

質問応答における言語的な知識と一般的な知識の飛躍

Gap between General Knowledge and Linguistic Knowledge in Question Answering

竹内 孔一*¹ 松尾 彰悟*¹

Koichi Takeuchi Shougo Matsuo

*¹岡山大学大学院自然科学研究科

Graduate School of Natural Science and Technology, Okayama University

In this paper we discuss what kind of knowledge should be needed in question answering task or textual entailment task. The QA we are targeting here is not assumed to answer general questions, but to extract intended information from a fixed text set. This is because we want to clarify the gaps of knowledge between questions and texts analyzed with lexical knowledge. The lexical knowledge we assumed here is Lexical Conceptual Structure based predicate-argument structure and Generative Lexicon based nominal-argument structure that we are developing in our laboratory. To see the gaps, we check the Japanese-Language Proficiency Test N1 and N2 as well as NTCIR RITE-2 tasks because these practical tests contain the answers judged by human. In the thought experiment, we reveal that chain of events in daily life or events accompanied by the main event should be needed to answer the questions.

1. はじめに

本研究では文書集合に対して、必要な情報を質問によって取り出す質問応答システムの構築を目指している。つまり、質問者はおおむね対象とする文書の内容を理解しており、その中から時間を掛けて読めばわかる情報を計算機によってより早く取り出すタスクである。こうした状況は実際の日常生活でも想定できると考えられる。例えば受信メールの中からうろ覚えで過去にあったメールを取り出すことや、お知らせの文書から必要とする情報を取り出すなどである。こうした質問応答は例えば近年のモバイルアプリに搭載されているような一般的な質問に答える質問応答エンジンと異なるが、実社会で要求される自然言語処理の応用の一つと考えられる。

このように限定した質問応答では質問対象の知識が書かれていないことによる質問応答のエラーは起こらないと想定できる。こうした状況の中で、言語的な知識を積み上げて質問応答を行う場合、さらにどのような知識が必要かを明らかにしたい。この問題意識の背景にあるのは、言語学における語彙意味論からの語の意味の提案である。具体的には述語に対する項構造と状態変化構造を取り込んだ概念構造 [Jackendoff 90, Jackendoff 03, 影山 11], 生成語彙論 [Pustejovsky 95] に基づく名詞の意味構造により「A の B」といった語用論的な意味 [西山 03, 西山 13, 庵 07] もかなり語の意味として記述できる可能性が出てきた [影山 11][竹内 14]。当然、言語学からの提案は語の分類と見方であるので、これらを言語処理で利用するには形式化が必要である。形式化の部分はまだ開発中であるが、先に形式化が出来たと仮定して、それでも解けない場合、どのような知識が必要で、それがどのくらい獲得に大変そうなのか見積もっておくことが必要である。また一方で DNN など統計的学習モデルが日々進歩していることから、語彙からの積み上げ式の解き方における限界を明らかにしておくことは、これからの開発方向を検討する上でも重要である。

語の意味構造の積み上げによる質問応答および含意認識システムの構築で期待できることは、回答を誤った場合、問題部分

をモジュール毎に切り分けることができるという点である。既に分かっているモジュールとしては語の意味辞書の部分では、静的な辞書の部分と分野依存として動的に変わる部分が必要である。例を下記に示す。

- この車を買った
- この車に決めた

この例では、「買う」と「決める」は語の意味からすると異なるが、一方で、購入という場面では、ほとんど同義として処理する必要がある。分野依存ではあるが、意味的な操作タイプでいえば、単に類義語集合を作成するだけであるので、分野毎に類義語を学習するなどして、類義語を構築することで対応する。

一方で、こうした分野依存を超える意味操作がどのようなものか見積もるのが本論文の課題である。そこで日本語能力試験 N1[田代 11a] および N2[田代 11b] における情報検索課題に注目し、その中でも表や属性構造が明確な問題に対する質問応答システムを仮定して語の意味構造を超えた知識がどの程度必要か考察する。語の意味構造が明らかな場合含意認識タスクと近くなることから RITE-2*¹ の含意関係についても一部考察する。この理由は、人または計算機に対する課題の構築に置いて、背景知識に依存せずに書かれてる文書に対して普通ならば解けるであろうという範囲の問題と解答のセットになっていると考えられるためである。つまり、知識が不要な課題という設定である。これらの分析の結果から、背景知識が不要であっても、語の意味の積み上げでは解けない事態間の関係が必要であることを明らかにする。さらにそれらを作り込んでシステムに取り入れた場合の質問応答システムの簡単な結果について述べる。

