情報処理学会論文誌 Vol.55 No.7 1695–1705 (July 2014)

名詞の類似表現拡張に基づくオープンドメイン音声質問応答システム用言語モデルの構築

ヴァルガ イシュトヴァーン[1] 大竹 清敬[1,a)] 鳥澤 健太郎[1]
デサーガ ステイン[1] 翠 輝久[2] 松田 繁樹[2] 風間 淳一[1]

受付日 2013年10月2日, 採録日 2014年4月4日

概要：本論文では，オープンドメイン音声質問応答システム「一休」で用いる音声認識言語モデル構築手法を提案する．「一休」は，幅広いトピックの比較的短い質問文をスマートフォン経由でユーザから受け取り，大規模な WWW コーパスから答えを探して出力する．オープンドメインの質問を正確に音声認識することを可能にする言語モデルの構築が課題となる．既存のドメインアダプテーションの手法と，名詞の分布類似度に基づくシードコーパスの拡張を組み合わせることで，低コストで高性能の言語モデルを作成した．500 文のシードコーパスと 6 億文の WWW コーパスから 41 万語を網羅する言語モデルを作成した．WWW コーパスからランダムに抽出した文によって構築したベースライン言語モデルを単語誤り率で 3.25%改善した．

キーワード：言語モデル，音声認識，質問応答，分布類似度

Open-domain Language Model Construction for Speech Driven Question Answering Employing Expansion with Similar Nouns

István Varga[1] Kiyonori Ohtake[1,a)] Kentaro Torisawa[1] Stijn De Saeger[1]
Teruhisa Misu[2] Shigeki Matsuda[2] Jun'ichi Kazama[1]

Received: October 2, 2013, Accepted: April 4, 2014

Abstract: This work presents a novel language model construction method for speech recognition, utilized with "Ikkyu", an open-domain speech-based question answering system. Ikkyu accepts relatively short spoken questions concerning a large variety of topics as input through a smartphone, providing the answers retrieved from a large scale Web archive. Our challenge is to construct a language model that can accurately perform speech recognition of open domain questions with smartphones as input devices. We tackle this problem by combining an existing domain adaptation method and distributional word similarity. From 500 seed sentences and a corpus of 600 million Web pages we constructed a language model covering 413,000 words. We achieved an average improvement of 3.25 points in word error rate (WER) over a baseline model constructed from randomly sampled Web sentences.

Keywords: language model, speech recognition, question answering, distributional similarity

1. はじめに

本論文ではオープンドメインの音声質問応答システム「一休」で用いる音声認識言語モデルを WWW から作成する手法を紹介する．「一休」は，幅広い分野の比較的短い質問文をスマートフォン経由でユーザから受け取り，大規模な WWW コーパスから答えを探して出力する．ここで

[1] 情報通信研究機構ユニバーサルコミュニケーション研究所情報分析研究室
National Institute of Information and Communications Technology, Souraku, Kyoto 619–0289, Japan
[2] 情報通信研究機構ユニバーサルコミュニケーション研究所音声コミュニケーション研究室
National Institute of Information and Communications Technology, Souraku, Kyoto 619–0289, Japan
[a)] kiyonori.ohtake@nict.go.jp

表 1 質問応答モジュールが回答できる質問文の例
Table 1 Answerable questions of Ikkyu.

(1)	デフレを引き起こすのは何ですか
(2)	ヤナーチェクが作曲したのは何ですか
(3)	閉塞性動脈硬化症を防ぐのは何ですか
(4)	河津川で何が釣れますか

問題になるのはそのようなトピック分野の範囲を制限しない，オープンドメインの質問を正確に音声認識できるかということであり，それを可能にする言語モデルの構築が課題となる．

従来研究のほとんどでは，ターゲットアプリケーションに合致したドメインおよびスタイルを持つ，人手で整備されたコーパスの存在を前提とし，そこに WWW から類似データを追加することで高性能な言語モデルを作成している [13]．

「一休」はオープンドメインだが，対応可能な質問文に制限がある．これは次の理由による．まず，現状の「一休」の質問応答モジュールは長い質問文に対して高精度で答えることができない．また，音声入力インタフェースで使用されている音声認識システムの長文に対する認識精度が低い．以上の理由から，我々は主に名詞 1 つ，疑問代名詞 1 つ，述語 1 つからなる短い質問文のみを対象とする（**表 1**）．以下では質問文のこのような制限を「スタイル」と呼ぶ．本研究の目的は，上記のスタイルに合致する質問文を，様々なドメインを網羅するように自動で収集してコーパスを構築し，オープンドメインの音声認識言語モデルを構築することである．オープンドメインでかつ大語彙であってもスタイルを限定することで，現在の音声認識器でも実用的な認識性能を達成できると考えた．しかしながら，スタイル限定であってもオープンドメインである以上，言語モデル構築に要するコーパスを手作業で大規模に構築するには大きなコストがかかるため，質問文の自動収集技術が必要となる．

本研究ではまずスタイルに合致する，様々なトピックを網羅する数百文からなるシードコーパスを手作業で構築する．次に大規模な WWW コーパス [15] からシードコーパスと類似している文を収集する．この類似している文の収集を名詞の分布類似度計算手法 [10] と言語モデルのドメインアダプテーション手法 [13] に基づいて行う．

オープンドメイン言語モデル用の大規模な学習コーパスを手作業で構築するのは不可能である．一方で手作業で作成した非常に小さいシードコーパスは少数のトピックしかカバーしておらず，また，それぞれのトピックの質問文数が少ない．既存のドメインアダプテーション手法でも，このようなスパースなシードコーパスを用いて高性能な言語モデルの学習コーパスを収集できるが，オープンドメイン，かつ，スタイル制限ありの設定で同じ効果が得られる保証はない．この問題に対して我々は，トピックの範囲を広げる目的で，シードコーパスにある名詞を統計的に求めた類似の名詞で自動的に置き換えることによってシードコーパスを拡張する．この拡張したコーパスに対してドメインアダプテーションを適用することでより多くのトピックを含み，なおかつ，我々の求めるスタイルに合致した質問文を大量に収集できる．

表 2 テストセットで正しく認識された質問文サンプル（質問応答モジュールが回答できない質問文も含む）
Table 2 Correctly recognised question examples from the test data (May contain questions that can not be answered by Ikkyu).

