3U-3

トピック推定と対話履歴の統合によるドメイン選択を行うマルチドメイン音声対話システム

池田 智志　　駒谷 和範　　尾形 哲也　　奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

マルチドメイン音声対話システムは一般に，独立に設計された単一ドメインを統合して構築されることが多い．このようなアーキテクチャでは，ユーザの要求に応答すべきドメインの決定処理 (**ドメイン選択**) が不可欠である．一方で，初心者の発話は，システムが受理できない**システム想定外発話**を多く含む．これらの発話は音声認識誤りの原因となり，ドメイン選択誤りを引き起こす．システムがユーザのあらゆる発話を全て言語理解できるよう，語彙や文法を網羅的に記述するのは不可能であるため，想定外発話は不可避である．神田らが開発した，対話履歴を用いたドメイン選択手法 [1] も，想定外発話からはドメイン選択に有効な情報を取得できなかった．

本稿では，トピック推定 [2] と対話履歴の利用を統合することで，想定外発話に頑健なドメイン選択を行う．トピック推定結果は想定外発話に対しても比較的信頼できるのに対し，対話履歴は想定外発話に起因する言語理解誤りの悪影響を受ける．一方，トピック推定は一発話のみに対して行われるが，対話履歴は文脈を考慮している．このように，トピック推定と対話履歴は相補的な情報であり，これらを統合することで，想定外発話に頑健なドメイン選択が可能となる．

2. 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

本研究では，マルチドメインシステム内のサブシステムをドメインと定義し，いずれのドメインにおいても受理できない発話の集合を **“システム想定外発話”** と定義する．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．トピック推定により，想定外発話に対してもドメイン選択に有効な情報を取得できる．

本研究では，文献 [1] の手法とトピック推定を統合し，想定外発話に頑健なドメイン選択を行う．ドメイン選択の概略を図 1 に示す．すなわち，文献 [1] で用いた特徴量に，トピック推定から得られる特徴量を加え，(I) ひとつ前の応答を行ったドメイン，(II) 言語理解に対して最尤のドメイン，(III) トピック推定に対して最尤のドメイン，(IV) それ以外のドメイン，の判別を行う．

本手法により可能となる対話例を図 2 に示す．U2 において，正解ドメインは観光であり，(I) の天気でも (II) のバスでもない．手法 [1] では，(I)，(II)，(IV') (= (I), (II) のいずれでもない) のみを考慮してドメインを選択するため，$S2_2$ (手法 [1]) では (IV') を選択する．このため，応答すべきドメインが一意に決定できず，具体的な応答ができない．これに対して本手法では，U2 のトピック推定結果が観光トピックであることから，(III) の観光ドメインを選択できる．これにより，$S2_3$ (本手法) のように，ドメインに応じた具体的な応答が可能となる．

Integrating Topic Estimation and Dialogue History for Domain Selection in Multi-Domain Spoken Dialogue System: *Satoshi Ikeda, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

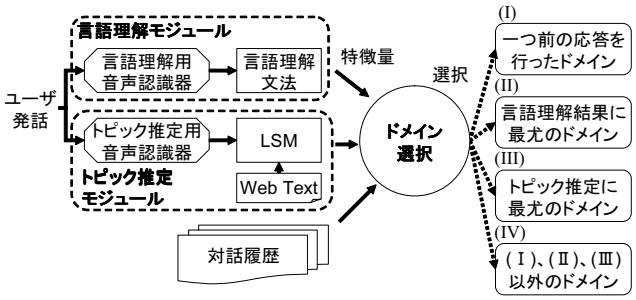

図 1: ドメイン選択の概略

U1: 明日の京都の天気を教えて (正解ドメイン: 天気)

S1: 明日の京都の天気は，晴れです．

U2: 京都の 夜景がきれいな場所 (正解ドメイン: 観光)
(下線部が文法外.「京都外大前より競馬場」と誤認識)

いずれか

- $S2_1$ (単純手法): 京都外大前から競馬場までですか？ (選択されたドメイン: バス)
- $S2_2$ (手法 [1]): 理解できませんでした． (選択されたドメイン: その他 (IV'))
- **$S2_3$ (本手法): 理解できませんでした．観光について，場所や施設タイプなどが指定できます.「祇園周辺の寺を検索」などとおっしゃって下さい.** (選択されたドメイン: 観光 (III))

