ドメインオントロジーと日本語Wikipediaオントロジーの統合に基づく質問応答ロボットの開発と評価

淺野 泰史† 森田 武史‡ 山口 高平‡

†慶應義塾大学大学院　理工学研究科開放環境科学専攻
‡慶應義塾大学　理工学部管理工学科

要旨

近年，サービスロボットの普及が進み，その社会貢献が期待されている．業務案内等のロボットサービスとしてユーザが満足する質問応答を行うためには，業務知識だけでなく，その周辺知識の両方が必要である．本論文では，日本語 Wikipedia から半自動的に構築された日本語 Wikipedia オントロジー（JWO）とドメインオントロジーを連携させることによりこれを実現し，業務案内ロボットのための質問応答システムを提案する．

1. はじめに

近年，AI インテグレーションを用いた QA システムの発展が報告されている．IBM Watson[2]は，多くの AI 技術を統合した質問応答システムであり，質問文に対する言語処理，仮説生成，それらの評価を経て非常に高精度な質問応答が可能である．他にも，自然言語による質問応答システムには[4]や[7]のような Web でのドキュメント検索を中心としているものや，[5]や[6]のようにドメインに特化したものがあるが，近年の質問応答システムは，回答が記述されている可能性の高いドキュメントの提示だけでなく，より直接的に回答を提示できることや，ドメインに関連する周辺知識についての質問応答ができることも大きな目的とされている．

2. 関連研究

本節では，既存 QA システムと本論文で提案するシステムの差異について述べる．

IBM Watson [2]は，2 億ページもの文書の中から与えられた質問に対する回答が記載されている可能性が高い文書ページ，もしくはページ内の文を提示する．この過程におけるページや文の選択には，CYC や DBpedia[1]を用いたランキングが行われている．

一方で，我々のシステムはオントロジーを知識源としており，質問文を適切な検索クエリに変換することができれば，データベースに問い合わせることで確実に回答することができる．これがオントロジーを知識源とする質問応答システムの強みであり，ファクトイド型の質問や，定義型質問に対して有効である．

一方で，ドメインオントロジーを用いた質問応答は，その構築コストや備えられる知識の量について問題がある．そこで本研究では[5]や[6]のようにドメインオントロジーのみを用いるのではなく，ドメインオントロジーと JWO[3]を組み合わせる．

これによって，ドメインオントロジーに記述された知識の周辺知識を JWO で補い，“周辺知識に対する質問への応答”を目指す．

3. システム構成

3.1. システム概要

図 1 に本システム構成を示す．本システムは，名詞句抽出モジュール，リソース取得モジュール，オントロジーアライメントを利用した解消モジュール，SPARQL に基づく QA モジュールから構成される．また，JWO とドメインオントロジーを質問応答に回答するための知識源として用いる．さらに，ユーザの質問を認識し，回答を行うインタフェースとして，業務案内ロボットを用いる．本研究では，業務案内ロボットとしてソフトバンクロボティクス社が開発した”Pepper”を用いる．

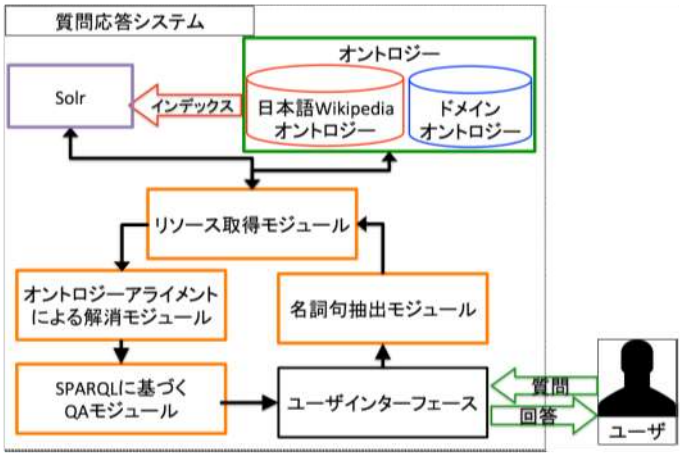

図１　システム構成

本システムでは，各モジュールを Web サービスとして実装しているため，業務案内ロボットが変更されたとしても，モジュールの再利用により，容易に対応可能である．

本システムでは，ユーザが業務案内ロボットに対して，3.6 で説明する質問パターンに従って質問をすると，５つのモジュールが質問文を分析し，JWO とドメインオントロジーを用いて回答文を生成し，業務案内ロボットが音声合成により回答する．ドメインオントロジーを各領域に用意することにより，業務案内ロボットは様々な領域に関する質問に回答することが可能となる．