(1)	はやぶさは何年ぶりに地球に帰還した
(2)	最近発売されたソニーの学習リモコンの型番は
(3)	板付遺跡はどこにありますか
(4)	源頼朝の弟の名前は何ですか
(5)	東京ディズニーランドの最寄り駅はどこですか
(6)	5 月の誕生石を教えて下さい
(7)	熱中症の初期症状は
(8)	国勢調査は何年おきに実施される
(9)	ステロイドの副作用にはどんな物がありますか
(10)	かいけつゾロリの作者はだれ
(11)	ウインブルドンで優勝した人はだれ
(12)	ルイ１４世の業績は何ですか
(13)	日本で iPhone はどれ位売れていますか
(14)	ポストモダンとは何ですか
(15)	Java の最新バージョンは

提案手法の言語モデルの語彙数は 41 万語で，単語誤り率は 15.49%であり，文誤り率は 54.73%である．この値は WWW コーパスからランダムに抽出した文によって構築したベースライン言語モデルより単語誤り率で 3.25%，文誤り率で 4.28%低い．

表 2 に我々の構築した言語モデルによって正しく認識された質問文の例をあげる．幅広い範囲のドメインに関する様々な質問文が正しく認識されていることが分かる．

2. オープンドメイン音声質問応答システム

本研究の提案手法で構築した言語モデルをオープンドメインの音声質問応答システム「一休」の音声認識モジュールで利用する．提案手法のタスク設定を明確にするため「一休」について説明する．

「一休」はスマートフォンにより音声で入力された幅広い範囲のドメインの質問文に対応可能な次世代情報システムである．**図 1** は入力質問文**「デフレを引き起こすのは何ですか」**の回答を表示しているスクリーンショットである．回答は 6 億ページの WWW コーパス [15] から数秒で自動的に抽出され，表示された回答にタッチするとそれぞれの情報抽出源である WWW 上の文が表示される．表 1 は実際に回答できる例を示している．

図 1 「一休」のスクリーンショット
Fig. 1 Screenshot of Ikkyu.

たとえば，図 1 の質問（「デフレを引き起こすのは何ですか」）に対して「一休」は，図 1 に含まれていないが，日本の代表的な大企業を回答した．情報抽出源のブログによると，この会社は数兆円の利益を上げたが，それを投資ではなく，貯蓄に回したため，日本全体の総需要が縮小しデフレが悪化した，とある．この回答を発見した後，著名な経済雑誌でほぼ同主旨の記事が掲載された．

「一休」の目的は，いつでもどこでも，上記の例のような，日常のふとした思いつきから，意外でありながら有用な情報を発見することを可能とし，ふだんの思考のオプションを広げることである．この目的のためには音声による容易な入力が重要であり，このための高性能な音声認識器が必要となる．

質問応答モジュールはパターンに基づく関係抽出手法 [7] を応用している．入力の質問文からパターンを抽出し，パターンとその自動推定されたパラフレーズを文書にマッチさせて回答を見つけ出す．たとえば，上記の質問文「デフレを引き起こすのは何ですか」から「X を引き起こすのは Y」というパターンが抽出され，「Y が X を引き起こす」や「X の原因は Y」のようなパラフレーズが推定される．変数「X」と「Y」はトピックに対応する名詞と疑問代名詞に相当する．これらのパターンの 2 つの変数の一方にトピックに対応する名詞を代入し（上記の例の場合は X =「デフレ」），大規模な WWW コーパスから作成しておいたパターンデータにマッチさせる．疑問代名詞に相当する変数 Y にマッチした名詞が回答として出力される．さらに，前もって自動学習した推論規則を適用することによって複数のパターンを組み合わせて一文に記載されていない回答も見つけ出す．より詳しくは文献 [6] を参照されたい．

「一休」はこのようなアーキテクチャを採用しているため，回答可能な質問文は，パターンで表されるものに限定される．なお，統計的手法で高精度にパラフレーズを推定するため，「一休」が利用可能なパターンは高頻度のものに限定されており，現在は前もって大規模な WWW コーパスから抽出された 7 千万個のパターンに限定されている．これらのパターンは幅広いドメインに関する質問に対応できると考えられるが，質問文中の名詞，述語などの数はこれらのパターンによって制約を受け，前述した「スタイル」の質問文しか受け付けられない．

本研究では人手で作成したシードコーパスを出発点として言語モデルを構築する．この方法は質問応答システムの構造と直接に依存していないため，他の質問応答システムにおいて扱うトピックの分野に広がりを持たせたい場合にも適用できる．

3. 既存手法

本章では，本研究で用いる統計的アダプテーション手法，および語の分布類似度について説明する．

3.1 統計的アダプテーション手法

本研究では語の分布類似度に基づくシードコーパスの拡張を Misu ら [13] の統計的アダプテーション手法と組み合わせる．Misu らはシードコーパス S から抽出した TF-IDF 値が高いクエリによって WWW から類似している文を収集した．次いで，収集された文の中から以下で定義される類似度の高い文をシードコーパスに追加する．類似度はシードコーパスに対する単語パープレキシティ（*score*）によって計算する．

$$score = 2^{-\frac{1}{n}\sum_{i=1}^{n}\log_2 p(w_i|w_{i-1},w_{i-2})} \quad (1)$$

score が θ 以下の文をシードコーパスに追加し[*1]，学習コーパス T を構成する．未知語が含まれている 3-gram の *score* はシードコーパス中で最小の 3-gram 確率とする．

3.2 語の分布類似度

分布仮説 [9] は「似た文脈に出現する語は似た意味をもつ」という仮説であり，これに基づいて計算した語の間の意味的な類似度を語の分布類似度という．これまで，この分布仮説に基づいて様々な分布類似度が提案されたが，本研究では，Kazama ら [10] が提案した分布類似度を用いた．それは，データスパースネス問題を軽減するために，ベイズ推定の手法を取り入れている．まず，元となる類似度として，Bhattacharyya 係数を考える．これは，確率分布間の類似度を測る係数の 1 つであり，式 (2) で定義される．

[*1] *score* の値が低い方が類似度が高い．

表 3 語の分布類似度による類似語の例
Table 3 Similar nouns constructed using distributional similarity.

語	上位 10 語の類似語
魚	花，野菜，物，動物，鳥，肉，卵，犬，水，猫
デフレ	インフレ，不況，少子化，円高，空洞化，円安，人口減少，混乱，危機，環境破壊
骨粗鬆症	骨粗しょう症，痛風，高脂血症，高血圧，動脈硬化，糖尿病，歯周病，生活習慣病，緑内障，心筋梗塞
金閣寺	銀閣寺，清水寺，姫路城，大阪城，六園，名古屋城，首里城，善光寺，法隆寺，平等院

$$BC(p_1, p_2) = \sum_{k=1}^{K} \sqrt{p_{1k} \times p_{2k}} \quad (2)$$

ここで，p_1，p_2 は，与えられた 2 つの語 w_1 と w_2 に対する条件付き文脈分布 $p(f_k|w_1)$，$p(f_k|w_2)$ である．文脈 f_k としては，各々の語に対して観測される係り受け関係を用いる．たとえば，「鮪」に対しては，「が 泳ぐ」「の ヒレ」などが文脈となる．Kazama ら [10] の手法では，条件付き文脈分布 $p(f_k|w_1)$ に対して Bhattacharyya 係数をそのまま適用するのではなく，ベイズ推定の手法を利用して，まず条件付き文脈分布自体の曖昧さを考慮した分布を求め，その分布の下で，元の Bhattacharyya 係数の期待値を計算する．Kazama ら [10] では，大規模な日本語 WWW データを用いた実験で，この分布類似度が多くの既存の類似度に比べて優れた性能を持つことが示されている．

