特集論文 「Linked Open Data とセマンティック Web」

LODを対象としたユーザ参加型質問応答サービスの開発

～ユーザフィードバックによる精度向上とコンテキスト自動登録手法の評価～

User Participatory Approach to Question Answering Service for LOD

– Accuracy Improvement by User Feedback and Automatic Context Registration –

川村 隆浩 (株) 東芝 研究開発センター
Takahiro Kawamura Corporate Research & Development Center, Toshiba Corp.

大須賀 昭彦 電気通信大学 大学院情報システム学研究科
Akihiko Ohsuga Graduate School of Information Systems, University of Electro-Communications, Japan

keywords: linked open data, question answering system, fieldwork

Summary

Open data is drawing attention for the creation of innovative services in recent years. For promoting a greater number of consumer services on the web, a search function that can reveal what kinds of data are available would be helpful. However, if we assume that the open data is described in a triple language like Resource Description Framework (RDF) in future, full-text search is not suitable for data fragments, and a formal query language is difficult for ordinary users. Therefore, we propose a question answering service based on Linked Open Data. As the problems of using the open data as a knowledge source, we then focus on mapping of question sentence to data schema, and data acquisition. And we propose improvement of accuracy based on user feedback and acquisition of new data by user context information. We also present 'Flower Voice' which is an application of the service for assisting with fieldwork and confirm the effectiveness.

1. 研 究 の 背 景

近年，新規 IT ビジネス創出の期待からオープンデータが注目され，政府系やバイオ系，またはスマートシティ向けにデータ公開が進められている．しかし，今後，コンシューマ向けサービスへの活用を促すには，Web 検索のようにオープンデータを簡単に検索できるサービスが必要だと思われる．特に，近い将来，オープンデータが RDF/XML のようなトリプルで表現されるようになることを仮定すると*1，全文検索では断片化したデータ項目は検索し難く，SPARQL (SPARQL Protocol and RDF Query Language) の活用は一般ユーザには容易ではないという問題がある．しかし，データが $< S,V,O >$ からなるトリプルであることから，ユーザの質問文をトリプルに起こしてノード・リンク間をマッチングできれば，簡単な検索の仕組みを実現できると考えた．そこで，以下ではオープンデータを対象とした質問応答サービスを提案する．同時に，ユーザの発話文を簡単にトリプルとして登録できる機能も提供する．本論文の主な貢献は，上記サービス実現にあたってスキーマのオープン性とデータ拡充という 2 つの問題に着目し，それぞれに対しユーザフィードバックによる精度向上とコンテスト自動登録手法を提案し，実サービスを試作して有効性を評価した点にある．

尚，質問応答システムの分野では，昨今，米 Apple の Siri や NTT ドコモのしゃべってコンシェルが注目されており，ユーザーがよく利用する機能としてアプリ呼出しと情報検索が挙げられている [Wired 12]．しかし，情報検索における，特にスマートフォンを用いたモバイルでの検索の問題点として，検索の結果返ってくるのが答えが書いてあるかもしれないページの URL リストであることが多いことが挙げられる．スマートフォンの画面でページ内から所望の情報を探し出すのは大変であり，場合によってはページ内でもう一度検索しなければならない．そこで，RDF からなるオープンデータ，Linked Open Data (LOD) を質問応答システムの知識源として活用すれば，大規模なグラフの中から必要なデータ項目をピンポ

*1 2012 年 12 月現在，総務省によるオープンデータ流通推進コンソーシアム技術委員会では，RDF を標準フォーマット案として検討が進められている．

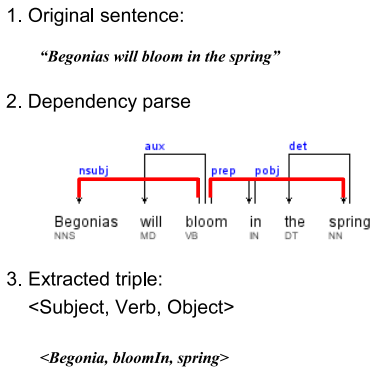

図 1 自然文のトリプルへの変換例

イントで返すことができる可能性があり，有用であるだろう[*2]．したがって，両者の観点から LOD と質問応答システムの組み合わせには可能性があると考えている．

以下，2 章ではオープンデータを用いた質問応答サービス実現における問題点とアプローチについて述べる．また，3 章では事例としてスマートフォンを用いた園芸・農作業向け情報検索・作業記録サービス「花之声」の開発について述べる．そして，4 章でいくつかの観点から関連研究と比較し，最後に 5 章でまとめと今後の課題について述べる．

2. LOD 活用の問題点とアプローチ

対話システムの系譜 [河原 13] に照らすと，今回提案のサービスは Siri やしゃべってコンシェルと同じ「一問一答エージェント」における DB 検索型質問応答 (QA) システムに分類される．しかし，これらが厳密にはクローズドな DB 検索型 QA システムと，オープンな Web 検索型 QA システムの組み合わせとなっているのに対し，提案のサービスはオープンな DB 検索型 QA システムと位置付けられる．具体的な構成は次章で述べるが，基本的な流れとしては，まず質問文から形態素解析，構文解析等により主語，述語，目的語といったトリプルを抽出する．構文解析木のトリプルへの変換例を図 1 に示す．そして，疑問詞や省略語を変数に置き換えて，LOD DB を検索する．概念的には，LOD DB 内の $< S, V, O >$ に $<?, V, O >, < S, ?, O >, < S, V, ? >$ といった質問文を unify させるものである．SPARQL はグラフパターンマッチングを基本としており，上記は基本グラフパターン，1 組のトリプルがマッチする場合に相当する．尚，登録時には S に該当するリソースと，V に該当するプロパティが存在した場合に，O を値として持つトリプルを DB に追加登録する．

対話制御を行わない DB 検索型 QA システムは数多く提案されているが，多くの DB 検索型 QA システムと異なり，データのスキーマがオープンであるために少なくとも以下の 2 つの問題が存在する．尚，スキーマがオープンとは，特定の個人・団体がスキーマを管理していないため，同じ意味のプロパティでも複数種類存在したり，勝手にプロパティが追加されたりすることを意味している．検索対象 LOD を DBpedia など 1 つに絞れば，このような問題はあまり見られないが，ここでは作成者の異なる複数の LOD セットを跨って検索することを前提としている．

2·1 質問文と LOD スキーマとのマッピング

検索にあたり多くの場合，質問文における述語と LOD スキーマにおけるプロパティとのマッピングが事前に必要だが，オープンデータを活用する場合，いずれも未知であり[*3]，マッピングの評価値，つまり回答候補のスコアを事前に定義することができない．

そこで，オントロジーアライメント技術における String Similarity と Semantic Similarity を用いて，述語とプロパティのマッピングを行うと同時に，ユーザフィードバックに基づくマッピングの改善を試みる．まず，シードとして一定量の { 述語, プロパティ } のマッピングを Key-Value Store (KVS) に事前に登録しておき，質問文の述語が未知であった場合，

