1N-3

半体系化 Web 情報源からの体系化知識ベースの構築

吉田知訓[1]　山本啓司[2]　栗田恵理[2]　高野敦子[3]　北村泰彦[2]

関西学院大学理工学研究科[1]　関西学院大学理工学部[2]　兵庫大学経済情報学部[3]

1.　はじめに

近頃 Web2.0 という言葉が注目されている．ブログや Wiki に代表されるように，これまで情報の受け手であるユーザーが情報を発信する立場になってきた．また，ユーザー同士がコラボレーションすることによって，より有益な情報が得られることがある．このようなものに「教えて goo!」や「yahoo!知恵袋」といった質問応答サイトがある．これらはユーザー同士が質問や回答をすることによって情報の蓄積を行うもので，情報源として有用である．しかし，断片的で未整理な情報であり，自然言語で書かれているため機械的な処理がしにくい．このような半体系的な情報源を，情報の属性や値が整理されている体系化知識ベースとして構築することができれば，情報の再利用が容易になる．　例えば，「大阪でおいしいお店があれば教えてください」という質問に対して「梅田に○○というお店があるよ」や「難波の△△というイタリアンのいいお店があるよ」というような回答のやりとりがあるとする．ここから，<レストラン名：○○，場所：大阪，梅田>，<レストラン名：△△，場所：大阪，心斎橋，料理ジャンル：イタリアン>のような形で情報を体系化することができれば，情報を再利用しやすくなり利便性が増す．

本研究では半体系化 Web 情報源である「教えて goo!」のレストランについての質問応答文を対象として，自然言語処理，オントロジ，様々なヒューリスティック[1]を組み合わせることによって体系化された知識ベースの構築をすることを目的とする．

2.　体系化知識ベースの構築

図 1 に示すように知識抽出と統合の二つの操作を行い，半体系化 Web 情報源から体系化知識ベースを構築する．まず半体系化 Web 情報源から断片知識の抽出を行う．断片知識とは，属性名と属性値対の集合である．例えば「三田でおいしいお店ありませんか？」という質問文と「フラワータウンのグルメシティというお店がいいですよ」という応答文から「店名：グルメシティ，地域：三田，フラワータウン」という知識を抽出する．次に，得られた断片知識を統合し，体系化知識ベースの構築を行う．　これらの知識抽出，統合の各処理については下記で詳しく述べる．

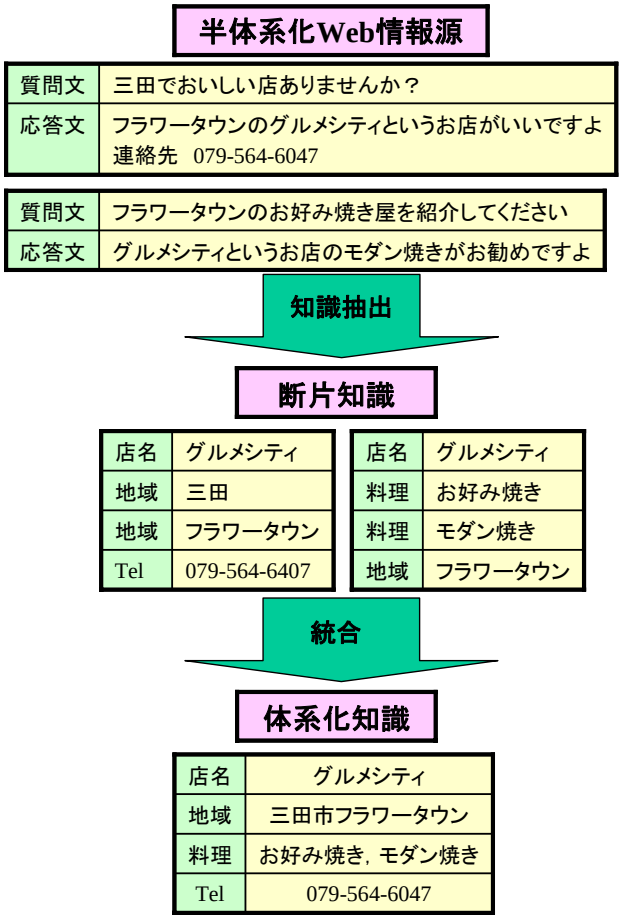

図 1　体系化知識ベースの構築

2.1　知識抽出

図 2 の手順で知識抽出を行う．まず，質問応答文を，辞書を用いて形態素解析を行い，用語を分類する．辞書に含まれている単語であれば既知語とし，辞書に含まれていない単語であれば未知語とする．辞書には，店名，地名，料理名，ランドマーク名などが登録されている．既知語であればそのまま断片知識として抽出し，未知語であれば未知語分類を行う．未知語分類ではヒューリスティックを用いて用語の分類を行う．レストラン名を分類するヒューリスティックの例を次に示す．レストランを紹介する場合，「○○というお店はどうですか」「△△と

