

図書館の自動リファレンス・サービス・システムの構築

平田 大志[†] 日笠 亘[†] 藤井 綱貴[‡] 黒橋 禎夫[†]
 Taishi Hirata Wataru Higasa Tsunaki Fujii Sadao Kurohashi

[†]京都大学大学院 情報学研究科 [‡]京都大学工学部
 {taishi, higasa, fujii, kuro}@pine.kuee.kyoto-u.ac.jp

1 はじめに

図書館においてリファレンス・サービス(参考調査)は重要な業務の一つであり、電子図書館などの普及にともない、その重要性は今後ますます高まると考えられる。我々は、これまでに京都大学総合情報メディアセンターの対話的ヘルプシステムを構築し、実際のサービスを行っている。このシステムは自然言語によって知識を記述しておけば、それを参照して質問に自動応答するシステムで、対話管理プログラムと知識ベースは独立したものとなっている。そこで、この対話管理プログラムを利用し、リファレンス・サービス用の知識を(自然言語で)与えることにより、リファレンス・サービス・システムを構築した。

2 システムの概要

本論文で述べる自動リファレンス・サービス・システムは、我々がこれまでに作成し、京都大学総合情報メディアセンターにおいて運用を行っている対話的ヘルプシステム [1] を拡張したものである。このシステムは、対話管理プログラムと知識ベースが独立したものになっており、今回行ったのは、プログラム(4節で述べるNDLC検索モジュール)の追加とリファレンス・サービスのための知識ベースの構築である。

システムの概要を図1に示す。このシステムの特徴は、ドメインの知識を辞典のような形式で自然言語で与えることができるという点にある(図1の知識ベース)。システムは、ユーザの質問と知識ベース中の記述との柔軟なマッチングを行い¹、質問に対して最も適切な知識を見つけて回答を行う。さらに、知識ベースを単に検索対象として用いるだけではなく、それによって文脈に応じた発話の解釈、種々の聞き返しなどの処理を実現している²。

リファレンス・サービスのための知識ベースは、京都大学附属図書館が発行している参考調査に関する手引書(ac-

cess.txt)などを参考にして人手で作成した。現在の見出し語数は約200項目である。各見出し語にはその上位語を与えており、その上位下位関係をまとめるとリファレンス・サービスのドメインのオントロジーをえることができる。現在の知識ベースから構成されるドメイン・オントロジーを図2に示す。

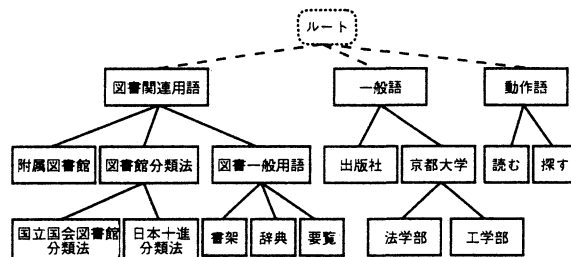


図2: リファレンス・サービス・システムのオントロジー

3 リファレンス・サービス

3.1 リファレンス・サービスの内容

図書館のリファレンス・サービス(参考調査係)には、非常に一般的で広範な質問がよせられる。たとえば次のような質問がある。

1. 日本の水力発電について調べたい
2. 江戸時代の農民一揆に関する本を探したい
3. ATMとは何の略ですか
4. 横浜国立大学の住所の調べ方
5. 伊藤博文はどんな人ですか
6. 磐梯山の高さはどのくらいですか
7. OPACの使い方を教えてください

リファレンス・サービスを実際に行っている図書館司書は、このような広範な質問のすべてに対して個別の知識を有しているわけではない。図書館司書が行っていること

¹文間のマッチングは、構文解析された木構造問で行い、自立語に3点(句・文節内で2語目以降は1点)、表解格等の一致に1点、文末からのノード位置の一致に1点という形で実装している。詳細は[1]

