

クエリログを利用した質問文カテゴリ分類

清武 寛¹ 徳永 陽子^{2,a)} 数原 良彦² 戸田 浩之² 鷲崎 誠司²

概要: 本論文では、音声入力による検索サービスにおける質問文のカテゴリ判定に取り組む。質問文の回答を生成する際、複数のカテゴリ別専門データベースの中から回答に関する情報を持つものを選択する必要がある。そこで質問文カテゴリをクラスラベルとし、教師あり学習を用いたマルチクラス分類の問題としてアプローチをした。この分類では質問文をカテゴリ判定する際、質問文中にその質問文のカテゴリを表す特徴的な単語があるにも関わらず学習データにその単語が出てこないためにカテゴリ判定に考慮できないという問題がある。以下、学習データの質問文群に含まれない単語を未知語と呼ぶ。本稿では、クエリログを用いて学習データに特徴を付与することで、未知語を減らす手法を提案する。具体的には、クエリログにおいて学習データに出現する単語と共起している未知語のうち、共起回数に基づいて求めた関連度が大きい単語を学習データに擬似的に特徴として加える。このような手法で質問文に新たに特徴を与え、単語の出現頻度ベースの手法などと比較した。

1. はじめに

近年、スマートフォンやタブレットを用いて音声入力によって検索するサービスが普及している。本稿では、ユーザがスマートフォンに質問を音声で入力し、答えを得る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がユーザの質問に答えるまでには、以下の3つのプロセスが存在する。

- (1) 入力された音声を認識してテキストの質問文に変換する。
- (2) 質問文のカテゴリを判定し、予め用意された複数のカテゴリ別専門データベースから検索するデータベースを選択する。
- (3) 検索条件を抽出して専門データベースを検索し、回答を生成する。

本研究では、(1)の質問文のカテゴリを判定するプロセスを対象とする。我々は、質問文のカテゴリを判定する問題を、質問文カテゴリをクラスラベルとし、教師あり学習を用いたマルチクラス分類の問題としてアプローチした。

音声対話エージェントに入力される質問文は、一般的な文書に比べて短文であるという特徴がある [1]。そのため、一般的なテキスト分類と同様に質問文中に出現した単語を特徴とした場合、分類に考慮できる特徴が少ない。このことから、出現単語を特徴として分類モデルを生成する場合、特に学習データ数が少ない状況において、分類精度が悪いという問題点がある。この問題の原因の1つに、カテゴリを推定したい質問文に含まれている単語が、学習データには一度も出現していないために、その単語の特徴を考慮できないことがある。例えば「ペペロンチーノの作り方」という質問文を分類する際、学習データ内に「ペペロンチーノ」という単語が存在しなかった場合を考える。学習データには「シュシュの作り方」「ホームページの作り方」など、「作り方」という語を含むレシピカテゴリ以外の質問文があるとする。このような場合、「ペペロンチーノ」というレシピカテゴリらしい単語を含むにも関わらず、学習に考慮できないため、「ペペロンチーノの作り方」という質問文を他のカテゴリに誤って分類してしまう可能性がある。

このように、一般的な文書分類と同様に、質問文中にその質問文のカテゴリを表す特徴的な単語があるにも関わらず、学習データにその単語が出現していない場合、その単語をカテゴリ分類に考慮できないため、分類を誤る可能性がある。本稿では、学習データに出現する単語を既知語、出現しない単語を未知語と呼び、未知語が原因となって分類精度が低下する問題を**未知語問題**と呼ぶ。[1]では、この問題への解決を今後の課題としており、本稿ではこれに取り組む。

¹ 九州大学大学院数理学府

Graduate School of Mathematics, Kyushu University

² 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
NTT Service Evolution Laboratories, NTT Corporation

a) tokunaga.youko@lab.ntt.co.jp

*¹ https://www.nttdocomo.co.jp/service/information/shabette_concier

*² <http://www.apple.com/jp/ios/siri/>

本研究では、まず音声対話エージェントを想定して作成した 56533 個の質問文データを用いて、実際に未知語問題が起きているのかどうか検証を行った。また、未知語問題を解決するために、学習データに対して未知語を擬似的に出現させてデータを追加することで、生成したモデルの分類精度にどのような変化が生じるのかを検証した。その結果、追加する元となる学習データ、追加する未知語、追加方法によっては、分類精度を向上することができることがわかった。

