エンティティのカテゴリ情報を用いたデータ拡張による質問カテゴリ推定

欅　　惇志†　　太刀岡勇気†

† 株式会社デンソーアイティーラボラトリ
E-mail: †{akeyaki,ytachioka}@d-itlab.co.jp

あらまし　本研究では，エンティティのカテゴリ情報を用いたデータ拡張を行うことで，対話システムの質問カテゴリの分類精度改善を目指す．自然言語が用いられる対話システムでは，同一の意図に基づくユーザの発話 (ただし本研究ではユーザの発話は質問に限定する) が多様な表現で投げかけられる．対話システムが多様な表現に対応するためには，多様な表現を含む学習データを用いてモデルを学習する必要がある．そこで本研究では，エンティティのカテゴリ情報を用いたデータ拡張を行う．学習データの自動生成のため，まずは Web からシードとなる少量の質問をクローリングする．次に，各質問をエンティティ·スロットとそれ以外のフレームに分類し，質問テンプレートを作成する．質問テンプレート中のエンティティを他のエンティティと置換することで新たな質問を生成する．この際，エンティティのカテゴリや質問のカテゴリが同一のエンティティに制限することで，不適切な質問が生成されることを抑制する．提案手法によって自動生成された質問の品質を評価するため，質問カテゴリ分類タスクの精度評価を行った．生成された質問を学習データとしてモデルを構築し，人手でアノテーションしたデータの分類を行った結果，データ拡張を行わなかった場合と比較して精度の向上が確認された．

キーワード　対話システム，データ拡張

1　はじめに

対話システムの発展に伴って，現在，さまざまな Web サイトにおいて，チャットボットと呼ばれる質問応答システムが提供されている．チャットボットの導入によってユーザへの即時レスポンスやオペレーター雇用コストの削減を実現している．チャットボットでは大量の質問-応答ペアからユーザの質問と類似質問を選別して回答を提示する用例ベースの手法が利用されることが主流である [1]．チャットボットのように自然言語が用いられる対話システムではユーザの多様な表現を吸収する必要があり，その結果，予め大量の質問-応答ペアを収集する必要がある．さらに，質問処理時には，それら大量データに対して探索を行う必要がある．

その一方で，タスク指向型と非タスク指向型が統合された構成の対話システムも一般的である [1]．特定のキーワードを含む発話にはタスク指向型，それ以外には非タスク指向型を適用するキーワード方式や，タスク指向型で対応できなければ非タスク指向型を適用する階層型などが存在する．チャットボットにおける用例ベースの手法とタスク指向型対話システムの組合せを考慮すると，予め質問のカテゴリの推定ができていれば，用例ベースの手法にて同一質問カテゴリに絞って類似質問を選別することができる．これによって誤った質問が選別されることを回避し，また，質問処理時に探索すべき質問-応答ペアの分量も削減される．これらを考慮し，本研究では，質問応答システムにおける質問カテゴリの推定に取り組む．

事前に質問カテゴリが定義可能な状況において，質問カテゴリの推定は分類タスクとして定式化される．分類問題としては教師あり学習の設定で解かれることが多いことから，本研究でも同様のアプローチに準拠する．教師あり学習の一般的な課題として，高性能な分類モデルを構築するためには多様な表現を含む大量のデータを用いる必要があるものの，それにはアノテーションコストの問題がつきまとう．大量の学習データを作成する際の人手によるアノテーション付与を行った場合にはコストが大きいため，人手の作業コストを抑制しつつ，多様な学習データを収集できる手法が期待される．また，データ拡張によるデータの自動生成はその解決策の一つであるものの，生成されるデータは実データから逸脱していては学習への悪影響を及ぼすことが懸念される．これらを踏まえて，人手によるアノテーションや書き換えルールの定義をすることなく，多様かつ大量な学習データの自動生成に取り組む．

提案手法では，質問中のエンティティは類似エンティティと可換であるという仮定のもと，質問生成を行う．なお，類似エンティティは同一のエンティティカテゴリかつ同一質問カテゴリのエンティティであるとみなす．具体的には，下記の手順を用いる．

（1）　質問テンプレートの生成
（2）　質問生成

質問テンプレートはエンティティ·スロットとそれ以外のフレームから構成され，エンティティに対してはエンティティ属性が付与される．また，質問生成では，質問テンプレート中のエンティティ·スロットに元エンティティと類似エンティティを挿入する．詳細は 3 節にて述べる．

