

音声対話システムにおける質問応答データベースの分類とその分析

井上僚介

黒澤義明

目良和也

竹澤寿幸

広島市立大学大学院情報科学研究科

[inoue, kurosawa, mera, takezawa]@ls.info.hiroshima-cu.ac.jp

1. はじめに

音声認識の主な用途はコンピュータとの会話や自動書き起こし等が挙げられ、応用分野は多岐に渡る。近年、インターネット検索大手の Google が Android や iPhone 向けに提供している音声検索、また、Apple が iPhone に搭載した対話エージェント Siri 等、これまで以上に音声認識の重要性が高まりつつある。本研究では音声認識を用いたシステムのうち、音声対話システムに焦点を当てる。

音声対話システムでは、ユーザがシステムに対して様々な質問を行う。システムは質問に対応した応答を返す。この動作の実現には音声認識技術だけでなく、対話処理技術、並びに大量の質問応答データベース（以下、QADB とする）が必要となる[1]。しかし大規模な QADB の管理には多大なコストが生じる。また、一般的な対話システムでは一つの適切な応答を返すことを目標としている。例えば、「広島の有名な食べ物は何？」という質問に対して「広島は知らないけど名古屋のきしめんなら知ってるよ」というような応答をシステムが返していたとする。従来のシステムでは正しく広島の有名な食べ物に関する応答が得られることが望ましい。しかし、会話の連続という観点からは違和感の生じる会話ではない。このように、本研究では会話の連続という観点から自然な対話を行うシステムの構築を目指す。

以上のような、より自然なシステムを実現するためにはデータベースのクラスタリングが必要であると考えられる。例えば、『A は B です』という文章と『A は C です』という文章から同じ文脈上において B と C は交換可能であると考えられる。すなわち、同種の文脈という観点で分析するならば、クラスタリングを行うことで関係性が獲得できる。この交換可能な関係性に着目することでより柔軟なシステムを構築することが可能である。そこで本研究では、Probabilistic Latent Semantic Analysis（以下、pLSA）を用いて QADB をクラスタリングすることを提案する。これにより、関係性のある質問や応答を抽出できることが期待され、自然に対話を行うことが出来る音声対話システムの構築を目指す。

2. 関連研究

本研究に関連する研究及び先行研究について紹介する。一点は一問一答型質問応答データベースにおける質問例選択の基本となる形態素マッチングによる手法、もう一点は、自然な対話を行うためにデー

タベースの分類を行う手法である。

2.1. 用例ベースの応答選択

用例ベースの応答選択では、質問例の応答データベースを用いて、入力に最も近い質問例から応答を生成する。具体的には、質問例と正解応答の対を QA ペアとしてデータベースに登録しておき、入力が生じるとデータベースから類似した質問例を選択することで、対応する正解応答をユーザに提示する[1]。

質問例の選択手法として、入力に対して最も類似した一例のみを質問例から参照し、これに対応した適切な応答例を取得するという簡便な最近傍法が用いられている。この際、文章間の類似度を求める手法として、文章を単語に分割した上で一致した単語数を比較するという単語マッチングの手法が考えられる。日本語の文章においては単語に分割することはすなわち形態素に分割することと同義となるので、これは形態素マッチングとも呼ばれる。この手法を拡張し、単語数の平均をとり、更に音声入力にはある一定の誤り傾向が存在すると仮定すると、この誤り傾向を質問例に含めることで、一定の誤り傾向に対して対応することが可能になる。これらの手法を用いてシステム応答の適切さを向上させる。

2.2. データベースの分類

Anton ら[2]は大規模な QADB をより小規模なクラスタに分類することで応答精度の向上を試みた。“質問例のみを分類”、“応答例のみを分類”という手法で QADB の分類を行った。

Q1	広島の名物は？	A1	川通り餅！
Q2	広島で有名なおみやげは？	A2	お好み焼きだよ
Q3	有名な食べ物は何？	A3	お好み焼きです
Q4	何が名物？	A4	食べ物ならもみじまんじゅうです

