対話による高度情報検索システムの構築

藤崎 博也[1] 亀田 弘之[2] 田島 研[1] 大野 澄雄[1]
[1] 東京理科大学 [2] 東京工科大学

1 はじめに

最近、インターネットの普及により、誰もが情報を発信することが可能になり、大量の情報が流通している。従って、これらの多種多様な情報から必要な情報を機械で検索して抽出・利用できれば、ユーザに多大な貢献をするであろう。しかし、真に必要なものを迅速かつ的確に抽出・利用することができなければ、それは無に等しいか、場合によっては却って有害でさえある。実際、現在の情報検索システムでは、ユーザの意図を正確に情報検索処理に反映することができず、検索洩れや、不要な情報の抽出が起こる。この問題を解決するためには、人間にとって最も自然かつ柔軟に意思を表現・伝達できるとともに、情報の相互交換の上で最も効率的な手段である音声言語による対話 [1] を用いて、ユーザの意図を正確に処理に反映できる情報検索システムを構築する必要がある。

本稿では、情報検索を目的とした対話を分析して、キーワード抽出における規則を調べた。その規則に従って、情報検索を行うシステムとユーザとの間の対話において、ユーザの発話からキーワードを抽出する実験を行った。次に、キーワードの表記のみに着目して処理を行うと、同表記異義・異表記同義の存在が検索性能の低下をもたらすため、キーワードの概念 (キー概念)[2] を導入し、検索性能の向上を目指した音声対話による新しい情報検索システムの構想を示し、その機能について述べる。

2 キーワード抽出

2.1 資料とした音声対話

本研究で資料とした音声対話は、日本音響学会研究用連続音声データベースに収録されている模擬対話「観光・旅行案内」(4対話、467発話、総文字9,389文字) と、Wizard-of-Oz 方式で作成したハンドボールに関する対話 (5対話、72発話、総文字2,142文字) とである。

2.2 ユーザ側の発話からのキーワード抽出のための規則

キーワードを抽出する規則を作成するために、まず、対話データから人間がキーワードを抽出し、これらのキーワードを選んだ理由について分析した。その結果、キーワードとして抽出されたものには、以下のいずれかの性質があることが明らかとなった。

- 固有的な表現のもの (地名、建造物、など)
- 一般的なものごとの名称を表すもの (ハンドボール、試合、歴史、日本、など)
- 時間、数量を表すもの (二日、17時、4人、など)
- キーワードに連なって、キーワードを修飾するもの (強い、有名な、など)
- 応答を表すもの (いえ、はい、そう、など)

これらの規則に従って選定されたものをキーワードとして自動抽出した結果を次に述べる。

2.3 実験結果

観光・旅行案内に関する対話と、ハンドボールに関する対話について、キーワード抽出実験を行った。キーワードの抽出個総数を表1に示す。ここで、適切とは、機械が抽出したキーワードと人間が抽出したキーワードが同じもの、不要とは、人間は抽出しなかったキーワードを機械が抽出してしまったもの、そして不足とは、人間が抽出したキーワードを機械が抽出できなかったもののことであり、表中の数値はこれらの延べ個数を示している。また、観光・旅行案内について、ユーザ発話と機械が自動抽出したキーワードとの対応の一部の例を表2に示す。

表1 キーワード抽出個数

対話データ	適切	不要	不足
観光・旅行案内	395	90	112
ハンドボール	100	29	13

表2 観光・旅行案内に関する対話のユーザ発話と機械が抽出したキーワード

ユーザの発話	機械が抽出したキーワード
二日ぐらいで、ちょっと京都をみてまわりたいんですけれども、なにかないですか。	二日, 京都, なにか
[うん] そうですね。社寺庭園コースでお願いします。	そう, 社寺, 庭園, コース
何か適当なのを捜して下さい。	何, 適当なの
[ええ、] さっきの寺の話ですけど、ぜんぶ説明してくれますか。	さっき, 寺, 話, ぜんぶ
そこでも良いです。	そこ, 良い
[ええ、] 二人で福永です。	二人, 福永
[ああ、] いいですね。で、さっきの、来迎院とか言うのは、何か有名なのがあるんですか。	いい, さっき, 来迎院, 何か, 有名なの
[ああ、] じゃあそのあたりをまわってみます。[ええ、] 昼ご飯を食べるのに何かよいところはあるんですか。	そのあたり, 昼, ご飯, 何か, よいところ

2.4 考察

L:機械が自動抽出したすべてのキーワード個数
M:人間が抽出した正しいキーワード個数
N:機械が自動抽出した適切なキーワード個数
とすると、これらの包含関係は以下のようになる。

