3H-1

Information extraction method suitable for helpdesk supporting system

池田 直哉† 藤原 祥隆‡ 吉田 秀樹‡

北見工大 SVBL† 北見工大情報システム工学科‡

緒言

ヘルプデスクは質問者と回答者の対話を通じて問題解決を支援するシステムである。しかし、問題点として労働集約型で労働力のリソースを大量に必要とするという課題がある。特に、労働力を確保することが難しい場合には、十分なヘルプデスクをおくことは困難となる。我々は文部省の支援の下に平成 16 年度より地域活用型の教育支援プロジェクトを推進している。本プロジェクトは文部省の「現代的教育ニーズ取組支援プログラム」の一環として「IT による地域活性化教育支援システム」の事業名称を持ち、地域の技術力向上と、本学学生の本プロジェクトへの積極的なかかわりを通した人間力育成を目的としている[1]。コンテンツサービスの普及にはヘルプデスクの充実が課題となるが、リソースの制約により理想とは遠いのが現状である。

本研究はヘルプデスクの機械支援を実現することでコンテンツサービスの普及に寄与することを目的としている。ヘルプデスクの機会支援を実現するためには、入力された質問から情報の検索に必要な情報を適切に抽出する必要がある。本研究の取り扱う教育支援システムのヘルプデスクにおいては相手がわかる、理解したというところが最終的なゴールであるので意味的な情報とともに相手の感情を抽出し状況に応じた臨機応変な応答が可能な問題解決システムの基盤の確立を図る。入力信号としては音声を想定しているので、情報としては言語的なものと非言語的[2]なものに大きく分けることができる。このうち、非言語的なものの取り扱いとしては時変サンプリング・間引き圧縮・木構造格納を特徴とする音声分析技術[3]を用いることを想定している。本発表では言語的な情報について考察する。

先行研究

言語的な情報の処理は従来から、自然言語処理[3]という形で多くの研究が蓄積されている。意味的な情報の理解、特に格フレームの構築については青山学院大学の原田研究室が開発した SAGE などが発表され、システムとしては本研究の目的と類似のものは、京都大学の河原研究室、マイクロソフトの共同研究によるダイアログナビ[4]などがある。ダイアログナビはマイクロソフトのサポート技術情報をバックエンドの事例ベースとして音声認識により技術情報の検索を可能としたシステムである。特徴としては、対話カードという情報を用いて聞き返すことで、不明瞭な質問を明確化する方法を持っていることである。また、いわゆるシソーラスの一種である上位・下位の概念などの辞書を持つことで単語そのものの曖昧性を軽減する能力を持っている。SAGE は形態素解析として JUMAN や茶筅[5]、構文解析として KNP[6]や南瓜[7]といった解析器を使用し格フレームを導出する意味解析器である。現在、知られている範囲では完成度の高い意味解析器であると考えられ多くの研究で利用されている。SAGE は構文解析器が出力した係り受け情報や EDR 辞書の情報から中心語の語意や深層格を探索し適切な解釈を行っている。

意味的情報に対して、主として感性工学の立場から感性的なイメージを抽出する研究も多くなされている。また、主として情報検索の立場から非熟練者により有効な情報の検索システムを提供するために感性検索技術といった提案も行われている。多くの場合にはイメージを表す言葉によって画像や楽曲、料理のレシピを検索すする例が知られている。

言語情報処理

本システムでは知識の内部表現を Triple で行う計画である。Triple はセマンティックWeb[8]]でも用いられる概念で情報を主語、述語、目的語の三つ組(Triple)で表現する。例えば、「私のマシンはデル製である。」という先ほどの文章はFig. 1のようなグラフで表現される。これを N-Triples 形式で表現すると

```
〈#MyHardware〉〈#Build〉〈#Dell〉.
```

となる。

Triple を格フレームの表現と比較すると、格フレームが動詞を中心とした表現であるのに対し、Triple は名詞をからスタートした表現となっている。このような違いはあるが格フレームで情報としては存在する主語を起点に情報を組み立て直せば Triple の表現は得られるのではないかと考える。

