機械学習を用いた観光対話システムのための自動知識獲得

Automatic knowledge extraction using machine learning for tourism dialogue system

與那城 寛　嶋田 和孝

Kan YONASHIRO and Kazutaka SHIMADA

九州工業大学 情報工学部 知能情報工学科

Department of Artificial Intelligence, Kyushu Institute of Technology

shimada@pluto.ai.kyutech.ac.jp

Abstract: Recently, dialogue systems for tourism have an important role due to a manpower shortage. Our target is a tourism dialogue system based on an enumeration strategy. For the purpose, we need to collect pairs of alternatives and explanations. In this paper, we propose an extraction method of keywords that are suitable as the alternatives for an explanation sentence. We apply a machine learning technique to the task and compare it with a simple baseline based on TF-IDF values. The experimental result shows the effectiveness of our method.

1　はじめに

近年，訪日外国人旅行者が増加しており，観光案内をするための人材が不足している．そこで人に代わる観光ガイドとして対話型の観光案内システムが作成されている．主なシステムとしては株式会社ビースポークの Bebot[1] や AssisTra[1] などが挙げられる．対話システムとしては Wikipedia を利用したもの [2] やニューラル言語モデルを用いたもの [3] などがある．しかし，ユーザが自由に入力できる対話システムではユーザの発話を正しく理解できなかった場合に対話破綻（ユーザが対話を継続できなくなる状態）[4] を引き起こすことが少なくない．

システムの頑健性を考慮すればシステムがユーザの入力を制御できる列挙型対話システム [5] が望ましい．図 1 に列挙型対話システムの例を示す．列挙型対話システムでは，システムが応答可能な選択肢を提示し，ユーザとのインタラクションを図る．したがって，列挙型の対話システムを構築するためには，列挙するべき選択肢とその説明文のペアが必要である．しかしながら，このようなペアを含むデータベースを手作業で作成することはコストの掛かる作業である．一方で，観光地に関する情報は Web などに広く存在し，説明文そのものの自動獲得は比較的に容易に可能である．

本研究では，Web から獲得された説明文とその選択肢の自動獲得に関する手法について提案する．説明文を対話システムの応答文だとしたときに，それを応答文として扱うための選択肢をその説明文内から抽出することで，説明文と選択肢のペアを自動獲得する．図 2 に提案手法の概要を示す．本論文では，説明文中の各単語に，それが選択肢として適切かどうかのタグ付けを行い，そのタグ付きデータを機械学習のアルゴリズムである Support Vector Machines (SVM) に適用し，各単語が選択肢（キーワード）となり得るかどうかを分類するモデルを作成する．また，SVM の出力に対してルールベースのフィルタリングを適用することで精度向上を目指す．

[1] https://www.be-spoke.io/jp/

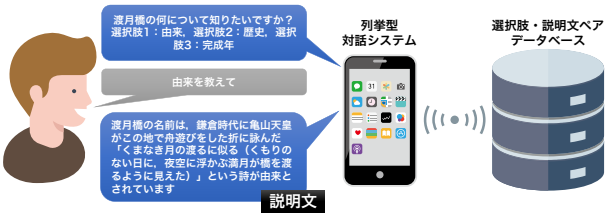

図 1: 列挙型対話システム．システムは選択肢・説明文ペアデータベースに基づき，選択肢を提示し，ユーザの選択に応じて説明文をシステムの応答文として出力する．

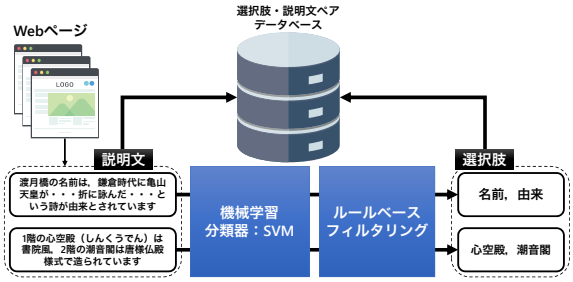

図 2: 提案手法の概要．Web から獲得した説明文から選択肢を自動抽出し，対話システムのための選択肢・説明文ペアデータベースを構築する．

2　関連研究

観光地に対する解説を行う対話システムの研究として，前述の翠らの研究 [2] や生田らの研究 [3] がある．翠らの研究では，Wikipedia の京都に関する文書を収集し，ユーザの発話に対して類似度の高い見出しとなる文書を検索し，システムの出力として，音声でその文書を読み上げる．また，ユーザから一定の時間入力がない場合，Wikipedia の階層構造またはユーザの表示履歴情報を利用してユーザの興味が近い文書を提示する．ユーザの発話に対して，システムが回答できない場合は，回答できないことをユーザに伝えて対処している．また，生田らは言語による情報集約を用い