²リファレンス・サービス・システムの対話例を本稿の末尾に示す。現在のシステムが対応できる種々の現象が含まれている。

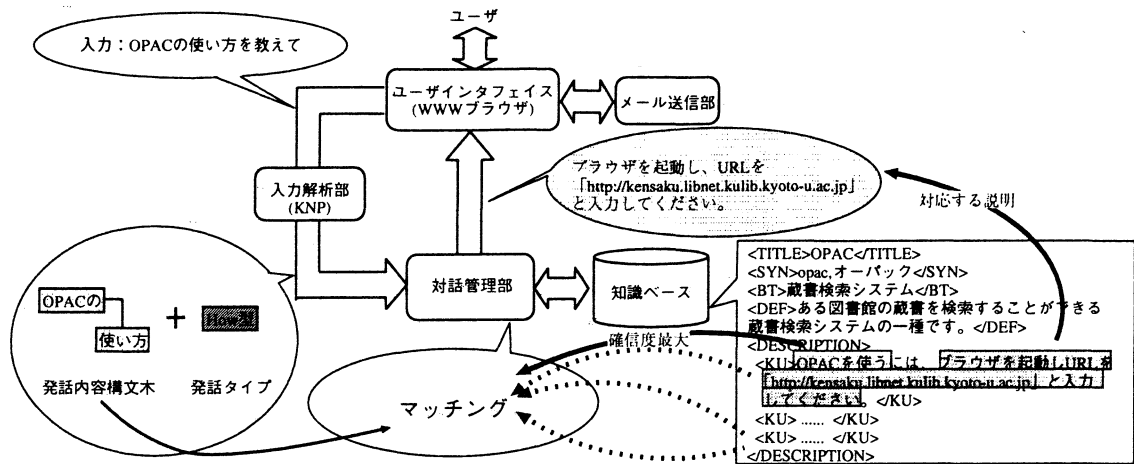


図 1: リファレンス・サービス・システムの概要

は、多くの場合、質問を一般化して解釈することにより、その質問内容の調べ方を回答することである。

典型的な例は質問 1,2 のような場合で、質問内容から図書分類のカテゴリーへの一般化を行い、図書館におけるそのカテゴリーの書架の位置を答える。たとえば、質問 1 に対しては「経済・産業の本は [D] の書架にありますので、そこを探してみてください」というように回答することができる。他の質問についても、略語、住所、人、地理情報の調べ方というように一般化して、略語辞典、全国学校総覧、日本人名典拠録、山名辞典などの存在を回答することができる。

もちろん、質問によってはより具体的な回答が必要となる場合もある。たとえば質問 7 の場合には、OPAC の具体的な使い方、すなわちその URL などを回答すべきであろう。

3.2 リファレンス・サービスの自動化の要点

このようなリファレンス・サービスの自動化を考えるうえでは、上記のような「質問の一般化した解釈」を実現することがキーとなる。我々が土台とする対話的ヘルプシステムは、知識を自然言語で記述しておき、これとユーザ質問との柔軟なマッチングを行うことにより回答を行う。「質問の一般化した解釈」は、この柔軟なマッチング機能においてある程度実現される。さらに、リファレンス・サービスにおいて頻繁に尋ねられる「あるトピックの本」という質問に答えるために、図書分類の情報を用いることによる質問の一般化解釈の機能を追加した。

図 3 にいくつかの例を示す。知識ベースの附属図書館の項目には、図書分類の上位下位構造のあるレベルのカテ

ゴリーに対して図書館中の書架の位置を記述しておく。一方、質問中の語句については、図書分類のデータベースを用いることによって、書架情報を与えたレベルへの一般化を行う。「日本の水力発電」の場合は、これが「経済・産業」に属することがわかるので、これと知識ベースとのマッチングをとることにより、質問に対して「図書館 2 階の [D] の書架にありますので、そこを探してみてください」という回答が可能となる。

一方、「化学系の事典類を見たい」の例では、「化学系の事典類」そのものについての情報が知識ベースにあり、これとのマッチングが一般化を行った「化学系・その他の工学」とのマッチングよりも優先される。これによって、より詳細な回答を行うことができる。

「横浜国立大学の住所の調べ方」の例では、住所に関する資料として「全国大学総覧」「官公庁要覧」「日本地名百科事典」などがあるが、この場合は複合名詞「横浜国立大学」と「大学」とのマッチングのスコア計算の柔軟さにより、「全国大学総覧」が選ばれる。

なお、現在のシステムでは、複合語間のマッチング、構文木間のマッチングなどに柔軟さがあるものの、自然言語表現の自由度、パラフレーズなどの関係を十分に扱うことはできない。そのため、知識ベースの表現は、質問の表現をある程度想定したものとなっており、不自然な表現もある。将来的に言語処理技術が進めば、表現間のずれをより柔軟に吸収することが可能となり、知識ベースの表現の不自然さの問題は解消していくと考えられる。

