

概念ベースを用いた場所に関する連想メカニズムの構築

5M-03 杉本 二郎 渡部 広一 河岡 司

同志社大学大学院 工学研究科

1. はじめに

コンピュータを人間にとってより使いやすいものにするには、人間同士が日常行っているコミュニケーション (意思疎通) の仕組みをモデル化し、コンピュータと人間とのインタフェースにもこれを取り入れることが望まれる。

人間の会話においては、場所、時間、感情、感覚などに関する知識を用いた常識的な判断が必要となる。例えば、刑事が犯人の取調べをしていると聞くと、その場所は「警察署」だと連想する。本稿では、「場所」に関する連想メカニズムの実現にあたり、ある一般名詞が場所に関する語であるか判断し、もし、それが場所に関する語であるなら、そこには何が存在し、何をやる場所なのかを連想する手法を提案する。

2. 場所連想

場所に関する一般名詞を「場所語」と呼ぶ。また、その場所語において、何が存在するかを示した具体名詞を「主体語」と呼び、人間にとって何をやる場所なのかを示した語を「目的語」と呼ぶ (表 1)。

表 1. 用語例

場所語	主体語	目的語
警察署	警官, 婦警, ...	捜査, 取調べ...

2.1 場所語知識ベース

すべての一般名詞について、場所語であるかどうかを知識として持たせ、かつ、すべての場所語に主体語と目的語を記憶させることは困難であり効率が悪い。そこで、よく使われる場所語を「代表語」として場所語知識ベース (以下 KB と記す) に持たせておく (443 語)。また、KB には、シソーラス^[1]のノードから選ばれた代表語を分類した「分類語」が存在し (120 語)、それらにも「主体語」と「目的語」を与える (表 2)。また、分類語と代表語は親子関係にあり、主体語と目的語の継承が可能となっている。

表 2. 知識ベースの例

代表語	主体語	目的語	親分類語
警察署	警官, ...	捜査, ...	司法官庁

分類語	主体語	目的語
司法官庁	公務員	司法

2.2 場所であるかの判断

あらゆる一般名詞が入力されたとき、それが場所語であるかどうかの判断について考える。人であれば、例えば「警察署」は、場所語であると判断できる。逆に、「林檎」と言えば、それは場所語でないと判断できる。このような判断を実現するためには、入力された一般名詞が KB の分類語に属するかどうかの判断が必要である (図 1)。

一般名詞が概念ベース^[2] (以下 CB と記す) とシソーラスには登録されているが、KB に登録されていない場合、これを未知語と呼び、この未知語に関して、KB のどの分類語に属するかどうかを特定するために、シソーラスと CB を用いる。シソーラスとは一般名詞の意味的用法を表す 2700 の意味属性 (ノード) の上位下位関係、全体部分関係が木構造で示されたものであり、約 13 万語 (リーフ) が登録されている。CB とは語 (概念) と意味 (属性) のセットを約 5 万語蓄積している、国語辞書等から自動構築された汎用データである (表 3)。概念間の関連性を定量的に評価する計算手法^[3] (以下、関連度計算と記す) によって関連のある概念を選択できる。

表 3. 概念ベースの例

概念	属性
警察署	役所, 警察官, 安全, 犯罪, ...
林檎	果物, 実る, 丸い, 赤い, ...

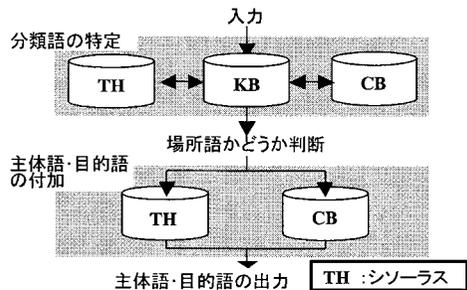


図 1. 処理の流れ

未知語が入力された場合、まず、シソーラスを参照し、未知語が属するノードを特定する。また、シソーラスに存在しない場合は、シソーラスに存在する全てのノードと関連度計算を行い、最も関連性が高いものをその語が属するノードとする。この特定されたノードが分類語の中にあれば、それは場所語であるとする。

2.3 主体語・目的語の付加

入力語が場所語であると判断された場合、次に、その場所語の主体語・目的語の付加を行う(図1)。

入力語が未知語の場合、まず、無条件にその入力語が属する分類語の主体語・目的語を取得する。次に、その分類語に属する代表語と関連度計算を行い、最も関連性が高い語に代替しその代表語の主体語・目的語を入力語の主体語・目的語として更に追加する。

