Vol.2013-SLP-97 No.4
2013/7/26

述語項構造を介したWebテキストからの文選択に基づく言語モデルの評価

吉野 幸一郎[1]　森 信介[1]　河原 達也[1]

概要：音声対話システムのための音声認識における言語モデル構築のために、Web 上で集積される文から適切なものを選択する手法を提案し、評価する。従来手法では文表層のパープレキシティを用いた文選択が一般的であったが、提案手法では音声対話において利用される文書集合（＝ドメイン）との意味的な類似度を定義し、これを文選択に用いる。具体的には、ドメイン固有の述語項構造パターンに着目し、統語的な尺度を定義する。この意味的な類似度と従来のパープレキシティに基づく手法を組み合わせることも検討する。2 種類の異なるドメインにおける音声認識実験によって、提案する文選択手法が有効であることが示された。この際、文選択を行うために複数の分類器を検討し、比較評価を行った。この結果、特に既存手法と、ナイーブベイズ法による提案手法を組み合わせた場合に有意な音声認識精度の向上が見られ、音声対話システムの意味レベルの理解精度も向上することが確認された。

1. はじめに

これまで数多くの音声対話システムが研究開発され、一部は実世界で利用されるようになっている。特に近年、スマートフォンなどで多様な要求に応答を行うことができるシステムも登場している。しかし、こうしたオープンドメインの対話システムにおいて、システムの応答は単純な一問一答にとどまっている。一方で、ユーザの複雑で曖昧な情報要求に対して、対象ドメインの知識を利用しながら、複数ターンにわたって対話を行うシステムも求められている。これは単純なキーワードベースの検索ではなく、観光地やレストランなどについてより詳細な情報の案内を行うものである。このようなアプリケーションは、対象とするドメインの知識を記述した文書を検索することによって実現することができる [1]。例えば、観光ガイドブックや Wikipedia 中の文書を利用して観光地のナビゲーションを行うシステム [2] が挙げられる。このような対象（ドメイン）は多様にあるので、音声対話システムに必要な要素を、対象ドメインの文書テキストから自動で構築できることが望まれる [3] 。

音声対話システムにおける音声認識モジュールは、ドメインと発話スタイルに適応した言語モデルを必要とする。既存の大語彙音声認識システムは、ドメイン特有の固有表現をカバーすることが難しいが、固有表現の認識誤りは、情報案内において致命的である。したがって、ドメイン毎に音声認識用の言語モデルを構築する必要があるが、そのための学習データが大量に用意できるという前提は現実的ではない。そこで、対象ドメインの文書を種として Web から関連した文章を収集する手法が検討されてきた [4], [5], [6], [7] 。これらは対象ドメインと話し言葉表現を間接的にカバーしようとするアプローチであるが、結果として多くの対象ドメイン以外の文を含んでしまうという問題点があった。これに対して、本論文では対象ドメインとの意味的な類似性に着目して、合致した文を選択する手法を提案する。

2. 提案手法の概要

提案手法の概要を**図 1** に示す。本研究では、対話システムが対象ドメインの文書集合 D を検索して情報案内を行うことを想定する。また、言語モデル学習のために Web から収集した文 q の集合を利用する。本論文では、Web から収集した文として Yahoo!知恵袋コーパス中の質問文を用いる。文書集合 D は書き言葉なので、言語モデルの学習データとして適当でなく、また Web から集めた文は対象ドメインに合致しないものが多い。従来手法では、ドメイン文書集合 D に対する、単語系列の表層的な類似度を定義し、Web から収集した文の選択を行う。この手法について 3 章で述べる。本研究では、述語項構造に基づく意味情報を利用することによって、深層的な類似度を定義する。

[1] 京都大学 情報学研究科
606–8501, 京都市左京区吉田本町
Kyoto University, School of Informatics
Yoshida-Honmachi, Sakyo-ku, Kyoto 606–8501, Japan