表 3 に Kazama らの手法で取得した類似語の例をあげる．

4. 提案手法

図 2 に提案手法のシステム構成を示す．提案手法は，シードコーパス S と WWW コーパス W を入力として受け取り，以下のように処理を進める．

ステップ 1 S のすべての文 s における名詞 w のうちストップワードリスト L に存在しないものを Kazama ら [10] の分布類似度上位 k 単語と置き換え，新しい文を作成しそれを S に追加する．文が追加されたシードコーパスを S' とする．

ステップ 2 S' と WWW コーパス W に Misu ら [13] の手法を適用し，学習コーパス T を構築する．

ステップ 3 学習コーパス T から音声認識用言語モデルを作成する[*2]．

たとえば，「痛風の症状は？」が S にあり，「痛風」も「症状」も L になく，かつ，「痛風」と「症状」それぞれの上位 k 類似語に「骨粗鬆症」と「原因」が含まれていると仮定する．この場合，ステップ 1 で「骨粗鬆症 の症状は？」や「痛風の 原因 は？」が S に追加される．ステップ 2 では，

[*2] 実験で利用した言語モデル作成ツールおよびパラメータについては，5.2 節を参照のこと．

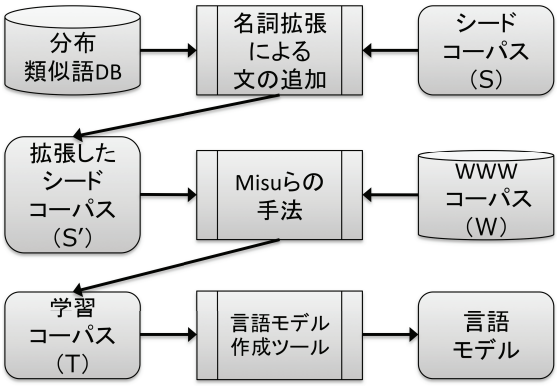


図 2 提案手法の概略
Fig. 2 Outline of our proposed method.

Misu 手法により，たとえば「骨粗鬆症の原因は？」が W から抽出され，T に追加される．この新しい疑問文は S に追加した文（「骨粗鬆症の症状は？」，「痛風の原因は？」）と共通の 3-gram（「〈文頭〉 骨粗鬆症 の」，「の 原因 は」，「原因 は ？」）を持っているため，比較的低い *score* で抽出される可能性が高い．一方で，オリジナルの「痛風の症状は？」は抽出された文「骨粗鬆症の原因は？」と共通の 3-gram をまったく持たないので，オリジナルの文からこうした抽出が直接起こる可能性は低い．「痛風」が「骨粗鬆症」，あるいは「症状」が「原因」に変換されることによって新しい文がシードコーパスに追加され，共通している 3-gram ができたため，この文は抽出される可能性が高い．これは本手法の利点を示している．

実験では約 500 文のシードコーパスを利用した．このコーパスは後述の指示（5.1.2 項を参照）によって手作業で構築した．WWW コーパスとして 6 億ページの大規模なコーパスを利用した [15]．ストップワードリストは約 2 千あり，WWW コーパスに出現する頻度 1 千万以上の名詞からなっている．これらの名詞（「もの」，「あなた」，「場」，「方」，など）は用法が多様であったり，非常に曖昧であったりするため，類似している単語と置き換えると意味的に不自然な文が作成される可能性が高い．

語の分布類似度は 1 億ページの WWW コーパスから計算した．

Misu ら [13] の手法を適用したときに以下の調節を行った．Misu らの手法では TF-IDF によって抽出したクエリ単語に基づいて WWW コーパスを検索し，文を抽出するが，「一休」が対応している質問文はドメインが限定されていないため，高頻度な口語調の単語も対象となる．TF-IDF に基づくクエリによる検索の結果にこのような高頻度の単語が含まれていない可能性が高いため，検索はせずに WWW コーパスのすべての文に対して *score* を計算した．また，シードコーパスの最小 3-gram 確率が高かったため，未知語が含まれている 3-gram の *score* を 10^{-10} に設定した．以下，本論文では，Misu らの手法において，上記 3-gram

値の設定で検索をせずに *score* を全文に対し求める手法を「Misu 手法」と呼ぶ．

5. 実験

5.1 コーパス

5.1.1 WWW コーパス

音声認識用の言語モデル学習のため以下の 2 種類の WWW コーパスを準備した．

www：WWW 6 億ページ [15] から日本語として許される文字・記号類以外の文字を含む文，あるいはアルファベットのみからなる文などを除いた約 13 億文（179 億形態素）のコーパス．

wwwq：www はオープンドメインだが，スタイルに関する処理は行っていない．したがって，疑問文だけでなく，肯定文なども含まれている．そこで文末が「か」，「かい」，「かしら」，「かな」，または疑問符「？」である文を疑問文として www から抽出した．さらに，ほぼ疑問文と同じ意味を持つ要求を表す文として，「下さい」あるいは動詞の連用形+「て」で終了する文も同様に抽出した．選択された文によるコーパス wwwq は約 1 億文（12 億形態素）からなる．

wwwq では「一休」が対応していない「なぜ」，「どうして」，「どうやって」や Yes/No 疑問文も含まれている．また，上記の wwwq の条件のみでは，疑問文に限定するのは困難なため，要求（「〜して下さい」，「〜を見て」など）を示す表現も wwwq には含まれる．一方，「一休」が対応できる，疑問代名詞が省略されている文が wwwq に少ない．wwwq は我々が求めているスタイルに www を限定しようとしたものであるが，このようなスタイルが合致しない文を含んでいる．

提案手法の利点は，www から wwwq を作成するように，求めるスタイルに近いものを集めるために不完全な制約をいろいろ考えるよりも，スタイルに合致する小さなシードコーパスと統計的方法によって求めるスタイルの学習用コーパスがより容易に大量抽出できる点である．

5.1.2 シードコーパスと評価セット

女性 25 名，男性 25 名，合計 50 名により，1 人あたり質問文約 50 文を自由に作成，発話してもらい，スマートフォンで収録した．上記の 50 名は，できるだけ様々なトピックを網羅する，名詞・述語・疑問代名詞（**何，誰，どこ，いつ**）から成り立つ比較的単純な文を作成する指示を受けた．また，疑問代名詞の「**教えて**」や「**教えて下さい**」のような要求へのいい換え，あるいは疑問代名詞の省略も可能とした．ただし，これらの指示に合致しない質問文も作成された．**表 4** に作成された質問文の例をあげる．

作成した質問文をランダムに 3 つに分離した．g0 に話者が 10 名，g1 と g2 にそれぞれ話者が 20 名の質問文が含まれている（**表 5**）．g0 を書き起こしたテキストをシードコーパスとして利用する．

表 4 発話者が作成した質問文の例
Table 4 Examples of the constructed sentences.

