5-2　音声質問応答システム「一休」

5-2 Speech-based Question Answering System "Ikkyu"

DE SAEGER Stijn　後藤　淳　VARGA István
DE SAEGER Stijn, GOTO Jun, and VARGA István

要旨

本稿では NICT 情報分析研究室で開発した音声質問応答システム「一休」を紹介する。一休はスマートフォンにより音声で入力されたユーザの多様な質問に対応可能な次世代情報システムである。「日本のデフレの原因」から「脳梗塞の予防策」まで、一休は 6 億 Web 文書に明示的に書かれた回答を網羅的にカバーし、さらに一見かけ離れた情報を組み合わせることで明示的に書かれていない仮説に基づいた回答も生成する。我々は、ふとした思いつきから思考、行動のオプションを広げることで人間の意思決定を支援する新しい情報検索基盤の提供を目指す。

This paper introduces the speech-based question answering system "Ikkyu" developed by the NICT Information Analysis Laboratory. Ikkyu is a next-generation information system that caters to users' various information needs, in the form of spoken natural language questions posed via smartphone. Ranging from causes of the Japanese deflation to preventive measures for strokes, Ikkyu exhaustively covers answers explicitly contained in our 600 million page Japanese Web archive, and furthermore is able to generate answer hypotheses that are not written explicitly but can be derived by combining seemingly unrelated information obtained from distinct documents. This system aims to provide a new search platform that enhances human decision making abilities by providing pinpoint information and relevant suggestions to questions people ask on a whim, thereby broadening their awareness of the various options available to them.

[キーワード]
質問応答，知識獲得，ビッグデータ，自然言語処理，音声認識
Question answering, Knowledge acquisition, Big data, NLP, Speech recognition

1　まえがき

Web の情報が爆発的に増え続ける現状では、検索キーワードにマッチする大量の文書をそのままユーザに提示する情報検索モデルの限界が明らかになってきた。検索結果の上位数十件の Web ページしか見ないユーザが多い現状では、人の知識やそれに基づいた判断が検索エンジンによって偏ってしまう可能性は否定できない。このような状況ではどの情報が見つかるか、どの情報が見つからないかは、偶然に支配されている。こうした状況を鑑みるに、広い視野に立った適切な意思決定に不可欠な情報収集を現在の検索エンジンで行うことは非常に困難であると考えられる。

以上のような問題意識に基づき、我々は「Web に答えさせる」音声質問応答システム「一休」を開発してきた。一休は 6 億文書の日本語 Web アーカイブの意味解析を行い、質問の対象領域を限定せずにテキストまたは音声で入力された様々な質問に対応し、意味解析された Web ページから回答を探し、ユーザの多様な情報ニーズに応えられるように列挙する。一休は、例えば日本のデフレの原因から脳梗塞の予防策まで、6 億 Web 文書に明示的に書かれた回答を網羅的にカバーし、さらに一見かけ離れた情報を組み合わせることで明示的に書かれていない仮説に基づいた回答も生成できる。一休は既存の検索エンジンにはない下記の特長を備えている。

1．検索漏れの抑制。一休は大量の Web 文書の高度な意味解析を行うことでユーザが入力

した質問の多様な言い換え表現を認識し、回答を発見する。そのため、単純なキーワードマッチングとは異なり、異なる表現で書かれた同じ内容の情報を網羅的に発見できる。

2．推論を用い、明示的に書かれていない回答まで仮説として提供する。現状ではこの世界の全ての有用な知識がWebに明記されているわけではない。一休は、Webに存在する断片的な知識を組み合わせることでWebに明記されていない新たな知識を仮説として生成し、ユーザに回答として提供することができる。

3．一覧性を重視する回答表示。回答ランキングに伴う情報の見落としを避けるために、一休は一覧性に欠けるリスト形式ではなく、ワードクラウド形式で回答を表示する（図3右）。**2.1**で説明するように、ワードクラウドの中心からの距離は回答の確からしさを表す。また、ワードクラウド内での他の回答からの距離は回答間の意味的類似度を示し、意味的に近いと思われる回答は近接して表示される。

一休にふとした思いつきを音声質問として入力することで、人間がそもそも把握できない情報量から意外でありながら有用な知識を発見できる可能性は高い。その一例としては、「デフレを引き起こすのは何ですか」という質問が挙げられる。この質問に対しての一休の回答には「リストラ」や「輸入製品」などという常識的なものが含まれるが、意外な回答も見つかる。例えば、日本のデフレの一因としては一休がある日本の大企業名を回答として提供した。その回答の根拠文を図1に示す。ブログから抽出された回答なので一見根拠が薄いと思われるかもしれないが、その結果を我々が発見した後に日経新聞にも同主旨の記事が掲載された。つまり、ある程度社会的に認められた、妥当な回答であると考えられる。この回答が抽出された根拠文と入力された質問文は「デフレ」以外単語のオーバーラップがなく、表層上大きく異なるので、単純なキーワードマッチングではこの回答を発見することは困難である（こうした回答を発見できる一休のメカニズムについては**2**で説明する）。この例が示すように、実際に質問文に含まれる表現に限定されずに未知なる有用な回答を網羅的に発見できることは非常に重要であり、ユーザの思考、行動のオプションを広げることにつながり、さらには広い視野に立った適切な意思決定の実現につながる。これこそが一休プロジェクトの最終目標である。

