研究課題名　　教師なし学習による対話エージェントの構築
所属研究機関名　北陸先端科学技術大学院大学
研究者氏名　　鳥澤　健太郎

Ⅰ．研究計画の概要

■ 研究の趣旨･目的

本研究課題における目的は、大量のテキストから教師なし学習によって学習された知識をもとに不特定多数のユーザーから日常に関する知識を収集し、その情報を整理した上で各々のユーザーに提供できる対話エージェントを構築することである。これにより、通常のインターネットでの情報発信を、その技術的問題を意識することなくユーザーに行わせ、多くの情報を組織化された形でネットワーク上で共有することを可能にする。より具体的には、これまでに研究してきたシソーラスおよび意味ネットワークの教師なし自動学習手法を大量のテキストに適用し、そこで生成されたシソーラス及び意味ネットワークを元に、可能な限り多くのトピック、分野についてユーザーと自然言語で対話し、情報収集、情報の整理、提供を行う open domain なシステムの研究を行う。

近年、インターネット上の情報は日増しに増加しているが、インターネット本来の利点であった個人ベースの情報発信は満足の行く形態にはなっておらず、企業ベースのものに圧倒されているのが現状である。真に豊かなネットワーク社会を実現するためには、このような現状は望ましくなく、本研究課題で目指すような一般個人の「草の根」からの情報発信を助けるシステムを開発することによって、回避すべきだと考える。

より具体的な例をつかって述べると、従来よりグルメに関するホームページが整備されてきているが、現状では、そのようなホームページを介して情報の発信をする個人はごく限られた人々であり、また、掲載されているレストランも量的にはごく少数である。本研究課題の主要なターゲットをこの例に即して述べれば、能動的に一般ユーザーに対してレストランに関する質問を発することにより、グルメページで提供しているような情報をより多くのユーザーから収集し、他のユーザーに提供するシステムを研究開発することである。具体的な技術に関して云えば、現在あるグルメページを見ると、レストランに関する情報として「必須」のものが存在するということがわかる。例えば、メニュー、場所、定休日、味に対する判断といったものがそれである。我々のターゲットを実現するのに必要な技術は、語、あるいはホームページに対してこのような必須の情報を見つけだし、その情報に関する質問を行う技術である。より詳しく云えば、ホームページ、あるいはホームページ中にあらわれる語をその意味に従って分類し、レストランという語に対するメニューという属性のように、それらの語やホームページの分類で必須の情報とされている属性、述語を計算し、その結果に従ってユーザーに質問、その解答を整理する技術である。

■ 研究計画の概要

研究に際しては、各年度毎に以下の研究目標を達成すべく研究、開発を進める。

平成13年度　教師なし学習の改良洗練とインターネット上のテキストへの適用
平成14年度　学習された意味ネットワークを利用したユーザー発話の意味解釈手法の開発
平成15年度の目標　ユーザーへの質問の生成手法の研究
平成16年度の目標　ユーザーから収集した情報の整理、提供手法の研究
平成17年度の目標　システムインテグレーション

II. 研究計画の詳細報告

（単位：百万円）

研究項目	所要経費					
	13 年度	14 年度	15 年度	16 年度	17 年度	合計
1　意味ネットワーク学習の改良、インターネット上のテキストへの適用	16					16
2　学習された意味ネットワークによるユーザー発話の意味解釈		5				5
3　ユーザーへの質問の生成の研究		6	13			19
4　ユーザーから収集した情報の整理、提供手法の研究						
5　システムインテグレーション／評価						
所要経費（合計）（間接経費を含む）	16	11	13			40

III. 研究成果の概要

■ 研究成果の概要

平成13年度-平成15年度に計画した研究項目、研究目標は以下の3点である。

(1) 意味ネットワーク学習の改良、インターネット上のテキストへの適用 （平成13年度）
(2) 学習された意味ネットワークによるユーザー発話の意味解釈（平成13年度、14年度）
(3) ユーザーへの質問の生成の研究 （平成14年度、平成15年度（予定））