表 1: 付与されるタグ例

単語	渡月橋	の	名前	は	、	鎌倉時代	に	亀山天皇	が	この
タグ	O	O	S	O	O	O	O	O	O	O
単語	1 階	の	心	空	殿	(	しん	くうでん	)	は
タグ	O	O	B	I	E	O	O	O	O	O

た観光案内システムを想定して，観光地に関するキーワードや観光状態など伝達したいコンテンツが断片的に定まっている際の文生成を行っている．文生成のために，観光案内文をクラウドソーシングを用いて収集し，ニューラル言語モデルで文の生成を行っている．これらの研究ではユーザは任意の入力を行うことが可能である．一方，本研究ではシステムが回答可能な事柄を提示することでシステムの頑健性が保証される列挙型対話システムを想定している．

本研究では回答となる文章から選択肢となりうるキーワードを抽出する．文章からキーワードを抽出する研究として，山田ら [6] や Lample ら [7]，遠藤ら [8] の研究がある．山田らは文中から SVM を用いて固有表現抽出規則の学習を単語の表層，品詞，文字種などを素性として行い，固有表現を抽出する実験を行っている．Lample らは LSTM-CRF モデルを用いて固有表現抽出を複数の言語で行っている．遠藤らは Web 上から観光情報収集を行い，収集した文から観光キーワードを抽出を行っている．その目的は辞書登録のない未知語を人手によるコストをかけず自動収集することである．本研究では文章から列挙型対話システムに利用する説明文と選択肢のペアデータを獲得することを目的にキーワードの抽出を行う．

3 データセット

本節では，本研究で使用するデータセットについて説明する．提案手法では機械学習を利用するため，訓練データが必要となる．訓練データ作成のためのアノテーション作業についても説明する．

3.1 対象データ

観光ガイドの書き起こしたデータおよび複数の京都の観光地について解説したサイトから説明文 1147 文を獲得した．収集に用いたサイトを以下に示す．

- KYOTOdesign[2]
- 京都観光オフィシャルサイト京都観光 Navi[3]
- 京都観光ネット[4]
- 京都の人気観光スポット TOP55[5]
- 京都観光のおすすめスポット・名所・人気コース京都旅行ガイド[6]

収集した観光地名とその説明文の例を以下に示す．

渡月橋： 渡月橋の名前は，鎌倉時代に亀山天皇がこの地で舟遊びをした折に詠んだ「くまなき月の渡るに似る（くもりのない日に，夜空に浮かぶ満月が橋を渡るように見えた）」という詩が由来とされています

銀閣寺： 1 階の心空殿（しんくうでん）は書院風，2 階の潮音閣は唐様仏殿様式で造られています

[2] https://kyoto-design.jp/
[3] https://ja.kyoto.travel/
[4] https://kyoto-kanko.net/
[5] https://tripnote.jp/kyoto/teiban-spot-kyoto
[6] https://tripnote.jp/kyoto/travel-guide

3.2 タグ付け

本研究の目的は，観光地の説明文から，その説明文をシステムの応答とした場合に提示すべき選択肢（キーワード）を自動抽出することである．たとえば，前述の「渡月橋」の例の場合，ユーザに「名前」や「由来」のような選択肢が提示され，その結果，その説明文が出てくることが自然である．以下はその対話の例である．

ユーザ： 渡月橋について知りたいです．

システム： 何が知りたいですか？
1: 由来，2: 歴史，3: 完成年

ユーザ： 由来について．

システム： 渡月橋の名前は、鎌倉時代に亀山天皇がこの地で舟遊びをした折に詠んだ「くまなき月の渡るに似る（くもりのない日に、夜空に浮かぶ満月が橋を渡るように見えた）」という詩が由来とされています．