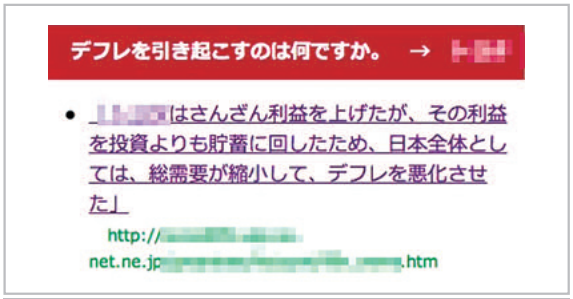


図1 日本のデフレの原因を問う質問に対して、意外だが有用な回答の一例の根拠文

本稿の構成は次の通りである。**2**で本システムの中心となる技術を紹介する。**3**でより多様な質問に対応するアプローチについて述べる。**4**では一休の音声入力とその関連技術を紹介する。**5**は近年注目されている質問応答研究における一休の位置づけを明確にする。最後に**6**は結論を述べる。

2 一休のコア技術

ここでは一休のコア技術を紹介する。一休では、質問応答を関係抽出問題として捉え、[1][2]で提案した意味的関係獲得手法をリアルタイム化したアルゴリズムで解く。例えば、「パリの名物は何？」という質問が入力された場合、一休は、「Xの名物はY」という言語パターンに表現される名詞間の意味的関係を獲得し、次に「パリ」という名詞と同じ意味的関係を持つ名詞を回答として取得する。以下に、質問応答プロセスの具体的な流れについて述べる。

2.1 質問応答アルゴリズムと処理の流れ

一休の質問応答アルゴリズムは図2で表示され、下記のステップからなる。本手法の技術的に鍵となるステップ3，4については**2.2**でより詳細に説明する。

1．質問の音声認識。テキストあるいは音声で

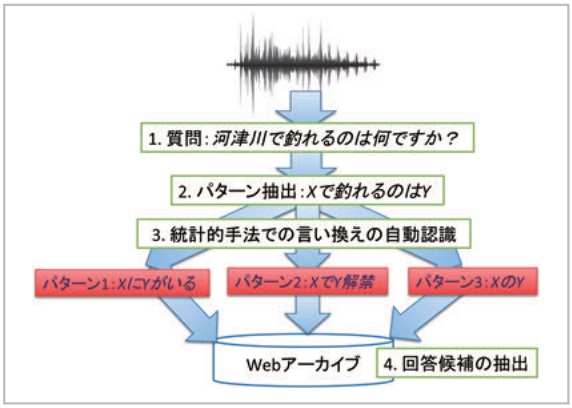


図２　質問応答アルゴリズムの概要

一休に質問を入力する。スマートフォンで入力された質問は、質問文に特化した音声認識モジュールを用い、テキストに変換する。音声認識モジュールについては **4** で紹介する。

2．構文パターンの抽出。次にルールベースの構文変形を行い、質問文を肯定文に置き換える。この肯定文を係り受け解析し、構文木から単語間の意味的関係を表すと思われる構文パターンを抽出する。構文パターンは構文木で２つの単語をつなぐ係り受け関係のパスにある単語から構成される。例えば、図２の「河津川で釣れるのは何？」という質問文は、まず「何は木津川で釣れる？」、「何が木津川で釣れる」という肯定文に置き換わり、「X で釣れるのは Y」、「Y は X で釣れる」、「Y が X で釣れる」（X ＝河津川、Y ＝何）という構文パターンが抽出される。以後、質問文から抽出した構文パターンは「クエリパターン」と呼ぶ。

3．言い換えパターンの獲得。次に質問文から抽出したクエリパターンの拡張処理に入る。このステップは Web に書かれている回答候補を網羅的に認識するためのキーとなる処理である。クエリパターンを自明な構文変形で拡張した後、大規模 Web アーカイブから得られた統計データから計算された構文パターン間の文脈類似度に着目し、質問文から抽出したクエリパターンの言い換え表現と思われる構文パターンを自動的に獲得し、クエリパターンを数十から数百程度の言い換えパターンに拡張する。この言い換えパターンは、クエリパターンとは表層上かなり異なっている場合もある。例えば、「X が Y を引き起こす」というクエリパターンの拡張パターンとして、「X が Y の原因となる」、「X は Y を誘発する」、「X による Y」、「X が招いた Y」などの言い換えパターンが獲得される。

4．回答候補の抽出。上記のパターン集合を用い、回答候補を Web コーパスから抽出し、ランキングして、ユーザに提示する。回答候補のランキングは、その単語の意味的クラス（**2.3** で説明する）、その回答候補を獲得した構文パターンとクエリパターンの意味的類似度を表す言い換え獲得スコア、回答候補とそれを抽出したパターンの関連度など、様々な部分スコアを統合した上でなされる [1][2]。Web ブラウザ及びスマートフォンでの回答の表示例を図３に示す。ブラウザ上の表示（図３右）では、有用な回答を発見しやすくするために、一覧性に欠けるリスト形式の表示ではなく、回答をワードクラウド（回答の「雲」）として表示している。回答の表示形式には次の２つの特徴がある。ワードクラウドの中心からの距離は回答の相対的なスコアを表す。信頼性の高い回答は中心の近くに表示される。また、画面上での回答間の距離は意味的類似度を示す。ワードクラウドの表示アルゴリズムは、意味的に類似する単語を互いに近くに表示する。一方、スマートフォン上では上記のように高度な回答表示には画面表示に十分なスペースがないため、回答をリスト形式で表示している（図３左）。