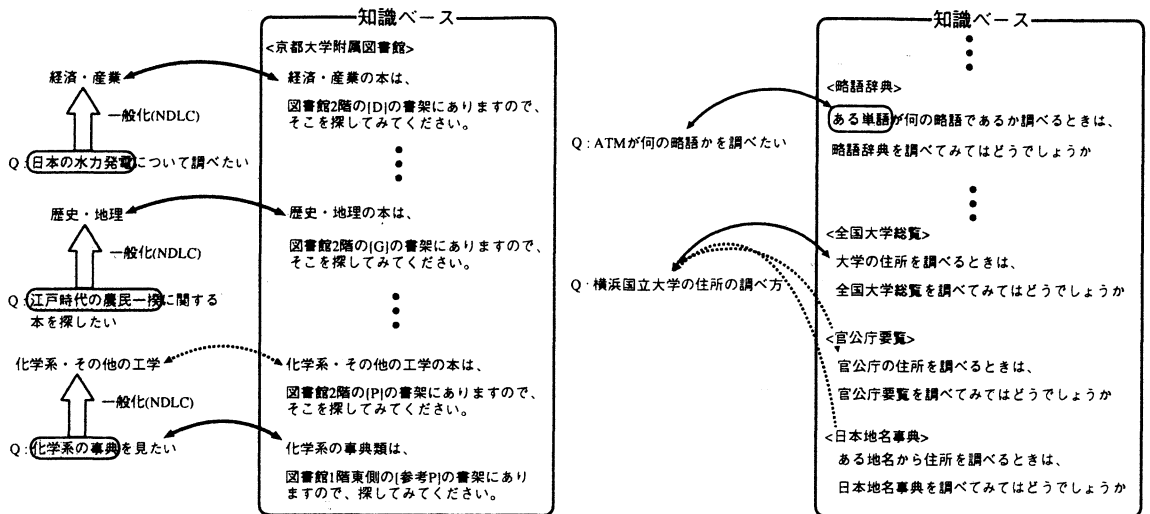


図 3: 入力と知識との柔軟なマッチング

4 NDLC 検索モジュール

4.1 国立国会図書館分類法

京都大学附属図書館では、図書館資料の分類・配置に国立国会図書館分類法 (NDLC)[2][3] を採用している。これらについて OCR(光学読み取り機) を利用して計算機可読な状態としたのち、人手で修正を行い、NDLC データベースを作成した。

このデータベースは NDLC の分類番号をノードとする木構造であり、木構造内の上位下位関係は図書分類の上位下位関係に対応する。ノードは、さらに見出し語、英表記、関連キーワードなどの情報をもつ。ノード数(分類番号数)は 7700、関連キーワード数は約 1 万、木構造の深さは最高 9 段である。

4.2 最上位カテゴリの推定

京都大学附属図書館における書架分類は、NDLC の最上位カテゴリ (15 分類) に対応して行われている。そこで、図 3 のような処理を行うために、入力質問の内容を NDLC の最上位カテゴリに対応付ける次のようなモジュールを作成した。

- NDLC 中のキーワードの重みを、キーワードの出現カテゴリ数を df 、総カテゴリ数を $N (=15)$ として、

$$-\log \frac{df}{N}$$

とする (これは情報検索で広く用いられている idf に相当する)。

- 入力質問中に含まれる自立語 (助詞・助詞相当語句以

外) について、NDLC のデータベースに出現するものだけを取り出す。

- 取り出されたキーワードのうち、半数以上が同時に出現するカテゴリを処理対象とする (半数以上が出現するカテゴリがなければ探索は失敗)
- 2 の各カテゴリについて、そのカテゴリで出現するキーワードの重みの総和をそのカテゴリに対するスコアとする
- 3 で求めたカテゴリのスコアで、最大値をとるカテゴリと、最大値のスコアから 90 % 以内のスコアをとるカテゴリを、結果として返す

図 3 の「日本の水力発電について調べたい」の場合は、水力発電は「経済・産業」と「機械系」の 2 カテゴリーに出現するので重み 2.9、日本は「経済・産業」「社会・労働」「教育」など 12 カテゴリーに出現するので重み 0.3 となる。この結果スコア最大のカテゴリ「経済・産業」と、その 90 % 以上のスコアをもつ「機械系」が推定結果となる。

