人工知能学会研究会資料
SIG-SLUD-9201-1(4/15)

音声対話システムにおける未知語の扱い
Processing Unknown Words in Speech Dialog System

伊藤克亘† 速水 悟‡ 田中穂積†
ITOU Katunobu HAYAMIZU Satoru TANAKA Hozumi
†東京工業大学工学部
Tokyo Institute of Technology
‡電子技術総合研究所
Electrotechnical Laboratory

Abstract: In a practical speech dialog system, it is almost impossible for a speaker to remember which words are in the vocabulary. Therefore, it is very inconvenient that the system forces a speaker to utter only with words in the vocabulary. And also, it is quite difficult to design a vocabulary and grammar for the system. So, it is very important for a speech dialog system to deal with unknown words. In this paper, we propose a new method to add unknown words to a vocabulary of a speech dialog system using interaction between a speaker and the system. In the proposed method, parts of unknown words are processed by a phonetic typewriter and other parts are processed with the dictionary and the grammar. The phonetic typewriter estimates a transcription of an unknown word with only a phone N-gram model which is a stochastic model of phone sequences. After detecting unknown words, the system add them to the vocabulary to interact with a speaker to obtain necessary information. Preliminary results indicate the estimated transcription by the phonetic typewriter is able to use as the lexicon.

1 はじめに

言葉は、社会や時代とともに変化する。逆に、社会や時代の変化を表現するために、言葉は変化せざるをえないともいえるだろう。したがって、未知の単語・言葉遣いを認識・理解・生成する能力は、自然言語をもちいるうえで、非常に重要である。また、現在の音声認識システムを含む自然言語処理システムでは、さまざまな制約から、さまざまなレベルで、あつかえる範囲に限界がある。したがって、未知語の処理は、ある意味で、人間が直面するよりも重要な問題となる。

たとえば、東京方面への交通案内をする音声対話システムに対して、そのシステムが知らない「所沢」までの案内を尋ねようとした場合を考える。このとき、未知語処理を全くおこなわないシステムでは、次のようにシステムが持つ知識(例えば語彙)だけを用いて、強引に認識してしまう。

システム：どこに行きますか。
利用者　：所沢まで行きたいんですけど。
(システムは、「東京までおねがいします。」と認識)
システム：東京ですか。東京へ行くには…
利用者　：ちがいます。
システム：すみませんでした。もう一度おねがいします。
利用者　：と、こ、ろ、ざ、わ、に行きたいんですけど。
システム：東京ですね…
利用者　：ちがうんだってば。

(以下略)

音声認識では認識誤りがつきものなので、こういった場合、利用者は、システムの返事が認識誤りにもとづく誤った対応なのか、未知語を含んでいるためにおこる誤った対応なのかがわからない。現実的には、目的地のような語であれば、あらかじめ、システムが答えられる範囲を明確にしておくといった対処も可能だろう。しかし、システムが値段と同等な語として「電車賃」という単語が使われることを知らない場合には、個人による語彙の違いがあるので、全ての利用者に「電車賃はいくらですか」という発話をしないようにさせるのは非常に困難であるし、利用者にも、かなりの負担を課すことになる。

また、実際のシステムを設計する上で、どのように語彙を制限するかというのも非常に難しい問題である。そういった語彙の自然な設計および拡張法のひとつとして、実際にシステムを使っていくうえで、人間が発話した未知語を語彙に足していく方法が考えられる。そのような拡張が、完全に自動的にできれば、何も問題はないが、人間が認識する場合でも、未知語がどのような音韻列かさえも推定できない場合もあり、未知語があらわす概念にいたっては、人間でも、正確に推定することはかなり困難である。

しかし、未知語を語彙に追加するためには、未知語の検出・未知語の音韻系列の推定・未知語の構文カテゴリの推定など様々な過程が必要となる。これらを全て自動化するのは無理ならば、自動化できない部分だけは利用者とのやりとりで補うという手法が考えられる。利用者とのやりとりをなるべく減らし、その手間だけで語彙を拡張できるなら、システムとしての適応能力もかなり十分だといえるだろう。また、特に、対話システムでは、一連の対話の中でおこなえるので、かなり自然な方法だといえるのではないだろうか。