次に、追加する未知語の候補を、一般のウェブ検索エンジンのクエリログにおける単語の共起関係を用いて抽出する手法を提案する。クエリは、ユーザの検索意図を表している点で質問文と類似している。また、少ない単語数で検索意図を表現しているクエリは、共起している単語同士の意味的関連が強いとされている [2]。よって、未知語の候補を選ぶ手がかりとする既知語として、各カテゴリに関連の強い単語を選ぶことで、そのカテゴリの質問文として出現する可能性が高い単語を候補に選ぶことができると考えた。これらの性質から、クエリにおける共起単語を用いて、学習データに未知語の特徴を補完する方法を検討した。

評価実験により、適切な追加方法ではない場合、未知語を追加していない学習データを用いた場合よりも分類精度が低下してしまう可能性があることがわかった。

本研究の貢献は以下の 2 点である。

- 音声対話エージェントで入力される質問文のカテゴリを推定する際、学習データに出現しない未知語による分類精度低下が起きていることを確認した。
- 外部データを用いて学習データに出現しない単語の特徴を追加することによる、分類モデルの精度の変化に関する知見を得た。

本稿の構成は以下のとおりである。まず、第 2 章で関連研究について述べる。3 章では、学習データに未知語を擬似的に出現させたときの分類精度について検証する。そして 4 章で提案手法について述べ、5 章で実験結果、評価を行う。6 章で評価に関する考察を行い 7 章で本論文をまとめる。

2. 関連研究

これまで、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 [3] 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可能であると述べられている。

音声対話エージェントの質問文は、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、出現する単語が

少ないため、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 [4] や Rao ら [5] 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 [6] や Cao ら [7]、Zhou ら [8] 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 [9] 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。

QA と質問文は性質が似ているが、特徴の工夫が必要。Qu ら [10] 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 [11] 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。しかし、これらは未知語問題について取り組んでおらず、これを解決することはできないと考えられる。本研究は、音声対話エージェントの質問文のカテゴリ分類を対象としており、マイクロブログやクエリ、QA サイトの質問文などとは異なるタスクである。我々は、[1] にて、質問文中に含まれている単語の分類候補となる全カテゴリにおける出現頻度に基づく特徴分布を用いることに

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	文書数
画像	1193	動画	1171
音楽	1143	天気	800
レシピ	6476	求人	2379
習い事・資格	2389	ショッピング	9963
テレビ	7645	イベント	8027
施設・店舗	4405	交通路線・ルート	3997
銀行・ATM	1203	コンビニ	1191
レストラン	4551	合計	56533

よって、出現する単語数が短くても精度高く分類できることを示した。一方、そもそも学習データに一度も出現しない単語については、その単語とカテゴリの関わりを測ることが出来ず、分類に考慮することが不可能であり、これによって分類精度が下がる可能性があることが示唆された。

本研究では、この未知語問題について取り組むことで、音声対話エージェントの質問文のカテゴリ分類精度を向上することを目的とする。

3. 検証

本章では、音声対話エージェントの質問文において、未知語問題が起きていることを確認するための検証を行う。また、未知語問題を解決することによる、分類モデルの精度の変化について検証する。

3.1 未知語問題の検証

まず、検証に用いるデータについて述べる。作業員 200 名に対し、音声対話エージェントを想定した質問文を生成し、テキストで入力してもらった。質問文は 15 種類のカテゴリについてそれぞれ作成し、一人あたり合計 300 個作成してもらった。ここから重複やカテゴリ誤りを除いた 56533 個の質問文を有効なデータとする。各カテゴリの質問文数は表 1 のとおりである。

この全質問文データを 2 分割し、半分を学習データ、残りをカテゴリを推定するテストデータとした。また、学習データを 100 分割し、1% の 283 個から 1% ずつ追加した場合の、テストデータに含まれる未知語の数を調べた。結果を図 1 に示す。この結果から、学習データに質問文を 25000 個用いても、未知語は 5000 個以上も存在することがわかる。

