「IBM Watson API」日本語化概説

2015年2月の日本IBMとソフトバンクテレコム株式会社（現ソフトバンク株式会社）の「IBM Watson」（以下、Watson）におけるパートナーシップ提携発表[1]以降、「IBM Watson API」（以下、Watson API）の日本語化が進み、2016年6月現在では9種類のAPIで日本語を扱うことができるようになりました。

Watsonの日本語化における重要な要素は、自然言語処理技術の開発と自然言語理解のための基礎知識（ベース・ナレッジ）の作成です。IBMは、世界各地の複数の研究開発拠点において、自然言語処理の技術を60年近くにわたり培ってきました。本稿では、その成果の一つであるWatson APIの日本語化について紹介します。

1. はじめに

2011年、Watsonは米国のクイズ番組「Jeopardy!」で世に登場しました。人間のチャンピオン2人をコンピューターが打ち負かしたニュースは、全世界を駆け巡り衝撃を与えました。それ以来Watsonが日本語で利用できる日を待ち望んでいた人は少なくないでしょう。だからこそ2015年2月に、ソフトバンクテレコム株式会社（現ソフトバンク株式会社、以下ソフトバンク）とWatsonの日本語化についてパートナーシップを発表したときは、大きな期待をもって迎えられました。

パートナーシップの発表から1年半。Watson APIの日本語化は順調に進み、2016年6月の時点でリリース済みの13種類のAPIのうち、以下の9種類のAPIで日本語を扱うことができるようになりました（**図1**）。

●Natural Language Classifier
●Retrieve and Rank
●Text to Speech
●Speech to Text
●Dialog
●Document Conversion
●Personality Insights
●Tradeoff Analytics
●Visual Recognition

多くのソフトウェア製品が英語をベースに開発されている現状において、一般的に製品の日本語化とは、英語しか扱えなかった製品で日本語も扱えるようにしたり、ユーザー・インターフェース（UI）を日本語にしたりすることを意味します。具体的には日本語の文字コードを適切に処理できるようプログラムを拡張し、画面に使用される文字列などのUIリソースを翻訳するといった作業です。

図1. Watson Cognitive APIと日本語化の対応状況（2016年6月現在）

一方、Watson APIの特徴の一つは自然言語によるコミュニケーションです。ユーザーは、人間に話すような言葉でWatsonとのやり取りを行います。このような製品の日本語化では、従来の製品のような日本語化の作業に加えて、自然言語処理（NLP：Natural Language Processing）の日本語への対応が必要となります。日本語を適切に処理できる自然言語処理の機能を実装することが、Watson APIの日本語化にとっての要となる作業なのです。

Watson APIの中にはVisual RecognitionやTradeoff Analyticsのように自然言語処理をしないものもあり、これらは先に述べたように従来通りのプロセスで日本語化されています。それ以外のAPIにおいては、自然言語処理機能を含めた日本語化が行われました。以降の章では、それらの日本語化の作業について詳しく紹介します。

2. 日本語化の基礎技術

Watson APIにおける自然言語処理は、ユーザーとのインターフェースのみに使用されているわけではありません。ユーザーが入力した文章を解析したり、ユーザーへの返答を作成したりするのに使用するのはもちろん、内部辞書や学習モデルの作成にも使用されています。自然言語処理はWatsonのコア技術の一つであり、多言語対応において最も重要な要素なのです。

IBMには60年以上続く自然言語処理に関する研究開発の歴史があり、目的に応じてさまざまな特性を持つ言語処理システムが開発されています。Watson APIではその成果を活用し、APIの特性に応じて複数の自然言語処理システムを使い分けています。

その中で最も多く使われているのが、「形態素解析」と「語義のあいまい性解消」を得意とする言語処理システムです。形態素解析と語義のあいまい性解消は言語固有の特性と密接に関連しており、言語ごとの考慮が必要になります。これらは日本語の処理において、特に重要な課題となります。

