オープンドメイン音声質問応答のための類似性に基づく音声認識言語モデル構築

Similarity Based Language Model Construction for Voice Activated Open-Domain Question Answering

ヴァルガ イシュトヴァーン † István Varga
大竹 清敬 † Kiyonori Ohtake
鳥澤 健太郎 † Kentaro Torisawa
デ・サーガ ステイン † Stijn De Saeger
翠 輝久 ‡ Teruhisa Misu
松田 繁樹 ‡ Shigeki Matsuda
風間 淳一 † Jun'ichi Kazama

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で用いる音声認識言語モデル構築手法を提案する．Web の大量の文から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．

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．ここで問題になるのはそのような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，そうした認識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及びスタイルを持つ，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．

「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」の質問回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，音声入力インター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，我々は主に名詞一つ，疑問代名詞一つ，述語一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，様々なトピックを網羅するように自動収集してコーパスを構築し，そのコーパスから高精度で対応できるトピックの広い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あり，したがって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 reasonable 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．

図 1: 「一休」のスクリーンショット

†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, Universal Communication Research Institute, Information Analysis Laboratory
‡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, Universal Communication Research Institute, Spoken Language Communication Laboratory

(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

表 1: 質問応答モジュールが回答できる質問文の例

(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)	板付遺跡はどこにありますか
(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

表 2: テストセットの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み）

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成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 [14] 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文脈類似度計算手法 [9] と言語モデルのドメインアダプテーション手法 [12] 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．我々の「スタイル」に比べて，扱うトピックの範囲がより狭いと考えられるドメインに関しては，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ことがわかっている．しかし，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な文脈の類似度が近い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．結果としてドメインアダプテーション手法は WWW コーパスからより効果的に幅広いトピックを含み，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．

音声認識器 ATRASR[10] を利用した実験で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 3.25 ポイント（単語誤り率），及び 4.28 ポイント（文誤り率）低い誤り率である．

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挙げる．幅広い範囲のトピック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

タスク設定を明確にするためにオープンドメインの音声質問応答システム「一休」について説明する．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マートフォンの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 [14] から数秒で自動的に抽出される．表示された回答にタッチするとそれぞれの情報抽出原である 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．

例えば，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」は日本の代表的な大企業を回答した．情報抽出原のブログによると，この会社が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．ちなみに，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．

オープンドメインの音声質問応答システム「一休」の目的は，いつでもどこでも，上の例のように，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，普段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，このための高い性能の音声認識器が必要となる．本研究の背景を一言で述べると以上のようになる．

質問応答モジュールはパターンに基づく関係抽出手法 [6] の拡張である．入力の質問文をパターンに変換し，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．例えば，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」はトピックと疑問代名詞に相当する．これらのパターンにトピックを代入し（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現される．

「一休」はこのようなアーキテクチャを採用しているため，適応可能な質問文は，パターンで表されてい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，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 WWW 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トピック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，述語等の数はこうしたパターンによって制約を受け，前述したような「スタイル」の質問文しか受け付けられないということになっている．

3 背景

本研究で利用している統計的アダプテーション手法，及び文脈類似度を説明する．

3.1 統計的アダプテーション手法

本研究では文脈類似度に基づくシードコーパスの拡張を翠ら [12] の統計的アダプテーション手法と組み合わせる．翠らはシードコーパス S から抽出した TF-IDF 値が高いクエリーによって 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．

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

$score$ が θ 以上の文をシードコーパスに追加し，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 [8] 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類似尺度が提案されてきているが，本研究では，風間ら [9] が提案された類似尺度を用いた．この類似尺度では，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ひとつであり，下の式で定義される．

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

ここで，p_1，p_2 は，与えられた二つの語 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．例えば，**「鮪」**に対しては，**「が 泳ぐ」「の ヒレ」**等が文脈となる．風間ら [9] 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ということを行う．風間ら [9] では，大規模な日本語 WWW データを用いた実験で，この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．

4 提案手法

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 に対してストップワードリスト L に存在しない名詞 w を風間ら [9] の文脈類似度上位 k 単語と置き換える．新しい文を S に追加する．