3.2 学習データへの未知語の補完による精度検証

次に、未知語問題を解決することによる精度の変化について調べる。

まず、未知語の追加方法について述べる。未知語を追加する方法としては、2 種類考えられる。1 つ目は学習データの質問文数を増やす方法である。質問文の数が増えることで、出現する単語の種類数が増え、未知語の数を減らす

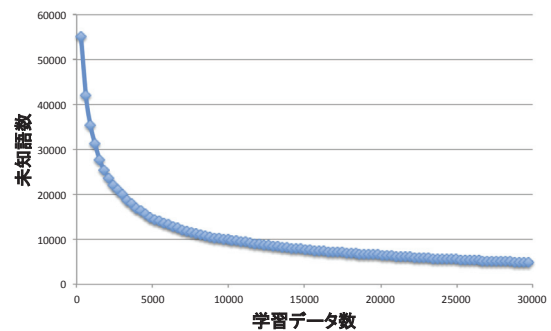


図 1 学習データ数と未知語数の関係

ことができる。2 つ目は、学習データに出現する単語の種類のみを増やす方法である。しかし、1 つ目の手法は正解カテゴリがラベル付けされた学習データを用意するのに非常にコストがかかる。よって我々は 2 つ目の手法について検討する。

学習データに現れる単語の種類を増やすためには、学習データに擬似的に未知語を出現させる必要がある。そこで、未知語を擬似的に出現させ、学習データに追加することによって、生成したモデルによる分類精度にどのような変化があるのかを検証する。

前述の質問文データのうち、学習データとした半分の 28266 個の質問文を用いて訓練誤差を比較した。まず、28266 個の質問文をランダムに 100 分割し、1% ずつ追加しながらそれぞれの段階で学習を行った。生成したモデルを用いて、学習データ 100% に対する学習曲線を測った。分類器には SVM を使い、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^{*3} を利用し、実験では線形カーネルを用いた。トレードオフパラメータ C の選択には、学習データを 5% 用いて 5 分割交差検定を行い、正解率が最大のものを選択した。このデータセットに対して、未知語を追加しない学習データを用いて学習した分類モデルをベースラインとして以下の二つの手法と分類精度を比較した。

ADD_1Word: 学習データを 1% 増やすごとに未知語の中からランダムに 1 語ずつ選び、その単語を一つの質問文としてとらえ学習データに加える。

ADD_1Sent: ADD_1Word と同様にランダムに未知語を一つ選び、その未知語が含まれている質問文をそのまま学習データに加える。

分類モデルの特徴には、各学習データを用いて、単語の出現回数を特徴とする Bag-of-Words を用いた。

3 つの手法を用いて学習した分類モデルによる全カテゴリの平均 F1 値の学習曲線を図 2 に示す。図 2 の場合少ない学習データにおいて、いずれの手法もベースラインより

