様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ７ 月 １５ 日現在

機関番号：９２７０７
研究種目：若手研究(B)
研究期間：2016～2019
課題番号：１６Ｋ１６１２０

研究課題名（和文）数値データを対象とした質問応答システムの構築

研究課題名（英文）Question Answering System for Numerical Data

研究代表者
横野　光（Yokono, Hikaru）

株式会社富士通研究所・その他部局等・研究員（移行）

研究者番号：６０５３５８６３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：本研究では統計情報などのような数値データを対象とした質問応答システムの開発を目指した．数値データから得られる情報にはデータに格納されている値だけでなく平均のようなデータの値を計算することで得られるものもあるため，質問応答システムにはデータ中の値を計算して回答する機能が必要となる．
そのようなデータを対象とした質問応答システムを開発するために，数値データと言語表現との対応を人手によってアノテーションしたデータセットを構築した．
また，関連研究として言語理解タスクにおいて，モデルに必要となる知的処理の分類を行い，それに基づいていくつかの言語理解タスクのデータセットの分析を行った．

研究成果の学術的意義や社会的意義
宣言的な知識が格納されたデータベースを知識源とするような一般的な質問応答システムでは質問に対する答えがデータベースに直接的に表現されているという仮定に基づいてモデルが構築される．一方で，数値データを対象とした質問応答システムでは，回答に必要な情報がデータ中に直接的に存在しておらず，回答のためにはデータ中の値を計算しなければならない，という状況が起こりうることを想定し，それらの計算のための機構を構築する必要がある．本研究ではその基礎となるデータセットの構築を行った．このデータセットを用いることで，言語表現による数値データの取り扱いに関する研究が行えると考えられる．

研究成果の概要（英文）：In this study, we aimed to develop a question answering system for numerical data, such as statistical information. Since some information obtained from numerical data can be obtained not only by referring the values stored in the data but also by calculating the value of the data such as the average, the question answering system for numerical data need the function to calculate the values in the data.
In order to develop a question answering system for such data, we constructed a dataset in which the correspondence between numerical data and linguistic expressions was manually annotated.
As a related study, in the machine reading task, we classify the intellectual processes required for the model, and analyze some datasets of the language understanding task based on that.

研究分野：自然言語処理

キーワード：自然言語処理　意味解析　コーパス構築

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年のオープンデータ化を受け，統計データやセンサの観測データなどを一般の人が使える形で公開するという流れができ，数多くの数値データを利用することが可能になった．しかし，データの規模が大きくなるにつれて，ユーザが必要とする情報を獲得するためのコストも大きくなる．この情報獲得のためのコストの増大は数値データに限った話ではなく，テキストデータなどでも発生する問題であり，このような問題に対して検索システムや自然文による質問を想定した質問応答システムが活用されるようになった．

一般的な質問応答システムは，自然文による質問の意味解析を行い，保有している知識ベースから回答となる情報を検索し質問者に提示する．このようなシステムの多くは質問の解答そのものが知識ベース中に存在しているという仮定に基づいてその情報の検索を行っている．

一方で，数値データから得られる情報には，数値自身の他に，平均などのような，ある基準値とその数値との大小関係や，前後における変動の様子などがあり，我々はこのような情報をデータ中の値からの計算によって獲得することができる．従って，数値データを対象とした質問応答システムでは，例えば気温データに対して「○月の気温は何度だったか」のように，一般的なシステムと同様に回答がデータ中に存在する場合だけでなく，「今月で最高気温を記録したのは何日か」のように回答するためにデータ中の値から計算する必要がある場合を想定しなければならない．そのため，数値データを対象とした質問応答システムでは質問文から必要な値を検索するだけでなく，データ中に回答となる値が直接記述されていない場合に必要な値を検索し計算して回答する機能が必要となる．

２．研究の目的

本研究では統計情報などのような数値データを対象とした質問応答システムの開発を行う．数値データから得られる情報にはデータに格納されている値だけでなく平均のようなデータの値を計算することで得られるものもあるため，質問応答システムにはデータ中の値を計算して回答する機能が必要となる．そこで本研究では自然文による質問の意味解析において計算を考慮した形式記述を定義し，それを用いて回答の検索と計算を行うことを目指す．また，ユーザにとって分かりやすい回答の出力として，実際の数値データをそのまま表示するだけでなく，その解釈となる言語表現の生成を行うことを目指す．

