

# 情報要求の曖昧性に基づき対話処理を行う情報検索対話システム

山崎 翔太<sup>1</sup> 中野 幹生<sup>2,3</sup> 船越 孝太郎<sup>3</sup> 菅野 重樹<sup>1</sup>

<sup>1</sup> 早稲田大学大学院 創造理工学研究科 <sup>2</sup> 早稲田大学 理工学術院

<sup>3</sup>(株) ホンダ・リサーチ・インスティテュート・ジャパン

## 1 はじめに

情報検索システムを使って何かを調べようとするとき、要求する情報のキーワードが自明でないことは多い。このとき、システム側から聞き返すことで意図を具体化することができれば、調べる対象に関する詳細な知識のないユーザでも必要な情報を見つけることができる。これは、対話によってシステムが知らない表現、抽象的な情報要求（クエリが不足しているので足りない情報を聞き返す）、状況やユーザの知識に関する情報不足（適合する検索結果の中でどの情報が適切かを聞き返す）などを解決することができるためである。特に、ユーザが詳細な知識を持っていない分野の情報収集を対象としたドメインで、Web などの未整理の情報資源を扱う場合には、曖昧性のある意図の表現が多くなり、対話によって意図を具体化することが有効だと考えられる。

このような、大規模な情報資源を対話的に検索する研究として、トラブルシューティングを扱った対話システム [5] やユーザに合わせた検索サービスを提供する対話エンジン [6] などがあるが、これらは定義されたルールに従って対話を行うため、大規模な情報資源を直接検索できる利点があるものの、不足情報の種類をある程度定義できる必要がある。従って、多様な情報を扱うドメインではルールの数を増やす必要があり、その結果複数のルールが重複して適合する場合にはその優先順位を逐一決めて処理を行う必要がある。一方で、対話のルールを定義しない方法では、コンテキストと検索結果から対話行動を出力する関数を機械学習する方法 [1] などが提案されているが、こうした方法では対話そのものの学習データが大量に必要な上に、対話行動の種類が制限されるため、やはり未知分野を扱うのは難しい。

本論文では、知識に基づいて多様な情報を扱う情報検索対話システム実現のために、階層的な知識構造に基づく確率的な意図推定により意図の曖昧性を解消する対話システムのアーキテクチャを提案する。

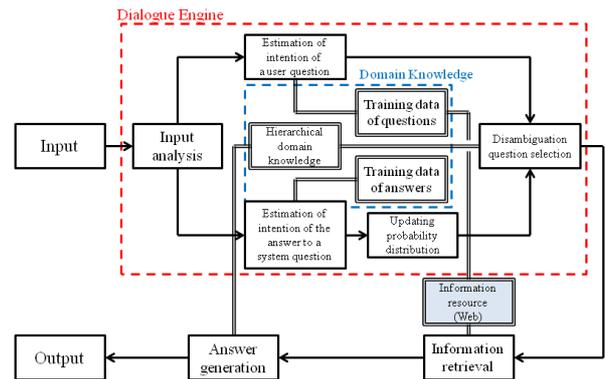


図 1: システム構成図

## 2 提案システム

図 1 は、提案システム全体の構成図である。ドメイン知識である階層構造の中には、ユーザが聞きたいこととその答え（または答えを検索するクエリ）がノードとして表現されている。

システムを構築するには、まず対象ドメインにおけるユーザの意図、すなわちユーザが聞き出したいことを階層的に構造化し、対応するユーザ質問文とシステム質問に対するユーザ応答文の学習データから意図の確率分類を機械学習する。このとき、用意した質問文から情報検索により学習データを集めることによって、大規模な情報資源を利用して確率分布を学習する。動作時には、自然言語の入力に対して意図の確率分布を推定し、曖昧である場合には聞き返すことで、その返答から階層構造を利用して確率分布を更新し、意図を具体化する。各発話ターンの対話状態に応じて、あらかじめ知識に組み込んである応答情報と、質問文の学習データから抽出したキーワードで情報資源を検索した結果を合わせた応答を出力する。以下に、知識構造、確率分布の学習と対話処理と確率更新について、それぞれの構成および処理を具体的に説明する。