(1) 述語を日本語 WordNet オントロジーを用いて同義語群に展開した上で，登録済みの述語との類似度を Longest Common Substring (LCS) に基いて計算する．
(2) また，述語を英訳して登録済みプロパティとの類似度を計算する．
(3) 更に，主語に該当すると思われるリソースが LOD から検索された場合は，そのリソースの全プロパティに対して英訳した述語，または日本語プロパティの場合は元の述語との類似度を計算し，上記 (1)(2)(3) の結果から未知の述語に対応すると思われるプロパティの暫定的なランキングを構成する（図 2 参照）．
(4) その上で，ユーザがどのプロパティの値を参照したかをサーバーにフィードバックすることで，未知の述語とプロパティとのマッピングを動的に獲得，KVS に登録する．
(5) また，登録済みのマッピングに関しても一意とは限らないため，ユーザからのフィードバック数に基いて信頼度を付与する．信頼度は整数値で 1 フィードバック毎に 1 追加され，一定以上の LCS 値を持つプロパティのリストは，信頼度の高い順，同一の信頼度であれば LCS 値の高い順にリランキングされる．これにより，N-best 精度の向上を図っている（4·1 節参照）．

*2 質問応答システム IBM Watson でも一部で LOD が用いられている．

*3 クローズドな DB であればスキーマは自ら規定するため既知である

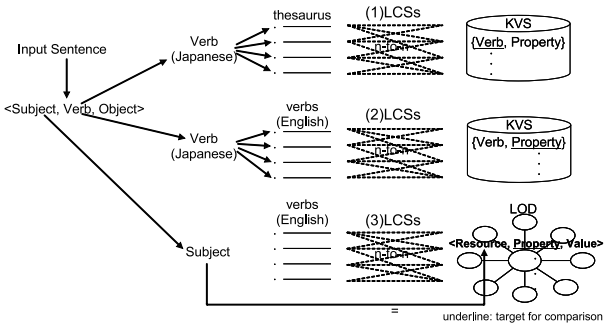

図 2 LCS の計算方法

2·2 LOD の追加登録・拡充

DB をオープンにしても，一般ユーザが SPARQL を用いてトリプルを DB へ登録するのは容易ではない．そこで，まず前述した自然文からのトリプル抽出処理により，自然文で追加登録できる仕組みを提供している．

また，ユーザによるデータ登録のサポートとしてコンテキストの自動付加機能を提供している．これはユーザがデータを登録した際に，センサーデータから必要な情報を収集し，セマンティック情報に変換してコンテキストとして自動的に付加するものである．現在でも，Twitter でのつぶやき時に緯度経度情報を付加する機能が存在するが，それをより多様なコンテキストに対応させたものである．これによって，ユーザが直接述べた情報だけでなく，背景となるさまざまな情報を自動的に登録することができる．センサーデータと付加コンテキストの例を次章に示す．これは，いわば Human Computation[*4] における人間計算資源の調達方法の 1 つ，副作用としてデータを収集するアプローチと言えるだろう．

一方で，人間計算資源の別の調達方法である，オープンデータ構築に関する価値観を共有するユーザに対し，貢献感を高める仕組みとして，データ登録者の Twitter ID を Creator として登録データに付加している[*5]．現在，こうした取り組みにより，ユーザ参加型のデータ登録の促進を計っている．

更に，一般的な情報や，反対に専門的な情報に関しては Web ページからの LOD 自動抽出・登録を行なっている [ミン 11]．これは Web ページ，またはブログを対象に CRF (Conditional Random Fields) を用いてトリプルを抽出するものであり，抽出対象によって異なるが 90%前後の精度で抽出可能であることを確認している．

3. フィールドワーク向け応用事例の開発

本章では，具体的な実装と応用事例について説明する．現状，有限状態トランスデューサなどによる対話制御は行なっていないため，問題解決型のタスク，例えば製品サポートなどは難しい．そこで，1 章でも挙げた情報検索に絞ると，以下のような応用が考えられる．

(1) 一般情報検索
　現在，DBpedia には 10 億トリプル超の情報が登録されており，Wikipedia で検索したくなるような情報の一部はピンポイントで検索することができる．

(2) フィールドでの情報検索・記録
　登録も可能であることから，フィールドワークの支援を目的として特定ドメインに関する情報検索や記録を行うことができる．例えば，農・園芸作業，エレベータ保守，プラント点検，山歩き，震災時避難，旅行などが挙げられる．

(3) Twitter と連動した情報の保存，検索
　情報共有にフォーカスした場合，例えば特定のハッシュタグでつぶやくと自動的にトリプルが抽出され LOD に登録されたり，ハッシュタグで質問すると LOD 化した過去のつぶやきを検索する，などの応用も考えられる．これはクチコミやライフログの記録，共有などに役立てることができるだろう．

はじめに述べた問題意識は上記 (1) であるが，以下ではドメインを区切って評価するために，(2) の観点から園芸・農作業時の各種疑問，病気，施肥，手入れ等に答える音声アシスタント「花之声」について紹介する．

3·1 花 之 声 と は

近年，環境意識やマクロビオティックへの関心の高まりから，都市緑化や都市農業が注目を集めている．しかし，都市の限られた空間で植物を育てるのは必ずしも容易ではない．園芸知識のない初心者は，植付けから収穫までさまざまな場面で悩みや疑問にぶつかるだろう．むろん，プロの園芸家や植木業者を呼んでも良いが，コストが掛かる上，都市部では探しにくい．また，こうした作業は環境に依存するため，前もって全てを計画することができず，現場で植物の状態を見て対応する必要がある．しかし，スマートフォンでインターネットを検索するにもキーワード入力が煩わしい上，検索結果のリストをタップ，スクロールして答えを見つけるのも不便である．そこで，園芸・農作業時の情報検索に向けてスマートフォンで動作する質問応答サービス「花之声」を開発した．また，作業時は周りに人がおらず，手が汚れていることも多いので音声操作も可能とした．更に，農水省によればデータの記録こそが精密な農業の基本とされているため，ユーザ作業を登録する機能も提供した．いわばスマートフォンを用いた園芸・農作業時の音声操作も可能な情報検索・作業記録ツールである．図 3 に「花之声」の概要を示す．ここでは入力文に基いて Question Type を自動的に以下の 4 つに分類している． 尚，Answer Types は literal，URI，画像の 3 つである．図 3 では分かりにくいが，スマートフォンのカバーは本体の端を抑える構

*4 人間と機械の協調問題解決，計算の一部を人間が行うアプローチ
*5 名前や e-mail，HP アドレスでもよいが，匿名性を担保したコミュニケーション手段としてここでは Twitter ID を用いた．

