2005－SLP－57（21）
2005／7／16

未知語を許容する対話システムにおける対話状態予測

安田宜仁　デネッケ・マティアス

日本電信電話 (株), NTT コミュニケーション科学基礎研究所
〒 619-0237 「けいはんな学研都市」精華町光台 2-4

yasuda@atom.brl.ntt.co.jp

あらまし

タスク遂行上の知識が欠落しているという意味での未知語を受け付けるような対話システムにおいて, 未知語を含めた対話状態を予測する方法を提案する. この対話状態の予測の使用例として, 対話システムの音声認識の言語モデルへの応用を提案する. 言語モデルとして, 未知語を取るスロットをクラスとするようなクラスベースの言語モデルを作成し, 新聞テキストコーパスと対話コーパスを用いて線形補間による適応を行った言語モデルを使用した場合の概念誤り率 47.0% に比べ, 提案法では 40.0%に改善することができた.

Predicting Dialogue State in a Dialogue System Capable of Handling Unknown Words

Norihito YASUDA and Matthias DENECKE

NTT Communication Science Laboratories, NTT Corp.
2-4 Hikaridai, Seika-cho, “Keihanna Science City” Kyoto, 619-0237, Japan

yasuda@atom.brl.ntt.co.jp

Abstract

In this paper, we describe a method for predicting whether the next user utterance will contain a word that is at the same time unknown to the system and relevant for achieving the users' intended goal. We extend this method by applying the prediction to the language models used in the speech recognizer. We built a class-based language model, in which one class represents unknown words. Experiments resulted in a concept error rate reduction to 40.0% from 47.0%, which was obtained by adaptation using linear interpolation of a newspaper corpus and collected dialgoue corpus.

1　はじめに

対話システムは定められたタスクにおける問題解決するため, タスクに応じた構造を持つデータベースをバックエンドに持つのが一般的である. たとえば, 列車の乗り換え案内システムであれば, 各駅の時刻表や列車の路線名に関するデータベースを持つと考えられる. このような構造化されたデータベースの変更はコストがかかるため, 語彙の追加は容易ではなく, 対話システムが受け付けることができる語彙を制限する要因になっている. しかし, ユーザにとってはこのようなバックエンドシステムの構造や, 語彙の制限を知ることは困難であるため, ユーザからのシステムへの入力がデータベース上で未知の言葉を含むことは避けられない.

我々は未知語を既知語と関連付けることにより, 問題解決のための知識のない未知語を含んだ入力に対しても問題解決を行えるような対話システムを提案している [8]. これは, 問題解決のための完全な情報をシステムが持たなくても, 未知語と関連づけられた既知語に関するデータベース上の各種知識を用いることによって問題解決を行う方法である.

例として, ユーザからの目的地と出発地に関する条件を要求として受けつけ, 列車の乗り換え経路を案内するような, 列車情報案内システムを用いて説明する. ユーザがシステムの知識の範囲内の語, たとえば「北新地」のような駅名を用いて案内を要求した場合には, 従来の対話システムは情報を案内することができる. 一方, 「浅草寺」のようなランドマーク名が指定れた場合, 浅草寺に関する複雑な知識を持たないため, 関連付けを試みる. 関連付けのための適切な候補を見つけるために質問応答技術を用いる. これにより, システムは浅草という候補を得ることができる. 質問応答によって常に正しい回答が得られるわけではないので, システムは関連付けが正しいかどうかをユーザに確認する. たとえば上記の例では「浅草寺の最寄駅は浅草ですか?」といった具合である. もし正しい関連付けを得ることができれば, システムはその情報に基づいて対話を進める. 一度このように未知語と既知語の関連付けを得ることができれば, 以後の対話においては, 速やかに対話をすることができる.

言語理解は正規表現によるパターンマッチによって行い, 未知語を含んだ入力から, 未知語を含んだ理解フレームが作成される. フレームに未知語を含むため, 受け付ける語彙が事前に定まっている通常の対話システムに比べ, 多様なシステム状態を取り得ることになる. しかし, そのような多様な状態すべてが均等に起こり得るわけではない. システムに入力される語は, システムにとっては未知であっても, ユーザにとってはそのタスクにおいて自然であるからこそ入力されるはずであり, 同じ役割で用いられる語の間には何らかの関連があると考えられる. このような関連は, 一般の文書中でも似た文脈を持つという仮定に立ち, 大規模テキストコーパスにおける出現方法の類似性を用いて, 入力される可能性が高い語を予測する方法を提案する. このような対話状態の予測ができれば, 対話システムの語彙を拡張する際の目安として用いたり, 予めよくある対話状態でのデータベース検索を事前にキャッシュしておくことにより, 対話システムの高速化などに活かすことができる.