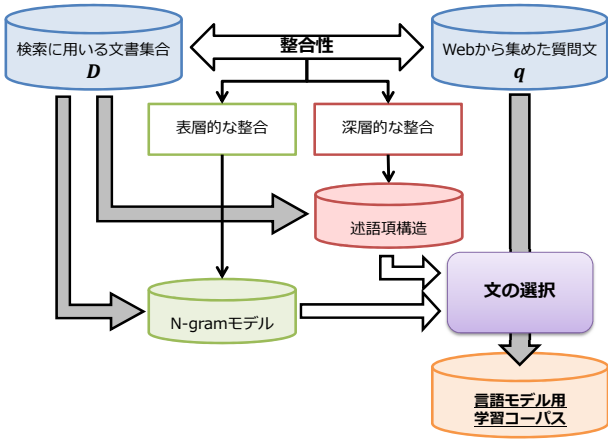

図 1　提案手法の概要

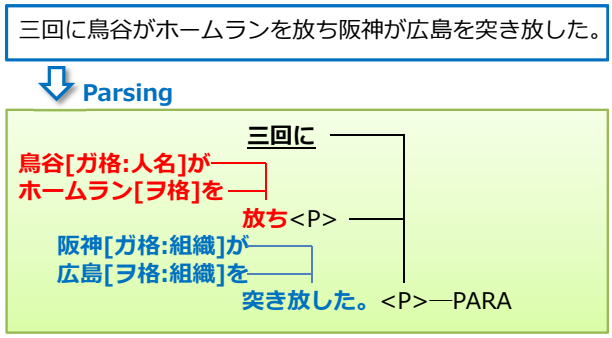

図 2　述語項構造の例

この新たな手法について 4 章で述べる。パープレキシティによる選択は生成モデルを仮定しているのに対して、提案手法では識別的なモデルによる文選択を行う。さらに、上記の 2 種類の文選択手法を併用する手法を検討する。これにより、2 つの手法の異なる特性が効果的に働くことが期待できる。これについて 5 章で述べる。また、これらの 2 種類の類似度を統合的に扱う確率モデルとして、ロジスティック回帰を用いた文の識別的選択を検討する。これについて 6 章で述べる。

3. N-gram モデルの表層的類似度

質問文 q が検索対象の文書集合 D にどの程度適合するかの表層的な指標として、N-gram モデルにおける KL 距離を導入する。KL 距離とは 2 つの確率分布の差違を測る尺度であり [8]、質問文 q と文書集合 D の KL 距離は以下のように定義される。w_i は質問文 q に含まれる単語である。

$$KL(q||D) = \sum_i P_q(w_i) \log_2 \frac{P_q(w_i)}{P_D(w_i)} \quad (1)$$

ここで P_D と P_q は D と q の言語モデルによって生成される確率であり、N-gram モデルによって与えられる確率で定義する。本研究では 3-gram によるモデルを用いる。q は質問文一文を想定しており、$P_q(w_i)$ は、q に含まれる N-gram が一意的である場合、$P_q(w_i) = 1$ となる。言語モデルの確率を 3-gram で与えると、短い一文の場合、多くの場合において一意となる。そこで、式 (1) を以下のように再定義することができる。

$$KL(q||D) \approx \sum_i \log_2 \frac{1}{P_D(w_i)} \quad (2)$$

$$= -\sum_i \log_2 P_D(w_i). \quad (3)$$

Web テキストを利用する先行研究において、検索対象の文書集合 D の言語モデルによる質問文 q のテストセットパープレキシティを利用した文選択が行われていた [6] が、テストセットパープレキシティ $PP(q, D)$ は以下のように定義される。

$$H(q, D) = -\frac{1}{n} \sum_{i=1}^{n} \log_2 P_D(w_i). \quad (4)$$

$$PP(q, D) = 2^{H(q,D)}. \quad (5)$$

これは文書集合と質問文の KL 距離を測ることと等しいと解釈できる。

4. 述語項構造に基づく意味的類似度

3 章で述べた文選択手法では、検索対象文書集合 D と質問文 q の表層的な一致度を利用していた。しかし、文の表層的な整合のみでは、検索対象文書中にある構造的・意味的な深層情報のレベルで合致した文を選択することは難しい。また、多くの対話システムでは意味的な情報を抽出することが求められている。そこで本論文では、述語項構造に着目した意味的類似度を定義し、深層的に合致する文を選択する手法を提案する。KL 距離あるいはパープレキシティに基づく手法では全ての単語の表層的な類似度を測っていたのに対して、本章では述語項構造における要素のみに着目した類似度の定義を行う。また、前章では N-gram モデルに基づく生成モデルを利用していたが、本章では識別モデルを導入する。

