社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

RDF情報源への自然言語質問手法

秋田　智慧†　　間瀬　心博††　　北村　泰彦†

† 関西学院大学理工学部
†† 関西学院大学理工学研究科
E-mail: ††{mase,ykitamura}@kwansei.ac.jp

あらまし　本研究では，自然言語による質問に対しWebページに付加されたRDFを情報源として返答を行う質問応答手法を提案する．質問文中に含まれるキーワードを基に，RDFを有向グラフとして表現したRDFグラフから，質問に該当するサブグラフを抽出し返答を探索する．研究室情報が記述されたRDF情報源を対象に提案システムの返答精度の評価を行った．

キーワード　自然言語質問，質問応答システム，RDFグラフ

Natural language Q&A method for RDF information resource

Chie AKITA†, Motohiro MASE††, and Yasuhiko KITAMURA†

† School of Science and Technology, Kwansei Gakuin University
†† Graduate School of Science and Technology, Kwansei Gakuin University
E-mail: ††{mase,ykitamura}@kwansei.ac.jp

Abstract　In this paper, we propose a questioning and answering system which responses to a natural language question about RDF information resource attached to Web pages. This system extracts a subgraph which covers the question from the entire RDF graph and searches the subgraph for an answer. We evaluate the precision of the system with an RDF information resource which describes the information about our research laboratory.

Key words　Natural Language Question, Questioning and Answering System, RDF Graph

1. はじめに

現在，Blogサイトを始めとしたインターネットにおける個人による情報発信が一般的になり，RSS(注1)やFOAF(注2)などのRDFによるメタデータを付加したWebページが増加している．RSSはBlogサイトだけでなくニュースサイト等でも配信される記事情報であり，FOAFは人間関係が記述されたメタデータである．これらのメタデータは，コンピュータや，エージェントにも処理することが容易であり，情報源として利用することが可能である．

例えば，RDFデータを情報源を利用することで，Webページの更新に追従して返答を行う対話エージェントが開発可能となる．従来の質問応答技術[1]では，ユーザからの自然言語による入力やシステムからの返答は，あらかじめシナリオやルールとして記述されていた．そのため，Webページの内容が更新された場合には，更新後の情報を反映した返答を行うことができないという問題がある．そのため，正しい返答を行うためには，内容が更新されるたびに開発者がプログラムやシナリオを変更して対応する必要があった．しかし，Webページに付加されたRDFを利用することで，システムは更新された情報を利用することが可能になるため，この問題点は解決できる．そのためには，RDFデータを対象にした自然言語による質問応答技術が必要となる．

そこで本研究では，RDF情報源を対象にした自然言語質問手法を提案し，質問応答システムを実装する．提案手法は，ユーザからの自然言語による質問文からキーワードを抽出し，RDFグラフからキーワードによる絞り込みを行い，返答探索を行う．これにより，Webページの更新に追従して質問応答を行うことが可能となる．

本論文では，第2章でRDFグラフと本研究に関連する自然言語質問手法に関する研究について説明する．次に第3章では自然言語質問に対する返答探索法の説明を行い，第4章では，システムに対する評価と考察を述べる．最後に第5章でまとめと今後の課題を述べる．

(注1)：RDF SiteSummary (RSS) 1.0,
http://web.resource.org/rss/1.0/spec
(注2)：The Friend of Friend (foaf) project,
http://www.foaf-project.org/

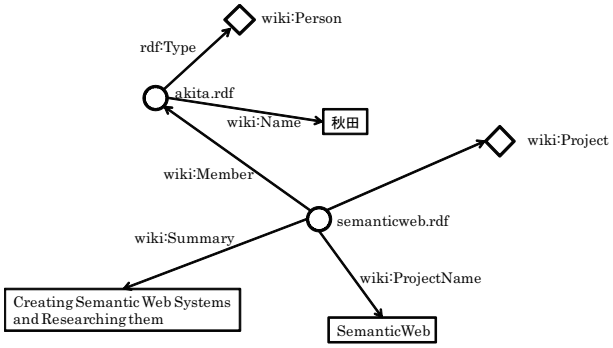

図 1　RDF グラフの例
Fig. 1　Example of RDF graph

2. RDF グラフと RDF 情報源への自然言語質問手法

2.1 RDF グラフ

RDF(Resource Description Framework)(注3)は，情報についての情報（メタデータ）の表現方法についての枠組みである [4]. RDF は主語と述語と目的語の三つ組み（トリプル）を基本単位として，トリプルの集まりとして定義される．

