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

2006－SLP－64 (42)
2006／12／22

ドメインとスタイルを考慮したWebテキストの選択による対話システム用言語モデルの構築

翠 輝久† 河原 達也†

† 京都大学 知能情報学専攻
〒 606-8501 京都市左京区吉田二本松町

あらまし 音声対話システムにおいて，ユーザの多様な発話を頑健に認識するためには，タスクドメインに合致した十分な量のテキストデータでN-gram言語モデルを学習することが望ましい．しかし，新たに音声対話システムを作成する際に，ユーザが入力すると想定される発話を大量に用意することは困難である．そこで本研究では，Webから学習データを収集・選択することにより効率的に言語モデルを構築する手法を提案する．Webの検索クエリは，対話システムが対象とするドメインについて記述された文書から作成して，Webを検索する．このようにして収集されたWebテキストの多くは，対話システムのユーザの発話スタイルとマッチしたものではなく，言語モデルの学習データとしてこれらのすべてを使用するのは適切でない．そこで，別の対話システムで収集されたユーザ発話コーパスを併用することで，発話スタイルの近い文を選択する．ソフトウェアサポートと観光案内の2つのドメインにおいて評価を行った結果，音声認識精度の有意な改善が得られた．また実験結果の分析により，Webテキストを選択する際に，テキストのスタイルを考慮することの重要性が確認された．
キーワード 音声認識，言語モデル，音声対話システム，Webテキスト選択

Efficient Language Model Construction for Spoken Dialog Systems by Web Text Selection Considering Domain and Utterance Style

Teruhisa MISU† and Tatsuya KAWAHARA†

† School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract This paper proposes a bootstrapping method of constructing statistical language models for new spoken dialog systems by collecting and selecting sentences from the World Wide Web (WWW). To make effective search queries that cover the target domain in full detail, we exploit the document set described about the target domain as seeding data. An important issue is how to filter the retrieved Web pages, since all of the retrieved Web texts are not necessarily suitable as training data. We induct an existing dialog corpus of different domain to prefer the texts of spoken style. The proposed method was evaluated on two different tasks of software support and sightseeing guidance, and significant reduction of the word error rate was achieved. We show that it is vital to incorporate the dialog corpus, though not relevant to the target domain, in the text selection phase.
Key words Speech recognition, Language model, Spoken dialog system, Web text selection.

1. はじめに

音声対話システムにおいて，ユーザの多様な発話を頑健に認識するためには，十分な量のタスクドメインに合致したテキストデータでN-gram言語モデルを学習することが望ましい．しかし，新たに音声対話システムを作成する際に，ユーザが入力すると想定される発話を大量に用意することは困難である．そのため，初期の言語モデルは人手により文法を記述するか，Wizard-of-Oz法によりデータを集めるなどの方法が用いられることが多い．しかし，これらの方法は文法作成やデータ収集にかかるコストが大きく，また，信頼できる文法が構築できるとは限らない．

そこで近年，Webテキストを利用して，学習データの不足を補う試みが行われつつある．Zhuら[1]は，音声の検索タスクにおいて，学習データ中の出現回数の少ない3-gramに信頼できる確率を付与するために，当該3-gramのWebでの出現回数を利用している．また，Bulykoら[2]はSwitchboardコーパスの認識タスクにおいて，頻出するn-gramエントリをWebの検索クエリとして用いてより多くの学習データを収集する手法を提案している．梶浦ら[3]は，講演の認識タスクにおいて，話題に特化した言語モデルを構築するために，初期認識結果から話題語を抽出してWebテキストを収集している．Sarikayaら[4]は，このようなアプローチを対話システムの音声認識タスクに適用している．具体的には，あらかじめ収集したユーザの発話コーパスのN-gramエントリを利用してWebを検索し，収集したテキストから発話コーパスに類似した文をBLEUスコアを用いて選択している．

