音声対話を用いた大規模 DB 向け情報提供システムにおける検索キーワードの自動生成について

渡辺　泰久　　森　弘之　　中嶋　宏

オムロン株式会社 IT研究所

1. まえがき

近年、ネットワークインフラの大規模化により、大量の情報がネットワーク上を行き交うようになった。しかし一方で、情報の効果的活用の観点から、大量の情報の中から所望の情報を効率よく取得できることが求められている。このことに関して、音声の有効性が認知されつつあり、昨今では、音声ポータルという、さまざまな情報を音声で検索できるサービスも始まっている。ところが、数十万件レベルの大規模なデータベースを対象としたものとなると、実用レベルのものはまだ発表されていない。

我々は、数十万件規模のデータベースから簡単かつ瞬時に情報を検索できる音声対話技術の開発に取り組んできた。本稿では、本技術を利用した情報提供サーバの構成、および、要素技術の一つである検索キーワード生成手法の事例について述べる。

2. 検索キーワード生成に関する課題

数十万件規模のデータベースへの対応に際しては、さまざまな課題が抽出された。本稿ではその一つとして、検索キーワードの自動生成を挙げる。

検索キーワード生成とは、データベースの各エントリに含まれる文字列から、そのエントリを音声で検索するためのキーワード文字列（読み）を抽出する処理を指す。部分一致検索を行うため、キーワードにはエントリ文字列の部分文字列も含まれる。

情報提供サービスでは、検索できないデータエントリがあることは許されない。一方、キーワード生成の中心となる形態素解析処理については現在のところ一定比率の解析誤りは避けられないため、解析誤りとなったデータエントリに対しては手入力にてキーワードを補う必要があった。ところが数十万件規模のデータベースを対象にした場合、解析誤りとなったエントリの手作業による修正がコスト的に見合わないことも容易に想定される。そこで、すべてのデータエントリに対して検索キーワードを必ず生成できるようにすることが望ましい。以上より、検索キーワードの自動生成を課題として設定した。

3. 情報提供サーバの概要

情報提供サーバの構成を**図 1** に示す[1]。入出力手段として電話音声を想定している。ユーザは通常の公衆回線を通じてサーバに接続する。サーバから流れる音声ガイダンスに従って、検索条件（書籍情報の場合、本のタイトルや著者など）を発話すると、検索が実行され、所望の情報を音声にて取得することができる。

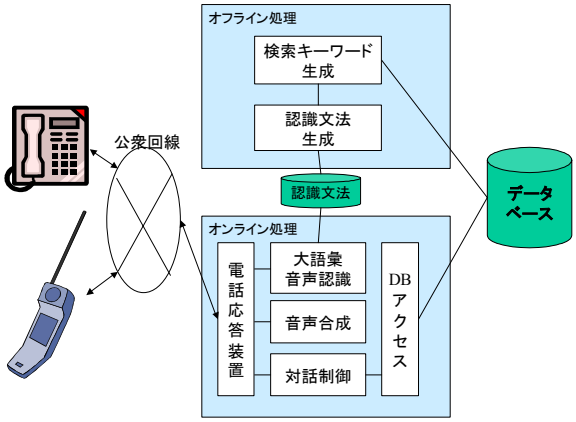

図 1　情報提供サーバの全体構成

図 1 において、検索キーワード生成および認識文法生成は、オンライン処理の事前に実行される。検索キーワード生成では、データベースの各エントリに含まれる文字列を抽出した後、そこから部分文字列を生成する。この部分文字列は、検索条件として発話の対象となるものである。認識文法生成では、生成された検索キーワードに、発音のゆれなどを吸収するための情報を付加した後、音声認識で用いる認識文法を生成する。

4. 課題の解決案とその考察

4.1. 解決案の検討

音声入力用の検索キーワードを得るためには、データ文字列を正しく単語分割し、かつそれぞれの単語に正しい読みを対応付ける必要がある。2 節で述べた課題を解決するため、(a)単語分割の手段である形態素解析の精度向上と、(b)正しく分割できなかったデータからもできるだけ多くのキーワードを抽出するという２つのアプローチを採った。以下、順に述べる。

(a)読み情報を利用した形態素解析

情報提供に用いられる楽曲や書籍などのデータベースは、形態素解析が困難である。理由は、処理対象の文字列が通常の文章に比べ短いことと、特殊な読みかたをするケースが多いことによる。特殊な読みは恣意的であるため、共起関係による類推も困難である。

Automatic Keyword Generation on the Voice-activated Information Server for a Large Database
By Yasuhisa Watanabe, Hiroyuki Mori and Hiroshi Nakajima
Information Technology Research Center, OMRON Corporation

そこで、原始的ではあるが、データベース自体に付与されている読み情報（以下、正解読みという）を参照して形態素解析する方法をとった。情報提供サービスで用いられるデータベースは、正解読みをもつことが多い。そこで、解析結果の集合（形態素解析によって付与された読みを含む）の中から、その読み（以下、生成読みという）が正解読みに一致したもの（一致しない場合は、不一致箇所が最も少ないもの）を結果として選択するようにした。

