8M-2 格文法解析による対話文からの知識収集

斎藤友伸　松居辰則　岡本敏雄
電気通信大学大学院 情報システム学研究科

1．はじめに

質問・応答システムの研究はTREC８やAAAIにおいても重要な問題として位置づけられている[1]。対話型の質問・応答システムを実装する際には、質問の対象となるドメインの知識を収集する必要がある。従来、このような知識の収集はおもに人手によって行われてきたが、ドメイン領域の拡大や、人的コストの面などを考えると人手による知識の収集には限界があるといえる。近年では、文章が盛んに電子化され、また音声認識技術の向上などもあり、まとまった量の情報を自然言語コーパスの形式で得る機会も多くなってきている。そのようなコーパスから知識を自動的に収集することが出来れば、より大きな質問・応答システムへの応用も考えられる。本研究では、自然言語で記述された対話文を入力とし、文中に含まれる知識を自動的に収集するシステムの作成を行う。

2．システムの概要

システムの概要を図１に示す。本システムは、自然言語で記述された対話コーパスを入力とし、知識を意味ネットワークの形式で出力する。意味ネットワークは、単語を接点とし、枝のラベルには深層格（意味ラベル）が付与されるという形式をとる（図２）。

入力されたコーパスはまず、jumanおよびKNPにより構文が解析される。構文解析の結果に構文解釈および事実検索の操作をほどこし、意味ネットワークの接点と枝を出力する（表層格ネットワーク）。さらに、係り受けの集計と深層格推定の結果から、上記ネットワークの枝に意味ラベルを付与したものが最終的な出力となる。

3．構文解釈と事実検索

構文解析の結果から、単語と単語の修飾関係を収集する。この関係は意味ネットワークにおける枝に相当する。

また、文中に含まれる事実を収集する、事実検索を行う。事実とは、たとえば「参加費は五万円です」という

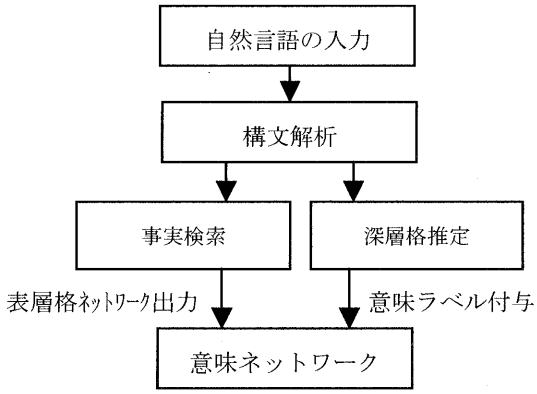

図１：システム概要

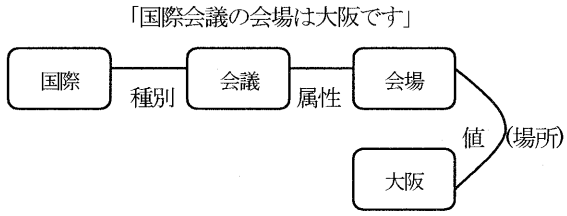

図２：意味ネットワーク

文の「参加費」と「五万円」の関係のように、「AハBダ」と表記できる語A、Bの組で表される。対話文中では事実はさまざまな形式で表現される。「AはBです」のような、直接の係り受け関係がある単純な例もあるが、複雑になると「A１の方と、それからA２に関しましては、Bになります」など、形式名詞「方」や、事実を伝える慣用表現の「～ハ～ニナリマス」などによる間接的な表現の結果、直接の係り受けがない場合もある。このような場合は、システムが持つ文型の辞書を利用し、事実を示す語の組を収集する。

4．係り受け収集と深層格推定

構文解析の結果から、単語間の係り受け、すなわち(係り語,係り方,受け語)の３つ組要素を収集する。たとえば、「林檎を食べる」という文からは(林檎,格助詞「を」,食

Gathering Knowledge from Dialog Sentence using Case Structure Analysis
Tomonobu Saito, Tatsunori Matsui, Toshio Okamoto
Graduate School of Information Systems, The University of Electro-Communications
1-5-1 Chofugaoka Chofu, Tokyo 182-8585, Japan.

