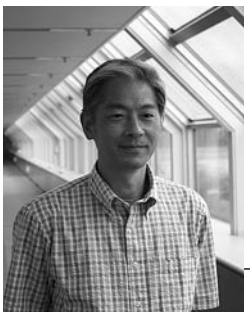
質問応答システムWatsonが示す未来

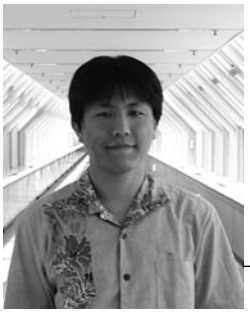
―質問応答技術がもたらす情報処理の新たな世界―



日本アイ・ビー・エム株式会社
東京基礎研究所
技術理事

武田 浩一
Koichi Takeda

【プロフィール】
1983 年、日本 IBM 入社。以後、東京基礎研究所において自然言語処理やテキスト・マイニングの研究開発に従事し、英日機械翻訳システムやテキスト・マイニング・ツールの研究開発に貢献。現在は電子カルテなど医療情報のマイニングや新しいビジネス・インテリジェンスの実現に取り組んでいる。2007 年 12 月より Watson プロジェクトに参加。



日本アイ・ビー・エム株式会社
東京基礎研究所
主任研究員

金山 博
Hiroshi Kanayama

【プロフィール】
2000 年、日本 IBM 入社。以後、東京基礎研究所において構文解析・意味解析など自然言語処理の基礎的技術や機械翻訳、テキスト・マイニング、文書校正などの応用の研究に従事し、現在に至る。2007 年 12 月より Watson プロジェクトに参加。

PROVISION 前号で速報として、IBM 基礎研究所が開発した Watson と命名された質問応答システムが、米国の有名クイズ番組「Jeopardy!（ジョパディ!）」に登場し、最高賞金額を獲得したことをご報告しました。この Watson で活用されている技術が「質問応答（question answering）」と呼ばれる技術です。

質問応答とは、知りたい対象についての記述を質問文として受け取り、その解答を提示するというタスクです。従来の情報検索では、質問文あるいは検索条件はキーワードの集合として扱われ、それらのキーワードを含む文書のリストが提示されるだけでした。質問応答は、情報検索よりもさらに踏み込んで、大量の情報源から、与えられた質問文を満足する解答を計算するところが大きく異なります。

Watson には IBM 基礎研究所の 25 名の研究者が 4 年の歳月を費やして世界最高の質問応答技術を搭載しました。本稿では、このような技術的革新がどんな意味を持ち、情報処理の世界をどのように変え得るのかについて詳しくご説明します。

Future of Information Processing Demonstrated by the Question-Answering System, Watson

- Question-answering Technology: a New World of Information Processing -

The question-answering system, Watson, won a human vs. computer match-up at the famous US TV quiz show Jeopardy! on February, 2011. The backbone of this system is question-answering technology, the main theme of this article.

“Question-answering” is defined as being a task that involves finding the entity (the answer) to a given description of the entity (the question). This task is very different from traditional information retrieval, where the search question is given as a set of keywords, and the response is simply a set of documents including the given keywords. Question-answering goes one step further, as the correct answer to a given question is computed using a large number of information sources.

More than 25 IBM researchers took over four years create the greatest question-answering system in the world. In this article we will introduce to you what this technical innovation means to us, and how it will change the world of information.

グランド・チャレンジ

まずは Watson が開発された経緯をごく簡単に振り返ってみましょう。IBM 基礎研究所には「グランド・チャレンジ」と呼ばれる、極めて難しい技術的な課題を設定し、その解決に向けた研究に投資するというプログラムがあります。グランド・チャレンジの主な意義は、学術的進歩への貢献であり、既知の手法を改良するような漸進的な研究開発では達成できそうにない課題に対して、新しいアイデアや解決方法を提示することを目指します。また、情報科学の専門家でない一般の人々にとっても課題の内容が理解できるようにテーマが設定されます。近年のグランド・チャレンジでは、ビジネスに活用できる基礎研究への投資という企業戦略を反映して、実世界への応用シナリオが同時に考案されるようになってきました。Watson の研究が重視された最大の理由は、近年の情報爆発時代を象徴するテキスト情報（非構造情報）の驚異的な増加を大きな価値に転換できる有望な技術と位置付けられたことにあります。