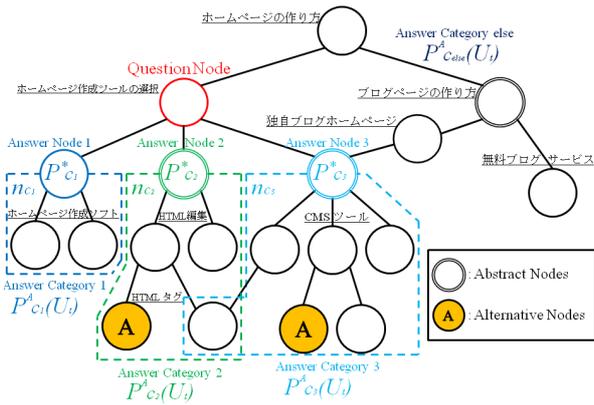


図 2: ドメイン知識の階層構造と確率更新

## 2.1 ドメイン知識

ドメイン知識として、ユーザの情報要求の意図を階層的に構造化し、意図をノードとして表現する(図 2 はドメインを「Web ページ作成」に関する情報とした例の一部である)。ノードは基本的に以下の要素から構成されている。

- <ID>: ノードの ID
  - <PARENT>: 上位ノードのリスト
  - <INFORMATION>: 知識に組み込むユーザへの  
 応答情報(テキストや URL など)
  - <DESCRIPTION>: 意図に対する応答情報の説明
  - <QUESTION>: 下位ノードを具体的に選択させる  
 ユーザへの聞き返し質問
- type="abs": 抽象ノード (Abstract Node) を表す属性

抽象ノードとは、例えば「ホームページの作り方を教えて」「HTML について知りたい」など、ノードが表現する意図が抽象的であり応答すべき情報が漠然として大きすぎるような場合を表す。

これらを XML で記述した知識の一例を図 3 に示す。この階層構造では、上位ノードの方が 1 つ下位のノードより知識を前提としない一般的な情報、時間的に前に当たる情報、抽象的な情報を持っている。上位ノードのリストに複数のノードを指定することを可能とすることで、知識の構造は完全な木構造ではなくネットワーク構造となる。また、各ノードの意図を表現するユーザ質問文と、上位ノードによりシステムが聞き返し質問をしてユーザから返答が得られた場合の意図を表現するユーザ応答文の学習データを用意し、これらを合わせてドメイン知識とする。

```

<NODE>
<ID> ホームページ作成ツールの選択 </ID>
<PARENT> ホームページの作り方 </PARENT>
<INFORMATION>
http://allabout.co.jp/gm/gc/31799/
</INFORMATION>
<DESCRIPTION>
どのような方法でホームページを作成するのがよいかを選ぶための
情報です。
ホームページ作成ツールを選択するための情報です。
</DESCRIPTION>
<QUESTION>
どのようなツールを使ってホームページを作っていますか？
・ホームページビルダーなどのホームページ作成ソフトを使っていますか？
・WordPress などの CMS ツールを使っていますか？
・直接 HTML を編集していますか？
</QUESTION>
</NODE>

```

学習データの元となる質問文の例

- ・ホームページを作るツールにはどんなものがありますか？
- ・ホームページを作るためのソフトを教えてください
- ・ホームページとブログの作成の仕方を教えてください
- ・ホームページの作り方を教えてください
- ・HP はどうやって作るんですか？
- ・Web ページを作るオススメの方法、ツール、ソフトは？