また，以降，本研究では，質問カテゴリが多様かつ複数存在する状況を想定する．具体例として観光地推薦カウンターセールス対話システムを取り上げる．質問カテゴリや設計については 2 節にて説明する．

1：用例ベースの手法は非タスク指向型対話システムに分類される．

表 1　施設ジャンルと取得質問件数

施設ジャンル名	質問の件数
博物館	137
科学館	88
寺社	51
工場見学	28
施設見学	91
滝	58
川	63
渓谷	34
公園	82
城	85

さらに，提案手法によるデータ拡張の有用性を評価するため，4 節にて質問カテゴリ推定の分類精度の評価実験を行い，最後に 5 節にて関連研究を報告する．

2　観光地カウンターセールス対話システム

本研究で想定する観光地推薦カウンターセールス対話システムについて説明する．観光地推薦カウンターセールスを行ううえで，どのような質問カテゴリが存在するのか調査するため，まずは「(観光地の主要な施設ジャンル名) + よくある質問」で Web 検索を行い，観光地サイトの FAQ ページ (質問の一覧ページ) から質問を収集した．施設ジャンル名と収集された質問も件数を表 1 に示す．合計で 10 ジャンル 717 件の質問が集まった．

次に，前述の質問を分類するうえで適切な質問カテゴリを議論し，定義した．質問カテゴリが決まれば，各質問に対して質問カテゴリのアノテーションを行った．質問カテゴリと質問の例を表 2 に示す．観光地と無関係な質問は「その他」として分類し，モデル学習の際にも除外した．これらのカテゴリ定義やアノテーション作業は著者 2 名にて行った．

表 2 の通り,「ミュージアムショップ」や「学校教育支援」,「周辺の施設」などの質問カテゴリでは割り当てられた質問の個数が極めて少なく，これらの質問カテゴリに対しては効果的な学習を行うことは困難であることが想定される．実際，後述の実験の結果，データ拡張なしでは質問件数が少ないこれら 3 個の質問カテゴリでは分類精度は 0.0 を示した．

この結果からも少量データからの効果的な学習モデル構築は困難であることが示唆された．従って，次節にて議論するデータ拡張によって学習モデルの改善を目指す．

以降，表 2 のデータセットをオリジナルデータセット，データ拡張によって生成されたデータセットを拡張データセットと呼ぶ [2]．

3　エンティティのカテゴリ情報を用いたデータ拡張

本研究のデータ拡張は，(1) 質問テンプレート作成，(2) 質問生成の 2 つの手順から構成される．以降，それぞれの処理について述べる．

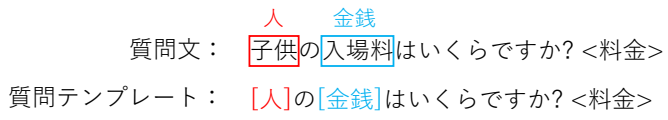

図 1　質問テンプレート作成

3.1　質問テンプレート作成

質問テンプレート作成の概要を図 1 に示す．まず，質問文からエンティティを抽出し，エンティティカテゴリを付与する．図 1 の例では，質問文：“子供の入場料はいくらですか？” から 2 個のエンティティ “子供” と “入場料” が抽出される．エンティティ “子供” にはエンティティカテゴリ “人”，エンティティ “入場料” にはエンティティカテゴリ “金銭” が付与されている．エンティティが抽出されれば，エンティティの存在していた箇所をエンティティ· スロットに置換し，エンティティカテゴリ情報を付与する．これにより，図 1 の質問テンプレートには 2 種類のエンティティ· スロットが埋め込まれ，それぞれ エンティティカテゴリ “人” とエンティティカテゴリ “金銭” のエンティティが挿入されることを意味する．なお，エンティティ· スロット以外の箇所 (黒字部分) はフレームであり，データ拡張の際に変化しない．

また，質問生成にて利用するため，エンティティカテゴリごとにエンティティリストを作成する (図 2 上部参照)．エンティティカテゴリは JUMAN++ [2] のカテゴリ情報 [3] を利用する．その際，質問生成にて利用するため，エンティティリストの各エンティティに対して質問カテゴリを付与する．なお，質問文や質問テンプレート末尾に記載されている情報は質問カテゴリであり，データ拡張によって生成された質問の質問カテゴリは質問テンプレートの質問カテゴリに倣う．

