# 言語学の知見に基づく関数オブジェクトを利用した言語理解システムの構成

竹内 孔一　　石原 靖弘　　竹内 奈央
岡山大学大学院　　言語アナリスト
koichi@cl.cs.okayama-u.ac.jp, ishihara@cl.cs.okayama-u.ac.jp

## 1　はじめに

本発表では言語学の理論的分析，ならびに近年の統計的手法を利用した言語理解システムの枠組みを提案する．従来人工知能で提案された [5] [4][13][1] 理解システムを発展させる形として近年のソフトウェア科学で整理された関数オブジェクト (つまり高階関数) の受け渡しによる質問応答システムを仮定することで，静的な情報である言語理論，ならびに分野依存の特徴を獲得できる統計的手法とリンクさせることで，言葉の正規化をベースとしてた発展的な言語理解アプローチの可能性について論じる．

## 2　言語学の分析を意味処理の基礎データとして応用する見方

近年の言語学における理論的分析により，述語の項構造 (例えば [2] [8]) や名詞の意味構造 [3]，さらに日本語の「A の B」や「A は B だ」に関する名詞の意味構造理論的な分析 [11][9] が進んでいる．よって命題を構成する基本要素である名詞句と動詞句に対する意味構造の具現化が期待される．こうした言語学からの意味構造の提案は，基本的には文の表現を手掛かりにしたいわゆる文法を説明するためのものであるが，本研究ではこれらを意味構造への基礎データと捉えて，工学的な言語処理 (意味処理) に利用することを考える．

言語学の提案を利用して言語処理を行う研究はかつて自然言語理解として行われていた [5][4][13][1]．文書から単に情報を引き出すだけで無く，実世界との対応付けを行い言語で問題を解いたり [5]，プランニング [1] に応用するなどさまざまなシステムが提案された．

しかしながら現在は単純な質問ならば Web 上の検索エンジンが広範囲に活躍してほぼ上記の質問は不要であり，またシステムそのものも狭い世界での具現化のみで汎用性が不明であり，過去の言語理解システムをそのまま具現化する必要せいは感じられない．一方で，統計的学習モデルの発展により，語義曖昧性解消や，固有表現抽出の精度が向上したにも関わらず，文書から確実に質問に対して答えを取り出すことや，画像から物体を理解して人の言葉とマッチさせる [7] といった，昔の言語理解システムが提示していた課題は解けていないように見受けられる．著者は語彙概念構造を基に述語項構造シソーラスを構築し，述語表現のタイプについて調べてきた [6]．こうした言語資源は WordNet や FrameNet など現在様々に提案されているが一方で，分析から出てきたものであり，言語処理のどこで使えるのかというのが部分的にしか分かっていない．よってこうした言語資源を処理の中で位置づける上で新たな言語理解システムの構築を検討できないであろうかというのが問題意識である．

## 3　意味処理までの問題の整理

まずここで想定している意味処理について説明する．ここでの意味処理は内部的に異なる表現をある分野で同じと見なす (つまり言い換え処理) だけでなく，言語と外部ロボットやセンサー情報との対応関係が可能であったり (例えば [5] や [7])，言葉の指示でデータを検索を実行するなど人工知能に近い形での意味処理を仮定している．

さて，仮定した意味処理システムに対して言語学ベースのデータを利用して構築しようとする場合，明らかな問題がある．既に先行研究の言語理解システムの段階で問題になっているものである．批判を下記に整理する

**1.** 言語の背景にある知識は書ききれないぐらい深いつまりなにか抽象化した意味を記号化した瞬間に背景の実世界から離れてしまい，人間がするような言語理解は記号化では不可能なように見受けら

れる．

**2.** 名詞や動詞の意味は組み合わせで理解されているいわゆる「名詞の構造」や「動詞の構造」と独立して綺麗に書けない．

**3.** 自然言語処理システムは分野・タスク依存の情報が必要
クイズを解くならクイズの知識が必要．言語学だけの中立的知識だけではシステムはできない．

**4.** 汎用的な言語理解システムは難しいように見える
上記 **3.** と同様で分野依存の知識をいれて初めて使えるようになるので汎用は難しい．

上記の批判に対する考え方

**1 に対して:** 完全な記号化はできないにしても問題を限定して記号化し処理システムを作ることは可能である．問題はどの程度現実的な問題をとけるかであり，そこを試す必要がある．

**2 に対して:** 名詞の意味構造に述語を入れる枠組みは既に GL の qualia 構造などで提案されている [3][9]．しかし著者らの考察では，telic の中でも，誰に対する telic かで複数有りさらに複雑であると考えられる．例えば「診療所」の telic は患者にとって治療するところであり，医者にとっては働く所である．よって名詞と動詞の組み合わせで名詞の意味構造を書く必要がある..