図 3: 知識構造の例

## 2.2 確率分布の学習

ユーザ質問文とユーザ応答文の学習データから、多クラス確率分類の手法を用いて、入力に対する意図の確率分布を学習する。今回は、Bag of Words を素性とし、ナイーブベイズの多項モデルを利用する。このとき、ユーザ応答文は、システムの聞き返し質問が知識の中にあり、ある程度ユーザの返答が予測できるため、用意したデータをそのまま学習に使う。一方で、ユーザ質問文の場合は、ユーザの意図を表すトピックを網羅できず確率値の精度が出ないため、人手で作った少数の質問文のみで確率分布を学習するのは適当ではない。そこで、擬似適合性フィードバック [3] を模した方法を利用して、大規模な情報資源から情報検索によってユーザ質問文の学習データを集める。

まず、用意した少数の質問文を形態素解析し、名詞集合  $S_n$  と他の自立語集合(今回は動詞と形容詞を用いる)  $S_v$  を抽出する。また、名詞が連続する場合は 1 つの複合名詞とみなして、 $S_n$  から複合名詞の集合  $S_c$  を作る。こうして、1 つの質問文について  $S_c \cup S_v$  から AND 検索のクエリを作り、情報資源(Web)に対してキーワード検索を行う。その検索結果の上位数件からタイトルとスニペットなどを抽出し、それらをユーザ質問文の学習データとする。このように学習データを集めることによって、トピックを広げること、また単語頻度の信頼性を上げることができる。また、1 つのクラス(意図)に属する質問文について、検索結果の重複を消去することで、用意した質問文の中に似たようなものが多くあることにより不適切に確率値が高くなることを防ぐことができると考えられる。

質問文と応答文の各学習データについて形態素解析を行い、 $S_n \cup S_v$  を抽出してこれを素性集合とする。これらにナイーブベイズの多項モデルを適用すると、ユーザの発話  $U$  が入力されたときの、各ノードが表現するユーザの意図  $I_i$  の確率分布  $P(I_i|U)$  は、

$$P(I_i|U) = \frac{P(I_i)P(U|I_i)}{\sum_{k=1}^N P(I_k)P(U|I_k)} \quad (1)$$

$$P(I_i) = \frac{1}{N} \quad (2)$$

$$P(U|I_i) = \prod_{w_j \in U} q_{w_j, I_i}^{n_{w_j, U}} \quad (3)$$

と表すことが出来る。ここで  $N$  は全ノードの数であり、 $n_{w, X}$  は文字列  $X$  に単語  $w$  が属する頻度である。この通り、今回は意図の事前確率を平均とし、事前分布を与えていない。また、 $q_{w, I}$  は意図  $I$  に属する単語の出現確率であり、 $|W|$  を総語彙数、 $\alpha$  を定数として、

$$q_{w, I} = \frac{n_{w, I} + \alpha}{\sum_I n_{w, I} + |W|\alpha} \quad (4)$$

を学習データから計算しておく（加算スムージング）。

### 2.3 対話処理と確率更新

求めた確率分布に基づいて対話処理を行い、曖昧性を解消するために得られたユーザ応答から確率分布を更新する処理について説明する。階層構造を知識として用いる研究は多く、例えば [2] では、木構造のベイジアンネットワークを用いて、対話によって割り当てられた属性を埋めていくモデルが提案されているが、これは全てのノードの値の組を求める問題であるため、中間層を含めて1つのノードを特定したい本問題とは性質が異なる。そこでこの対話システムでは、階層構造を利用してノードのクラス分類することで、確率分布を更新するアルゴリズムを提案する。

ユーザの質問入力に対して学習結果から意図の確率分布  $P(I_i|U_1)$  を推定し、その分布から曖昧性を評価する。最大確率のノードとその他のノードの確率値の差が大きい場合には、最尤と推定されたノードで応答するが、それが抽象ノードである場合には聞き返し質問を返し、意図を具体化する。一方、差が小さい場合には、確率値の高いノード群 (Alternative Nodes : AN) を定め、それらの上位ノードを辿ることで共通する上位ノードを聞き返しノード (Question Node : QN) として特定し、このノードが持っている聞き返し質問を返す。それに対するユーザの返答結果から、聞き返しノードの子ノードの確率分布  $P_{C_m}^*(U_t)$  を推定し、以

