DBSJ Japanese Journal, Vol.13-J, No.1
October 2014

単語の出現度合いを考慮した質問文マルチタスク分類

Multi-class Question Classification Based on Word Occurrence Rates

徳永 陽子♡　数原 良彦♢
戸田 浩之♠　鷲崎 誠司♣

Yoko TOKUNAGA　Yoshihiko SUHARA
Hiroyuki TODA　Seiji SUSAKI

本論文の概要は以下のとおりである．音声によって入力された質問文に対して回答を返す検索サービスでは，カテゴリごとに用意されたデータベースを用いた検索方式が取り入れられている．検索の際には，予め質問文がどのカテゴリに関するものなのかを判定することで，検索対象とするデータベースを絞ることが出来，回答精度を向上することができる．本研究では，まず，約 **45,000** 文の質問文の分析を行い，単一のカテゴリのみに出現する頻度が高い単語を含む質問文とそうでない質問文が含まれていることがわかった．そこで，後者のような質問文についてカテゴリを判定する課題を，教師あり学習を用いたマルチクラス分類問題として解く．質問文は短文である場合が多く，単語の出現を特徴とする方法では分類性能に限界がある．そこで本稿では，全クラスにおける単語の出現度合いを数値化した拡張特徴表現を提案し，これを用いることで分類性能を向上し，高精度なカテゴリ判定を実現する．

Search services such as those performed by an intelligent personal assistant provide search results based on user input (questions) by relating them in synthesized speech form and/or by displaying them on a screen. Search services of this type select an appropriate database in advance to generate better responses. When the system receives a question, it selects a database by judging the category under which the question falls. Through an analysis of 45,000 questions, we find that questions can be divided into two types. One contains informative words that fall under a certain category, and the other does not contain any such words. In this paper, we tackle the category classification problem of the latter type of questions. We formalize the problem as a multi-class classification with supervised learning. Because questions usually consist of only a few words, the shortage of linguistic features results in misclassification if merely a machine learning technique is used. To avoid this problem, we propose a novel feature extraction method that reflects the word occurrence in each category. Experiments show that the method can improve classification accuracy.

1. はじめに

近年，スマートフォンやタブレット端末の普及とともに，ユーザが知りたい内容について音声で入力して情報を得るような音声入力による検索サービスが普及している．ユーザが音声で入力した質問に対して回答を提示するようなサービスを音声対話エージェントと呼ぶ．音声対話エージェントは，従来のキー入力を敬遠していた情報リテラシの低い高齢者や，運転中などキー入力が不可能な状況で情報を得たいユーザに有効であり，今後益々利用者が増えると推測される．本研究では音声対話エージェントの品質向上に取り組む．音声対話エージェントのサービス例としては，NTT ドコモのしゃべってコンシェル[1]や Siri[2]などが挙げられる．

音声対話エージェントでは，入力された音声データをテキストデータに変換し，それに基づいてクエリを生成して検索エンジンに問い合わせる．次に，検索エンジンから得られた結果から回答として出力する文等を生成し，画面表示または音声に変換してユーザに提示する．入力された音声データをテキストデータに変換したものを質問文と呼び，ユーザに提示するために検索結果から得られた回答に基づきシステムが生成した文を回答文と呼ぶ．

音声対話エージェントにおいて回答を得るための検索方法としては，予め用意されたデータベースから探す方法が広く取り入れられている．この検索方法は，大きく 3 段階に分かれている．1 つ目のステップでは，質問文について問い合わせるデータベースを判定する．次に，質問文からキーワードを抽出し，データベースに問い合わせる検索式を生成する．最後に，生成した検索式による検索結果から，該当するレコードの中から回答となる部分を抜粋し回答文を生成する．本研究では 1 つ目のステップのデータベース検索の部分に注目する．

音声対話エージェントでは，施設検索，レシピ検索，画像検索など，様々なカテゴリに関する質問文の入力を全て 1 つのインタフェースから受け付けている．そのため，音声対話エージェントはユーザの入力に応じて，回答文を生成する情報が格納されたデータベースを選択する必要がある．そこで，質問文が入力された際にまず質問文のカテゴリを判定し，そのカテゴリのデータベースを選ぶことで回答精度の向上につながる．我々は，質問文のカテゴリ判定の精度向上について検討する．

本研究では，最初に，音声対話エージェントを想定して生成した質問文約 45,000 件について傾向を分析し，カテゴリに分ける特徴となる要素が質問文にどれほど含まれているかを検証した．その結果，多くの質問文については一定以上の確率で単一カテゴリのみに出現する単語が含まれており，容易にカテゴリを判定できることがわかった．一方，一部の質問文についてはいくつか特徴を持つ単語が出現しているものの，質問文全体では特徴が表れにくく分類が難しいことがわかった．

そこで，一定以上の確率で単一カテゴリのみに出現する単語を含まない質問文について，分類候補となるカテゴリをクラスと捉え，マルチクラス分類の問題として解く．この問題に対し，正解クラスのラベルを付与した学習データを使って教師あり学習で予測モデルを作り，未知の質問文のクラスを予測する機械学習のアプローチで取り組む．質問文を始めとする様々な文書分類におい

