6T-3

データベース検索音声対話システムにおける対話を通じた店舗属性取得

大塚 嗣巳 † 駒谷 和範 ‡ 佐藤 理史 ‡ 中野 幹生 §

† 名古屋大学 工学部 ‡ 名古屋大学大学院 工学研究科 § ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

本研究はデータベース検索音声対話システムにおいて，未知語の属性の対話を通じた取得を目指す．未知語とは，検索対象データベース (以下 DB) 外の単語のことである．本論文では，愛知県内のレストラン検索 DB を対象としており，今回は未知語として店舗名を対象とする．属性とは DB 内の項目のことで，本論文ではジャンル，具体的には居酒屋，中華，カフェなど 16 種類とする．

本研究は賢い質問による効率的な対話を通じて店舗のジャンルを取得する．賢い質問とは，単純に何でも質問するのではなく，システムが可能な範囲で内容を推定した後に行う，より具体的な質問を意味する．図 1 の上部の対話例 1 では，何も推定せず，単純な質問を行っている．この場合，ユーザの応答には何の制限もないため，新たな未知語をユーザが発話する可能性がある．これに対し，対話例 2 のような質問に対しては，ユーザの応答は肯定または否定表現に絞られる．つまり，発話内容の候補が少ないほど，ユーザが DB の内部情報と関連しない単語を言う可能性が低く，効率的である．したがって対話例 2 のような具体的な質問を生成することを目指す．推定及び応答文生成の流れを，図 2 に示す．本研究では以下の 2 つの推定を行う．

1. 検索対象 DB を利用した機械学習に基づく推定
2. Web ページ上のジャンル名の共起頻度に基づく推定

2 つの推定結果に基づき，推定の信頼度を算出し，ユーザからジャンルを効率的に取得できる，より具体的な質問を生成する．この際に，以下の 2 つの課題を解決する必要がある．1 点目は，検索対象 DB を利用した機械学習において過学習を避けることである．2 点目は，2 つの推定結果から，具体的な質問生成に有効な信頼度を得ることである．本稿では，推定手法の実装と，この 2 点の評価実験について述べる．

2. 店舗属性の取得

2.1 DB 内の情報に基づく推定

本研究では，DB 内の店舗名 $s_i \in S$ を入力とし，ジャンル $g_j \in G$ を出力する．例えば，図 1 の対話例 2 の店舗「オステリア リュウ」に関して人間が推定する際，単語「オステリア」の文字の並びから，ジャンルが「イタリアン」だと推定できる．これを機械学習により行う．

機械学習には Maximum Entropy Model(=ME)[1] を用いる．ME により求まる事後確率 $p(g_j \mid s_i)$ を DB 推定における信頼度 $CM_D(g_j)$ とし, 式 (1) で表す．

$$CM_D(g_j) = p(g_j \mid s_i) = \frac{1}{Z} \exp\left[\vec{\lambda} \cdot \vec{\phi}(s_i, g_j)\right] \qquad (1)$$

Acquiring Restaurant Attribute by Generating More Concrete Questions in Spoken Dialogue Systems for Database Search Task: *Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato (Nagoya Univ.), and Mikio Nakano (Honda Research Institute Japan Co., Ltd.)*

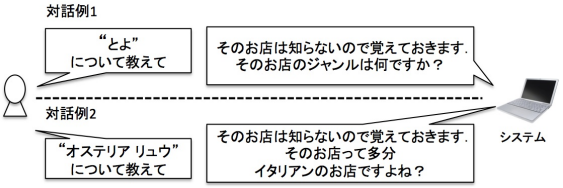

図 1: 対話例

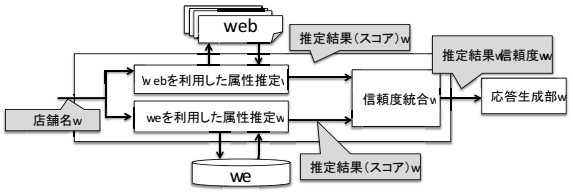

図 2: システム全体像

ここで，$\vec{\phi}(s_i, g_j)$ は g_j と店舗名 s_i に関する素性ベクトルである．$\vec{\lambda}$ は素性ベクトルに対する重みであり，Z は $\sum_{g_j} CM_D(g_j) = 1$ を保証する正規化係数である．