Watson は、2007 年からこのグランド・チャレンジとして研究開発をスタートしました。そして、2011 年 2 月 14 ～ 16日の3日間にわたり、米国の有名クイズ番組「Jeopardy!」で、最多連勝記録を達成したケン・ジェニングス氏と、過去最高の累積賞金を獲得したブラッド・ラター氏の 2 名と対戦し、獲得賞金総額で両名を抑え首位となる快挙を成し遂げたのでした。IBM は、1956 年にアーサー・サミュエルが学習機能を持つチェッカーのプログラムを発表して以来、1997 年にはフェン・シュン・スーらの開発した Deep Blue がチェスの世界チャンピオンに勝利するなど、ゲーム分野への挑戦に不思議な縁があるようです。

図 1. Jeopardy! の設問パネル

表 1. Watson が対戦で正答した問題の例

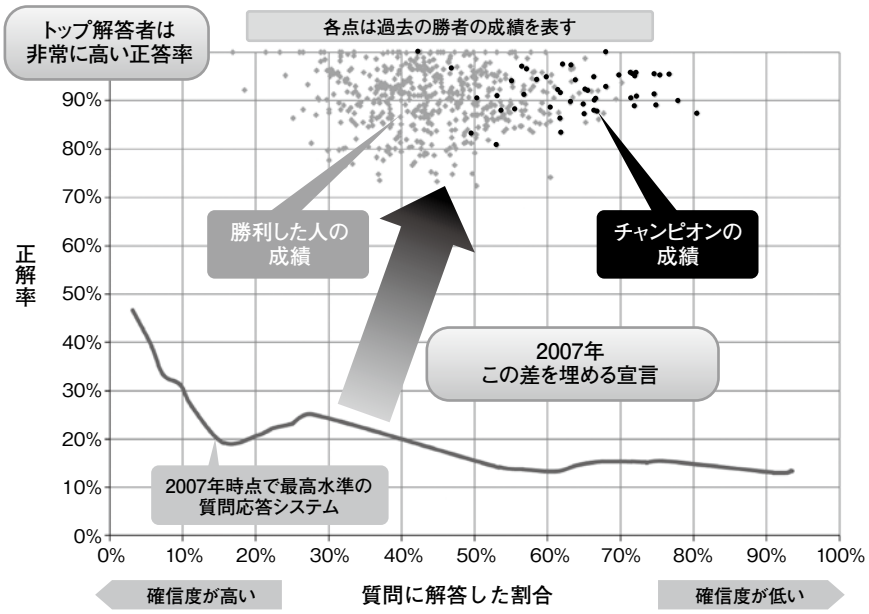
1	カテゴリー： Dialing for Dialects（方言について答えよう） 問題文：While Maltese borrows many words from Italian, it developed from a dialect of this Semitic language（マルタ語はイタリア語から多くの語彙を借りているが、それはこのセム語系言語の方言から発展した） 答え： Arabic（アラビア語）
2	カテゴリー： Alternate Meanings（2 つの意味を持つ単語） 問題文：4-letter word for the iron fitting on the hoof of a horse or a card-dealing box in a casino（馬のひづめに付ける金具、またはカジノでカードを入れる箱を表す 4 文字の語） 答え： Shoe

Watson が挑戦した Jeopardy! は、米国で 40 年以上の歴史を持つクイズ番組で、歴史、科学、文学、スポーツなどの幅広い知識を問われます。3 人の回答者が獲得金額を競う形式で、最初に一人が**図 1** のようなパネルの中から、最上段に表示されたカテゴリー（出題分野）と金額（難易度に応じて 5 段階）を指定すると、質問文が表示されます。**図 1** にあるように 1 ラウンドで 6 つのカテゴリーから各 5 問が出題され、計 30 問の解答を競います。ボタンを最初に押した人が解答し、正解すれば該当する金額が得られ、不正解なら同額を減らされます。このため、自信のないときには解答しないことが獲得賞金額を稼ぐ上で重要な判断となります。