ステップ 2 新しいシードコーパス S と WWW コーパス W に翠ら [12] の手法を適用し，学習コーパス T を構築する．

ステップ 3 既存ツールを利用して学習コーパス T から音声認識用言語モデルを作成する．

例えば，**「痛風の症状は？」**がシードコーパス S にあり，**「痛風」**も**「症状」**もストップワードリスト 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，翠らの手法により，例えば**「骨粗鬆症の原因は？」**が WWW コーパスから抽出され，学習コーパス T に追加される．この新しい疑問文はシードコーパスに追加した文（**「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**「通風の症状は？」**は抽出された文**「骨粗鬆症の原因は？」**と共通の 3-gram を全く持たないので，こうした抽出が起きる可能性は低い．**「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たため，共通している 3-gram がある．そうでない場合は，共通している 3-gram がないてめ，この文は抽出される可能性が低い．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 S を利用した．このコーパスは後述の指示（5.1.2 節を参照）によって手作業で構築した．WWW コーパス W として 6 億ページの大規模なコーパスを利用した [14]．ストップワードリストは約 2 千あり，WWW コーパスの頻度 1 千万以上の名詞から成っている．これらの名詞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．

名詞間の分布類似度は 1 億ページの WWW コーパスから計算した．

翠らの手法 [12] を適用した時に以下の調節を行った．翠らの手法では TF-IDF によって抽出したクエリーに基づいて WWW コーパスから類似している文が抽出されるが，「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 による検索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，TF-IDF によるクエリー抽出を行わなかった．その代わり検索結果ではなく，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．

5 評価

5.1 評価設定

5.1.1 コーパス

音声認識用の言語モデル学習のため以下の二つの WWW コーパスを準備した．

- **www** WWW6 億ページ [14] 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フィルタリングした後約 179 億形態素（13 億文）のコーパスを得た．この第一のコーパスを `www` と呼ぶ．
- **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．従って，疑問文のみならず，肯定文など，質問応答システムが対応していない文も含まれている．第二のコーパスとして，簡単なフィルターによって `www` から疑問文のみを抽出した．このフィルターでは**「か」**，**「かい」**，**「かしら」**，**「かな」**，または疑問符**「？」**で終了する文を疑問文として抽出する．

  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」**で終了する文も抽出した．選択された文によるコーパス `wwwq` は約 12 億形態素（1 億文）から成っている．

`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が，疑問代名詞が省略されていることが多いためすべてのこのような文の削除は困難である．提案手法によって，このような問題のある文が削除されるものと期待する．

5.1.2 評価セット

女性 25 名，男性 25 名，合計 50 名により，一人当たり質問文約 50 文を自由に作成してもらい，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」**のような要求への言い換え，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．

作成してもらっ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 3）．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．g1 と g2 を利用して提案手法のパラメータ推定あるいは評価を 2 分割交差検定で行う．

グループ	**g0** (S)	**g1**	**g2**
発話数	498	1000	999
形態素数	4043	7671	8322
平均 形態素/発話	8.118	7.671	8.330

表 3: シードコーパスと評価セット

5.2 実験結果

5.2.1 最適な文脈類似度ランク 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文脈類似語で置き換えたが，文脈類似度の上位 k 位までの文脈類似語で置き換えた．最初に最適な文脈類似度ランク k を推定した．k の値は 1，2，3，4，5，10，15，20，100 に設定した．それぞれの k によるシードコーパスの拡張を行い，翠らの手法のしきい値 θ をすこしずつ増やして行きながら，言語モデル構築のための学習コーパスの量を 1 千万形態素から徐々に増加させた．図 2 は `www` コーパスで学習した文脈類似度ランク k を変化させた時の単語誤り率を示している．提案手法の最も低い単語誤り率は g1，g2 のいずれにおいても学習 1.6 億形態素分の学習コーパスで $k = 10$ の時に得られた（図 2）．この言語モデルの語彙数は 41 万語であった．その上，$k = 10$ の文脈類似度ランク設定の言語モ