(b) 解析結果の精度に応じた生成ルールの切り替え

上記で得られた解析結果の信頼度に応じ、適用する検索キーワード生成ルールを動的に切り替える構成をとった。今回は、解析結果精度の判定方法として、生成読みと正解読みとの不一致箇所の位置を用いた。検索キーワード生成ルールはあらかじめ複数個用意しておく。概略処理フローを**図 2** に示す。ルール１では通常の生成を行う。ルール２では、不一致箇所を名詞相当の未知語とみなした後、ルール１を適用する。ルール３では、区切り文字でデータエントリを分割した後、分割後の各文字列に対して再度図２に記した処理を実行する。区切り文字は中黒（「・」）やダッシュ（「ー」）などで、あらかじめデータベースごとに定めておく。分割するための区切り文字がデータエントリに含まれない場合は、文字列全体を一つのキーワードとして扱う。なお、正解読みの分割位置推定は、正解読みと生成読みとの文字単位での比較によった[4]。以上で述べた方法により、解析に失敗したエントリに対しても、できるだけ多くのキーワードを抽出することが可能になる。

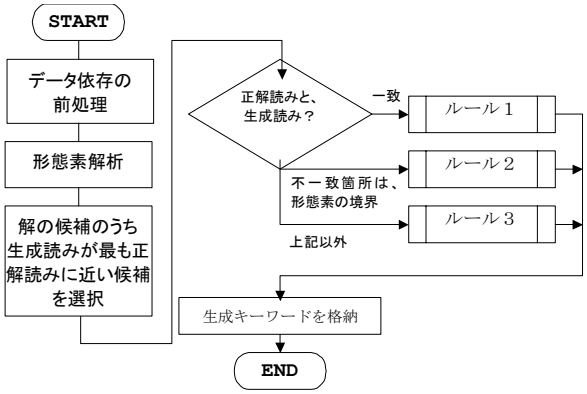

図２　概略処理フロー

4.2. 結果と考察

本稿では前項(a)の結果について記す。

対象としたデータベースを**表 1** に示す。楽曲情報データベースには戦後国内で発売されたすべての楽曲、書籍情報データベースには 1999 年から 2000 年の間に発行された書籍がそれぞれ収録されている。

形態素解析エンジンは茶筌 ver2.2.1 をベースに、読み情報を与える機能を独自に追加して使用した。[2][3]

表 2 に、データ種別ごとの正解率を示す。読みの比較を行わない場合に比べ、曲名で正解率が 12 ポイント向上している。これは約 56,000 件分に相当する。

表 1　対象データベース

データ種別		件数	生成KW数
楽曲情報	曲名	468,998	1,237,771
	人名	109,861	204,479
書籍情報	書名	99,312	576,807

表 2　単語分割の正解率

データ種別	正解率	
	正解読み非参照	正解読み参照
楽曲曲名	81%	93%
楽曲人名	83%	91%
書籍題名	77%	90%

表 3　解析失敗の原因分析

原因種別	人名	曲名	書名
未登録語(日本語)	56%	29%	75%
未登録語(外国語)	29%	27%	3%
特殊な読み	1%	13%	2%
記号の読み	8%	14%	12%
特殊な略記	1%	4%	0%
その他	4%	13%	8%

ここで正解率とは、読みが一致しかつ正しく分割されているものの比率を指す。

なお、解析失敗したエントリについて、想定される原因を分析した結果を**表 3** に示す。未登録語が多く、日本語・外来語を合わせると半数以上を占める。データ種別に着目すると、曲名は、特殊な読み（例えば「楽園」と書いて「パラダイス」と読ませるなど）が使われているケースが他の種別に比べ多い。これはデータ種別の特質上、やむをえないものと考えられる。

5. まとめ

音声を用いて数十万件規模のデータベースを検索できる「情報提供サーバ」を開発した。検索キーワードを自動生成するための手法を確立した。さらに、応用例として、アプリケーション２例を実装した。今後、実フィールドによる実証実験を通じて、新たな課題の抽出を行う予定である。

6. 参考文献

[1] 大本、牛田、中嶋、石田：「電話音声で情報提供を行うアプリケーションの UI 設計と評価の実践報告」。ヒューマンインタフェース学会研究報告集、Vol.3 No.4 pp. 35-40 (2001)

[2] 山下：「パトリシア木を用いた形態素解析のための辞書検索」. ChaSen Technical Report, CTR-1 (1996)

[3] 松本、浅原：「IPA 品詞体系に基づく日本語辞書説明書」http://chasen.aist-nara.ac.jp/(2000)

[4] S. Wu, et al.: “An O(NP) Sequence Comparison Algorithm”, Information Processing Letters, 35, pp. 317-323 (1990)