2.2　言い換えパターンの自動獲得

ここでは一休の言い換えパターン認識アルゴリズムを紹介する。一休の特長の１つは、Web コーパスから回答候補を抽出する際にユーザの質問文の多様な言い換え表現を認識できる点にある。一休は単純なキーワードマッチングでは得る事ができない、ユーザの質問から表層上大きく異なる言い換え表現で書かれている回答まで網羅的に抽出できる。これらの言い換え表現の自動獲得

図３ 一休の回答表示（スマートフォン左、PC ブラウザ右）

は文献 [1][2] で提案した、クラス依存の言い換えパターンを用いた意味的関係獲得手法に基づいている。別の言い方をすれば、一休の言い換えパターン認識は文献 [1][2] のパターン学習アルゴリズムをリアルタイム化したものである。

ある構文パターンの言い換えパターンは、大規模な Web コーパスからパターンの変数に当てはまる単語対を検出し、それらの単語対の相対的なオーバーラップを計算することで獲得できる。例えば、「X で Y が治る」と「X で Y を治療する」という 2 つのパターンは X と Y の変数に頻出する共通の単語対（例えば、「ステロイド剤、アトピー」）が多ければ多いほど、これらの構文パターンがお互いの言い換え表現となっている可能性が高いと考えられる。似た文脈に出現する語は似た意味をもつというのは、分布仮説（Harris [3]）と呼ばれる言語学におけるよく知られた仮説である。

一方、クラス依存の構文パターンとは、変数として取れる単語の意味クラスに制約を掛けた構文パターンである。構文パターンにクラス制約を掛けることでパターンの多義性が解消できる。例えば、「Y のための X」というパターンは「Y：病名のために X：薬品」のように、X が病名、Y が薬品の意味クラスの単語の場合は、X と Y の治療関係を表し、上記のパターン「X：薬品で Y：病名が治る」の言い換えパターンとみなせるであろう。一方、「X：作業のための Y：道具」の場合は手段／道具という意味的関係を表現する。このようにしてパターンの類似度を計算する際に共起する単語を特定の意味クラスに限定することで、パターンの曖昧性が大きく減らされ、高頻度で曖昧なパターンが活用可能になり、より大量の関係インスタンス（単語対）を獲得できる。

このような意味クラスは、文献 [1][2] と同様に、[4] で提案された単語クラスタリング法によって自動獲得する。この手法では大規模 Web コーパスから得られる名詞と動詞の係り受け関係の統計データを用いて、名詞の隠れクラスへの事後確率の分布を求める。ある名詞の所属確率が 0.2 以上の隠れクラスを、その名詞の意味クラスとする。現状では一休は 100 万名詞を 500 クラスに分類したクラスタリングデータを用いる。これらの意味クラスは言い換えパターンの認識以外でも、例えば有望な回答候補の意味的クラスの推定にも活用されている。

2.2 推論により生成した仮説により回答

2.1 では一休が Web 上に書かれている回答を網羅的に抽出するための言い換え獲得について説明した。それを用いることでユーザの質問から表層上大きく異なる表現で書かれている回答が抽出可能になる。しかし、Web 文書がどれほど大量でも、一文内で明記されていない有用な知識は当然あるであろう。そのために一休は、2 つの異なる Web ページから得られた情報を組み合わせる

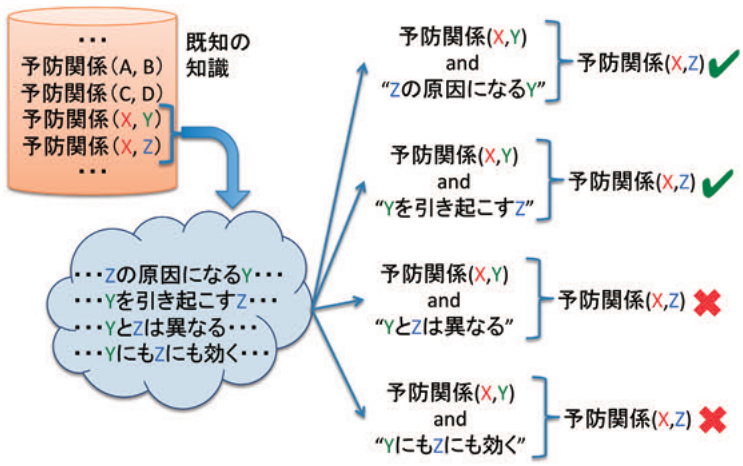


図4　推論規則の自動学習方法（例）

ことで、「人間なら導ける」というような回答も模索する。言い換え表現の自動獲得に基づいた回答抽出アルゴリズムが発見に失敗した回答に関しては、文献[5]で提案された推論過程を用い、仮説として多数生成している。例えば、“if X が Y の原因であり、Z が X を予防する then Z は Y の予防に繋がる可能性がある”といった推論規則を自動発見し、異なる Web ページに見つかった「ダークチョコレートが動脈硬化を予防する」と「脳梗塞の原因となる動脈硬化」という情報から「ダークチョコレートが脳梗塞の予防に繋がる」ことをダークチョコレートに関する好ましい副作用の仮説として生成している。この副作用は、一休の入力となった Web アーカイブでは比較的知られていなかったが、現在では多くの Web ページが取り上げている。