３．研究の方法

本研究における課題は，数値データを対象とした質問応答システムについて，計算操作に関するオントロジーの構築，計算過程を考慮した質問文の意味解析(基本的な質問文の解析，計算を必要とする質問文の解析)，また，回答出力に関して，数値データの言語的解釈のための分析，数値データの意味解析，である．このうち，基礎となるデータ構築と意味解析器の構築を行う．データ構築に関しては，気象データや交通量データなどいくつかのドメインの数値データを収集し，その内容を表す言語表現とのセットを用意し，言語表現が指し示す数値データの箇所をアノテーションすることでデータセットを構築する．また，テキストの意味解析器として述語項構造解析器の開発を行う．述語項構造解析の問題の一つに省略の対応が挙げられるが，これに対してはこれまでも研究を進めている一貫性を考慮する手法(a)の改良での対応を行う．

４．研究成果

(1) データセットの構築

研究開発のためのデータセットとして，気象庁の日本の天候の特徴と見通し(https://www.data.jma.go.jp/gmd/cpd/longfcst/)を対象とした．この中の天候のまとめでは，各季節毎の天候についての概況や各月の気候の変化がテキストで記述されている．また，同サイトでは過去の気象情報が掲載されている．この天候のまとめにあるテキストと該当する期間の日毎の気温，降水量，日照時間などの気象データを１組として，テキスト中にある気象データに基づいて記述された表現の同定と，その表現の根拠となる数値データの箇所との対応を人手によりアノテーションしたデータを構築した．

元となる気象データは観測値からなる数値データであるが，概況として記載される情報は数値だけではなく，例えば，「平年や前の月と比べて高かった」，といったような言語表現で表されていたり，「月の半分は曇りだった」のように，日照時間の少なさが曇りを表すといったような間接的な表現，また，月全体に対してその半分といった範囲に対する表現などとして表されることが多い．そのため，アノテーションにおいては，テキスト中の該当する言語表現に対して，数値データの点だけでなく範囲の指定，また複数の範囲の指定などを行い，さらにそれらに対してどのような演算処理(例えば，平均を求める)が行われてその言語表現が生成されたと考えられるか，といった情報を付与した．

アノテーション作業の際に挙がった問題の一つとしては「平年よりも気温の高い日が多かった」のように「平年よりも気温の高い日」でまずある月の日のうち条件を満たす日を見つけると

いう操作が行われ，次に「その日が多かった」という表現で数えるという操作が行われる，といったように多段の演算を行う必要がある表現の存在が挙げられる．今回のアノテーションではそのような表現は対象外としたが，実際にはある演算の結果に対してさらに演算を行う，ということは柔軟な質問応答には必要であると考えられるため将来的な課題とする．

作成したデータセットは研究期間終了時ではまだ未公開であるが，アノテーション対象となった気象庁のデータと作成したアノテーションデータの分離を行い，分離したアノテーション情報と気象庁のサイトから各自で取得したデータから実際のアノテーションデータを復元するスクリプトを整備した後に公開する予定である．

(2) 意味解析器の構築

一貫性を考慮することにより文間にある述語項関係の同定の性能を向上させることを目指したが，使用している一貫性モデル自体の性能があまり良くなかったため，それを利用する述語項構造解析の精度の改善は見られなかった．

現在では BERT や XLNet のようなニューラルネットベースの言語モデルを用いることで意味解析の精度の大幅な改善が見られているため，今後はこれらのモデルを考慮に入れた意味解析器の構築を行う．

(3) 言語理解タスクの分析

自然言語処理の目的の一つとして言語の理解が挙げられる．これに対して様々な言語理解タスクが提案され，それに対応するデータセットが構築されてきた．近年ではニューラルネットベースの言語モデルの隆盛により言語理解モデルの性能は飛躍的な向上を見せている．

多くの言語理解タスクではテキストに対して与えられた問いに答えられるかという総合的な評価を行っているが，この方法ではモデルは様々に存在する知的処理のどれが苦手なのかが分からず，モデルの改善を行うことが困難となる．

そこで言語理解で必要とされる知的処理にはどのようなものがあり得るかについて整理を行い，これに基づいて言語理解タスクのデータセットの分析を行った．その結果，解答に網羅的な知的処理能力を必要とするようなデータセットは少なく，データセットによっては必要となる知的処理に偏りがあることが明らかになった．

<引用文献>

(a) 横野光, “テキスト一貫性の特徴を用いた述語項構造解析結果の精度比較モデル”， 言語処理学会第 22 回年次大会，2016．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Saku Sugawara, Yusuke Kido, Hikaru Yokono and Akiko Aizawa
２．発表標題 Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability
３．学会等名 Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)（国際学会）
４．発表年 2017年

１．発表者名 Saku Sugawara, Hikaru Yokono, Akiko Aizawa
２．発表標題 Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems
３．学会等名 Thirty-First AAAI Conference on Artificial Intelligence（国際学会）
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考