表 1 に、カテゴリーと問題、それに対する答えの例を挙げます。この例を見て分かるように、カテゴリー自体が巧妙で、単に問題のジャンルを指定するというよりは、何らかのヒントや示唆を間接的に与えるようなものになっています。質問文も「米国の初代大統領は?」といったたぐいの単純なものではなく、娯楽性の高いクイズ番組として、視聴者が読んで謎解きを楽しめるような英文で表現されており、事前に質問文を予測して答えを用意しておくことはまったく不可能です。また一度使われた質問文が再度番組で使われることはまずないため、過去の問題とその解答を用意しても新たな対戦でそのまま使える可能性はゼロといっていいでしょう。

質問応答技術

Jeopardy! で人間と対戦するために必要なのは、「カテゴリー」と「質問文」とを入力として提示したときに「解答」を計算して出力するシステムです。冒頭で述べたよ



上部の点は過去の番組における勝者のゲームごとの成績を表す。曲線は、2007 年時点のシステムの性能。

図 2. 回答率と正解率のグラフ

うに、このようなタスクは従来から「質問応答（question answering）」と呼ばれており、さまざまな研究がなされてきました。Jeopardy! のように質問応答の対象となる分野を事前に制限しないものは、オープン・ドメイン質問応答と呼ばれ、1999 年に初めて公式のタスクとして国際会議などで技術的な評価が試みられるようになりました。

Watson を開発した研究者たちは、このようなタスクを評価する国際会議で長年研究を続けており、タスク参加チーム中では上位の成績を収めていました。

図 2 には、当時のシステムの性能と、Jeopardy! で勝利した出場者の成績を比較したチャートを示しています。ここで、横軸には「回答率（どれだけの質問に答えようとしてボタンを押すか）」を、縦軸には「正解率（押した時にどれだけの割合で正解するか）」という 2 軸の指標が対応しています。図中央上部の薄い点は、1 回のゲームで勝利した人の成績を、右上部の濃い点は最多連勝記録保持者のケン・ジェニングス氏の各ゲームでの成績を示し、図の下部にある曲線が、2007 年当時のシステムの性能を示しています。このチャートからは、システムが最も自信をもって解答できる質問文（全体の 5%）に対しても正答率は 40%程度に過ぎず、全問題に解答しようとすれば正答率は 10 数%程度しか達成できないという悲惨な事実が明らかになりました。このように、当時のシステムの改良では出場者のレベルに到達することは不可能だと判断されたため、Watson は従来とはまったく異なるアプローチを取ることになったのです。

Watson のために新たに設計された手法が、**図 3** に示す、質問応答のアーキテクチャー「DeepQA」で、「情報源と統計情報を基に、解の候補の生成と根拠の探索を行う」という点が特徴になっています。**図 3** では、質問文とカテゴリーを入力して、解答と確信度を出力するまでの DeepQA の処理の流れが左から右に図示されています。以下では、この各部分を順に解説します。

(1) 質問文解析

英語で書かれた質問文から、何が問われているかを判断します。質問文の構文的な多様性はほぼ無限にあるため、特定の型に当てはめて解析することはできず、質問文ごとに正確な構文の解析が必要です。

