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—

清田 陽司
京都大学 大学院情報学研究科
kiyota@kc.t.u-tokyo.ac.jp

黒橋 禎夫
東京大学 大学院情報理工学系研究科
kuro@kc.t.u-tokyo.ac.jp

木戸 冬子
マイクロソフトアジアリミテッド
fkido@microsoft.com

1 はじめに

何かを調べたいとき，一番よい方法はよく知っている人 (その分野の専門家) に直接聞くことである．多くの場合，自分の調べたいこととその答えの間には，具体性のズレ，表現のズレ，背景の認識の不足などがあるが，専門家は質問者との対話を通してそのようなギャップをうめてくれるのである．

現在，WWW 等の大規模な電子化テキストが存在するようになり，潜在的にはどのような質問に対してもどこかに答えがあるという状況が生まれつつある．しかし，今のところ WWW を調べても専門家に聞くような便利さはない．その最大の原因は，上記のようなギャップを埋めてくれる対話的な能力が計算機にないからである．本稿では，大規模なテキスト知識ベースに対する対話的な問い合わせシステム，ダイアログナビについて述べる．

2 ダイアログナビの概要

現在のマイクロソフト「話し言葉検索」[1] は，入力は自然言語であるが，そこからキーワードを抽出し，マッチするテキストのリストを返すという従来型の情報検索システムである．ここに対話の要素 [1] を取り入れていく研究を，東大とマイクロソフトの共同研究として進めている．ダイアログナビにおいて使用するテキスト集合とその規模，具体例を表 1，図 1 に示す．

このテキスト集合の大きな特徴は，テキストが質問応答用に構造化されておらず，かつ非常に大規模であるという点である．このような既存の大規模テキストの上で「曖昧な質問に対する聞き返し」を実現することを第一の目標としている．

本システムの構成を図 2 に示す．ユーザインタフェースの主な特徴は，図 3 に示すように対話の履歴が表示されることである．ユーザはシステムが提示した選択肢をクリックして選ぶことができるほかに，随時キーボードから質問等を入力することもできる．

入力解析部とテキスト検索部は，ユーザの質問文を解析しテキスト集合の検索を行う (3 章)．症状説明文抽出部は，テキスト検索でマッチした文から症状を説

表 1: テキスト集合

テキスト集合	件数	文字数	マッチング対象
用語集	4,707	約 70 万	見出し語
ヘルプ集	11,306	約 600 万	タイトル
サポート技術情報	23,323	約 2,200 万	文書全体

音声認識ソフトウェアがインストールされた環境でページ違反が発生する

最終更新日: 1999/08/18
文書番号: J049655

この資料は以下の製品について記述したものです。

- Microsoft(R) Internet Explorer Version 5 (以下 Internet Explorer 5)
- Microsoft(R) Windows 98 (以下 Windows 98)

概要
この資料は、Windows 98 上に Internet Explorer 5 がインストールされた環境で、音声認識ソフトウェアが起動されていると、Internet Explorer 5 を起動した際に、ページ違反が発生する現象について説明したものです。

内容
以下の条件を満たすときに Internet Explorer 5 を起動すると、ユーザー補助プログラムの OLEACC.DLL が不正なメモリ領域を参照することにより、ページ違反が発生する場合があります。

- Windows 98 にユーザー補助プログラムがインストールされている
- 音声認識ソフトウェアが起動している

回避方法
Windows 98 システムアップデートモジュールをインストールします。システムアップデートモジュールには、新しい OLEACC.DLL が含まれており、この不具合が修正されていることを確認しております。これは Windows 98 Service Pack 1 に含まれるモジュールとなっており、Windows Update からダウンロードすることができます。

入手方法

1. [スタート] メニューから [Windows Update] をクリックします。
2. 画面の指示に従い "Windows Update へようこそ" が表示されたら、"製品の更新" をクリックします。
3. "ソフトウェアの選択" 画面にて、"Windows 98 System Update" にチェックをつけ、" ダウンロード" ボタンを押します。
4. 画面の指示に従い、モジュールをインストールします。