(1)	頭痛に効く薬は何ですか
(2)	世界で一番標高の高い湖はどこですか
(3)	１００メートル走で一番速い人は誰
(4)	オーロラがよく見られる所を教えて
(5)	ワールドカップで優勝した国はどこ
(6)	ドイツの時計メーカーは
(7)	お腹が空くとお腹が鳴るのはどうしてですか

表 5 シードコーパスと評価セット
Table 5 Seed corpus and evaluation set.

グループ	**g0** (S)	**g1**	**g2**
発話数	498	1,000	999
形態素数	4,043	7,671	8,322
平均 形態素/発話	8.118	7.671	8.330
1-gram 数	1,096	1,834	1,784
2-gram 数	2,438	4,350	4,330
3-gram 数	2,848	5,293	5,403

5.2 音声認識器 ATRASR

実験では質問応答システム「一休」が用いる音声認識器 ATRASR [11] を利用した．ATRASR は，音響モデルとして隠れマルコフモデル（Hidden Markov Model: HMM）を用い，第 1 パスで単語 bi-gram による単語ラティスの生成，第 2 パスで単語 tri-gram によるリスコアリングおよび最終認識結果の探索を行う．本論文の実験では語彙サイズの制限は行っていないが，bi-gram と tri-gram のカットオフ値を 1 に設定した．また，ATRASR の実装上の制約により www コーパスすべてから言語モデルを作成できなかった．処理可能な最大学習コーパスサイズは 31 億形態素であった[*3]．

5.3 実験結果

5.3.1 最適な分布類似度ランク k の推定

実験では，シードコーパス中の単語をその単語の分布類似度上位 k の単語で置き換えた．最初に最適な k を推定した．k の値を 1，2，3，4，5，10，15，20，100 に設定し，それぞれの k において文を追加したシードコーパスを用意し，本研究で呼ぶところの Misu 手法によって学習コーパスを作成するが，Misu 手法による *score* が大きい文から順に学習コーパスに含めることとし，学習コーパスの量を 1 千万，2 千万，4 千万，8 千万，…，12.8 億形態素と学習コーパスサイズを倍増させて複数の学習コーパスを用意した．したがって，1 つの k に対して異なるサイズの複数の学習コーパスが用意される．k の値と性能との関係を g1 と g2 を用いて評価した．当初は，g1 を開発セット，g2 を評価セットとする場合とその逆の組合せの両方の設定で実験を行う予定であったが，各コーパスが発話単位ではな

[*3] 本研究の実験ではメモリが 72 GB のマシンを利用．

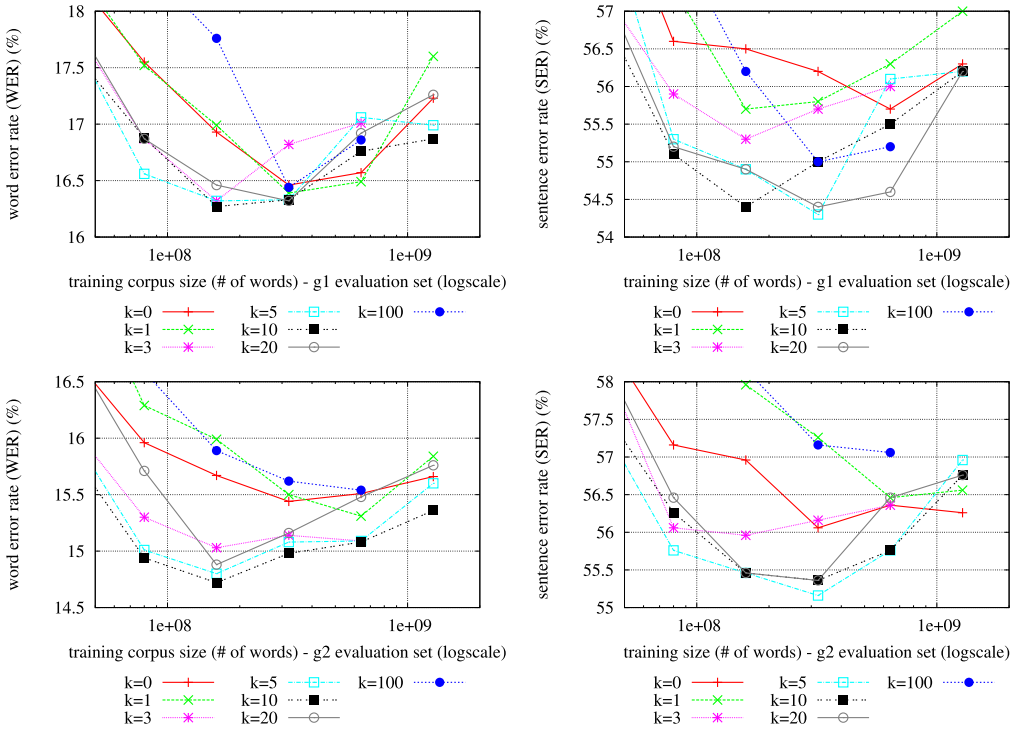


図 3 分布類似度上位 k 語によって拡張し，www から学習用コーパスを作成した場合の単語誤り率と文誤り率

Fig. 3 WER (word error rate) and SER (sentence error rate) curves with various distributional similarity rank k, trained on the www corpus.

く，話者を単位として分割されていることからばらつきが一定程度あると考え，パラメータである k と学習コーパスサイズについて組合せすべてを実験し，評価した．**図 3** は各 k に基づいてシードコーパスを拡張し，www コーパスから文を抽出し，学習コーパスを大きくしたときの単語誤り率と文誤り率を示している．提案手法の中で単語誤り率が最も低いモデルは g1，g2 のいずれにおいても学習コーパスサイズが 1.6 億形態素で $k = 10$ のときに得られた（図 3 左側）．この言語モデルの語彙数は 41 万語であった．その上，$k = 10$ の言語モデルは他のほとんどの学習コーパス量においても最適な設定となっている．この一貫性は提案手法の有効性を示している．