2. 日本語能力試験 N1 と N2 および RITE-2 データの分析

まず日本語能力試験の情報検索に関する課題について分析し、後に含意認識タスク RITE-2 の課題について分析する。前

*¹ <http://www.cl.ecei.tohoku.ac.jp/rite2/>

大学名	日時	学部	備考
みやこ経済大学	①7/31(土) ②8/21(土) ③8/28(土)	経済	キャンパスツアー実施 ①②③11:00～ ③13:00～ 自由参加
ひがし大学	①7/17(土) ②8/1(日) ③8/22(日)	経済・経営・工	入試問題解説実施(①②のみ) 希望者は大学ホームページから 申込
野口工科大学	①8/22(日) ②8/29(日) ③9/5(日)	工	入試説明会(②③11:00～) 予約不要・入退場自由 詳しくは大学ホームページへ

図 1: 日本語能力試験(情報検索)の表の例(一部)

者が人間に対する課題であるのに対し後者は機械に対する課題であるが後に示すように語の単独の意味では解けない知識が必要であることが分かる。

2.1 日本語能力試験(情報検索)の分析

日本語能力試験は選択式の回答であるが、質問は与えられた文書(表や属性リスト)から論理的な解けるべき問題が提示されている。以下具体例を示そう。

図 1 に示すように各要素の中身は複数段で構成されており、要素の中で丸数字を利用した関係などが定義されている。仮想的な課題ではあるが実生活でよく見受けられる形式と考えられる。この表に対する質問は下記の通りである。

Q1 「右のページは、東京地区の大学のオープンキャンパス日程表である。下の問いに対する答えとして最もよいものを一つ選びなさい。」

Q1-1 「ヤンさんは、経済学部のオープンキャンパスに参加したいと考えている。一日ですできるだけたくさん回りたいが、いちばんたくさん回れるのは何月何日か。」

Q1-2 「クリスさんは、8月20日にこの表を見た。工学部のオープンキャンパスへ行こうと考えているが、クリスさんがしたほうが良いことはどれか。」

まず **Q1** でこの表についての説明があった後、2問の質問が出されている。この表に対する質問はこの2問だけが仮定されているが、図 1 に示すとおり、備考欄にかなり詳細な情報が有り、これらに関する質問も仮定すると簡単では無いことが分かる。また、**Q1-1** の質問では最も多く回れる日を表から(実際には 10 大学の登録がある)読み出すには人間でも時間のかかる作業である。さらに、問題はこれだけであるが(1)この表が検索結果に対するリストであったならば、より多くの項目が得られる可能性があり、(2)実際の大学を回る際には、時間との関係があることから、住所を調べて、Google API などを利用して、自動的に大学の回る最短経路まで必要である。つまり仮想的であるが、実問題として解くことが出来れば、応用範囲の広がる問題であることがわかる。

まずここで語彙的な意味として、また処理として日程などの表記「7/17(土)」などが月、日、曜日であることは既知として無視したい。また、表には必ず最初の行に各列の属性の説明があるとす。これにより例えば「学部」に対する問い合わせはその「学部」の列の要素の文字列とマッチングすれば良いこととする。また、**Q1-2** の質問ではすべき動作を問うているが、選択肢が用意されているためその選択肢が条件(8月20日かつ工学部)に合致するかどうかを見れば良いとする。

ここまで語彙意味の積み上げでとけない飛躍は次の通りである(ただし係り受け解析や照応解析など言語処理は正しく行われるものと仮定する。)

■休日等夜間救急診療所

【診療科目】内科, 小児科, 外科, 耳鼻咽喉科

○平日夜間 = 午後7時30分～10時30分

○土曜日 = 午後5時～10時

○日曜日・祝日・年末年始 = 午前9時～午後10時

図 2: 日本語能力試験(情報検索)の課題の例(一部)

G1 「X が Y (=オープンキャンパス(大学)) を回る」ということは「Y に行く」ということを含む

G2 「Y (=イベント) に行く」ということは開催されている必要がある

G3 「オープンキャンパス」というイベントは「日時」に書かれている日に開催されている

まず **G1** は動詞の意味の含意関係に関する意味構造である。Goo の辞書^{*2}を見ると「回る」の語義の中に移動に関するものが含まれている。

- 「2. 物の周囲に沿って、円を描くように移動する。」
- 「3. 順々に決まった場所などをめぐる。」

しかしながら、「Y に行く」という表層では書かれていないため、文字列ベースの処理であれば語義を調べてもマッチしない。一方、述語項構造ソーラス^{*3}では「回る」は