表 1: 対話状態とシステム応答

Dialog State	状態の説明	応答処理
Initial	初期値または対話履歴消去後	質問の受付
Unknown	該当する意図なし	分からない旨を返答
Specified	該当する意図が唯一つで抽象ノードではない	意図に応じて返答し対話終了
Abstract	該当する意図が唯一つで抽象ノードである	意図に応じて返答し聞き返し質問を返す確率分布をリセット
Ambiguous1	Alternative Nodes として複数の意図が該当 (AN に QN は含まれない)	Question Node の聞き返し質問を返す (情報返答なし)
Ambiguous2	Alternative Nodes として複数の意図が該当 (AN に QN が含まれる)	Question Node の聞き返し質問を返す (情報返答あり)

下の式により確率分布を更新する。

$$P_{I_i \in C_j}^A(I_i|U_t) = \frac{1}{\sum_{m=1}^{|C|} n_{C_m} \frac{P_{C_m}^*(U_t)}{P_{C_j}^*(U_t)}} \quad (5)$$

$$P(I_i|U_{t+1}) = \frac{P(I_i|U_t)P^A(I_i|U_{t+1})}{\sum_{k=1}^N P(I_k|U_t)P^A(I_k|U_{t+1})} \quad (6)$$

確率更新では、知識構造を利用してノードを各子ノードの下位に属すカテゴリー  $C$  に区分する (図 2)。ここで、 $U_t$  はターン  $t$  におけるユーザの発話、 $I_i$  は各ノードが表現するユーザの意図を表し、 $n_{C_m}$  はカテゴリー  $C_m$  に属するノードの数である。

各発話ターンにおいて、対話状態 (Dialog State) を設定し、それぞれシステムの応答の仕方を変更する。その設定を表 1 に示す。意図に応じてシステムが返答をするとは、要求された情報の応答情報と検索結果を返すことを意味し、聞き返し質問は返答とは独立したシステムの発話として出力される。

### 3 動作検証

提案アーキテクチャに基づき、ドメインとして「Web ページ作成」に関する知識を構造化してシステムを構築した。図 4 は、抽象的な質問に対するシステムの対話動作の例である。また図 4 は、情報不足な質問により曖昧性が生じた場合のシステムの対話動作の例であり、図 6 はその時の意図の確率分布の変化である。図 6 における  $H$  は確率分布のエントロピーであり、対話によりばらつきが小さくなることを示している。

このように提案システムでは、「ホームページの作り方を教えて」といった抽象的な情報要求に対してホームページの形式や作成ツールの選択を聞き返したり、「フォームを作りたい」などの情報不足な要求に対して必要情報のレベルや使用ツールなど詳細な情報を聞き返すことで、意図を具体化することが出来る。

U : ユーザ S : システム

- U : ホームページの作り方を教えて  
 S : 応答情報「ホームページの作り方」(右上图)  
 作りたいホームページの形式は決まっていますか?  
 ・静的なホームページ(あまり更新しない)  
 ・ブログ形式のホームページ(更新が容易)  
 あるいは、どのようなツールを使ってホームページを作るのかを選択するための情報が必要ですか?  
 U : ブログ形式のホームページ  
 S : 応答情報「ブログホームページ」(右下图)  
 アメーバブログなどの無料ブログサービスを利用しますか?  
 WordPressなどのCMSツールを使って自分だけのブログホームページを作成しますか?  
 U : 無料のサービスを利用する  
 S : 応答情報「無料ブログサービス」



図 4: 対話動作例(抽象的な質問)