われわれは、これまでに、音声対話システムへの応用を考慮した連続音声認識システムでの未知語処理の手法を提案している [1]。本論文では、そのシステムを用いて、音声対話システム [2] で未知語を扱う手法について提案する。

2 音声対話システムの概要

まず、本稿で提案する手法の実験に用いるシステムについて述べる。このシステムは、実際のシステムを相手にした会話音声の収集を目的に構築された。したがって、以下のふたつの点を重視している。

- 実時間で応答できること
- 相手の発話をある程度正しく認識でき、対話をすすめられること
- 利用者に自然な発話をさせること

文法や語彙の規模は、この条件を満たすように設計した。また、機械からの応答はテキストでおこなわれる。

シナリオは、つくばから東京までの交通案内である。案内できる行き先は 7 か所で、一度の対話では、1 か所に関する案内をおこなう。システムが答える項目は、路線名・乗り換え駅・所要時間・費用だけである。対話例を次に示す。

システム：どこに行きますか。
利用者　：新宿に行きたいんですけど。
システム：新宿ですか。新宿へ行くには、常磐線と山手線に乗ります。
　　この経路について、のりかえ・所要時間・費用をお尋ね下さい。
利用者　：どこでのりかえるんですか。
システム：上野です。
　　他に何かございますか。
利用者　：どのくらいかかりますか。
システム：値段ですか、それとも時間をお尋ねですか。
利用者　：時間です。

システム：1 時間 20 分です。
　　他に何かございますか。
利用者　：上野から新宿までは何分かかりますか。
システム：上野から新宿までは 20 分です。
(以下略)

文法は、28 の文テンプレートで構成される。文テンプレートの例をしめす。

<目的地名> <まで> <行く> <行く (語尾)>

<目的地名> = 東京 | 新宿 | …
<まで> = まで | に | へ
<行く> = 行き
<行く (語尾)> = ます | たいです | …

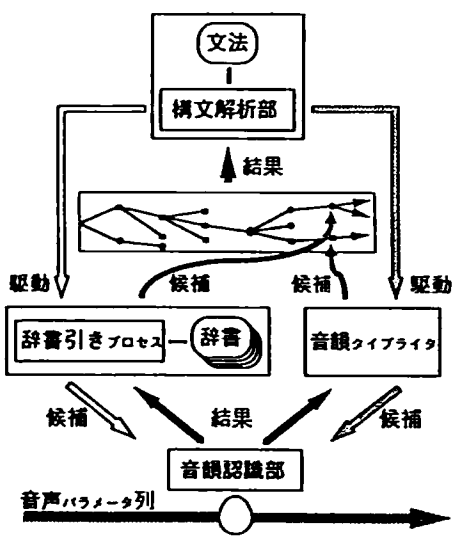
かなり制限の厳しい文法だが、様々な種類の表現に対応するものを用意している。

以上述べたように、本システムでは、対話できる内容はかなり限られている。しかし、内容が同じでも、様々な表現による発話を受け付けられるように設計されている。また、実際にワークステーション上で、利用者が発話してから数秒で回答するので、対話の進行はかなり自然におこなえる。

3　未知語の検出と音韻列の推定

3.1　概要

未知語処理のために、以下のように音声認識システムを拡張する。



この手法の特徴は、構文解析部が音韻タイプライタと辞書引きプロセスを並行して駆動しているところである。構文解析部は、未知語となりうるカテゴリを予測した場合には、辞書引きプロセスを駆動するのと同時に音韻タイプライタも駆動する。音韻タイプライタが認識した未知語は、駆動時にその構文カテゴリを決定しているので、構文解析部に送られたあとは、辞書引きプロセスが認識した語と同じように扱われる。したがって、未知語を一部だけ含んで、あとはシステムが知っている語からなる認識結果をえることもできる。以下、詳細を述べる。

3.2 構文解析部

構文解析部には、未知語の検出をおこないながら形態素・構文解析をすすめることができる堀内らの構文(形態素)解析手法 [3] を用いる。この手法は、構文など上位の情報を利用しながら、バックトラックせずに未知語の検出をおこなえるので、フレーム同期で処理をすすめる niNja のような音声認識システムとの親和性が高い。