*3 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

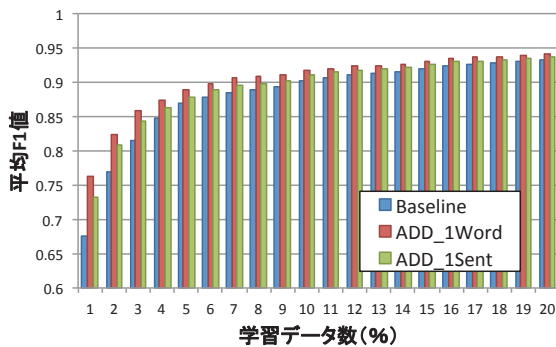


図 2 未知語を 1 語補充した場合の全カテゴリの F1 値のマイクロ平均

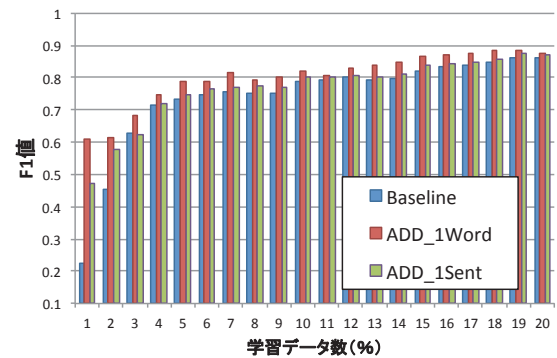


図 3 未知語を 1 語補充した場合の音楽カテゴリの F1 値

高い分類精度となった。特に単語を一文としてみてくわえる ADD_1Word が特に精度が高かった。また、図 3 は音楽カテゴリでの F1 値の変化を比較してみたものである。この場合もやはり学習データ数が少ない場合において顕著に分類精度に差が出ている。次に、追加する未知語を変えて同様の実験を行った。結果を図 4 に示す。この場合は二つの手法においてベースラインとの精度の差はほとんど見られなかった。このように、付与する単語を適切に選ばなければ精度に影響が与えられない可能性がある。

さらに、未知語を大量に追加してモデルを生成し、精度を比較した。ランダムに選んだ 100 個の未知語について、ADD_1Sent と同じ方法で質問文を学習データに追加し、モデルを生成した。これを ADD_100Sent とする。ADD_100Word とベースラインの分類精度を比較したものを図 4 に示す。学習データが少ない場合、Bag-of-Words に比べて精度が低くなっていることがわかる。このことから、未知語を多く付与することによって、多くの単語が学習に考慮されるが、分類精度が下がってしまう可能性があることがわかる。

これらのことから、未知語問題を解決することができれば、分類精度向上につながる事が確認できた。ただし、学習データに未知語を付与する場合、以下の 3 点に注意する必要がある。

- 追加する未知語の種類
- 追加する未知語の個数
- 未知語の追加方法

4. 提案手法

前章の実験から、未知語問題を解決することで、質問文のカテゴリ分類精度を向上させることができることがわかった。本章では、未知語問題の解決方法について述べる。

本研究では、既存の学習データに含まれている既知語の正解カテゴリのラベルを共起関係にある未知語に付与することで、学習データの補充を試みる。ここで、補充する未知語の候補を取得するために、一般のウェブ検索エンジン

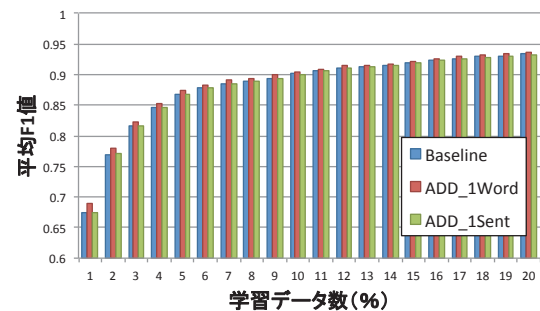


図 4 異なる未知語を 1 語補充した場合の全カテゴリの F1 値のマイクロ平均

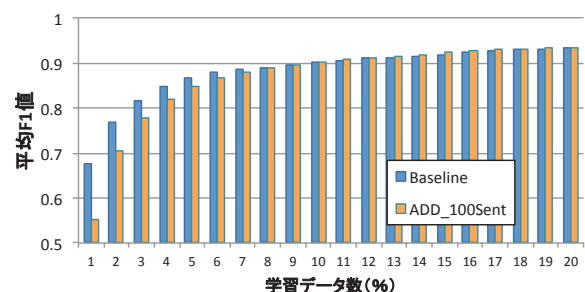


図 5 未知語を 100 語補充した場合の全カテゴリの F1 値のマイクロ平均

のクエリログを用いることとする。クエリに入力される単語は、ユーザの検索意図を表しており、質問文に含まれている単語と似た傾向があると考えられる。また、音声対話エージェントによる質問文には、自然文の形式のものと、単語の羅列によって検索意図が表されているものが混在しており。後者の単語の羅列は、一般のウェブ検索エンジンのクエリと類似している。さらに、田ら [2] は検索エンジンに入力されるクエリに含まれる単語の数が 2 個以上である場合、単語間の意味的関連が強いと述べている。このことから、クエリに既知語と未知語が共起している場合、この 2 単語の間の意味的関連が強いことから、そこで、この既知語が出現する質問文のカテゴリは、未知語が表すカテゴリとみなすことができると考えられる。