$k = 0$ の場合は分布類似度を用いた名詞拡張による文の追加がされないため，シードコーパスに Misu 手法を適用しただけである．$k = 10$ という設定の提案手法の最も小さい単語誤り率は，Misu 手法（$k = 0$）のみを適用した場合の最も小さい単語誤り率より g1 においては 0.19%，g2 においては 0.72%改善した．McNemar テスト [12] でこの差が統計的に有意であることを確認した（$p < 0.05$）．g1 において差が比較的小さいが，(1) g1，g2 のいずれにおいても提案手法（$k = 10$）の最も低い単語誤り率の学習コーパスサイズ（1.6 億形態素）は，Misu 手法のみを適用した場合（$k = 0$）に最も低い単語誤り率を達成した際の学習コーパスサイズ（3.2 億形態素）の半分であった．(2) 同一の学習コーパスサイズという観点から，提案手法（$k = 10$）において単語誤り率が最も低かった場合のサイズ（1.6 億形態素）において $k = 10$ の結果と $k = 0$ の結果を比較すると，その差は，g1 においては 0.66%，g2 においては 0.95%になった．これは我々の提案手法が Misu 手法単体をシードコーパスに適用するより効果的であることを示している．

言語モデルの学習コーパスに含める文を抽出するコーパスを www から wwwq に変更した場合も同じ傾向が見られた．最も低い単語誤り率は分布類似度ランク $k = 10$ で得られた（g1 においては 16.35%，g2 においては 15.28%）が，www コーパスを用いた場合に比べて性能の改善は見られなかった．

5.3.2 提案手法とベースラインの比較

我々の最適な設定（$k = 10$，学習コーパスサイズが 1.6 億形態素）の提案手法によるモデルとランダムサンプルによるモデルをベースラインとして比較した．

- **www.X**：www コーパスと提案手法で作成した言語モデル．
- **wwwq.X**：wwwq コーパスと提案手法で作成した言語モデル．
- **www.R**：www コーパスからランダムサンプルして作成した言語モデル．
- **wwwq.R**：wwwq コーパスからランダムサンプルして作成した言語モデル．

図 4 は実験結果を示している．

ベースラインの最も低い単語誤り率は，学習コーパスサイズが用いることができる最大のときに得られた．提案

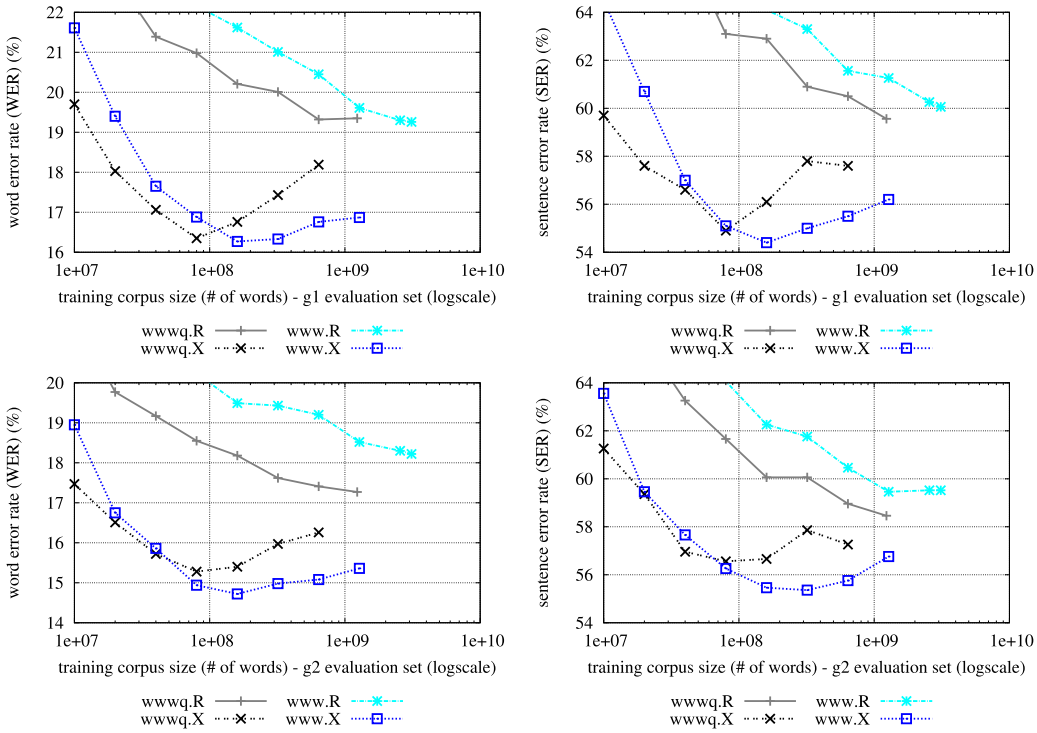


図 4 提案手法とベースラインの比較
Fig. 4 Our proposal compared with the baselines.

手法の単語誤り率（g1 においては 16.27%，g2 においては 14.72%）は最も性能が良いベースライン（www）の単語誤り率を g1 と g2 の平均で 3.25%改善した．この差も統計的に有意である（$p < 0.01$）．文誤り率の場合も，両テストセットともに提案手法の言語モデルが最も低い値を示している．提案手法の g1 においては 54.30%，g2 においては 55.16%の文誤り率は最も性能が良いベースライン（wwwq）の値を g1 と g2 の平均で 4.28%改善する．この差も統計的に有意である（$p < 0.01$）．

学習コーパスを大きくしてもベースラインの性能が鈍化する傾向を示していないため，学習コーパスをさらに大きくすることによって性能が向上する可能性がある．しかし，学習コーパスを大きくすることによって言語モデルの語彙数や n-gram 数も増加するため，実行速度に大きく影響する可能性が高い．**図 5** は学習コーパスのサイズと語彙数および bi-gram 数の関係を表している．ベースラインモデルの bi-gram 数は提案手法の bi-gram 数より 2 倍以上多いことを示している．これはスマートフォン上の音声質問応答システムとしては致命的である．たとえば，単語誤り率が最も低いベースラインモデルの学習コーパスサイズは，提案手法のそれよりも 8 倍程度大きいため，音声認識が非常に遅くなる可能性が高い．**図 6** は RTF 値[*4]を示している．学習コーパスサイズが最も小さい時点で，提案手法によるモデルが 1.0，ベースラインモデルは 1.7 の RTF であるが，学習コーパスが大きくなるにつれて，その差が開くことが見てとれる．つまり，ベースラインモデルでは，音声認識の性能を上げるために大きな学習コーパスを用いることで，単語誤り率は改善するが，それ相応の速度を犠牲

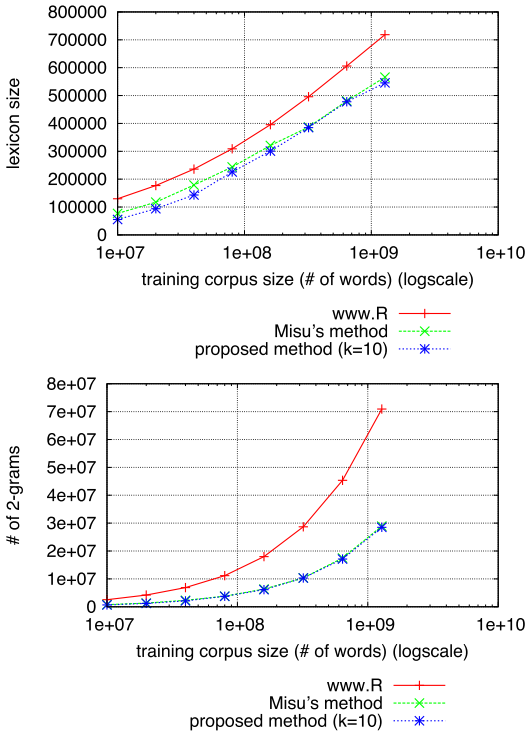


図 5 語彙数および bi-gram 数と学習コーパスサイズの相関関係
Fig. 5 Lexicon size and 2-gram count in function of training corpus size.