しかし，これらの研究においては，当該ドメインタスクのベースライン言語モデルや，ある程度まとまった量の発話の書き起こしデータが利用可能であることが前提とされている．実際に[4]においては，約二千発話が初期データとして使用されている．これに対して本研究では，言語モデルの作成にドメインに特化した初期学習データを必要としない方法を考える．初期データとしては，ユーザ発話の代わりに，対話システムが対象とするドメインについて記述された文書集合(知識ベース)を利用する．多くの情報検索タイプの音声対話システムは，対象とするドメインについて記述された文書集合を前提としていることから，この仮定は妥当である．たとえば，レストラン検索システムは，レストランについて記述されたWebページを検索することにより実現される．質問応答(QA)システムにおいても，バックエンドの知識ベースとして大量の文書が利用されている[5]．また，このような対象ドメインに関する文書を収集することは，ユーザ発話を収録するよりも非常に容易である．このような，文書集合をもとに検索クエリを作成することにより，大量のWebテキストを収集する．

しかし，このようにして収集されたWebテキストの多くは，文書スタイルで記述されたものが多く，対話システムのユーザ発話のスタイルとマッチしたものは必ずしも多くない．そこで本研究では，別のドメインにおいて収集された既存の音声対話コーパスを利用することで，言語モデルの学習データとしてふさわしいテキスト，すなわち対話システムのドメイン・発話のスタイル共にマッチしたデータを効率的に選択することを目指す．

提案手法の評価のために，情報検索と質問応答の異なる2つのドメイン・タスク用に言語モデルを構築して，音声認識実験を行った結果について報告する．

2. Webテキストを利用した統計的言語モデルの構築

近年，音声認識技術と情報検索技術の進展に伴い，音声対話システムの検索対象は単純なデータベース[6]から，マニュアル[7]や，新聞記事[8]といった一般的な文書検索に広がりつつある．これらのシステムにおいては，ユーザ発話と文書集合とのマッチングが行われ，検索結果をもとに応答が生成される．このようなシステムにおいて検索対象の文書集合(知識ベース)を利用することで，効率的に音声認識用の言語モデルを構築することを考える．

まず，知識ベース中の各文書を特徴づける単語を抽出し，検索クエリを作成することにより，システムが対象とする話題のすべてを網羅した検索クエリを作成して，Webテキストの収集を行う．このようにして収集したテキストの中から，言語モデルの学習データとしてふさわしいものを選択する必要がある．選択の基準として，知識ベースとの類似度を利用することも考えられるが，知識ベースのテキストは文書スタイル(書き言葉)で記述されているため，話し言葉(検索要求の口調)が多い対話システムの学習データとしては不整合を生じる．そこで，本研究では，別のドメインにおいて収集された音声対話コーパスと，知識ベースを混合して作成したベースライン言語モデルを使用する．この言語モデルによりゆう度を計算することで，Webテキストがドメイン・スタイル双方の観点から学習データとしてふさわしいものである