一休の推論過程は、推論規則の自動学習フェーズとその適用による推論フェーズからなる。下記に各ステップの概要を簡単に紹介する[5]。

1．推論規則の自動学習。推論規則の自動学習のため、文献[1][2]の意味的関係獲得手法を用い、「因果」、「予防」などという特定の意味的関係のインスタンスを正例として用意する。これらの関係インスタンスは単語対から成り、ある単語を共有する単語対に着目する。例えば、予防関係の場合に正例のインスタンスは「コーヒー、眠気」と「カフェイン、眠気」という単語対を含むとする。そうしたら「コーヒー」と「カフェイン」という単語もある種の意味的関係を持つと仮定し、その関係を記述するかもしれない構文パターンをコーパスから抽出する。図4が示すように、そうした構文パターンから「if 予防関係（A，B）and “C が A を含む” then 予防関係（C，B）」などという、予防関係に関する推論規則の候補を大量に自動生成する。これらの推論規則は、入力となった意味的関係の正解データをどれだけ再現できるかにより自動評価し、スコアリングを行う。スコアの高い推論規則は信頼できる確からしいものと見なす。

2．推論規則の適用による推論。次に、自動学習した推論規則を Web コーパスに適用することで、ターゲットの意味的関係の新規インスタンスを仮説として生成する。図5が示すように、多くの信頼できる推論規則から生成される仮説は確からしいと考え、仮説の信頼度をその仮説を生成した推論規則のスコアの和として計算する。例えば図5では、「ダークチョコレート」と「脳梗塞」が予防関係にあるという仮説は、「予防関係（X＝ダークチョコレート，Y＝動脈硬化）＜Y＝動脈硬化が Z＝脳梗塞を起こす＞→予防関係（X＝ダークチョコレート，Z＝脳梗塞）」、「予防関係（Y＝ポリフェノール，Z＝脳梗塞）＜Y＝ポリフェノールを含む

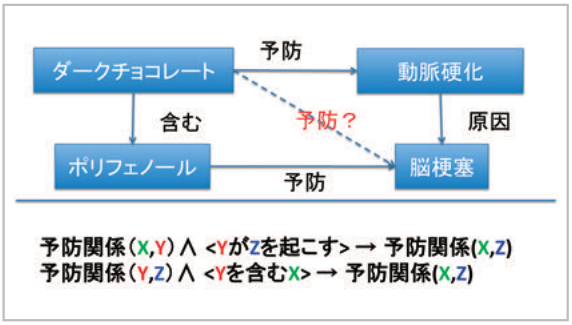


図 5 推論規則の適用による推論（例）

X＝ダークチョコレート＞→予防関係（X＝ダークチョコレート，Z＝脳梗塞）」など、複数の推論規則に生成され、信頼性の高い仮説と見なされる。ちなみに、以上で述べたダークチョコレートと脳梗塞の関係は、我々が入力として使った Web コーパス（2007 年当時のもの）では広く書かれている関係とは言えず、一文で直接的に記載されることはなかった。つまり、現在の一休の構文パターンによって抽出することは不可能であった。一方で、その後、この関係はマスコミ等でも大きく取り上げられ、現在では Google 等によってこの関係を記載した文書を大量に見つけることができる。つまり、このダークチョコレートと脳梗塞の関係を我々の手法は「先取り」していたと言っても良いかもしれない。

こうした推論に基づいた意味的関係獲得手法は、まだ初歩的なレベルとはいえ、上記のダークチョコレートの例が示すように一文内に明記されていないけれど有用な回答を仮説としてユーザに提供することが既にできる。しかしながら、現在の仮説生成技術は単語の間の関係など極めて限定された対象にしか有効でない。今後、適用範囲を広げるべく、現在研究を進めている。1 つの可能性としては、単語間の関係ではなく、フレーズ間の意味的関係からユーザにとり有用な情報を得る手法がある。これに関しては、すでに成果が出始めているところである [6]。

3 多様な質問への対応

ここでは、言い換えパターンの自動認識技術を利用した一休のコア技術を、より多様な質問に対応させるための取り組みについて紹介する。

3.1 クエリパターンを複数含む質問

これまでに説明してきた手法（一休コアシステムと呼ぶ）を単純に利用した場合、回答可能な質問は、「デフレを引き起こすのは何ですか」のように、1 つの名詞（デフレ）と疑問詞（何）の間に述語等で表される関係を持つものに制限されてしまう。このような制限下では大量の文書に対してパターン、名詞等で検索をするコストも低く、億単位の Web 文書であっても高速に回答を抽出できる。一方でそうした単純な意味的な関係と捉えることができない質問に対してはそのような高速の検索をすることが困難になる。以下では、このような状況を踏まえ、複雑な質問に対して一休コアシステムを利用して高速に回答を得る手法について説明する。

基本的なアイデアは、複数の名詞間の意味的関係を含む質問が入力された場合、質問を一休コアシステムで回答可能な部分質問に分割して、それぞれの部分質問の回答を求め、それらの統合により質問が意図していた回答を獲得する。処理の流れを図 6 に示す。

(1) 部分質問の生成

入力された質問文を、クエリパターンが 1 つだけ含まれる質問文に分割する。質問文の分割では構文解析の結果を利用し、名詞と疑問詞の間の係り受け関係のパスから成る構文パターン全てを部分質問として取得する。例えば、「日本が中国から輸入しているのは何ですか」という質問は、(A)「日本が輸入しているのは何ですか」、(B)「中国から輸入しているのは何ですか」という 2 つの部分質問に分割することができる（図 7）。

(2) 部分質問の回答獲得

部分質問の回答の獲得には、**2** で説明した一休コアシステムを利用する。まず、質問文を分割して得られた部分質問から、回答を取得するための基となるクエリパターンを生成する。次に、得られた部分質問のクエリパターンから、同じ文脈で用いられる可能性のある拡張パターンを取得する。これらの拡張パターンを利用して意味的関係のインスタンスを検索し、部分質問の回答を獲得する。図 8 に部分質問の回答例を示す。

