大規模知識ベースの検索を行う音声対話システムの確認戦略の評価

駒谷 和範　　翠 輝久　　河原 達也　　奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻

komatani@kuis.kyoto-u.ac.jp

1　はじめに

音声による自然言語テキストに対する検索システムの研究が行われつつある [1, 2, 3]．音声対話システムでは，発話の音声認識結果から，ユーザの意図を解釈する手続きが不可欠である．従来の音声対話システムでは，ユーザ発話から事前に人手で定めたキーワードを抽出することで意図を解釈し，それが同定できなければ確認するという方法論を採ることができた．しかし，マニュアル [4] や Web ページなど，テキストで記述された大規模な知識ベースを検索する際には，タスク達成に必要なキーワードの集合を事前に定義することが不可能である．したがって発話の音声認識結果からキーワードを取り出して意味解釈結果とするのではなく，自然言語文として解釈する必要がある [5]．

このようなシステムを実現する際に，単純に音声認識結果をそのまま検索システムの入力とするには，以下の 2 つの問題点がある．

1. 音声認識誤りへの対処
2. 音声言語表現に含まれる冗長性

音声認識誤りに対しては，キーワードが明確に定義されている場合には，それらに対して確認を行うことができる．しかし，自然言語に対する検索システムの場合には，キーワードを適切に定義することは難しい．また，音声は入力が容易であることが利点として挙げられるが，これは同時にドメイン外発話や多様な文末表現など，タスク遂行には直接関係がない冗長な部分が入力されやすいことを意味する．したがって，ユーザの発話の音声認識結果の中から，タスク遂行に有用な部分を自動的に判別する枠組みが必要である．

本研究では，検索対象のみから学習した統計的言語モデルによる尤度と，音声認識結果の N-best 候補に対する検索結果の両方を用いて，音声認識結果の各文節が検索に有用かどうかを判定する．この判定を行う際に必要な情報は，検索対象である自然言語の知識ベースから，*tf·idf* 値や統計的言語モデルという形で機械的に抽出している．これらを用いることにより，事前に人手でキーワードを定義することなく，検索に有用な部分とそうでない部分を切り分け，効率よく確認を行うことができる [6]．本稿では，さらに多人数の被験者により行った評価実験について報告する．

表 1: マイクロソフト社ソフトウェアサポート用知識ベース

知識ベースの種類	件数	文字数
用語集	4707	約 70 万
ヘルプ集	11306	約 600 万
サポート技術情報	23323	約 2200 万

2　大規模知識ベースに対する検索システム

本研究では，対象タスクドメインとして，マイクロソフト社のソフトウェアサポート用知識ベースに対する検索を扱う．これは，表 1 のように知識ベースが大規模であることが特徴の一つである．

この知識ベースに対して，ユーザのテキスト入力文により検索を行う質問応答システムとして，ダイアログナビが東京大学で開発されている [7]．ダイアログナビは，自然言語入力文と知識ベースを柔軟にマッチングするために，係り受け関係や同義表現を考慮して解釈する．すなわち，自立語の一致だけではなく，木構造の文節の深さの一致や，係りタイプ等も評価し，マッチングの尺度として用いている．

本研究では，バックエンドとしてダイアログナビを使用して，音声入力により検索を行うシステムを作成する．この際に問題となる音声認識誤り，ドメイン外発話といった音声言語の問題に対して，頑健にユーザ発話の理解を行うための確認戦略を提案する．

3 検索整合度と検索重要度を用いた対話戦略

3.1 検索整合度の定義と検索への重み付け

ユーザ発話の中から，検索に有用な部分とそうでない部分を切り分ける基準の一つとして，各文節に対する単語パープレキシティを用いる．単語パープレキシティの計算に用いる言語モデルは，検索対象である知識ベースから学習したもので，音声認識時に用いるものとは異なる．したがってここでの単語パープレキシティPPは，検索対象との整合性を示す尺度である．

このパープレキシティの値が小さいということは，知識ベースにおけるその単語列の出現頻度が高いことを意味する．逆に，音声認識結果中の認識誤りである箇所は文脈的に不自然である場合が多く，またドメイン外発話の単語列も知識ベースの中では出現確率が低いため，パープレキシティの値は大きくなる．これにより，認識誤り部分や，認識したが検索には重要でない部分を同時に検出できる．