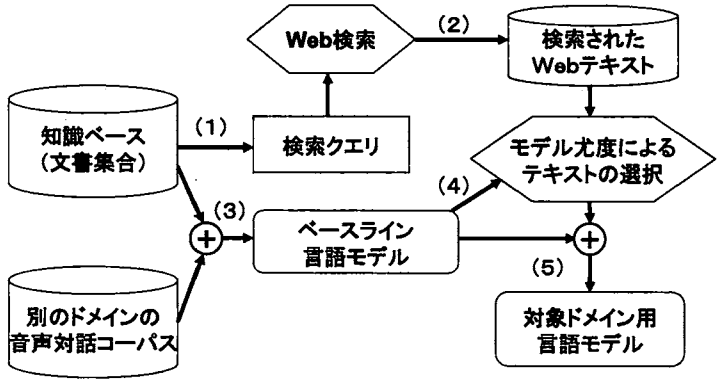

図 1　提案手法の概要

かを判定する．

提案手法の概要を以下に示す．また，手法の流れを図1に示す．

（1） 知識ベースからキーワードを抽出し，Web の検索クエリを作成する．

（2） Web を検索する．

（3） 知識ベースと別のドメインの既存の音声対話コーパスからベースライン言語モデルを作成する．

（4） ベースライン言語モデルのゆう度により，収集した Web テキストから学習データとして利用する文を選択する．

（5） 知識ベース，対話コーパス，選択した Web テキストを用いて，対象ドメイン用の言語モデルを作成する．

本研究では音声対話システムの対象ドメインとして，ソフトウェアサポートタスクと観光案内タスクを想定する．ソフトウェアサポート用の知識ベースとして，マイクロソフト社から提供されたソフトウェアサポート用知識ベースを利用する．この知識ベースは，著者らが開発してきた音声による文書検索システム「音声版ダイアログナビ [9]」において使用されている．観光案内用の知識ベースとしては，Wikipedia(注1) の京都に関する文書と，京都市観光局による京都情報データベース(注2)を使用する．これらの文書は，著者らが京都大学博物館において運用を行った音声による観光情報案内・質問応答システム「京都版ダイアログナビ [10]」において，バックエンド知識ベースとして使用されている．また，別ドメインの既存の音声対話コーパスとして，名古屋大学で収録された CIAIR 車内音声対話コーパス [11] を利用する．このコーパスの対話のドメインはレストラン検索である．

(注1)：http://ja.wikipedia.org/
(注2)：http://www.raku.city.kyoto.jp/sight.phtml

表 1　学習コーパスの概要

テキストの種類	総文数	総単語数
ソフトウェアサポートタスク		
マイクロソフト社ソフトウェアサポート用知識ベース	88,440	1.7M
観光案内タスク		
Wikipedia の京都に関する文書	10,081	0.11M
京都情報データベース	2,903	0.06M
別のドメインで収集された対話コーパス		
CIAIR 車内音声対話コーパス	24,701	0.24M

人間のオペレータ，WOZ システム，音声対話システムの3種類の対話相手との対話から構成されており，ユーザの発話スタイルの多くをカバーしていると考えられる．今回，このコーパスからユーザ側の発話を選択して，発話スタイルがマッチしたコーパスとして利用する．これらのコーパスの概要を，図1に，例を図2に示す．

3. Web テキストの収集

Web から学習データを収集する際には，検索エンジンに入力するクエリの生成方法が重要となる．本研究では，知識ベースから，各文書を特徴づける語句を TF*IDF スコアを用いて自動抽出して検索クエリを生成する．具体的には，文書毎に単語の頻度情報から TF*IDF ベクトルを計算し，その値が大きい単語 (一文書あたり 5 単語程度) を選択してそれらの AND 結合を検索クエリと

○ソフトウェアサポート用知識ベース

Windows XP で音声認識を使用する方法

概要　この資料では，Windows XP で音声認識を使用する方法について説明しています．Microsoft Office XP の音声認識をインストールしているか，または，Office XP がインストールされたコンピュータを新たに購入...

○ Wikipedia テキスト

銀閣寺

概要　銀閣寺は，京都府京都市左京区にあり，東山文化を代表する臨済宗相国寺派の寺院．通称銀閣寺，山号は東山．開基は，室町幕府 8 代将軍の足利義政，･･･

○ CIAIR 対話コーパス

- あー，この近くにコンビニとかないかなあ。
- じゃあ，一番近いケンタッキーでお願いします。
- えーと，駐車場がある店で。

図 2　用いたテキストの例

表 2　収集された Web テキスト

	ページ数	総文数	総単語数
ソフトウェアサポート	3.4M	88M	1,870M
観光案内	0.3M	12M	250M

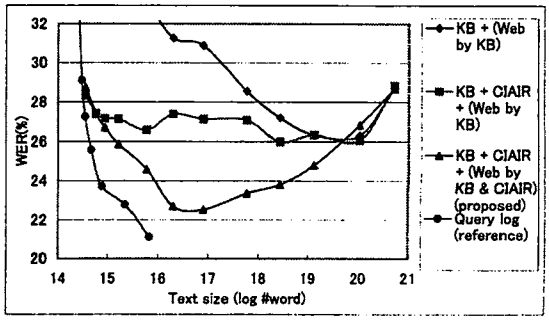

図 3　学習テキストサイズと単語誤り率 (WER) との関係 (ソフトウェアサポート)