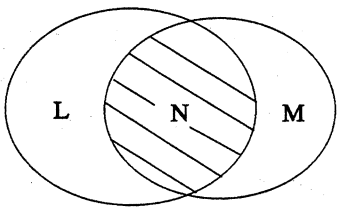

図1 包含関係

これらの値から、以下のような評価を行った。

$$\text{キーワード適合率 [\%]} = \frac{N}{L} \times 100 \quad (1)$$

$$\text{キーワード再現率 [\%]} = \frac{N}{M} \times 100 \quad (2)$$

キーワード適合率は不要なキーワードの抽出の少なさを、キーワード再現率はキーワードの抽出洩れの少なさを示す指標となる。

観光・旅行案内に関する対話では、キーワード適合率81.4%、キーワード再現率77.9%、ハンドボールに関する対話では、キーワード適合率77.5%、キーワード再現率88.5%と、第2.2節に示した規則を用いて、キーワードを抽出することが有効であることが示された。

本稿で用いた規則が有効であった理由について考察する。ここで用いた対話データが情報検索に限られており、ユーザは情報を要求する立場、システムは情報を提供する立場、というような原則がある。これは、述部の情報によらずに、ユーザの意図が推測できることを示している。例えば、

「ハンドボールの特徴を( )」

というユーザの発話では、( )の記述を見なくても、ユーザの目的が情報を引き出すことにあることから、「知りたくない」などの否定要素や「教えます」といった情報の要求と提供の立場の交代は原則的に起こらない。つまり、( )の記述には「調べている」や「教えて欲しい」などが当てはまり、システムはハンドボールの特徴を呈示するための作業をすればよいことが明らかである。

しかし、キーワードがでてこなかった例として、

回答者：ここも、見ますか。

質問者：車ではなくて、歩いていけますか。

回答者：バス停から２０分ぐらいですね。

質問者：そうですか。[ええ] それと、あと何でしたっけ。

という下線部の発話があった。

下線の発話では、キーワードとして「そう」と「ここ」が抽出された。しかし、この発話で重要なのは「歩いていけるか」どうかである。このような述部的な情報のみにキーワードが含まれている場合には、キーワードとして述部を抽出する必要がある。

さらに、キーワードが「それ」や「ここ」などの指示語である照応や、キーワードの省略が起きる場合がある。これは音声対話の迅速性による制約から、発話が冗長でないように起こる現象である。従って、ユーザ発話からキーワードを抽出する際には、ユーザの発話における省略・照応現象が指し示す語や句を特定する処理を行なうことが必要となってくる。

以上のように、ユーザの発話からキーワードを抽出する実験を行った。しかし、キーワードによる従来の情報検索では、検索洩れや不要な情報の抽出が起こり、ユーザの意図を正確に情報検索処理に反映することができない。従って、第3章において、このような問題を解決するため、音声対話を用いた新たな情報検索システムを提案する。

3 高度情報検索システムの提案

キーワードによる従来の情報検索では、キーワードの表記のみに着目して処理するため、同表記異義・異表記同義の存在が検索性能の低下をもたらす。これを避けるには、キーワードの概念(キー概念)を用いることが有効であるが[2]、さらにキーワードがシステムの辞書に登録されていない語、すなわち未知語の場合には、その処理が必要となる。また、情報検索のより一層の高度化を実現するためには、検索効率を向上させるための様々な知識を、システムが自動的に獲得する必要がある。このような見地に基づき、本稿では、キー概念の抽出・未知語処理・知識獲得を組み合わせた情報検索の新しい方式[3]の構想を示す。

3.1 情報検索の高度化

一般に語は表記(表層表現)と概念(深層表現)から構成される。語に同表記異義・異表記同義がある場合の表記と概念の関係を図2に示す。

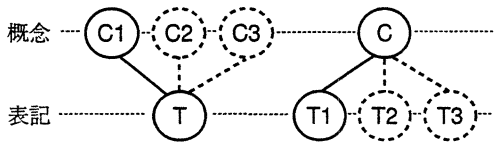

図2 同表記異義・異表記同義がある場合の表記と概念の関係

従来のキーワード検索では、語の表記のみに着目するため、キーワードに同表記異義が存在する場合には、ユーザが呈示した表記Tにより検索し概念C1に関する情報を取り出そうとすると、概念C2, C3に関する不要な情報まで取り出してしまう(図2(a))。また、キーワードに異表記同義が存在する場合には、ユーザが呈示した表記T1では、概念Cに関わる情報のうち、表記T2, T3の形式に言語化されたものは取り出すことができない(図2(b))。すなわち、同表記異義の存在は不要な検索をもたらし、異表記同義の存在は検索洩れをもたらす。これらを避けるには、キー概念のレベルにまで遡った検索が必要である。

しかし、キーワードが未知語の場合には、キー概念を抽出することができない。したがって、未知語の処理も合わせて行う必要がある。

また、的確で効率の良い検索を行うためには、ユーザやデータベースの特徴に関する知識が必要となるが、それらをシステムに予め与えておくことは不可能であり、システムに自動獲得させる必要がある。