♡ 正会員　日本電信電話株式会社 NTT サービスエボリューション研究所
tokunaga.youko@lab.ntt.co.jp
♢ 正会員　日本電信電話株式会社 NTT サービスエボリューション研究所
suhara.yoshihiko@lab.ntt.co.jp
♠ 正会員　日本電信電話株式会社 NTT サービスエボリューション研究所
toda.hiroyuki@lab.ntt.co.jp
♣ 日本電信電話株式会社 NTT メディアインテリジェンス研究所
suzaki.seiji@lab.ntt.co.jp

[1] https://www.nttdocomo.co.jp/service/information/shabette_concier/
[2] http://www.apple.com/jp/ios/siri/

て，各質問文の単語の unigram を特徴とする bag-of-words を特徴として学習する手法が多く取り入れられてきた．しかし，このような質問文については，bag-of-words を特徴として生成した予測モデルでは分類誤りが発生する可能性がある．

本研究では，文中に含まれる各単語がどのような特徴を持つ単語であるかという観点から，各単語とそれぞれのクラスとの関連の強さの分布を用いた特徴抽出手法を提案する．また，bag-of-words を特徴とした予測モデルと提案手法を用いて抽出した特徴を用いた予測モデルを比較し，有効性を検証する，特に学習に用いる質問文データ数が少ない場合に，bag-of-words を特徴とする手法に比べて高精度に分類可能であることを示す．

本研究の貢献は以下の 2 点である．

- 質問文データを分析し，各カテゴリに出現する単語の頻度分布を分析することで，特定の単語を含む文は特定のカテゴリに属するというルールを予め作って分類できるものと，そうでないものが存在することがわかり，これらを判別する方法に関する知見を得た．
- 単語のカテゴリ別の出現頻度を考慮する特徴抽出手法を提案し，ルールを用いた判別が難しい質問文に対して有用であることを確認した．また，学習データが少ない場合でも高い精度で分類できることがわかった．

以下，本稿の構成を述べる．2 章では，関連研究と本研究の位置づけについて述べる．3 章では，質問文データの傾向を分析した結果について説明する．4 章では，質問文の特徴抽出手法の提案を行い，5 章では提案手法の評価実験の方法と結果について述べる．6 章では実験で得られた知見について述べ，7 章でこれらをまとめる．

2. 関連研究

これまでも，機械学習を用いたテキスト分類に関する研究が進められてきた．

Sebastiani ら [8] の調査によると，テキストに出現する単語に対し，出現頻度に基づいた重要度を付与したものを特徴として，機械学習を用いて分類することで，テキストのカテゴリを精度高く推定することができる．この中では，マルチクラス分類の場合にも機械学習を用いた分類が利用可能であると述べられている．

質問文は，一般の文書と比較するとテキストの長さが短い場合が多く，出現する単語も多くはないため，抽出できる特徴が少ないと考えられる．そこで，予測モデルを生成する学習の際に用いる特徴を工夫する必要がある．質問文のように短い文書の分類としては，Sriram ら [9] や Rao ら [7] がマイクロブログのカテゴリ分類手法を提案している．ここでは，ユーザが登録している居住地などの属性情報や，マーク付きのメンションやリツイートなど，ツイッター独自の特徴を用いている．しかし，音声入力による検索サービスの質問文はツイッターとは性質が異なり，このような情報が付加されていないため，適用不可能である．

また，質問文のようにユーザの検索要求を表したテキストのカテゴリ推定手法も多く提案されている．Li ら [5] や Cao ら [2], Zhou ら [10] は，一般の検索エンジンに入力されたクエリをカテゴリに分類するために，クリックログやクリックスルーログを用いて，ユーザが閲覧した文書の情報を利用した手法を提案している．また，Kang ら [4] はクエリのカテゴリ分類に文書の URL やリンク情報を用いた手法を提案している．検索結果に提示される文書やユーザがクリックした文書には文字数の多い文書も含まれており，クエリのカテゴリに関する情報が豊富に含まれている可能性があり，これらを学習に用いることで元のクエリのカテゴリを推定することができる．質問文でも同様に，質問文と回答文のペアを用いて，回答文に含まれる単語や，その単語に関連するウェブ上の文書などを学習に用いてカテゴリを推定できる可能性がある．しかし，分類の候補となる全カテゴリについて，このようなペアのデータを大量に用意するにはコストがかかる．よって，本研究では，回答文を用いることなく質問文のカテゴリを推定する手法を検討する．