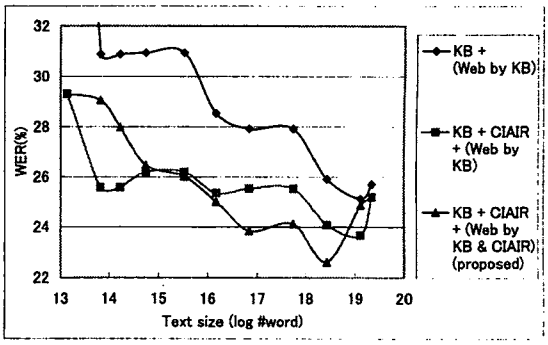

図 4　学習テキストサイズと単語誤り率 (WER) との関係 (観光案内)

する．ただし，観光案内タスクの知識ベースは，ソフトウェアサポート用知識ベースと比較して，文書数が少なく，TF*IDF により特徴的な単語を抽出することが困難であったため，文書のタイトル (観光地名，人名等) を検索クエリとして使用した．

このように作成した検索クエリを用いて，検索エンジン Goo(注3)により，各クエリごとに 500 文書を上限として Web ページを収集した．また，ファイルサイズを考慮して，HTML ファイルのみをダウンロードした．収集された Web テキストから，HTML タグを取り除き文単位に区切り，形態素解析を行った．

このようにして収集されたテキストのサイズを表 2 に示す．

4.　Web テキストからの学習データの選択

収集した Web テキストから，言語モデルの学習データとして適当な，ドメイン・発話スタイルともにマッチした文を選択する．Web テキストを言語モデルの学習に利用する先行研究の多く [3], [12] は，基本的に収集した Web ページ全体を学習データとして使用している．しかしながら，Web テキストのすべてが言語モデルの学習データとしてふさわしいとは限らない

そこで本研究では，ドメイン・発話スタイルに基づく類似度を利用して，学習データとして使用するかを判定する．類似度として，ベースライン言語モデルによるモデルゆう度を利用する．具体的には，知識ベースと既存の CIAIR 対話コーパスから作成した 3-gram モデルにより，Web から収集したテキストの文単位でのモデルゆう度 (単語パープレキシティ) を計算する．このゆう度がしきい値 θ より大きい (パープレキシティが小さい) 文を学習データとして選択する．ここで，CIAIR 対話コーパスを混合することは，発話スタイルがマッチした文を選択するために重要と考えられる．なお，テキスト選択用の言語モデルにおける未知語には，ペナルティとしてベースライン言語モデル中の最小の確率を与える．

以上の手順で選択した Web テキストに，知識ベース，CIAIR 対話コーパスを加えて対象ドメイン用の言語モデルを作成する．

(注3)：http://www.goo.ne.jp/

表 3　音声認識精度による言語モデルの比較 (WER%)

	ソフトウェアサポート	観光案内
KB+CIAIR (baseline)	28.5	28.4
KB+(Web by KB)	27.0	25.1
KB+CIAIR+(Web by KB)	26.7	24.4
KB+CIAIR+(Web by KB&CIAIR) (proposed)	22.8	22.6