4.1 意味的な類似度の定義

意味的な類似度の定義を行うために、述語項構造に着目する。述語項構造は、意味解析によって得られる情報構造の 1 つであり、古くから自然言語処理において利用されてきた。述語項構造の抽出例を**図 2** に示す。述語項構造は最小単位である述語項として述語と、それに対する格要素、その意味表現を持ち、1 つの述語に対して 1 つから複数の格要素と意味表現が付与されている。図 2 の例では「放つ」という述語に対し、格要素「鳥谷」が意味表現「ガ格」で、格要素「ホームラン」が意味表現「ヲ格」でかかっている。この最小単位である「鳥谷（ガ格）-打つ」、「ホームラン（ヲ格）-打つ」を述語項と呼ぶ。以降では、意味表現を含めた述語を w_p、格要素を w_a と表記する。パーザ

JUMAN/KNP [*1] により、検索対象文書集合からこのような情報構造を自動で得ることができる。しかし、文書集合から得られる述語項構造全てが情報案内に有用なわけではなく、ドメインに依存した有用な情報構造のパターンがあることが知られている [9], [10]。例えば、野球ドメインにおいては「A（ガ格）, B（ニ格） -勝つ」「A（ガ格）, B（ヲ格） -打つ」といったパターンが重要であるが、経済ドメインでは「A（ガ格）, B（ヲ格）-売る」「A（ガ格）, B（ヲ格）-買収」などが重要なパターンとなる。こうしたドメイン依存の情報構造は、手動で定義することが一般的であったが [9] 、これを自動で抽出する手法が提案されている [11]。このドメイン依存の情報構造を自動で定義する手法では、 ナイーブベイズ法 を用いた手法が TF-IDF 法 よりも有効であることが報告されている。

そこで、ナイーブベイズ法によって定義されるドメインらしさを表す確率を用いて意味的な類似度を定義する。単語 w_i が与えられたとき、そのドメインが文書集合 D と一致する確率は、ベイズの定理を用いて以下のように定義できる。

$$P(D|w_i) = \frac{P(w_i|D) \times P(D)}{P(w_i)} \tag{6}$$

$$\simeq \frac{C(w_i, D) + P(D) \times \gamma}{C(w_i) + \gamma}. \tag{7}$$

ここで γ は中華料理店過程を用いて推定されたディリクレ過程に基づくスムージング係数である [12]。識別的なアプローチを取っているため、学習データとしてドメイン外データである $\bar{D}$ が必要となるが、これに関しては文書集合 D と同じ出典から無作為に抽出したドメイン外コーパスを用いる。また、これにより $P(D)$ を推定する。述語 (w_p)、格要素 (w_a) から構成される述語項構造 PA_i に対して、$P(D|PA_i)$ を以下のように定義する。

$$P(D|PA_i) = \sqrt{P(D|w_{p,i}) \times P(D|w_{a,i})}. \tag{8}$$

4.2 固有表現のクラス化

統計的手法においては、データスパースネスの問題が学習セットとテストセットの不整合により生じ、特に固有表現において大きな問題となる。そこで、固有表現をクラス化して式 (8) の確率を計算することでこの問題の解決を図る。固有表現は、述語項構造と同様に意味解析によって得られる情報構造の 1 つであり、図 2 に示されるように、人名や組織名などの固有名詞に自動でタグを付与したものである。今回は、固有表現を捨象した場合に同じパターンとなる述語項構造の確率を合計する。この例を図 **3** に示す。例では同じ「人名」クラスの格要素を持ち、意味表現「が」と述語「完投」が同一のパターンがまとめられている。固有表現で捨象した場合に同じパターンとなる述語項構造が

[*1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

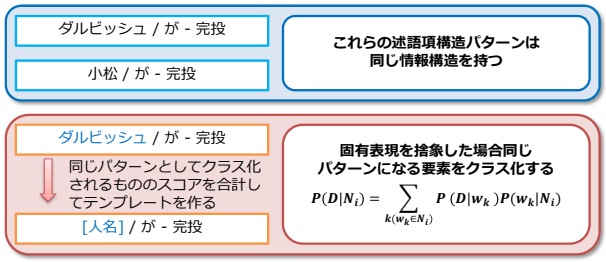

図 3　固有表現のクラス化

述語項構造

q = “イチローは今オフにＦＡ権を行使して他球団に移籍すると思いますか。”
PA = [“[人名]/ガ格/移籍:する”, “今:オフ/ニ格/行使:する”,
“ＦＡ:権/ヲ格/行使:する”, “他:球団/ニ格/移籍:する”]