(3) 部分質問の回答の統合

部分質問から得られた回答候補を統合して、元の質問の回答を取得する。最も簡単な方法は、それぞれの部分質問の回答集合の積集合を求めることである。しかし、部分質問が同じ回答を持つ場合でも、回答の根拠は別の文書の文脈から得られた可能性があり、元の質問の回答として必ずしも正しいとは限らない。例えば、「製品Pを日本が輸入している」という記述が文書1にあり、「中国が製品Pを輸出している」という記述が文書2にあれば、両国間で貿易があっても不思議ではないが、政治情勢やその他の要因で、直接取引しているとは限らない。

そのため、それぞれの部分質問の回答が得られた文書と文を特定することにより、回答とする優先順位を決める。もし、部分質問の回答が同一文から得られていれば、その優先度を最も高くする。例えば、図8の部分質問（A）「日本が輸入しているのは何ですか」と（B）「中国から輸入しているのは何ですか」が同じ回答「電化製品」を出力していた場合、それぞれの回答の根拠が、同一の文、同一文書、異なる文書のいずれから得られたかによって、順に回答の優先度を高くする。さらに、それぞれの部分質問に同じ回答が見つからない場合でも、ある部分質問の回答の根拠となる文の周辺に、他の部分質問に含まれる名詞が出現していれば回答リストに加える。例えば、回答「レアメタル」が部分質問（A）の回答リストだけに存在していたとしても、その回答の根拠文の周辺に（B）のクエリパターンの引数の名詞「日本」が現れていれば、回答「レアメタル」を最終回答に追加する。このようにして部分質問の回答候補を統合することにより複数のクエリパターンを含む複雑な質問に回答することができ

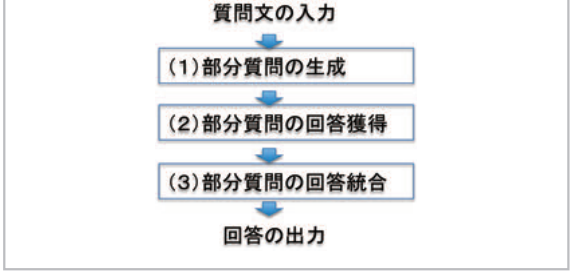


図6 クエリパターンを複数含む質問の処理フロー

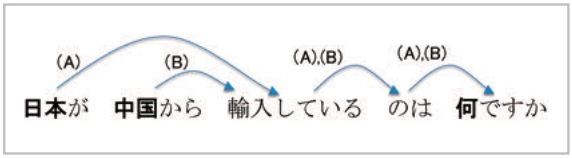
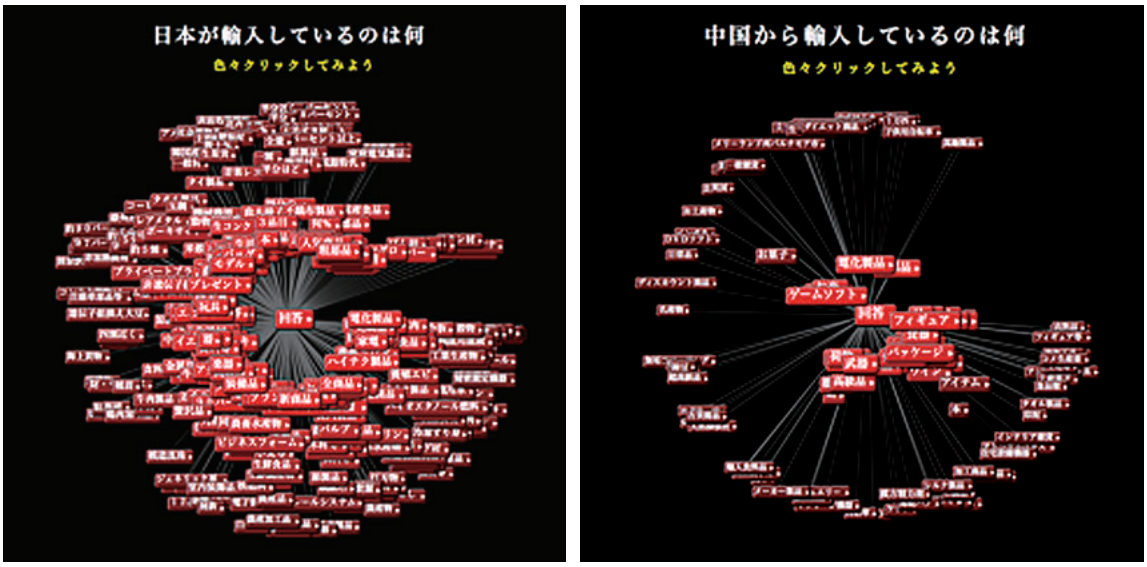


図7 構文解析結果による質問文の分割



部分質問（A）　　部分質問（B）

図8 部分質問の回答例

図 9　統合した質問の結果例

る。図 9 にその回答結果を示す。

3.2　主題語による回答フィルタリング

「日本が輸入しているものは何ですか？」のような回答の種類を問わない質問が入力された場合、一休の回答には、「椎茸」から「濃縮ウラン」まで幅広い名詞が含まれる。しかし、あるカテゴリに関する回答のみが欲しい場合、人は「日本が輸入している**食べ物**は何ですか」のように回答を限定した質問を入力するであろう。このように回答候補のカテゴリを指定する質問にも対応するため、ここでは、上記の質問の「食べ物」のようにユーザが欲している回答を制限する語（以後、主題語と呼ぶ）を用いた回答のフィルタリング処理について説明する。