提案手法のプロセスは下記のとおりである。

- (1) 未知語の候補を取得する手がかりとする既知語を選択する。
- (2) 選択した既知語を手がかりとし、意味的関連の強い未知語をクエリログから抽出する。
- (3) 抽出した未知語を、手がかりとした既知語のカテゴリをラベル付けして学習データに追加する。

まず(1)では、カテゴリ C_i において、当該カテゴリの質問文に最も多く出現している単語の上位 n 個を手がかりの既知語とする。なぜならば、クエリ内共起する単語の間には修飾関係があり [2], 係り元となる単語は係り先となる単語よりも出現回数が多くなると考えられる。この傾向は質問文でも同じであると考え、出現頻度を基準に選択した。

次に、(2)について説明する。ここでは、 n 個のうちいずれかの既知語を含むクエリを抽出し、既知語とクエリ内共起している未知語を抽出する。この抽出された未知語が、カテゴリ C_i の学習データを補完する特徴の候補となる。この未知語の候補のうち、手がかりとなった既知語と最も共起回数が多いものから順に上位 k 個を選択し、この単語の特徴を学習データに補完する。

最後に、(3)で未知語を学習データに追加し、未知語の特徴を補完する。追加する方法は複数考えられるが、ここでは下記の2種類を示す。1つ目は、手がかりとなった既知語と選択した未知語がそれぞれ1回ずつ出現する質問文を仮定し、カテゴリ C_i を正解ラベルとして学習データに追加する方法である。2つ目は、選択した未知語のみが1回出現する質問文を仮定し、カテゴリ C_i を正解ラベルとして学習データに追加する方法である。これを全てのカテゴリについて行う。

このように、各カテゴリについて手がかりとなる既知語を決め、クエリログから未知語を抽出することで学習データを補完する。このデータを用いて、分類モデルを生成することで、補完する前には考慮できなかった未知語をカテゴリ判定に反映することができると考えられる。

5. 評価実験

提案手法の有効性を調べるための評価実験を行った。

5.1 データセット

実験には、3章の検証で用いた質問文データを用いた。全質問データ 56533 個のうち、半分を学習データ、残りをカテゴリを推定するテストデータとした。未知語の解決に用いる外部データとして、一般のウェブ検索エンジンにおける 2013 年 1 月～6 月の半年分のクエリログを用いた。

5.2 比較する手法

本実験では、分類器を固定し、特徴のみを変えて精度を比較する。比較する特徴抽出手法は、追加する未知語を決

表 2 特徴抽出手法

抽出手法名	既知語数 k	未知語数 u	追加する単語 N
Ext($k=1, u=1, N=2$)	1	1	既知語+未知語
Ext($k=1, u=10, N=2$)	1	10	既知語+未知語
Ext($k=10, u=1, N=2$)	10	1	既知語+未知語
Ext($k=10, u=10, N=2$)	10	10	既知語+未知語
Ext($k=1, u=1, N=2$)	1	1	未知語
Ext($k=1, u=10, N=2$)	1	10	未知語
Ext($k=10, u=1, N=2$)	10	1	未知語
Ext($k=10, u=10, N=2$)	10	10	未知語
Baseline	1	1	なし

める手がかりとなる既知語の数、学習データに補完する未知語の種類数、学習データへの未知語の追加方法の3点について異なる値を用いた4手法と、未知語を追加せず、既存の学習データのみを用いた手法の5つである。5種類のデータの詳細を表に示す。質問文の特徴としては、全データについて、単語の出現回数を特徴とする Bag-of-Words を用いた。分類器には、3章の検証実験と同じアルゴリズムを用いた。