Qu ら [6] は，QA のカテゴリ分類において単語の出現を特徴とした bag-of-words と n-gram を特徴とし，Naive Bayes，Maximum Entropy，Support Vector Machines(SVM) の 3 種類の分類器を用いた場合の比較を行っている．これによると，学習・推定それぞれに必要な時間，分類精度を合わせて総合的に考えると，SVM が最も実用的であると結論づけている．また，Aikawa ら [1] は，コミュニティー型の QA サイトの投稿における質問文のカテゴリ推定を行っている．QA の質問は自然文で検索要求が表されており，音声対話エージェントでの質問文にも類似したものが含まれていることから，類似した特徴を持っているものも多いと考えられる．一方，音声対話エージェントの場合，タブレットやスマートフォンを用いて質問をするため,「音楽を聞く」「写真を見せて」というような端末内にあるデータベースからの検索を支持するものも含まれている．しかし，QA サイトでは端末内のデータに関する操作を要求する質問はほとんどない．また，外出先での質問や，天気や放送中のテレビ番組情報に関する質問など，現在の時間と場所に依存した質問文が入力されることも多い．このような質問文には，時間を表す際に特定の日時を表現せず,「今」「これからの」といった表現が用いられたり，場所を表す際に「近所の」「このあたりの」というような現在地からの相対的な表現が用いられたりすることも多い．QA サイトにおける質問では，回答者から適切な情報を得るために，調べたい時間や場所の情報は具体的に説明されることが多い．このように，音声対話エージェントの質問文には QA の質問文と性質が異なるものも多く含まれている．そのため，QA サイトでの質問のカテゴリ分類手法は，そのままでは適応できないと考えられる．

3. 質問文の分析

本章では，音声対話エージェントの質問文を分析して得た知見を述べる．本研究で用いた質問文は，作成者 200 人が，音声入力による検索サービスを想定し，カテゴリ別に調べたいことを問い合わせる文を一人あたり 300 個作成したものである．そこから重複，誤字，カテゴリ誤り等を除いたものを質問文データと呼び，これを利用する．質問文は 1 文のみの自然文か単語の羅列とする．作成する質問文のカテゴリは，日常で検索する機会が多いと思われるものを 15 種類選んで用いた．質問文のカテゴリと作成数を表 1 に示す．

作成者には，質問文が考えやすいように，例えばレシピカテゴリについてはクックパッド[3]といったように，各カテゴリのデータベースに対応した商用の専門検索サービスの例を示した．また，カテゴリによっては，調べる際に指定できる検索条件のパターンが多い．例えばレシピカテゴリの場合は，料理名，材料，カロリー，時間など，ショッピングカテゴリの場合は，商品の名前，値段，商品ジャンル，購入目的などの様々な検索条件があり，この組み合わせも考えられるため，聞き方のバリエーションが多いと考えられる．そのため，作成する質問文の検索条件にばらつきが出るように，例えば「料理名を検索条件に含むレシピの質問文」,「材料を検索条件に含むレシピの質問文」というように必ず含むべき検索条件とカテゴリの組み合わせを設定し，その組み合わせ毎に一定件数以上の質問文を作成するよう指定した．

これにより，質問文の検索条件の種類が多いカテゴリは作成する質問文も増えるため，バリエーションが多いカテゴリについては質問文の数が他のカテゴリより多くなっていることから，一般の音声対話エージェントに入力される質問文のバリエーション数に近いデータであると考えられる．

作成した質問文が含む全単語数は平均 7.8 個（標準偏差 9.3 個），名詞・形容詞・動詞のみに限ると平均 4.2 個（標準偏差 2.7 個）であり，質問文のカテゴリの特徴を表すことが期待される単語の

[3] http://cookpad.com/

表 1: 質問文のカテゴリとデータ数

カテゴリ	文書数	カテゴリ特徴語数
画像	1103	275
動画	1054	245
音楽	2086	1198
天気	731	58
レシピ	6158	1297
求人	1915	351
習い事・資格	2122	279
ショッピング	6335	2427
テレビ	6938	1561
イベント	5078	1156
施設・店舗検索	1693	270
交通路線・ルート検索	2955	486
銀行・ATM	1078	102
コンビニ	1086	148
レストラン	4391	453
合計	44723	10306

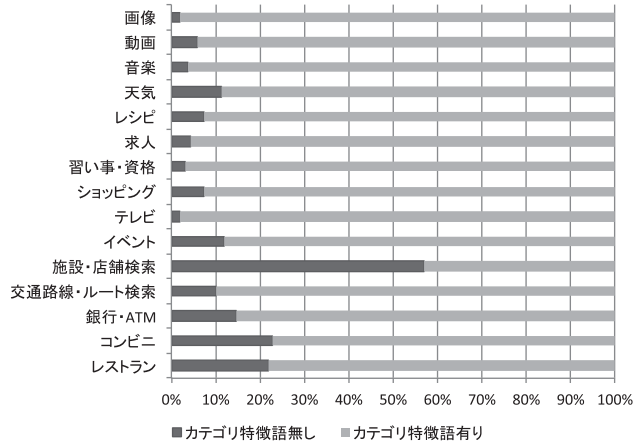

図 1: カテゴリ特徴語を含む質問文の割合.