すなわち，説明文から選択肢となり得るキーワードを抽出することが本研究のタスクである．文中のすべての単語がキーワードになり得るわけではなく[7]，選択肢として自然なものを選ぶ必要がある．そこで，説明文中の各単語に対して，それぞれが選択肢として適切かどうかのタグ付けをする．単語の区切りには形態素解析器 MeCab[8] を利用する．

区切られた単語について，BIOES 法 [9] に基づき，各単語が選択肢として適切かのタグ付けを行う．BIOES 法によるタグ付けの例について表 1 に示す．BIOES 法では，S，B，I，E，O の 5 つのタグを用いる．S はある一つの単語がキーワードである場合に用いる．たとえば，表 1 の例では「名前」がそれにあたる．B は二つ以上の単語でキーワードが構成される場合，その先頭単語に割り振り，E はその場合の最後の単語に割り当てられる．加えて，キーワードが 3 単語以上で構成される場合，B と E に挟まれた単語には I が付与される．表 1 では「心空殿」がこれにあたる．O はそれ以外の場所（すなわちキーワードではない単語）に付与される．すなわち，本研究では O タグが付いたもの以外を抽出することが目的となる．前述の 1147 文に対してこのタグ付け作業を本論文の第一著者 1 名で行った．

4 提案手法

前節で述べたように，説明文中のすべての単語が選択肢（キーワード）として適切であるわけではない．したがって，対話システムに利用する知識群を構築するためには，説明文中からキーワードを自動で抽出することが必要になる．本研究では，機械学習を利用し，このキーワード抽出のモデルを構築する．ここでは，各単語に 5 つのタグのどれかが割り当てられているという前提から，各単語のタグを分類

[7] たとえば上の例で「鎌倉時代」はこの「渡月橋」の説明文としては明らかに不適切．
[8] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

する問題として定式化する．また，機械学習モデルの出力に対してヒューリスティックなフィルタリングルールを適用することで精度向上を目指す．

4.1 分類モデル

分類器には Support Vector Machines (SVM)[10] を用いる．SVM に用いる素性は以下の通りである．

- word2vec[11] を用いて得られた単語の分散表現

  word2vec とは，大規模コーパスを用いた教師なし学習によって，各単語を表す固定長のベクトルを獲得するためのツールである．本研究では skip-gram モデルを用いて日本語 Wikipedia で学習した 200 次元の単語分散表現を利用する．単語分散表現がない単語に関しては 200 次元のゼロベクトルに設定した．

- 単語の品詞細分類

  固有名詞である単語は選択肢になる可能性が高く，助詞や動詞は選択肢になる可能性が低いと考えられる．そこで抽出される選択肢と品詞情報には関係性があると考え，素性の一つとして加える．説明文を MeCab を用いて形態素解析を行った際に得られる品詞細分類を利用した．品詞細分類とは，名詞-普通名詞，動詞-自立語-サ変などの品詞の分類である．

- 自立語に付く付属語の品詞情報

  文章の主語はその文の主題であり，選択肢として抽出される傾向にあった．そこで，自立語となる単語に付属する「は」「が」など付属語の品詞情報は有用な情報になると考え，素性の一つとして加える．

- 前後 2 単語の単語ベクトル

  文脈情報を考慮するため文章の i 番目の単語のタグの推定をするとき，$i-2$ から $i+2$ 番目までの各単語のベクトルを素性の一つとして加える．

- 推定した前 2 単語のタグ情報

  推定される単語のタグは，その直前の単語のタグ情報に依存する場合が多い．例えば抽出対象となる選択肢の先頭の単語に付与される B タグの直前のタグはかならず O タグが付与される．直前の単語のタグ情報を考慮するため文章の i 番目の単語のタグを推定するとき，SVM が推定した $i-2$，$i-1$ 番目の単語のタグ情報を素性の一つとして加える．

- 単語の位置情報

  文章中で重要な単語は文章の開始位置に近い位置で記述されることが多いと考えられる．その単語の出現順番を総単語数で割った値を素性の一つとして加える．

- 文中の各品詞の数

  文中に名詞が一つしかない場合はその単語が選択肢となる可能性が高いと考えられる．逆に，文中に大量の名詞が出現する場合，すべての名詞が選択肢となるとは限らない．このように文中の品詞の数は抽出される選択肢に影響を与えると考えられる．文中に出現する各品詞の数を素性の一つとして加える．

