LE-007

マルチドメイン音声対話システムにおける想定外発話への対処のための Web を用いたシステム知識の拡張

Expanding System Knowledge by Using Web Texts to Manage Out-of-Grammar Utterances in Multi-Domain Spoken Dialogue Systems

池田 智志† 駒谷 和範† 尾形 哲也† 奥乃 博†
Satoshi Ikeda **Kazunori Komatani** **Tetsuya Ogata** **Hiroshi G. Okuno**

1. はじめに

マルチドメイン音声対話システムは，構築に多大な労力がかかる．そのため，ドメインの追加，修正が容易であること (ドメイン拡張性) が必要であり，独立に設計されたドメインを統合するアーキテクチャが提案されている [1]．このようなアーキテクチャでは，想定外発話が重要な課題となる．まず，各ドメインの文法が一貫しておらず，ユーザが各ドメインで受理できる発話を推測しづらい．さらに，システムの扱うタスクが広く，ユーザの発話が多様になる．このため，ユーザ発話が想定外発話となりやすい．本研究では，『ユーザが本来意図していたドメイン』をトピックと定義し，これを推定する．これにより，当該ドメインにおけるヘルプ生成など，想定外発話への適切な応答が可能となる．

このトピック推定において課題となるのは，ドメインの拡張性である．関連研究 [2] では，あらかじめ用意された対話コーパスを用いたトピック推定に基づき，システムの扱わない話題を検出している．しかし，対話コーパスが存在しない場合，ドメインを追加するには対話コーパスを新たに収集する必要がある．これは，ドメイン拡張性を満たさない．本稿では，ドメイン拡張性を損わずに，想定外発話に対処する方法を新たに開発したので，これについて述べる．

2. ドメイン拡張性を備えた想定外発話の対処

本研究では，マルチドメインシステム中の 1 つのサブシステムを“ドメイン”と定義し，ドメインで受理・解釈できる発話の集合を“ドメイン内発話”と定義する．また，システムのいずれかのドメイン内発話に含まれる発話をシステム想定内発話とする．一方，システム想定外であっても，あるドメインの内容を意図した発話の集合を“トピック”と定義する．ドメインとトピックの関係及びその具体例を図 1 に示す．

本研究では，学習データの Web からの収集と Latent Semantic Mapping (LSM)[3] を用いることで，ドメイン拡張性を備えたトピック推定を実現する．Web から学習データを収集することで，あらゆるトピックに対して大量の文書を容易に収集できる．これにより，ドメイン拡張性が満たされる．しかし，収集された文書には求めるトピックに強く関連する文書のみが含まれるとは限らず，ノイズが混在する．そこで，単語と文書の関係を低次元の空間に写像することで，文書の潜在的な意味を表現できる LSM を用いる．これにより，学習データのノイズの影響を取り除いたトピック推定を可能とする．

トピック推定に基づきヘルプを提示する対話例を図 2 に示す．U1 はシステムにとって想定外発話であるため，U1 の言語理解結果を棄却[1]し，トピック推定を行う．この結果，U1 のトピックが観光案内と推定され，システムは S1 で観光案内ドメインの詳細なヘルプを提示できる．

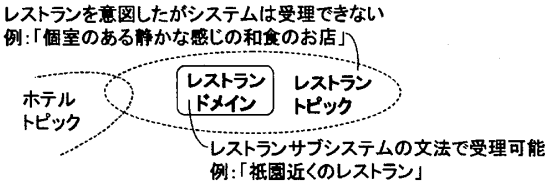

図 1: ドメインとトピックの関係及びその具体例

U1：	京都のお勧めスポットを教えて． (『お勧めスポット』が言語理解文法語彙外．言語理解結果を棄却し，観光案内トピックと推定．)
S1：	理解できませんでした．観光案内については寺社，公園，博物館，名所などの検索ができます．例えば，「祇園近くの寺社」のように答えてください． (トピック推定に基づき，観光案内のヘルプを提示．)
U2：	左京区にある観光名所．