図 1: マイクロソフト・サポート技術情報の例

[1] http://www.microsoft.com/japan/enable/nlsearch/

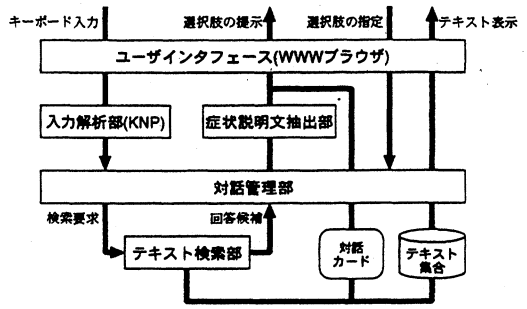

図 2: システム構成

明している部分を抽出する (4.1 節). 対話管理部は, ユーザから典型的な曖昧な質問がなされた際に, 対話カードを利用して, 選択肢を示して聞き返しを行う (4.2 節).

3 テキストの検索

質問応答システムにおいてまず重要なことは, 質問の答えを含むと思われるテキストを十分な精度で検索できることである. そのために, 質問タイプやプロダクト名によるテキスト集合のしぼりこみを行う. また, 表現の揺れを吸収するために同義表現辞書 (図 4) を利用したマッチングを行う. さらに, スコア計算において,「ファイルを→解凍する」のような係り受け関係に加点している.

テキスト検索部は, 以下に述べるアルゴリズムによって各テキストのスコアを計算し, スコアにもとづいてユーザに提示するテキストの候補をしぼりこむ.

3.1 テキスト集合のしぼりこみ

入力解析部は, 質問文を KNP によって構文解析し, さらに文末に質問タイプの分類規則 (文末表現に対するパターン) を適用することによって, 質問文の内容表現と質問タイプを抽出する. 質問タイプは Symptom, How, What の 3 種類である. どの分類規則にも一致しなかった場合は「タイプなし」として扱う. テキスト集合については表 2 に示すように分類する. 質問文より抽出された質問タイプによって, 検索対象のテキスト集合を表 3 に示すようにしぼりこむ.

サポート技術情報・ヘルプ集についてはプロダクトによるしぼりこみも行う. 質問文にプロダクト名 (Windows NT, Word, Excel など) が出現する場合

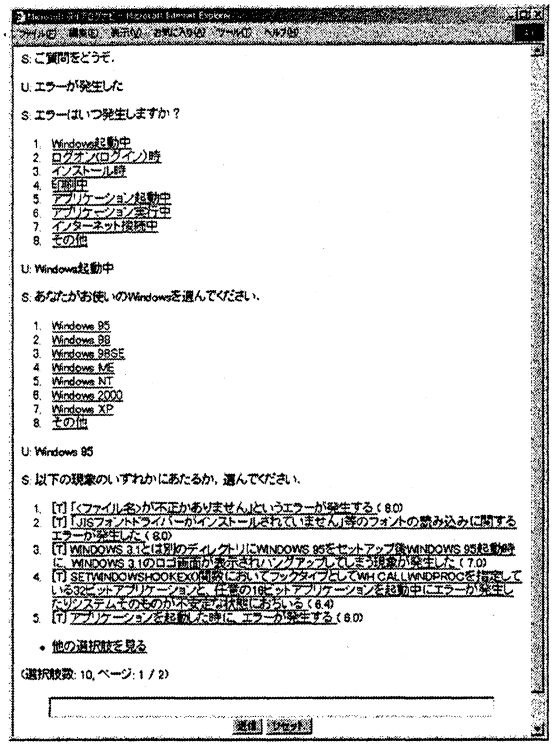

図 3: ユーザインタフェース

表 2: テキスト集合のタイプ分類

テキスト集合		タイプ
用語集		What
ヘルプ集	タイトルが「～の概要」	What
	上記以外	How
サポート技術情報	タグ (BUG,FIX,PRB)	Symptom
	タグ (HOWTO)	How
	上記以外	タイプなし

は, そのプロダクト名がタグ付けされているテキストを検索対象とする.