表 1 の 1 番目の例の場合は、質問文を構文構造に変換した上で、以下の手掛かりを抽出します。

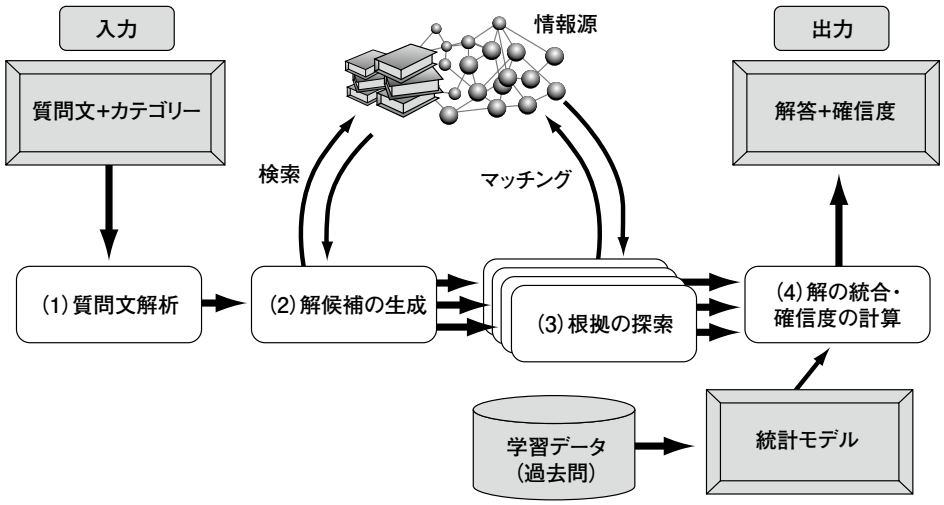


図 3. DeepQA フレームワークの概略

- マルタ語がイタリア語から多くの語彙を継承したこと（Maltese borrows many words from Italian）
- it がイタリア語ではなく、マルタ語を指すこと
- マルタ語がセム語系言語の方言から発展したこと（it developed from a dialect of this Semitic language）

さらに、指示詞 this で修飾された名詞句から、回答すべきものが “this Semitic language” であること、“borrows many words from Italian” という従属節の部分は解答に直接関係がないことなどを認識します。ここで、構文の解析だけでなく、文中の代名詞 it が マルタ語（Maltese）を指していると正しく解釈（照応解析）することも問題を解くために極めて重要であることにご注意ください。これらはすべて非常に高度な処理であるため、常に正しい解析ができるわけではないのですが、IBM 基礎研究所における機械翻訳などの自然言語処理分野における長年の研究成果の蓄積が、質問の正確な解析に大きく貢献しています。

(2)解候補の生成

次に、質問文に対する解答の候補を、大量の情報源の中から探して列挙します。質問の解析によって得られた手掛かりと同じ文脈（コンテキスト）に解答も現れやすいだろう、と仮定して候補を選定するのです。情報源には、ニュース記事、百科事典やその他のテキスト文書（聖書全文や歌の歌詞など）や、英語の語彙体系などの辞書といったものが含まれています。なお、Watson はクイズ番組の出場時にはインターネットに接続しないので、情報源はあらかじめ処理されてシステムに蓄えられています。

この段階で候補として正しい答えを見落としてしまうと、後段の処理で取り返すことができません。このため、質問文に含まれている語句と同時に現れやすい語句を検索したり、質問で問われている事物に該当する語を辞書から列挙したり、複数の手段で候補を補完しつつ選定します。その結果、以下の（3）で検討される候補の数は数百程度に達します。

(3)根拠の探索

候補の中から正しい答えを選択するために、各候補を元の質問文に「代入」して、仮説を生成し、これを情報源を利用して検証します。ある候補が質問文に対する解答であるなら、その「根拠」が情報源の中に見つかるはずだ、という前提に基づいています。もちろん、情報源の中に質問文と一語一句変わらない表現が含まれていることはまれなので、質問文を分解して重要な部分を抜き出したり、候補の語が持つべき意味的性質を列挙したりして、それらと適合する情報源の記述を探す必要があります。その合致（マッチング）の評価のための次元を「観点」、マッチしたものを「根拠」と呼んでいます。正しい解答に対して多くの根拠を見つけられるように、新たな観点を増やしたり、マッチングのアルゴリズムを改良したり、情報源を充実させたりすることが、性能向上に向けて取り組んだ研究開発の核となりました。最終的に観点の数は百以上に上りました。