3.2 新しい情報検索システムの提案

3.2.1 システムの概要

このシステムは、図3に示すように、ユーザとのインターフェースを担当する1個のAgent1と、データベースとのインターフェースを担当する一般に複数のAgent2を持つ。Agent2には各々専門分野があり、インターネット上に分散する多種多様のデータベースに適切に対応する。以下では、これらのAgentの機能を具体的に説明する。

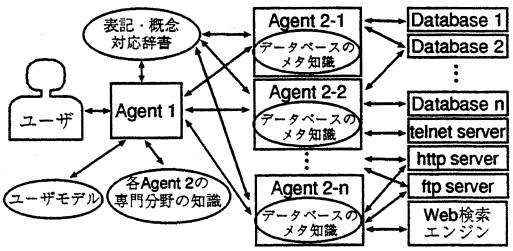

図3 提案する情報探索システムの概要

3.2.2 Agent1の機能

(a) 意図推定とそれに即した検索式の生成

ユーザが必要とする情報を的確に検索するために、ユーザが呈示したキーワードとユーザとの対話に基

づいてユーザの意図を推定し、これに即してキー概念からなる検索式を生成する。さらに、この対話およびユーザの検索歴に基づいて、ユーザが要求するキー概念の重みづけを行う。なお、この際の対話は、ユーザの思考の流れを妨げないように、主として音声により行う。

(b) 表記・概念対応辞書の管理・拡張と知識獲得

対話は自然言語を介して行うため、意図を推定するには、表層表現 (表記) と深層表現 (概念) との対応づけが必要となる。Agent1 は、表記と概念の双方から参照できる表記・概念対応辞書を用いて、この対応づけを行い、さらに、辞書の管理と拡張を行う。最初は、システム設計者から与えられた小規模な辞書から始まり、以後、辞書拡張の知識を自動的に獲得する。

(c) 未知語の処理

ユーザが呈示したキーワードが未知語の場合には、対話によりその概念を明確化し、システムにおける既知概念との対応づけを行う。もし明確化された概念に対応するものがシステムになければ、新概念としてシステムの辞書に登録する。

(d) ユーザのシステム利用法に関する知識の獲得

検索精度と検索速度を向上させるために、ユーザのシステム利用法に関する特徴をデータ (ユーザモデル) として蓄積し、ユーザに適した検索ルートを自動的に獲得する。

(e) 検索依頼

抽出したキー概念の分野を判断し、適切な Agent2 (一般に複数) に検索を依頼する。

(f) 検索結果の検討

Agent2 から受けた検索結果とユーザの意図との整合性を上記 (a) で求めたキー概念の重みづけに基づいてチェックし、検索結果の順位づけを行う。さらに、その結果をユーザに呈示し、検索結果の妥当性をユーザとの対話によって確認する。結果が妥当であれば処理を終了する。ユーザの意図と合致しない場合には、キー概念の重みづけを変更し、既に求めた検索結果の順位を変えて再度ユーザに呈示する。また、対話により新たなキー概念が呈示された場合には、検索式を修正し、再び検索を Agent2 に依頼する。以後、満足な結果が得られるまでこの処理を繰り返す。

3.2.3 Agent2 の機能

(a) データベースからの情報検索

Agent1 から検索要求を受け、データベースの内容とその記述形態に関する知識 (データベースのメタ知識) を利用しながらキー概念検索を行う。

(b) データベースに関するメタ知識の獲得

各データベースを定期的にチェックし、データベースの内容に変更がある場合には、それに応じてデータベースのメタ知識も変更する。また、検索時のヒット件数を記録することにより、データベースの有用性に関する知識も、メタ知識として自動的に獲得する。

4 おわりに

本稿では、情報検索を行うシステムと、ユーザとの間の対話において、ユーザの発話からキーワードを抽出するために、情報検索を目的とした対話を分析して、キーワード抽出のための規則を調べた。その規則に従ってユーザの発話からキーワードを抽出する実験を行った。その結果、この規則がキーワード抽出に有効であることを示した。

次に、キー概念の抽出・未知語処理・知識獲得を組み合わせた新しい情報検索システムを提案した。ユーザの意図推定と情報検索とを分離して取り扱うことにより、検索精度・検索効率の一層の向上が期待される。

参考文献

[1] 藤崎博也：“対話音声の言語学的モデルの構築,” 文部省科研重点領域研究『音声対話』研究報告書, pp. 321-324 (1996).

[2] 藤崎博也, 亀田弘之, 河井恒：“新聞記事情報の階層構造に基づく記事分類・検索システム,” 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[3] 藤崎博也, 亀田弘之, 大野澄雄, 阿部賢司, 伊東卓哉, 佐久間聖仁：“キー概念の抽出と未知語の処理に基づく情報検索方式の高度化,” 情報処理学会第 54 回全国大会 (1997 年 3 月) 発表予定.