数が少ないことがわかる．なお，形態素解析には JTAG を用いた [3]．このため，bag-of-words では値を持つ特徴が少なく，質問文が持つ特徴の数が少なくなると考えられる．

次に，カテゴリ別に質問文を分析したところ，単一カテゴリでの出現回数の割合が大きい単語を含む質問文があることがわかった．例えば，“レシピ” という単語を含む質問文の 99%はレシピカテゴリの質問文である．また，“画像” という単語を含む質問文の 99%，“写真” という単語を含む質問文の 95%は画像カテゴリの質問文である．このように，一定以上の確率で単一カテゴリのみに出現する単語をカテゴリ特徴語と呼ぶ．ここでは，単語 w が，全カテゴリの中で 1 つのカテゴリ C に出現する確率が 0.9 以上の場合，すなわち $P(C|w) \geq 0.9$ である場合に単語 w はカテゴリ C のカテゴリ特徴語とする．各カテゴリの質問文に含まれているカテゴリ特徴語の種類数を表 1 の右列に示す．質問文データにおいて，カテゴリ特徴語を含む質問文の割合をカテゴリ別に調べた結果を図 1 に示す．画像，音楽，テレビ，習い事・資格のカテゴリではカテゴリ特徴語を含む質問文が多い一方，施設・店舗検索では半分以上が含まない質問文であることがわかる．このようなカテゴリ特徴語を含む質問文の場合，予め「単語 w が出現する質問文はカテゴリ C の質問文と判断する」という分類ルールを記述することで，高い精度で正しいカテゴリに分類することができる．

次に，カテゴリ特徴語を含まない質問文について，文中に含まれる各単語の特徴を調べた．例えば，「海が見える鎌倉のイタリアンは」という質問文について考える．この質問文に含まれている単語のうち，“海”，“見る”，“鎌倉”，“イタリアン” の 4 つはカテゴリ特徴語ではないものの，ある程度カテゴリの特徴を表す単語であると考えられる．この 4 つの単語の全質問文データにおけるカテゴリ別の出現割合を表したものを図 2 に示す．“イタリアン” はレストランカテゴリで出てくる割合が最も多く，全カテゴリ中ではレストランカテゴリと関わりが強い単語であると推測される．また，“海” もレストランカテゴリにおける出現割合が最も大きい．“見る” は 2 割程度がテレビカテゴリに出現しているが，“イタリアン”とレストランカテゴリほど強い関わりではない．“鎌倉” は，場所を表す単語であるため，場所と関わりがある施設・店舗検索やレストラン，天気，イベント，交通路線・ルート検索などのカテゴリで出現しているが，特定のカテゴリとの強い関わりはない．よって，レストランカテゴリと関わりの強い “イタリアン” と “海” が出現していることから，この単語の出現を特徴とすることで，質問文を正しくレストランカテゴリに分類することができると考えられる．このように，カテゴリ特徴語が含まれていない質問文でも，カテゴリごとに出現する割合が比較的大きい単語が含まれている場合もあり，この単語が出現するという特徴を用いることで，質問文自体の特徴を表すことができるものがある．

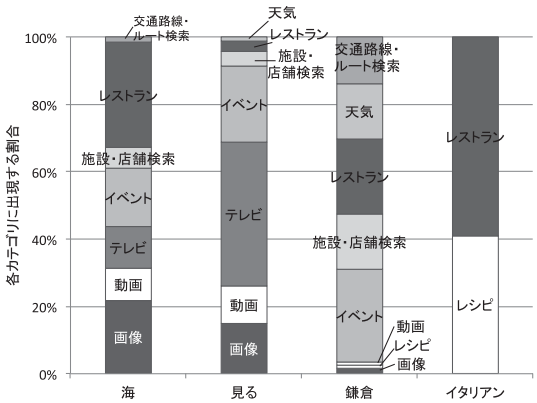

図 2: 単語のカテゴリ別出現割合.

しかし，単語とカテゴリの関わりの強さを実世界における関わりの強さに近似した値で表すためには，質問文とカテゴリのペアが大量に必要となる．質問文の分類候補となるカテゴリが増えるほど必要なデータ数は増え，データ作成に膨大なコストがかかる．例えば，“海” について，各カテゴリとの関わりの計算に用いる質問文数を変えて出現割合を求めた．質問文数と出現割合の関係を図 3 に示す．質問文数が 2,240 個のとき，“海” は画像カテゴリと動画カテゴリでのみ出現しており，レストランカテゴリとの関わりはない．また，4,480 個に増やしても，レストランカテゴリよりも画像カテゴリで多く出現しており，画像カテゴリとの関わりの方が強いことがわかる．よって，カテゴリのラベルが付いた質問文数が少ない場合には，レストランカテゴリとの関わりの強さを表す特徴が少なくなり，質問文を誤ったカテゴリに分類してしまう可能性がある．

このことから，少量の質問文データしか用いることができない場合，本来の単語とカテゴリの関わりの強さを正確に表すことができない可能性がある．このような場合，文中に出現する単語と関わりの強い特定のカテゴリのみを考慮すると判定を誤ってしまう可能性が考えられる．