このマッチングの処理において、(2) で列挙した全候補に対して根拠を探す必要があり、処理時間の高速化が不可欠でした。このために、各候補に対する処理を並列化して同時に実行しています。クイズ番組に登場した本番システムの環境では POWER7 アーキテクチャーの 2,880 コアを用いた超並列の環境で動作させ、出題から 3 秒以内に解答を計算するという高速化が実現されました。

(4)解の統合・確信度の計算

質問文から生成された各解候補に対して、(3) で見つかった根拠に応じた得点付けをします。正答につながりやすい強い根拠を持つ候補に大きな値が割り当てられるよう、それぞれの観点に「重み」を付与します。過去の Jeopardy! の問題と解答のデータ（過去問）数万件を用いて機械学習と呼ばれる手法により最適な重みの配分を計算しました。やや技術的すぎる説明かもしれませんが、過去問をその時点のアルゴリズムと情報源に基づいて解こうとした時に、正解率が最大となるように、各観点への重みがロジスティック回帰という手法により計算されています。このようにして、過去の問題に最高の正解率を達成した（観点の重みの配分を調整した）システムであれば、本番でも最も良い成績を上げることができるだろうと期待したのです。

各候補について、有効な根拠の重みを足し合わせて、その候補の確信度とします。答えの確信度の最大値が事前に設定されたしきい値以上の時に、Watson はボタンを押して回答をするように設計されています。

解答導出の例題

上記のプロセスでどのように問題を解いているのかをより直観的に理解するために、具体的な例を挙げて解説して

みます。実際に Watson が対応しているのは英語の質問文と情報源のみですが、ここでは理解を容易にするために、日本語の例を用いて説明することにします。

質問文:「本州の中で最も西に位置するこの県は、1871 年に発足した」
正答:「山口（県）」

Watson の処理の流れを説明する前に、人間ならこの問題にどのように解答するかをちょっと考えてみてください。

普通の日本人にとっては、上記の質問文は誤りなく理解でき、「本州」という言葉から、これが日本についての質問で、解答は日本にある 47 都道府県のうち、43 ある「県」の1つだということが分かるでしょう。仮に 1871 年に発足したかどうか自信がなくても、日本の地形的な特徴と、本州の最も西にあるという地理的条件を強い根拠として、すぐに「 山口 」県と答えられるか、あるいは解答を思い付かないか判断できるでしょう。

残念ながら、このような処理を直接システムで実現することはとても難しいのです。人間のような思考を再現するためには、例えば概念的に高度に組織化された情報と連想記憶による解答の発見のような仕組みが必要です。本州という単語の示す概念が、日本という国の主要 4 島の1つであり、県という単語の示す概念が日本の行政上の区分であり、その位置や隣接性などの地理的な属性を持つ、といった膨大な情報の体系化が不可欠です。コンピューター・システムにこのような体系を実装するだけでも、人手による定義と形式化、入力の作業を要し、現実的な期間やコストではとても完成できないでしょう。

また、仮にインターネット検索エンジンのような情報検索ツールがあったとしても、質問応答の実現には程遠いことが分かるでしょうか。上記のような質問文は、インターネット検索エンジンでは一般に『本州』『西』『位置する)』『県 』『1871』『 発足 』といったキーワードの集合として解釈され、これらのキーワードを含む文書が検索されます。ただし、文書から正解を見つける根本的な手段が提供されないため、後は人手で文書を順に読みながら正解を探し出さなければなりません。この意味で、従来の情報検索では質問応答を技術的に解決する手段となっていません。さらに、質問文中のキーワードを多く含む文書が必ずしも正解の記述を含む保証はありません。「1871 年に発足したこの県の西に本州が位置する」というまったく語順の違う質問文でもほぼ同じキーワード集合が含まれるため、検索される文書が質問文と関連の深い内容を含む可能性は一般にあまり高くないのです。