3.2.1　同義語辞書による主題語拡張

主題語は、質問に対してユーザが欲している回答の範囲を定める働きがある。「日本が輸入している食べ物は何ですか」という質問に対して、「タオル」や「電化製品」という回答は適切でないことがわかる。そこで、質問文を構文解析し、疑問詞に直接係る名詞がある場合には、その名詞を主題語として取得する。上記の質問では、名詞「食べ物」から疑問詞「何」への直接の係り受けが存在するため、「食べ物」を主題語として取得する。次に、高度言語情報融合フォーラム ALAGIN（www.alagin.jp）で公開されている同義語辞書（言語資源 A-9：基本的意味関係の事例ベース）を利用して、主題語と同義の名詞を獲得する。例えば、「食べ物」からは、異表記の名詞「食べもの」「たべもの」と、同義語「食物」を得ることができる。同義語辞書についてのより詳しい説明は、本特集号の論文 **5-5**「基盤的言語資源」を参照されたい。

3.2.2　主題語による回答フィルタリング

同義語辞書により拡張した主題語をもとに回答の絞り込みを行う。回答のフィルタリング処理には、単語の文脈類似度、単語の上位下位概念辞書を利用する。主題語とこれらの言語資源を利用し、回答となるべき範囲を定めることで、適切な回答のみを出力する。現状では、過剰な回答のフィルタリングを防ぐため、下記の処理のいずれにも合致しない場合にのみ、回答リストからの削除を行う。

(1) 文脈類似度を利用した回答フィルタリング

Web 6 億文書から得られた係り受けの確率的クラスタリングを用いた文脈類似度を利用する。分布仮説 [3] とは「似た文脈に出現する語は似た意味をもつ」という、言語学におけるよく知られている仮説であり、これに基づいて計算した語間の意味的類似度を語の文脈類似度という。本研究では、文献 [7] が提案した類似尺度に基づいて構築され、高度言語情報融合フォーラム ALAGIN で公開された「文脈類似後データベース」を用いる（本特集号 **5-5**「基盤的言語資源」参照）。例えば、「食べ物」に対して分布が高い名詞には、お菓子、酒、魚、肉、ワイン、コーヒー、ビール、チョコレート、バナナ、キノコなどが含まれている。これらの名詞に比べて、食べ物と関連性の薄い「タオル」や「電化製品」の文脈類似度はかなり低い。このように文脈類似度を利用することにより、主題語とは異なる文脈で使用される、主題語と関連性の薄い回答を省く事ができる。

(2) 上位下位概念辞書を利用した回答フィルタリング

Wikipedia から獲得した上位下位概念辞書を利用する（ALAGIN の言語資源 A-4：上位語階層データ、[8]）。例えば、「食べ物」の下位概念の「果物」「キノコ」「魚」「海産物」「日本酒」「ケーキ」などを取得し、更に「果物」の下位概念より、「サクランボ」「イチジク」「アールスメロン」などの具体的な果物の名前を取得することができる。このように主題語の下位概念の名詞を再帰的に取得することで、主題語の下位概念の回答のみ

を取得することができる。

4　一休の音声インターフェースについて

一休は経済、健康、哲学から趣味、観光、アニメ、料理などまで、幅広いドメインの質問文をスマートフォン経由で質問を受け取り、Web 6億文書から回答を探し出す。スマートフォンなどを入力デバイスとして利用する場合、オープンドメインの質問を正確に音声認識できる言語モデルの構築が重要かつ解決が必須な課題となる。ここでは、**2**で説明した一休のコアシステムに対応する音声インターフェースについて説明する。

言語モデル構築の従来研究のほとんどは、対象アプリケーションのドメイン（例えば観光や医療）や入力される文の形式で制限された、人手で構築されたコーパスの存在を前提とし、そこにWebから類似データを追加することで言語モデルを作成している[9]-[11]。一休が対象としている質問文の形式は、**2**で説明したようにクエリパターンを1つ含む形式（以後、こうした形式を単に入力のスタイルと呼ぶ）であるものの、観光、医療などのドメインによっては縛られず、つまるところオープンドメインである。一休の言語モデルを作成する際には、まずスタイルに合致する質問文を人手で集めてシードコーパスとし、それに類似する文をWebから自動収集して新たなコーパスを構築し、そのコーパスからオープンドメインの言語モデルを構築した。

こうした手法はドメインを限定した音声認識器の言語モデル構築ではある程度有用であることが分かっているが、一方で一休の場合のようにスタイルは限定されているものの、オープンドメインである場合に有用であるかどうかは分かっていなかった。まず1つ予想される問題は、シードコーパスに現れる語彙はオープンドメイン、すなわちWeb全体に現れる語彙に比べて極めて少数であり、いくらシードコーパスに類似する文をWebから自動収集すると言っても、結果として得られるコーパスがカバーする語彙には限界があるのではないかということである。この問題に対処するために、語彙の範囲を広げる目的でシードコーパスにある名詞を意味的に類似すると思われる名詞[7]で自動的に置き換えることによってシードコーパスを拡張する。その結果としては、Webコーパスから得るより効率的に幅広い語彙を含む、なおかつ一休が求めるスタイルに合致した質問文を大量に収集でき、またそうして得られたシードコーパスに既存のドメインアダプテーション手法[9]を適用することで、低コストで高性能な言語モデルを作成できた。