4. 単語の出現度合いを考慮した質問文分類

本章では，単語の各カテゴリでの出現度合いを表す特徴を用いることで，単語の特徴を考慮した予測モデルの生成を目指す．単語の各カテゴリにおける出現度合いを表す特徴を単語特徴分布と呼ぶこととする．

本手法では，質問文中に出現したある単語について，その単語と各カテゴリとの単語特徴分布の値を全て用いて特徴ベクトルを構築する．従来のように bag-of-words を特徴とする場合，ある

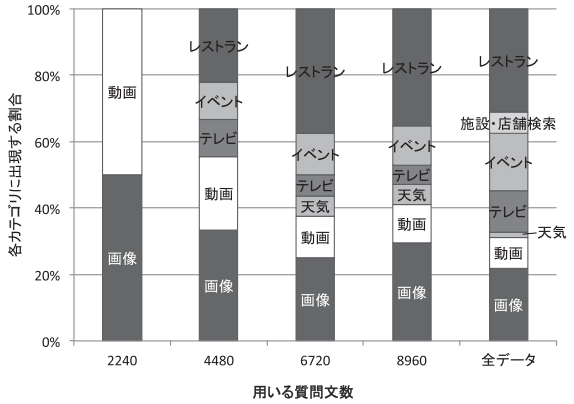

図 3: “海” の出現カテゴリ割合と用いるデータ数の関係.

単語 w が出現すると，対応する特徴の値を 1 とする．ここで，単語 w の i 番目のカテゴリ C_i における単語特徴分布を $r(w, C_i)$ と置く．ただし，カテゴリは全 $|C|$ 種類とし，$i = 1, ..., |C|$ とする．本手法では，単語 w の出現に対して，$r(w, C_1), ..., r(w, C_{|C|})$ の $|C|$ 種類の特徴を用いる．よって，質問文中に単語が t 種類出現した場合，特徴ベクトルは $(t \times |C|)$ 個の要素が値を持つ．

本稿では，単語特徴分布として，以下の 3 つの方法の利用を検討し，実験で有効性の評価を行う．

$P(C|w)$:
単語 w が全学習データ中の n_w 個に出現し，そのうち k_{wC} 個がカテゴリ C の質問文である場合，$P(C|w) = k_{wC}/n_w$．これは，単語 w が出現するときのカテゴリ C の条件付き確率を表す．

$P(w|C)$:
学習データにおいて，カテゴリ C の質問文が m_C 個あり，そのうち k_{wC} 個に単語 w が出現する場合，$P(w|C) = k_{wC}/m_C$．これは，カテゴリ C である質問文の単語 w の条件付き確率を表す．

$qTFIDF(w, C)$:
学習データにおいて，カテゴリ C の質問文のうち k_{wC} 個に単語 w 出現する場合，$qTF = k_{wC}$ とする．また，学習データ数 N 個のうち w が出現する質問文数が n_w 個ある場合，$qDF = \log(n_w/N)$ とする．このとき，$qTFIDF = qTF/qDF$ と表す．これは，カテゴリ C の質問文のうち，w を含む質問文数を TF，全質問データにおいて w を含む質問文数を DF とした時の $TFIDF$ の値と同じ性質の値であり，そのカテゴリに特化して出現する単語では値が高くなる．

5. 評価実験

本章では，提案手法の評価手法とその結果について述べる．まず，前章で挙げた単語特徴分布の 3 つの計算方法のうち，最適な方法を検証する（実験 1）．次に，bag-of-words と単語特徴分布を特徴とした場合の予測モデルによる分類精度を比較する（実験 2）．

5.1 実験 1: 単語特徴分布に用いる値による分類精度の比較

4 章で述べた 3 種類の単語特徴分布を特徴とした予測モデルを生成し，分類精度を比較する．まず，全質問文データを半分に分け，その片方についてカテゴリ特徴語を含む質問文は除外し，テストデータとした．これは，3 章で述べたように，カテゴリ特徴語を用いた分類のルールを定めることでカテゴリを推定可能な質問文は対象外とするためである．カテゴリ特徴語を含む質問文を除外すると，テストデータは 2,359 個となった．データの内訳を表 2 に示す．

表 2: テストデータの正解カテゴリ.

カテゴリ	文書数
画像	11
動画	31
音楽	39
天気	40
レシピ	243
求人	39
習い事・資格	31
ショッピング	237
テレビ	61
イベント	291
施設・店舗検索	473
交通路線・ルート検索	163
銀行・ATM	77
コンビニ	120
レストラン	503
合計	2359

表 3: 単語特徴分布として用いる値による分類精度の違い.