手がかりとなる既知語としては、各カテゴリについて最も出現回数が多い名詞または動詞の上位 n 個の既知語について、その単語を含む質問文を対象とした。学習データに補完する未知語の種類数は、クエリログにおいて、 n 個の既知語のそれぞれと最も共起回数が多い未知語の上位 k 個とした。未知語の追加方法は、共起関係にある既知語と未知語がそれぞれ1回出現する質問文を学習データに追加する方法と、未知語のみが1回出現する質問文を学習データに追加する方法の2種類とした。また、未知語の補完を行わない場合を Baseline とした。これらをまとめたものを2に示す。

5.3 実験結果

提案手法によって4種類の方法で未知語を補完した学習データと、補完しない学習データの5つについて、学習データ数を100分割し、1%ずつ増やしながら学習曲線を作成した。既知語と未知語が1回ずつ出る質問文を追加した場合の結果を図6、未知語のみが1回出る質問文を追加した場合の結果を図7に示す。図6に示した中では、学習データを1%～6%利用したときに、Ext($k=1, u=1, N=2$)が最も高い値となった。図7に示した中では、学習データを1%～9%利用した時に、Ext($k=1, u=1, N=1$)が最も高い値となった。この2つの手法は、学習データへの追加方法のみ異なるが、手がかりとする既知語の数も補完する未知語の数も同じである。このことから、学習データが少ない場合、各カテゴリで最も出現頻度が高い既知語を手がかりとし、クエリログにおいてその単語と最も多く共起する未知語の特徴を既知語の正解カテゴリをラベルとして用いて学習データに補完することで、最も効果的に未知語問題を解決することが出来、従来手法よりも高い精度で分類できる

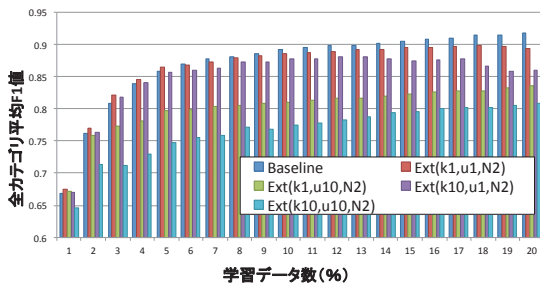


図 6 既知語+未知語を1質問文として追加した場合

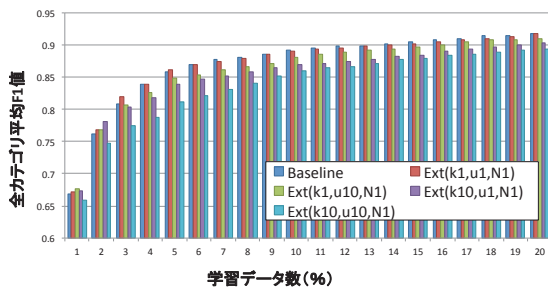


図 7 未知語のみを1質問文として追加した場合

ことがわかった。

一方、図 6 に示した $\text{Ext}(k=10, u=10, N=2)$ は全手法の中で常に最も低い値となった。同様に、図 7 に示したように未知語のみを質問文として追加する手法の中では、 $n = 10, k = 10$ とした $\text{Ext}(k=10, u=10, N=1)$ が最も低い値となった。この2つの手法は、学習データを1%増やす毎に、手がかり語を10種類追加し、手がかり語1つにつき未知語を10種類補完するため、合計で100種類の未知語を補完している。しかし、3.3.2節で検証した結果と同様に、未知語をあまりに多く補完すると、分類精度が下がってしまうことが確認できた。

また、学習データが増えるに従って、その差は逆転していることがわかる。これについて、次章でカテゴリ別に詳細を分析し、原因を調べる。

6. 考察

前章の実験結果について詳細な分析を行った。

まず、提案手法の中で精度が高かった、手がかりとなる既知語を1種類ずつ増やして追加する手法について、カテゴリ別に見る。特に音楽、ショッピング、銀行・ATMのカテゴリで、学習曲線の立ち上がり方が早くなっている。音楽カテゴリにおいて、 $k = 1, u = 1$ とした場合の分類精度を図 8 に示す。特に学習データが2%~14%と少ない場合に、提案手法の2手法のどちらにおいても Baseline を上回っている。音楽カテゴリは、今回対象とした15カテゴリの中で最も単語のバリエーションが多いことから、未知語問題が発生しやすいカテゴリであると言える。実際に補完された単語を見てみると、「曲」という既知語から「歌詞」とい