3.2. オントロジー

本システムでは，JWO とドメインオントロジーの２種類のオントロジーを質問応答に回答するための知識源として利用している．

JWO は，WordNet などの従来型の汎用オントロジーと比較すると，即時更新性や語彙網羅性が優れている．本研究では，日本語における質問応答を想定しているため，日本語 Wikipedia から半自動的に構築した JWO を用いる．本システムでは，JWO におけるリソースの URI，ラベル，ルビ，クラスのインスタンス数，インスタンスが持つプロパティ数，Wikipedia 記事のアクセス回数などを Apache Solr[1]に登録し，リソースの抽出やランキングを行う際に活用している．

ドメインオントロジーは，各領域文書などから，質問応答に役立つクラス階層，プロパティ，インスタンスネットワークを手作業で構築する．

3.3. 名詞句抽出モジュール

本モジュールは，文字列を入力として，形態素解析ライブラリ[2]を用いて，名詞句のリストを出力する．この時，連続する名詞は名詞句として１つにまとめる．また，名詞と名詞の間に別の品詞が一つある場合，これらを一つの名詞句と仮定する．この名詞句がオントロジー中のラベルとして存在するかを Solr により確認し，存在する場合は，固有名詞として扱い，存在しない場合には，個々の名詞に分割して扱う．これにより，「となりのトトロ」のような固有名詞と「机の上」といった語を区別できる．

3.4. リソース取得モジュール

本モジュールは，名詞句抽出モジュールにより取得した各名詞句を入力として，JWO におけるリソース（クラス，プロパティ，インスタンス）のラベルおよびルビと照合を行い，リソースを取得する．ラベルについては，完全照合しない場合には部分照合を行う．また，別名や略称のような関係を表すプロパティの値になっているリソースも照合する．さらに，JWO には，クラスとインスタンスで同一のラベルやルビを持つ場合があるため，そういった場合は①「リソースについての辞書的な定義が値となるデータタイププロパティを持っているかどうか」②「クラスならばインスタンスを持っているか，インスタンスならば「ふりがな」といったような Wikipedia 固有の質問応答に使えないプロパティの他に，プロパティを持っているかどうか」を考慮し，「①と②を満たすもの＞どちらかを満たすもの＞いずれも満た

[1] http://lucene.apache.org/solr/
[2] https://github.com/lucene-gosen/lucene-gosen

さないもの」といったように優先度を決める．同一のラベルやルビを持つクラス，インスタンスをこの優先度で比較した時，どちらも同じならば，クラスを優先して質問応答に用いる．

さらに，このモジュールでは「ドメインオントロジーの統合」を行う．ドメインオントロジーの統合とは，図２に示すように，ドメインオントロジーには存在しないトリプルを JWO により統合することを意味する．ドメインオントロジーの統合は，ラベルが一致するドメインオントロジーと JWO のインスタンスについて，それらのプロパティとプロパティ値を統合し，回答文を生成する際に用いられる．