図 2: トピック推定により可能となる対話

3. Web からの大量データの自動収集と LSM を用いたトピック推定

以降，レストラン，観光案内，バス，ホテル，天気の 5 ドメイン音声対話システムを例として説明を進める．これらの各ドメインに加えて，「はい」など，どのドメインにも共通した発話をコマンドとする．トピック推定の概略を図 3 に示し，以下で順に説明する．

3.1 学習データの収集

コマンド以外の 5 つのトピックに関して，ツール [5] を用い，学習データを収集する．各トピックごとに人手で 10 個前後のキーワードを指定し[2]，5000 程度の Web ページから学習データとして 10 万文を収集した．コマンドの学習データは Web から収集せず，175 文を人手で準備した．また，システムの言語理解用文法からは各トピックにつき 1 万文を生成し，学習データに加えた．以上の作業で収集した文書をトピックごとに d 個に分割し，学習文書を構成した．ここで d はトピックごとの学習文書の数である．

3.2 Latent Semantic Mapping を用いたトピック推定

各トピックに対する学習文書集合と入力発話との近さを LSM[3] を用いて計算することで，トピック推定を行う．具体的な手法を以下に示す．

各学習文書に対する単語の頻度をもとに得られる $M \times N$ 共起行列を求める．ここで，M は学習文書集合に現れる異なり単語数，N は学習文書数である．また，トピッ

†京都大学大学院 情報学研究科 知能情報学専攻

[1]発話の受理，棄却の判断には検証用言語モデルとの音声認識の対数音響尤度差などを用いる [4].

[2]本研究では，各トピックごとに，キーワードを数セット試し，主観的に最も適切なキーワードを採用した．

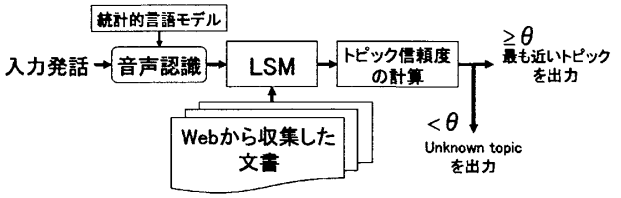

図 3: トピック推定の概略

クの数を n，トピックごとの学習文書数を d とすると，$N = n \times d$ と表される．その共起行列に対して特異値分解と次元縮約を行い，共起行列の階数を k に減じる．また特異値分解をもとに，N 個の学習文書それぞれに対して k 次元空間でのベクトル表現を得る．本研究で作成した共起行列は，$M = 67533$，$N = 120$，$n = 6$，$d = 20$ である．次元縮約に関しては $k = 50$ とした．

入力発話の音声認識には，LSM の学習データから学習した言語モデルを用いる．その音声認識結果に対して，単語の頻度と特異値分解の際に算出した行列を用いて k 次元ベクトルを求める．このベクトルと学習時に求めた N 個の k 次元ベクトルとのコサイン距離を計算することで，トピック推定を行う．ここで，トピックに属する d 個の学習文書と入力発話とのコサイン距離の最大値を，トピックと入力発話の近さと定義する．

3.3 トピックが不明な発話の処理

入力発話のみからでは，トピックが不明な場合がある．例えば，「上限予算五千円」という発話は，その発話のみではホテルとレストランの両方のトピックの可能性があり，文脈に依存している．トピックが不明な発話を **"unknown topic"** とし，以下の手順に従って推定する．

まず，入力発話に最も近いトピック T を求める．次に，トピック T の信頼度を $CM_T = closeness_T / \sum_t closeness_t$ として CM_T を求める．ここで，$closeness_t$ はトピック t と入力発話の近さである．$CM_T > \theta_1$ ならトピックの推定結果を T とする．それ以外の場合，トピックの推定結果を unknown topic とする．unknown topic には，文脈情報等をもとに，対話管理や発話誘導を行っていく．

4. 評価実験

4.1 評価用対話データ

評価には教示あり対話データと教示なし対話データの 2 種類を用いた．教示あり対話データは，10 分間の練習後，実際のシステム [6] を用いて収集した話者 10 名 2129 発話からなる．これは，「あさっての祇園の天気」などの想定内発話を多く含む．教示なし対話データは，模擬対話により 8 名の被験者から 272 発話を新たに収集したもので，初心者ユーザによるシステムの使用条件に近い．このため，「京都で紅葉のきれいなお勧めスポットはありますか.」などの想定外発話を多く含む．

