E-022

対話型観光地推薦システムにおける未知語の属性推定
Estimating Attributes of Unknown-Words in an Interactive Sightseeing Spot Recommendation System

渡辺 雄介† 杉本 徹†
Yusuke Watanabe Toru Sugimoto

1. はじめに

対話システムとは，コンピュータが人と対話を通してタスクを遂行するシステムである．タスクを遂行するためシステムはタスクごとの専門分野に関する知識を保有し，その知識を活用して対話のやり取りを行っていく．しかし，保有できる知識の量には限界があるため，ユーザから与えられる入力の中にシステムにとっての未知の語が含まれることは避けられない．未知語に対し何らかの処理を行わないと，対話がそこで途切れてしまいタスクを遂行することができない．これに対し，未知語に関する知識を対話によって獲得することができれば未知語への対応が可能となる．このような対話による知識獲得を的確に行うためには，未知語の属性を判別する必要がある．ここで，ユーザから観光地の希望を聞き出し対話を通してユーザの旅行プランの作成を支援する対話型観光地推薦システム[1]を例に説明する．このシステムではユーザの入力を「観光地」「地域」「カテゴリ」「目的」の 4 つの属性に分類し，その属性の情報に基づいて応答文を生成する．ユーザから入力された未知語に対してもこれらの属性を推定することにより，属性に応じたその未知語に関する知識獲得の対話が円滑に進められると期待できる．

本稿では，対話型観光地推薦システムにおける未知語の属性を推定する手法の提案を行う．高橋ら[2]は蓄積した実例から未知語の現れた例に最も類似する実例を選び，それに基づいて未知語の属性推定を行っている．それに対し本稿では，形態素の品詞や単語の概念という特徴から属性を推定する手法を提案する．

2. 対話型観光地推薦システム

対話型観光地推薦システム[1]は，自然言語によるユーザとの対話を通して観光地を推薦するシステムである．ユーザからの入力を形態素解析と係り受け解析を用いて解析し，その結果に応じてスロットを埋めていく．スロットには「観光地」「地域」「カテゴリ」「目的」の 4 つの属性があり，システムは埋められたスロットの状態によって応答を決定する．スロットの属性についての説明を表 1 に示す．

表 1：スロットの属性

属性	説明	具体例
観光地	観光地の名前	清水寺
地域	地域の名前	京都府
カテゴリ	観光地の特徴	寺
目的	観光の目的	初詣

次にシステムの対話例を示す．

ユーザ：千葉に行きたい
（「地域」スロットに「千葉」を格納）
システム：千葉県の中で何か希望はありますか?
ユーザ： 登山がしたい
システム：入力文を理解することができませんでした

「千葉」という語はシステムのデータベースに存在し，地域のスロットへと格納される．また「登山」という語は未知語であり，システムは理解できない旨の応答を返している．ここで「登山」という語はデータベースに存在していないが，「山」や「自然景勝地」というような近い意味を持つ語が既存のデータベース内に存在する．未知語に対し属性の付与を行うことで，知識の拡張，あるいは既存の知識との関連付けの処理を行う際の手がかりとすることができると考えられる．例えば「登山」という未知語に対し，目的という属性を付与した場合「登山ができるような場所はどこですか？」というような質問応答を用いることで，既存の場所を表す語の知識との関連付けを行うことができる．同様に，観光地の未知語に対してはその観光地についての情報を尋ねる質問，カテゴリに対してはそれに近い意味を持つカテゴリを尋ねる質問，地域に対しては地域の位置を尋ねる質問を行うことで対応する．このように属性に応じた処理を行うことで，未知語に対して効率的な知識獲得が行えると考えられる．

3. 未知語の属性推定

入力された未知語に属性を付与するため，4 種類の分類器を作成した．この分類器は，それぞれの属性を持つ単語は属性ごとに固有の特徴を持っているという仮定に基づいている．ここでいう特徴とは，品詞の並び，文字列の並び，単語の持つ概念の 3 つである．

3.1 SVM を用いた観光地の分類

既知の観光地の属性を持つ単語を用い，品詞の出現数，文字列の並びの情報を特徴量とし，サポートベクターマシン(SVM)を利用して推定を行う．

3.1.1 学習手順

SVM の学習は以下の手順で行う．

1. 既知の観光地名を形態素解析し，観光地名ごとに品詞の出現数，単語の終端の形態素の文字列を取得する．なお，品詞は IPA 品詞体系で定義されている 68 種類，形態素の文字列は出現回数が 10 回を越えるもののみを用いる．この単語を正例として定義する．
2. ブログからランダムに収集した観光地名以外の単語に対しても 1 と同様の処理を行う．この単語を負例として定義する．

†芝浦工業大学 Shibaura Institute of Technology

3. 1，2 で取得した情報を用い，特徴ベクトルを定義する．
4. 定義した特徴量について SVM で機械学習する．
5. 4 で学習したモデルをもとに未知語が観光地か否かの推定を行う．