3.2　質 問 生 成

続いて，質問テンプレートとエンティティリストを用いて，質問生成を行う．エンティティリストのエンティティのうち，中には質問カテゴリに特徴的なエンティティが出現する可能性がある．一例を示すと，図 2 のエンティティカテゴリ “人” のエンティティリストに含まれるエンティティの “ガイド” は，質問カテゴリ “展示” に特化したエンティティであり，質問カテゴリ “料金” の “子供” と代替すると不適切な質問文が生成される可能性がある．従って，置換候補のエンティティを選択する際に，質問テンプレートの質問属性と同一のエンティティを選択する．

これらを踏まえて，質問テンプレート中のエンティティ· スロットに，エンティティカテゴリと質問カテゴリが一致するエ

2：拡張データセットには，オリジナルデータセットと，データ拡張によって生成された質問が含まれる．

3：JUMAN++ のカテゴリは下記の 22 種類である：人，組織・団体，動物，植物，動物-部位，植物-部位，人工物-食べ物，人工物-衣類，人工物-乗り物，人工物-金銭，人工物-その他，自然物，場所-施設，場所-施設部位，場所-自然，場所-機能，場所-その他，抽象物，形・模様，色，数量，時間

表 2　観光地推薦対話システムにおける質問カテゴリ一覧

カテゴリ名	質問の件数	質問の例	分類精度
カフェ＆レストラン	24	七五三が終わってからおじいちゃんと会食がしたいのですが？	1.00
バリアフリー対応	54	車椅子で館内を移動することはできますか？	1.00
ミュージアムショップ	10	図録やグッズを購入できますか。	0.00
学校教育支援	10	学校のクラス等で申し込みできるイベントはありますか？	0.00
館内・施設の質問・ルール	125	ツアーに不要な荷物を預けることはできますか？	0.47
見学・体験可能時間	47	夜でも行ける？	0.78
交通アクセス	55	駅から徒歩何分くらいですか？	0.92
周辺施設	4	博物館の外で食事がとれる場所はありますか？	0.00
団体でのご来館	23	大型バスは駐車できますか？	1.0
展示・体験の質問・ルール	247	博物館をみるのに、時間はどれくらいかかりますか？	0.47
予約	34	キャンセル待ちはできますか？	0.40
料金	29	高齢者に対する割引はありますか？	0.00
その他	53	「中吉」と「吉」はどっちがいいですか？	–

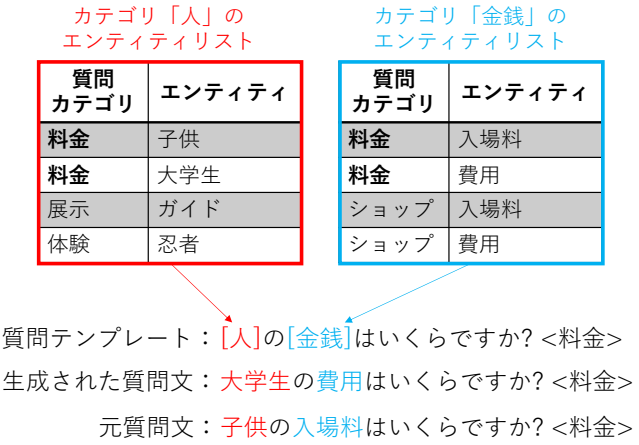

図 2　質 問 生 成

ンティティをランダムに選択して挿入する．図 2 では，エンティティカテゴリ “人” には質問カテゴリ “料金” の “大学生”，エンティティカテゴリ “金銭” には同様に質問カテゴリ “料金” の “費用” が挿入された．その結果，元の質問文 “子供の入場料はいくらですか?” から “大学生の費用はいくらですか?” が生成された．このとき，多様性向上のため，元の質問文と同一のエンティティは選択しないこととする．また，質問生成によって既に生成された質問と同一の質問文が生成された場合には，再度質問生成を行って同一の質問文が存在しないようにする．

4　評 価 実 験

提案手法の有効性評価のために行った評価実験について述べる．

4.1　比 較 手 法

まず，データ拡張を行わない状況でのオリジナルデータセットの潜在能力を明らかにする．評価実験を行うために，オリジナルデータセットを学習用の train と評価用の test に分割した．その際，train:test は 5:5 となるように調整し，それぞれに分類されるデータは無作為抽出を行った．test を用いてモデルを構築して test に対して評価を行った．