	$P(C\|w)$	$P(w\|C)$	$qTFIDF$
平均適合率	0.760	0.699	0.703
平均再現率	0.725	0.589	0.586
平均 F1 値	0.733	0.608	0.607

分類器には SVM を用いた．マルチクラス分類にはペアワイズ法を用いた．実装には LIBSVM[4]を利用し，実験では RBF カーネルを用いた．トレードオフパラメータ C の選択には，学習データにおいて 5 分割交差検定を行い，正解率が最大のものを選択した．

質問文データが少ない場合を想定し，モデルの生成に用いる学習データは残り半分のデータのうち 20%の 4,459 件とした．学習データは，全質問文データと同じ割合で各カテゴリの質問文を含むようにした．3 種類それぞれの方法で，学習データを用いて単語特徴分布を計算し，それを特徴とした予測モデルを用いてテストデータのカテゴリ分類を行った．評価指標は，カテゴリ別に適合率，再現率，F1 値を計算し，マクロ平均を取ったものである．

結果を表 3 に示す．$P(C|w)$ を単語特徴分布の値としたとき，適合率・再現率ともに，最も高い数値を示した．このことから，例に挙げた 3 種類の単語特徴分布の値のうち，$P(C|w)$ が最も適した値であることを確認した．

5.2 実験 2: bag-of-words との比較

単語の unigram の出現を表した bag-of-words を特徴として学習した予測モデルをベースラインとし，提案手法である単語特徴分布を特徴として学習した予測モデルの分類精度を比較した．提案手法では，前節で最も高い分類精度を示した $P(C|w)$ を単語特徴分布として利用する．また，ベースラインでは単語の出現をバイナリで表したものを特徴とする．

質問文データが少ない状況を想定し，学習データの数を減らした場合の分類精度の傾向を調べた．学習データ全体の 1%の数の 223 個から 1%ずつ質問文を増やす毎に予測モデルを生成し，テストデータの分類を行った．事前に単語特徴分布を求める際には，学習データに含まれる質問文のみを利用することとし，学習データを増やす毎に単語特徴分布を計算して値を更新して用いる．テストデータは，前節と同様に，全質問文データの半分からカテゴリ特徴語を除いた 2,359 個とし，これをベースラインと提案手法による予測モデルで分類した．実験条件は 5.1 と同じである．評価指標には，カテゴリ別の F1 値のマクロ平均を用いた．

全カテゴリでの F1 値の平均値の推移を図 4 に示す．学習データ数が 223 個から 5,000 個前後の間で，ベースラインを上回って

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

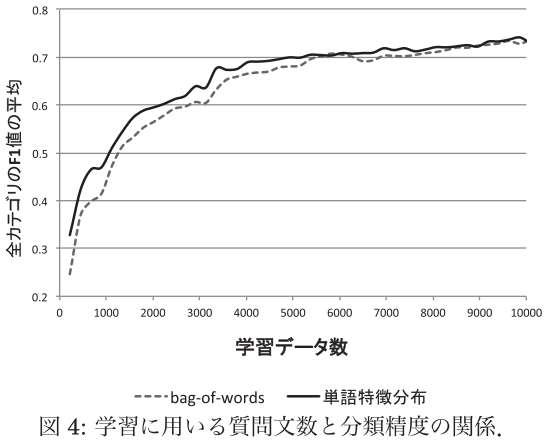

図 4: 学習に用いる質問文数と分類精度の関係.

いる．その後は，学習データ数を増やしても F1 値はあまり変化せず，ベースラインと提案手法の間でもほとんど差がないように見える．学習データを 10,000 個以上用いた場合，ベースラインによる予測モデルでのみ正しく分類できた質問文は 77 個あったのに対し，提案手法による予測モデルのみで正しく分類できた質問文は 209 個あった．このことから，単語特徴分布を用いた場合，学習データが最も少ない場合にベースラインより高い精度で分類出来たことがわかった．

次に，カテゴリ毎に比較し，特にベースラインと提案手法の間で違いが見られた 6 つのカテゴリについて図 5 に示す．天気カテゴリでは，学習データ数が 1,000 個前後を除き，それ以降から 6,000 個を超えるまではベースラインを上回る結果となった．また，6,000 個を超えたあたりから，ベースラインでは学習データ数が増えるたびに F1 値が上下に振れているが，提案手法では安定して高い値を示している．求人カテゴリでは，学習データ数が最も少ない場合，ベースラインでは全てのテストデータで分類を誤り，適合率・再現率ともに 0 となったが，提案手法では F1 値が 0.34 となり，精度は高くはないものの正しいカテゴリに分類できた質問文があることがわかる．ベースラインで F1 値が 0.34 を超えるためには，1,344 個以上の学習データが必要であった．テレビカテゴリでは，ベースラインでは学習データ数を最少の 224 個から 3,808 個まで増やす間に F1 値が 0.3 から 0.6 まで線形に上昇しているが，提案手法では 2,016 個の時にすでに 0.6 を超えており，学習データ数の増加とともに精度が急激に向上している．下の段の施設・店舗検索カテゴリ，交通路線・ルート検索カテゴリ，レストランカテゴリでは，どれも曲線のカーブはベースラインと提案手法で似ており，学習データを増やすと同じ値に収束しているように見える．学習データ数が最も少ない時に着目すると，提案手法の F1 値がベースラインを上回っており，施設・店舗検索カテゴリでは 0.18，交通路線・ルート検索カテゴリでは 0.35，レストランカテゴリでは 0.19 の差がある．