図 2: 本手法により可能となる対話例

3. 想定外発話に頑健なドメイン選択

本研究では，ユーザ発話と Web から収集した文書との近さを Latent Semantic Mapping (LSM)[3] を用いて計算し，トピックを推定する．トピック推定の詳細は文献 [2] にある．さらに，対話履歴とトピック推定から得られる特徴量に基づき，ドメイン選択器を構築する．

3.1 Web からの大量文書の自動収集と LSM に基づくトピック推定

まず，ツール [4] を用いてトピックごとに 10 万文の Web 文書を収集する．さらに，システムの言語理解用文法から各トピックにつき 1 万文を生成し，収集した Web 文書に加える．以上の作業で各ドメインごとに収集した文書をランダムに d 個に分割し，学習文書を構成する．ただし，この学習文書は自動的に収集されるため，当該トピックとは無関係な文書がノイズとして混在する．

次に，LSM を用いて学習文書のノイズの影響を除去する．具体的には，まず，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．共起行列に対して次元圧縮を行い，各学習文書に対応する k 次元ベクトルを得る．本研究で作成した共起行列は，$M = 67,533$，$N = 120$，$d = 20$，$k = 50$ である．トピックに属する d 個の学習文書の k 次元ベクトルと，入力発話の k 次元ベクトルとのコサイン距離に基づき，トピックと入力発話の近さを求める．

表 1: トピック推定に関する特徴量

T1:	(Ⅲ) に対応するトピックと発話の認識結果との近さ
T2:	(Ⅲ) に対応するトピックの信頼度
T3:	(I) に対応するトピックと発話の認識結果との近さ
T4:	(I) に対応するトピックの信頼度
T5:	ユーザ発話と (I), (Ⅲ) の近さの差 (=T1 - T3)
T6:	(I) と (Ⅲ) のトピック信頼度の差 (=T2 - T4)
T7:	(Ⅲ) と (Ⅱ) が一致するか
T8:	(Ⅲ) と (I) が一致するか
T9:	(Ⅲ) がコマンドトピックかどうか
T10:	トピック推定用音声認識器による認識結果の長さ (音素数)
T11:	トピック推定用音声認識器による認識結果の音響スコア
T12:	T11 と U1 の一音素あたりの音響尤度差 (=(T11 - U1)/T10)
T13:	T11 と U4 の一音素あたりの音響尤度差 (=(T11 - U4)/T10)

3.2 トピック推定と対話履歴の統合

本研究では，文献 [1] で使用した特徴量に加えて，トピック推定に関する特徴量 (表 1) を新たに導入する．これらの特徴量を入力とし，応答すべきドメインを出力とする決定木を，対話データから学習する．以下では，新たに導入した特徴量について述べる．T1〜T6 は，トピック推定結果がどの程度信頼できるかの指標である．ここで，トピック T の信頼度は，$CM_T = closeness_T / \sum_t closeness_t$ として定義する．t はシステムに存在するトピックであり，$closeness_t$ はトピック t と入力発話の近さである．次に，ドメイン (I), (Ⅱ), (Ⅲ) の関係を表すために，T7〜T9 を導入した．例えば，(I) と (Ⅲ) が一致する場合は，このドメインの信頼性は高いと判断できるからである．T10 は，音声認識結果が短い発話のトピック推定結果は信頼できない場合が多いという傾向があるため定義した．T11〜T13 は，ユーザ発話が想定外発話かどうかの情報を表す．ユーザ発話が想定外発話であると判定されれば，(Ⅱ) よりも (Ⅲ) の方が信頼性が高いと判断できる．