図 3 花之声の概要

造となっており，表面にビニールシート等はついていない．また，マイクのある下部は露出する形となっており，キー入力，音声入力の性能に影響はない．

(1) 定性検索
予め LOD DB に登録された植物に関する情報の検索．

(2) 定性登録
LOD DB に登録されていない植物情報の新規登録，または登録済み植物に関する情報の追加．

(3) 記録登録
作業の記録および共有．農業では作業の記録が重要とされているため，作業登録時のセンサー情報の記録は有用性が高いと思われる．但し，登録情報は事前に定義された LOD スキーマに限定される．次節にて詳述する．

(4) 記録検索
以前の作業の思い出しや，他者の作業を参考とするための作業記録の検索．

具体的なユースケースとしては，以下のような形を考えている．

§1 芋づる式検索

ユーザが園芸・農作業中にその場で知りたくなった内容に，「花之声」がピンポイントで回答していくケースである（図 4）．

§2 ユーザ参加型活用

データを追加登録できる仕組みを利用し，ユーザが互いに知り得た情報を共有するケースである．主に，生態調査的活用（図 5 上段）と知恵袋的活用（図 5 下段）があると考えている．ユーザが特定の環境項目，例えば希少種の発見等についての調査・報告に協力したり，知識コミュニティを形成することを意図している．尚，登録情報には登録者の Twitter ID が付記される．

図 4 芋づる式検索

3·2 Plant LOD の概要

「花之声」で対象としている LOD は，DBpedia における Plant Class 下の 10000 超の植物種に国内種 104 を追加した Plant LOD である．プロパティとしては既存の

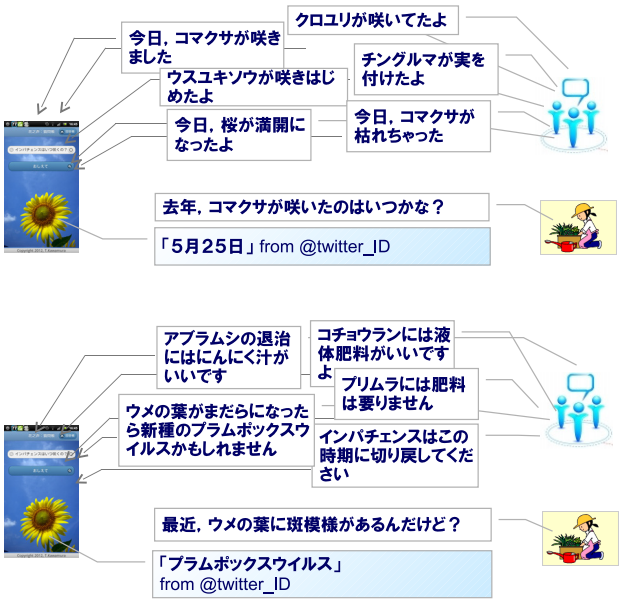

図 5 生態調査的活用（上段）と知恵袋的活用（下段）

300 に加えて，植物栽培の観点から新たに 67 を追加した．記録用のスキーマとしては，開花日や施肥日，収穫日などを記録するためのプロパティを追加した．図 6 に Plant LOD の概要を示す．Plant LOD は先に植栽設計にフォーカスして開発した「花咲かめら」[Kawamura 12] で使用したものを拡張しており，Dydra.com にて格納，公開されている*6．

3·3 サービス構成

「花之声」のサービス構成を図 7 に示す．著者らは，今後，農業・園芸分野以外へ応用する場合も，対象 LOD，および { 述語，プロパティ } といったデータの内容を変更するのみで，構成としては図 7 のままで横展開できると想定している．したがって，図中，花之声に固有の部分は LOD DB として利用している Dydra.com と KVS として利用している Google Big Table のコンテンツであり，他は共通部分，フレームワークとして位置付けられると考えている．一般形としては，(1) 自然言語による質問文または登録文の入力，(2) 形態素解析エンジンを用いた自然言語による自然文からのトリプル抽出，(3)LOD DB の検索・登録，(4){ 述語，プロパティ } ペアを登録した KVS，翻訳エンジン，オントロジー検索を用いた，プロパティと述語とのマッピング計算，(5) プロパティとプロパティ値の出力，(6) ユーザフィードバックによるマッピング信頼度の更新または登録プロパティの決定，という流れとなっている．但し，本論文は 2 章で述べたユーザフィードバックによる精度向上とコンテキスト取得によるデータ獲得手法を実サービスを通して評価することを目的としており，フレームワークの最適な構成について十分な検討を行ったわけではなく，構成自体の議論は今後の課題としたい．

以下，具体的な処理について示す．自然言語による質問文の入力はキー入力，または Google 音声認識による音声入力が可能である．本論文において入力手段は本質ではないが，前述したように利用シーンを考慮して音声入力も可能としている．次に，Yahoo!API による形態素解析結果を内部の構文解析ルールに掛けてトリプルを抽出する．SPARQL クエリーはテンプレートベースの手法を用い，S や V などの空白スロットを埋める形式で生成される．また，回答は XML 形式で受け取っている．その上で，Google Big Table に登録した { 述語, プロパティ } ペアの検索，Microsoft 翻訳，NICT 日本語 WordNet オントロジー検索を用いて 2.1 節で示した方法で各マッピングの LCS を計算する．マッピングの順番としては，まず ‘sameAs’ や ‘wikiPageRedirects’ などのリンクをトレースしながら S がリソースにマッピングされ，次に V とプロパティとのマッピングが探索される．そして，LCS が高いプロパティと述語の組み合わせから，回答候補を作成する．尚，回答候補はクライアント UI の制約から 3 つまで作成される．同時に，クライアント UI 下部には Google での通常の検索結果も併せて示す．これは両者の比較により質問応答システムの利点や限界を示すためである．ユーザフィードバックはクライアントにおけるアコーディオン UI の開閉，つまりプロパティ値の閲覧有無によってサーバに伝えられる．より明示的にフィードバックを得る方法として,「いいね」ボタンを付けるなどの方法も検討したが，今回はユーザに余計なアクションを求めない方法として本方式を採用した．また，複数のアコーディオンを開けてしまった場合でもフィードバックされるのは初めの 1 つのみである．これはユーザの行為としてランダムに開ける，または関係なさそうなところから開けるという選択は自然ではないと考え，逆説的に初めの 1 つを取得するようにした．尚，これらインターフェイスの評価については今後の課題としたい．ユーザフィードバックは検索時は登録済み { 述語，プロパティ } マッピングへの信頼度付与，未登録 { 述語，プロパティ } ペアの登録，登録時には O(値) の登録先プロパティの決定という役割を持っている．また，出力は画面表示であり，音声応答は未実装である．

コンテキストの自動付加機能は，センサーデータの取得と対応スキーマに応じたセマンティック変換からなる．センサーデータは，おサイフケータイの履歴を除いて，端末の JavaScript にて取得している．表 1 にセンサーデータとコンテキストの対応例を示す．尚，時計やおサイフケータイはセンサーではないが，コンテキストとの対応を示すため便宜的に表に入れている．また，POI や天気は GPS 情報を基に Yahoo! Open Local Platform や気象庁のデータを参照することで取得している．