述語項構造テンプレート

評価値	格要素	意味表現	述語
0.98201	フォーク	ヲ格	はじく:返す
0.98202	ホームラン	ヲ格	放つ
0.96353	ＦＡ:権	ヲ格	行使:する
0.95954	他:球団	ニ格	移籍:する
0.92919	今オフ	ニ格	行使:する
0.78062	[人名]	ガ格	放つ
0.68310	[人名]	ガ格	移籍:する
0.09994	株価	ガ格	下落:する
0.09994	負債	ガ格	拡大:する
	…		

図 4　意味的な類似度の例

持つ確率の合計を、捨象された固有表現を持つ述語項構造 N_i のスコアとする。図 3 の例では同じ「人名」クラスの格要素を持ち、意味表現と述語が同じパターンがクラス化されている。まとめられた固有表現のクラス N_i に対する確率は以下のように求める。

$$P(D|N_i) = \sum_{k(w_k \in N_i)} P(D|w_k) P(w_k|N_i). \tag{9}$$

4.3 意味的類似度による文の選択

選択対象文 q 中に存在する述語項構造 PA_i 全ての $P(D|PA_i)$ の平均を取り、$P(D|q)$ とする。この評価の例を図 **4** に示す。例では入力文 q は 4 つの述語項構造を持っており、各構造についての評価値の平均を計算することで、質問文の評価値を決定する。この評価値 $P(D|q)$ が高いものを選択して言語モデルの学習データとする。これにより、検索対象の文書集合に、意味的に関連があるユーザ発話を認識しやすい言語モデルの構築が期待できる。

5. 文選択手法の併用

3 章及び 4 章で異なる類似度を述べたが、これらを組み合わせて文選択を行うことも検討する。組み合わせの方法として、文の順位に基づく手法と、文のスコアに基づく手法を検討する。

5.1 文の順位に基づく手法

3 章と 4 章で各文にドメイン文書集合に対する類似度を付与する手法を示したが、この評価値によって選択対象文

表 2　評価セットの詳細

タスク	話者数	発話数
京都観光案内	4	219
ニュース案内	10	2,747

q を並び替え、順序（PP_{rank} と PA_{rank}）を付与する。この順位の合計（$PP_{rank} + PA_{rank}$）によって文を並び替え、文の選択に用いる。

5.2 文の評価値に基づく手法

3 章と 4 章で示した評価値を組み合わせて、新しい評価値を定義する。この際、2 つの評価値の値域を揃えるため、パープレキシティ PP を以下のシグモイド関数によって変換する。

$$PP(q,D)' = \frac{1}{1+e^{-PP(q,D)}}. \tag{10}$$

PP' と $P(D|q)$ の混合比は試行の結果、3:7 と定めた。

6. ロジスティック回帰を用いた文の識別的選択

4 章で文選択に識別モデルを導入して意味的な類似度の定義を行い、5 章で従来手法である表層的な類似度との併用について議論した。本章ではこれらを統合的に識別的アプローチで扱うため、ロジスティック回帰を用いた手法を提案する。4 章ではナイーブベイズ法を用いて述語項が与えられた場合の条件付き確率を定義したが、これをロジスティック回帰に拡張して、質問文 q が与えられた場合のドメイン確率

$$P_{LR}(D|q) = \frac{1}{Z_{q,\omega}} \exp(\omega \cdot \phi(q,D)) \tag{11}$$

を定義する。ここで $\phi(q,D)$ はクエリ q を素性ベクトルに拡張したものであり、ω はその重み、$Z_{q,\omega}$ は正規化項である。この素性ベクトルとして、3 章で用いた単語と、4 章で用いた述語項を用いる。

7. 評価実験

前章までに述べた手法によって文選択を行い、選択された文から言語モデルを構築した上で、音声認識による評価を行った。また、提案した手法による音声認識結果からの意味理解・対話応答についても評価した。