形態素解析とは、文章を単語単位に分割し、品詞を推定する処理です。**図2**は「日蝕が13時に横浜で見られた」という文章を形態素解析した例を示しています。スペースで区切られているヨーロッパ言語と異なり、日本語は単語に分割するためにも複雑な文法的処理が必要となります。このように単語に分割し注釈を付けることで、本来は非構造である自然言語の文章を構造化データのように扱うことができます。また、品詞の推定に加えて、固有表現の抽出や単語の原型への正規化も、その後の処理のために重要です。

語義のあいまい性解消とは、形態素解析で抽出された単語もしくは文節が、その文章中でどのような意味で使われているのかを推定するものです。例えば「CD」という単語は、「キャッシュ・ディスペンサー」を意味することがあれば、「コンパクト・ディスク」を意味することもあります。この違いを理解することが語義のあいまい性解消です。ある単語がどの単語と一緒に使われることが多いかという共起関係を機械学習することで推定しています。

Watson APIの日本語化においてもう一つ重要な作業は、日本語版「ベース・ナレッジ」の作成です。ベース・ナレッジとは言語を理解するための基礎知識で、それぞれのWatson APIでの機械学習の基となるデータです。新しい言語に対応するときには、対応するAPIと言語ご

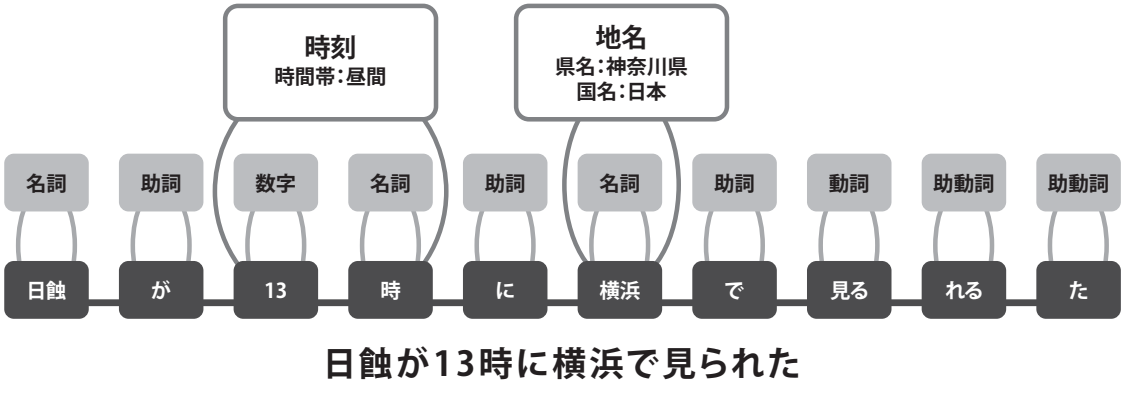

図2. 形態素解析の例

とにベース・ナレッジを用意しなくてはなりません。日本語化においてもWeb上や社有データから多くの日本語リソースを収集し、先に述べた自然言語処理を使用してその内容を分析し、それぞれのWatson APIのベース・ナレッジを作成しています。

3. 各APIの日本語化

日本語化において実施された作業は、APIによって異なります。いくつかのWatson APIで行われた日本語化の作業を紹介しましょう。

●Natural Language Classifier(NLC)

NLCは入力された文章の意図を判別するAPIです。日本語化においては、自然言語処理システムに組み込む日本語のベース・ナレッジ作成が重要なポイントでした。NLCにおけるベース・ナレッジの一つは、単語の特徴ベクトルです。大量の文章データのコンテキストを調べることにより、単語をベクトルとして表現して、語と語の関連性を捉えやすくします。日本語化においては大量の日本語データを収集、自然言語処理システムを使って解析し、NLCのためのベース・ナレッジを作成しました。また、日本語化の完成度を確認するため、ソフトバンクより実際に使用する日本語データをご提供いただき、テストを実施しました。

●Speech to Text(STT)