- 単語の TF-IDF 値

  TF-IDF 値は，文書内の単語の出現頻度を表す TF 値と，その単語が現れる文書頻度の逆数である IDF の積である．解析対象となる文章に特徴的な単語は選択肢となりやすいと考えられる．説明文 1 文を 1 文書として考えて TF-IDF 値を計算し，推定する単語の TF-IDF 値を素性の一つとして加える．

SVM は二値分類器であり，本タスクは BIOES の 5 つのタグを対象とした多値クラスの問題設定である．したがって，二値から多値への拡張が必要である．多値への拡張方法はいくつか存在するが，本論文では，one-vs-rest 法を採用する．one-vs-rest 法は，あるクラスかそれ以外かを分類する分類器をクラスの数だけ用意し，それぞれの結果（超平面からの距離）から総合的に最終的なクラスを決定する，という手法である．

4.2 フィルタリング

一般論として選択肢（キーワード）を考えた場合，それにはいくつかの前提や制約が考えられる．たとえば，接続詞など特定の品詞のみで構成されるキーワードというのは考えづらく，キーワードの先頭や末尾に助詞がくるというのも考えづらい．前節の分類モデルでは品詞や前後の単語の情報を利用しているが，それで十分に対応できる保証はない．一方で，これらの前提や制約は比較的明確であり，ルール化が可能である．そこで，SVM の出力に以下の処理（フィルタリング）を適用することで，出力結果の精錬化を図る．

- 代名詞，副詞，接続詞のみからなるキーワード

  たとえば,「単に/S 大文字/O 焼き/O と/O 呼ばれて/O ....」[9] のような例を考える．このとき「単に」は内容語でない．このように，代名詞，副詞，接続詞のみに S タグが付与される場合は適切ではない．その場合は S タグを O タグに変更する．すなわち，この例では 「単に/**O** 大文字/O 焼き/O と/O 呼ばれて/O ....」と変更する．

- 終了位置の単語の品詞が非自立の名詞または助詞

  たとえば,「当初/B の/E 寺地/O は/O ....」で「当初の」という単語群がキーワードとして抽出されたとする．この例では助詞である「の」がキーワードの末端であると判断されており，キーワードとしては不適切である[10]．このような場合，関連するすべての箇所を O タグに変更する．すなわち，この例では「当初/**O** の/**O** 寺地/O は/O ....」と変更する．

- 先頭に存在する接続詞・副詞

  たとえば,「ちなみに/B 三十三間堂/E は/O ....」という例を考えた場合,「ちなみに」に B タグが付与されるのは不自然である．このような場合，B タグを O タグに変更し，その次の単語が I タグの場合は B タグに，E タグの場合は S タグに変更をする．すなわち，この例では「ちなみに/**O** 三十三間堂/**S** は/O ....」と変更する．

5 実験

3 節で説明したデータセットを用いて提案手法を評価する．実験では 1147 文のうち 1019 文を訓練データとして使用し，128 文をテストデータとした．提案手法の有効性を検証するために，TF-IDF を用いた手法をベースラインとして比較

[9] SVM によって「単に」という単語に S タグ,「大文字」という単語に O タグが付与されたことを意味する．

[10] 非自立名詞（たとえば,「もの」など）については議論の余地が残るが，今回は経験的に非自立名詞もルールに組み込んだ．

表 2: 実験の結果.

手法	再現率	適合率	F 値
ベースライン	0.591	0.109	0.184
提案手法 (SVM のみ)	0.721	0.280	0.404
提案手法 （フィルタリングあり）	**0.732**	**0.290**	**0.416**

表 3: 実験結果の例.