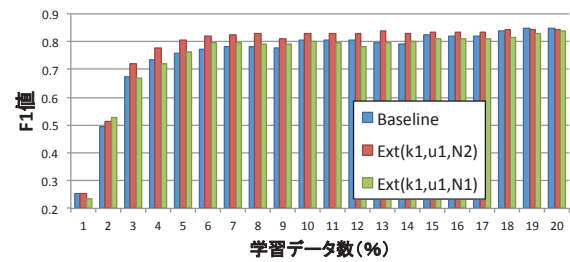


図 8 音楽カテゴリの F1 値

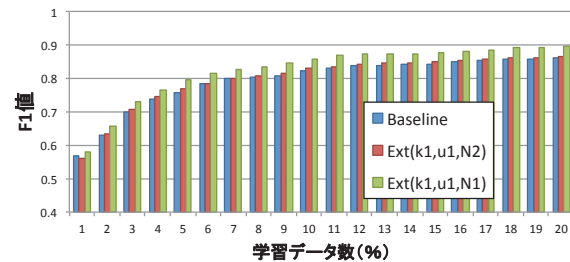


図 9 ショッピングカテゴリの F1 値

う音楽カテゴリで広く用いられている未知語が補完されていたり、「歌う」「シングル」などの既知語から歌手名が補完されていたりと、音楽カテゴリに関連する様々な単語が追加されている。このことから、本提案手法は、質問文中の単語バリエーションが多いカテゴリにおいて有効であることがわかった。

次に、ショッピングカテゴリについて同様の結果を図 9 に示す。学習データを2%~20%に増やしても、提案手法の精度が上回っていることが分かった。ショッピングカテゴリでは、「プレゼント」「贈る」「ギフト」という既知語から「父の日」「友達」「卒業」などといった、プレゼントを贈る目的を表す未知語が補完されている。これによって、プレゼントのために商品を探している質問文において未知語問題が解決し、分類精度向上につながっていると考えられる。銀行・ATMは、銀行名を表す既知語から、「手数料」「振込」などの未知語が補完されていたり、逆に「バンク」という既知語から銀行名を表す未知語が補完されていたりなど、銀行やATMでの操作に関連する単語や固有名詞などが学習データに追加されることで、未知語解決につながっている。

これらのことから、特定のカテゴリで特徴的な単語を手がかりの既知語とした場合、関連する様々な未知語を補完することができた。これによって、関連の強いカテゴリがラベル付けされた未知語を学習データに追加できたため、未知語問題が解決し、分類精度向上につながったと考えられる。

一方、精度が最も低かった、既知語を10種類ずつ増やし、既知語+未知語で1文として追加する手法 $\text{Ext}(k=10, u=1, N=2)$ において、詳細に分析する。この

学習データによるモデルでは、学習データ数が5%を超えた付近から精度が上がりなくなっている。

カテゴリ別に見てみると、特に画像、動画、天気、習い事・資格カテゴリで、Baselineや既知語と合わせて追加した場合に比べて精度が悪くなっていることがわかる。ここで追加する未知語を決める元となる既知語に、どのような単語が使われているのかを調べた。

画像と動画のカテゴリでは、追加する未知語を決める既知語として、「12月」などの撮影日時に関連する単語が用いられていた。また、同様に撮影日時を表す単語として「お正月」「クリスマス」といった特定の時期や月日を表す単語も手掛りの既知語として用いられていた。このような単語は、端末内に保存されている、ユーザが撮影した写真や動画を検索する意図の質問文に含まれている。これは、従来の検索エンジンのクエリやQAサイトの質問文では起こらない、スマートフォンやタブレットで用いられる質問対話エージェントの質問文特有の事象である。今回は外部データとして検索エンジンのクエリログを用いたため、音声対話エージェントにおける質問文と性質が異なることから、適切に単語を補完することができなかつたと考えられる。