べる)という係り受けが得られる。それら個々の係り受けに対して深層格を１つ割り当てる。深層格の推定は(係り語,係り方,受け語)の組と深層格の対応が記述された教師データ（辞書）の情報をもとに行う。ここで問題となるのは、上記のような辞書情報をコーパス中に存在する全ての係り受けについて用意するのは、事実上不可能だということである。この問題は一般的には「素の問題」と呼ばれており、本研究では、係り受け間の距離を定義することにより素の問題を解決する方法を提案する。

４．１ 係り受け間の距離

例えば、「京都へ行く」「大阪へ行く」という係り受けがあった時、我々はこれら２文の用法が類似していると主観的に感じることができる。ここでは、用法の類似性を客観的な観点から定義する。すなわち、係り語同士の類似度、受け語同士の類似度を定義し、その値から２組の係り受け間の距離を(1)式を用いて計算する。

$$d((k_1,u_1),(k_2,u_2))=\sqrt{d(k_1,k_2)^2+d(u_1,u_2)^2}\ldots(1)$$

ただし、$d(w_1,w_2)=\dfrac{1}{r(w_1,w_2)}-1$

また、$r(w_1,w_2)$は単語 w_1,w_2 の類似度、$d(w_1,w_2)$は単語 w_1,w_2の距離、$d((k_1,u_2),(k_1,u_2))$は係り受け $(k_1,u_2),(k_1,u_2)$ の距離である。

単語同士の類似度は、シソーラスから得られる６桁の意味ラベルの一致桁数から０～１尺度で算出する。シソーラス上に存在しない単語に関しては(2)式を用いて係り受け共起回数から類似度を算出する[2]。

$$r(x,y)=\frac{\sum_z freq(x,z)\,freq(y,z)}{\sqrt{\sum_z freq(x,z)^2}\sqrt{\sum_z freq(y,z)^2}}\ldots(2)$$

ただし、式中の $freq(x,y)$は単語 x が y に係った回数を示している。

４．２ 深層格の推定

深層格の推定は、コーパス中に存在する係り受けを、係り受け間距離によりクラスタリングすることにより行う。クラスタリングには貪欲なアルゴリズム(greedy algoritm)を用い、クラスタ A,B 間の距離の算出には式(３)を用いる。

$$d(A,B)=avg\{d(a,b)\mid a\in A,b\in B\}\ldots(3)$$

ただし、式中の avg は引数内の要素に対して平均をとる関数、$d(a,b)$はクラスタに含まれる係り受け a,b 間の距離を示す。

教師データと同じクラスタに分類された係り受けには、教師データと同じ深層格が付与される。深層格の不整合を避けながらクラスタリングを進めてゆき、最終的に同一クラスタに存在する係り受けには、同一の深層格が割り当てられる。このようにして全ての係り受けに深層格を割り当てる。

５．実験

国際会議案内タスクを扱った対話文を入力として、意味ネットを出力する実験を試験的に行った。図３は入力文「参加登録料の方ですけれども、一般参加者の方は八万五千円、あと、国公立大学、研究所、そちらの関係の方は五万円というふうになっております」に対する解析結果の意味ネットワークを示している。図上の<情報>ラベルは、特に単語が事実関係にあることを示している。図３より、参加登録量が八万五千円または五万円であること、八万五千/円という関係は金額を示すこと、などが解析されていることがわかる。

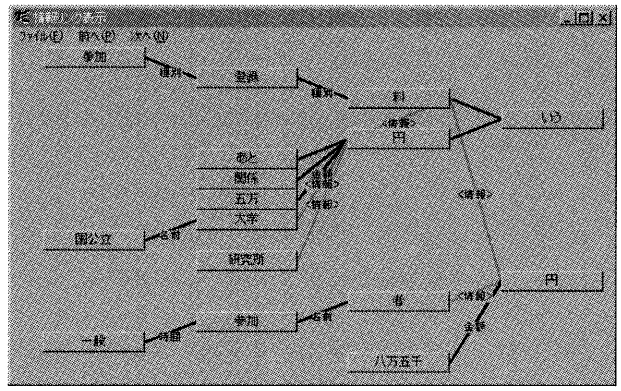

図３：解析結果

６．まとめ

現状では比較的妥当な結果を得られていると考えられる。今後はより多くの対話文を解析し、対話文中の知識に対する出力された意味ネットワークの再現率、精度などを求め、本手法の有効性を検証する予定である。

参考文献

[1]村田真樹,内山将夫,井佐原均, “類似度に基づく推論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 135-24 (2000)

[2]北研二,中村哲,永田昌明 共著, “音声言語処理” ,森北出版,1996

[3]持橋大地,松本祐治, “連想としての意味” ,情報処理学会研究報告 自然言語処理 134-21 (1999)