課題としては、先ほどの例の場合のようにデル製という文法論に従えば補語に当たる句と #Build なる述語を結びつけるために一種の辞書

が必要なことである。また、類義語も多数存在するので計算機処理可能なシソーラスも同様に必要となる。英語の場合には WordNet[9]のようなウェブオントロジーが存在するが日本語の場合は WordNet と日本語の訳語を結びつけるような試み[8]はあるものの日本語と英語の根本的な違いもあり難しさもある。日本語を処理する場合には EDR 辞書のような辞書を使う必要があると考える。

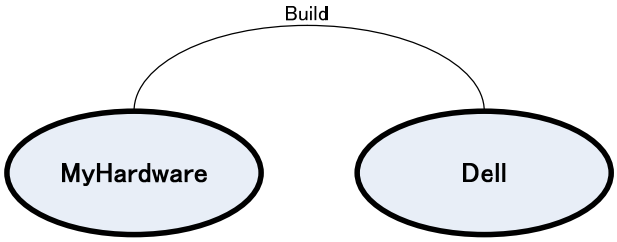

Fig. 1 RDF graph example

課題はあるものの本研究においては、格フレームを構築した上で Triple の表現で情報を取り扱う計画である。理由としては、格フレームは従来より自然言語処理の研究の上で蓄積がされていることが挙げられる。また、Triple を用いる理由は、セマンティック Web という一つの枠組みの中で技術開発が急速に進展していることが挙げられる。

言語情報は意味的な情報と感性的な情報を含んでいると考えるが、意味的な情報ではない感性的な情報を言語情報から抽出するには通常の辞書で行うのは難しいと考えられる。言語情報から感性的な情報を得るための手法として我々はカウンセラーが用いるテクニック[10]に着目している。言語情報から感性的な情報を取り出すことで、音声の非言語情報と組み合わせて、より高度な情報の検索が行えるようになると考える。

感性情報を言語情報から取り出す方法としては、中野[11]で見られるように感性語に着目してそのイメージを取り出すといった研究がなされている。この場合、課題としてはイメージは文章の種類に依存すると考えられることからどのようなサンプルを集めるかに依存すると考える。本研究の場合には、教育システムでの利用が目的なので教育システムの利用者のアンケートにサンプルを求める必要があると考える。

結語

意味的な情報と感性的な情報を組み合わせることで、より良質な情報の検索が行えると考える。また、過去の同様の事例を事例ベースとして用いることで実用性を高めることも重要であると考える。

参考文献

[1] Fujiwara, Okada, Suzuki, Ohnishi, Yoshida “Self-Adaptive Java Production System and Its Application to a Learning Assistance System”, IEICE TRANS. INF. & SYST, VOL. E87-D, No.9 September 2004

[2] 後藤 “非言語情報を活用した音声インタフェイス”,,<http://staff.aist.go.jp/m.goto/PROJ/speechinterface-j.html>

[3] 吉田, Xie, 藤原 “音響構造モデルの提案とスペクトル推定成分との対応”, FIT2004 415-417, 2004

[4] 清田, 高橋, 木戸 "大規模テキスト知識ベースに基づく自動質問応答 -ダイアログナビ-“

[5] 松本，北内，山下，平野，松田, 高岡, 浅原 "日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書",2000,<http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1-j.pdf>

[6] 黒橋, "日本語構文解析システム KNP", http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html

[7] Kudoh, Matatsumoto "Japanese Dependency Analysis Based on Support Vector Machines", 2000, EMNLP/VLC 2000

[8] 神崎 "セマンティック・ウェブのための RDF/OWL 入門", 2005, 森北出版

[9] Fellbaum “WordNet An Electronic Lexical Database”, MIT Press, 1998

[10] 宗像 "SAT カウンセリング技法", 1997, 広英社

[11] 中野 “感性語を用いた単語イメージの多次元表記手法”, < http://www.nlp.its.hiroshima-cu.ac.jp/thesis/h13g_nakano.html>