6P-02

限定された学習データ量での未知レストラン名の検出

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 ‡ 大阪大学 産業科学研究所

1. はじめに

音声対話システムでは未知語の出現が避けられない．未知語とは，関係データベース（RDB）を検索する音声対話システムでは，RDB 外の単語にあたる．ユーザ要求を満たす応答をするためには，図 1 に示すように，入力されたユーザ発話内の単語に対して，意味ラベルを付与する必要がある．付与すべき意味ラベルは，RDB 検索型音声対話システムでは，一般に検索対象の RDB のフィールド名に対応する．

未知語が入力された場合でも，その単語が RDB にないことを考慮した応答をするには，未知語に対するラベル付与が必要である．単語列に対するラベル付与には機械学習を用いる手法が一般的であるが，学習データを大量に用意する必要がある [1]．意味ラベルに用いるフィールド名は RDB ごとに変わるため，一般性がない．例えば，図 1 の RDB では,「ラーメン」は GENRE フィールドに属するが，別の RDB では FOOD フィールドに属することもある．したがって，RDB ごとに大量の学習データの存在を前提とするのは現実的に厳しく，限定された学習データ量でも意味ラベルが付与できる必要がある．

本研究ではレストラン検索ドメインにおいて，限定された学習データ量でも，発話内の未知レストランを検出する手法を提案する．未知レストラン名とは，学習データや RDB に現れないレストラン名のことである．レストラン名の周辺文脈の特徴とレストラン名自身の特徴を新たに設定し，用いることで，未知レストラン名に対するラベル付与を行う．

さらに，二つの異なる特徴セットを用いたラベル付与器の結果を併用することで，発話内のレストラン名の検出性能の向上を図る．具体的には，当該単語の表記を特徴に用いる既知レストラン名重視の特徴セットと，当該単語の表記を特徴に用いない未知レストラン名重視の特徴セットのそれぞれを用いた二つのラベル付与器の結果を併用する．これにより，未知レストラン名と既知レストラン名の両方を検出する．

2. 用いた特徴量

本研究では，レストラン名の検出に条件付き確率場（CRF）を用いる．ラベル付与は IOB2 フォーマットに基づいて行う．CRF などの識別モデルにおいて，一般的によく用いられる特徴は，単語 n-gram やその品詞である [2]．本稿では，限定された学習データ量でレストラン名を検出するために，これら以外にも文脈情報や単語自身の情報を用いる．用いた特徴を表 1 に示す．

文脈情報は，前後 4 単語ずつの単語表記や品詞の他にも，同格の有無など計 7 個を用いる．図 2 のように，レストラン名を抽象化した表現（カレー屋）は，レストラン名（サンマルク）の後に同格を表す文字列（という，ていう，っていう，って）を伴って現れやすい．このことを利用し，同格を表す文字列の後にレストラン名を抽象化した表現が現れたとき，これを表す特徴を同格の文字

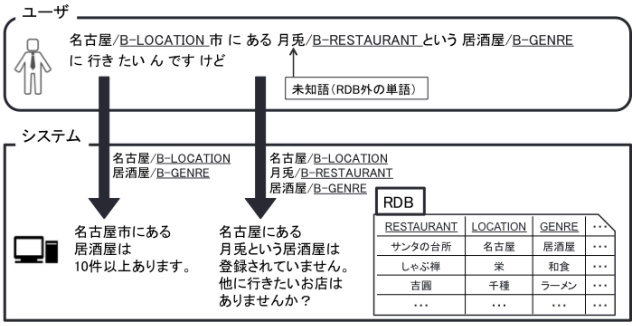

図 1: ユーザの発話に対するシステム応答の例

表 1: 使用した特徴

文脈情報	単語自身の情報
・前後 4 単語ずつの単語表記 ・前後 4 単語ずつの単語の品詞	・当該単語の表記 ・当該単語の品詞
・同格を表す文字列の有無 ・レストラン名特有の接尾語の有無 ・店（表記）の後がカタカナか否か ・語頭がジャンル名 ・地域名が前後に現れるか否か	・学習データ中での 頻度上位 3,000 語の 非レストラン名か否か ・mecab の辞書に 含まれる否か ・文字長 ・文字種