それでは、Watson ではこのような問題をどう処理しているのか確認してみましょう。

まず、質問文の中のキーワード、この場合「本州」「最も」「西」「県」「1871」などを検索条件として、情報源の中を検索し、それと一緒に出現しやすいキーワードを列挙します。すると、「広島」「山口」「鳥取県」「中国地方」「奥多摩」など、解候補が得られます。問われているもの（これを「質問の型」と呼びます）が「県」だということが分かっても、Watson には最初から日本の 43 の「県」だけを考えればよいと結論付けることができません。解答は日本の県に限るという知識が質問文には明示されておらず、Watson には本州に位置する県というものが実質的に日本の県を意味するということは容易には判断できないからです。このほかにも、質問の型が「作曲家」だったらどの集合を調べればよいか、「液体」なら、または「形式」なら一体何を調べるか…と考えていくと、質問の型とその解答になる語句の組み合わせには際限がなく、関連しそうな語句を大量に調べてみるほかはありません。

次に、これらの候補が答えとして適切かどうかを調べるため、情報源の中から根拠を探します。根拠を調べる観点には、「候補が、質問の型である『県』であるか?」「候補が、質問文中の制約『最も西にある』と記述されているか?」「質問文中と同じ時間表現と共に現れるか?」、「該当する語句への参照（リンク）が幾つあるか?」などが挙げられます。それぞれの候補について、これらの観点か

表 2. 解候補ごとの根拠の探索

観点 ＼ 解候補	広島	山口	鳥取県	中国地方	奥多摩
候補と質問で型が一致する？（「県」である）	○	○	○	×	×
条件の一部が一致？（最も西にある）	×	○	×	○	○
時間表現が共通？（1871 年の記述を含む）	×	○	×	○	×
該当する語句へのリンクの数（多い方がよい）	1300	500	200	150	10
総合点（確信度）	2%	92%	20%	6%	0%

ら根拠を見いだせるかを**表 2** に示します。

すべての観点で根拠を見いだせる解答は存在しないことが多いので、過去の問題から学習した重み付けに基づいた確信度を計算します。この例の場合、正解である「山口」に最も高い確信度が与えられました。なお、このほかに「山口県」という候補もあった場合は、それらを統合した上で確信度を計算します。最終的な確信度が十分に大きければ、ボタンを押し、それが対戦相手よりも速ければ回答ができます。

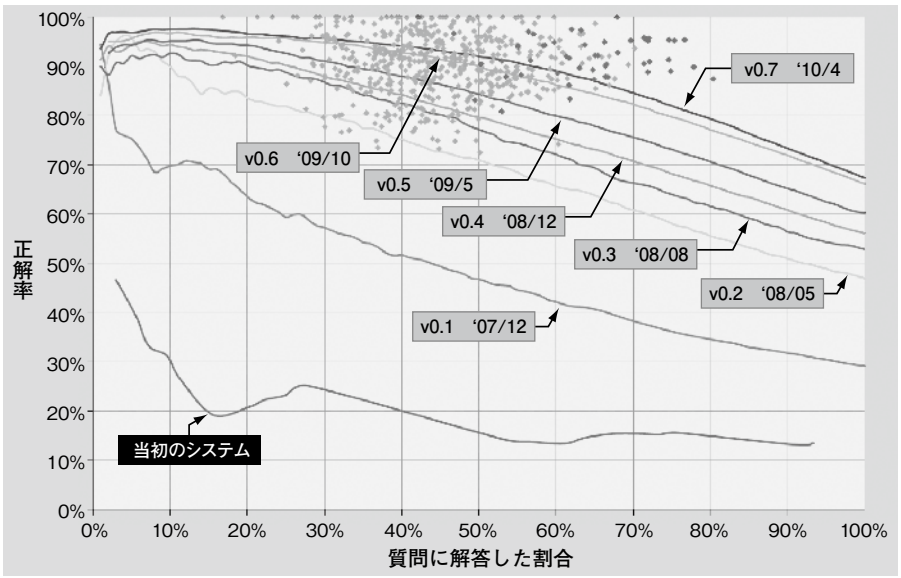


図 4. 性能向上の軌跡

このように、Watson は人間とはまったく違う処理の流れではあっても、人と同等の速度や精度で質問に応答するタスクを実現してみせたのです。

着実な性能向上

図 2 に示したように、開発当初は人間の能力には遠く及ばなかったものの、DeepQA の仕組みを設計・実装して、幾度もの実験を繰り返し、新しいアルゴリズム、必要なデータの検討を重ねることによって、**図4**に示すように性能は着実に向上していきました。2008 年の末には、過去のトップ回答者の一部のレベルを上回り、チャンピオンとの対戦のめどが付いたため、2009 年 4 月に IBM は Jeopardy! の対戦を公式に発表したのです。