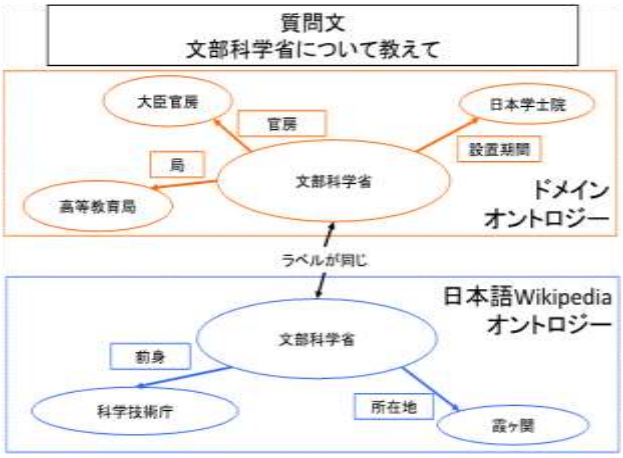

図２　オントロジー統合の様子

この利点は，お互いのプロパティ定義を統合し合うことにより，より多くの質問に回答できることである．JWO は語彙網羅性が高い反面，集合知的に構築されているため，一般的なプロパティ定義となりがちである．一方，ドメインオントロジーはドメインの専門家が構築するため，専門的なプロパティ定義は多くなされるが，一般的なプロパティについては定義がされないことがある．

3.5. オントロジーアライメントを利用した解消モジュール

このモジュールは，リソース取得モジュールで得られたリソースをドメインオントロジーの内容を参照して絞り込むモジュールである．まず，質問応答に用いるリソースとしてもっとも優先されるのは，ドメインオントロジーから取得されたリソースである．これが取得された場合，ユーザの質問はドメインに関するものだと考えられるから，より詳しい情報を元に返答するべきであり，JWO から取得されたリソースは質問応答には使わない．

ドメインオントロジーから取得されたリソースがない場合，質問文中の名詞句から完全照合されたリソースならびにその類義語であるとみなされたリソースを用いるが，これも取得されなかった場合には名詞句から部分照合されたリソースを用いる．

さらに，JWO 由来のリソースを使用する場合は，ドメインオントロジーのクラス階層を用いて，取得されたリソースに絞り込みを行う．

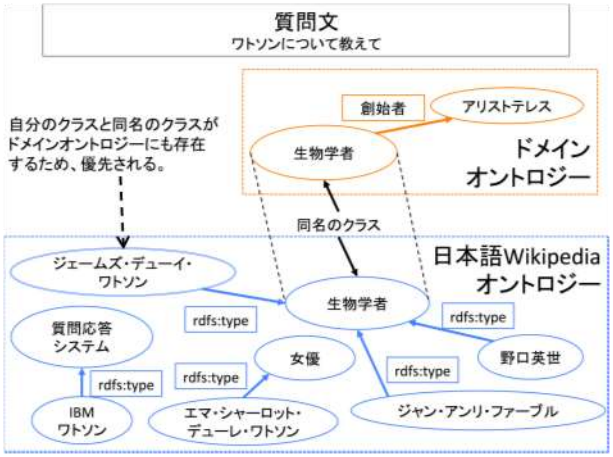

図３　オントロジーアライメントを利用した解消の様子

図３は，ドメインオントロジーを用いて上記の絞り込みを行なっている様子である．「ワトソンについて教えて」という質問文から得られる「ワトソン」という名詞句に対応するリソースは複数考えられる．しかし，ドメインオントロジーに「生物学者」というクラスがあれば，それを利用してリソースを絞り込むことができる．

3.6. SPARQL に基づく QA モジュール

本モジュールでは，質問文が以下の質問文パターンのいずれに該当するかを同定し，その質問文パターンごとに，対応する回答文を生成する．

本システムが想定している質問文パターンは質問文中の名詞句（疑問詞を除く）が２つ以下」であることが前提であり，名詞句が一つの場合と二つの場合で，以下の質問パターンを用意した．

(1)　名詞句が 1 つの場合：名詞句を X とした時,
「X とは何か」を問うもの

(2)　名詞句が 2 つの場合：名詞句を A，B とした時,
「A の B は何か」
「A にとって B はどのような関係か」
「A という関係のもの（属性値）が B であるものは何か」
「A と B は同じものか」
「A は B に属するか」