表記	品詞	追加した特徴の例			正解ラベル
名古屋	名詞	freq_noun	0	0	O
市	名詞	freq_noun	0	0	O
に	助詞	0	0	0	O
ある	動詞	0	0	0	O
サンマルク	名詞	0	0	unk	B-RESTAURANT
という	助詞	0	0	0	O
カレー	名詞	freq_noun	GENRE-FOOD	0	O
屋	名詞	freq_noun	0	0	O
に	助詞	0	0	0	O
行き	動詞	0	0	0	O
たい	助動詞	0	0	0	O

図 2: 特徴を追加した学習データ

列に対して付与する．また，レストラン名特有の接尾語の有無を特徴量に用いる．これは,「亭，屋，園，堂」のようにレストラン名に付属する接尾語に対して特徴を付与する．

単語自身の情報は計 6 個を用いる．これにより,「名古屋にあるサンマルクに行きたい」と「名古屋にある居酒屋に行きたい」という発話がある場合,「サンマルク」と「居酒屋」のように周辺文脈では区別できない文字列がレストラン名か否かを分類する．まず，学習データ中での頻度上位 3,000 語の非レストラン名か否かを特徴とする．次に，形態素解析器 mecab（ipadic）の辞書に含まれない単語に特徴を付与する．これは，未知レストラン名の検出に有効な特徴であり，検出したいレストラン名が辞書にない場合に有効である．

3. 二つのラベル付与器の併用

未知レストラン名と既知レストラン名それぞれに重点をおいた特徴セットを用いた検出結果の併用により，レストラン名の検出性能を向上させる．ここでは，これら

Detecting Unknown Restaurant Names with Limited Training Data: *Hirotsugu Fujimaki (Nagoya Univ.), Kazunori Komatani (Osaka Univ.), and Satoshi Sato (Nagoya Univ.)*

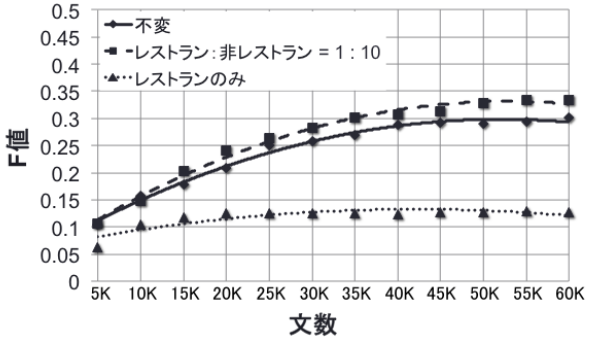

図 3: 学習データ量ごとのラベル付与器の性能

のラベル付与器のどちらかがレストラン名とした単語をレストラン名とする．二つの特徴セットを以下に示す．

未知レストラン名重視（未知重視）：

表 1 に示す特徴から「当該単語の表記」を取り除いたもの（計 12 個）．

既知レストラン名重視（既知重視）：

表 1 に示す特徴から「mecab の辞書に含まれるか否か」を取り除いたもの（計 12 個）．これを取り除いた理由は，既知レストラン名の検出に悪影響を及ぼす可能性があるためである．

4. 評価実験

特徴追加前後でのラベル付与性能を評価する．学習データには，yahoo 知恵袋のカテゴリ「料理，グルメ，レシピ」に属する文から必要最小限の量を見積もり，用いる．実験では，二つのデータに対するラベル付与結果（precision，recall，F 値）を評価した．一つ目は，学習データと同じ yahoo 知恵袋コーパス（10,000 文，レストラン数：629 語）である．二つ目は，研究室内の学生から収集したユーザ発話（120 文，レストラン数：120 語）である．

4.1 必要最小限の学習データの作成

必要最小限の学習データの量を見積る．学習データの量を 5,000 文ずつ増やしたときのラベル付与性能を図 3 に示す．

図 3 の結果では，学習データ中のレストラン名を含む文の割合を変えた．これは，割合の変化により事前確率 P(レストラン名 | 全単語) が変化することで，文字列に対するラベル付与数が変わるためである．具体的には，レストランを含む文の数を固定とし，レストランを含まない文を減らして割合を調整した．

図 3 の，レストランを含む文の割合が不変の学習データを用いた結果において，学習データの量を 5,000 文から 10,000 文に増やしたとき，F 値は約 0.05 向上している．その後の F 値の変化量は，徐々に小さくなっていき，55,000 文から 60,000 文に増やしたときの F 値の変化量は 0.01 に満たない．そこで本稿では，F 値の変化量が十分に小さくなったときの学習データ 60,000 文を必要最小限の学習データの量とした．以降の実験では，この 60,000 文の学習データを用いた．