これらの結果から，学習データが少ない場合，単語特徴分布を特徴とした予測モデルを生成することで，単語の出現のみを特徴とした場合よりも高い精度で分類できる可能性があることがわかった．

5. 考察

個々の質問文において，正しいカテゴリに分類するために必要な学習データ数を比較した．3 章で例に挙げた，「海が見える鎌倉のイタリアンは」という質問文を正しくレストランカテゴリに分類するためには，ベースラインでは 2,912 個以上の学習データが必要だったが，提案手法では 448 個のみで達成することができた．カテゴリを推定した質問文のうち 369 個の質問文については，提案手法を用いることで，ベースラインを用いて正しく分類するのに必要な学習データ数の半分以下で分類することができた．このことから，単語特徴分布を用いることで，単語が最も多く出現するカテゴリとの関わりだけでなく，各カテゴリとどのくらい関わりがある単語であるかを十分に考慮することができ，学習データ数が少なくても正しく分類することができたと考えられる．

一方，画像カテゴリにおいては，ベースライン，提案手法ともに学習データの数を増やしても正しいカテゴリに分類することができなかった．他のカテゴリでも，学習データを最大限用いて予測モデルを生成しても誤ったカテゴリに分類されてしまう質問文があった．このような質問文について分析した結果，カテゴリを表していると考えられる特徴的な単語を含んでいても，その単語が学習データ中に一度も出現していないため，分類の際にその単語の出現を特徴として反映することができないことが原因の多くを占めているとわかった．例えば，「泳ぐ子供　イラスト」という質問文では，“イラスト” が画像カテゴリらしさを表していると捉えることができる．しかし，“イラスト” を含む質問文は学習データに存在しないため，“泳ぐ” と “子供” の 2 つの単語からカテゴリを推定しなければならない．このような特徴分布の偏りが，適切な分類を行う分類器の生成を妨げている．また，全質問文データにおいて，“泳ぐ”を含む質問文は，画像カテゴリ 2 個と動画カテゴリに 1 個の合計 3 個しか出現していない．このデータから単語特徴分布を求めると，“泳ぐ” は画像カテゴリで高い値となるが，実世界においては画像カテゴリとそれほど関わりの強い単語であるとは考えにくい．

これらのことから，学習データ中に含まれない未知の単語や，出現する質問文が極端に少ない単語については，実世界に近似した単語特徴分布を計算することができず．分類に適した特徴を付与することができない．また，質問文データをどれだけ膨大に用意しても，実世界に存在する全ての単語を網羅し，その特徴分布を表すデータを作成することは困難である．そのため，入力された質問文に未知の単語や出現頻度が低い単語が含まれているような場合，質問文データだけでなく，その単語を含むウェブ文書や検索クエリログなど，外部のデータを分析することで，実世界に近似した単語特徴分布を求められる可能性があると考えている．

6. まとめ

本研究では，音声対話エージェントに入力される質問文のカテゴリを推定する方法について述べた．まず，質問文データを分析した結果，特定のカテゴリでの出現頻度が高い単語を含む質問文とそうでない質問文に分かれることがわかった．後者のような質問文のカテゴリ分類は，教師あり学習によるマルチクラス分類の問題として捉え，質問文に含まれる各単語の全カテゴリそれぞれにおける単語の出現度合いに基づいた単語特徴分布を抽出し，これを特徴として学習に用いる方法を提案した．評価実験を通して，特に学習データが少ない場合，単語特徴分布を特徴とした予測モデルは単語の出現のみを特徴とした場合よりも高い精度で分類できることがわかった．今後は，学習データに出現しない未知の単語や，出現頻度が少ない単語の特徴を外部データを用いて補うことで，より分類精度の高い予測モデルの生成を目指す．

[文献]

[1] Naoyoshi Aikawa, Tetsuya Sakai, and Hayato Yamana. Community QA question classification: Is the asker looking for subjective answers or not? *IPSJ Online Transactions*, Vol. 4, pp. 160–168, 2011.

[2] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. Context-aware query classification. In Proc. SIGIR ’09, pp. 3–10, 2009.

[3] Takeshi Fuchi and Shinichiro Takagi. Japanese morphological analyzer using word co-occurrence: JTAG. In Proc. COLING-ACL ’98, pp. 409–413, 1998.

[4] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In Proc. SIGIR ’03, pp.