表 1 センサーとコンテキスト情報の対応

センサー	取得可能なコンテキスト情報
時計	日時
GPS	住所，周辺 POI
（上記 2 つの組み合わせ）	温度，湿度，天気
照度	場所 { 屋内，屋外 }
	歩行状況 { 移動中，静止中 }，
加速度	歩行時間・距離
おサイフケータイ履歴	購入金額，交通機関利用履歴

また，コンテキストに対応するスキーマは予め用意しておき，スキーマ毎の DataType に応じて，センサー情報を Literal や Integer に変換する．例えば，ユーザが開花を登録した場合，gtcprop:flowerAddress（住所），gtcprop:flowerDateHighTemp（最高気温），gtcprop:flowerDateLowTemp（最低気温），gtcprop:flowerSpace（場所）などが自動的に登録されるが，住所は Literal，気温は Integer，場所は { 屋内，屋外 } のいずれかに変換される．そして，対象となる植物名に相当するリソースとのトリプルが生成され，LOD DB に登録される．

更に拡張機能として，本サービスにはクライアント UI

*6 dydra.com/takahiro-kawamura/fv

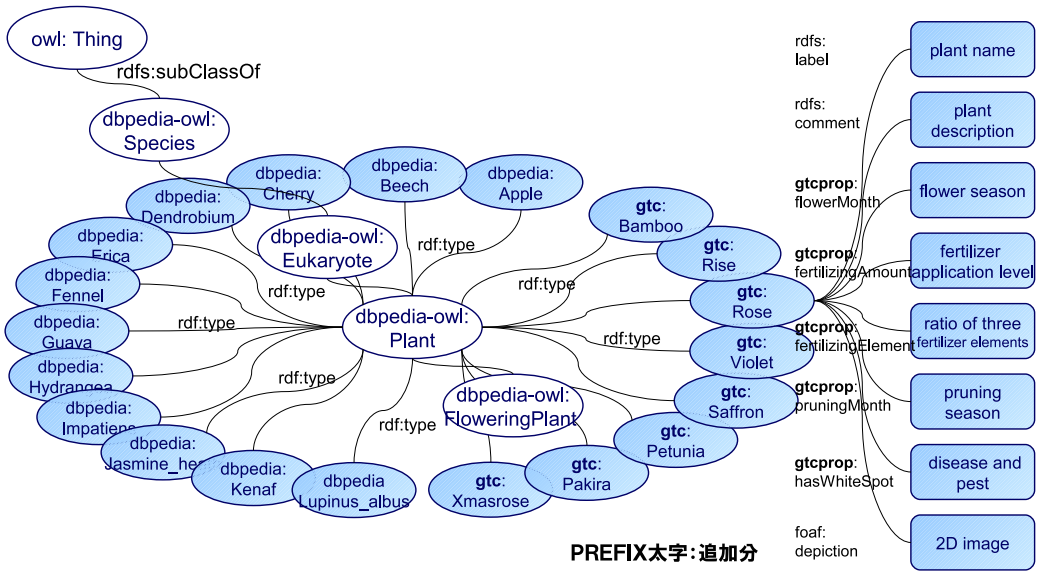

図 6 Plant LOD の概要

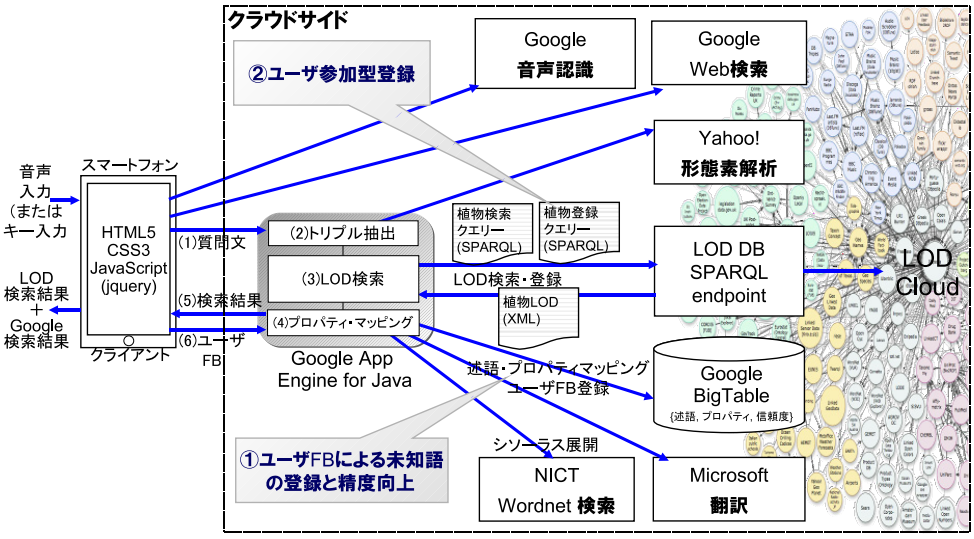

図 7 サービス構成

のインプットフィールドを通してユーザが検索対象とする SPARQL endpoint を変更できる仕様となっている．但し，検索のみであり登録はできない．また，検索式が固定されており，返却型として XML を用いているため，非対応のサーバーも存在する．更に，サーバーによってはレスポンスが遅いため注意が必要である．動作確認済み endpoint としては，日本語 DBpedia[*7]，Data シティさばえ[*8]，ヨコハマ・アート・LOD[*9] などが挙げられる．また，述語とプロパティのマッピングは，2.1 節のアルゴリズムに基いて自動的に候補として表示され，ユーザフィードバックによって登録される仕組みとなっているが，検索したいプロパティが候補内に出てこない場合は，ユーザが明示的に述語とプロパティのマッピングを登録することも可能である．{ 述語, プロパティ } の組みをインプットフィールドを通して KVS に登録すると次回検索時より反映される．いずれも LOD に関する多少の知識を持つユーザを対象としたものであるが，ユーザに機能をオープンにすることでユーザ参加型の新しい使い方を見いだすことを期待している．

尚,「花之声」には公開サイト[*10]よりアクセス可能である[*11]．

4. 評 価 実 験

本章では，2 章で挙げた 2 つの問題点に沿って，ユーザフィードバックによる精度向上とコンテキスト取得によるデータ獲得について応用事例を通して評価する．

∗7 ja.dbpedia.org/sparql
∗8 lod.ac/sabae/sparql
∗9 archive.yafjp.org/test/inspection.php

∗10 www.ohsuga.is.uec.ac.jp/~kawamura/fv.html
∗11 本サービスは LOD チャレンジ Japan2012 にて審査員特別賞を受賞した．