但し，質問文中の A と B の順序や，質問文の表現は問わない．

各パターンに，SPARQL クエリテンプレートを用意し，リソース取得モジュールにより取得した名詞句に対応するリソースをテンプレートに埋め込み, JWO またはドメインオントロジーを SPARQL クエリにより検索し，検索結果から答えるべき質問パターンを決定する．

質問文中の名詞句が一つの場合には，対応するリソースのタイプ（クラス，プロパティ，インスタンス）により，生成される回答文が異なる．リソースのタイプがクラスの場合には，定義文を述べた後，クラスのインスタンスを列挙し，ユーザに，さらに知りたいインスタンスを選択させる．リソースのタイプがプロパティの場合には，そのプロパティの主語リソースが何であるかをユーザに質問する．リソースのタイプがインスタンスの場合には, 定義文を述べた後, インスタンスが持つプロパティを列挙し，ユーザに，さらに知りたいプロパティを選択させる．

質問文中の名詞句が二つの場合には，回答は一意に定まるため，質問文パターンに対応する回答文を生成する．

4. 実験と評価

4.1. 実験概要

ケーススタディとして文部科学省の案内を支援するシステムを構築して評価を行った．実験内容は，ドメインオントロジーの統合とオントロジーの組み合わせによる正答率についてである．

そのために，文部科学省の組織構成と業務内容が主に記載されたパンフレットより，手動で文部科学省に関するドメインオントロジーを構築した．なお，このパンフレットは公開データであり，業務案内のための情報資源ではないため，本実験は実評価ではなく仮の評価である．

4.2. ドメインオントロジー統合に関する実験

ドメインオントロジーの統合がどの程度なされているかを定量的に計測するために，JWO が統合したプロパティの抽出を行った．まず，文部科学省のパンフレットからドメインオントロジーのインスタンスと対応する名詞句を抜き出す．表１に示すような，これらの名詞句について問う質問にリソース取得モジュールを適用させることで，ドメインオントロジーに補われたプロパティの数を数えた．

表１　質問リストの一部

文部科学省について教えて
生涯学習政策局について教えて
初等中等教育局について教えて
高等教育局について教えて
…

表２　ドメインオントロジー統合の実験結果

実験した 名詞句の数	ドメインオントロジーが 有しているプロパティ数	JWOから補われた プロパティの数
131	84	182

表２にドメインオントロジー統合の実験結果を示す．ドメインオントロジーのリソースに対し，十分なプロパティの統合が行われていることがわかる．この一般的な知識をプロパティ統合により再利用することで，ドメインオントロジーの構築においては専門的な知識部分に特化すればよく，ドメインオントロジーの構築コストが下げられると考えられる．

4.3. オントロジーの組み合わせによる正答率の向上に関する実験

本実験の目的は，「X について教えて」型の質問に対して，ドメインオントロジーと JWO を組み合わせたことにより，正答率が上がるか否かを調べることである．

まず，文部科学省のパンフレットの文章中から「X について教えて」という質問の X にあたる名詞句を 1529 個抽出した．次に，こうして作られた名詞群に，「リソース取得モジュール」と「オントロジーアライメントを利用した解消モジュール」を適用し，「X」という名詞句に対して，回答候補として出力されたリソースの数を数え上げた．さらに，こうして得られた結果のうち，回答候補となるリソースが１つに絞れていて，かつリソースが意味する概念が題意に一致している場合を正解とした．

また今回の実験では比較のために，JWO のみを参照データにした場合と，単にドメインオントロジーを先に検索し，候補が見つからなければ JWO を検索するという場合の結果も用意した．