RDF のトリプルである主語，述語，目的語はグラフで表すことができる．主語と目的語をノードとして，述語をこの 2 つのノードを結ぶエッジとして表現する．また，RDF トリプルは複数のトリプルを連結することが可能である．このような RDF 記述よって，メタデータをグラフとして表現できる．このグラフを RDF グラフという．

この RDF グラフは，リソース，クラス，プロパティ，リテラルから構成される．リソースとは，RDF で記述される対象であり，URI で指定される．クラスとは，リソースのタイプもしくはカテゴリを表す．プロパティとは，リソースを記述するために用いる属性である．リテラルとはプロパティの内容を表す記述である．図 1 に Web ページの内容を表した RDF グラフを示す．RDF グラフ内では，リソースは丸形のノード，クラスは菱形のノード，プロパティはエッジ，リテラルは長方形のノードとして表現されている．

図 1 の RDF グラフでは,「秋田」という人物のリソースが存在している．このリソースのクラスは「wiki:Person」であり，これは人物を表すクラスである．また，このリソースの名前を表す属性は,「wiki:Name」というプロパティで表されている．また,「SemanticWeb」プロジェクトのリソースからは,「wiki:Member」というプロパティによって,「秋田」が「SemanticWeb」プロジェクトのメンバーであることが示されている．

2.2 RDF 情報源への自然言語質問手法

RDF を情報源とした質問応答を行うためには，RDF データに対する検索手法が必要となる．現在，RDF で表現されたデータに対する検索手法として，クエリ言語が提案されている．W3C では，SPARQL(SPARQL Protocol and RDF Query Language)(注4)という RDF データに対する汎用クエリ言語の標準化が行われている．このクエリ言語を用いて検索条件をクエリとして記述することで，RDF データから必要な情報を抽出することが可能である．しかし，クエリを作成するためには，RDF データの構造が把握している必要がある．また，データ構造がわかったとしても適切な検索条件を設定しなければならず，ユーザにとってクエリ作成は困難な作業である．それに対して，本研究で提案する手法は自然言語質問文を用いて RDF データから必要な情報を抽出することが可能であり，ユーザは容易に使用できる．

自然言語入力による質問応答システムとしては，Lopez らが開発した AquaLog [6] がある．AquaLog では自然言語による質問文からオントロジーを用いてトリプルに変換して，返答探索を行っている．提案システムは，質問文中の単語を用いて RDF グラフを絞り込み，返答候補と質問文の単語とのグラフ構造における距離を調べることで，返答の探索を行っている．そのため，AquaLog では質問に該当するデータがない場合には返答ができないのに対して，提案システムではグラフ上で距離の近い質問文に関連するデータあれば返答することが可能である．

グラフ構造を利用した質問応答システムとしては，倉田ら [5] が開発したシステムがあげられる．このシステムでは，テキスト情報を対象に係り受け関係に基づくグラフ構造を用いた質問応答を行っている．まず質問の返答候補となる文章をテキストから抽出し，その文章の係り受け関係をグラフ構造として表現する．このグラフから返答候補と検索文字列の距離による順位付けを行い返答候補の絞り込みを行っている．

また，対話ルールを利用した質問応答システムとしては，Goh らの開発したシステムがあげられる．Goh らは，組み込み対話エージェント AINI をインタフェースとして用いたクライシスコミュニケーションについてのポータルサイトを構築している．AINI は，XML ベースの対話ルールである AIML(Artificial Intelligence Markup Language) を用いて質問応答を行う対話エージェント ALICE(注5)を拡張したものであり，政府やニュース等の信頼性の高い情報源から抽出した爆発感染の領域知識についての質問応答を行うことができる．