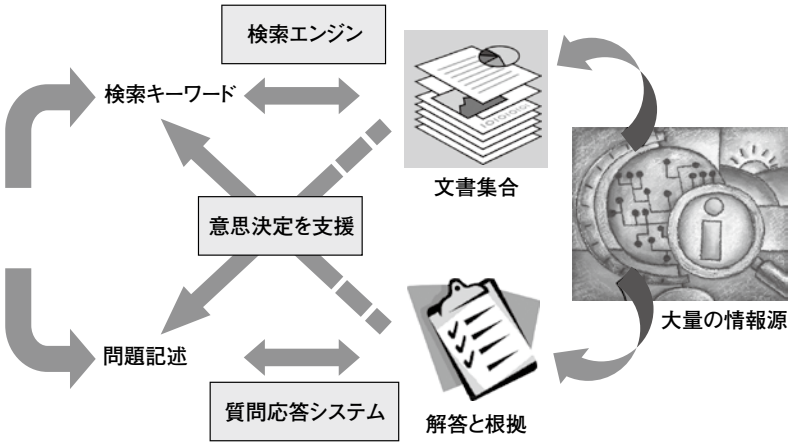


図 5. 従来の情報検索と質問応答との統合

質問応答技術の実用化

Watson によって培われた技術は、さまざまな分野での応用が期待されています。特に、従来の情報検索の手法と質問応答の組み合わせによって、これまで以上に多様な情報へのアクセスが可能になるでしょう。**図 5** にこのような考え方を示しました。

Watson が実現した質問応答技術を通じて、記述に基づいて対象物を検索する手法が確立されたといえます。つまり、症状と疾患、不具合と障害部位、テスト・ケースとテスト対象などの関連情報の検索に役立てることができますし、日常的な記述表現はそのまま検索質問として利用可能だと考えられます。これは情報アクセスの手段がさらに高度化されたということを意味します。従来の検索エンジンが、情報源へのアクセスを可能にしたとすれば、Watson は記述に合致する対象を求めて解答するという特徴を持ちます。この両方のアプローチを統合することによって、エンタープライズ向けの情報検索・意思決定の支援に役立てられると考えています。特定の記述に合致する対象が複数個存在し、判断に迷うような場合にも、Watson が計算した根拠・確信度の上位候補に基づいて、より信頼できる意思決定を支援できる適用事例もあるでしょう。このように情報にアクセスする手段が多様化・高度化することで、情報処理課題の解決能力が大きく向上したといえるでしょう。