まず、項目１の「**意味ネットワーク学習の改良ならびにインターネット上のテキストへの適用**」では、従来より行ってきた意味ネットワークの学習手法で使われていた確率モデルを改良し、それを従来より大規模なテキスト（新聞33年分）に適用し、学習された結果を項目2、3の研究に利用した。また、意味ネットワークの学習と同時に、格フレームや単語意味クラスが得られるが、特に、この単語クラスを人手でチェックし、その精度を算出した。また、意味ネットワークの学習手法は Expectation Maximization 法とよばれる統計的アルゴリズムを利用したものであるが、このアルゴリズムをインターネット上で頻繁に使われる「表」に適用して知識獲得学習を行い、その結果をもとにネット上のテキストから情報抽出を行う手法についても研究した。さらに学習の前処理として必要な構文解析手法の洗練に関する研究も行った。

項目2「**学習された意味ネットワークを利用したユーザー発話の意味解釈に関する研究**」では大量のテキストコーパスから学習された意味ネットワーク、格フレーム、単語意味クラスを用いて、文間の意味的論理的関係をコーパスから抽出する手法を研究した。これにより、自然言語理解、ないしは対話エージェントの構築に必要であると従来より考えられてきた「スクリプト」に類似した知識あるいは「推論規則」をある程度の精度でコーパスから抽出することが可能となった。実際にこれまでの実験では、「もしXがビールを飲めばビールに酔う」あるいは、「もし、Xが本を書けば X が本を出版する」などの推論規則が大量のテキストから学習、抽出されている。なお、本項目に関する研究では第9回言語処理学会年次大会優秀発表賞を受賞した。

この研究をとおして得られた知見としては、推論規則などの高レベルの知識において、名詞が大きな役割を果たすということが挙げられる。従来より、推論規則やスクリプトの研究においては、名詞ではなく、動詞の果たす役割が重視される傾向があった。今回得られた知見は、このような見方に一石を投じるものと考えられる。今のところ、名詞が推論において大きな役割を果たす理由としては、我々人間が推論において、名詞を介した文の間の相互作用を重視しているからではないかという仮説をたてている。例えば、「もしXがビールを飲めば、ビールに酔う」という推論規則は常識に照らし合わせて妥当であると考えられる。この規則は二つの文からなっているが、それらの二つの文では名詞「ビール」が共有されており、さらには、二つの動詞、「飲む」「酔う」はビールに限らずアルコールを含んだ飲料一般を共有する可能性が高い。別のいい方をすると、「飲む」と「酔う」の間には、アルコール飲料を介して相互作用が生じている可能性が高いということである。一方で、「風が吹けば、桶屋が儲かる」という文、あるいは推論が奇妙に聞こえるのは、「吹く」と「儲かる」の間でなんら名詞が共有されないからであると考えることができる。つまりは、「吹く」という出来事と「儲かる」という出来事の間に名詞の共有を介した直接的な相互作用がなんらないために、奇妙に聞こえるわけである。

以上の仮説は、現状では狭い範囲の推論、あるいは推論規則の学習に対してしか適用されていない。今後さらに広い範囲の、またより複雑な推論規則に関して以上の仮説が成立するのかどうかを検討、実験していく予定である。また、現段階では言語の枠の中だけで議論をしているが、名詞はそれが参照する概念、あるいは具体的存在物であり、動詞はそれが指し示すイベント、出来事であると読み替えれば、以上の仮説が言語の枠を離れて人間の認識一般に対して適用可能となろう。これまでに述べてきた仮説で行ったように、名詞が参照するような対象の相互作用によって、推論、認識一般をとらえ直すのは興味ある研究課題であると考えている。

また、研究項目3の「**ユーザーへの質問の生成の研究**」は、来年度（平成15年度）に達成すべき目標に関する研究項目であるが、これまでのところ研究項目 2 で得られた推論規則や、研究項目1で自動学習を行った意味ネットワークや、意味クラスを元に質問を生成する手法、アルゴリズムに関して研究を行った。現状では、例えば、「レスト