3. 自然言語質問に対する返答探索

本研究では，ユーザからの自然言語による質問文からキーワード抽出を行い，抽出したキーワードを用いて Web ページの内容を表す RDF グラフから質問サブグラフを抽出することで，ユーザからの自然言語による質問に対する返答探索を行う．質問サブグラフとは，質問文に含まれるキーワードとマッチするクラス，プロパティ，リテラルを最も多く含む最小のグラフを RDF グラフから抽出したものである．

(注3)：Resource Description Framework(RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/rdf-concepts/

(注4)：SPARQL Query Language for RDF, http://www.w3.org/TR/2005/WD-rdf-sparql-query-20050721

(注5)：ALICE, Artificial Linguistic Internet Computer Entity, http://www.alicebot.org

表 1　クラス辞書の例
Table 1　Example of class dictionary

クラス名	単語
wiki:person	人，人物，学生
wiki:project	プロジェクト，研究
wiki:event	イベント

表 2　プロパティ辞書の例
Table 2　Example of property dictionary

プロパティ名	単語
wiki:Name	名前，誰
wiki:Hobby	趣味
wiki:Member	所属，参加
wiki:Summary	内容
wiki:Grade	学年

まず，質問文に対して形態素解析を行うことでキーワード抽出を行う．次に，キーワードとクラスの対応が定義されたクラス辞書と，キーワードとプロパティの対応が定義されたプロパティ辞書を抽出されたキーワードと照合し，質問サブグラフ抽出に必要なクラス，プロパティ，リテラルを得る．これらのクラス，プロパティ，リテラルを用いて質問サブグラフを抽出する．そして，質問サブグラフから回答候補を探索し，キーワードと返答候補の距離による絞り込みを行い，返答を決定する．以下に，各手順の詳細を述べる．

3.1　質問文からのキーワード抽出

本手法は，質問文中に出現する単語を基に RDF グラフから返答を探索する手法であるため，質問文中の単語と RDF グラフ上の情報との対応関係が必要となる．そこで，予め質問文中に出現するであろう単語と，RDF グラフ上のクラス，プロパティとの対応関係を表すクラス辞書，プロパティ辞書を作成しておく．表 1 にクラス辞書の例，表 2 にプロパティ辞書の例を示す．抽出された単語に対応するクラス名，プロパティ名がある場合には，それらをキーワードとする．また，辞書に対応関係のない単語については，リテラルと判断しキーワードとして扱う．

疑問代名詞の「何」が含まれている質問文では，質問の対象としてリソース自体が指定される場合がある．例えば,「秋田さんが所属しているプロジェクトは何ですか？」という質問文では，質問の対象は秋田が所属しているプロジェクトであるが，返答としてはプロジェクトの名前が妥当である．このように，質問の対象がリソースである場合には，そのリソースを適切に表現する代表値で返答することが望ましい．

そこで，本研究ではあらかじめクラスごとに代表値を指定するデフォルトプロパティを設定する．表 3 にクラスごとのデフォルトプロパティを示すデフォルトプロパティ辞書に示す．質問文に疑問代名詞「何」が含まれる場合には，構文解析を行い質問対象を特定する．構文解析の結果より,「何」が含まれる文節と係り受け関係にある文節内に，クラスと対応関係にある単語が含まれている場合には，そのクラスの代表値を示すデフォルトプロパティを質問サブグラフ抽出のためのキーワードとして用いる．また,「何」ついてはキーワードとして扱わないこととする．以上の手順により，質問文から質問サブグラフを抽出するためのキーワードを得る．

表 3　デフォルトプロパティ辞書の例
Table 3　Example of default property dictionary

クラス名	デフォルトプロパティ
wiki:person	wiki:Name
wiki:project	wiki:ProjectName
wiki:event	wiki:EventName