このパープレキシティを，以下のシグモイド関数により 0〜1 の間の値に変換したものを検索整合度 (relevance score; RS) とする．

$$RS = \frac{1}{1 + \exp(\alpha * (\log PP - \beta))}$$

本研究では，部分的な認識誤りを棄却するために文節単位で検索整合度を求める．その手順を以下に示す．

1. 音声認識結果を構文解析ツール KNP[8] を用いて文節単位に区切る．
2. 区切られた文節にそのコンテクストとして前後1単語を付け加える．
3. 知識ベースのみから作成した言語モデルでパープレキシティを計算し，その文節の検索整合度に変換する．

計算例を図1に示す．この例では，文頭の「新しく買った」という部分は検索には直接関係がない．また，文末に近い部分は誤って認識されている．これらの部分に対するパープレキシティの値は大きくなっている．これらの値は，検索整合度に変換され，3.3 節で述べる重要語句の確認や，知識ベースとのマッチング時の各文節に対する重みとして用いられる．

このように，音声認識時の言語モデルと検索文書のみから学習した言語モデルを使い分けることにより，音声認識時の頑健性を向上させながら，検索に重要な部分を検出することができる．

ユーザ発話 ：
「新しく買ったＸＰのパソコンでＦＡＸ機能を使うにはど うしたらいいですか？」

音声認識結果 ：
「新しく買ったＸＰのパソコンでＦＡＸ機能を使うにそのＡ以降」

構文解析により文節単位に分割 ：
「新しく／買った／ＸＰの／パソコンで／ＦＡＸ機能を／使うに／その／Ａ／以降」

前後 1 形態素を追加し，パープレキシティを計算 ：

<S>**新しく**買った	$PP =$	499.57
新しく**買った**ＸＰの	$PP =$	2079.83
買った**ＸＰの**パソコン	$PP =$	105.64
の**パソコンで**ＦＡＸ	$PP =$	185.92
で**ＦＡＸ機能を**使う	$PP =$	236.23
を**使うに**その	$PP =$	98.40
に**その**Ａ	$PP =$	1378.72
その**Ａ**以降	$PP =$	144.58
Ａ**以降**</S>	$PP =$	27150.00

<S>，</S>はそれぞれ始端記号，終端記号

図 1: 検索整合度 (パープレキシティ) の計算例

3.2 検索結果を用いた検索重要度の計算

音声認識結果の N-best 候補に対する検索結果を用いて検索重要度を定義する．音声認識結果の N-best 候補には音声認識の過程で曖昧であった部分が現れるが，この部分が検索に重要かどうかを規定している．

まず音声認識結果の N-best 候補間の相違箇所を同定し，この相違部分に対して検索重要度を計算する．N-best 候補それぞれについて実際に検索を行い，検索結果の相違の大きさを検索重要度 (significance score; SS) とする．第 n 候補と第 m 候補間の検索重要度 $SS(n,m)$ は，第 n 候補に対する検索結果を $res(n)$，その数を $|res(n)|$ として，以下のように定義する．

$$SS(n,m) = 1 - \frac{|res(n) \cap res(m)|^2}{|res(n)||res(m)|}$$

3.3 検索整合度と検索重要度を用いた対話管理

文節ごとの音声認識誤りによる損失，つまり検索結果に対する重要度を考慮して，確認の方法を切り替える．

音声認識誤りが検索に決定的な影響を与えることが予測される語句は，検索を実行する前にユーザに確認する．これらの語句は，知識ベースにおいて計算した

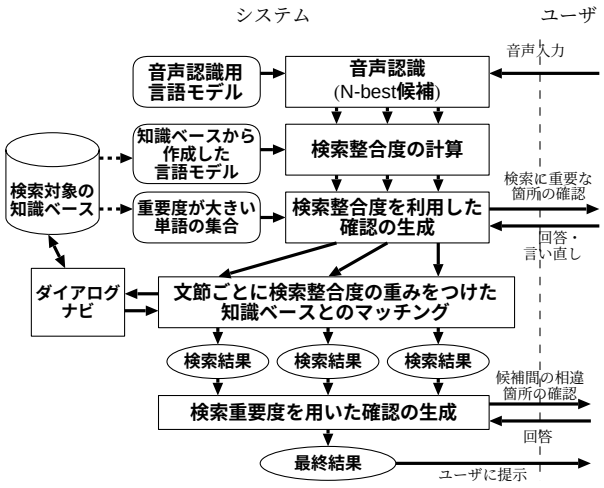

図 2: 確認対話戦略を導入した検索システムの概要

tf·idf 値により規定する．この事前確認を行うか否かの判断に検索整合度を利用し，あるしきい値 θ 以下のものには確認を行わない．ユーザからの回答により認識誤りであるとされた文節は，音声認識結果から取り除かれ，残りの部分が次のマッチングモジュールに渡される．