図1：広島に関する QADB

図1を例に説明する。Anton らの手法により、広島というキーワードのクラスタ{Q1, Q2}及び、お好み焼きというキーワードのクラスタ{A2, A3}が出来る。また、Q2 と A2 のペアを見ると、“広島”と“お好み焼き”というキーワードには密接な関係があると判断できる。分類結果のクラスタとこれらのキーワードの関係から、Q1 に対して A2 の応答が得られる。このように Anton らの手法では、複数の応答候補を持つ。しかし、日本語においては適切な結果

が得られない場合がある。この問題について次章で述べる。

3. 提案手法

3.1. pLSAの適用によるクラスタリング

データベースの分類には pLSA を用いる。pLSA とは、確率的潜在意味解析 (Probabilistic Latent Semantic Analysis) のことで、基本的には LSA と同様に次元の圧縮を行うだけでなく、次元圧縮を確率的に行う手法である[3]。

潜在変数 $z \in Z$ を考えて、文書 d における単語 w の生起確率は以下のように表せる。

$$P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z)$$

また、潜在変数モデルにおける最尤推定のために、EM アルゴリズムにより以下のように定式化できる。まず、E ステップとして次式が定式化される。

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

次に M ステップとして以下が定式化される。

$$P(w|z) \propto \sum_{d \in D} n(d, w)P(z|d, w)$$

$$P(d|z) \propto \sum_{w \in W} n(d, w)P(z|d, w)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d, w)P(z|d, w)$$

ここで、 $n(d, w)$ は文書 d における単語 w の出現回数とする。

pLSA では E ステップと M ステップを反復させ、生起確率 $P(d, w)$ を最大化させるようなモデルが作成される。また、E ステップの右辺全体を β 乗するような温度パラメータ β ($0 < \beta \leq 1.0$) を与える。 $\beta = 1.0$ に近づけば近いほど、生成される確率モデルの確率分布は鋭いピークを持つようになり、逆にこの値を小さくすると、平滑化される。

本研究で pLSA を採用した理由を述べる。第一に、データベースの分類が自動的に行われるため、システムの利用者もしくは開発者から何らかのパラメータ付与等の作業が必要でないということが挙げられる。厳密には次元数や温度パラメータといった設定は必要であるが、クラスタリングの際のクラスタ数やその分布の決定のためには必須のパラメータである。しかし、これら以外ではパラメータ付与等の作業が必要ない。そのため、利用者もしくは開発者の主観が入ることなく、より客観的なデータベースの分類が行えることが期待される。また、pLSA は複数クラスタに属することを許容する分類結果 (ソフトクラスタリング) を返し、各クラスタへの所属確率を得ることができる。ゆえに pLSA を採用した。

所属確率についてより詳細に説明する。

今日の天気を教えて	0.00000	1.00000
今日の天気は	0.00856	0.99144
こんにちは	1.00000	0.00000
元気ですか	1.00000	0.00000
予定はありますか	0.99727	0.00273
運勢を教えて	0.00000	1.00000

図 2 : pLSA 学習結果の一例

図 2 を例にあげると、例えば、“今日の天気は” という文章は 1 つ目のクラスタに所属する確率が 0.00856、2 つ目のクラスタに所属する確率が 0.99144 という解釈をすることが可能である。クラスタリングする際のクラスタ数、すなわち次元数や、温度パラメータ β の値によっては、1 つ目のクラスタに所属する確率が 0.4、2 つ目のクラスタに所属する確率が 0.6 というような場合もある。このようにして、pLSA では複数クラスタに対する所属確率が与えられる。