4. 評価実験

4.1 評価対象の対話データ

評価には，文献 [1] において，5 ドメインシステムを用いて収集された話者 10 名 2191 発話を用いた．言語理解用音声認識には Julian[5] を用いた．音声認識文法は，各ドメインの言語理解文法からの自動生成により得た．トピック推定用音声認識には，Julius[5] を用いた．言語モデルは，トピック推定の際に使用した学習データを用いて構築した．語彙サイズは，それぞれ 7,373，56,453 であった．音響モデルは 3000 状態不特定話者 PTM トライフォンモデル [5] を用いた．単語正解率はそれぞれ 63.3%，67.3%であった．また，対話データには，発話ごとに 2. 章で述べた (I)〜(Ⅳ) のいずれかを人手でラベル付けした．ただし，一つの発話に複数のラベルが付けられる場合には，(I)，(Ⅱ)，(Ⅲ)，(Ⅳ) の順に優先した．

決定木の構築には，C5.0[6] を用いた．特徴量は，文献 [1] と表 1 の 43 個から，Backward stepwise selection により選択された 27 の特徴量を用いた．評価は 10-fold cross validation を用いて，発話ごとに行った．また，最高スコアのドメインが複数存在した場合，その中から無作為にドメインを選択して正解判定を行った．

4.2 ドメイン選択精度の評価

以下をベースライン手法として，比較評価を行った．

ベースライン手法: 文献 [1] に基づきドメインを選択する．ドメイン選択器の構築には，文献 [1] の 30 個から本手法と同様に選択された 16 の特徴量を用いた．

表 2: ベースラインと本手法の比較 (正解数/発話数)

正解ラベル \ 手法	ベースライン	本手法
(I) 一つ前	1303/1442	1348/1442
(Ⅱ) 言語理解最尤	238/380	258/ 380
(Ⅲ) トピック推定最尤	0/0	37/131
(Ⅳ) その他	131/369	84/238
計	1672/2191	1727/2191

正解ラベルごとのドメイン選択の正解数を表 2 に示す．ベースライン手法におけるドメイン選択誤り数は 519 である．ここで，ベースライン手法におけるドメイン選択の正解基準を本手法のドメイン選択結果に適用した場合，本手法のドメイン選択誤り数は，445 ($= 2191 - 1727 - 13 - 6$) である‡．このとき，ドメイン選択誤り削減率は 14.3% (= 74/519) となる．また，正解ラベルが (I) や (Ⅱ) の場合にも正解率の改善が見られる．これは，T7 や T8 が，(I) と (Ⅱ) の判別に効果的な情報となるためである．

具体的なドメインが正しく選択できた発話数は，ベースライン手法において 1541($= 1303 + 238$) である．一方で，本手法では 1643 ($= 1348 + 258 + 37$) であり，ベースライン手法から 102 発話増加している．このうち，正解ラベルが (Ⅲ) である 37 発話は，従来手法では本質的に具体的なドメインを選択できない発話である．これらの発話には，「京大正門前へ行くバス は動いていますか」(下線部が文法外) などの想定外発話が含まれていた．これは，本手法が，より広範囲のユーザ発話に対して具体的なドメインを決定できることを示している．

5. おわりに

本研究では，マルチドメイン音声対話システムにおいて，システム想定外発話に対しても，応答すべきドメインを頑健に選択する手法について述べた．今後は，(Ⅲ) や (Ⅳ) が選択された場合など，言語理解結果が得られない際の対話戦略の検討を進める．

謝辞 トピック推定の学習データは京都大学河原研究室開発のツール [4] を用いて収集した．評価データは，HRI-JP の中野幹生氏らとの共同研究において，神田直之氏らとともに構築したシステムにより収集した．本研究の一部は，科研費，GCOE，SCAT 研究助成の援助を受けた．

参考文献

[1] 神田直之, 他. 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. 情処学論, Vol. 48, No. 5, pp. 1980–1989, 2007.

[2] S. Ikeda et al. Topic estimation with domain extensibility for guiding user's out-of-grammar utterance in multi-domain spoken dialogue systems. In *Proc. Interspeech*, pp. 2561–2564, 2007.

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

[4] 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.

[5] 河原達也, 李晃伸. 連続音声認識ソフトウエア Julius. 人工知能学会誌, Vol. 20, No. 1, pp. 41–49, 2005.

[6] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. http://www.rulequest.com/see5-info.html.

‡本手法におけるドメイン選択結果には，(Ⅲ) を (Ⅳ) に誤って識別した発話が 6，(Ⅳ) を (Ⅲ) に誤って識別した発話が 13 存在する．