次に，拡張データセットを用いた評価実験方法について述べる．まず，train に含まれる質問に対してデータ拡張を行う．拡張データセットには train とデータ拡張によって生成された質問が含まれ，これらを用いてモデル構築を行う．その後，データ拡張を行わない状況と同様に，test に対して評価を行う．これによって，自動生成された質問のアノテーションデータとしての品質を評価することが可能である．その際，エンティティの挿入候補の選択において，エンティティカテゴリのみに着目したデータ拡張 (E カテゴリ) と，エンティティカテゴリと質問カテゴリ両方に着目したデータ拡張 (EQ カテゴリ) それぞれの評価を行う．これは，質問カテゴリごとに特徴的なエンティティによって不適切なデータ生成が行われる可能性を確認することを目的としている．

4.2　分類モデル構築

分類手法としては SVM を用いたが，任意の教師あり設定の学習手法が適用可能である．質問から特徴量を抽出する過程は下記の通りである．

（1） 形態素解析を行い，内容語のみを取り出す．動詞については原形化を行う．

（2） 各質問文に対して one-hot encoding を行う．

4.3　データ拡張の実験結果

表 3 に質問テンプレートのエンティティ·スロット数と頻度の関係を掲載する．質問テンプレートに対して平均 2.97 個のエンティティ·スロットが含まれていた．エンティティ·スロットが多いほどオリジナルの質問文と異なる多様な表現の質問文が生成されることになる．

また，エンティティカテゴリとその出現頻度は表 3 の通りである．中小物の出現頻度が高く，具体的なエンティティとしては，例えば，“予約” や “飲食”，“案内” など多様なエンティティが含まれる．

続いて，表 4 にデータ拡張を行ったことでカテゴリごとに何件質問生成が行われたのかを示す．E カテゴリでは，置換可能なエンティティの条件として質問テンプレートの質問カテゴリ

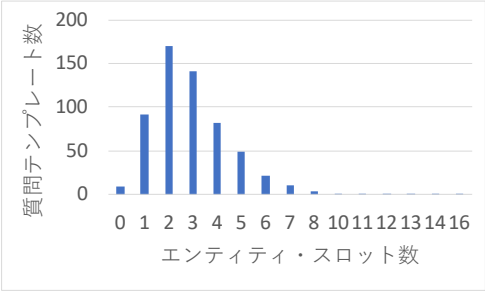

図 3　質問テンプレートのエンティティ· スロット数と頻度

表 3　エンティティカテゴリの出現頻度

エンティティカテゴリ	頻度
抽象物	712
場所-施設	198
人工物-その他	161
時間	143
人	79
場所-その他	72
人工物-乗り物	56
人工物-食べ物	46
数量	46
場所-自然	37
人工物-金銭	33
動物	32
組織・団体	31
人工物-衣類	26
場所-機能	19
自然物	14
植物	13
場所-施設部位	12
植物-部位	4
形・模様	3
動物-部位	2

表 4　各手法のカテゴリごとの学習データ件数

カテゴリ名	オリジナル	E カテゴリ	EQ カテゴリ
合計	341	3,236	2,258
カフェ＆レストラン	13	130	11
バリアフリー対応	29	270	13
ミュージアムショップ	6	60	8
学校教育支援	6	60	8
館内・施設の質問・ルール	63	605	286
見学・体験可能時間	24	230	124
交通アクセス	28	237	83
周辺施設	3	30	0
団体でのご来館	12	120	1
展示・体験の質問・ルール	124	1,144	1,676
予約	18	180	19
料金	15	150	37

の制約がないため，EQ と比較して生成された質問の件数が多くなった．

表 5　各手法のカテゴリごとの分類精度

カテゴリ名	オリジナル	E カテゴリ	EQ カテゴリ
全体	0.51	0.51	0.55
カフェ＆レストラン	1.00	0.25	1.00
バリアフリー対応	1.00	0.47	1.00
ミュージアムショップ	0.0	0.00	0.00
学校教育支援	0.0	0.00	0.00
館内・施設の質問・ルール	0.47	0.42	0.55
見学・体験可能時間	0.78	0.78	0.91
交通アクセス	0.92	0.46	0.75
周辺施設	0.00	0.00	0.00
団体でのご来館	1.00	1.00	1.00
展示・体験の質問・ルール	0.47	0.57	0.50
予約	0.40	0.29	0.33
料金	0.00	0.38	0.46

4.4　分類精度計測結果

表 5 に実験結果を掲載する．提案データ拡張 EQ カテゴリは，オリジナルと比較して全体的な傾向としては分類精度が向上した．これら結果より，データ拡張によって生成された質問は一定の品質を持ち，学習モデル構築において好影響を及ぼしたと考えられる．

