4S-1

音声対話システムのための 周辺の文脈に着目したドメイン固有語のカテゴリ特定

藤巻 寛継 † 駒谷 和範 † 佐藤 理史 †

† 名古屋大学大学院 工学研究科 電子情報システム専攻

1. はじめに

本研究はデータベース検索型音声対話システムにおいて，ユーザ発話内のドメイン固有語のカテゴリの特定を目指す．ドメイン固有語とは，ドメインに現れる固有名詞のことであり，例えばレストラン検索ドメインでは，店名や駅名，食べ物などである．カテゴリは，検索対象データベースのフィールドに対応している．例えば，レストランデータベースにはレストラン名を表すフィールドがあるため，そのフィールドに属する単語に対してカテゴリ（例えば NAME）が定義できる．データベース検索型音声対話システムでは，ユーザからの検索要求を SQL コマンドで表現するため，ユーザの質問に答えるにはカテゴリの特定が必要である．

本研究では，未知のドメイン固有語（未知語）がユーザ発話に含まれている場合でも，それに正しくカテゴリを付与することを目標とする．近年開発されている超大語彙音声認識器（Google 音声認識など）を使用する場合，このような未知語が音声認識結果に含まれることになる．この未知語のカテゴリが分からない場合，システムは適切な応答ができない．例えば，図 1 に示すように，未知語（桜坂）のカテゴリが分からない場合，システムは金山のみで検索し，ユーザの要求を満たす応答を返すことができない．これに対して，未知語のカテゴリがわかることで，仮にデータベースに桜坂がなくても，システムは適切に応答を返すことができる．

本稿では，機械学習を用いてドメイン固有語のカテゴリを特定する．特に，学習データにおいて正解ラベルが与えられていないドメイン固有語（未知語）に対してもカテゴリを特定する手法を検討する．

2. 機械学習によるカテゴリ付与

本研究では，Conditional Random Fields (CRF)[1] を用いた機械学習により，ドメイン固有語に対してカテゴリ付与を行う．

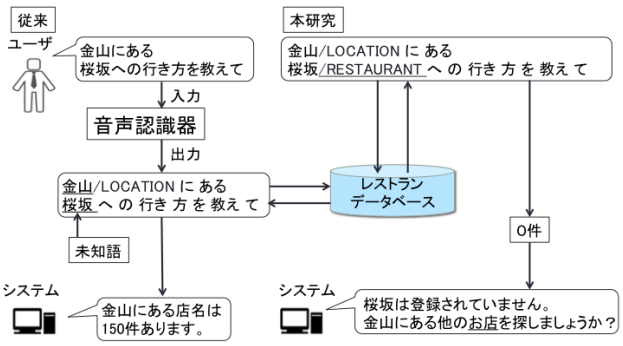

図 1: ユーザとシステムの対話例

Identifying Categories of Domain-Specific Words by Focusing on Those Contexts for Spoken Dialogue Systems: *Hirotsugu Fujimaki, Kazunori Komatani, and Satoshi Sato (Nagoya Univ.)*