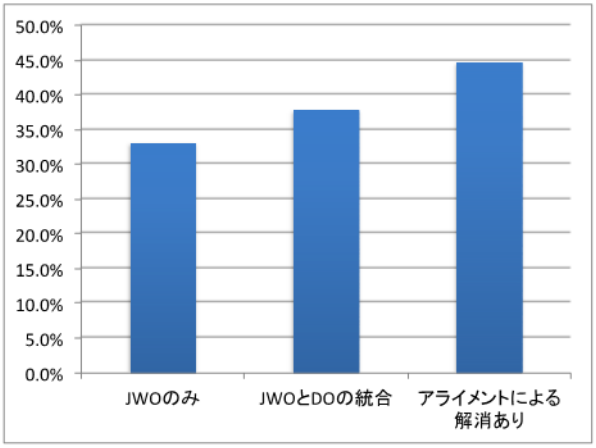

図４　オントロジーの組み合わせによる正答率の変化

図４より，「JWO のみ」と「JWO と DO の統合」を比較することで，パンフレットに書かれている，専門的な名詞に対してはドメインオントロジーの知識を使って答えられるようになっており，単に知識の増加に伴って正解率が上がっている．さらに，ドメインオントロジーに検索結果の候補となるものが存在する場合にはドメインオントロジーの結果のみを返すようになっているため，JWO に対する検索で，複数候補が存在していたものが提示されなくなり，それが回答候補の絞り込みになっている．

次に，「JWO と DO の統合」と「アライメントによる解消あり」の結果を比べると，100 件近くの JWO のリソースに対してオントロジーアライメントを利用した解消による絞込みが行われていたことがわかった．具体例をあげると，「行政機関」という名詞句からは「行政機関」と「行政府」というインスタンスが抽出される．これらのリソースには“似たような意味である”という意味のプロパティがついているからである．しかし，ドメインオントロジーと JWO がともに「行政機関」クラスを持っているため，「行政機関」インスタンスが取得される．

これによって本実験における正解率が上がっているが，正解率としてはまだまだ低いのは確かである．今後は，公開データではなく適切な情報資源を用いることと，ユーザが実際にする質問データを収集し

分析をすることによって，ドメインオントロジーを充実させる．こうすることによって本システムの正答率は向上することが予想される．

5. 結論

本論文では，JWO とドメインオントロジーを連携することにより，対象業務と業務に関連する周辺知識の両方に対する質問応答が可能な，業務案内ロボットのための質問応答システムを提案した．評価実験により，ドメインオントロジーと JWO の連携が，「周辺知識に対する質問への応答」に有効であることを示した．

業務案内ロボットにおける質問応答システムとしては，業務案内ロボットが不適切な回答をした場合の対応，回答に時間がかかる場合の対応，音声認識の精度が低い場合の対応，ユーザへの効果的な回答方法などの問題がある．これらは今後の課題である．

6. 謝辞

本研究は，科学技術振興機構(JST) 戦略的創造研究推進事業(CREST)「実践知能アプリケーション構築フレームワーク PRINTEPS の開発と社会実践」の支援によって実施した．

参考文献

[1] Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives, “DBpedia: A Nucleus for a Web of Open Data”, 6th International Semantic Web Conference, Vol. 4825, 2007, pp. 722-735.

[2] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty, “Building Watson:An Overview of the DeepQA Project”, AI Magazine, Vol. 31, 2010, pp.59-79.

[3] Susumu Tamagawa, Shinya Sakurai, Takuya Tejima, Takeshi Morita, Noriaki Izumit and Takahira Yamaguchi, “Learning a Large Scale of Ontology from Japanese Wikipedia”, 2010 IEEEIWIC/ACM InternationalConference on Web Intelligence and Intelligent Agent Technology, Vol.1, 2010, pp. 279-286.

[4] Maria Vargas-Vera and Enrico Motta, “AQUA - Ontology-based Question Answering System”, MICAI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, Vol. 2972, 2004, pp. 468-477.

[5] Lakshmi Palaniappan and Dr. N. Sambasiva Rao, “An Ontology-basedQuestion Answering Method with the use of Query Template”, International Journal of Computer Applications, Vol. 9, 2010, pp. 23-27.

[6] Jibin Fu, Jinzhong Xu and Keliang Jia, “Domain Ontology Based Automatic Question Answering”, 2009 International Conference on Computer Engineering and Technology, Vol. 2, 2009, pp. 346-349.

[7] Samir Tartir, Bobby McKnight, and I. Budak Arpinar, “SemanticQA:Web-Based ontology-driven question answering”, Symposium on Applied Computing, 2009, pp. 1275-1276.