STTは入力された音声を文字列に変換する、いわゆる音声認識APIです。日本語の巨大音声コーパスを作成し、そこから音響モデルを生成・学習させることで、高い認識率の音声認識モデルを作成しています。さらに、同音異義語が多い日本語で正しい漢字かな混じり文を生成するには、文章中の単語の語義を正しく判別する必要があるため、自然言語処理システムに期待される処理が大きくなります。そのため、さまざまなシチュエーションや発話スタイルをカバーする言語モデルを作成し、学習させています。

●Text to Speech(TTS)

TTSは文字列を音声に変換する、いわゆる音声合成APIです。入力された漢字かな混じり文から、文脈を考慮した読みとアクセントを予測することと、音節ごとの高低アクセントを予測することで、滑らかで自然な発話の音声を生成しています。また日本語はモーラ(音節)の等時性という特徴があります。これは、音節一つひとつの発声がほぼ等間隔であるということです。モーラの等時性を保持することで、日本語らしい発音を実現します。さらに、日本語に特化した独自のパラメーターも導入したチューニングを行っています。

●Dialog

Dialogは自然言語での対話を実現するAPIです。その

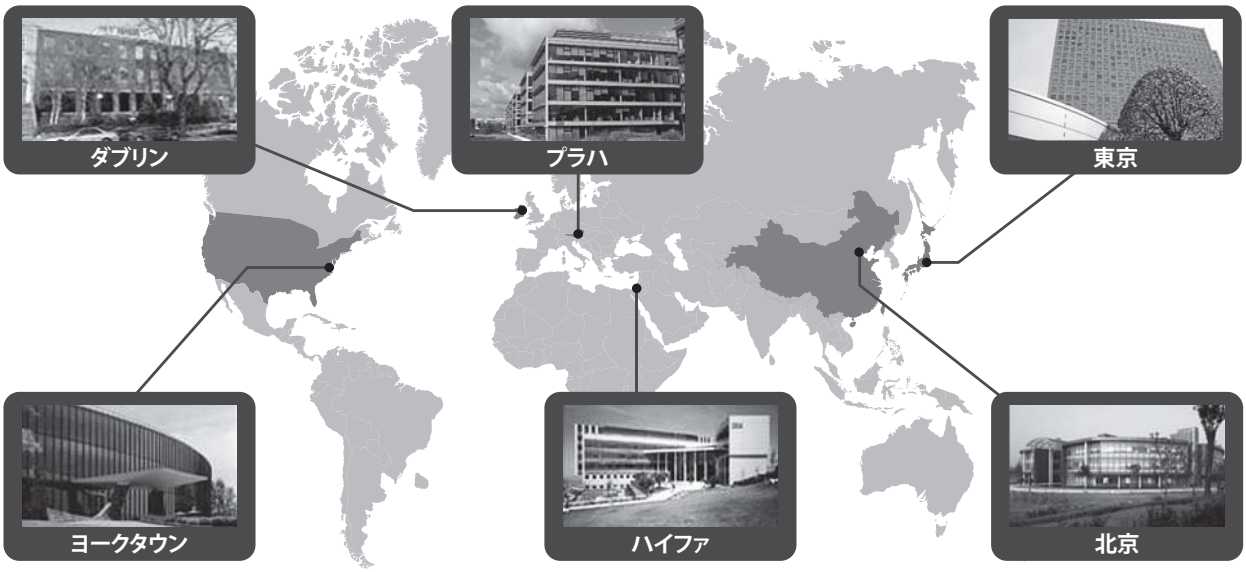

図3. Watson日本語化に関わる主な研究開発拠点(2016年1月当時)

日本語化においては、動詞の活用形などの日本語の文法や、数字の書き方や全角・半角の違いなどの記法を考慮して、表記揺れを吸収できるようにしました。また、人名・地名・通貨記号や、日付と時刻・電話番号のパターン処理など、日本文化に依存した知識を加えることで、対話のコンテキストを正しく理解できるようにするとともに、会話に含まれるデータを抽出できるようにしています。