5. 音声認識実験による言語モデルの評価

提案手法の評価を行うために，2. 章で示した 2 つのドメインを対象として，それぞれのシステム用の言語モデルを作成して音声認識実験を行った．評価データとして，ソフトウェアサポートタスクでは 30 名の話者による 499 発話，観光案内タスクでは 4 名の話者による 220 発話を用いて単語誤り率 (WER) を求めた．音声認識には，音声認識エンジン Julius 3.5 [13]，音響モデルには話者非依存 PTM Triphone モデルを使用した．それぞれのドメインでの評価結果を図 3，図 4 に示す．Web テキスト選択のためのしきい値 θ を変化させることで学習に用いる Web テキストの量を調整して言語モデルを作成し，WER を学習データの単語数 (の自然対数) に対してプロットした．表中の (KB+CIAIR+(Web by KB & CIAIR)) が提案手法である．比較対象として，Web テキストを選択するためのベースライン言語モデルを作成する際に，CIAIR 対話コーパスを使用しない場合 (=知識ベースのみを利用して学習した言語モデルを用いた場合) との比較を行った．ここで，CIAIR 対話コーパス利用することの効果を詳細に分析するために，2 通りの方法を試した，ひとつは，対象ドメインの言語モデルを作成する際の学習データとして CIAIR コーパスを加える場合 (KB+CIAIR+(Web by KB)) であり，もうひとつは，学習データに加えない場合 (KB+(Web by KB)) である．おおむね前者の方が高い性能を示しているが，WER が最小となる点での性能には大きな差がない．これに対して提案手法では，これらのモデルに対して，より高い性能を示していることが確認された．

また，どちらのタスクにおいても，すべての Web テキストを学習データに利用した場合に性能の低下が見られた．この結果は，学習データとして利用するテキストを選択することの必要性を確認するものである．また，知識ベースのみから学習した言語モデルにより Web テキストを選択する場合と比較して，提案手法の方が WER が最小となる点で使用したテキストの量が少ないことから，ユーザの発話にドメイン・スタイルともにマッチした文が適切に選択されていると考えられる．

さらに，Web テキスト選択のしきい値 θ を決定するために，クロスバリデーションによる評価を行った．すなわち，評価データをテストセット 1，テストセット 2 に分割して，一方の認識結果に基づいて他方のしきい値を決定した．また，ベースライン手法として，知識ベースと CIAIR 対話コーパスを混合して作成したベースライン言語モデル (混合重みもクロスバリデーションにより決定) との比較を行った．これらの結果を表 3 に示す．(KB+CIAIR+(Web by KB)) と，(KB+(Web by KB)) の違い (ソフトウェアサポートタスクにおいて 0.3%，観光案内タスクにおいて 0.7%) は，発話スタイルがマッチした CIAIR 対話コーパスを学習データに加えることの効果である．この改善には有意な差はなかった．また，(KB+CIAIR+(Web by KB)) と，提案手法の違い (ソフトウェアサポートタスクにおいて 3.9%，観光案内タスクにおいて 1.8%) は CIAIR 対話コーパスを Web テキストの選択に用いることの効果である．ソフトウェアサポートタスクにおける改善は有意水準 1%で有意である．この結果は，Web テキストから学習データを選択する際に，スタイルがマッチした CIAIR 対話コーパスを利用することが重要であることを示している．

ベースラインからの音声認識精度の改善は，ソフトウェアサポートタスクにおいて絶対値で 5.7%，観光案内タスクにおいて 5.8%であった．これらの改善は有意水準 1%で有意である．

さらに参考として，ソフトウェアサポートタスクにおいて，想定入力発話が大量に利用できる場合との比較を行った．想定発話として，テキスト入力によりマイクロソフト社のソフトウェアサポート知識ベースを検索する

システム [14] で収集されたログデータ(注4)を利用した．このデータはユーザの音声による発話の書き起こしではないが，ドメイン・スタイルともにほぼマッチしたデータである．ログデータからランダムに選択することにより，テキストサイズを変えながら，知識ベースを混合して言語モデルを作成した．この言語モデルによる音声認識結果を表3に併記する (Query log)．提案手法により作成した言語モデルは，想定発話が約 25 万文 (300 万単語) 得られた場合 (図中の○で示す点) とほぼ同等の性能であることがわかる．