- 月が [対象] 地球の周囲を [経路] 回る【移動動作/循環】

として分類しており、「を格」を経路として通ることが分類できている^{*4}。このあたりはどちらかという語の意味構造の記述でまだ処理できる範囲である^{*5}。

一方で **G2** と **G3** は日常的な行動に関する知識であり、語の意味では記述できない知識である。これが個別問題に依存すると、処理システムとして取り込むことは難しいが、一般的に考えて、「イベントに行く/参加する」場合、開いていないと意味を成さないことが多い。実はこうした例は多数見受けられる。

図 2 は属性と属性値で整理されたお知らせの例である。これに対して下記のような質問が与えられている。

Q2 「右ページは、市役所からのお知らせである。下の問いに対する答えとして最もよいものを一つ選びなさい。」

Q2-1 「日曜日の午後 8 時頃、祖母が急に高熱を出した。医者に見てもらってはどうか。」

Q2-1 の質問に対して「日曜日」と「午後 8 時」がポイントとなるが、これも医療機関が開いているかどうかの問題で、上記のイベントの場合と同様に、日時の部分の解析とのマッチン

*2 <http://dictionary.goo.ne.jp/leaf/jn2/210283/m0u/回る/>

*3 <http://pth.cl.cs.okayama-u.ac.jp/vth/vths/search/見出し語/回る>

*4 ここで [] は意味役割, {} は動詞の共通語義概念を表している。

*5 **Q1-1** の前の方で「参加する」という動詞があるのでこれと同様であることが処理で扱えれば、「イベントに参加」→「イベントは開催されている必要がある」→「開催の日程や日時を見て開いているかどうかが見る」という処理を行うことは可能である。

グを行わなくてはならない。こうした処理は語の意味ではなく、イベントやサービスが「使える」時間であり、Generative Lexicon の枠組では Qualia Structure における Telic Role に相当すると考えられる。通常 GL は語に関する概念構造をの記述を仮定しているがこの例に示すようにインスタンス (この場合は「診療機関」そのもの) を表す意味構造としても有効であると考えられる。こうしたサービスに対する「利用」(つまり Telic role 相当) において必要となる情報は実社会においても提供されていなければ人間も利用することができない。よって社会として提供される可能性が高く、これらを GL の Telic role として記述することで、意味処理として扱える可能性がある。

2.2 RITE-2 の課題の分析

含意認識タスク (RITE-2) は t1 の文書に t2 の文書が含意されているかどうかを判定するタスクであり、含意される場合とされない場合のデータが構築されている。様々な種類がある中で下記の事例を取り上げたい。

t1 「伊坂幸太郎は直木賞候補になった 2003 年の『重力ピエロ』で一般読者に広く認知されるようになった。」

t2 「『重力ピエロ』は伊坂幸太郎による小説で直木賞候補作品だった。」

この例の正解判定は yes, つまり t1 が t2 を含意するという判断である。人が読んだ場合でも普通は含意すると考えられるが、この文の関係ですでに知識の飛躍が存在する。これを GL に近い項構造の形式で記述してみると下記ようになる。

tg1 「認知する (Agent=一般読者, Theme=伊坂幸太郎, Method=重力ピエロ)」という述語の意味は、「作品 (Agentive=wirte (Agent=伊坂幸太郎, Theme=重力ピエロ))」との関係が直接言及しているわけではない

つまり、t2 で伊坂幸太郎が「重力ピエロ」を書いたことが指摘されているが、t1 では伊坂幸太郎が有名になった手段として「重力ピエロ」が関係付けされているだけであるため間接的である。語の意味構造からしても、「知られる」ようになった手段が何か対象 (Theme) と関連があることは間違いないが、作者であるとまで書くことは出来ない。

このように言語処理システムを試験するための単純化した問題にまで、辞書ベースの処理システムからすると知識の飛躍が含まれており、人間の言語認識は積み上げだけでは不足する部分が少なくないことを示している。

3. 表に対する質問応答システムの試作

試作として日本語能力試験 N2 の図 1 に取り上げたオープンキャンパスの表に対する質問応答システムを構築した [松尾 15]。限定されているが、表を表層以上の内部表現を豊にすることでどの程度質問応答ができるか、またどのような知識が必要かを明らかにすることが目的である。