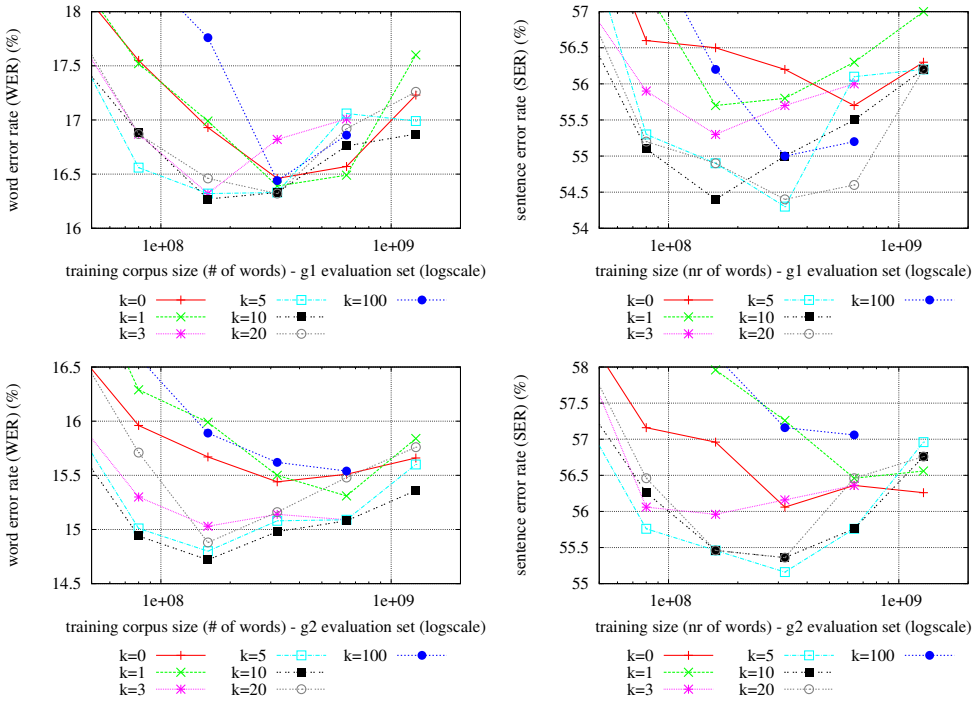

図 2: www コーパスで学習した脈類似度ランク k による単語誤り率

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．さらに，最適な文脈類似度ランクが比較的小さいことは，提案手法が名詞の置き換えに比較的敏感であることを示している．

$k = 0$ の場合は文脈類似度を考慮しないため，ベースとなった翠らの手法の我々の適用である．$k = 10$ という設定の提案手法は翠らの手法より g1 においては 0.19 ポイント，g2 においては 0.72 ポイント改善した．McNemar テスト [11] でこの差が統計的に優位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) 提案手法の最も低い単語誤り率の学習量は翠らの手法の学習量の半分であった．(2) 翠らの手法にも提案手法の最も低い単語誤り率モデル（$k = 10$）の学習量と同量の学習データを利用すると，差が g1 においては 0.66 ポイント，g2 においては 0.95 ポイントになった．(3) $k = 10$ の最も低い文誤り率は翠らの手法の最も低い文誤り率を g1 においては 0.90 ポイント，g2 においては 1.90 ポイントで改善した．これは我々の提案手法が翠らの手法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．

5.2.2 提案手法とベースラインの比較

我々の最適な設定の提案手法（文脈類似度ランク $k = 10$，www コーパスの 1.6 億形態素）とベースラインを以下のモデルで比較した．

- **www.X：** www コーパスで学習した提案手法．
- **wwwq.X：** wwwq コーパスで学習した提案手法．
- **www.R：** www コーパスからランダムサンプル．
- **wwwq.R：** wwwq コーパスからランダムサンプル．

図 3 は実験結果を示している．

メモリー制限のため音声認識器 ATRASR[10] は www コーパス全体を処理できなかった．(メモリーが 72GB のマシンを利用．) 処理できる最大学習量は 31 億形態素分であった．

提案手法以外のベースライン手法の最も低い単語誤り率は最大の学習量で学習する際に得られた．提案手法の単語誤り率（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 ポイントで改善した．この差も統計的に優位である（$p < 0.01$）．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 ポイントで改善する．この差も統計的に優位