音声認識には Julius[3] を用いた．言語モデルは LSM の学習データから構築し，音響モデルは性別非依存 3000 状態 PTM を用いた．単語正解率は，教示あり対話データに対しては 69.6%であり，教示なし対話データに対しては 67.3%であった．

4.2 トピック推定の評価

ベースライン手法として，各ドメイン文法により生成された文章のみを学習データとして共起行列を作成し，次元縮約を行わずにトピックを推定する．これは，Web からの学習データの収集 (以後，Web 収集) と LSM の両方を用いずにトピック推定を行った場合に相当する．

表 1: 各手法におけるトピック推定の正解率

	教示あり対話データ	教示なし対話データ
ベースライン	53.5%	37.9%
+(1) Web 収集	51.2%	44.5%
+(2) LSM	60.8%	45.6%
+(1)+(2) (本手法)	**60.4%**	**61.0%**

ベースライン手法，Web 収集のみを用いた手法，LSM のみを用いた手法，本手法の 4 手法を比較評価した．音声認識結果に対する，各手法のトピック推定の正解率を表 1 に示す．ここで，正解ラベルは発話ごとにレストラン，観光案内，バス，ホテル，天気，コマンド，unknown topic のいずれかに人手で付与されている．

ベースライン手法では，教示なし対話データに対する正解率が教示あり対話データより 15.6 ポイント低い．これは，教示なし対話データには想定外発話が多く含まれるからである．Web 収集を用いた手法では，Web から大量の学習データを収集し，システムの知識を補強することで，教示なし対話データに対して正解率が 6.6 ポイント改善した．一方で，Web 収集と LSM を用いる本手法ではさらに 16.5 ポイント改善した．この大幅な正解率の改善は，Web から収集したデータに含まれるノイズを，LSM を用いて取り除いているためである．これは，Web 収集と LSM の 2 つのアプローチを同時に行うことの有効性を示している．

正解率が全体的に低いのは，本研究が一発話のみからより多くの情報を得ることを目的としており，文脈情報などを一切用いていないからである．実際，対話データから unknown topic を除いた場合，本手法では，教示あり対話データ (1663 発話) に対しては 70.7%，教示なし対話データ (202 発話) に対しては 67.8%の正解率でトピックを推定した．音声認識における単語正解率が 70%程度であることを考慮すると，これは妥当な値であり，本手法が想定外発話に対しても頑健なトピック推定を実現したといえる．対話における文脈情報 [6] との統合は今後検討を進める．

謝辞: 本研究の一部は科研費，SCAT の支援を受けた．Web からの文書収集には京都大学河原研究室で開発されたツール [5] を用いた．

参考文献

[1] B. Lin, H. Wang, and L. Lee. A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history. In *Proc. ASRU*, pp. 345–348, Keystone, USA, 1999.

[2] I. R. Lane, T. Kawahara, T. Matsui, and S. Nakamura. Topic classification and verification modeling for out-of-domain utterance detection. In *Proc. ICSLP*, pp. 2197–2200, 2004.

[3] J.R. Bellegarda. Latent semantic mapping. *IEEE Signal Processing Mag*, Vol. 22, No. 5, pp. 70–80, 2005.

[4] 福林, 駒谷, 尾形, 奥乃. 音声対話システムにおける発話検証を利用したシステム想定外発話の誤受理抑制. 情処研報, SLP-65-12, 2007.

[5] T. Misu and T. Kawahara. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Proc. Interspeech*, pp. 9–12, 2006.

[6] 神田, 駒谷, 中野, 中臺, 辻野, 尾形, 奥乃. マルチドメイン音声対話システムにおける対話履歴を利用したドメイン選択. 情報処理学会論文誌, Vol. 48, No. 5, pp. 1980–1989, 2007.

[3] http://julius.sourceforge.jp/