例えば,「秋田さんが所属しているプロジェクトは何ですか？」という質問文の場合では，形態素解析を行い得られる単語は,「秋田」,「所属」,「プロジェクト」,「何」となる．次に，疑問代名詞「何」が含まれるため，質問文の構文解析を行う.「何」が含まれる文節「何ですか？」と係り受けにある文節「プロジェクトは」には，クラス「wiki:project」と対応関係にある「プロジェクト」が含まれているため，デフォルトプロパティ辞書を参照して「wiki:project」のデフォルトプロパティである「wiki:ProjectName」をキーワードとして追加する．他の単語についても，クラス辞書，プロパティ辞書を参照し対応関係にある単語をキーワードして追加する．単語「所属」がキーワード「wiki:Member」となり，対応関係のない「秋田」はリテラルを示すキーワードとする．最終的に質問文から抽出されるキーワードは，リテラル「秋田」，プロパティ「wiki:Member」,「wikiProjectName」，クラス「wiki:project」となる．

3.2　RDF グラフからの質問サブグラフの抽出

質問文から得られたキーワードを用いて，RDF グラフから質問サブグラフを抽出する．質問サブグラフの抽出手順を以下に示す．

（1） 質問文から得られたキーワードに含まれるリテラルをランダムに 1 つ選択し，探索開始キーワードとする．リテラルが含まれていない場合には，クラスから同様に選択する．

（2） 探索開始キーワードに一致する RDF グラフのノードを探索開始ノードとする．探索開始ノードが見つからない場合には探索を終了する．

（3） 探索開始ノードを始点として幅優先探索を行い，キーワードに含まれる全てのリテラル，クラス，プロパティを探索する．全てのキーワードを最低 1 つずつ発見したら，その時点の深さのノードを全て探索し，終了する．

（4） 探索されたサブグラフから不要なノードとエッジを削除し，質問サブグラフを抽出する．最後に探索されたノードを始点として全てのノードとエッジを確認する．キーワードに含まれるリテラル，クラス，プロパティと，ノードとそれに接続するエッジのいずれも一致しない場合には，そのノードとエッジを削除する．ただし，子孫ノードやそれらのノードに接続するエッジがキーワードと一致する場合には削除しない．

例えば，図 2 に示す RDF グラフから「SemanticWeb の研究をしている B4 は誰ですか？」という質問文の返答を探索することを考える．この質問文からはキーワードとしてクラス「wiki:project」，プロパティ「wiki:Name」，リテラル「Seman-