[*4] RTF (“real time factor”) は，音声認識における性能指標の 1 つで認識する音声の長さを x 秒とし，その音声認識にかかる時間を y 秒とするとき y/x で計算される．

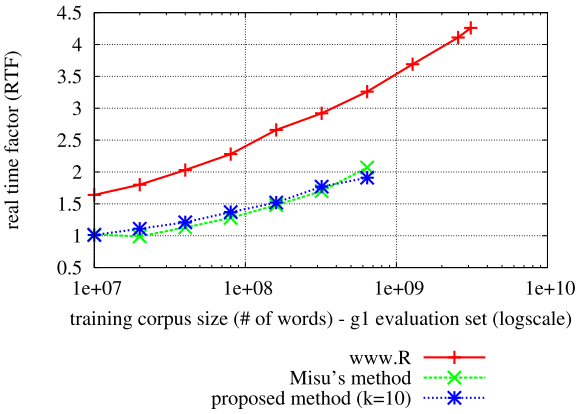


図 6 RTF と学習コーパスサイズの相関関係
Fig. 6 RTF in function of training corpus size.

図 7 「一休」のラティス構造を用いるエラー回復機能の例
Fig. 7 Ikkyu's word lattice based error recovery interface with "What causes deflation" as input.

にすることになる．

5.3.3 N ベスト評価

音声質問応答システム「一休」はエラー回復機構として N ベスト結果から擬似的なラティス構造を作成し，そこから正しい音声認識結果を効率的に選択するインタフェースを備えている．図 7 は「デフレを引き起こすのは何ですか」のエラー回復の例を示す．そのため，実際の使用感覚に近い評価指標として 100 ベストの中に正解があったかどうかを上記の文誤り率の最も低かった条件で計算すると次のようになった．g1 においては 62%の入力に対して 100 ベスト中に正解があり，g2 においては 59%であった．したがって，実用上は約 6 割で完全な認識結果を容易に入力できる．

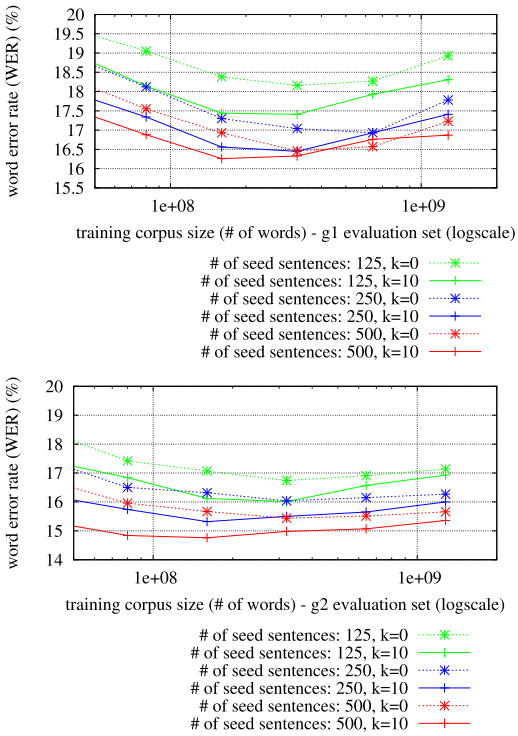


図 8 シードコーパスのサイズを変化させた場合の単語誤り率
Fig. 8 WER curves with various sized seed corpora.

5.3.4 シードコーパスの量

提案手法を適用するよりシードコーパスの量を増やすことによって性能がより向上する可能性がある．そこでシードコーパスの量を変化させた場合の性能を確認した．提案手法，つまり名詞拡張を行ったうえで Misu 手法を適用する場合（$k = 10$）と，シードコーパスに Misu 手法のみを適用する場合（$k = 0$）それぞれについて WWW コーパスには `www` を用いて，5.3.1 項の実験同様に複数の学習コーパスを作成し，評価した．図 8 はシードコーパスのサイズが 125，250，500 文の場合の単語誤り率を示している．まず，g1，g2 のいずれにおいてもほぼすべての学習コーパスサイズで，提案手法を適用した場合に性能向上するといえる．一方，学習コーパスサイズ 1.6 億形態素の場合に限っていえば，$k = 0$ でシードコーパスのサイズを倍にした場合より，同じサイズのコーパスで $k = 10$ とした場合の方が高い性能を示している場合があり，提案手法の一定程度の効果，つまり，シードコーパスのサイズを倍にした場合と同程度の性能向上が見込めることが分かる．

一方で，シードコーパスの質について，今回用いたシードコーパス（g0）とは異なるコーパスを用いた場合の結果については，規模が小さい予備実験の中でその性能に大きな差がなかったことを確認している．