3.1.2 特徴ベクトルの定義

- 68 種の品詞

品詞は単語に現れる形態素の数だけ存在する．1 単語中の形態素の数を N とし，それぞれの品詞の出現回数をω_iとするとそれぞれの品詞のスコアS_iは以下のような式で定義される．(i：品詞の種類，$1 \leq i \leq 68$)

$$S_i = \frac{\omega_i}{N}$$

- 文末の形態素

単語内の終端の形態素の文字列の種類の総数を T とし，そのそれぞれの出現回数をω_iとするスコアS_iは以下のような式で定義される．(i:文字列の種類，$1 \leq i \leq T$)

$$S_i = \frac{\log_{10}\omega_i \times T}{\sum \log_{10}\omega_i}$$

これらを特徴量として学習データに用いる．なお SVM としては libsvm[3]を採用し，カーネルには radial basis function(rbf)を用いる．

3.2 SVM を用いた地域の分類

既知の地域の属性を持つ単語を用い，観光地の分類と同様に品詞と単語の終端の形態素の文字列の情報を特徴量とし，サポートベクターマシンを利用して推定を行う．

3.3 概念を利用した目的の分類

EDR 概念辞書[4]の持つ概念体系の情報を利用して目的の分類を行う．EDR 概念辞書は根に抽象的な概念を持ち，葉へと辿るにつれて具体的な概念となるように構造化されたツリー構造を持っている．入力された語が上位概念に「行為」を持つかどうかで目的の推定を行う．

3.4 概念を利用したカテゴリの分類

カテゴリという属性の推定は，目的と同様に概念を用いて行う．入力された語が上位概念に「場所」を持つかどうかでカテゴリの属性を持つかどうかの推定を行う．

4. 評価

既知のデータベースに存在し，学習データとして利用していない観光地（1000 語），地域（3824 語），カテゴリ（205 語），目的（221 語）の属性を持つ単語と，それらの属性を持たないその他の単語（1020 語）を用いて分類器の精度の比較を行った．なお，その他の単語はブログよりランダムに抽出した単語である．単体の分類器の精度を表 2 から表 5 に示す．

表 2：目的の分類の精度

分類	精度(%)
観光地(1000)	98.1
地域(3824)	99.9
カテゴリ(205)	90.7
目的(221)	95.9
その他(1020)	71.8

表 3：観光地の分類の精度

分類	精度(%)
観光地(1000)	87.4
地域(3824)	79.1
カテゴリ(205)	47.3
目的(221)	93.7
その他(1020)	98.5

表 4：地域の分類の精度

分類	精度(%)
観光地(1000)	72.5
地域(3824)	96.9
カテゴリ(205)	94.6
目的(221)	100.0
その他(1020)	99.4

表 5：カテゴリの分類の精度

分類	精度(%)
観光地(1000)	27.7
地域(3824)	43.4
カテゴリ(205)	78.5
目的(221)	98.6
その他(1020)	95.0

次に，これらの分類器を図 1 のように組み合わせて分類器を作成する．

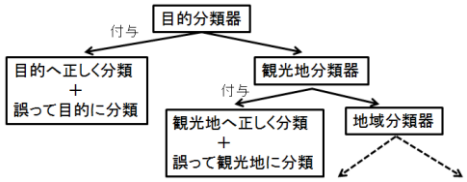

図 1：組み合わせた分類器による分類の方法

4 つの分類器の組み合わせは，全部で 24 通り存在する．その中で最も高い精度の組み合わせの結果を表 6 に示す．

表 6：全体の分類の精度

正解 / 実測	目的	観光地	地域	カテゴリ	その他	合計
目的	212	19	4	19	289	149
観光地	1	859	797	104	14	749
地域	0	13	2951	0	1	3994
カテゴリ	0	54	26	63	40	154
その他	8	55	46	19	676	1224
合計	221	1000	3824	205	1020	6270
分類精度	95.9	85.9	77.2	30.7	66.3	

この分類器は目的，観光地，地域，カテゴリの順に分類器を組み合わせたものである．目的は，最初に分類を行っているため単体の精度と同じ精度となる．観光地の精度は目的の付与誤り率が低いため，単体の精度とほぼ同じ精度となっている．それに対し，地域とカテゴリは，前段の分類器で誤って分類されることが多く，単体の分類器に比べ精度が低下するという結果となった．

5. おわりに

本稿では，単語の形態素と概念の情報に基づいて「観光地」「地域」「カテゴリ」「目的」の 4 つの属性を付与する手法を提案した．本手法により，「目的」「観光地」「地域」の分類はそれぞれ 96%，86%，77%という精度が得られた．「カテゴリ」の精度は約 30%と低くなったが，これは「観光地」との区別が正しく行えないこと原因として考えられ，この 2 つを区別する手法を検討する必要がある．また，今後はこの付与された属性を利用した属性ごとの知識獲得手法を考案するとともに，対話システムを用いて，ユーザからの入力に対する属性付与の精度の評価も行っていきたい．

参考文献

[1] 磯崎紘,杉本徹, “対話型観光地推薦システムへの状態遷移モデルの導入と評価”， 第 74 回情報処理学会全国大会，2012
[2] 高橋康博,堂坂浩二,相川清明，”音声対話における実例に基づく未知語属性推定”， 情報処理学会研究報告．自然言語処理研究会 2001(69), 199-204, 2001
[3] Chin-Chung Chang，LIBSVM，http://www．csie．ntu．edu．tw/cjlin/libsvm
[4] 日本電子化研究所:EDR 電子化辞書第二版，2001