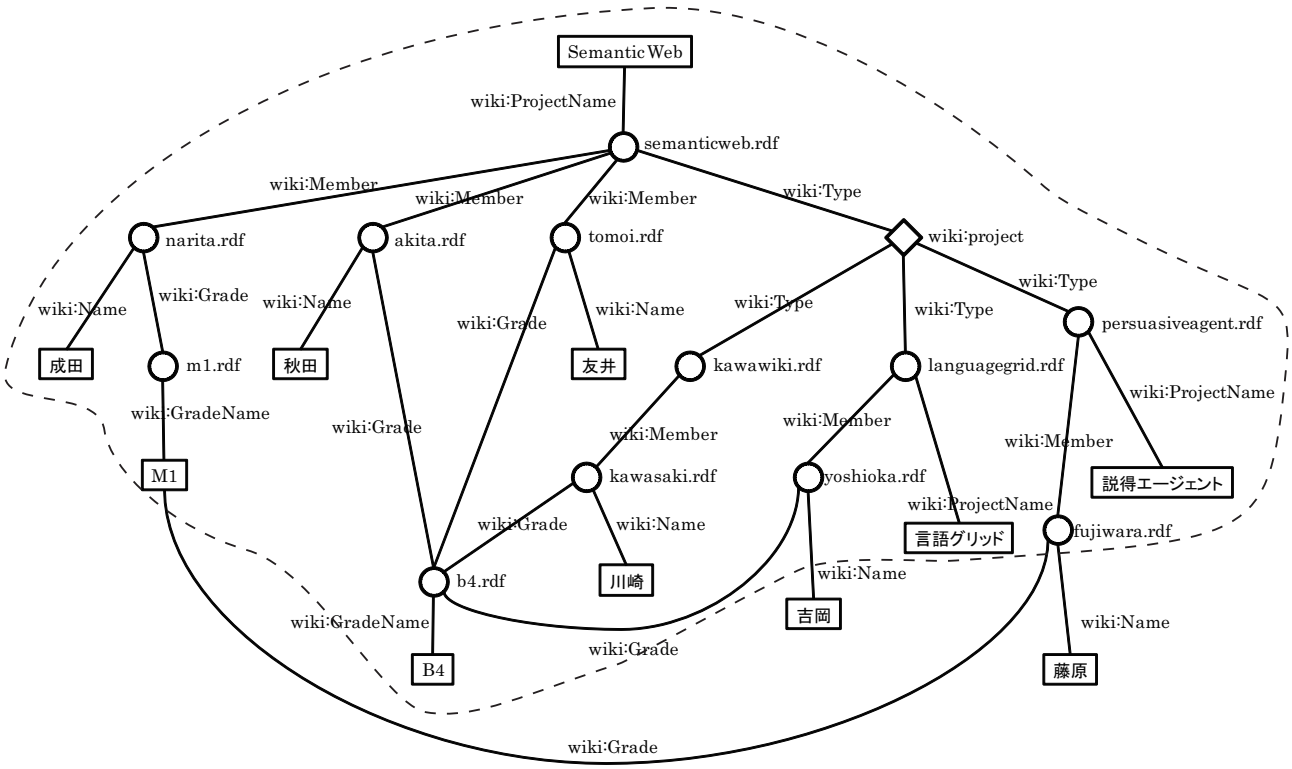

図 2　RDF グラフの探索例
Fig. 2　Example of searching RDF graph

ticWeb」,「B4」が得られる．ここではリテラル「SemanticWeb」を開始ノードして探索を行う．

質問文から得られたキーワードを全て含み，かつその時点での深さにあるノードを全て探索してサブグラフを抽出する．これは，複数の返答候補がある場合にはそれらのノードは RDF グラフ上の同じ深さに存在すると考えられるため，探索漏れを防ぐために行っている．図 2 の破線で囲まれた部分が探索されたサブグラフである．このサブグラフには，返答候補以外のノードが含まれている．そのため，質問文のキーワードを基に，不要なノードやエッジを削除する．まず，サブグラフの右下のノードとエッジから判定していく．リテラル「説得エージェント」とそれに接続しているプロパティ「wiki:ProjectName」は，キーワードとは一致しないため不要なノード判断し削除する．このような判定を繰り返していき，サブグラフから不要な情報を削除したものが質問サブグラフとなる．図 3 に抽出された質問サブグラフを示す．

3.3　質問サブグラフからの返答探索

抽出された質問サブグラフから返答探索を行う．返答として適切なのは，質問の対象となっているリソースの何らかのプロパティを表すリテラルである．そこで，質問文中には出現しないリテラルを返答の候補とし，返答候補と各キーワードとのグラフ上での距離（最短パス長）の総和を調べ，最小値のリテラルを答とする．

「SemanticWeb の研究をしている B4 は誰ですか？」という質問文から得られた質問サブグラフ中のリテラルは「SemanticWeb」,「秋田」,「友井」,「成田」,「B4」である．これらのうち，質問文中に出現しないリテラル「秋田」,「友井」,「成田」を返答候補とする．これらの返答候補のリテラルと，キーワードに含まれるリテラル「SemanticWeb」,「B4」との最短パス長の総和を調べる．図 4 に返答候補のリテラル「秋田」の最短パス長を示す.「SemanticWeb」までの最短パス長が 3,「B4」までが 3 となり，最短パス長の総和は 6 となる．同様に，返答候補「友井」,「成田」の最短パス長の総和はそれぞれ，6，8 となる．よって，最短パス長の総和が最小値 6 の「秋田」,「友井」が「SemanticWeb の研究をしている B4 は誰ですか？」という質問文の返答となる．