7.1 対象ドメインとシステムの構成

評価対象としてプロ野球ニュースに関するニュース案内システム [3] と京都観光案内システム [2] でのユーザ発話音声を利用した。プロ野球案内システムには毎日新聞記事データベース（CD-毎日新聞データ集 2000–2009）のうち、日本プロ野球のタグが付与された記事を用いた。また、京都観光システムには Wikipedia における京都関連文書文を用いた。これらを用いて $PP(q,D)$ 、$P(D|q)$ と $P_{LR}(D|q)$ の学習を学習した。音声認識用言語モデルの学習用テキストとしては Yahoo!知恵袋 [*2] から、質問タグが付与されているものを用いた。ニュース案内システムにはエンターテイメント-野球カテゴリのものを、京都観光案内は旅行-国内のカテゴリのものをそれぞれ用いる。音声認識用言語モデルの学習用テキスト内の各質問文を $PP(q,D)$ 、$P(D|q)$ と $P_{LR}(D|q)$ によって評価・並び替えし、それぞれのスコアによって選択された文から音声認識用 3-gram の学習した。選択する文数を変化させることにより、それぞれの文選択手法の評価を行った。**表 1** に学習セットの詳細を、**表 2** に評価セットの詳細を示す。

音声認識精度の評価尺度として単語誤り率 (WER) を用いる。また、参考のために補正パープレキシティ (PP) を示す。補正パープレキシティでは学習テキスト (q) 全体から語彙を構築し、この中から、選択された学習テキストに出現しない単語の数によって未知語 (<UNK>) の確率を割る。これにより、語彙数が異なる言語モデル同士のテストセットパープレキシティの比較が可能になる。

音声認識用デコーダとしては Julius[*3] [13] を用いる。また、音響モデルとして JNAS-IPA99-Testset に付属する binhmm,s2000mix16.gid、及び logicalTri.added を用いる。

7.2 音声認識精度による評価

ニュース案内システムの音声認識タスクにおける単語誤り率を**図 5** に、京都観光案内システムの音声認識タスクにおける単語誤り率を図 **6** に示す。いずれも、学習に利用した質問文の割合を横軸に示す。PP が KL 距離・パープレキシティに基づく類似度を用いて文選択を行った場合、PA が述語項構造に基づく類似度を用いて文選択を行った場合、$PP + PA$ が文の順位に基づく文選択の併用を行った場合、LR がロジスティック回帰に基づく類似度を用いて文選択を行った場合である。単語誤り率 (WER) においては、提案手法である述語項構造に基づく類似度を利用した場合 (text=7/10)、選択を行わない場合と比較して有意な性能差が認められた（有意水準 $p < 0.05$）。述語項構造に基づく類似度を用いる場合と、述語項構造に基づく類似度と表層的な一致に基づく指標を併用する手法では有意な差が見られず、2 種類の併用手法の間でも有意な差は見られなかった（よってここでは $PP_{rank} + PA_{rank}$ のみを示す）。さらに、いずれのドメインでも意味的な類似度を用いる提案手法が、表層的な整合のみを用いる既存手法よりも、おおよその場合において有効であることがわかる。この結果により、意味的な類似度を利用する提案手法で一定の音声認識精度向上があることが示された。また、いずれのドメインにおいても、学習テキスト量を 70% 程度まで

[*2] このコーパスは Yahoo!JAPAN と国立情報学研究所から提供を受けた。
[*3] http://julius.sourceforge.jp

表 1　学習セットの詳細

用途	タスク	コーパス名	文数
文選択器の学習	京都観光案内	Wikipedia	35,641
	ニュース案内	毎日新聞データベース	176,852
音声認識用言語モデルの学習	京都観光案内	Yahoo!知恵袋コーパス:旅行-国内	679,588
	ニュース案内	Yahoo!知恵袋コーパス:エンターテイメント-野球	403,602

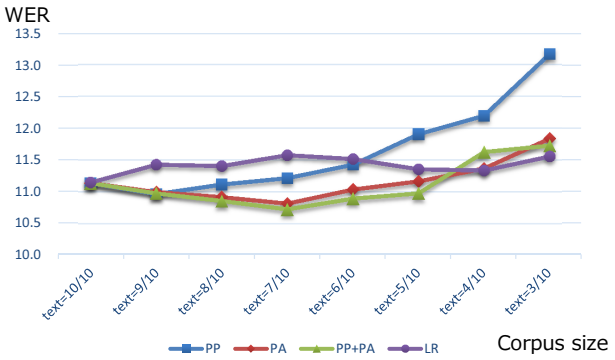

図 5　ニュース案内システムタスクにおける単語誤り率 (WER)