また, このような対話状態の予測は, 未知語を受理するような音声対話システムを実現するためにも有用である. 近年, 超大語彙での連続音声認識を実時間で行うような技術が開発されており, 大規模テキストコーパス中の語を網羅的に認識語彙とすることも可能になっている [4]. 音声対話システムで用いる場合には, 適切な言語モデルが必要であり, システムとユーザとの対話を収集し, その書き起こしを用いて言語モデルを作成することが多い. しかし, 我々が提案しているような未知語を受け付けるシステムのためには, このような閉じたコーパスでは不十分である. このため, 大規模な語彙を持つ新聞等のテキストコーパスと収集した対話コーパスを組合せることが考えられるが, 統計に基づいた言語モデル適応では, どのようなをスロットがありそれがどのような役割かといった, 対話システムが持つタスクの構造を用いることが困難である. 一方, 本稿で提案する対話状態予測は, 各スロットで想定される語を予測するため, スロットをクラスとしたクラス言語モデルを自然な形で作ることができる. このため, 対話システムが持つタスクの構造を有効に用いることができ, より精度の高い認識が可能になると期待される.

2 関連研究

対話システムでの対話状態の予測という観点では, 文献 [3] において, 音声対話システムに対するユーザの発声の韻律情報を用いることにより, 認識器の出力の補完的な役割として, 認識誤りを予測する方法が提案されている. ドメインに合致した小規模なコーパスと大規模汎用コーパスを用いた言語モデルの適応という観点では, さまざまな統計的な言語モデルの適応手法が研究されている [7, 1, 2]. 対話システムでの言語モデルの場合, 特定の小さなドメイン上の特徴を統計的に活かすだけでなく, 対話システムが持つタスクの構造を積極的に用いることが有効であると考える.

本稿では対話システムが持つ問題解決能力を保ちつつ未知語を含めた入力を受け付けるという意味で対話システムの拡張を図っているが, 一方で我々は, 質問応答技術を制限ドメインで行い, 対話型な処理能力を持たせる研究を行っている [6]. 対話システムが限られた

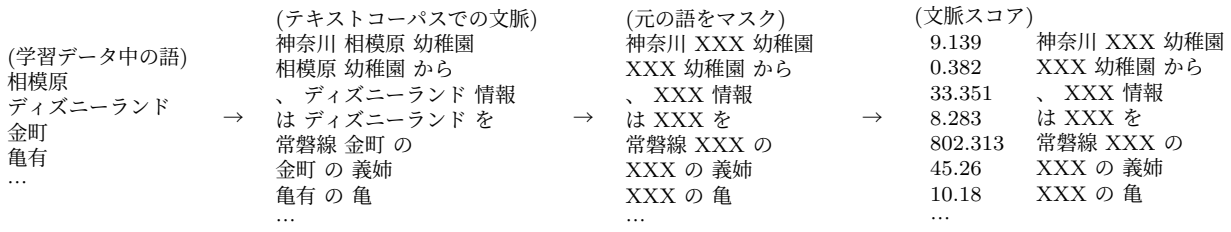

図 1: 対話状態予測のための文脈スコア付けの流れ

語彙ではあるが問題解決能力を持つのに対し, 質問応答は対話システムのような問題解決能力は持たないが語彙の制限がないという意味で, 2 つの極からのアプローチと言うことができる.

3 対話状態予測

対話状態の予測とは, 対話システムがもつ各スロットにどのような語が入りやすいかを予測することである. 訓練コーパス中であるスロットに入ると判断された語を元に, このような語と一般のテキストコーパスにおいて類似した文脈で出現するような語は, 対話システムでの当該スロットに入りやすいと考える. ここでの文脈とは, 具体的には n-グラムや bag of words のように, 機械によって自動的に重複なく抽出したものを考える.

このために, まずどのような文脈が, より良くこのような語の特徴をよく表しているかという文脈のスコアを定義し, それに基いて各語のスコアを定義する.