**3 に対して:** 本質的に起こる問題で，逆に Web 上の Wiki や統計的学習モデルなど利用して分野・タスク依存知識を集め，それらをシステム内で利用して質問応答など応用レベルで評価できる枠組みを構成する必要がある．

**4 に対して:** 適用する有用な分野・タスクを探して構築することで回避したい．次節で述べるように，文学作品の解釈などは現段階では可能性が低いと考えられるが一方で，掲示物や案内，広告に対する質問応答は広く浅い社会構造の知識と言語の知識が必要で，これらを蓄積しつつ処理する言語理解システムを構築することは有効ではないかと考える．またシステムがカバーしていない知識に対しても，節 4.2 に示すように関数プログラミングを利用して method だけを押さえておき，具体的な名詞などは人間 (文書を書いた人) と人間 (質問者) が同義表現を使うことでマッチするようなシステム構成で書き切れない表現の扱いを押さえたい．

# 4 質問を関数オブジェクトで渡して解を得る質問応答フレームワーク

まず対象とする分野について述べた後，具体的な質問応答の構成全体像について記述する．

## 4.1 質問応答を対象とする分野

前節の批判 **3.** と **4.** より分野を絞る必要がある．構築して有効な質問応答システムは事実や技術といった背景が明確で，質問する内容がはっきりしており回答が正しいかどうかも理解しやすいものを対象とする[1]．具体的には疑似問題であるが「日本語能力試験 N2」[10] の情報抽出問題を対象とする．問題の例として，例えば診療所の空いている時間に関する広告に対して，ある症状の時どうすべきかなどを質問する．生活に即した少し深い知識 (「診療所は治してもらう場所」や「内科は風邪のときに見てもらう分野」など) を必要とする．もちろん問題は選択式なので選択すればなにか答えられるが，本研究の立場ではこれをテストベッドとして構文的な処理，言い換え，社会情報に関する知識を適用しつつ理解システムを構築していく立場を取る．

## 4.2 関数オブジェクトを利用した質問応答の枠組み

問題の設定として，質問対象となる文書が存在し，質問文もあるとする．このとき，基本的な枠組みとしては概略下記のようになる[2]．

**(1)** 質問対象の文を規格化しておき，今後呼び出される特徴量や質問文書から取り出されるべき内容を構造化 (オブジェクト化) しておく．

**(2)** 質問文は文書に対して何をどういう条件，どういう範囲で引き出すかを指示している．よってその言葉を関数オブジェクトとして検索方法を指定する高階関数に変えて (1) の DB に渡し，(1) 内の持つオブジェクトの属性に対して条件マッチしたものを取り出す

これにより質問応答システムを具現化する．つまり中間モデルはプログラムそのものであり，言語からプロ

[1] 反対は小説を対象にした主人公の気持ちの理解などの質問応答タスク．同じく N2 内の課題であるが現状では人の感情に関する知識をいれる必要がありとける見込みが無い．

[2] この考えに沿う提案が本会議で既になされている [12].