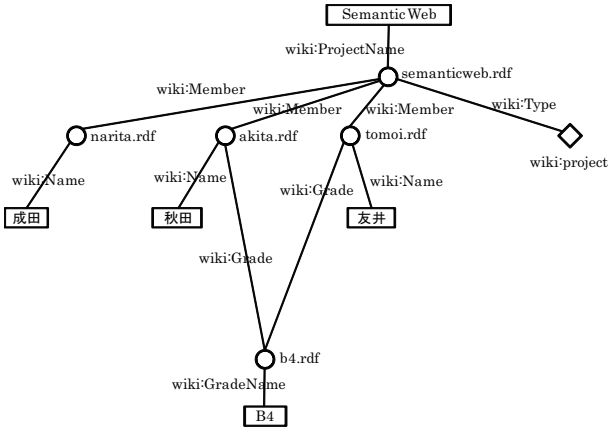

図 3　質問サブグラフ
Fig. 3　Question subgraph

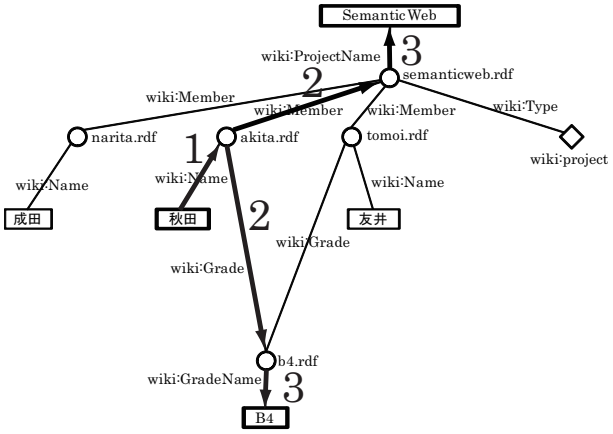

図 4　返答候補の最短パス長

Fig. 4　The length of shortest path of the candidate for the answer

表 4　正しく回答されなかった質問文の種類

Table 4　Type of query sentences which are wrongly answered

質問文の種類	文章数
情報源に回答が記載されていない質問文	2
否定表現が含まれる質問文	2
複数の質問項目がある質問文	1
諾否を問う質問文	7
数量を問う質問文	5
その他	7

4. 質問応答手法の評価

提案システムが自然言語による質問文に対し正しく返答できるかどうかを評価するために以下のような実験を行った．

4.1 実験内容

10 名の大学生，大学院生を対象に評価実験を行った．被験者には，研究室情報が掲載されている Semantic Wiki を閲覧してもらい，Wiki 上に記載されている情報が答えとなるような質問文を 8 文作成してもらった．Semantic Wiki としては KawaWiki [3] を用いた．Wiki 上には，研究室のメンバーの情報（名前，学年，趣味，参加している研究プロジェクト），研究プロジェクトの情報（プロジェクト名，プロジェクト内容，参加しているメンバー）が記載されている．Wiki 上の全ての Wiki ページにはメタデータとして RDF が付加されており，これらの RDF から RDF グラフを作成した．この RDF グラフを質問応答のための情報源として用いて，質問文に対する返答探索を行った．

4.2 実験結果

Wiki ページに付加された RDF を情報源として，提案システムを用いて被験者の作成した質問文に対する返答探索を行った．80 文の質問文に対して，正しく返答が行えた質問は 56 文であった．表 4 に正しく返答が行われなかった質問文の種類とその内訳を示す．

4.3 考　　察

実験の結果より，提案システムが正しく返答することのできなかった質問文について検証を行う．

表 4 より，情報源に該当する答えが記載されていなかった質問文は 2 文であった．これらの質問文では正しい返答は行われなかったが，関連する情報を提示することができていた．「KawaWiki の研究をしている M1 は誰ですか？」という質問文では，情報源には該当する人物名が記述されていなかったが，KawaWiki の研究をしている人物と学年が M1 である人物の名前が答えとして提示された．提案システムは質問文から得られるキーワードによる絞り込みを行い，質問サブグラフを抽出している．そのため，質問サブグラフに該当するデータがない場合でも，グラフ上でキーワードの近くにある関連するデータを提示することが可能となっている．