●Personality Insights

Personality Insightsは、文章を解析してそれを書いた人の性格（パーソナリティー）を推定するAPIです。日本語化においてはサンプルとなる被験者に「性格診断アンケート」を実施してその人の性格を分析するとともに、その被験者のTwitterでのツイート履歴を収集するところから始まりました。そして収集したツイートを自然言語処理で解析して文章の特徴を抽出し、性格と文章の相関関係を学習させることで性格を推定する解析エンジンを最適化しています。また、解析エンジン内部の単語のカテゴリー分類は、生活習慣や文化に依存しているため、日本語向けにチューニングしました。詳細は、PROVISION No. 89をご参照ください[2]。

4. 日本市場に向けた技術者の養成

Watson APIの日本語化においては、製品の日本語化だけではなく、そのAPIを利用したソリューションを作成してAPIを日本市場に広める日本人技術者が必要です。ソフトバンクとのパートナーシップにおいて、Watsonのビジネスを一刻も早く立ち上げるためには、ソフトバンク社内の技術者をトレーニングによって十分なスキルを持った専門家へと育成することが急務でした。しかし、Watson APIはまったく新しい技術であり、英語でも研修資料が揃っていません。そんな中で、海外の開発者と連携しながらAPIの日本語版研修資料を作成し、ソフトバンクの技術者に対して研修を実施しました。全世界に先駆けて日本語にて製品の研修が実施されるという、IBMでは珍しいケースとなりました。

5. 開発体制

IBMでは複数の拠点にまたがったグローバル・チームで開発チームを構成し、製品を開発します。Watson APIの日本語化も例外ではありません。Watsonの日本語化に関わった主な拠点を、**図3**に示します。東京をはじめ、米国のヨークタウン、中国の北京、アイルランドのダブリン、チェコのプラハ、イスラエルのハイファなど多くの拠点が関与しており、各拠点ごとに得意とする技術分野があります。ヨークタウンの研究所は「Watson 研究所」として知られ、「Jeopardy!」で使用されたWatsonをはじめとするWatson APIの基礎技術を持っています。プラハはDialogの開発の中心地です。そして東京基礎研究所は、文章の検索と分析の基本的な技術、そして優れた音声技術を持っています。その力を結集したのが今回の日本語化なのです。

6. 今後の展望

Watson APIは今後もその種類を増やし、続々と日本語化されていく予定です。その中で自然言語処理の重要性はさらに増し、より高い精度が求められていきます。自然言語処理の発展がWatson APIの多言語対応の発展につながっているのです。

また、自然言語は文化と密接に関係しています。Personality Insightsで実装されたとおり、文化の違いを学習モデルに組み込むことで、自然言語処理の精度は向上します。文化の違いを組み込むことは、自然言語でやり取りをするAPIばかりではなく、画像解析や感情抽出といったAPIにおいても重要です。文化を理解することでWatsonの新たな可能性が開かれていくのです。

[参考文献]
[1] 日本IBM：日本IBMとソフトバンクテレコム IBM Watsonを日本で共同展開, http://www-03.ibm.com/press/jp/ja/pressrelease/48336.wss?vm=r&s=1
[2] 北村 英哉, 那須川 哲哉, 上條 浩一：文章を解析し、書いた人の性格を推定「IBM Watson Personality Insights」の可能性, PROVISION No.89 2016年5月, http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=CO113416JPJA

日本アイ・ビー・エム株式会社
ソフトウェア＆システム開発研究所 ワトソン技術開発
シニア・ソフトウェア・エンジニア

難波 かおり
Kaori Namba

1997年日本IBM入社。以来ソフトウェア・エンジニアとして、リレーショナル・データベース関連製品を中心にソフトウェア開発に従事。2015年よりソフトバンクとのWatson日本語化パートナーシップ・プロジェクトに、日本側の開発リーダーとして参画。