グラムに変換する部分がいわゆる言語処理技術を活用する部分である．変換先はプログラムであるため，正しいか間違っているかはっきりわかる．これにより，例えば従来，意味役割ラベルは中間表現として存在するだけであったが，プログラムに変換するための文の正規化システムの一部として利用し，その解析精度についても評価することが出来る．

具体的な例を図 1 で示そう．簡単のために文書ではなく表に対する質問の場合を示した．また，関数オブジェクトの簡易表記として Scala による表記を採用している．

図 1 の設定として大学のオープンキャンパスを実施する大学名，日時，学部，注意事項が記載された表があったとする．この表に対して，「経済学部のオープンキャンパスはどれか?」という質問を行うとする[3]．このとき，表データは内部に属性と属性値の item として集合 (Seq) として記録しておく．その時の属性名は表のデータにある「日時」などの言葉をそのまま利用する．Map は連想配列である．1 大学分の情報を Elem 型のインスタンスで格納しているとしよう．表全体のオブジェクトは集合で内部の集合に対してなにか与えられた条件 (関数オブジェクト k) を受け取ってループを回すメソッド find が定義されているとする．この時，質問文をプログラムの質問条件 (マッチ条件) isX(e: Elem):Boolean を言葉から作成できれば正しく，経済学部の大学のみ取り出せる．ここで isX は Map 型を受け取って Boolean を返す関数オブジェクトである．

ここでの例は大変単純なものであるが，関数オブジェクトは基本的にはどんな手続きも構築できるので自然言語の質問とと関数オブジェクトの対応関係をくみ上げていくことでより複雑な質問に対応させることが可能であると考えられる[4]．

## 5　提案システムに対する考察

さて，3 節の問題に対して提案手法はどのように対処しているか整理する．問題 **1.** と **2.** に対しては名詞の構造化の部分で対応することになると考えられる．名詞は動詞に比べていくつもの側面を持っており，名詞だけで使い方や注目すべき部分などが明らかになる (例えば「借金」というだけでいろいろな状況が仮定される)．よって名詞の意味構造が本質的な難しさと考えられる．

問題 **3.** は基本的に単語やフレーズ間の同義関係 (分野依存) の計算に関連する．今回質問文は「経済学部」という言葉を使い，表では「経済」という省略を用いている．ここでは「経済学部」と「経済」は同義として扱って良いが，こうした知識が問題依存で関係してくる．手法としては知識というより統計的学習モデルを利用したアプローチの導入が必要となると予測される．

問題 **4.** は提案システム自身は設計したとおりにしか動かない closed なプログラムであるが，一方，文書と質問文はシステムの外側の情報である．よって 3 節でも述べたとおり，人間同士の言葉の一致でマッチングが成功する場合もある．また統計的学習モデルは学習データも外側の情報であるため，積極的に統計的学習モデルを利用することで，幅広い分野に対して言語理解システムが動作するように構築する必要がある．

## 6　まとめ

本稿では言語学的知見を意味処理のための基本概念として捉え，現実に近い質問応答システムと結びつけることで評価および，言語知識をプログラムという形で蓄積する手法を提案した．言語学的知見とはここでは言語の正規化 (言い換え) に関する処理であり，プログラムとして組み込むことで，言語処理として扱いやすい形となる．また質問とのマッチングにおいて，社会常識に近い知識が必要となるため統計的学習モデルと接続することが出来る．またこの点においても，うまく解析できた単語やフレーズ間類似度を質問応答システムに取り込むことで，知識を蓄積することが可能となる．また，質問文を関数オブジェクトとして言語表現と質問を構造化した中でマッチさせることで，用意していた属性外の入力に対して質問応答ができる可能性を示した．今後は具体的に上記の質問応答システムを構築する予定である．

## 参考文献


[1] James Allen. *Natural Langauge Understanding (2nd Edition)*. Addison-Wesley, 1994.

[2] Ray Jackendoff. *Semantic Structures*. MIT Press, 1990.

[3] J. Pustejovsky and Martha P.and A. Meyers. Merging propbank, nombank, timebank, penn