3.3 音韻タイプライタ

音韻タイプライタは、辞書や文法を利用せずに、音声を認識するプロセスである [4]。基本的には、まず全ての音韻モデルを入力と照合し、そのモデルの照合が終了したら、つぎに全ての音韻モデルを連結して入力と照合する、といった手順で認識をすすめる。しかし、常に、すべての音韻の組み合わせを生成していたのでは、日本語として体をなさない音韻系列がたくさん生成されてしまう。そこで、できるだけ無駄な処理をさけるためにふたつの知識をつかう。まず、生成するときには、日本語として体をなさない候補を作らないように、子音は連続しないなどの日本語の音韻の連鎖に関する知識を用いる。しかし、その知識だけでは、日本語らしくない候補もたくさん生成されてしまうので、さらに、音韻系列の日本語らしさを評価する知識として、音韻連鎖の N グラムモデル [5] を用いる。

音韻連鎖の N グラムモデルは、統計的な言語モデルの一種である。統計的な言語モデルでは、発話 $\mathrm{P}(=p_1, p_2 \cdots, p_n)$ (ここで p_x は音韻とする) が発話される確率 $P(\mathrm{P})$ は、次の式であらわされる。

$$P(\mathrm{P}) = \prod_{i=1}^{n} P(p_i | p_1, p_2, \cdots, p_{i-1})$$

ここで、$P(p_i|p_1, p_2, \cdots, p_{i-1})$ は音韻系列 $p_1, p_2, \cdots, p_{i-1}$ が発話され、次の音韻として p_i が発話される確率である。理論上は n が無限大になりうるため、$P(p_i|p_1, p_2, \cdots, p_{i-1})$ の値をあらかじめ用意するわけにはいかない。そこで、次の式のように近似して計算するモデルが N グラムモデルである。

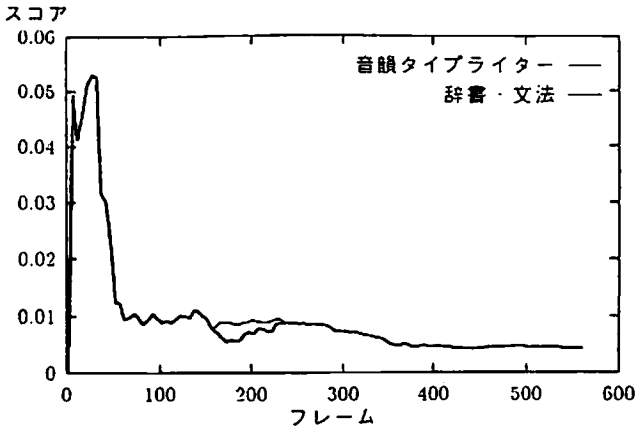
$$P(\mathrm{P}) = \prod_{i=1}^{n} P(p_i | p_{i-N+1}, p_{i-N+2}, \cdots, p_{i-1})$$

このように連続する N 個の要素だけを考慮するように近似するため、非常に局所的な制約となってしまうが、音韻・音節レベルで音声認識の精度を向上させるのに役立つことが知られている [6, 4]。

3.4 制御

このふたつの知識を用いても、音韻タイプライタの処理は、現在のかなり制限された語彙を対象とする辞書引きプロセスと比べるとかなり大きな計算量を必要とする。音声認識システムに対する発話に未知語が含まれる割合はそれほど大きくないと考えられるので、その未知語を処理するためのプロセスの計算量が常に大きすぎたり、候補をたくさん生成したりすることは、効率の面でも精度の面でも現実的ではない。

そこで、本システムでは、未知語の部分で重点的に音韻タイプライタが動作するように制御する [1]。辞書引きプロセスが生成する候補のスコアと、音韻タイプライタが生成する候補のスコアの推移を次に示す。



この図は、100 フレームのあたりに未知語が存在している発話を認識した例だが、未知語の部分では、辞書引きプロセスのスコアだけが小さくなっている。したがって、辞書引きプロセスのスコアをしきい値として、音韻タイプライタが生成する候補は、それ以上のスコアを持つものだけを残すようにすれば、未知語以外の部分では、ほとんど音韻タイプライタは候補を生成しなくなる。