まず文脈のスコアについて述べる. 訓練コーパスでの言語理解結果中でスロット s に入ると判断された語を一部に含むようなテキストコーパス中の文脈をすべて抽出する. これらの各文脈について, スロットに入っていた語を, 他のすべての単語と同一視した (マスクした) 文脈を作成する. このマスクした各文脈の, スロット $slot$ でのスコア $cscore(c, slot)$ を以下のように定義する.

$$cscore(c, slog) = \sum_{t \in T_c} \frac{F_{(t,s)}}{F_{(t,ALL)}} \times \frac{N_{(t\&c)}}{N_c}$$

ただし, T_c は文脈 c の一部として含まれるすべての語であり, $F_{(t,slog)}$, $F_{(t,ALL)}$ はそれぞれ語 t のスロット $slot$ での出現回数, 語 t のテキストコーパス全体での出現回数を表し, また, $N_{(t\&c)}$, N_c はそれぞれ文脈 c のマスクの値として t を与えた文脈のテキストコーパス全体での出現回数, マスクされた部分を任意の語とした場合の文脈 c のテキストコーパス全体での出現回数を表す. 文脈のスコアの考え方は, 一般の文書よりもより多く対話コーパスで出現するような語を含むような文脈が, 他の語よりもより多くその語を含んでいる場合が高いスコアになるべきだという考え方である. 例として, 評価に用いた列車の乗り換え案内システムでの対話において, 「目的地」スロットに入った語から文脈スコアを出すまでの過程を図 1 に示す.

次にこの文脈スコアを用いて, テキストコーパス中の任意の語についてのスコアを定義する. 語のスコアは, スコアの高い文脈の中でよりたくさん出てくる語は, 高いスコアになるべきだという考えに基づいている. スロット $slot$ における語 t のスコア $score(t, slot)$ を以下のように定義する.

$$score(t, slot) = \sum_{c \in C} cscore(c, slot) * \frac{N_{(t+c)}}{I_t}$$

ただし, C は t を含むようなすべての文脈を表し, $N_{(t+c)}$ は文脈 c のマスクの値として t を与えた文脈の出現回数を, I_t は t が出現した文脈の異なり数を表す.

以上のように定義したスロットにおける語のスコアによって, 未知語を受け付け得るような各スロットにどのような語が入力されやすいかを予測する. 予測を行うためには, 訓練データから得られる理解フレームが必要となる. この理解フレームは, 人手によって正解を与える場合と, 訓練コーパスの書き起こしからシステムの言語理解によって自動的に生成する場合が考えられる. 次節では, これら両方の場合での評価を行っている.

ユーザ: 成城学園前からよみうりランドへはどうやって行けばいいですか
システム: 出発地は 成城学園前 ですね?
ユーザ: はい
システム: どちらまで行かれますか
ユーザ: よみうりランド
システム: 目的地は よみうりランド ですね?
ユーザ: そうです
システム: えーっと, 私はよみうりランドについて詳しくないので, ちょっとお待ちください.
システム: もしかしてよみうりランドの最寄り駅は 京王よみうりランド ですか?
ユーザ: はい
システム: えー、お調べする間、少々お待ちください
システム; 小田急小田急線で登戸へ行って, 南武線で稲田堤に行って, 徒歩で京王稲田堤に行って, 京王相模原線で京王よみうりランドで降りれば目的地です.
システム; ご案内は以上です. ご利用ありがとうございました.

図 2: システムとの対話例

入力: 花小金井駅から早稲田駅まではどうやって行ったらいいですか
理解フレーム: ((from . 花小金井駅) (to . 早稲田駅) (intention . how))
クラス化した入力文: [FROM] から [TO] まではどうやって行ったらいいですか

入力: グラバー邸まで行きたいのですけれども
理解フレーム: ((to . グラバー邸))
クラス化した入力文: [TO] まで行きたいのですけれども

入力文: えっとー、東京大学から慶応大学に行きたいんですけど
理解フレーム: ((from . 東京大学) (to . 慶応大学))
クラス化した入力文: えっとー、[FROM] から [TO] に行きたいんですけど

図 3: 言語理解結果例

4 評価

4.1 コーパス

提案法を使った言語モデルを作成した. テストデータに対してこの言語モデルを用いた音声認識および言語理解を行い, 結果得られた理解フレームの概念誤り率によって評価を行った.