学習データとして，DB 内の店舗名とそのジャンル，店舗名から生成される以下の素性を用いる．

- 店舗名の文字 n-gram (n = 1, 2, 3)
- 店舗名の形態素列
- 店舗名の文字の種類

本研究は形態素解析に Mecab を使用し，その辞書は IPADIC を用いる．文字の種類はひらがな，カタカナ，漢字，アルファベットとする．例えば「IB カフェ」という店舗の場合「カタカナ+アルファベット」となる．対象 DB の登録店舗数は 2,398 件で，25,418 種類の素性が生成される．

ここで，相互情報量に基づく素性選択により過学習を回避する．機械学習に用いる学習データは DB 内の店舗 2398 件であり，これに対して，素性の種類は 10 倍以上もあるため，過学習を引き起こす可能性がある．相互情報量は式 (2) で表される．

$$I(f_k; G) = \sum_{g_j \in G} p(f_k, g_j) \log \frac{p(f_k, g_j)}{p(f_k)p(g_j)} \qquad (2)$$

ここで $p(f_k)$，$p(g_j)$ は学習データから作成される素性 f_k とジャンル g_j それぞれの生起確率，$p(f_k, g_j)$ は同時確率を表す．$I(f_k; G)$ のスコア降順の順位から上位 x %のみを用いる．

2.2 Web 上の頻度に基づく推定

人間は全く未知の店舗を調べた時でも，検索ページ内の文章から店舗のジャンルについて予想できる．DB に存在しない店舗に関するページが Web 上には存在する場合があり，Web を利用した属性値取得に関する研究

表 1: 質問項目数 num に応じた応答形式とその例

num	応答形式	応答例
1	Yes/No 形式	A とは g_1 ですよね？
2	2 者択一形式	A とは g_1，g_2 どちらですか？
3	3 者択一形式	A とは g_1，g_2，g_3 どれですか?
4 以上	5W1H 形式	A のジャンルは何ですか?

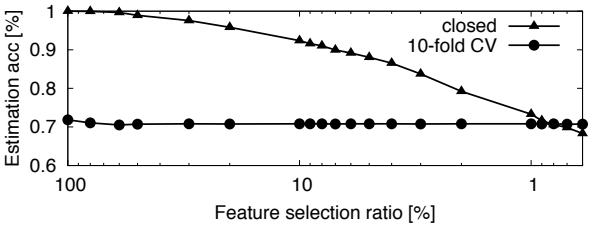

図 3: closed と 10 分割交差検定との比較

[2][3] も多数存在する．本研究でも DB に存在しない情報を補うために Web を利用する．

推定の手順を述べる．まず，検索クエリを「< 検索店舗名 >　愛知県　レストラン」とする．次にこの検索結果に関して，タグを取り除いた HTML ファイルを取得する．取得した HTML ファイルに対して DB 内のジャンル g_j の頻度 $h(g_j)$ を求める．式 (3) により正規化し，Web に基づく信頼度 CM_W のスコア降順のリストとして出力する．

$$CM_W(g_j) = \frac{h(g_j)}{\sum_{g_l \in G} h(g_l)} \tag{3}$$

2.3 推定の信頼度に基づく応答文生成

推定結果が 1 位のジャンルの信頼度が高ければ，その 1 つのみを，上位 2 位のジャンルの信頼度が高い場合は，その 2 つのみを質問項目とする方がより具体的である．より具体的な質問を生成するために，2 つの推定結果を統合した信頼度 $CM_I(g_j)$ を用いて，質問の項目数を決定する．本論文では，信頼度 $CM_I(g_j)$ は $CM_D(g_j)$ と $CM_W(g_j)$ の相加平均として定義する．質問項目数 num は信頼度 $CM_I(g_j)$ を用いて式 (4) で決定する．j は $CM_I(g_j)$ を降順に並べたときの順位を表す．

$$num = min(n)\ \ s.t. \sum_{j=1}^{n} CM_I(g_j) > \theta \tag{4}$$

ここで，θ は定数である．例えば，$n = 1$，つまり第 1 位のスコア $CM_I(g_1)$ だけで θ を超えているとき，その候補 1 つだけを質問項目とする．ある店舗 A の推定結果ジャンル g_i に関する質問項目とその形式の例を表 1 で示す．この表 1 に従って質問文を選択できる．