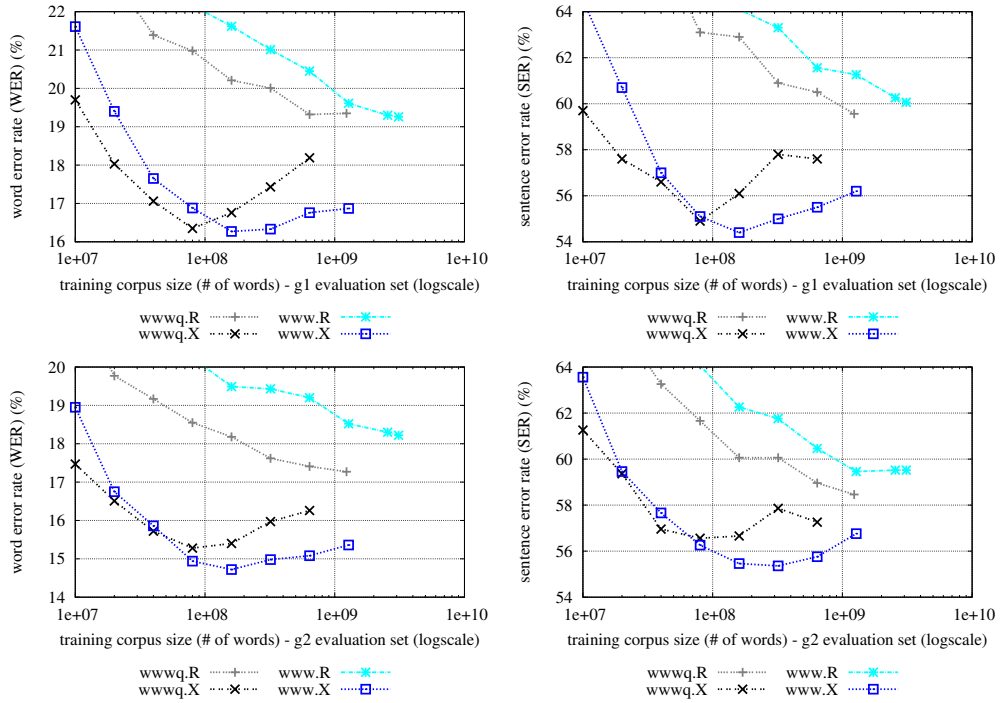

図 3: 提案手法とベースラインの比較

である（$p < 0.01$）．

データの量を増やして行っ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，学習量をさらに増やすことによって性能が向上する可能性がある．しかし，現在の単語誤り率が最も低いベースラインモデルの学習量が我々の提案手法の最も性能が良い設定の学習量より 8 倍程度多いため，音声認識が非常に遅くなる可能性が高い．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．図 4 は RTF 値[1]を示している．現在の設定ですでにベースラインのモデルは提案手法より 2.8 倍程度遅いため，我々の手法が我々の目的にはより適切なことを示している．

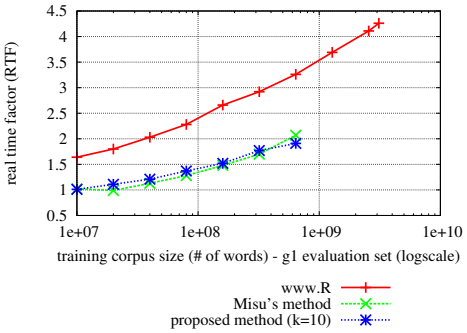

図 4: RTF と学習量の相関関係

[1]RTF (“real time factor”) は，音声認識における性能指標の一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 秒とする時 y/x で計算される．

5.2.3 N ベスト評価

音声質問応答システム「一休」に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．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．従って，実用上は約 6 割で完全な認識結果を容易に入力できる．

5.3 考察

以上，提案手法が翠らの手法のみを適用した場合に比べ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．一つの可能性として，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，例えば，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，繰り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