音韻認識率が 100% でないために、音韻タイプライタが生成した誤った候補が、正しい解析結果よりも大きいスコアを持つことがありえる。したがって、何も対策を講じないと、最悪の場合には、発話全体を覆うような長い未知語を生成することもある。そこで、それぞれの未知語の候補について、そのスコアがどれくらいの長さ (音韻数) の間、辞書引きプロセスのスコアと接近しているかを調べて、長い間接近しつづけている候補については枝刈りする。このようにすれば、未知語の長さに制限を加えることなく、正しい解析をおこなっている区間を覆うような長い未知語を抑制できる。

音韻タイプライタは、なるべく正しい推定ができるように、内部的には、非常に多くの候補を生成している。したがって、同じフレームで、同じカテゴリについて、未知語の部分以外は同じ解析履歴を持つ音韻系列の異なる未知語がいくつも生成される。これらの未知語は、これ以後、全く同じ解析をおこなうので、これらの未知語のうち、最もスコアの高い候補だけを構文解析部にかえし、それ以外のものは捨てれば、無駄な計算を減らせる。

3.5 評価

この手法を用いて、本論文でのべている音声対話システムで用いているタスクと同程度の難しさのタスクを用いて、未知語が入力の 15% 程度になるような条件で実験したところ、未知語の検出率 75% 、未知語部分の音韻認識率 51% という結果が得られている [7]。

この実験では、最終的な認識結果を選択する段階では、未知語を含む結果が未知語を全く含まない結果より適当なスコアだけしか大きくない場合には、未知語を全く含まない結果を優先するようにした。

4 未知語の登録

4.1 語の音韻表記の決定

音声認識用の辞書に新たに語を追加するためには、その語がどのような音韻系列からなるかがわからなければならない。日本語の場合、ある語がどのような音韻系列であるかは、だいたいつづりと一致しているが、例えば、「高校生」のように「コーコーセー、コウコウセー、…」と、厳密にはどのように発話するのか、よくわからなかったり、人によってかなり違うものもある。このような問題は、英語ではさらに深刻で、つづりだけからは、どのように発話するのか推定するのはかなり難しい。

英語の場合に、新しく語を追加する場合の音韻系列の決定を自動化した手法として、利用者がつづりだけを与えると、あとは、システムが自動的に音声合成装置を用いて音韻系列を決定する手法が提案されている

[8]。しかし、この手法では、人間がキーボードを用いてつづりを教えなければならないうえ、当然ながら、音声合成装置が知らない単語の音韻系列は推定誤りをおこす可能性がある。また、Asadi らの論文では、人間が与える音韻系列ですら、4% 程度の単語に誤りがみられると報告されている。

ここでは、新しく語を追加するのに、前節で述べた未知語を検出するときに、推定した音韻系列を用いる。このようにすれば、システムがある語を登録すると決定したら、あとは完全に自動的に新しい語を登録することができる。当然、音韻系列は、誤りを含む。しかし、この手法では、正確ではないにしろ、いちど発話された語の音響的な特徴をふまえた音韻系列を用いるので、人間が追加しようとしたときにどのように発音するのか判断に迷うような語を登録する場合には有利である。また、システム内の閉じた世界では、推定した音韻系列が誤っている場合でも、同じシステムで認識をおこなう場合には、「正しい」音韻系列よりも、システムの音韻モデルが認識しやすい音韻系列となっている可能性もある。

4.2 語の登録

ここでは、なるべく自然で利用者をわずらわせない方法で未知語を登録していく方法を考える。こういった方法として、Gorin らが、単語を認識の単位とするシステムを用いて、全く語をしらない段階から、語彙を構築していく手法を提案している [9]。このシステムでは、離散発声された単語列を入力、行動 (3 か所への内線案内) を出力とするニューラルネットを用いて、その行動に対する利用者からの反応をもとに、単語列とそれに対する行動を学習していく。この手法は、音響的なレベルから意味的なレベルまで学習している。しかし、単語を認識の単位としていると、語彙が拡張されればされるほど、効率・精度の悪化が問題になる。また、日本語で、単語を離散発声することを利用者に強要することは、利用者の負担がかなり重くなる。