3. 評価実験

3.1 相互情報量に基づく素性選択

DB に基づく推定において，素性選択が過学習の回避に有効であることを確認する．素性選択により closed テストの正解率と open テストの正解率の差が小さくなることを過学習回避の評価指標とする．ここで正解率は，$CM_D(g_j)$ が最大値となる g_j と，正解ジャンルが一致した件数を，2398 件で割った値とする．open テストには，対象 DB 内の 2,398 件の店舗に対する 10 分割交差検定を用いた．

結果を図 3 に示す．横軸は式 (2) の $I(f_k; G)$ のスコア降順にソートされた全素性のうち，上位 x %を使用した場合を表す．縦軸は正解率を表す．x が減少するにつれて closed テストの正解率が減少することから，本研究の使用する DB では過学習が起こっていたことが確認できた．また $x = 0.8$(203 種類) の時に closed テストと 10 分割交差検定の正解率がほぼ同じとなり，過学習が解消したと言える．

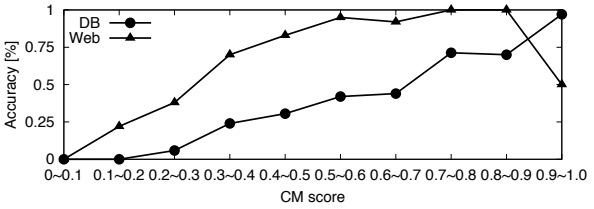

図 4: CM スコア別の正解率

表 2: CM スコアの分布表

CM	CM_D		CM_W	
	正解数	失敗数	正解数	失敗数
0.0 ～ 0.1	0	0	0	34
0.1 ～ 0.2	0	0	5	18
0.2 ～ 0.3	1	16	28	46
0.3 ～ 0.4	6	19	47	20
0.4 ～ 0.5	11	25	62	13
0.5 ～ 0.6	21	29	69	4
0.6 ～ 0.7	22	28	35	3
0.7 ～ 0.8	41	16	13	0
0.8 ～ 0.9	21	9	1	0
0.9 ～ 1.0	131	4	1	1
合計	254	146	261	139

3.2 質問生成に用いる信頼度の分布

信頼度 CM が正解を示す尺度としての有効性を示す．DB 内から 400 件の店舗を抽出して評価を行った．DB に基づく推定では，残りの 1998 件を学習データとした．素性は前節の結果の，203 種類を用いた．$CM(g_j)$ が最大値となるジャンル g_j と正解ジャンルが一致した場合を正解とする．

結果を図 4 と表 2 に示す．図 4 は CM_D, CM_W それぞれの分布をグラフにしたものである．横軸は CM の区間を表す．縦軸は表 2 で示した各 CM 区間内の正解数/(正解数+失敗数) を表す．図 4 の CM_D，CM_W 両方においてグラフが右肩上がりであること及び表 2 の結果から，CM が正解の尺度として有効だと確認できる．

表 2 において CM_D は 0.5〜1.0 の範囲に正解の多くが分布しているのに対し，CM_W は 0.2〜0.7 の範囲に正解の多くが分布している．また，両者は別々の情報として見ることができる．例えば,「しゃぶしゃぶ温野菜 金山駅前店」という店舗は，DB からは「和食」だと推定できたが，Web からは推定できなかった．一方,「高矢禮火 (ゴシレ ファ)」という店舗は，Web からは「焼肉 韓国料理」だと推定できたが，DB からは推定できなかった．これらの例から，推定の成功率をより向上させるためには両者の情報が必要であると言える．今後は 2 つの推定結果の統合方法の検討を行う．

参考文献

[1] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, Vol. 22, No. 1, pp. 39–71, March 1996.

[2] 山本あゆみ, 佐藤理史. ワールドワイドウェブからの人物情報の自動収集. 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 99, No. 534, pp. 93–100, 2000.

[3] 吉永直樹, 鳥澤健太郎. Web からの具体物の属性・属性値情報の自動獲得. 言語処理学会第 13 回年次大会発表論文集, 2007.