いうお店があります」という形式で紹介されることが多い．ここから，形態素解析の結果が「X（未知語）＋という(助詞)＋お店(名詞)」という並びの場合，その未知語 X はレストラン名と判断することができる．他にも「Y(地域を表す名詞)＋の(助詞)＋X(未知語)」や「Z(ランドマークを表す名詞)＋に(助詞)＋ある(動詞)＋X(未知語)」といった並びのときに X をレストラン名と認識するヒューリスティックを用いる．このようなヒューリスティックを複数組み合わせる事によって未知語の認識率を上げていく．知識抽出の例を挙げる．「フラワータウンのグルメシティというお店がいいですよ」という応答文は形態素解析の結果，地名として「フラワータウン」が得られ，未知語として「グルメシティ」が得られる．先ほどのヒューリスティックによりこの未知語はレストラン名であるということが分かり，この応答文から〈レストラン名：グルメシティ〉〈地域：フラワータウン〉という断片知識が得られる．未知語分類の結果は，辞書に反映させることで，次回からの形態素解析に役立てることができる．

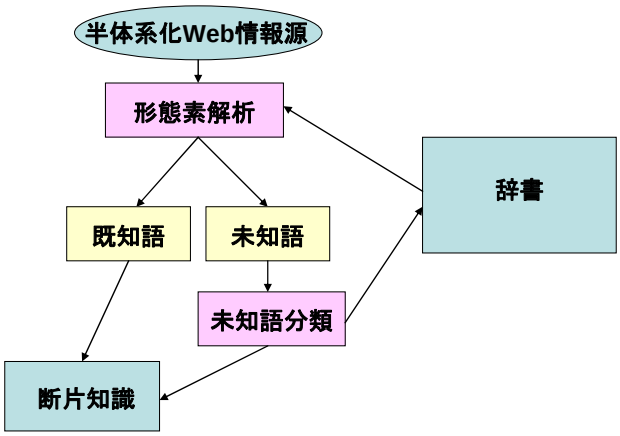

図 2　知識抽出の手順

2.2　統合

上記の方法で抽出された知識は断片的で重複する情報も多く見られる．そこで，固有情報による統合，同義語の統合，上位下位関係の関係付けの 3 つを行い，情報を統合する．

・固有情報による統合

固有情報が同じものがあれば統合をする．固有情報とは，断片知識を識別するのに必要な情報で，レストラン情報においては，電話番号やレストラン名と地域の組といったレストランを特定できる情報のことである．このとき重複するデータは一つにまとめる．図 1 の断片知識はこの操作で

〈店名：グルメシティ〉
〈地域：三田〉
〈地域：フラワータウン〉
〈料理：お好み焼き〉
〈料理：モダン焼き〉
〈Tel：079-564-6407〉

とまとめることができる．

・同義語の統合

同じ意味で使われている語句を，同義語辞書や分類語彙表を用い一つにまとめる．分類語彙表とは，品詞ごとに分かれた類語辞典である．例えば，「お好み焼き」と「お好み」は同じ意味で用いられているので一つにまとめる．

・上位下位関係の関係付け

質問文と応答文の関係や地名辞書・分類語彙表を用いることにより，上位下位関係を関係付ける．例えば，「フラワータウンのお好み焼き屋を紹介してください」という質問文に対して「グルメシティというお店のモダン焼きがお勧めですよ」という応答文があるとする．質問文にある料理は応答文に含まれる料理名の上位概念であるこというヒューリスティックを用いて「お好み焼き」が上位概念であり，「モダン焼き」が下位概念であるとし，「モダン焼き」を「お好み焼き」のサブクラスとして関係付ける．また，地名辞書の都道府県と市町村の上位下位関係を記した情報を用いて，上記の〈地域：三田〉と〈地域：フラワータウン〉は〈地域：三田市フラワータウン〉とまとめることができる．

3．まとめと今後の課題

本論文では，半体系化 Web 情報源から断片知識抽出と統合を行うことによって，体系化知識ベースを構築する手法について提案した．今後はシステムの評価を行う予定である．知識抽出においては，どのヒューリスティックが有効か，また，知識抽出を行った結果を精度と再現率により評価する．統合では，同値データの統合，同義関係の統合，上位下位関係の関係付けがどの程度正しく行われているかを評価する．また，「教えて goo!」のレストラン情報に特化して開発をしているため，他の分野や他の Web サイトでも同様の結果が得られるように汎用性を持たせる必要がある．

参考文献

[1]木谷強,固有名詞の特定機能を有する形態素解析処理,情報処理学会 研究報告「自然言語処理」NL-90 pp73-80,1992.