[3] 日本語 N2 の問題を簡略化したもの．

[4] N2 の元々の質問は「経済学部を 1 日で最も多く回れる日は何月何日か」でありもう一段階ループが必要な質問であった．

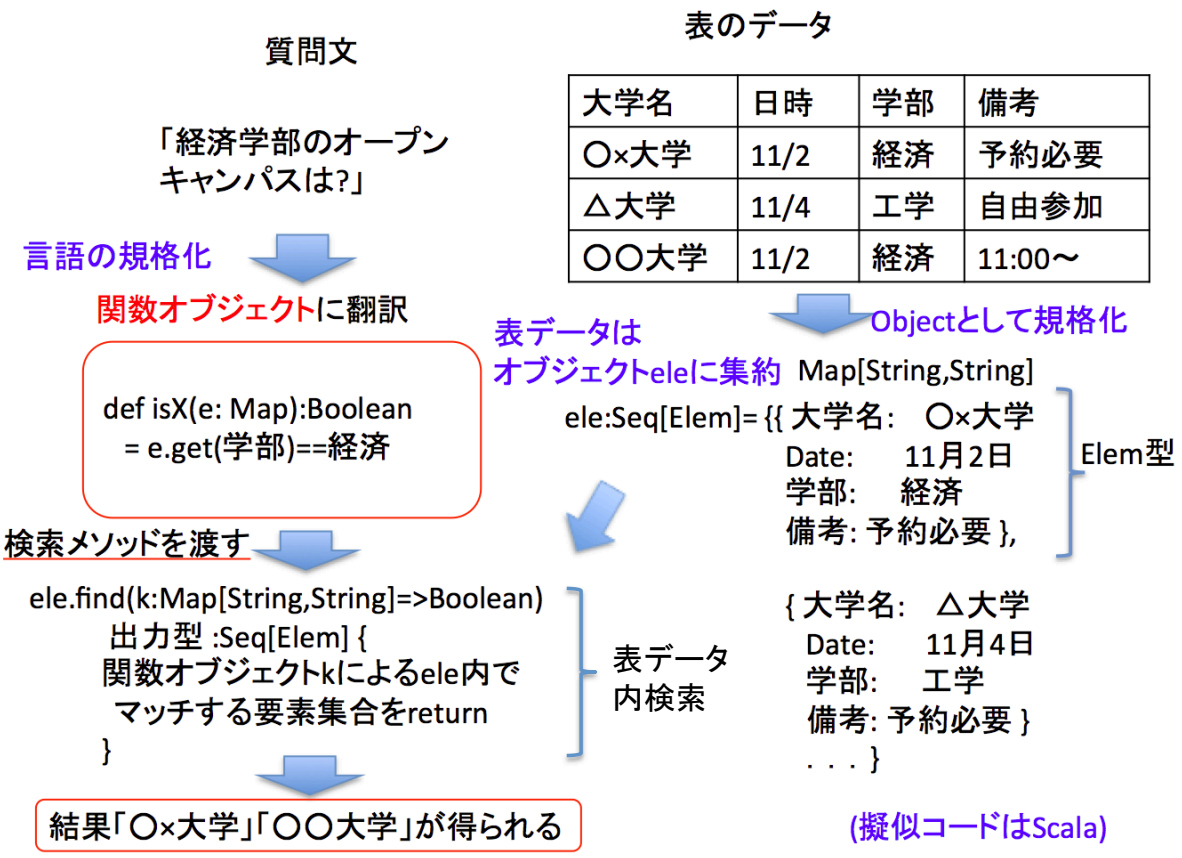


図 1: 関数オブジェクトを受け取り回答を返す質問応答システムの例

discourse treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 5–12, 2005.

[4] R. C. Schank. Conceptual dependency: A theory of natural language understanding. *Congnitive Psychology*, Vol. 3, pp. 552–631, 1972.

[5] Terry Winograd. *Understanding Natural Language*. Academic, 1972.

[6] 竹内孔一, 竹内奈央, 石原靖弘. 述語項構造のシソーラス分類と意味役割の設計について. 人工知能学会全国大会, pp. 2D4–OS–03a–1, 2013.

[7] 小林瑞季, 小林一郎, 麻生英樹. 動画像中の人の動作を表現する確率的言語生成に関する取組み. 人工知能学会全国大会, pp. 2D5–OS–03b–3, 2013.

[8] 影山太郎. 動詞意味論. くろしお出版, 1996.

[9] 影山太郎. 日英対照 名詞の意味と構文. 大修館書店, 2011.

[10] 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.

[11] 西山佑司. 日本語名詞句の意味論と語用論. ひつじ書房, 2003.

[12] 山田隆弘. 語彙概念構造のオブジェクト指向化について. 言語書学会第 20 回年次大会, 2014.

[13] 高木朗, 伊東幸宏. 自然言語の理解. 丸善出版, 1987.