まず問題を扱いやすい形にするために下記の制約を設けた。

r1 表はテキストの形式で入力する

r2 表の最初の行には各カラムの属性名の記述がある

r3 なんの表であるかキーワードはあらかじめ与える

r1 であるが、実システムとしては画像から取り込むことも考えられるが [磯崎 15], 本研究はテキストに対する質問応答に焦点をあてているためここでは扱わない。テキスト形式の入力を図 3 に示す。

```
大学名 & 日時 & 学部 & 備考\n
○ ○ 大学 & 17/22 & 工・経済 & 入試説明会あり\n
& & 28/21 & & 13:00~\n
& & & 自由参加\n
\line\n
○ × 大学 & 8/12 & 工 & キャンパスツアー\n \line\n
× × 大学 & 19/5 & 農・経営 & 模擬授業実施\n
& 210/15 & & ホームページから予約\n
\line\n
...
```

図 3: オープンキャンパスの表の入力データ

また r2 を利用して、例えば「工学部のオープンキャンパスを教えてください」といった質問の際、「学部」のカラムで「工」の文字列がある行を取り出すという操作が可能になる。また質問文は独自に作成し、(1) システムの構成から抽出することが想定される質問文、および (2) 表を見て自由に作成した質問文、が解けたかどうかで評価する。

基本的な質問応答システムの構成は、(I) 表データの構造化、(II) 質問文の形式化、(III) 形式化した質問と表データをマッチさせて回答を返す。プログラミング言語として Scala を利用し、この中で、自然言語の質問文を形式化した関数リテラルに変換し、高階関数を利用して構造化した表データから解を獲得している。図 5 にこの概形を示す。

図 5 における isX は変数であるが中身は TableLine 型のデータを受け取って True/False を返す関数で有り、その条件式は自然言語の文から埋め込まれて固定化される。一方、表データ Table 型のインスタンス内において find 関数でこれを受け取り、マッチしたレコードの集合である Table 型データを返す。出力がまた Table 型になることから、複数の条件文、すなわち「8 月以降の工学部のオープンキャンパスを教えてください」など日時に対する条件と学部に対する条件の AND 操作をいくらかでも繰り返すことができる仕様になっている。

作り込みについて

ここでは 11/2 などは月/日と認識して Scala 関数における Date 型に変換する。Date 型は絶対的な日にマップするため何年か指定する必要があるが現在では、システムを実行した年を与えている。これにより、「8 月以降」など期間など指定した柔軟な質問が可能となっている。また「オープンキャンパス」をキーワードとしており、この言葉を係り先に持つ係り元の文節を条件文と見なして処理を行う。この部分が他の表の質問においてどの程度拡張可能化については検討が必要である。

質問文の形式化

質問文の形式化手法を簡単に述べる。係り受け解析器 CaboCha を利用して質問文を文節と係り受けに展開する。その後、各文節部分が、単に属性に関する文字列マッチの質問か、比較に関する質問かでタイプ分けを行う。

次に、属性に関する文字列マッチならば、表データの 1 行目に記述されている項目名を利用して、「大学名」「日時」「学部」を探し*6、係り元を属性値とする。

*6 備考の部分は全ての質問に関係するので属性としては処理しない。

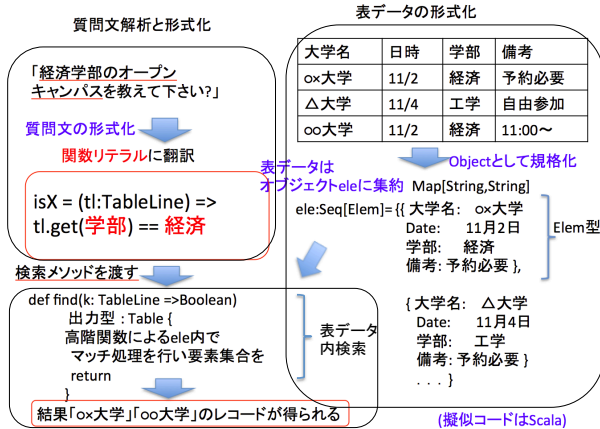


図 4: 表データの構造化と質問文の形式化による質問応答システム

例えば「8月以降の工学部のオープンキャンパスを教えてください」の場合、係り受け木から文節を2つ同定して、それぞれ処理を行う。

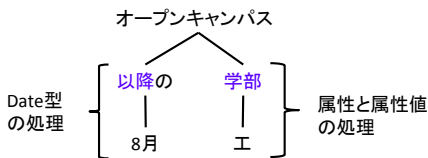


図 5: 係り受け木 (複合名詞も分解) から属性と属性値の取り出し

日時の場合は Date 型とのやりとりから時間操作に関する単語と時間操作の定義を記述した。例を下記に示す。ここでは、

時間に関する表現	日・時間に関する条件の生成
「X 以降」「X から」	条件 date.compareTo(X) >= 0
「X 以前」「X まで」	条件 date.compareTo(X) <= 0