U : ユーザ S : システム

- U : フォームを作りたい  
 S : どのようなツールを使ってホームページを作っていますか?  
 ・ホームページビルダーなどのホームページ作成ソフトを使っていますか?  
 ・WordPressなどのCMSツールを使っていますか?  
 ・直接HTMLを編集していますか?  
 U : ワードプレスを使っている  
 S : 応答情報「WordPressで問い合わせフォームを設置する」(右図)



U : ユーザ S : システム

- U : フォームを作りたい  
 S : どのようなツールを使ってホームページを作っていますか?  
 ・ホームページビルダーなどのホームページ作成ソフトを使っていますか?  
 ・WordPressなどのCMSツールを使っていますか?  
 ・直接HTMLを編集していますか?  
 U : HTMLを編集する  
 S : 応答情報「HTMLタグ<form>に関する情報」(右図)



図 5: 対話動作例(情報不足な質問)

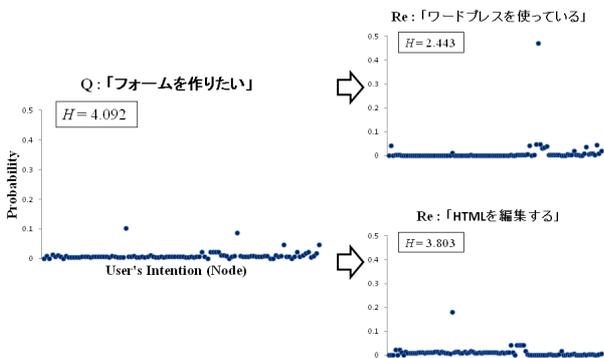


図 6: 確率分布の変化

## 4 まとめ

本論文では、ユーザが詳細な知識を持っていないようなドメインで多様な情報を扱える情報検索対話システム実現のために、階層的な知識構造に基づく確率的な意図推定によって発話意図の曖昧性を解消する対話システムのアーキテクチャを提案した。このアーキテクチャの利点は、緩い関係を定義としてドメインの知識をおおまかに構造化できること、また確率的な表現を用いることで曖昧性に応じた様々な対話処理を簡単に実装できることである。

今後は、実際にユーザに利用してもらうことによる比較評価や、より人手で作る知識を少なくするために大規模な情報資源から直接曖昧性を計算したり、自動的に聞き返しを生成する枠組みの構成などを行う。また、意図の確率分布の学習精度の向上に関して、ユーザの履歴を利用する方法や、WebのQ&Aサイトからの類似質問検索[4]などを応用する方法を検討する。さらに、ドメイン外の質問に対する応答も検討する。

## 参考文献

- [1] Matthias Denecke and Norihito Yasuda. Does this answer your Question? Towards Dialogue Management for Restricted Domain Question Answering Systems. In *SIGDial*, pp. 65–76, 2005.
- [2] Neville Mehta, Rakesh Gupta, Antoine Raux, Deepak Ramachandran, and Stefan Krawczyk. Probabilistic Ontology Trees for Belief Tracking in Dialog Systems. In *SIGDial*, pp. 37–46, 2010.
- [3] Jinxi Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis. In *SIGIR*, pp. 4–11, 1996.
- [4] Xiaobing Xue, Jiwoon Jeon, and W. Bruce. Retrieval Models for Question and Answer Archives. In *SIGIR*, pp. 475–482, 2008.
- [5] 清田陽司, 黒橋禎夫, 木戸冬子. 大規模テキスト知識ベースに基づく自動質問応答-ダイアログナビ-. 自然言語処理, Vol. 10, No. 4, pp. 145–176, 2003.
- [6] 北村美穂子, 下畑さより, 介弘達哉, 池野篤司, 坂本仁, 折原幾夫, 村田稔樹. ラダリング型検索サービスのための対話エンジンの設計・開発. 情報処理学会自然言語処理研究報告, Vol. 2008, No. 67, pp. 97–102, 2008.