4.2 特徴追加後のラベル付与性能

特徴追加前後でのラベル付与性能を表 2，未知レストラン名と既知レストラン名のそれぞれに対するラベル付与数を表 3 に示す．baseline1 では，前後 4 単語の単語表記と品詞，当該単語の品詞を特徴量に用いた．baseline2 では，baseline1 に当該単語の表記を追加した．表 2 より，baseline1 では，recall が低いのがわかる．また，baseline2 の recall は 0.69 であるが，表 3 より，検出されたのは，ほとんどが既知レストラン名であったことがわかる．

表 2: 特徴追加前後および統合前後のラベル付与性能

		precision	recall	F 値
yahoo 知恵袋	baseline1	0.70 (130/ 185)	0.21 (130/ 629)	0.32
	baseline2	0.96 (433/ 450)	0.69 (433/ 629)	0.80
	未知重視	0.22 (355/1602)	**0.56** (355/ 629)	0.32
	既知重視	0.87 (452/ 517)	0.72 (452/ 629)	0.79
	統合	0.29 (516/1800)	**0.82** (516/ 629)	0.42
ユーザ発話	baseline1	0.91 (20/ 22)	0.17 (20/ 120)	0.28
	baseline2	0.97 (38/ 39)	0.32 (38/ 120)	0.48
	未知重視	0.93 (64/ 69)	**0.53** (64/ 120)	**0.68**
	既知重視	0.91 (53/ 58)	0.44 (53/ 120)	0.60
	統合	0.92 (78/ 85)	**0.65** (78/ 120)	**0.76**

表 3: 未知と既知に対するラベル付与数（recall）

		未知レストラン名	既知レストラン名
yahoo 知恵袋	baseline1	0.06（ 7/123）	0.24（123/506）
	baseline2	0.05（ 6/123）	0.84（427/506）
	未知重視	**0.39**（48/123）	0.61（307/506）
	既知重視	0.10（12/123）	**0.87**（440/506）
	統合	0.40（49/123）	0.92（467/506）
ユーザ発話	baseline1	0.12（10/ 82）	0.26（ 10/ 38）
	baseline2	0.10（ 8/ 82）	0.79（ 30/ 38）
	未知重視	**0.55**（45/ 82）	0.50（ 19/ 38）
	既知重視	0.28（23/ 82）	**0.79**（ 30/ 38）
	統合	0.55（45/ 82）	0.87（ 33/ 38）

表 2 より，yahoo 知恵袋において，特徴を追加した未知重視の recall が baseline1 より 0.35 向上した．一方，baseline2 と比較すると，recall が低い．これは，評価データである yahoo 知恵袋中のレストラン名に既知レストラン名が多くあるからである．表 3 より，baseline2 では検出されたレストラン名の多くが既知レストラン名であり，未知レストラン名にはほとんどラベルが付与できていない．これに対し，未知重視では未知レストラン名にもラベルが付与できた．

未知重視の precision が baseline1，2 に比べて低いのは，mecab の辞書に含まれない非レストラン名の多くに対し，誤ってレストラン名だと検出しているためである．また表 2 より，ユーザ発話では，baseline1，2 に比べて未知重視の recall が向上した．それに伴い，未知重視の F 値が baseline1 に比べて 0.40，baseline2 に比べて 0.20 向上した．

4.3 二つのラベル付与結果の併用

表 3 より，yahoo 知恵袋とユーザ発話の両方において，未知重視では未知レストラン名が多く検出でき，既知重視では既知レストラン名が多く検出できた．それに伴い，併用後の recall も向上した．このことから，未知重視と既知重視のそれぞれが，未知レストラン名と既知レストラン名の検出に有効に働いたことがわかる．

また表 2 より，yahoo 知恵袋において，未知重視の precision が低いため，併用後の F 値が既知重視より低い．ユーザ発話においては，二つのラベル付与器の結果を併用することで，併用前の二つの特徴セットを用いた結果よりも F 値が向上した．

参考文献

[1] 橋本泰一，乾孝司，“拡張固有表現タグ付きコーパスの構築”，情報処理学会研究報告，NL-188-17，2008.

[2] 内元清貴，馬青，“最大エントロピー法と書き換え規則に基づく日本語固有表現抽出”，自然言語処理，Vol.7，No.2，pp.63-90，2000.