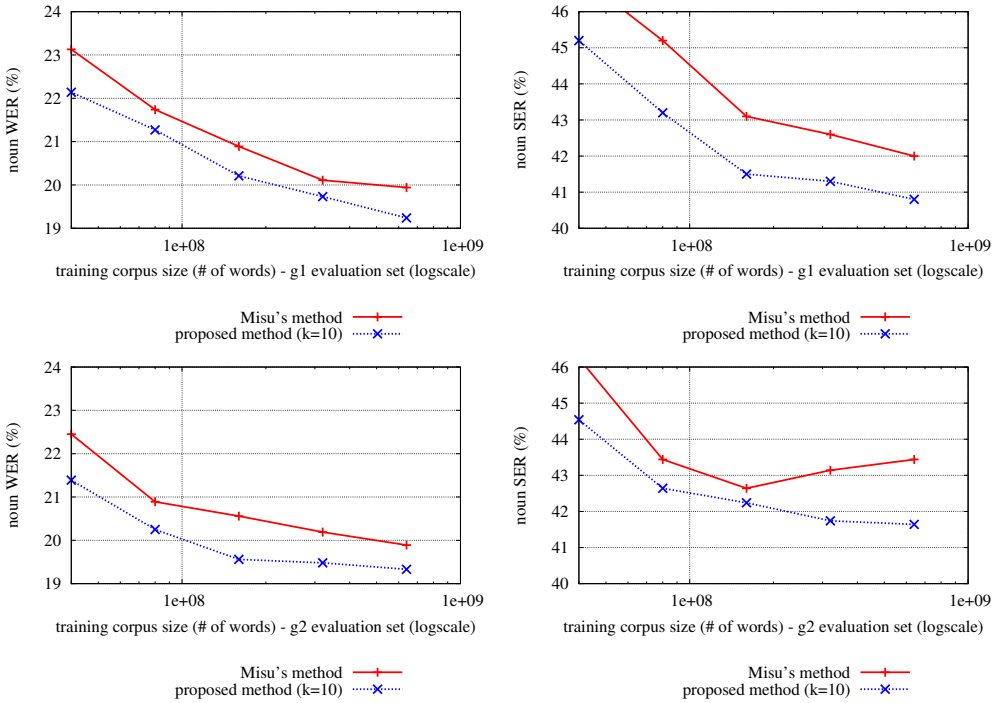

図 5: www コーパスで学習した名詞評価

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 5 は提案手法の最も低い単語誤り率モデル ($k = 10$) と翠らの手法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翠らの手法（$k = 0$）と g1 においては 0.70 ポイント，g2 においては 0.56 ポイントの差になった．文誤り率の場合は，g1 においては 1.20 ポイント，g2 においては 1.80 ポイントの差になった．この差も統計的に優位である（$p < 0.05$）．従って，名詞置き換えが効果があると言える．

また，提案手法に類似した手法として，名詞置き換えをシードコーパスではなく，学習コーパスに適用することも考えられる．この手法も実際に実験で試したが，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現れないため，性能の低下に直接的にはつながらない．別の言い方をすれば，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，そうした奇妙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盛んである．Berger ら [1] は入力文の内容語をクエリーとして利用し，抽出したテキストで学習コーパスを改良した．Zhu ら [15] 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 をクエリーとして利用した．Sarikaya ら [13] 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 [12, 5]．翠らの手法のみならず，上記の他の適合モデルに我々の名詞置き換えフレームワークに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 [16, 17, 4, 7]．翠らの手法の 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文脈類似度ランクが比較的に小さかった ($k = 10$) 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．文脈類似度計算手法 [9] による名詞置き換えと言語モデルのドメインアダプテーション手法 [12] に基づいて行う方法を提案した．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 3.25 ポイント（単語誤り率），及び 4.28 ポイント（文誤り率）改善した．

参考文献

[1] Adam Berger, Robert Miller. 1998. Just-in-time language modeling. In *Proceedings of ICASSP-98*, pages 705–708.

[2] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4), pages 467–479.

[3] I. Bulyko, M. Ostendorf, A. Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of Human Language Technology 2003 (HLT2003)*, pages 7–9.

[4] Stanley F. Chen, Stephen M. Chu. 2010. Enhanced Word Classing for Model M. In *Proceedings of Interspeech 2010*, pages 1037–1040.

[5] Mathias Creutz, Sami Virpioja, Anna Kovaleva. 2009. Web augmentation of language models for continuous speech recognition of SMS text messages. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 157–165.

[6] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata. 2009. Large Scale Relation Acquisition using Class Dependent Patterns. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'09)*, pages 764–769.

[7] Ahmad Emami, Stanley F. Chen, Abraham Ittycheriah, Hagen Soltau, Bing Zhao. 2010. Decoding with shrinkage-based language models. In *Proceedings of Interspeech 2010*. pages 1033–1036.

[8] Zellig Harris. 1954. Distributional Structure. In *Word* 10(23), pages 142–146.

[9] Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, Kentaro Torisawa. 2010. A Bayesian Method for Robust Estimation of Distributional Similarities. In *Proceedings of ACL 2010*, pages 247–256.

[10] S. Matsuda, T. Jitsuhiro, K. Markov, S. Nakamura. 2006. ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles *IEEE Transactions on Information and Systems* vol. E89-D(3), pages 989–997.

[11] I. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. In *Psychometrika* 12, pages 153–157.

[12] Teruhisa Misu and Tatsuya Kawahara. 2006. A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proceedings of Interspeech 2006*. pages 9–13.

[13] R. Sarikaya, A. Gravano, Y. Gao. 2005. Rapid Language Model Development Using External Resources for New Spoken Dialog Domains. In *Proceedings of ICASSP 2005*, vol I, pages 573–576.

[14] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, Sadao Kurohashi. 2008. TSUBAKI: An open search engine infrastructure for developing new information access. In *Proceedings of IJCNLP*, pages 189–196.

[15] Xiaojin Zhu, R. Rosenfeld. 2001. Improving trigram language modeling with the world wide web. In *Proceedings of ICASSP*, pages 533–536.

[16] Hirofumi Yamamoto, Yoshinori Sagisaka. 1999. Multi-class composite n-gram based on connection direction. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 533–536.

[17] Hirofumi Yamamoto, Shuntaro Isogai, Yoshinori Sagisaka. 2001. Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters. In *Proceedings of ACL-2001*, pages 6–11.