4·1 ユーザフィードバックによる精度向上の評価

検索・登録精度，および 2 章で述べたユーザフィードバックによる精度向上を確かめるため，評価実験を実施した．尚，2 つ以上のトリプルを含む複雑な質問は単文に分割して質問されることを想定している．また，Question Type の判別は疑問詞の有無や助詞の活用から判断しており，抑揚は認識していないため，肯定文か疑問文かは陽に表現する必要がある．実験では，ガーデニング経験者数名に日頃の作業でよく疑問に思うことを選んでもらい，99 の質問文と望ましい回答を収集した．以下に質問文の一部を示す.「昨日，インパチェンスを買ったよ」／「いつランタナに水をあげたの？」／「インパチェンスに肥料をあげたほうがいい？」／「インパチェンスってどんな花？」／「インパチェンスの肥料は何をあげられればいいの？」／「インパチェンスは半日影が好き？」／「ワイルドストロベリーは室内で育ててもいいの？」／「プミラの日当たりはどのくらい？」／「インパチェンスの葉が黄色くなるのは？」／「インパチェンスの原産地はアフリカです」／「インパチェンスに必要なのは Mg です」文自体に重複はないものとする．但し，主語，述語個別での意味レベルの重複はよしとした．次に，それらをランダムに 11 文ずつ 9 つのセットに分け，ランダムに 1 セットを選択し，評価した．各質問文の回答候補に対しては，都度，正しい回答 1 つへユーザフィードバック，つまり { 述語, プロパティ } マッピングの登録，または信頼度の付与を実施した．そして，2 セット目を評価し，2 セット目の後にユーザフィードバックによる影響をリセットし，再度，1 セット目から上記を繰り返した．つまり，1 セット目と 2 セット目の精度の差がユーザフィードバックによる精度の向上を意味している．実験結果を表 2 に示す．尚，Google 音声認識では認識結果文のリストの中からユーザが正しいものを選択するか，やり直すことができるため，質問文は正しく入力されたものとする．

ここで，“リソースなし” とは該当するリソース，つまり植物インスタンスが存在しなかったことを意味し，“プロパティなし” とは該当するプロパティが存在しなかったことを表す．これらの場合，結果は 0 件となるか，別の植物，プロパティに関する情報が表示される．また，文解析誤りとは質問文が長い場合など，正しくトリプルを抽出できなかったことを表す．この場合，現状ではエラーメッセージが返される．また，N-best 精度の計算は以下の式に従う．

$$N-best\ precision = \frac{1}{|D_q|}\sum_{1\leq k\leq N} r_k$$

ここで，$|Dq|$ は質問文に対する正解数を表し，r_k は k 番目の回答が正しければ 1 を返し，それ以外は 0 を返すものとする．つまり，$N=3$ であれば，3 候補以内に正解があれば，成功と見なす．

今回，登録外の植物に関する質問が約 2 割あったが，登録プロパティはほぼ充実していることが分かった．また，現状，構文解析はルールベースであり，これに当たらない質問が約 1 割存在した．短文に限っているという意味では Controlled Natural Language ではあるが，現状，9 割程度の質問文を解釈可能と言える．今後は，ルールの充実と [ミン 11] で示した CRF の活用を検討している．

N-best 精度はデータ，植物種や { 述語, プロパティ } マッピングを登録すればするほど良くなることが期待できるため，1 セット目のベース精度はあまり重要ではない．しかし，1 セット目と 2 セット目を比較することで，ユーザフィードバックによる N-best 精度向上の効果を確認することができた．実験では，信頼度が向上した述語の個数は 1 セット 11 文中，平均して約 2 つであった．これにより，1 セット目から 2 セット目にかけて，2.1 節のアルゴリズムに沿って { 述語, プロパティ } マッピングのランキングが変動し，1-best が 54.4%から 72.7%へ向上している．ここで，1-best=3-best とは正解候補が全て 1 位であったことを表す．尚，今回の実験では意図的に質問内容を分散させることはせず，経験者から得られた質問文をデータセットとしている．その結果，データセット内には主語に当たる植物名や述語の具体的な表現などはそれぞれ異なるが，内容としては頻出作業である水遣りや施肥，日照に関する質問文が多く見られる．そのため，各文から正しくトリプルを抽出できた場合，各データセット間では平均 36.4%のプロパティが共通するものとなっている．これは利用シーンに沿ったデータセットの特徴であるため，本手法の有効性を損なうものではないが，精度向上に有利な要因となっている点は注意が必要である．

更に，本実験では検索対象 LOD を Plant LOD，つまり DBpedia ＋追加 LOD に固定しており，スキーマが完全にオープンな状態とは言えないが，全プロパティ 367 種の内，42 種のプロパティに対して平均約 5 つの述語とのマッピング，計 201 個を事前に登録し，他を未知とすることでオープンな状態を模擬している．事前登録したマッピングの例を表 3 に示す．これらは，開発者 2 名が想定される質問をリストアップし，必要な { 述語，プロパティ } の組みを作成，サービスデプロイ時に一括して登録したものである．{ 述語, プロパティ } マッピングの獲得数は，事前登録数と試行回数に応じてドメインによって異なる一定数への飽和曲線を描くことが予想されるが，本ドメインに関しては，201 のマッピングから 1 回の試行につき平均 0.09 個の未知の述語の獲得が確認された．マッピングの獲得例としては，例えば施肥に関して { 肥料やり, fertilizingMonth} というマッピングは事前に登録されているが，施肥は時期により，元肥（もとごえ）やお礼肥（おれいひ），寒肥（かんごえ）といった言い方もするため,「インパチェンスに元肥はいつやればいいの？」といった質問には直接的にはマッチしない．しかし，2.1 節の類似度計算から fertilizingMonth プロパティが候補として提示され，ユーザがそのプロパティ値

表 2 検索精度

	失敗			成功	
	リソースなし	プロパティなし	文解析誤り	1-best	3-best
1 セット（平均）	18.2%	0%	9.1%	54.5%	72.7%
2 セット（平均）				72.7%	72.7%

を参照することで，{ 元肥やり，fertilizingMonth} というマッピングが新たに獲得された．今後，より詳細に分析し，他分野展開時のブートストラップの参考にしたい．

一方で，今後，プロパティの種類が増えると 1 述語に複数のプロパティがマッチし，競合解消のオーバーヘッドが大きくなることが予想される．現状，上記 42 種のプロパティを Google App Engine 1.8.4，1CPU，55.1MBytes メモリで計算した場合，LCS 値に基いて 1 述語から対応するプロパティの順位付きリストを出すまでにかかる計算時間は平均 660ms 程度となっている．これは，図 7 における Google Big Table，Microsoft 翻訳，NICT Wordnet 検索へのアクセス時間を含んでいる．尚，1CPU は 1.0–1.2GHz 2007 Opteron に相当するとされている．これに対して，1 プロパティ増える毎に LCS 値の計算回数と，プロパティ間での LCS 値の比較回数は増えていくが，実験ではいずれも 1 回 1(ms) 程度であり，かつ，2.1 節のアルゴリズムよりこれらの計算量は単純増加であり，$O(N)$ で収まると予想される．したがって，プロパティ増加による競合解消のオーバーヘッドは，実際上はクラウドプラットフォーム上の CPU 数を増やすことで対応可能であると考えられる．また，プロパティは基本的に情報の種類に対応しているため，際限なく増えることは想定しにくい．そのため，結果的には登録情報の多様性から来る検索精度向上のメリットのほうが大きくなるだろう．