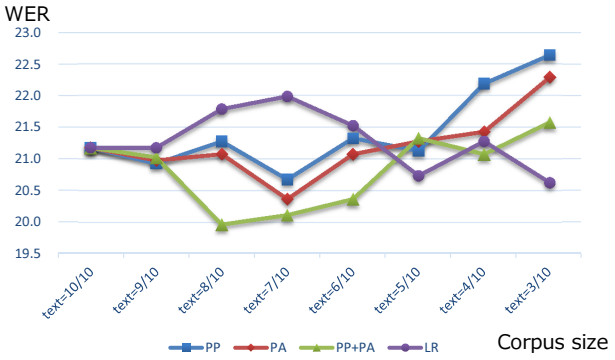

図 6　京都観光案内システムタスクにおける単語誤り率 (WER)

削減する場合に単語誤り率が最小となっており、経験的に学習用テキストの選択量が示されている。

一方、ロジスティック回帰を用いたモデルでは WER を学習用テキスト量に従って削減できていない。この原因としては、単なるロジスティック回帰では文書集合 D に対する過学習を起こしている可能性が挙げられる。これに対してナイーブベイズ法を用いた提案手法では、ディリクレ過程に基づく事前分布を導入することでこの問題を回避している。

参考のためにそれぞれの補正パープレキシティを**図 7**、**図 8** に示す。

7.3　音声対話における意味理解・対話タスクの精度

次に、音声対話における評価としてニュース案内システムタスクと京都観光案内システムタスクにおける意味理解精度、及びニュース案内システムタスクにおける対話タスクの達成精度を評価した。音声対話の評価としてニュース案内システムタスクの精度と、述語項誤り率 (PAER) を用いる。ニュース案内システムタスクとは、検索対象の

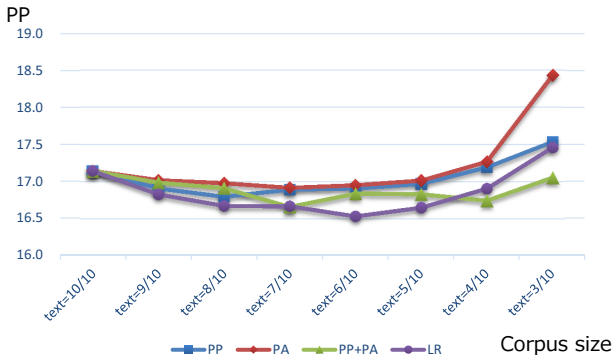

図 7　ニュース案内システムタスクにおける補正パープレキシティ (Adj. PP)

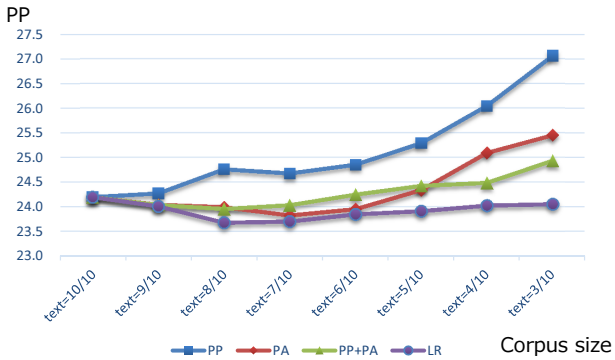

図 8　京都観光案内システムタスクにおける補正パープレキシティ (Adj. PP)

ニュース文書集合から、ユーザの質問意図に従った文を提示するタスクである。述語項誤り率とは、認識対象文中の述語項構造における「要素/格/述語」の三つ組の抽出精度を示す。これらの 3 つ組が全て正しく認識できていれば、正解とする。

ニュース案内システムタスクにおける述語項誤り率を**図 9**、京都観光案内システムタスクにおける述語項誤り率を**図 10** に示す。ニュース案内タスクにおける述語項認識精度（表 9）では、text=7/10 の際に述語項認識精度が最も高くなっており、選択を行わない場合 (text=10/10) の 21.5%から 20.4% まで誤り率が改善している。この結果、述語項誤り率においては既存手法に対する提案手法の有効性が顕著に見られ、深層的な情報の利用が述語項の誤り率改善に寄与しているということが言える。