4.3 分類番号の推定

また、入力質問の内容によって、最上位カテゴリだけでなくカテゴリ中でのより詳細な分類番号を推定できれば、質問者が実際に書架を探す際のよりよい手助けとなる。そこで、入力キーワード集合に対する適切な分類番号を以下の方法で推定する (図 4)。

推定された最上位カテゴリの中で、そのカテゴリに出現するキーワードの異なりを考える (図

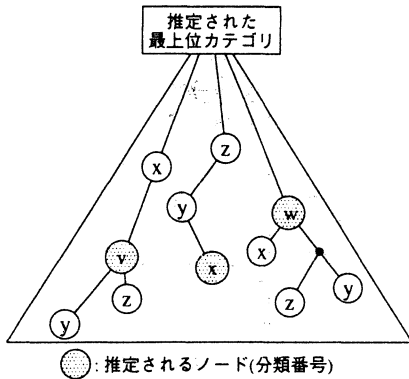


図 4: 分類番号の推定の例 (キーワード数 3)

4では x, y, z の 3 つ)。先祖ノードまたは子孫ノードにそれらをすべて含むノードのうち、最も下位の (深い位置にある) ノードを求める。

図 4 の例では、 v, w, x のノードが求められ、それらのノードに与えられた分類番号が推定結果となる。

図 3 の「江戸時代の農民一揆に関する本を探したい」の場合には、「江戸時代」「農民一揆」から、既に最上位カテゴリ「歴史・地理」が推定されているので、そのカテゴリ内で探索を行うと、「農民一揆 (分類番号: GB367)」が探索結果として一意に得られる。この結果、「図書館 2 階の [D] の書架にありますので、そこを探してみてください。特に分類番号 GB367 を探してみてくださいか」というような回答することができる。

なお、本節で述べた図書分類に関する処理は、個別の分類法に依存したものではないので、他の分類法に基づいた他の図書館に対しても適用可能である。

5 おわりに

本稿では、対話的リファレンス・システムの構築について述べた。現在、このシステムは研究室内で試験運用されているが、今後、京都大学附属図書館において実際に運用していくことを予定している。システムを実際に運用していくことで、対話データを解析・評価して新たな対話の現象を捉え、システムを徐々に高度なものにしていくのが今後の課題である。

また、3.2 節のように、調査方法等のみを回答するだけでなく、質問内容について WWW 上のテキストを自動検索し、質問の意図に対応した回答を生成するような機能の実現も行っていく予定である。

個別事例とのマッチング

- U1: OPAC の使い方を教えてください。
 S1: 端末のブラウザを立ち上げて、URL を「http://kensaku.libnet.kulib.kyoto-u.ac.jp/」と入力してください。
 U2: カードの引き方がわからない。
 S2: 和書は書名の五十音順で配列、同一書名の中は著者の五十音順、同一書名、同一著者の場合は出版者の五十音順で並べられています。洋書は著者名から引いてください。
 U3: カードの赤い印の意味は？
 S3: 赤印は廃棄・除籍・焼失・管理換などのため、現在は所蔵していないことを示しています。

柔軟なマッチング

- U1: 日本の水力発電について知りたいのですが。
 S1: 経済・産業の本は [D] の書架にありますので、そこを探してみてください。
 U2: 磐梯山の高さを調べたい。
 S2: 山の高さを調べるならば、山名辞典を調べてみてはどうでしょうか。

曖昧な質問に対する聞き返し

- U1: 京都大学の学位論文を探しています
 S1: “博士論文” “修士論文” から : 学位論文を選んでください。
 U2: 博士論文です。
 S2: 附属図書館に所蔵していますので、カード目録コーナーの学位論文のボックスを検索してください。

文脈に応じた発話の理解

- U1: 附属図書館はどこにありますか？
 S1: 京都大学本部キャンパスの南西端にあります。
 U2: 採用されている図書館分類法は何ですか。
 S2: 京都大学附属図書館では、基本的に国立国会図書館分類法に基づいて書籍等の分類を行っています。

参考文献

- [1] 日笠亘, 藤井綱貴, 黒橋禎夫: 入力質問と知識表現の柔軟なマッチングによる対話的ヘルプシステムの構築, 情報処理学会 第 134 回自然言語処理研究会, 1999.10
- [2] 国立国会図書館図書部 (編): 国立国会図書館分類法, 国立国会図書館 (1987).
- [3] 国立国会図書館図書部 (編): 国立国会図書館分類法 総索引, 国立国会図書館 (1987).