提案システムでは対応できなかった質問文は大きく 4 種類に分類できる．まず 1 つ目は否定表現を含む質問文である．提案システムでは，質問文中の否定表現を考慮せずにキーワードを抽出し，返答を探索している．そのため,「M1 で LanguageGrid の研究をしていない人は誰ですか？」と「M1 で LanguageGrid の研究の研究をしている人は誰ですか？」では，どちらの質問文からも「M1」,「LanguageGrid」,「研究」,「人」,「誰」というキーワードが抽出されるため，同じ返答が提示される．この問題を解決するためには，質問文中の否定表現の対象を特定し，質問サブグラフの探索時における制約条件として利用する必要がある．

2 つ目は，質問項目が複数ある質問文である．例えば，「ChieAkita の学年と趣味は何ですか？」という質問では,「趣味」の「sleeping」のみが返答され，学年については返答されなかった．これは，提案システムでは回答候補となるリテラルのノードと質問文に出現するキーワードのノードとの距離を求め，最小値の候補を回答として選択していることが原因である．そのため,「趣味」の候補である「sleeping」の方が「学年」の候補である「B4」よりも質問文のキーワードとの距離が小さかったため，回答として「sleeping」が出力されていた．この問題に対応するためには，質問文に出現する並列関係にある単語が，クラス，プロパティのいずれかであった場合には，質問対象が複数あると判断してキーワード間の最短パス長による絞り込みを別個に行う必要がある．

3 つ目は,「ChieAkita は言語グリッドプロジェクトのメンバーですか？」のような正否を問う質問文である．提案システムでは，情報源となる RDF グラフ上に記載された情報を探索し回答することを前提としているため，対応することは難しい．また，4 つ目の数量を問う質問文についても同様である．その他にも，口語表現による質問文などでは，係り受け関係が適切に解析できないなど，質問対象を特定できないため，正しい返答が行われなかった．

5. ま と め

本研究では，RDF 情報源を対象にした自然言語による質問応答手法を提案し，質問応答システムを開発した．提案したシステムは，ユーザからの自然言語による質問文からキーワード抽出を行い，抽出したキーワードを用いて Web ページの内容を表す RDF グラフから質問サブグラフを抽出することで，ユー

ザからの自然言語による質問に対する返答探索を行う．

研究室情報を情報源とし，被験者から集めた質問文に対し返答探索を行ったところ，80 文中 56 文について正しく返答することができた．今後の課題として，否定表現が含まれる質問文や，質問項目が複数ある質問文に対応できない問題に取り組む必要がある．

文　献

[1] E. Andre, T. Rist, and J. Muller. WebPersona: A Life-Like Presentation Agent for the World-Wide Web. Knowledge-Based Systems, Vol. 11, No. 1, pp. 25-36, 1998.

[2] O. S. Goh, C. C. Fung, K. W. Wong, and A. Depickere. An Embodied Conversational Agent for Intelligent Web Interacton on Pandemic Crisis Communication, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, pp.397–400, 2006.

[3] K. Kawamoto, M. Mase, Y. Kitamura, and Y. Tijerino. KawaWiki: Semantic Wiki Where Human and Agents Collaborate, Proceedings of the 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology Workshops(IWI08), pp. 147-151, 2008.

[4] 神崎正英．セマンティック・ウェブのための RDF/OWL 入門，森北出版株式会社，2005.

[5] 倉田岳人，岡崎直観，石塚満．係り受け関係に基づくグラフ構造を用いた質問応答システム，情報処理学会研究報告，自然言語処理研究会報告, No. 108, pp. 69-74, 2003.

[6] V. Lopez, M. Pasin, E. Motta. AquaLog:An Ontology-Portable Question Answering System for the Semantic Web, ESWC 2005, LNCS 3532, pp.546-562, 2005.