また、選択を行わない場合と、述語項認識精度が最も高くなった点の 2 点においてニュース案内システムタスクにおける回答精度の評価を行った。その結果 0.8% の音声対話精度が向上し、提案手法による音声認識用言語モデルの

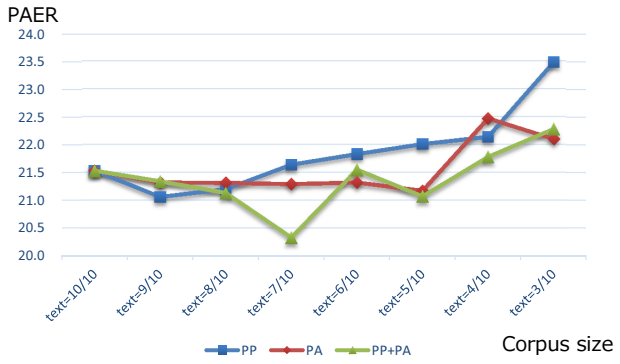

図 9　ニュース案内システムタスクにおける述語項誤り率 (PAER)

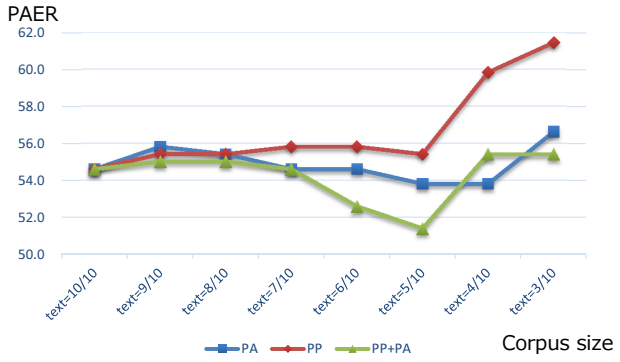

図 10　京都観光案内システムタスクにおける述語項誤り率 (PAER)

ための学習データ選択が、音声対話精度にも寄与することが確認された。

8. 関連研究

Web から獲得した言語資源を音声認識の言語モデル構築に活用することは、Web の普及に伴って研究されてきた。例えば、Web から獲得した 3-gram の頻度を用いる手法 [14] や、タスクごとに固有のキーワードを手動で設定して文を収集する手法 [15] が挙げられる。最も一般的な手法 [4], [7], [16], [17] では、ドメインで特徴的な N-gram を Web の検索クエリとして文を収集する。こうした手法では種となる N-gram を獲得するためのコーパスが必要である。このような検索クエリを生成するために、対話システムの検索対象となる文書集合 [6] や、講演におけるスライド [18], [19]、初期的な音声認識結果 [20] などを用いることが検討されている。

一方で、Web から取得した文集合から言語モデル学習に適切な文を選択する手法も研究されてきた。選択のために最も一般的に用いられるのは、種として用いたテキストから構築した言語モデルに対するパープレキシティである [6], [7] が、BLEU スコアを用いる手法 [4] や、トピックモデルの利用 [5] なども検討されてきた。Masumura ら [21] はナイーブベイズ法を用いた文選択が提案しているが、これらの先行研究における指標はいずれも文表層に対するもので、文の意味的な類似度まで考慮されてはいない。

9. まとめ

本研究では、音声対話システムのための言語モデル構築に利用する、文の選択手法について提案した。既存手法であるパープレキシティを用いた手法と比較して、述語項構造を用い意味的な類似度を利用することで、より音声認識システムが背後に持つ知識ベースに整合した文を選択し、言語モデルを構築することができる。この手法により、パープレキシティを用いた従来手法よりも有意に高精度な音声認識を実現できることが確認された。また、この手法を異なるドメインに適応することによって、手法の一般性を確認することができた。さらに音声対話実験を行うことによって、音声対話における応答精度が向上することも確認された。提案手法は背後に知識ベースや文書集合を持つような音声対話システムに対して適用することができ、ドメイン固有の知識を用いるような、複雑なタスクを行う音声対話システムの精度向上に寄与することが期待される。