4·2 コンテキスト取得によるデータ獲得機能の評価

コンテキスト取得の有効性を確認するため，以下の評価実験を実施した．実験では，ガーデニング経験者数名から 44 の登録文例を収集し，登録と同時に各種コンテキストを取得した．尚，先の実験と同様に，音声認識の誤りはないものとする．また，登録先のプロパティは正しく選択されるものとする．実験結果を表 4 に示す．

また，ユーザ登録用のプロパティとセンサーで取得可能なコンテキストの組み合わせ例を図 8 に示す．例えば，gtcprop: flowerDate – Weather は，開花日に天気も同時に記録することを意味している．尚，今回はユーザ行動系のコンテキストである NumOfStep:歩数，WalkDist:歩行距離，WalkTime:歩行時間等は使われていない．また，POI (Point of Interest) はその場所を含む周辺のビル，会社，駅などの施設名称を意味している．

表 4 中，“プロパティなし” は該当するプロパティが存在しなかったことを表し，文解析誤りとは質問文から正しくトリプルを抽出できなかったことを表す．但し，登録に当たっては該当するリソースが存在しなかった場合は自動的に追加登録されるため，“リソースなし” は発生しない．また，付加コンテキスト数とは，登録に成功したデータ 1 件に付き，平均何トリプルのコンテキスト情報が自動的に付加されたかを意味している．実際には，GPS が取れなかったりするため，必ずしも図 8 のコンテキストが全て取れるわけではない．また，有効コンテキスト数とは，自動的に付加されたコンテキストの内，ガーデニング経験者によって有効な情報とされたトリプルの数を表す．以下に，有効なコンテキストの例を示す．

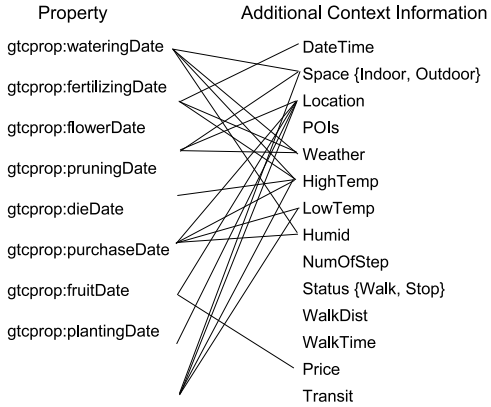

図 8 プロパティとコンテキストの対応

例えば，wateringDate – Address, HighTemp, Space の組み合わせからは，環境と季節毎に異なる水遣り間隔の相関を解析するデータを集めることができる．

また，flowerDate, fruitDate, dieDate – Address, Weather, HighTemp, LowTemp, Space の組み合わせから，地域毎の天候の変化と開花から結実，更に枯れるまでの経過に関するデータを集めることができる．

pruningDate, flowerDate, fruitDate – Address, HighTemp, LowTemp からは，剪定時期と花つき，実付きとの相関を調べることができる．

hasWhiteSpot – Humid からは，乾燥からハダニ発生のリスクを予想することができるだろう．

結果として，ユーザ登録の副作用として登録時にコンテキスト情報を自動的に収集することで，登録されたデータ 1 件につき，有効性が認められるトリプルを平均で 3.4 個登録することができた．こうしたデータが RDF 化され，かつ LOD クラウドを介して共有されれば，さまざまな分野の人がそれぞれの観点から関連を辿って分析することが容易になるだろう．

表 3 述語とプロパティの組み合わせの例

定性情報用プロパティ	
{ 述語，プロパティ}	対応する質問例
{ 水やり, gtcprop:wateringAmount}	（植物名）に水やりしたほうがいい？
{ 肥料やり, gtcprop:fertilizingAmount}	（植物名）に肥料はどのくらいあげればいいの？
{ 咲く, gtcprop:flowerMonth}	いつ（植物名）は咲くの？
{ 剪定, gtcprop:pruningWay}	（植物名）はどう剪定すればいい？
{ 白い斑点ある, gtcprop:hasWhiteSpot}	（植物名）に白い斑点があるのはなぜ？
作業記録用プロパティ	
{ 述語，プロパティ}	対応する質問例
{ 買った, gtcprop:purchaseDate}	昨日,（植物名）を買ったよ
{ 植えた, gtcprop:plantingDate}	２日前,（植物名）を植えました
{ 枯れた, gtcprop:dieDate}	（植物名）が枯れちゃった

表 4 付加コンテキストの有効性

失敗		成功		
プロパティなし	文解析誤り	登録成功	付加コンテキスト数	有効コンテキスト数
0%	9.1%	90.9%	9.3 件（1 登録当たり平均）	3.4 件

5. 関 連 研 究

DB および質問応答システムの研究分野においては，一般ユーザの支援や専門家の理解を助けるために，自然文による問い合わせを SQL や SPARQL といったクエリー言語へ自動変換する試みは数多い．反対に，クエリーや検索結果を自然文へ戻す研究も行われている [Simitsis 09, Ell 12]．また，1 章にて断片化したデータ項目に対する全文検索の困難さを指摘したが，この点に対してキーワード列から論理的なクエリーへ変換する研究も行われている [Haase 09, Shekarpour 11, Tran 09].

ここで検索対象のデータ構造とクエリー言語を Linked Data と SPARQL に限定すると，自然文をクエリー言語に変換して行う質問応答システムに関する研究は，Deep な言語解析を行うものと，Shallow な解析に留めるものに大きく分類できるだろう．

Deep な解析を行うものとしては，まず ORAKEL[Cimiano 04, Cimiano 07] が挙げられる．これは自然文を LTAGs (Lexicalized Tree Adjoining Grammars) に基いて文法木に変換した後，一階述語論理 F-logic，または SPARQL に変換する．高い表現力を持った変換を可能とするが，その分，自然文に厳密性，規則性が求められるアプローチである．一方，[Wendt 12] では主にイベント情報に関するオントロジーの設計と併せて QA システムの検討がなされており，構文解析木の作成時に時制の解釈や N-ary が考慮される点に特徴がある．オントロジーに基いて定義された semantic description と呼ばれる制約内のスロットを満たすように単語を割り当ててゆき，最終的に semantic description を再帰的に SPARQL へ変換する．但し，これにはオントロジー構造を事前に知っていることが重要である．

しかし，一般ユーザ向け音声操作可能な質問応答システムとしては，これらのアプローチは音声認識誤りや質問文の文法的な誤り，更に言語処理の解析誤りなどを考慮すると実用性に難があるだろう．また，対象がオープンデータであることから，2 章でも指摘したようにオントロジーのスキーマ等を既知とすることには問題がある．そこで，対象データからの Portability，Schema Independent を狙って，より Shallow な解析を行うアプローチが行われている．本論で提案のアプローチはこちらに分類される．