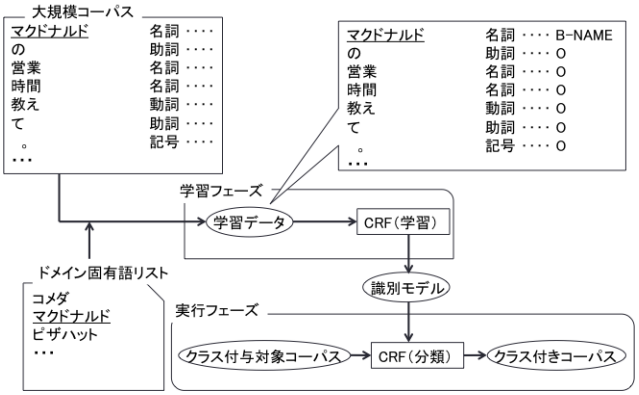

図 2: 全体像

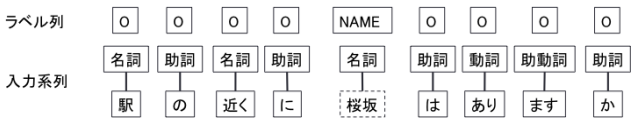

図 3: 使用する特徴量

2.1 当該単語に関する特徴量の除去

使用する特徴量の例を図 3 に示す．本稿では，当該単語の前後の単語の表記および品詞を特徴量に用いる．文脈から未知語を推定するためには，ドメイン固有語の周辺単語の情報を利用する必要がある．CRF は単語列全体のパターンの出現確率からカテゴリを推定するため，学習データに多く現れる当該単語の表記がカテゴリ付与に強い影響を与えることが考えられる．このため，周辺単語の特徴量がカテゴリ付与に有効に働かないと考えられる．当該単語の特徴量とは，NAME が付与されるドメイン固有語自身の表記のことであり，図 3 の例では「桜坂」を指す．

未知語に対するカテゴリ付与に有効な特徴量を調査するために，当該単語の特徴量を用いた場合と除去した場合のそれぞれで CRF を学習し，ドメイン固有語に対するカテゴリ付与数を比較した．

3. 評価実験

評価実験では，文脈に重点をおいた学習をするための有効な特徴量の検証を行う．

3.1 実験条件

CRF によるカテゴリ付与の全体像を図 2 に示す．大規模コーパスから店名を人手で抜き出し，ドメイン固有語リストを作成した．本稿で用いたドメイン固有語リストには 50 単語の店名を登録した．ドメイン固有語リストと大規模コーパスとの文字列の一致により，正解ラベルが付与されたデータを学習データとした．大規模コーパスには，Yahoo!知恵袋の「料理，グルメ，レシピ」カテゴリに属するコーパスを二分割し，それぞれを学習データとカテゴリ付与対象コーパスに用いた．

学習データの詳細を表 1 に示す．カテゴリ付与対象

表 1: 学習データの詳細

文数	語彙サイズ	NAME カテゴリ付与単語数	
		のべ	異なり
314,292	766,485	5,530	27

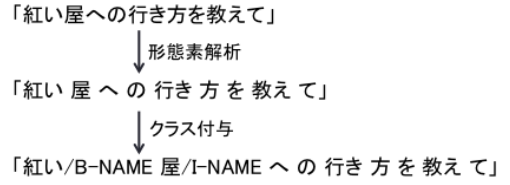

図 4: IOB2 を用いたカテゴリ付与例

表 2: NAME カテゴリが付与された単語数

	のべ	異なり	正解数（異なり）	
			既知語	未知語
当該単語あり	4935	22	22	0
当該単語なし	1402	329	20	31

コーパスに用いた Yahoo!知恵袋コーパスは 314,230 文である．形態素解析器には Mecab を用いた．

店名には複数の形態素からなるものが多くある．例を図 4 に示す．図 4 より，「紅い屋」を形態素解析すると 2 つの形態素に分割される．このような店名に対して「紅い/B-NAME 屋/I-NAME」のように二つのカテゴリを与える．このため機械学習には，一つ前のカテゴリを考慮することができる CRF を用いる．CRF を用いることにより，カテゴリ系列全体としての確からしさを学習する．

3.2 NAME カテゴリ付与単語数の検証

カテゴリ付与結果を表 2 に示す．このとき用いた特徴量は当該単語の前後 4 単語までの単語表記および品詞である．表 2 より，のべ数は当該単語ありで学習した場合のほうが多くの単語に NAME が付与されていることがわかる．この単語はすべて学習データに NAME が付与されている単語（既知語）であり，学習データに NAME が付与されていない単語（未知語）は，カテゴリ付与対象コーパスにも NAME が付与されない結果が得られた．つまり，当該単語ありで学習した場合，学習データに正解カテゴリが付与されている単語にしかカテゴリが付与されない．このことから，当該単語ありで学習した場合，文字列の一致でカテゴリ付与する方法と同じ結果になり，未知語に対してカテゴリ付与できないことがわかる．これは，当該単語の特徴量がカテゴリ付与に強い影響を与えているためだと考えられる．