評価のための対話コーパスとして, 乗り換え案内を行う対話システムと被験者の対話を作成した. 図 2 にこのシステムとの対話例を示す. 対話は Wizard of Oz (WOZ) 形式にて収集した. ただし, 通常の WOZ 形式の実験とは異なり, ウィザード役が行うのは, ユーザの音声をキーボードでタイプするだけである. それ以外の処理, たとえば, 入力に対する処理や, 応答生成, TTS を使ったシステム発話生成などはシステムが行った. システムの言語理解は, 正規表現によるパターンマッチによって理解フレーム表現へ変換する. ルール数は 22 である.

30 名の被験者によって, 1007 対話を収集し, 各対話の最初のユーザ発話 1007 発話について, 正しい理解フレームを人手によって作成した. この 1007 発話のうち, 603 発話を訓練に, 202 発話を開発に, 202 発話をテストに用いた. 書き起こしはウィザードが対話実験中に聞き取ってタイプしたものをそのまま使用した. そのため, 音声の聞き取り誤りや, タイプミスを含んでいる.

実験システムは, 「出発地」, 「目的地」, 「出発時刻」, 「到着時刻」, 「意図」, 「手掛りを持たない語」の 5 つのスロットを持つ. このうち, 「出発時刻」, 「到着時刻」についてはユーザが未知語を発声することはなかった. また, 「意図」についてはシステムの制限から記号化された値しか取ることはできない. また, 「手掛りを持たない語」についてはシステムが一時的に値を格納するために用いている. 以上から, 未知語を含んだスロットは出発地, 目的地であった. このため, 未知語の解決の対象としてはこれらの出発地と目的地のみを取り扱うことにした.

音声認識には, NTT で開発された音声認識システ

	書き起こし	線形補間	提案法 システムの言語理解で学習	提案法 人手による正解で学習
概念誤り率	13.1%	47.0%	42.9%	40.0%
(部分一致を正解とした場合)	(12.0%)	(43.6%)	(39.6%)	(36.7%)

表 1: 各言語モデルでの概念誤り率

ム SOLON[4] を用いた. 音響モデルは, IPA 読み上げ音声データ 104 時間分から作られた性別非依存モデルである. 状態共有型トライフォン HMM の 3061 状態で, 1 状態あたりの混合数は 16, 特徴量は 26 次元 (12 MFCC + 12 Δ MFCC + エネルギー + Δエネルギー) である.

言語モデルの適応および提案した対話状態予測に用いるテキストコーパスとして, 毎日新聞 1991 年から 2001 年までの 11 年分を用いた. 単語への分割には, SVM に基づく固有表現抽出 [5] を用いて, 1 つの固有表現が 1 つの単語になるようにした.

4.2 言語モデル作成

以下の手順で提案法を適用したクラス 3-グラムモデルを作成した. 未知語を考慮するスロット (出発地, 目的地) に入る語は, スロット名に置き換え, それ以外の語は語自体を 1 つのクラスとした. この際の, 理解フレームの作成は, 訓練データ中の書き起こしをシステム自身のパターンマッチによる言語理解よるものと, 人手によって作成したものの 2 種類を準備した. 図 3 に, 理解フレームおよびクラス名に置き換えた文の例を示す. クラス内での生起確率は, 未知語の生起確率と, 訓練データ中でスロットの値となった既知語の生起確率の重み付き和として定めた. 既知語の生起確率は, 訓練コーパスでの生起確率を用い, 未知語の生起確率は, 3 節で述べた手法を用いて算出したスコアを比例配分したものを用た. スコア算出の際の文脈には 3-グラムを用いた. 未知語の生起確率と既知語の生起確率の重みは, 開発データ中で最も概念誤り率が小さくなる重みを用いた.

比較手法として新聞テキストコーパスと訓練コーパスの線形補完による適応を行った言語モデルを準備した. 提案手法同様, 重みは, 開発データ中で最も概念誤り率が小さくなる重みを用いた.

4.3 評価指標

評価の指標として概念誤り率を用いる. 音声認識のための言語モデルを作成した場合には, 単語誤り率を用いるのが一般的であるが, 単語誤り率の改善は, 対話システムの動作の改善には直接結びつかない. なぜなら, 理解フレームの形成に関与しない語を正確に認識できたとしても, 理解フレームの形成に関与する語を認識できなければ意味がないからである. このため本稿では, より対話システムの動作を反映した指標として概念誤り率を用いる.