ラン」あるいは「飛行機」といった単語に対して、

「そのレストランは X まで営業するか？」（X は時間を示す変数であり、レストランの営業時間を訊いている）

「そのレストランで X を食べるか？」（X は食物を示す変数であり、レストランでのメニューを訊いている）

「その飛行機で X に向かうか？」（X は空港などの場所を示す変数であり、飛行機の行き先を訊いている）

「その飛行機が X を出発するか？」（X は空港などの場所を示す変数であり、飛行機の出発地を訊いている）

といった、一般ユーザーにとって重要な情報に関していわば「擬似的な質問」を生成できるようになった。（現在までのところ、変数 X を「何」「何時」「どこ」といった通常の表現で置き換えるにはいたっていない。）アルゴリズムは、入力された単語がさす概念の「用途」「利用法」に関して質問をするように設計されており、厳密な実験はまだ行っていないが、プロトタイプシステムの出力は、確かに「用途」「利用法」といった意味的情報を反映していた。開発した手法は、ある仮説に基づき、基本的に頻度、確率などの統計的情報のみによって質問を生成しているが、このような統計的情報から「用途」といった高次の意味的情報に関連した質問が生成されるというのは興味深い。

■ 波及効果、発展方向、改善点等

以上に研究成果の概要について述べてきたが、現在の研究、あるいはこれまでに開発してきた学習手法は、残念ながら今すぐに実用化できるような精度を達成しているとはいえない。例えば、前述の推論規則の学習にしても、一定の割合で、我々の常識に反する推論規則も学習されてしまう。今後第一に改善しなければならない点はこの精度である。改善策の第一はより大量のテキストを入力として自動学習を行うということである。これは今後より大量のテキストをインターネットからダウンロードすることにより対応する。また、改善策の第二として考えられるのは、単語意味クラスの自動学習手法さらに拡張発展し、単語意味クラスが低頻度語を含むようにし、精度を低くする原因である低頻度語に関して推論や質問の生成を行う場合には、対応する単語意味クラスで低頻度語を置き換え、一般化するようにすることである。これまでに開発してきた手法でもそのような処理は行っているが、より広い範囲の低頻度語にたいして上述の処理ができるように拡張する必要がある。

Ⅳ. 研究成果発表等の状況

⑴ 研究発表件数

	原著論文による発表	左記以外の誌上発表	口頭発表	合　　計
国 内	0 件	0 件	4 件	4 件
国 際	2 件	0 件	1 件	3 件
合 計	2 件	0 件	5 件	7 件

⑵ 特許等出願件数

合計　　0 件 （うち国内 0 件、国外 0 件）

⑶ 受賞等

1 件 （うち国内 1 件、国外 0 件）

1. 第9回言語処理学会年次大会最優秀発表賞（「常識的」推論規則のコーパスからの自動抽出，鳥澤 健太郎，第9回言語処理学会年次大会，2003）

(4) 主な原著発表論文による発表の内訳

* 発表者氏名：,「発表題目」,文献名,巻(号),頁,（掲載年）の順

国内誌

該当なし

国外誌

1. Kentaro Torisawa, Kenji Nishida, Yusuke Miyao and Jun’ichi Tsujii, CFG Filtering and Parsing Strategies,in Collaborative Language Engineering (Stephen Oepen et al., eds), CSLI Publications, pp. 81-104, (2002)
2. Takaki Makino, Yusuke Miyao, Kentaro Torisawa, and Jun’ichi Tsujii, Native-code Compilation of Feature Structures, in Collaborative Language Engineering (Stephen Oepen et al., eds), CSLI Publications, pp. 19-47, (2002)

(5) 主要雑誌への研究成果発表

国外誌として掲載した文献はスタンフォード大学で発行している lecture note であるため、該当なし

教師なし学習による対話エージェントの構築