5.4 考察

以上，提案手法が Misu 手法のみを適用した場合に比べ

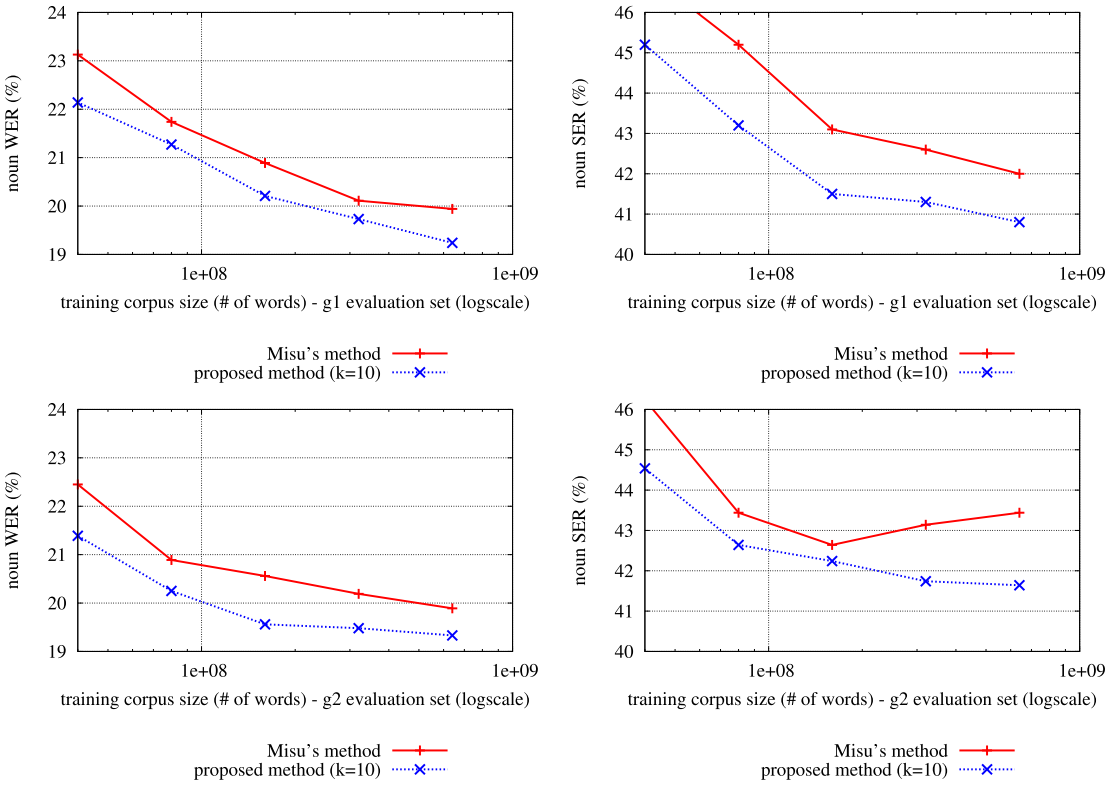


図 9 www コーパスで学習した名詞評価
Fig. 9 Noun evaluation on www.

て性能が向上することは示せた．しかしながら，この性能向上が本当に名詞の置き換えによるのかはより詳細な検討が必要である．1 つの可能性として，名詞の置き換えではなく，名詞の置き換えによって生じた他のタイプの表現，たとえば，疑問代名詞や「を教えて下さい．」などの文末表現の学習コーパス中での頻度向上，繰返しが性能向上の原因である可能性が，これまでの実験結果からは否定できない．以下では，こうした可能性を異なる実験結果の解釈によって否定する．実験結果の異なる解釈によって，これまでに得られた認識結果から名詞のみを考慮した単語誤り率と文誤り率を計算した．もしこれらの値が改善されているのであれば，名詞の置き換えが実際に効果があった可能性がより高くなる．なお，文誤り率の場合は，ある文に現れるすべての名詞が正しく認識されているならば，正解とした．ここで，疑問代名詞は名詞として扱っていない．**図 9** は提案手法の最も低い単語誤り率モデル（$k = 10$）と Misu 手法（$k = 0$）の名詞評価を示している．単語誤り率の場合は，両テストセットともに提案手法から作成した言語モデルが最も低い単語誤り率を示していた．www コーパスで学習した Misu 手法と g1 においては 0.70%，g2 においては 0.56%の差になった．文誤り率の場合は，g1 においては 1.20%，g2 においては 1.80%の差になった．この差も統計的に有意である（$p < 0.05$）．したがって，名詞置き換えが効果的であるといえる．

また，提案手法に類似した手法として，名詞置き換えを

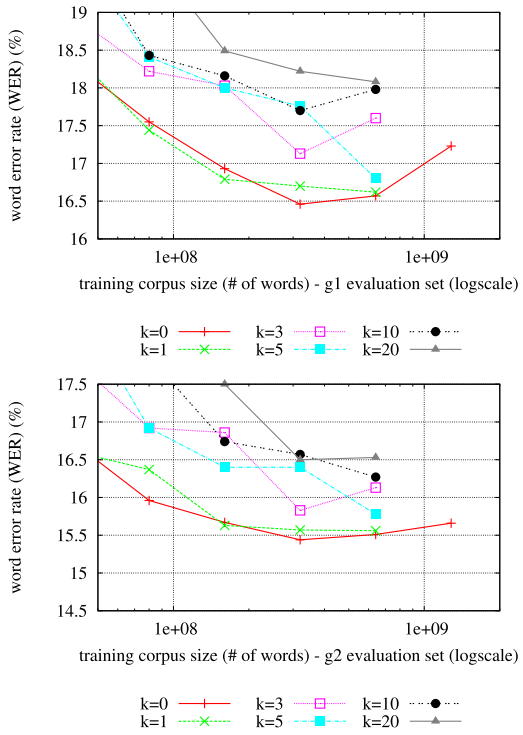


図 10 www コーパスに名詞置き換えを適用した場合の単語誤り率
Fig. 10 WER curves with noun replacement performed on the www corpus.

シードコーパスではなく，www コーパスに適用することも考えられる．**図 10** は名詞置き換えを www コーパスに適用した単語誤り率を表している．ほとんどの学習コーパス

量および分布類似度ランク k においてもベースとなったMisu 手法の性能を下回った．学習コーパス中に名詞置き換えによって非文法的な文，意味的に奇妙な文が大量に生成され，認識の実験においても高い性能は得られなかった．シードコーパスに名詞置き換えを適用する提案手法においても，奇妙な文が生成される可能性はあるが，それは最終的には言語モデルの学習コーパスに含まれないため，性能の低下に直接的にはつながらない．別のいい方をすれば，提案手法の場合は，学習コーパスに現れる文はあくまで人が書いた自然な文だけであり，名詞置き換えによる非文法的，あるいは意味的に奇妙な文は文選択で使われるスコアの計算でのみ使われるため，そのような文の悪影響は直接的には現れない．

6. 関連研究

近年，WWW コーパスを言語モデル学習に用いる研究がさかんである．Berger ら [1] は入力文の内容語をクエリとして利用し，抽出したテキストで学習コーパスを改良した．Zhu ら [16] は学習コーパス中の 3-gram の確率を WWW での出現頻度で学習し直した．Bulyko ら [3] はシードコーパス中の高頻度の 3-gram をクエリとして利用した．Sarikaya ら [14] は類似している WWW テキストを BLEU スコアで選択していた．パープレキシティもよく利用されている類似度計算方法である [5], [13]．Misu らの手法だけでなく，上にあげた他の手法に我々の名詞置き換えフレームワークを適用することも今後の課題となる．