誤りの判定は, 人手によって作成された正解と比較し, 完全一致するかどうかによって行った. このため, たとえば, システムの出力が「JR 横浜駅」であり, 正解が「横浜駅」である場合や, 「阿佐ヶ谷」に対し「阿佐ケ谷」である場合などいずれも誤りと判定されるので, 厳しい判定基準であると言える.

4.4 評価結果

表 1 に各手法での概念誤り率を示す. 比較のため, 音声認識を通さずに書き起こし, つまりウィザードが実験中に聞き取ってタイプした結果に対して言語理解のみを行った結果も示す. なお, 書き起しの結果では訓練の必要はないので, 1007 発話は全ての結果を示している. 表の通り, 新聞テキストコーパスと対話コーパスとの線形補間による適応を行った場合の概念誤り率が 47.0% であるのに対し, 提案法では, システムの言語理解を用いた場合の概念誤り率が 42.9%, 人手によって作成された理解フレームを用いた場合の概念誤り率が 40.0% と, 4.1%あるいは 7.0%の改善をすることができた.

前節で述べた通り, 正解の判定は文字列の完全一致で行っているため, 「横浜駅」に対し「JR 横浜駅」という出力でも誤りと判定される. しかしこのような例は, 実用上は正解とみなしても問題ないと思われる. こ

のような例の参考のために, 回答と正解のどちらかがどちらかの部分文字列になっていれば正解として概念誤り率を出した場合の結果も表 1 に示す. ただし, この値は「阿佐ヶ谷」と「阿佐ケ谷」のようなものは相変わらず誤りとして判定され, 「中京競馬場」と「けどねえ中京競馬場」のようなものは正解として判定された結果の値である.

5 まとめと今後の課題

本稿では, 未知語を許容するような対話システムにおける対話状態の予測を提案した. さらに提案した対話状態の予測を用いた言語モデルを用いた音声認識と言語理解の実験を行った. 結果, 新聞テキストコーパスと対話コーパスとの線形補間による適応をした言語モデル用いた場合の概念誤り率が 47.0%であったのに対し, 提案法では 40.0%に改善することができた.

提案法で, 人手による正解理解フレームを用いた場合には, 書き起し以外に正解理解フレームを作成する必要があり, 余分な手間がかかることになる. しかし, この追加の手間は, 書き起しの作成と比べた場合非常に小さい. なぜなら, 書き起こしでは, フィラーや言い直しも記述する必要があるが, 理解フレームの作成の場合は, 事前に定められた少数のスロットの値を埋めるだけで済むからである.

提案した対話状態予測では, スロットの値として単一の語のみを想定している. しかし, 収集した対話コーパスにおいては, 目的地として「上野の博物館」といった値が入る場合があった. 複合語や複数の文節も含めた予測は今後の課題である.

謝辞 音声認識システムの使用にあたりご協力いただいた, 堀貴明氏, 渡部晋治氏, NTT コミュニケーション科学基礎研究所音声オープンラボの皆様に感謝します.

参考文献

[1] Bacchiani, M. and Roark, B. Unsupervised language model adaptation, Proc. ICASSP (2003).

[2] Florian, R. and Yarowsky, D. Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation, Proc. 37th ACL (1999).

[3] hirschberg, J., Litman, D. and Swerts, M. Prosodic and other cues to speech recognition failures, *Speech Comm.*, **43** (2004), 155–175.

[4] Hori, T., Hori, C. and Minami, Y. Fast On-The-Fly Composition for Weighted Finite-State Transducers in 1.8 Million-Word Vocabulary Continuous Speech Recognition, Proc. ICSLP (2004).

[5] Isozaki, H. and Kazawa, H. Efficient Support Vector Classifiers for Named Entity Recognition, Proc. 19th COLING (2002).

[6] Matthias, D. and Yasuda, N. Does this answer your question? Towards Dialogue Management for Restricted Domain Question Answering Systems, Proc. 6th SIGdial Workshop on Discourse and Dialogue (2005), (to appear).

[7] Seymore, K., Chen, S. and Rosenfeld, R. Nonlinear interpolation of topic models for language model adaptation, Proc. ICSLP (1998).

[8] Yasuda, N., Dohsaka, K. and Denecke, M. Handling Unknown Words in Dialogue Systems Using Unannotated Large Text Corpora, Proc. Konvens Workshop on Advanced Topics in Modeling Natural Language Dialog (2004).