次に，結果として検索結果に影響を与える箇所を検索重要度により同定し，検索後に確認する．検索重要度は，音声認識の N-best 候補の相違部分に対して計算されるが，ある部分に対する検索重要度がしきい値を超えている場合には，その相違部分をユーザに対して提示し確認する．今回音声認識で出力する候補数 N は 3 とし，検索重要度のしきい値は 0.5 とした．ユーザがその中から適切な候補を選択すると，対応する検索結果をユーザに提示する．検索重要度がしきい値以下の場合には確認を行わず，第 1 候補による検索結果をそのまま提示する．候補が全て適当でないとされた場合には，現在の候補を棄却し再発話を促す．

これらの確認戦略を含めた全体の流れを図 2 に示す．

4 評価実験

評価用データは本システムを利用したことのない 30 名の被験者により収集した．設定した想定場面に基づいて各 11 課題，これとは別に自由に 3 課題に対して検索を行ってもらった．ただし，検索結果としてふさわしい候補が提示されない場合には，各課題につき 3 度まで言い直しを許した．その結果，合計 420 課題，651 発話を得た．30 人の発話の音声認識率は平均で 76.8%であった．

収集した音声データに対して，以下の 3 つの条件で検索実験を行った．なお，システムがユーザに提示した候補の中に，最初の質問の答えとなる候補が含まれていた場合を検索成功としている．

表 2: 検索成功率

総発話数	書き起こし入力	認識結果を入力	提案手法
651	520 (79.9%)	421 (64.7%)	457 (70.2%)

1. ユーザ発話の正確な書き起こし (人手で作成) を用いて検索した場合［書き起こし入力］
2. 音声認識結果の第 1 候補をそのまま検索した場合［認識結果を入力］
3. 検索整合度と検索重要度の両方を用いて検索を行い，生成する確認に対してユーザが適切に応答したとする場合［提案手法］

検索の成功率を表 2 に示す．書き起こし入力は音声認識率が 100%の場合に相当し，音声認識部における改善の上限を表す．提案手法では，音声認識結果の第 1 候補をそのまま用いて検索を行った場合よりも検索の成功率が上昇している．提案手法による検索成功数の改善は 36 であった．このうち確認によるものが 30，検索整合度を用いたマッチングにより検索が改善したものが 14 であったが，逆に 8ヶ所で，マッチングにより適切な検索結果が得られなくなっている．

次に，システムが生成した確認の回数に関して検証を行った．提案手法により生成された確認の回数は 221 回であった．これは，おおよそ 2 課題に 1 回強，確認が行われたことになる．このうち，検索整合度を用いた事前確認の回数は 66 回あり，検索重要度を用いた事後確認が 155 回であった．検索整合度を用いた検索前の確認により，確認を行わない場合と比べて 3 発話分，検索成功回数が増加した．また検索重要度を用いた検索後の確認により，27 発話で検索成功回数が増加した．

提案手法の確認回数を評価するために，音声認識結果の N-best 候補から計算される信頼度 [9] を用いた確認手法との比較を行った．確認を行うための信頼度の閾値θ_1として，0.4，0.6，0.8 の 3 通りを用いた．信頼度が閾値θ_1以下の自立語に対して確認を行うものとし，それが誤りであった場合には，その単語を含む文節を取り除いて検索した．

この結果を表 3 に示す．提案手法は，従来手法の信頼度の閾値θ_1が 0.8 の場合に比べて，確認回数を半分以下に抑えながら．より高い検索成功率を得ている．

表 3: 音声認識の信頼度を用いた確認戦略との比較

	提案手法	信頼度 ($\theta_1 = 0.4$)	信頼度 ($\theta_1 = 0.6$)	信頼度 ($\theta_1 = 0.8$)
確認回数	221	77	254	484
検索成功数（成功率）	457 (70.2%)	427 (65.6%)	435 (66.8%)	445 (68.4%)