また，我々の手法は単語クラスタリングを用いた確率的言語モデルとも類似している．このアプローチは，Brown ら [2] によって最初に提案され，現在はその改良がいくつか提案されている [4], [8], [17], [18]．Misu らの手法の 3-gram をクラス n-gram に置き換えることによって我々の名詞置き換えと同じ効果が得られる可能性があるが，我々のフレームワークにおいては最善の性能を得た分布類似度の上位 k 語が比較的小さかった（$k = 10$）ことを考えると，クラスタリングベースの確率的言語モデルではクラスタリングの粒度調整が難しくなることが予想される．今後はこうした問題にも決着をつけるべく，比較を行いたい．

7. まとめ

本論文ではオープンドメインの音声質問応答システムで用いる音声認識言語モデルを WWW から作成する手法を紹介した．それは，求めるスタイルで記述された少数のシードコーパスを語の分布類似度計算手法 [10] による名詞置き換えによって拡張し，既存のドメインアダプテーション手法 [13] を適用するものである．最適な設定で構築した学習コーパスから作成した言語モデルは WWW コーパスからランダムに抽出した文によって構築したベースライン言語モデルを単語誤り率で 3.25%，文誤り率で 4.28%改善した．

参考文献

[1] Berger, A. and Miller, R.: Just-in-time language modeling, *Proc. ICASSP 1998*, pp.705–708 (1998).

[2] Brown, P.F., Della Pietra, V.J., de Souza, P.V., Lai, J.C. and Mercer, R.L.: Class-Based n-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467–479 (1992).

[3] Bulyko, I., Ostendorf, M. and Stolcke, A.: Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures, *Proc. HLT2003*, pp.7–9 (2003).

[4] Chen, S.F. and Chu, S.M.: Enhanced Word Classing for Model M, *Proc. Interspeech 2010*, pp.1037–1040 (2010).

[5] Creutz, M., Virpioja, S. and Kovaleva, A.: Web augmentation of language models for continuous speech recognition of SMS text messages, *Proc. EACL 2009*, pp.157–165 (2009).

[6] De Saeger, S., Goto, J. and Varga, I.: Speech-based Question Answering System "Ikkyu", *Journal of the National Institute of Information and Communications Technology*, Vol.59, No.3/4, pp.83–96 (2012).

[7] De Saeger, S., Torisawa, K., Kazama, K., Kuroda, K. and Murata, M.: Large Scale Relation Acquisition using Class Dependent Patterns, *Proc. ICDM 2009*, pp.764–769 (2009).

[8] Emami, A., Chen, S.F., Ittycheriah, A., Soltau, H. and Zhao, B.: Decoding with shrinkage-based language models, *Proc. Interspeech 2010*, pp.1033–1036 (2010).

[9] Harris, Z.: Distributional Structure, *Word*, Vol.10, No.23, pp.142–146 (1954).

[10] Kazama, J., De Saeger, S., Kuroda, K., Murata, M. and Torisawa, K.: A Bayesian Method for Robust Estimation of Distributional Similarities, *Proc. ACL 2010*, pp.247–256 (2010).

[11] Matsuda, S., Jitsuhiro, T., Markov, K. and Nakamura, S.: ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles, *IEICE Trans. Information and Systems*, Vol.E89-D, No.3, pp.989–997 (2006).

[12] McNemar, I.: Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, Vol.12, pp.153–157 (1947).

[13] Misu, T. and Kawahara, T.: A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts, *Proc. Interspeech 2006*, pp.9–13 (2006).

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

[15] Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: TSUBAKI: An open search engine infrastructure for developing new information access, *Proc. IJCNLP 2008*, pp.189–196 (2008).

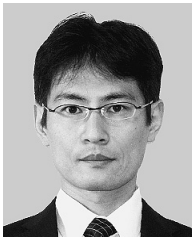
[16] Zhu, X. and Rosenfeld, R.: Improving trigram language modeling with the world wide web, *Proc. ICASSP 2001*, pp.533–536 (2001).

[17] Yamamoto, H. and Sagisaka, Y.: Multi-class composite n-gram based on connection direction, *Proc. ICASSP 1999*, pp.533–536 (1999).

[18] Yamamoto, H., Isogai, S. and Sagisaka, Y.: Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters, *Proc. ACL 2001*, pp.6–11 (2001).

ヴァルガ イシュトヴァーン

2009 年山形大学大学院理工学研究科博士課程修了．同年より情報通信研究機構専攻研究員．2014 年から日本電気株式会社主任研究員．博士（工学）．



大竹 清敬 （正会員）

2001 年豊橋技術科学大学大学院博士後期課程修了．博士（工学）．同年より株式会社 ATR 音声言語コミュニケーション研究所．2006 年より独立行政法人情報通信研究機構を経て現在，同機構情報分析研究室主任研究員および情報配信基盤研究室室長を兼務．音声言語処理，自然言語処理の研究に従事．言語処理学会，人工知能学会各会員．

鳥澤 健太郎 （正会員）

1992 年東京大学理学部卒業．1994 年同大学大学院修士課程修了．1995 年同大学院博士課程中退．同年同大学院助手．1998 年科学技術振興事業団さきがけ研究 21 研究員兼任（2002 年まで）．北陸先端科学技術大学院大学助教授を経て，2008 年より独立行政法人情報通信研究機構言語基盤グループ，グループリーダ．現在，同機構情報分析研究室室長．博士（理学）．自然言語処理の研究に従事．日本学術振興会賞等受賞．言語処理学会，人工知能学会各会員．

デサーガ ステイン

2006 年に北陸先端科学技術大学院大学博士課程修了後，2007 年に情報通信研究機構専攻研究員を経て，2012 年より 2013 年まで同機構主任研究員．知識の自動獲得の研究に従事，言語処理学会第 16 回年次大会優秀発表賞等受賞．博士（知識科学）．

翠 輝久 （正会員）

2008 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．2008 年から 2013 年まで情報通信研究機構専攻研究員．この間，2005 年から 2008 年まで日本学術振興会特別研究員（DC1）．2011 年から 2012 年まで南カリフォルニア大学（USC）Institute for Creative Technologies（ICT）客員研究員．現在は，Honda Research USA, Inc. Scientist．音声言語情報処理，特に対話システムの研究に従事．2007 年度情報処理学会山下記念研究賞受賞．2010 年度日本音響学会粟屋潔学術奨励賞受賞．2012 年度ドコモ・モバイル・サイエンス賞奨励賞受賞．IEEE，電子情報通信学会，人工知能学会，日本音響学会，言語処理学会各会員．

松田 繁樹 （正会員）

2003 年北陸先端科学技術大学院大学博士後期課程修了．同年（株）国際電気通信基礎技術研究所音声コミュニケーション研究所研究員．2009 年情報通信研究機構研究員．2014 年より株式会社 ATR-Trek に勤務．博士（情報）．音声認識に関する研究に従事．日本音響学会会員．2010 年 4 月文部科学大臣表彰受賞．

風間 淳一

2004 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年北陸先端科学技術大学院大学情報科学研究科助教．2008 年から 2013 年まで情報通信研究機構．