FREyA[Damljanovic 11] は，もともとオントロジー検索のための自然言語インターフェイスを LOD 検索に適用したものである．単語とリソース，プロパティとのマッチングに文字列の類似性と WordNet による同義語を用いている点や，ユーザフィードバックによる精度向上の仕組みを使っている点など本アプローチとの類似点が多い．但し，オントロジーベースの制約を用いて論理的な表現への変換を行なっており，対象データが使用しているオントロジーが完全であることを前提としている．但し，semantic description と異なり原文の文法等は考慮しない．一方，DEQA[Lehmann 12] では，TBSL (Template based SPARQL Query Generator) と呼ばれるアプローチ [Unger 12] を取っている．これは事前に SPARQL クエリーのテンプレートを用意しておき，オントロジー上の制約ではなくテンプレートのスロット埋めるように変換していく．また，本アプローチと同様に特定ドメイン，不動産検索への適用を進め，一定の精度を上げている．PowerAqua[Lopez 06, Lopez 09, Lopez 10] も元はオントロジーのための自然言語インターフェイスを LOD 検索に適用したものであるが，Query-Triples と呼ばれる基本グラフパターンへのシンプルな変換や，単語とリソース，プロパティとのマッピングに文字列の類似性と WordNet を用いている点，ユーザフィードバックを使っている点で本アプローチと多くの類似点が見られる．尚，PowerAqua はオープンドメイン化にあたって，クエリーの文脈に応じたヒューリスティクスを導入し速度低下に対応してい

る点に特徴がある．

質問応答システムのサーベイとしては [Lopez 11] が参考になるだろう．本アプローチは多くの点で先行研究を参考しているが，主には検索と登録のシームレスな結合によるユーザ参加型の精度向上やデータ獲得など，ソーシャル活用を目指している点に特徴がある．また，フィールドワーク支援への応用や日本語対応という点でも同様の研究は見られない．

商用化している音声アシスタントとしては，先に挙げた Siri やしゃべってコンシェル [辻野 13]，仏 xBrainSoft の Angie などがよく知られている．いずれも高い音声認識精度を持ち，タスク種類の判定が容易な定型的な端末機能・アプリ呼び出しに強い．情報検索においても，おそらくキーワード毎に事前に設定された Wikipedia 等の検索先が，表形式等で構造化されている場合は正しく一答で答えることができる．しかし，構造化されていない教えて ! goo 等の検索先からの抽出には失敗が見られ，1 章で述べたように答えが書いてあると思われるページの URL をクリックする必要がある．但し，Angie には Facebook と連携する機能や開発キットなども提供されている．本サービスと直接的に精度を比較することは難しいが，本サービスは情報検索にフォーカスした上で，ユーザフィードバックによる精度向上とコンテキスト取得によるデータ獲得に特徴がある．

また，スマートフォンの農業応用としては，富士通 (株) よりスマートフォンで栽培中の作物の写真を撮影し，簡単なボタン操作で作業を記録・登録するシステム [富士通 12] や，日本電気 (株) よりセンサー情報の見える化や，営農日誌の作成支援を目的とした M2M (Machine to Machine) サービス [NEC 12] が提供されている．いずれも問題意識は「花之声」と同様に作業の記録，見える化にあるが，本アプローチは音声操作可能な質問応答システムの形態を取り，データ記録と参照を統合している点や，オープンデータを用いてソーシャルな活用を目指している点などに違いがある

センサーと意味情報の組み合わせに関しては，Semantic Sensor Network に関する研究において，センサーデータが意味情報でアノテートされ，環境モニタリングや意思決定に用いられている．例えば，SemSorGrid4Env[Castro 11] では，洪水による緊急時の避難計画に適用している．しかし，これらが収集されたセマンティック・センサーデータ内を検索し，推論しているのに対し，我々はネット上に LOD の存在を仮定し，センサーデータをそれらと結びつけている．一方，ソーシャルセンサーに関する研究においては，RFID のような物理的な存在，および GPS データが SNS と統合されユーザ間の協働作業やコミュニケーション促進のために用いられている．例えば，Live Social Semantics[Szomszor 10] は，学会参加者間で関心のありそうな内容を示唆するために用いられている．目的は異なるものの，RFID で得られた F2F でのコンタクト情報が SNS 上のソーシャル情報と結び付けられるという構成は我々の研究とよく似ている．今後，システムのスケールアップに当たって参考にしていきたい．

昨今，さまざまな LOD 活用が提案され，多岐にわたる社会実装例が報告されている．一例としては，欧米を中心とした Open Government[*12] に対する取り組みや，都市全体の社会インフラ情報のオープン化を目指した Dublinked[*13]，LOD を用いた各種メディア情報の連携プロジェクト EventMedia Live[*14]，または複数分野の医療情報を Linked Data 化し，より幅広く，深い診断を目指した Mayo Clinic[*15] の取り組みなどが有名である．他の事例に関しては，国際会議 ISWC (International Semantic Web Conference) の併設チャレンジや，国内では LOD チャレンジなどが参考になるだろう．いずれにおいても LOD を用いて特定の傾向を分析したり，可視化したりといった使い方がなされているが，我々は 1 章で述べたように，こうした LOD 活用の基盤として LOD の検索手法を提案するものである．全世界では既に 550 億トリプルの Linked Data が公開されており，今後，一般ユーザが直接，データを検索したり，あるいは開発者がどんな種類のデータがあるのかを知りたいといったニーズも出てくるであろう．LOD の総合的な検索を目指したサイト[*16] も存在するが，やはり SPARQL の知識が必要となる．それに対し，本論文では LOD の自然言語による検索手法を提案している点に特徴がある．

6. まとめと今後の課題

本論では，コンシューマ向けサービスにおけるオープンデータ普及を目指し，オープンデータを知識ベースとするユーザ参加型質問応答サービスを提案した．また，応用事例としてフィールドワーク支援向けのアプリケーションを開発し，評価を行った．ここでは，ユーザフィードバックを用いた精度向上や，ユーザ参加型の未知データ獲得の仕掛けなど Human Computation 的要素を特徴としている．

尚，今回はデータ登録を増やすために自動的にセンサー情報からコンテキスト情報を付加する手法を実現したが，これは検索時に活かすこともできるだろう．ユーザコンテキストに基づく検索結果の部分グラフマッチングによる絞り込みである．声で検索するに当たって，全ての情報を言わなくてもユーザの現在と過去の状況から自動的に絞り込んでくれる機能は有効だろう．今後，今回のアプリケーションの反応を探ると共に，ガーデニング以外への展開を検討して行きたい．

∗12 data.gov
∗13 www.dublinked.ie
∗14 eventmedia.eurecom.fr
∗15 www.mayoclinic.org
∗16 sparql.sindice.com

◇ 参 考 文 献 ◇

[Castro 11] R. Garcia-Castro, K. Kyzirakos, M. Karpathiotakis, JP. Calbimonte, K. Page, J. Sadler, O. Corcho, M. Koubarakis, D.D. Roure, K. Martinez, A. Gomez-Perez: “A Semantically Enabled Service Architecture for Mashups over Streaming and Stored Data”, Proc. of 8th Extended Semantic Web Conference (ESWC), 2011.