また、天気カテゴリでは、「今日」「来週」など、知りたい天気予報の日時を相対的に示す単語が特に多く含まれている。このような単語は、画像・動画カテゴリの他に、イベントカテゴリでも多く見られた。日時を相対的に表す単語は、一般のウェブ検索エンジンではあまり入力されない。このような単語を手がかりとしてクエリログを探すと、その既知語を含むテレビ番組名や曲名などがあり、その中で共起する未知語が学習データに追加される。その結果、天気カテゴリ以外のカテゴリに関連する単語が、天気カテゴリの質問文として追加されてしまい、カテゴリ判定を誤っていると推測される。

このように、カテゴリによっては、同じ検索意図を持っていても、ウェブ検索エンジンのクエリとして入力されることが少なく、音声対話エージェント特有の単語があると考えられる。これらの単語を手がかりの既知語として用いた場合、データの性質の違いから、異なるカテゴリに関する未知語が共起語として用いられる可能性がある。これによって、関連の低いカテゴリを正解ラベルとして未知語が学習データに追加されてしまうことから、分類精度低下につながっていると推測される。

これらの結果から、音声対話エージェントには、端末内の検索や相対的な日時表現による検索など、従来のウェブ検索では用いられない単語が出現しやすいカテゴリがあることが確認できた。一方、カテゴリによっては、未知語を補完するために用いる外部データと共起語の傾向が類似しているもあり、予めカテゴリ別にその分析を行った上で手がかりとなる既知語を選定することで、未知語問題を解決

し、精度向上につなげることができると示唆された。

7. 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検討した。特に学習データが少ない時に、学習データに含まれない未知語が多く存在することから精度低下につながるという未知語問題に焦点を当てた。実際に音声対話エージェントによる質問文データを分析し、未知語問題が生じていることを確認した。また、ウェブ検索エンジンのクエリログを用いて、この未知語問題を解決する手法を提案した。評価実験の結果、カテゴリによっては分類精度が向上していることが確認できた。一方、ウェブ検索エンジンと音声対話エージェントの使われ方の違いから、正しく未知語を補完できないカテゴリも存在することがわかった。今後は、サービス間の違いを踏まえ、手がかりとなる既知語の選択方法を改善し、より多くの未知語問題の解決を目指す。

参考文献

- [1] 徳永陽子, 数原良彦, 戸田浩之, 鷺崎誠司: 単語の出現度合いを考慮した質問文マルチクラス分類, *DEIM2014* (2014).
- [2] 田馳, 手塚太郎, 小山聡, 田島敬史, 田中克己: 質問キーワードの意味的関連と近接性に着目したウェブ検索の精度改善, *DEWS2006* (2006).
- [3] Sebastiani, F.: Machine Learning in Automated Text Categorization, *ACM Comput. Surv.*, Vol. 34, pp. 1-47 (2002).
- [4] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: Short Text Classification in Twitter to Improve Information Filtering, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 841-842 (2010).
- [5] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying Latent User Attributes in Twitter, *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pp. 37-44 (2010).
- [6] Li, X., Wang, Y.-Y. and Acero, A.: Learning Query Intent from Regularized Click Graphs, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pp. 339-346 (2008).
- [7] Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J.-T., Chen, E. and Yang, Q.: Context-aware Query Classification, *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 3-10 (2009).
- [8] Zhou, K., Cummins, R., Halvey, M., Lalmas, M. and Jose, J. M.: Assessing and Predicting Vertical Intent for Web Queries, *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pp. 499-502 (2012).
- [9] Kang, I.-H. and Kim, G.: Query Type Classification for Web Document Retrieval, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SI-

- GIR '03, pp. 64–71 (2003).
- [10] Qu, B., Cong, G., Li, C., Sun, A. and Chen, H.: An evaluation of classification models for question topic categorization., *JASIST*, Vol. 63, No. 5, pp. 889–903 (2012).
- [11] Aikawa, N., Sakai, T. and Yamana, H.: Community QA Question Classification: Is the Asker Looking for Subjective Answers or Not?, *IPSJ Online Transactions*, Vol. 4, pp. 160–168 (2011).