以下にこの手法をより具体的に説明する。ドメインアダプテーション手法[9]はシードコーパスから得られたN-gramを基にWeb中の文のperplexityを計算してシードコーパスと傾向が類似している文をWebから収集する。言語モデルを作成した際にはまずスタイルに合致する、様々なトピックをカバーする500文から成るシードコーパスを手作業で構築した。次にこのシードコーパスとWebコーパスを入力として受け取り、以下のように処理を進める。

1．**3**と同様にALAGINで公開された「文脈類似後データベース」を用い、シードコーパスのすべての文に対して、その文が含む名詞を類似度の高い単語上位k個と置き換える。新しく得られた文をシードコーパスに追加する。
2．拡張されたシードコーパスとWebコーパスに文献[9]の手法を適用し、学習コーパスを構築する。
3．既存ツール[12]を利用して学習コーパスから音声認識用言語モデルを作成する。

音声認識器ATRASR[12]を利用した評価実験で、提案手法の言語モデルの語彙数は41万語であり、単語誤り率は15.49%であり、文誤り率は54.73%である。この値はWebコーパスからランダムに抽出した文によって構築したベースライン言語モデルより3.25ポイント（単語誤り率）、及び4.28ポイント（文誤り率）低い誤り率である。表1は正しく認識される質問文の例を示す。本手法で構築された音声認識言語モデルを用いることで、高精度な音声質問応答システムが得られることが分かった。なお、この表1には一休のコアシステムが回答できるよりも複雑な質問も含まれている。より詳しくは文献[13]を参照されたい。

表 1 正しく認識された質問文の例

はやぶさは何年ぶりに地球に帰還した？
最近発売されたソニーの学習リモコンの型番は？
板付遺跡はどこにありますか
東京ディズニーランドの最寄り駅はどこですか
5 月の誕生石を教えて下さい
熱中症の初期症状は？
国勢調査は何年おきに実施される？
ステロイドの副作用にはどんな物がありますか
かいけつゾロリの作者はだれ？
ウインブルドンで優勝した人はだれ？
ルイ 14 世の業績は何ですか
日本で iPhone はどれ位売れていますか
ポストモダンとは何ですか
Java の最新バージョンは？

5 質問応答研究における一休の位置づけ

近年では、検索エンジンや質問応答システムなど、情報へのアクセス手段の進歩が目覚ましい。たとえば、質問応答システムとしては IBM 社の Watson [14] が注目を浴びている。

Watson は米国において Jeopardy というクイズ番組の人間のチャンピオンに圧勝し一躍有名になったが、その稼働にはスパコンが必要とされる、さらには Jeopardy の質問に対してチューニングがなされるなどと言われている。一方で、一休は、データの更新さえ考慮しなければ、サーバー 1 台でほぼリアルタイムで回答を億単位の Web ページから抽出することが可能な非常にシンプルな構造を持つシステムであり、また、特定のタイプの質問へのチューニングなどは行っていない。今後、我々はこうした一休の特色を活かし、すでに NICT で公開されている情報分析システム WISDOM（www.wisdom-nict.jp）のサブモジュールとして数十億単位の Web ページからユーザの多様な質問に答える役割を担うべく開発を継続している。また、同じく NICT の耐災害 ICT 研究センターにおいて、災害時に発生する膨大なネット情報の中から、孤立している被災地、必要とされる物資、提供されている物資、支援情報などを迅速に抽出し、支援、復興に役立てるシステムとして公開する予定である。

また、本稿の冒頭で述べたように、ユーザのふとした疑問を意外でありながら有用な情報の発見に結びつけることが一休開発の最終目標であった。この目標は、人間には回答が難しいとは言え、一意の回答がある質問に対して一意の回答を正確に出力するという Watson の設定とは異なり、回答があるかないか不明な質問に対して仮説としての回答候補を出したり、見つかる限りの回答をすべて列挙するなどの機能が必要になる。こうした機能を実現すべく、一休は開発されてきた訳である。現在は、こうした目標の実現にさらに近づくべく、さらに強力な仮説生成機能やユーザがある時点までにした質問や回答の閲覧の履歴等から、さらに有用な質問、回答を推薦する機能、さらには現状 Watson でも回答することのできない「文章による回答を要する」質問、すなわち、いわゆる「Why 型質問」や「How 型質問」への対応をすべく研究を進めている。この中でも Why 型質問への対応についてはすでに一休にプロトタイプが組み込まれており、例えば「ガダルカナル島で米軍に負けたのはなぜですか」といった音声の質問に対して、兵力の逐次投入、基地からの距離など、多数の「負けた理由」に言及している Web ページのパラグラフを回答として提示できる。さらに、本稿中でも軽く触れたが、「円安になる」⇒「輸出が増える」などの文、フレーズ間の因果関係など、いわゆる世界知識と呼ばれる知識も Web ページなどから大量に獲得、蓄積が可能になってきており、今後はこうした知識を活用して、より有用な仮説の提示や、より有用な質問、回答の推薦を行う機構を開発して行く予定である。

最後にこうした有用な仮説の具体例を 1 つ挙げる。2010 年に中国政府は日本との間の領土問題に関連して日本に対するレアアースの輸出を停止した。我々は別の物質の輸出停止が続くものと予想し、日本が中国に依存している原料、その原料を含む製品、さらにその製品を作っている企業を一休で調査することにした。この結果、日本がタングステンを中国から輸入しており、ある日本の大企業がタングステンを超硬工具の製造に使用していることが判明した。このことから、中国がタングステンの対日輸出を停止し、例の企業が超