[Cimiano 04] P. Cimiano: “ORAKEL: A Natural Language Interface to an F-Logic Knowledge Base”, Proc. of 9th International Conference on Applications of Natural Language to Information Systems (NLDB), 2004.

[Cimiano 07] P. Cimiano, P. Haase, J. Heizmann, M. Mantel: “Orakel: A portable natural language interface to knowledge bases”, Technical Report, University of Karlsruhe, 2007.

[Damljanovic 11] D. Damljanovic, M. Agatonovic, H. Cunningham: “FREyA: an Interactive Way of Querying Linked Data using Natural Language”, Proc. of 1st Workshop on Question Answering over Linked Data (QALD-1), 2011.

[Ell 12] B. Ell, D. Vrandecic, E. Simperl: “SPARTIQULATION: Verbalizing SPARQL Queries”, Proc. of Interacting with Linked Data (ILD), 2012.

[Haase 09] P. Haase, D. Herzig, M. Musen, D.T. Tran: “Semantic Wiki Search”, Proc. of 6th European Semantic Web Conference (ESWC), 2009.

[Kawamura 12] T. Kawamura, A. Ohsuga: “Toward an ecosystem of LOD in the field: LOD content generation and its consuming service”, Proc. of 11th International Semantic Web Conference (ISWC), 2012.

[Lehmann 12] J. Lehmann, T. Furche, G. Grasso, A.N. Ngomo, C. Schallhart, A. Sellers, C. Unger, L. Buhmann, D. Gerber, K. Hoffner, D. Liu, S. Auer: “DEQA: Deep Web Extraction for Question Answering”, Proc. of 11th International Semantic Web Conference (ISWC), 2012.

[Lopez 06] V. Lopez, E. Motta, V. Uren: “PowerAqua: Fishing the Semantic Web”, Proc. of 3rd European Semantic Web Conference (ESWC), 2006.

[Lopez 09] V. Lopez, M. Sabou, V. Uren, E. Motta: “Cross-Ontology Question Answering on the Semantic Web - an initial evaluation”, Proc. of 5th International Conference on Knowledge Capture (KCAP), 2009.

[Lopez 10] V. Lopez, A. Nikolov, M. Sabou, V. Uren, E. Motta, M. d’Aquin: “Scaling Up Question-Answering to Linked Data”, Proc. of 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW), 2010.

[Lopez 11] V. Lopez, V. Uren, M. Sabou, E. Motta: “Is question answering fit for the semantic web?, a survey.”, Semantic Web J., Vol. 2, No. 2, pp.125-155, 2011.

[NEC 12] NEC: “M2M を用いた農業 ICT ソリューション”, http://ascii.jp/elem/000/000/637/637743/

[Simitsis 09] A. Simitsis, Y.E. Ioannidis: “DBMSs Should Talk Back Too”, Proc. of 4th biennial Conference on Innovative Data Systems Research (CIDR), 2009.

[Shekarpour 11] S. Shekarpour, S. Auer, A.N. Ngonga, D. Gerber, S. Hellmann, C. Stadler: “Keyword-driven SPARQL Query Generation Leveraging Background Knowledge”, Proc. of International Conference on Web Intelligence (WI), 2011.

[Szomszor 10] M. Szomszor, C. Cattuto, W. V. Broeck, A. Barrat, H. Alani: “Semantics, sensors, and the social web: The live social semantics experiments”, Proc. of 7th Extended Semantic Web Conference (ESWC), 2010.

[Tran 09] D.T. Tran, H. Wang, P. Haase: “Hermes: Data Web search on a pay-as-you-go integration infrastructure”, J. of Web Semantics Vol. 7, No. 3, pp. 189-203, 2009.

[Unger 12] C. Unger, L. Buhmann, J. Lehmann, A.N. Ngomo, D. Gerber, P. Cimiano: “Template-based question answering over RDF data”, Proc. of World Wide Web Conference (WWW), 2012.

[Wendt 12] M. Wendt, M. Gerlach, H. Duewiger: “Linguistic Modeling of Linked Open Data for Question Answering”, Proc. of Interacting with Linked Data (ILD), 2012.

[Wired 12] WIRED.jp: “『Siri』はどう使われているか：調査結果”, http://news.livedoor.com/lite/article_detail/6414414/, 2012.

[河原 13] 河原 達也: “音声対話システムの進化と淘汰-歴史と最近の技術動向-”, 人工知能学会誌, Vol. 28, No. 1, pp. 45–51, 2013.

[辻野 13] 辻野 孝輔, 栄藤 稔, 礒田 佳徳, 飯塚 真也: “実サービスにおける音声認識と自然言語インターフェース技術”, 人工知能学会誌, Vol. 28, No. 1, pp. 75–81, 2013.

[富士通 12] 富士通: “スマホと農業”, http://app10.jp/blog/?p=7084

[ミン 11] グェン ミン ティ, 川村 隆浩, 中川 博之, 田原 康之, 大須賀 昭彦：“条件付確率場と自己教師あり学習を用いた行動属性の自動抽出と評価”, 人工知能学会論文誌, Vol. 26, No. 1, pp. 166–178, 2011.

〔担当委員：森田武史〕

2013 年 8 月 12 日 受理

著 者 紹 介

川村 隆浩(正会員)

1994 年早稲田大学大学院 理工学研究科 電気工学専攻 修士課程了. 同年 (株) 東芝入社. 現在, 同社 研究開発センター主任研究員. 2001–2002 年米国カーネギ・メロン大学 ロボット工学研究所 客員研究員 兼任. 2003 年より電気通信大学大学院 情報システム学研究科 客員准教授 兼任. 2007 年より大阪大学大学院 工学研究科 非常勤講師 兼任. 工学博士 (早稲田大学). 主としてセマンティック Web, エージェント技術の研究・開発に従事. 人工知能学会理事. 情報処理学会会員.

大須賀 昭彦(正会員)

1981 年上智大学 理工学部 数学科卒. 同年 (株) 東芝入社. 同社 研究開発センター, ソフトウェア技術センター等に所属. 1985-1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向. 2007 年より電気通信大学 大学院情報システム学研究科 教授. 2012 年より国立情報学研究所 客員教授 兼任. 工学博士 (早稲田大学). 主としてソフトウェアのためのフォーマルメソッド, エージェント技術の研究に従事. 1986 年度情報処理学会論文賞受賞. IEEE Computer Society Japan Chapter Chair, 人工知能学会理事, 日本ソフトウェア科学会理事を歴任. 情報処理学会, 電子情報通信学会, 日本ソフトウェア科学会, IEEE Computer Society 各会員.