6. おわりに

本研究では，音声対話システムを構築する際に，Web から収集した大量のテキストデータから，当該音声対話システムにマッチした文を選択することで，言語モデルを効率的に構築する手法を提案した．提案手法の効果を確かめるために，ソフトウェアサポートと観光案内の 2 つの異なるドメイン用の言語モデルを作成して音声認識実験を行った結果，提案手法により，ドメインにマッチしたデータなしに，数百万単語の学習データが利用可能な場合と同等の性能が得られることを確認した．

さらに，音声認識を利用したことのない本学の学生に提案手法を利用して本手法に基づいた言語モデルを作成するツールを提供して，国内の天気情報の検索とサッカーに関する情報収集をタスクとして対話システムの作成を行ってもらった．その結果，音声認識技術に対する習熟がない者でも，「台風二号のアジア名は何ですか？」，「京都の今日の最高気温と最低気温を教えて？」や「コートジボワールのドログバの背番号は何ですか？」，「ボランチはどのようなポジションですか？」といったある程度複雑な発話を，精度よく認識できる言語モデルが作成できることが確認できた．

(注4)：計約 50 万文，650 万単語

文　献

[1] X. Zhu and R. Rosenfeld. Improving trigram language modeling with the world wide web. In *Proc. ICASSP*, Vol. 1, pp. 533–536, 2001.

[2] I. Bulyko, M. Ostendorf, and A. Stolcke. Getting more mileage fromweb text sources for conversational speech language modeling using class-dependent mixtures. In *Proc. of Human Language Technology 2003 (HLT2003)*, 2003.

[3] 梶浦泰智, 鈴木基之, 伊藤彰則, 牧野正三. Web テキストを利用した言語モデル教師なしタスク適応. 電子情報通信学会技術研究報告, SP2006-18, 2006.

[4] R. Sarikaya, A. Gravano, and Y. Gao. Rapid Language Model Development Using External Resources for New Spoken Dialog Domains. In *Proc. ICASSP*, Vol. 1, pp. 573–576, 2005.

[5] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, and S. Furui. Deriving disambiguous queries in a spoken interactive ODQA system. In *Proc. IEEE-ICASSP*, 2003.

[6] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. Di Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. The AT&T-DARPA Communicator mixed-initiative spoken dialogue system. In *Proc. ICSLP*, 2000.

[7] 伊藤亮介, 駒谷和範, 河原達也. 機器操作マニュアルの知識と構造を利用した音声対話ヘルプシステム. 情処学論, Vol. 43, No. 7, pp. 2147–2154, 2002.

[8] E. Chang, F. Seide, H. M. Meng. Z. Chen, Y. Shi, and Y. C. Li. A system for spoken query information retrieval on mobile devices. *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 8, pp. 531–541, 2002.

[9] T. Misu and T. Kawahara. Dialogue strategy to clarify user's queries for document retrieval system with speech interface. *Speech Communication*, Vol. 48, No. 9, pp. 1137–1150, 2006.

[10] 翠輝久, 河原達也. 限定されたドメインにおける質問応答機能を備えた文書検索・提示型対話システム. 情報処理学会研究報告, 2006-SLP-62-13, 2006.

[11] N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda, and F. Itakura. CIAIR In-Car Speech Database. In *Proc. ICSLP*, Vol. IV, 2004.

[12] R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, and K. Shikano. Automatic N-gram language model creation from Web resources. In *Proc. Eurospeech*, 2001.

[13] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano. Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository. In *Proc. ICSLP*, Vol. IV, 2004.

[14] 清田陽司, 黒橋禎夫, 木戸冬子. 大規模テキスト知識ベースに基づく自動質問応答 ーダイアログナビー. 自然言語処理, Vol. 10, No. 4, pp. 145–175, 2003.