3.2. データベース分類手法の提案

Anton らの手法は、英語等の主語省略を行わない言語では情報の欠落が起きないため有用であると考えられる。例えば、“Do you know favorite souvenir?” という質問に対し“The souvenir is *Okonomi-yaki*.” という応答を行う。しかし日本語では、“お好み焼き！”と単語のみで応答する場合が多い。このように、日本語では主語省略が頻繁に行われる。ゆえに、情報が欠落しデータベースの分類が難しい。そこで本研究では“質問と応答の両方を用いて分類”という手法を提案した。これにより、図 1 の Q3 と A4 が“食べ物”というキーワードを持つクラスタが得られ、Q3 の応答候補に A4 が追加される。更に{Q1, Q2} クラスタから名物とお土産が新たに結びつくため、質問と応答の両方でクラスタリングすることで A2 もしくは A3 と Q1 で、“お好み焼き”と“名物”の関係性を得る。このように、日本語では本研究の提案手法により更に応答候補を得ることができる。

また、この手法を採用した場合には、これまでのように QADB は一問一答型に最適化されたものではなく、「一つの質問に対して、常に最適な応答を返すのではなく、二番目、三番目の候補も許容するような対話システム」となることを想定している。pLSA を採用したことによって、3.1 節で述べたように、複数クラスタへの所属確率を表現することが可能であるため、複数の候補を許容し得る。

4. 実験

本研究では QADB の拡張に次のような手順で質問例の収集を行った。Twitter を利用して収集することで、人手で質問例を作成する手間を省いた。また、収集された質問例に対する応答の作成を人手で行った。しかし人手で作成する場合には、文章の性質 (も

しくは傾向) がほぼ同じ文章ばかりになるという問題がある。そこで客観性保持のため複数の人に対し、丁寧語やくだけた文体等、様々な文体を指定した上で、質問に対する応答を人手で作成してもらうこととした。これにより、質問例に限らず応答例もより客観性を持った大規模な QADB の構築が可能となる。一般的に、大規模な QADB を用いた場合にはシステムの応答精度が低下する場合があると知られている。そこで、大規模な QADB という条件下で応答精度が低下するという問題点を解決した上でシステムが適切に応答することが可能かどうかを評価する。

本研究の実験条件は次の通り。

- ・対象：“名物”に関する Tweet
- ・収集期間：2011 年 11 月 1 日～2011 年 11 月 7 日
- ・質問例総数：2182 文（ラベル数 852）
- ・応答例総数：2182 文（ラベル数 769）
- ・質問応答対総数：2182 文（ラベル数 1313）

本研究では“質問例のみを分類 (Q)”，“応答例のみを分類 (A)”，“質問と応答を分類 (QA)”の 3 通りの実験を行う。特に，“質問と応答のみを分類”した場合の結果を以下の図 3 に示す。図中の横軸の Conditions は pLSA でクラスタリングする際のクラス数、すなわち次元数を示す。また、縦軸の Precision は適切なクラスタに分類されているかどうかを示す精度を表す。従来のシステムでは、一つの適切な応答を返すことを目標としているため、質問に対して適切な応答一つを返せたかどうか精度の値となる。しかし本研究では、対話を続けることができるような、すなわち会話の連続という観点からは違和感が生じないような応答をしているかどうか精度の値となる。ゆえに、従来の精度とは意味合いが異なる。

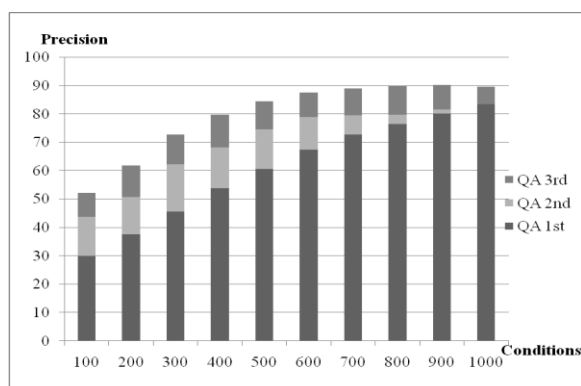


図 3：質問と応答の両方を用いて分類した結果

図 3 の結果から、精度が向上していることが分かる。特に上位 3 クラスに属することを許容すると更に向上する。しかし、pLSA の性質から、次元数を上げると各クラスが小さくなるため、この結果は妥当と言える。これだけでなく、更に詳細な分析を行う。