事例 1	
説明文	傍らに薬師如来を安置し、当日は如来の水として一般に提供
正解	薬師如来, 如来の水
SVM のみ	傍ら, **薬師如来**, 安置, 当日, 如来, 水
フィルタリングあり	傍ら, **薬師如来**, 安置, 当日, 如来, 水
事例 2	
説明文	また土日祝のみ公開される、法堂天井の「八方睨みの龍」も必見です
正解	土日祝のみ公開, 八方睨みの龍
SVM のみ	**土日祝のみ公開**, 法堂天井の, 龍
フィルタリングあり	**土日祝のみ公開**, 龍

する．TF-IDF はキーワード抽出などで広く使われる重み付け手法である．このベースラインでは，TF-IDF 値が閾値以上であるものをキーワードとして抽出することとした．閾値は訓練データを用いて決定した．具体的にはいくつかの閾値を比較し，訓練データで最も精度が高くなった 4 とした．提案手法の実装には scikit-learn を用いた．SVM には scikit-learn の SVC を用い，カーネルは rbf，コストパラメータは 1 とし，gamma 値は 0.001 とした．one-vs-rest 法の実装にも同様に scikit-learn の OneVsRestClassifier を利用した．

実験結果を表 2 に示す．評価には再現率，適合率，F 値を用いた．実験結果より，単純なキーワード抽出手法（ベースライン）では十分に対応できないことがわかる．また，フィルタリングを用いることにより，精度向上が見られ，ルールベースの後処理が有効であることが確認された．ただし，全体の精度は F 値で 0.416 で十分とはいえない．一方で，本研究の目的は説明文からの選択肢の自動抽出であり，柔軟な対話システムの構築を考えた場合，できるだけ多くの選択肢を用意できることが望ましい．このような観点から考えれば，再現率が高いことが望ましく，提案手法には一定の有用性があると考えられる．

実験結果の具体的な例を表 3 に示す．事例 1 では，正解に「如来の水」が含まれているが，提案手法の出力では単語の分割処理によって 2 つに分けられた「如来」と「水」を正しく合わせて抽出できなかった．形態素解析の後処理や辞書などの工夫が必要である．事例 2 は，SVM は誤って「法堂天井の」という単語列をキーワードとして抽出しているが，フィルタリングの適用により，提案手法で正しく抽出された例である．このような事例が提案手法（フィルタリングあり）の精度向上に貢献している．

6 おわりに

本論文では，列挙型対話システムのための知識の自動獲得のために，機械学習に基づく説明文からの選択肢（キーワード）の抽出手法について提案した．提案手法はベースラインよりも有意に高い精度を得たが，まだ十分とはいえない．さらなる精度向上が必要である．また，今回の正解データの構築は著者一人の主観によって行われており，複数人による正解データの構築によるデータの頑健性についても今後の課題である．

謝辞

本研究は株式会社コンピュータサイエンス研究所からの受託研究「チャットボットにおける満足度向上のための選択肢提示方法の検討およびプロトタイプ研究開発」の成果の一部です．

参考文献

[1] 翠輝久, 水上悦雄, 柏岡秀紀ほか. 音声対話による観光案内システム assistra. 研究報告ヒューマンコンピュータインタラクション (HCI), No. 8, pp. 1–2, 2013.

[2] 翠輝久, 河原達也ほか. 限定されたドメインにおける質問応答機能を備えた文書検索・提示型対話システム. 情報処理学会研究報告音声言語情報処理 (SLP), No. 73 (2006-SLP-062), pp. 69–74, 2006.

[3] 生田和也, 品川政太朗, 吉野幸一郎, 鈴木優, 中村哲. 観光案内におけるニューラル言語モデルを用いた説明文の生成. 人工知能学会全国大会論文集 第 32 回全国大会, pp. 1333–1336, 2018.

[4] Bilyana Martinovsky and David Traum. The error is the clue: Breakdown in human-machine interaction. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[5] 松山匡子, 駒谷和範, 武田龍, 尾形哲也, 奥乃博ほか. バージイン許容音声対話システムにおけるユーザ発話の分析と指示対象同定への応用. 研究報告音声言語情報処理 (SLP), No. 21, pp. 1–6, 2010.

[6] 山田寛康, 工藤拓, 松本裕治ほか. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pp. 260–270, 2016.

[8] 遠藤雅樹, 横山昌平, 大野成義, 石川博. 特定地域に限定しない観光キーワードの自動抽出. In *DEIM Forum*, 2014.

[9] Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. Named entity extraction based on a maximum entropy model and transformation rules. In *Proceedings of the 38th ACL*, pp. 326–335, 2000.

[10] Vladimir N Vapnik. The nature of statistical learning. *Theory*, 1995.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.