われわれは、音声対話システムが認識や対話をおこなうために用いる知識を最大限に利用して、新しい語の登録をおこなう。ここで用いる未知語の検出手法では、未知語がどの文法カテゴリとしてあらわれたかを推定できる。したがって、システムが「本郷三丁目」という語を知らなくても、以下にあげるようなやりとりが可能になる。

利用者　：本郷三丁目まで行く行き方を教えて下さい。
(システムは「ホンゴウサンキョウエンまでどのくらいかかりますか」と認識、
「までどのくらいかかりますか」の部分から「ホンゴウサンキョウエン」が場所をあらわす言葉であると判断する。)
システム：残念ながら、ホンゴウサンキョウエンまでの行き方は知りません。

また、システムが、「料金」という言葉を知らない場合に、「料金はどのぐらいかかりますか。」と発話された場合は、「はどのくらいかかりますか。」の部分の解析結果からは、「料金」が「値段」をあらわすものであるか「時間」をあらわすものであるという曖昧性が残る。こういった場合には、以下のようなやりとりを利用者とかわすことで、曖昧性を解消すればよい。

利用者　：料金はどのくらいかかりますか。
(システム は「ヨキンはどのくらいかかりますか」と認識)
システム：値段のことをお尋ねですか。それとも時間ですか。
利用者　：値段です。

これで、「ヨキン」という未知語を「値段」をあらわす言葉として登録すれば、次に「料金」という単語が発話された場合には、以下のようなやりとりがかわされる。

利用者　：料金はいくらぐらいでしょう。
(システムは「ヨキンはいくらぐらいでしょう」と認識し、ヨキンは「金額」をあらわす言葉だとわかっているので、以下のように応答する)
システム：1230 円です。

また、実際に、検出・推定された未知語を登録して、語彙を拡張するやり方では、認識誤り・未知語の検出誤りから生じる正しくない未知語の登録をさけるため、ある語を登録する前に、その語をシステムに登録するかどうかも利用者に確認するという段階が必要だろう。

5 実験

実験には、交通案内用の文法と辞書を用いた。まず、この辞書には含まれていない「所沢」「池袋」という地名を含んだ以下の発話を認識した。

所沢まで行きたいんですけど。
池袋です。

結果を次に示す。

t o k o r o z o N まで行きたいんですけど。
i ch i u k o r o です。

このように、未知語の部分の音韻系列の推定には誤りが含まれるが、未知語の検出には成功している。そこで、この「トコロゾン」「イチウコロ」という音韻系列がそれぞれ「所沢」「池袋」であるとして辞書に追加した場合と、正しい「トコロザワ」「イケブクロ」という音韻系列を辞書に追加した場合を比較してみる。そのための認識実験を、以下にしめす 3 種類の条件のもとでおこなった。

1. 推定に用いた発話
2. 推定に用いた発話と同じ発話を、もういちど話したもの
3. 推定に用いた発話と同じ発話を、別の話者が話したもの

このときの認識結果の HMM のスコアを以下にしめす。(— は、認識に失敗したことをあらわす。)

未知語	条件	推定結果を追加	正しいものを追加
所沢	1	1.37	1.22
	2	1.45	1.37
	3	0.74	0.80
池袋	1	2.39	2.25
	2	—	1.69
	3	1.22	—

「所沢」の結果をみてみると、推定するのに用いた発話と同じ話者の場合は、推定するのに用いたのとは違う発話でも、正しい表記よりもむしろスコアがよい。また、違う話者の場合でも、ほとんど同じスコアになっており、推定した音韻系列を用いた語彙の拡張の可能性をしめしている。さらに興味深いのが「池袋」の結果である。推定するのに用いた発話の場合に、正しい表記より推定したものの方がスコアがよいのは、「所沢」のときと同じである。しかし、同じ話者の違う発話では、推定結果を用いた方は、認識の途中で枝刈りされてしまう。しかし、違う話者の場合は、逆に正しい表記を用いた場合に、認識の途中で枝刈りされてしまっている。このように、常に同じ音韻モデルを用いる場合には、正しい表記を使えばよいというものでもないことがわかる。