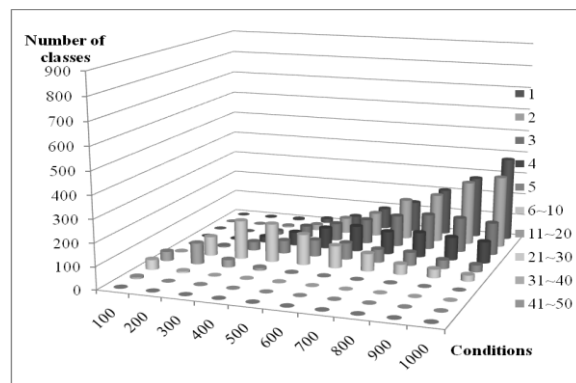


図 4：質問と応答の両方で分類した際のヒストグラム

図 4 は、各クラスを構成するテキスト数、すなわちクラス内に質問応答対がいくつあるかを示した。各クラスのテキスト数が 1 であるクラスは多い。しかし、テキスト数が 6~10 文であるクラスに着目すると、300 次元でピークを迎えている。図 3 を参照すると、精度は 70% 程度である。これは雑談など厳密さを要求されない対話システムの場合、この値は有効であると言える。

図 4 より、クラスタの分布が非常に広範囲になっていることが分かる。このうち、200~300 次元で最初のピークを迎えている、1 つのクラスが 6~10 文で構成されているようなクラスに注目する。このクラスの中で、pLSA によってどのようにクラスタの存在確率を与えられているか、その分布及びピークを調査する。

表 1：各テキストの存在確率分布

Probability	Conditions					
	100	200	300	400	500	600
0.9 ~ 1.0	1	11	16	11	12	4
0.8 ~ 0.9	2	27	59	46	42	22
0.7 ~ 0.8	2	73	122	81	83	34
0.6 ~ 0.7	5	95	191	181	128	90
0.5 ~ 0.6	4	123	275	188	171	99
0.4 ~ 0.5	4	129	214	236	157	110
0.3 ~ 0.4	6	102	227	223	148	131
0.2 ~ 0.3	7	132	286	245	222	188
0.1 ~ 0.2	3	161	334	262	211	151
0.0 ~ 0.1	3	38	50	39	20	11

表 1 のヒストグラムは、質問と応答の両方を用いてクラスタリングした結果のうち、6~10 文で構成されているようなクラスを対象として得られた結果である。例えば、あるテキストが各クラスに対して 0.6~0.7 の存在確率を返された数が 200 次元の際には 95 件ある、という見方をする。

この表より、どの次元においても、0.1~0.3 で最も多いテキスト数となっている。これは pLSA のパラメータがより厳密なクラスタリングされるように設定されていたためであると考えられる。この結果が

ら、存在確率の閾値を 0.2 として、質問に対して応答の選択候補数がどれだけあるかを示すために、Multi の指標を導入する。

$$Multi = \sum_{6 \leq \text{Text in Cluster} \leq 10} \frac{(\text{count}(\text{Text} \geq p = 0.2))}{8}$$

この計算の例を示す。例えば、あるクラスタのテキスト数が 8 文ある場合には分母は 8 となる。それに対してあるテキストの各クラスタの存在確率 0.2 を超える値を示したのが各テキストに 1 つずつしか見つからなかった場合、分子は 8 となり、Multi=1.0 となる。この場合、そのクラスタは厳密に分類されていると考えられ、ある質問に対するクラスタ内での応答の選択候補は 1 つしかないと考えられる。また、Multi=1.4 など、より大きい値が得られれば、応答の選択候補が増えたことを示すため、クラスタリングの有効性を示す指標になるものと考えられる。