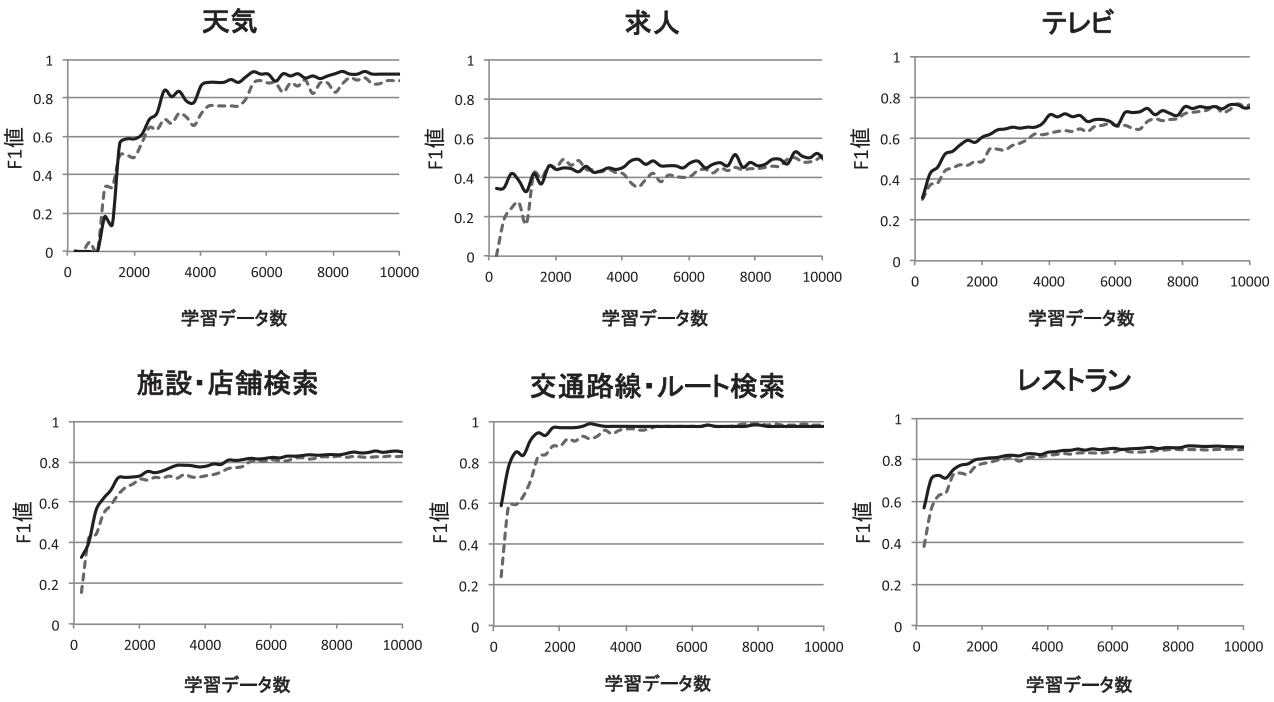

図 5: カテゴリ別の分類精度と学習データの質問文数の関係.

64–71, 2003.

[5] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In Proc. SIGIR '08, pp. 339–346, 2008.

[6] Bo Qu, Gao Cong, Cuiping Li, Aixin Sun, and Hong Chen. An evaluation of classification models for question topic categorization. *JASIST*, Vol. 63, No. 5, pp. 889–903, 2012.

[7] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In Proc. SMUC '10, pp. 37–44, 2010.

[8] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, Vol. 34, pp. 1–47, 2002.

[9] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in Twitter to improve information filtering. In Proc. SIGIR '10, pp. 841–842, 2010.

[10] Ke Zhou, Ronan Cummins, Martin Halvey, Mounia Lalmas, and Joemon M. Jose. Assessing and predicting vertical intent for web queries. In Proc. ECIR'12, pp. 499–502, 2012.

徳永 陽子 **Yoko TOKUNAGA**
2008 年 同志社大学工学部知識工学科卒業. 2010 年 京都大学大学院情報学研究科社会情報学専攻修士課程修了. 同年, 日本電信電話株式会社入社. 現在, NTT サービスエボリューション研究所に所属. 主として情報検索に関する研究開発に従事. 情報処理学会会員.

数原 良彦 **Yoshihiko SUHARA**
2006 年 慶應義塾大学理工学部管理工学科卒業. 2008 年 同大学院理工学研究科開放環境科学専攻修士課程修了. 同年, 日本電信電話株式会社入社. 現在, NTT サービスエボリューション研究所に所属. 主として情報検索, 機械学習に関する研究開発に従事. 情報処理学会, 人工知能学会, 言語処理学会各会員.

戸田 浩之 **Hiroyuki TODA**
1997 年名古屋大学工学部材料プロセス工学科卒業. 1999 年同大大学院工学研究科材料プロセス工学専攻博士課程前期課程修了. 同年, 日本電信電話株式会社入社. 現在, NTT サービスエボリューション研究所に所属. 以来, 情報検索, データマイニングの研究開発に従事. 2007 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻博士後期課程修了. 博士 (工学). ACM, 情報処理学会各会員.

鷲崎 誠司 **Seiji SUSAKI**
1988 年名古屋大学理学部数学科卒業. 同年, 日本電信電話株式会社入社. 自然言語処理, 情報検索に関する研究開発に従事. 現在, NTT メディアインテリジェンス研究所, 主幹研究員. 情報処理学会会員.