一方，当該単語の特徴量を除去して学習した場合，正しく NAME が付与された単語の，異なりののべ数は 51 単語であり，その内未知語に対して 31 単語付与できた．

未知語に対してカテゴリ付与した例を図 5 に示す．図 5 より，ジョナサンという未知語に対してカテゴリ付与できた．これは，学習データから NAME が付与される単語の後の「の クーポン 券」という文字列の出現確率を学習した結果である．この結果より，当該単語の周辺の文字列の出現確率を学習することで，未知語に対してカテゴリ付与が可能になったといえる．

また，カテゴリ付与した 1402 語のうち，正解数は 972 語（69.3 %）という結果が得られた．この結果より，NAME 付与数が当該単語ありと比べて少ないことがわかる．

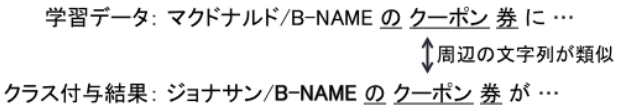

図 5: 未知語に対する NAME カテゴリの付与例

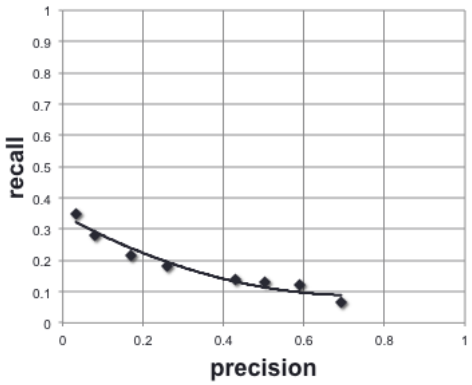

図 6: Precison-Recall 曲線

3.3 NAME カテゴリ付与精度の検証

カテゴリ付与対象コーパス中の店名数を推定し，カテゴリ付与精度を検証する．まず，カテゴリ付与対象コーパス 314,230 文から 1,000 文を抽出した．その結果，1000 文に含まれる店名数が 77 語であることを人手で確認した．従って，カテゴリ付与対象コーパス中の店名数を 15,000 語と推定できる．

評価指標には，Precision-Recall を用いる．Precision と Recall はトレードオフの関係であるため，両者を評価指標として用いる．そこで，学習データ中の NAME カテゴリが含まれる文の割合を 0.01 % ～100 % まで変化させたときの各点をプロットした．NAME カテゴリが含まれる文は約 5000 文である．これに対して，NAME カテゴリが含まれない文を学習データに追加することにより，学習データが含まれる文の割合を調整した．

結果を図 6 に示す．図 6 より，全体的に Recall が低いことがわかる．これは，カテゴリ付与対象コーパス中の店名の前後の文字列が，学習データ中に現れていないことが問題であると考えられる．今回用いた学習データは，人手で作成したドメイン固有語リストと大規模コーパスとの文字列の一致により，正解ラベルを与えた．この学習データには，本来店名であるはずの単語に対して NAME が付与されていない単語が多く存在すると考えられる．このため，店名の周辺の文脈の学習が不足すると同時に，店名を誤って学習したと考えられる．

4. おわりに

実験結果より，精度の高いカテゴリ付与を行うためには，今回の学習データでは店名の周辺の特徴量が不足していると考えられる．そこで，カテゴリ付与対象コーパス中の店名に対してより多くのカテゴリを付与を行うことにより，精度の高いカテゴリ付与を目指す．具体的には，人手で作成した小規模な学習データから学習したモデルを大規模コーパスに適用し，大規模な学習データへ拡張する．

参考文献

[1] 高村大也：言語処理のための機械学習入門 (自然言語処理シリーズ), コロナ社 (2010).