これを、対象の全クラスタ (6~10 文で構成されているクラスタ) に対して算出し、そのうち最大値を以下に示す。

表 2：最大選択候補数

	Conditions					
	100	200	300	400	500	600
Question	1.50	2.50	2.29	2.57	2.60	2.57
Answer	2.13	2.00	1.89	2.13	2.20	2.11
Q and A	2.22	2.30	2.67	3.00	2.90	2.43

表 2 より、質問と応答の両方を用いた場合では、全てにおいて Multi=2.0 を超えている。次元数が 400 の際に Multi=3.0 となっている。これらから提案手法では、複数の応答候補を持つことができるため、より柔軟なシステムを開発することができるのではないかといえる。

また、質問のみ、応答のみでクラスタリングを行った際には、質問と応答を合わせた場合よりも低い値が得られるため、質問と応答を合わせて pLSA よりクラスタリングすることで、質問のみ、応答のみでクラスタリングする場合に比べてより柔軟な対話システムへ繋げることが可能であることが示される。ゆえに、本研究の提案手法の有効性を示したものである。

補足として、ラベルに関して解説を行う。本実験で用いた 2000 文以上の質問応答対には、1313 種のラベルを人手で与えた。このラベル数は非常に多いといえるが、理由として 2 点挙げられる。

まず一つ目は、Twitter から抽出する際の手続き上の問題が挙げられる。類似語に限定した上でそれらを含む Tweet を収集していることが原因である。例えば、お土産、名物といったキーワードは多く収集できるが、広島などといった地名に関しては収集対象として特に制限を設けていない。そのため、広島

に限らず様々な地名を含む Tweet も同時に収集された。そのため、付与される地名ラベルが増加する。

二つ目の理由としては、日本語の文章の特性が挙げられる。前述の通り日本語では英語と異なり主語が省略されることが多い。その結果、文章中のキーワードのいくつかが省略されている場合がある。例えば、“お好み焼き”という単語だけの応答文があった場合には、既存のラベルを与えることが出来ず、“お好み焼き”という新たなラベルを与えることになる。その結果、よりラベル数が増加する。

これらの理由から、各質問応答対に付与するラベルが増加することとなる。このラベルの付与の仕方は pLSA によるクラスタリングと類似する点があり、ラベルの増加によって 1 つの文で構成されるクラスタが増加する可能性がある。今後、このようなクラスタの調査も必要であると考えられる。

5. おわりに

本研究では質問と応答の両方を用いて QADB を分類することにより、適切な質問例を取得するための実験を行った。上位 3 クラスタを許容するように精度を算出した場合には 90% という高い値となった。精度は pLSA の特性上、次元数を上げればその向上は見込めるため、更に分類結果を緻密に分析した。新たに Multi の指標を導入することにより、質問に対する応答の選択候補数が増えたことが明らかになった。これにより、提案手法を用いることで柔軟な対話システムを構築することが可能であると言える。

今後は各単語に何らかの重みを付けることを検討する。たとえば 名詞や動詞といった品詞によって単語の重みを大きくすることが挙げられる。また、与えられた文章のラベル付けを利用した分類を考慮することも、より柔軟な音声対話システムの構築には必要であると考えられる。また、実際に分類された結果を用いて、実環境の音声対話システム上でその動作を評価する必要がある。

謝辞

研究の一部は、平成 23 年度 広島市立大学特定研究費 (一般研究) の補助を得ている。関係各位に感謝申し上げる。

参考文献

- [1] S. Takeuchi, T. Cincarek, H. Kawanami, H. Saruwatari, and K. Shikano, "Construction and Optimization of a Question and Answer Database for a Real-environment Speech-oriented Guidance System," in *Oriental COCOSA*, pp.149-154, 2007.
- [2] L. Anton, and T. David, "Practical language processing for virtual humans," in *Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*, pp.1740-1747, 2010.
- [3] Thomas Hofmann, "Probabilistic Latent Semantic Analysis," *Uncertainty in Artificial Intelligence*, 1999.