また，E カテゴリと EQ カテゴリの比較では，全て EQ カテゴリが同等もしくは高精度を示した．この結果からも，エンティティ置換候補の選択肢として質問カテゴリの制約を設けることは妥当であると考えられる．

これらの実験結果から，質問中のエンティティは類似エンティティと可換であり，類似エンティティ選択の際には，エンティティカテゴリに加えて質問カテゴリにも着目することがより効果的であるという結果が示唆された．

5　関 連 研 究

タスク指向型対話システムの中には，ユーザの発話意図 の推定結果に基づいてシステムの挙動 · 応答が決定されるものが存在する．これら意図推定を行う際に，事前に定義された対話行為カテゴリへの分類を行うタスクは対話行為分類 (dialogue act classification) [3] と呼ばれる．文献 [4] では対話行為カテゴリは「STATEMENT」,「OPINION」,「YES-NO-QUESTION」など 42 種類に分類される．ユーザ発話の意図推定は対話制御において重要な要素であると知られており [5]，その正確さはタスク遂行に成功率にも関与するため，教師あり学習の設定で多数取り組まれている [6], [7]．本研究では，対話行為として「質問」が投げかけられる状況を想定しており，推定の対象は質問カテゴリである．この通り推定対象の分類ラベルカテゴリが異なるものの，その点を除くと類似タスクであるため，対話行為分類におけるデータ拡張を行った関連研究について紹介する．

河野ら [8] は条件付き敵対的生成ネットワークを用いたデータ拡張を行った．当該研究では文脈を考慮する重要性を主張しており，条件付き Encoder-Decoder モデルによって文脈を考慮したデータ拡張を行った結果，実際に性能改善が見られた．

これらの知見は，本研究において着目した質問カテゴリやエンティティカテゴリの同一性は文脈の一種であると捉えることができ，性能向上が実現したという知見も一致する．

また，塚原ら [9] はフレーズ単位で書き言葉から話し言葉への言い換え変換を行うことでデータ拡張を行った．変換規則は，挨拶や相槌など 8 種類のフレーズに対して 173 種類の言い換え規則を定義した (例: おはようございます → おはよう，ですよね → ですもんね)．当該研究で分類したカテゴリにおいては名詞を中心としたエンティティにおいて多様な表現を獲得することが効果的であったが，対話行為の多様性向上においては言い回しにおける多様な表現の吸収が効果的であったと推測される．ただし，これらの言い換え規則によって本研究におけるフレームの書き換えが可能となるため，より多様な表現の質問文を生成できる可能性がある．従って，今後これら言い換え規則を利用した質問文生成を検討する．

6 おわりに

本研究では，質問中のエンティティを同一のエンティティカテゴリや質問カテゴリのエンティティと置換することでデータ拡張を行った．質問分類精度の評価結果，提案手法を用いて作成した質問を用いることで分類精度が向上することが明らかになった．

今後の課題としては，今回は質問カテゴリの定義を著者 2 名の主観にて行ったため，今後はその妥当性の検証や必要に応じて修正を行う．また，今回の提案手法が汎用的に分類性能改善に貢献できることを検証するため，既存のデータセットを用いた定量評価を行う．

謝　辞

対話システムの分類や構成・実装に関して理化学研究所吉野幸一郎氏に有益な教示を受けた．ここに記して謝意を表す．

文　献

[1] 東中 竜一郎, 稲葉 通将, and 水上 雅博. *Python* **でつくる対話システム**. オーム社, 2020.

[2] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. In *Proceedings of EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015.

[3] Kristy Boyer, Eun Y. Ha, Robert Phillips, Michael Wallis, Mladen Vouk, and James Lester. Dialogue Act Modeling in a Complex Task-Oriented Domain. In *Proceedings of the SIGDIAL 2010 Conference*, 2010.

[4] Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26, 2000.

[5] Conversational System for Information Navigation Based on POMDP with User Focus Tracking. *Computer Speech Language*, (34), 2015.

[6] Ji Young Lee and Franck Dernoncourt. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Proceedings of NAACL-HLT 2016*, 2016.

[7] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *Proceedings of the 26th International Conference on Computational Linguistics*, 2016.

[8] 河野 誠也, 吉野 幸一郎, and 中村 哲. 条件付き敵対的生成ネットワークを用いたデータ拡張による対話行為分類法の検討. In **情報処理学会研究報告音声言語情報処理** *(SLP)*, 2018.

[9] 塚原 裕史 and 内海 慶. 言い換えを利用した対話行為推定の汎化性能向上. In **言語処理学会 第** *22* **回年次大会 発表論文集**, 2016.