表 4: 各発話ごとの音声認識率と検索成功率

認識率 (%)	発話数	認識結果を入力	提案手法	改善数
−30	23	6 (26.1%)	7 (30.4%)	1 (4.3%)
−40	14	3 (21.4%)	4 (28.6%)	1 (7.1%)
−50	34	16 (47.1%)	20 (58.8%)	4 (11.8%)
−60	39	17 (43.6%)	22 (56.4%)	5 (12.8%)
−70	91	50 (54.9%)	56 (61.5%)	6 (6.6%)
−80	103	66 (64.1%)	73 (70.9%)	7 (6.8%)
−90	132	84 (63.6%)	91 (68.9%)	7 (5.3%)
−100	215	179 (83.3%)	184 (85.6%)	5 (2.3%)
合計	651	421 (64.7%)	457 (70.2%)	36 (5.5%)

従来手法の信頼度 [9] は，音声認識結果の音響的・言語的尤度のみを反映したものである．これに対して本手法での確認は，検索対象から学習した言語モデルや，音声認識結果の N-best 候補に対する検索結果を用いて行うため，確認を行うかどうかの判断にドメイン知識が反映されている．実験結果により，これが適切であることが示されている．

さらに，表 2 の結果を詳しく分析した．各発話ごとの音声認識率と検索成功率の関係を表 4 に示す．各音声認識率における，発話数に対する改善率をみると，音声認識率が 40%から 80%の発話における改善率が比較的大きい．これは，本手法が，音声認識率が高い部分のみに対してだけでなく，音声認識率が 50%程度の発話に対しても有効であり，これらの発話に対しても適切に確認を行うことで，検索成功率を向上させていることを示している．

5 まとめ

音声により，大規模な知識ベースを検索するタスクにおける確認戦略を提案した．音声認識結果に対して検索整合度と検索重要度の 2 つの尺度を導入し，音声認識誤りや余分な入力を含む部分に対して効率よく確認を行う．自然言語テキストを対象した検索では，タスクを達成するのに必要なキーワードを定義するのが困難であり，確認の対象となるべき語句を決定できない．本研究では，検索対象の知識ベースから得られる統計的言語モデル，*tf·idf* 値や，実際の検索結果などの情報を用いて，確認対象箇所を決定する．30 人の被験者による評価実験により，その有効性を確認した．

謝辞

本研究は，東京大学の黒橋禎夫助教授，清田陽司氏，マイクロソフト株式会社の木戸冬子氏との共同研究である．各氏の貴重な貢献に感謝します．

参考文献

[1] Harabagiu, S., Moldovan, D. and Picone, J.: Open-Domain Voice-Activated Question Answering, *Proc. COLING*, pp. 502–508 (2002).

[2] 藤井敦: 音声による言語バリアフリーな他言語情報アクセス, 情報処理学会研究報告, SLP-44-33 (2002).

[3] Hori, C., Hori, T., Isozaki, H., Maeda, E., Katagiri, S. and Furui, S.: Deriving Disambiguous Queries in a Spoken Interactive ODQA System, *Proc. IEEE-ICASSP* (2003).

[4] 伊藤亮介, 駒谷和範, 河原達也: 機器操作マニュアルの知識と構造を利用した音声対話ヘル プシステム, 情報処理学会論文誌, Vol. 43, No. 7, pp. 2147–2154 (2002).

[5] 駒谷和範, 河原達也, 清田陽司, 黒橋禎夫, Fung, P.: 柔軟な言語モデルとマッチングを用いた音声によるレストラン検索シ ステム, 情報処理学会研究報告, 2001-SLP-39-30 (2001).

[6] 翠輝久, 駒谷和範, 河原達也, 奥乃博: 音声対話によるソフトウェアサポートタスクのための確認戦略, 情報処理学会研究報告, SLP-47-11 (2003).

[7] Kiyota, Y., Kurohashi, S. and Kido, F.: "Dialog Navigator": A Question Answering System based on Large Text Knowledge Base, *Proc. COLING*, pp. 460–466 (2002).

[8] 黒橋禎夫, 長尾真: 並列構造の検出に基づく長い日本語文の構文解析, 自然言語処理, Vol. 1, No. 1, pp. 35–57 (1994).

[9] 駒谷和範, 河原達也: 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理, 情報処理学会論文誌, Vol. 43, No. 10, pp. 3078–3086 (2002).