ただし、Watson があれば世の中の質問

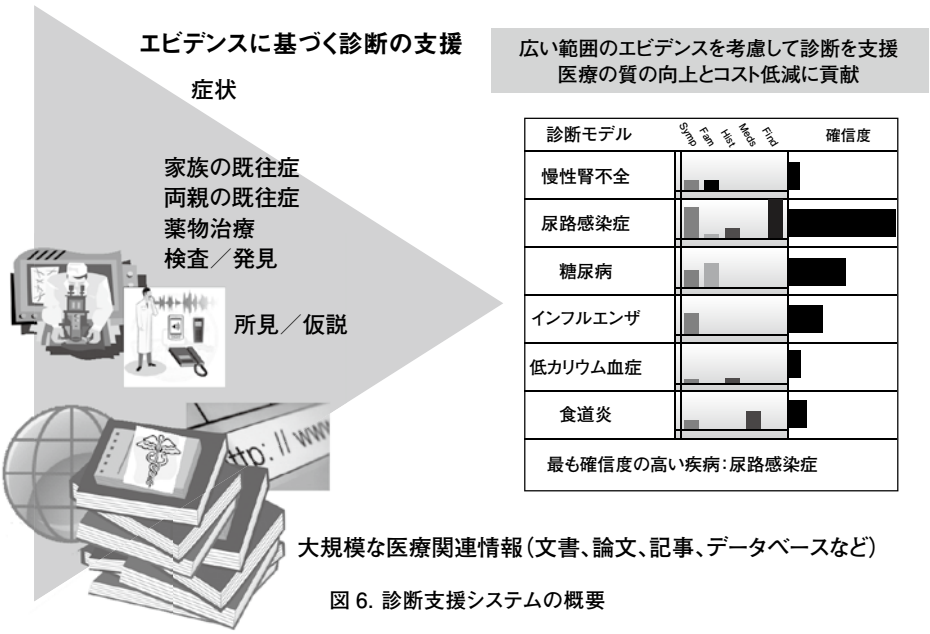


図 6. 診断支援システムの概要

に何でも答えられるというわけではない点には注意が必要です。クイズ番組への勝利で立証されたものは、「答えが 1 つに定まるような質問文が与えられた時に、一般的な知識が書かれたテキストを参照して解答を導く」という Watson の能力です。Watson が、明日の天気や、独自性を持った政治についての意見について答えられるわけではありません。逆に、決定的に解ける問題、例えば掛け算や辞書引きに対しては DeepQA のような仕組みは不要です。また、Watson が人間ではあり得ないような過ちを犯すことも分かっています。

Watson の本質を理解すれば、質問応答技術を活用してこそ解決できる、真に役に立つ課題が見つかるはずです。その例の 1 つが医療分野への応用です。

患者のカルテの情報、本人や親の病歴、血圧などの数値が入力として与えられた時に、その患者がどの病気であるかを推測するという課題がこれに当たります。情報源としては、過去のカルテ、医学に関する文献などが利用できます。このとき、1 つの病気を言い当てる必要はなく、複数の病気とその確信度を出力すればよいでしょう。

実際の医療の現場で、医師は自分の知識を基に診断を行っていますが、その際に本来の病気を見落としているという状況が存在するといわれています。このように、質問応答システムが持つ情報アクセスにより人間の活動を補助できる場面は、今後も多数考え出されるでしょう。**図 6** に現在研究されている、Watson を利用した診断支援システムの概要を示しました。

その他、Watson に使われた自然言語処理の要素技術は広い範囲に応用が可能です。例えば、英語の構文解析器は、Watson の開発を経て大幅に性能が改良されました。全体の質問応答システムのごく一部として働くものであっても、その達成度が客観的な数値で測れたおかげで、改良を進めることができました。このような構文解析や、その他の意味の解析を用いたテキスト・マイニングにより、人間では読み切れない量の文書から知識を抽出することや、事物の関係を知ることができるようになり、応用の幅はさらに広がるものと期待されています。

テキスト・マイニングの分野では、1990 年代にコールセンターの応対記録から顧客の声を分析するソリューションが大きく発展しました。その後、インターネット上の評判分析（例えば本誌 48 ページ以下の解説②にある風評の分析もこれに含まれます）など、ソーシャル・メディアと呼ばれる多様な特性を持つメディアが分析されるようになりました。ここでは、時系列的な情報、人と人とのつながりといったリンク情報などに対応できるように、より意味的な情報や、時間・位置・ネットワークなどの多次元的な分析能力を向上させています。Watson が実現した高度な情報アクセス能力と、テキスト・マイニング技術の向上により、多くの産業分野で、構造情報と非構造情報を統合した知的なエンタープライズ・ソリューションが登場することが予見できます。また、従来のデータベースに加えて、信頼できるテキスト情報を豊富に蓄積することが、業務の大きな差別化につながるという事例が周知されるようになるでしょう。IBM 基礎研究所がこのような技術の進展に今後も寄与できることを願っています。