はそれほどみられなかった．しかし，関連コーパスを学習に用いることで，認識性能が改善している．これは，ブログ記事は話し言葉の要素を多分に含むため，対談のような自由発話音声に対して特に改善が大きかったものと考えられる．また，評価用音声データ全体の平均文字正解精度を，各言語モデルに表 2 に示す．Web キーワード，さらにその関連コーパスを併用することで，評価用音声データ全体の性能も改善できることがわかった．

表 2: 評価用音声データ全体の平均文字正解精度

言語モデル	文字正解精度 (%)
ベースライン	63.47
キーワード	64.74
ブログ	64.64
キーワード+ブログ	**66.11**

表 3: 評価用音声データ全体の平均単語正解精度

言語モデル	単語正解精度 (%)
ベースライン	57.24
キーワード	57.83
ブログ	57.65
キーワード+ブログ	**59.32**

最後に，参考として評価用音声データ全体の平均単語正解精度を表 3 に示す．厳密な比較ではないが，単語正解精度においても，提案手法が比較的良好な性能を示すことがわかる．

5 おわりに

本稿では，ポッドキャスト音声認識を改善するための言語モデリング手法について検討した．ポッドキャストのように，幅広い話題，タスクの音声データが日々増え続けるという特徴を持つ音声データに対し，不特定多数のユーザによって整備，更新されている「Web キーワード」を有効活用した言語モデリング手法を提案した．ニュース，コラム，対談の 3 種類のポッドキャスト音声データを用いた評価実験により，提案手法が認識性能を向上できることを示した．

我々が開発している音声情報検索システム「PodCastle」では，不特定多数のユーザが音声認識誤りを

効率的に訂正できる機能[14]を提供している．このような訂正インタフェースにおいては，形態素のような短い単位に対する訂正はユーザにとって煩わしいものとなる可能性があるため，認識結果としては比較的長い単位が望ましいといえる．また，Webキーワードのように，ある話題において重要な意味を持つ可能性の高い単語が比較的精度良く認識できることは，認識結果を閲覧，訂正する際に有用なものとなる．したがって，本稿で提案した言語モデリング手法は，音声認識自体の認識性能を改善するだけでなく，PodCastleにおけるユーザインタフェース面にも波及効果があると考えられる．

今後の課題としては，高次N-gramの利用や他のカテゴリのポッドキャストに対する評価，Webキーワードベース形態素解析の改善などが挙げられる．

謝辞

本研究の一部は，科研費(19300065)の助成を受けた．

参考文献

[1] 伊藤 克亘, 相川 清明, 秋葉 友良, 伊藤 慶明, 河原 達也, 南條浩輝, 西崎 博光, 安田 宜仁, 山下 洋一: “音声ドキュメント検索評価のためのテストコレクションの試作” 情処研報, 2006-SLP-64-24, pp.137-142, 2007.

[2] 緒方 淳，後藤 真孝: “PodCastle: ポッドキャストをテキストで検索，閲覧，編集できるソーシャルアノテーションシステム”, WISS 2006, 論文集, 2006.

[3] 後藤 真孝, 緒方 淳, 江渡 浩一郎: “PodCastleの提案: 音声認識研究2.0を目指して” 情処研報, 2007-SLP-65-7, 2007.

[4] 緒方 淳, 後藤 真孝, 江渡 浩一郎: “PodCastleの実現: Web2.0に基づく音声認識性能の向上について” 情処研報, 2007-SLP-65-8, 2007.

[5] T.Shinozaki and S.Furui: “Error analysis using decision trees in spontaneous presentation speech recognition”, Proc. of ASRU, 2001.

[6] 西村雅史, 伊東伸泰, 山崎一孝, 荻野紫穂: “単語を単位とした日本語の大語彙連続音声認識”, 情処研報, 1998-SLP-20-3, pp.17-24, 1998.

[7] はてなダイアリーキーワード: http://d.hatena.ne.jp/keyword/

[8] アットペディア: http://atpedia.jp/

[9] 工藤 拓, 山本 薫, 松本 裕治: “Conditional Random Fields を用いた日本語形態素解析” 情処研報, 2004-NL-161-13, pp. 89-96, 2004.

[10] ETSI ES 202 050 v1.1.1 STQ; ”Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms”. 2002.

[11] C.L.Leggetter and P.C.Woodland：“Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language,Vol.9, pp.171-185, 1995.

[12] 緒方淳, 有木康雄: “大語彙連続音声認識における最ゆう単語back-off接続を用いた効率的なN-best探索法”, 信学論 (D-II), Vol.84-D-II, No.12, pp.2489-2500, 2001.

[13] L.Mangu, E.Brill and A.Stolcke: “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network” Computer Speech and Language,Vol.14, No.4, pp.373-400, 2000.

[14] 緒方 淳，後藤 真孝: “音声訂正：選択操作による効率的な誤り訂正が可能な音声入力インタフェース”, 情処学論, Vol.48, No.1, pp.375-385, 2007.