3.2 同義表現辞書によるマッチング

検索の一般的な問題として表現のずれの問題がある. 表現のずれは語のレベルだけでなく,「パソコンを起動する」「Windows を起動する」「電源を入れる」のようなフレーズレベルの同義表現も多数存在する. そこで, 頻出する同義表現について辞書を作成し, これを用いて同義表現のマッチングを行う. 作成した同義表現辞書の一部を図 4 に示す.

表 3: 質問タイプによるしぼりこみ

テキスト集合		質問タイプ What	How	Symptom	なし
用語集	What	○	×	×	×
ヘルプ集	What	○	×	×	○
	How	○	○	×	○
サポート技術情報	Symptom	○	×	○	○
	How	○	○	×	○
	タイプなし	○	○	○	○

・発生する，起きる，おきる，起こる，おこる
・メール，メイル，電子メール，電子メイル，Mail，E-Mail
・メールを読む，メールを受信する，メールを見る，メールを受ける，メッセージを受信する，メッセージを受ける
・パソコンを起動する，Windows を起動する，電源を入れる，電源をオンする，ブートする，パソコンを立ち上げる，スイッチを入れる

図 4: 同義表現辞書の例

システムは，質問文の内容表現・テキストと同義表現辞書のマッチングを行い，同義表現を抽出する．フレーズレベルの同義表現については自立語同士の係り受け関係が一致した場合に抽出する．例えば「メールを受信する」という同義表現については，「メール→受信」という係り受け関係がある場合に抽出する．

3.3 テキストのスコア計算

各テキストのスコアは質問文とテキスト内の文との類似度として計算する．

まず，質問文とテキスト内の各文との間で**文スコア**を計算する．文スコアは，質問文に出現するキーワード[2]と係り受けペアのうち，テキスト文にも出現するものに点数を与え，合計したものである．キーワードの一致については 1 点，係り受けペアの一致については 2 点を与えている．同義表現として一致するキーワード・係り受けペアについても同様に点数を与える．

最後に，各テキスト中でもっともスコアの大きな文をテキストの**代表文**とし，その文スコアをテキストのスコアとする．

なお，サポート技術情報については記述が長く，一つの事象を複数文で説明している場合があるので，近接マッチについても考慮する．具体的には，近接する文 (前後の文) との一致についても，正規の 1/2 の点数を与える．また，タイトル・概要・現象・症状セクションと，それ以外のセクションの間には重要度に差があるので，文の存在するセクション名に応じてスコアに以下の係数を掛け合わせる．

- タイトル・概要　1.0 倍
- 現象・症状　0.8 倍
- 上記以外　0.6 倍

[2]普通名詞・サ変名詞・固有名詞・カタカナ語・英数字・形容詞・形容動詞・動詞をキーワードとして扱っている．

表 4: 回答候補のしぼりこみのパラメータ

テキスト集合	最大候補数 n	スコア閾値 t
用語集	2	0.8
ヘルプ集	10	0.6
サポート技術情報	10	0.2

3.4 回答候補のしぼりこみ

3 つのテキスト集合 (用語集・ヘルプ集・サポート技術情報) ごとに，テキストのスコアに基づいてユーザに提示する回答候補をしぼりこむ．

テキストをスコアの大きい順に整列し，上位 n 個までをユーザに提示する回答候補とする．ただし，スコアが閾値 t を下回るものは対象外とする．n, t の値はテキスト集合ごとに表 4 に示すように定めた．

複数のテキスト集合から回答候補が得られた場合は，用語集，ヘルプ集，サポート技術情報の順で提示する．

4 ユーザのナビゲート

多くの場合，ユーザの質問と答えの間には様々なギャップが存在する．例えば，「Windows でエラーが発生した」のような曖昧な質問では該当するテキストが多すぎて，それをそのまま提示したのでは，ユーザ自ら答えを見つけるのは非常に困難である．ユーザの求める答えにたどりつくためには，「エラーが発生したのはいつですか」「使っている Windows のバージョンは何ですか」「どんなエラーメッセージが表示されましたか」などといった聞き返しを行って，ユーザを求める答えにナビゲートする必要がある．

4.1 症状説明文の抽出

回答候補中のユーザ質問とマッチした文の，マッチした部分の周囲には，ユーザが遭遇する問題の具体的な症状が書かれている．そこで，各回答候補からそのような症状説明文を取り出してユーザに一覧として提示する．

たとえば,「Windows 98 上でページ違反が発生する」という質問文に対しては,「Windows 98 で IE5 を起動した際にページ違反が発生する」や「Windows 98 でディスククリーンアップを実行するとページ違反が発生する」という文がマッチするが，ここから具体的な状況を説明している「IE5 を起動した」「ディスククリーンアップを実行した」の箇所を抽出する.