硬工具の製造で問題を抱えるという仮説を作成し、研究報告で取り上げた。一週間後、日経新聞は「タングステン、レアアースの二の舞も」というタイトルの記事で中国政府がタングステンの値上げを通告したニュースが、例の企業の代表者とのインタービュー付きで記載された。我々が考えた仮説が現実に一定程度追認されたことになるが、こうした高度な仮説に至るために人間による手作業が必要であった。つまり、一休は「日本がタングステンの輸入で中国に依存している」、「タングステンは超硬工具の製造に使用される」、「超硬工具はどの企業が製造しているか」など、その仮説の生成に必要な部分情報は取得できるが、この部分情報を統合して仮説を生成するには人間の手助けが必要であった。今後、こうした仮説を全く人手を介さずに自動生成できるように研究開発を進めていく予定である。タングステンのストーリーのように、現実と合致する仮説をピンポイントで自動生成することは極めて困難であるが、多数のありそうな仮説をユーザに提供することは可能であろう。さらには、望ましくない仮説に対してユーザが取りうる様々な対策案の自動生成も検討していきたい。例えば、超硬工具に依存している企業には、他の製造業者と代替の入手経路の交渉をするなどの対策が考えられる。このようにして、一休を単なる質問応答という機能にとどまらず、広くユーザの相談役、ガイド役へと進化させていくことが研究の次のステップと考えている。

6　おわりに

本稿ではユニバーサルコミュニケーション研究所情報分析研究室が開発してきた「Web に答えさせる」音声質問応答システム「一休」を紹介した。一休は、人間がそもそも把握できない量の情報の意味解析を行い、得られた情報を柔軟に組み合わせることで未知なる有用な仮説を回答として生成でき、ユーザの多様な情報ニーズに応えることができる。

いわゆる情報爆発が収束する兆しが見えない現状では、有用な情報、知識へのアクセスを改善することが直接、個人と社会の適切な意思決定の質の向上につながると考えられる。一方、現在の検索エンジンのように、ユーザの与えたキーワードを含む大量の Web 文書を単純にリストアップするだけのシステムはそういった意思決定で必要な情報収集を十分に行えないことが明らかになってきた。我々は一休の開発を通して、そうした意思決定の質の向上、さらには意思決定のための情報収集の効率化に貢献して行きたいと考えている。

参考文献

1 Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata, “Large Scale Relation Acquisition using Class Dependent Patterns,” in Proceedings of the IEEE International Conference on Data Mining (ICDM'09), pp. 764–769, Miami, Florida, USA, Dec. 2009.

2 De Saeger Stijn，鳥澤健太郎，風間淳一，黒田　航，村田真樹，“単語の意味クラスを用いたパターン学習による大規模な意味的関係獲得，” 言語処理学会第 16 回年次大会．2010.

3 Zellig Harris, “Distributional Structure. In Word 10(23),” pp. 142–146, 1954.

4 Jun'ichi Kazama and Kentaro Torisawa, “Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations,” In ACL08-HLT: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 407–415, 2008.

5 Masaaki Tsuchida, Kentaro Torisawa, Stijn De Saeger, Jong Hoon Oh, Jun'ichi Kazama, Chikara Hashimoto, and Hayato Ohwada, “Toward Finding Semantic Relations not Written in a Single Sentence: An Inference Method using Auto-Discovered Rules,” In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, Nov. 2011.

6 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama, "Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web," Proceedings of EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning, 2012.

7 Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa, "A Bayesian Method for Robust Estimation of Distributional Similarities," In Proceedings of ACL 2010, pp. 247–256.

8 Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida, "Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures," EMNLP'09, 2009.

9 Teruhisa Misu and Tatsuya Kawahara, "A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts," In Proceedings of Interspeech 2006, pp. 9–13.

10 R. Sarikaya, A. Gravano, and Y. Gao, "Rapid Language Model Development Using External Resources for New Spoken Dialog Domains," In Proceedings of ICASSP 2005, Vol. I, pp. 573–576.

11 Mathias Creutz, Sami Virpioja, and Anna Kovaleva, "Web augmentation of language models for continuous speech recognition of SMS text messages," In Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 157–165.

12 S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," IEEE Transactions on Information and Systems vol. E89-D(3), pp. 989–997.

13 Istvan Varga, Kiyonori Ohtake, Kentaro Torisawa, Stijn De Saeger, Teruhisa Misu, Shigeki Matsuda, and Jun'ichi Kazama, "Similarity Based Language Model Construction for Voice Activated Open-Domain Question Answering," In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, Nov. 2011.

14 Ferrucci et al., "IBM Research Report: Towards the Open Advancement of Question Answering Systems," http://domino.watson.ibm.com/library/CyberDig.nsf/papers/D12791EAA13BB952852575A1004A055C/$File/rc24789.pdf

（平成 24 年 6 月 14 日 採録）

DE SAEGER Stijn
ユニバーサルコミュニケーション研究所
情報分析研究室主任研究員
博士（知識科学）
自然言語処理、知識獲得
stijn@nict.go.jp

後藤（ごとう） 淳（じゅん）
ユニバーサルコミュニケーション研究所
情報分析研究室専門研究員
自然言語処理、情報抽出
goto-j@nict.go.jp

VARGA István
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（工学）
自然言語処理、情報抽出
istvan@nict.go.jp