参考文献

[1] Kawahara, T.: New perspectives on spoken language understanding: Does machine need to fully understand speech?, *Proc. IEEE-ASRU*, pp. 46–50 (2009).
[2] Misu, T. and Kawahara, T.: Bayes Risk-based Dialogue Management for Document Retrieval System with Speech Interface, *Speech Communication*, Vol. 52, No. 1, pp. 61–71 (2010).
[3] 吉野幸一郎，森　信介，河原達也：述語項の類似度に基づく情報抽出・推薦を行う音声対話システム，情報処理学会論文誌，Vol. 52, No. 12, pp. 3386–3397 (2011).
[4] Sarikaya, R., Gravano, A. and Gao, Y.: Rapid Language Model Development Using External Resources for New Spoken Dialog Domains, *Proc. ICASSP*, Vol. 1, pp. 573–576 (2005).
[5] Sethy, A., Georgiou, P. G. and Narayanan, S.: Building Topic Specific Language Models from Webdata Using Competitive Models, *Proc. Interspeech*, pp. 1293–1296 (2005).
[6] 翠　輝久，河原達也：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音声対話システム用言語モデルの構築，電子情報通信，Vol. J90-D, No. 11, pp. 3024–3032 (2007).
[7] Bulyko, I., Ostendorf, M., Siu, M., Ng, T., Stolcke, A. and Çetin, O.: Web resources for language modeling in conversational speech recognition, *ACM Trans. Speech Lang. Process.*, Vol. 5, No. 1, pp. 1:1–1:25 (2007).
[8] Kullback, S. and Leibler, R. A.: On information and sufficiency, *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86 (1951).
[9] Grishman, R.: Discovery Methods for Information Extraction, *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 243–247 (2003).
[10] L.Ramshaw and R.M.Weischedel: Information Extraction, *IEEE-ICASSP*, Vol. 5, pp. 969–972 (2005).
[11] Yoshino, K., Mori, S. and Kawahara, T.: Spoken Dialogue System based on Information Extraction using Similarity of Predicate Argument Structures, *Proc. of*

SIGDIAL, pp. 59–66 (2011).

[12] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M.: Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, Vol. 101, pp. 1566–1581 (2006).

[13] Lee, A. and Kawahara, T.: Julius–an open source real-time large vocabulary recognition engine, *Proc. EuroSpeech*, pp. 1691–1694 (2001).

[14] Zhu, X. and Rosenfeld, R.: Improving trigram language modeling with the world wide web, *Proc. of IEEE-ICASSP*, Vol. 1, pp. 533–536 (2001).

[15] Nisimura, R., Komatsu, K., Kuroda, Y., Nagatomo, K., Lee, A., Saruwatari, H. and Shikano, K.: Automatic n-gram language model creation from web resources, *Proc. EuroSpeech*, pp. 5181–5184 (2001).

[16] Wan, V. and Hain, T.: Strategies for language model web-data collection, *Proc. IEEE-ICASSP*, Vol. 1, pp. 1069–1072 (2006).

[17] Tsiartas, A., Georgiou, P. and Narayanan, S.: Language model adaptation using www documents obtained by utterance-based queries, *Proc. IEEE-ICASSP*, pp. 5406–5409 (2010).

[18] Munteanu, C., Penn, G. and Baecker, R.: Web-based language modelling for automatic lecture transcription, *Proc. INTERSPEECH*, pp. 2353–2356 (2007).

[19] Kawahara, T., Nemoto, Y. and Akita, Y.: Automatic lecture transcription by exploiting presentation slide information for language model adaptation, *Proc. IEEE-ICASSP*, pp. 4929–4932 (2008).

[20] Suzuki, M., Kajiura, Y., Ito, A. and Makino, S.: Unsupervised language model adaptation based on automatic text collection from WWW, *Proc. INTERSPEECH*, pp. 2202–2205 (2006).

[21] Masumura, R., Hahm, S. and Ito, A.: Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition, *Proc. INTERSPEECH*, pp. 1465–1468 (2011).

[22] Akbacak, M., Gao, Y., Gu, L. and J.Kuo, H.-K.: Rapid transition to new spoken dialogue domains: Language model training using knowledge from previous domain applications and web text resources, *Proc. INTERSPEECH*, pp. 1873–1876 (2005).

[23] Hakkani-Tur, D. and Gilbert, M.: Bootstrapping language models for spoken dialog systems from the world wide web, *Proc. IEEE-ICASSP*, Vol. 1, pp. 1065–1068 (2006).