症状説明文抽出部は，各回答候補の代表文から以下のアルゴリズムで症状説明文を抽出する.

- 頻出する冗長な表現 (「この資料では，(〜)」,「以下の」,「(〜する) 問題について説明しています.」など) を削除する.
- 文をパートに分割する．分割する箇所は以下の通りである.
  - 連用修飾節 (複合辞は除く)
  - 〜とき，〜際，〜場合，〜最中など
  - デ格 (読点あり)
- 質問文の一部と完全に重なるパートを削除する.
- 末尾のパートと，それを直接修飾するパートのみを状況説明文として出力する.

4.2 対話カードの作成

しかし，ユーザの質問が曖昧すぎる場合にはユーザ質問とマッチする部分の検出そのものが難しく，うまく症状説明文を抽出することができない．また，このような場合には大量のテキストがマッチしてしまうので，たとえ上記の方法を適用しても多くの選択肢の中から選ばなければならず，ユーザの負担は膨大である.

そこで，曖昧な質問に対して段階的な聞き返しを行うための対話カードを作成した．この例を図 5 に示す．対話管理部は，特定の曖昧な質問がなされたときに対話カードに従って，聞き返しを行う.

対話カードが利用された場合の対話例を図 3 に示す．まずユーザが「エラーが発生した」という質問をすると，質問文と各対話カードの <UQ> の部分とのマッチングを 3.2 節，3.3 節で述べたアルゴリズムにより行う．この結果，[エラー] の対話カードが選ばれる．システムはこのカードに従って,「エラーはいつ発生しますか」という聞き返しを，選択肢を示して行う．ユーザが「Windows 起動中」を選ぶと，システムは [エラー/Windows 起動中] の対話カードに移って,「あなたがお使いの Windows を選んでください」という聞き返しを行う．ユーザが「Windows 95」を選んだ

```
[エラー]
<UQ> エラーが発生する
<SYS> エラーはいつ発生しますか？
<SELECT>
  Windows 起動中    goto [エラー/Windows 起動中]
  ログイン時        goto [エラー/ログイン時]
  印刷中            goto [エラー/印刷時]
                      ...
</SELECT>
```

```
[エラー/Windows 起動中]
<UQ>Windows を起動中にエラーが発生する
<SYS> あなたがお使いの Windows を選んでください.
<SELECT>
  Windows 95        retrieve「Windows 95 で起動時
                    にエラーが発生する」
  Windows 98        retrieve「Windows 98 で起動時
                    にエラーが発生する」
                      ...
  Windows XP        retrieve「Windows XP で起動
                    時にエラーが発生する」
</SELECT>
```

図 5: 対話カードの例

結果,「Windows95 の起動時にエラーが発生する」という症状に対する解決方法を求めていることがわかるので，これを質問文としてテキスト検索部に渡す.

さらに，得られた回答候補から状況説明文の抽出を行う．図 3 の対話例では，一部の選択肢 (図中では 1, 2, 4 番の選択肢) で冗長な部分が削除されている．このように，状況説明文の抽出と対話カードは，いずれもユーザに選択肢を示して聞き返しを行うものなので，併せて用いることができる.

5 まとめ

今後，本システムを実際に運用し，収集した対話ログを評価してシステムの有用性を確認する予定である．また，収集した対話ログを利用して，複数テキスト間の差異を検出して聞き返しを行うための手法や，ユーザモデルの扱いについて研究を進める予定である.

参考文献

[1] Sadao Kurohashi and Wataru Higasa. Dialogue helpsystem based on flexible matching of user query with natural language knowledge base. In *Proceedings of 1st ACL SIGdial Workshop on Discourse and Dialogue*, pp. 141-149, HongKong, 2000.