表データの開催日 date と指定した時間 X との比較による操作を定義している。

実験結果

想定内の質問文 13 問を作成し、そのうち 9 問を正しく答えることが出来た。自由質問に対する回答では、備考欄に対する質問が多く、10 問中 2 問のみの正解となった。正解できたもの事例としては「予約のいない工学部のオープンキャンパスに参加したい。どの大学に行けばよいか」である。想定した質問での誤りでは係り受け解析の処理で『オープンキャンパス』の係り元』に限定したことが原因であった。例えば「9月より前でオープンキャンパスが行われるのは..」の場合、「前で」の文節が「行われる」に係ってしまい「オープンキャンパス」を飛び越えてしまったため、解析器が働かなかった。

4. まとめ

本稿では実問題に近い質問応答の事例や含意認識タスクにおける課題を取り上げて、語の意味構造の積み上げではとけな

い知識の飛躍部分を取り上げて分析した。その結果、イベントと時間に関する常識的な操作が必要性を明らかにした一方で、RITE-2 にあるように語の意味構造の積み重ねでは記述されていないが、「知られる」ようになった要因などはそれを生み出した対象と強い結びつきがあるはずだという推論を働かせて、含意すると判定する可能性があることを示した。今回は数例のみを取り上げたが、今後より分析を進めると共に、こうした知識の開きを獲得する方法について検討していく予定である。また、表データを対象とした質問応答システムを試作した。現段階では質問文の形式化では作り込みによる処理が扱える文の表現の幅を狭めている。しかしながら近年 Dependency Based Copositional Semantics (DCS) が提案されており [Liang 13], 本研究でも適用を試みたい。

謝辞

本研究の遂行に当たって科研費「言語処理及び言語分析を指向した大規模コーパスを利用した述語ソーラスの拡張」(26370485) の助成を受けた。ここに記して感謝する。

参考文献

[Jackendoff 90] Jackendoff, R.: *Semantic Structures*, MIT Press (1990)

[Jackendoff 03] Jackendoff, R.: *Foundations of Language: Brain, Meaning, Grammar, Evolution*, Oxford University Press (2003)

[Liang 13] Liang, P., Jordan, M. I., and Klein, D.: Learning Dependency-Based Compositional Semantics, *Computational Linguistics*, Vol. 39, No. 2, pp. 389-446 (2013)

[Pustejovsky 95] Pustejovsky, J.: *The Generative Lexicon*, MIT Press (1995)

[庵 07] 庵 功雄: 日本語におけるテキストの結束性の研究, くらしお出版 (2007)

[磯崎 15] 磯崎 秀樹: 論文QAのための画像処理〜表を読む〜, 言語処理学会第 21 回年次大会, pp. 139-142 (2015)

[影山 11] 影山 太郎: 日英対照 名詞の意味と構文, 大修館書店 (2011)

[松尾 15] 松尾 彰悟: 表やリストに対する質問応答システムの構築, 岡山大学大学院自然科学研究科修士論文 (2015)

[西山 03] 西山 佑司: 日本語名詞句の意味論と語用論, ひつじ書房 (2003)

[西山 13] 西山 佑司 (編): 名詞句の世界, ひつじ書房 (2013)

[竹内 14] 竹内 孔一, 竹内 奈央, 石原 靖弘: 述語項構造ソーラスによる述語と名詞の構造化, 人工知能学会全国大会, 2I5-OS-08b-1 (2014)

[田代 11a] 田代 ひとみ, 中村 則子, 初鹿野 阿れ, 清水 知子, 福岡 理恵子: 新完全マスター読解日本語能力試験 N1, スリーエーネットワーク (2011)

[田代 11b] 田代 ひとみ, 中村 則子, 初鹿野 阿れ, 清水 知子, 福岡 理恵子: 新完全マスター読解日本語能力試験 N2, スリーエーネットワーク (2011)