6 むすび

音声対話システムにおいて、未知語が含まれる発話が入力された場合に、その検出・推定をおこない、その結果をもとに、利用者と自然なやりとりをおこなって、自動的にその未知語をシステムの知識に追加していく手法について提案した。

認識実験の結果から、複雑さの小さいタスクでは、推定結果の音韻系列をそのまま用いて半自動的に語彙を拡張することもある程度は可能だろう。しかし、統計的な言語モデルを利用しているため、どんどん未知語を登録して拡張していった場合に、新しく登録された単語の弁別性が保たれるのかなど、まだまだ明らかにしなければならない問題点がある。

また、究極のインターフェイスとしては、ときおり、利用者とやりとりしながら、言葉遣いやその利用者の発声の音響的な特徴など様々な面でのふるまいにたいして適応していくインターフェイスが考えられる。そういったシステムの実現のためには、本稿でおこなった実験のように、利用者のふるまいとして語彙に含まれない語を発話する可能性の他に、システムのもつ音韻モデルとはほど遠い発声をする可能性や、システムのしらない言葉遣い (文法など) をする可能性などを考慮する必要がある。

本稿で提案した未知語の音韻系列の推定に関しては、発声の変動を考えると、一度の発話で音韻系列を固定してしまうのは、余り信頼性が高いとはいえない。したがって、登録された後も、他の知識からえられる情報などを利用したり、その語に発声されたときの情報を考慮に入れて修正していけるような枠組が必要だろう。

今後は、実際に音声対話システムを用いた実験とデータ収集をおこないながら、タスクの自然な拡張・設定方法や、タスクやシステムの利用者への適応について考えていく予定である。

謝辞

本稿で使用した N グラムモデルを作成するのに利用したテキストデータベースのうち、日本経済新聞の記事については、NTT 情報通信処理研究所メッセージシステム研究部から、国際会議に関するものについては、ATR 自動翻訳電話研究所から提供していただきました。貴重なデータを使用させていただいたことを感謝いたします。音韻モデルを作成するのに用いた連続音声資料は、日本音響学会の研究用連続音声データベースの一部であり、関係各位のご尽力に感謝いたします。

また、日頃御討論いただく、東工大田中研の皆様、並びに電総研知能情報部音声研究室の皆様に感謝します。

参考文献

[1] 伊藤克亘, 速水悟, 田中穂積. 連続音声認識における未知語の扱い. 信学会技術報告, Vol. SP91-96, pp. 41-47, 12 1991.

[2] 速水悟, 伊藤克亘, 田中和世. 音声対話システムの構築とそれを用いた会話音声収集. 信学会技術報告, 12 1991.

[3] 堀内靖雄, 伊藤克亘, 田中穂積. 拡張 LR 構文解析アルゴリズムによる未定義語を含む日本語文の構文解析. 情報処理学会 第 40 回全国大会, pp. 325 - 326, 1990.

[4] T. Kawabata, T. Hanazawa, K. Itou, and K. Shikano. Japanese phonetic typewriter using HMM phone recognition and stochastic phone-sequence modeling. *IEICE Transactions*, Vol. E 74, No. 7, pp. 1783-1787, 1991.

[5] 伊藤克亘, 速水悟, 田中穂積. 連続音声認識システム NINJA での音韻連鎖統計情報の利用. 人工知能学会 第 5 回全国大会, pp. 599-602, 1991.

[6] 村上仁一, 荒木哲郎, 池原悟. 2 重マルコフ連鎖確率モデルを使用した単音節音声入力の改善. 信学会技術報告, Vol. SP88-29, pp. 63-70, 1988.

[7] 伊藤克亘, 速水悟, 田中穂積. 連続音声認識システム ninja への未知語処理の導入. 日本音響学会講演論文集, pp. 115-116, 3 1992.

[8] A. Asadi, R. Schwartz, and J. Makhoul. Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. In *Proc. ICASSP-91*, pp. 305-308, 1991.

[9] A. L. Gorin, S. E. Levinson, and A. N. Gertner. Adaptive acquisition of speken language. In *Proc. ICASSP-91*, pp. 805-808. IEEE, 1991.