しかし、入力語と代替された代表語とは、あくまで関連性が高い語であって、同義・類義であるとは言えない。そのため、その代表語の主体語・目的語の全てを入力語の主体語・目的語として追加したのでは、不適切な単語が多々含まれる。

そこで、CBに含まれるある概念X(200語)それぞれについて、極めて密接な関係、密な関係、疎な関係にある語の平均関連度の実験を行った結果、それぞれ0.3425, 0.1323, 0.215 という値が得られた。この実験結果を利用して入力語とその代表語の主体語・目的語の関連度が0.1323以上の場合、意味的に関連が強いと判断し、それらの主体語・目的語のみを追加を行い、不適切な主体語・目的語は除外する。

3. 評価

3.1 場所であるかどうかの評価

テストデータとして、一般名詞645個について調査を行った。人の判断と同じようにシステムが正しく判断した名詞(○)、人の判断がそもそも曖昧なためシステムが場所と判断しても、場所でないとしてもよい名詞(△)、それ以外の正しく判断できなかった名詞(×)の3段階に評価した(図2)。

○・△あわせて、89.1%の名詞に対して矛盾のない判断ができた。

また、テストデータの内の未知語(530語)についても同様の調査を行った(図3)。○・△あわせて、86.8%の名詞に対して矛盾のない判断ができた。

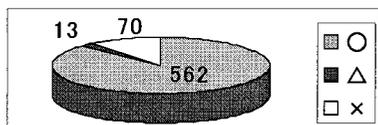


図2. 場所であるかどうかの評価結果

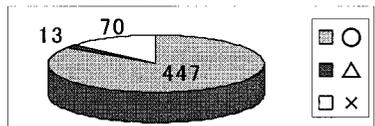


図3. 未知語についての評価結果

3.2 主体語・目的語付加の評価

図2で○であり、かつ、それが場所である一般名詞は258語あり、代表語(115語)と未知語(143語)の主体語・目的語の数と1つの名詞に付加される平均の数を比較した(表4)。

未知語に関して、主体語の平均数は代表語の38.3%であり、目的語に関しては47.1%であった。

表4. 代表語と未知語との比較

種類	主体・目的語	総数(個)	平均数(個)
代表語	主体語	1227	10.7
	目的語	610	5.3
未知語	主体語	591	4.1
	目的語	364	2.5

また、未知語の場所である名詞の主体語・目的語に関して、正しい情報(○)、矛盾しない情報(△)、正しくない情報(×)の3段階にそれぞれ評価した(図4)(図5)。

主体語については、○・△をあわせて、88.5%が矛盾のない情報であり、目的語については、○・△をあわせて、87.4%が矛盾のない情報であった。

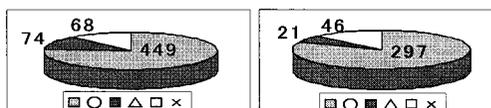


図4. 主体語の評価結果



4. 考察

場所であるかどうかの評価結果の失敗例を見ると、70語のうち、32語はCBにもシソーラスにも存在しない語であり、CBの更なる精錬を行うことでより精度を向上させることができると考えられる。

未知語に関する主体語・目的語の付加については、失敗例は少ないものの、半分以上の主体語・目的語が抜け落ちており、今後はCBから自動的に主体語・目的語を更に取得する手法を検討する必要がある。

5. おわりに

本稿では、独自のKBを用い、CBとシソーラスを利用した場所に関する連想モデルを提案した。これにより、例えば、「警察署」は、警官が取調などをする場所として連想可能となった。

今後、一般名詞だけでなく固有名詞にも範囲を広げ、より知的な場所に関する連想を目指す。

尚、本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行ったものである。

参考文献

- [1]NTT コミュニケーション科学研究所監修、「日本語語彙体系」,(岩波書店, 東京, 1997)
- [2]真鍋康人, 小島一秀, 渡部広一, 河岡司:「概念間の関連度やシソーラスを用いた概念ベースの自動精錬法」,同志社大学理工学研究報告, Vol.42, No.1, pp.9-20, 2001.
- [3]渡部広一, 河岡司:「常識的判断のための概念間の関連度評価モデル」,自然言語処理, Vol.8, No.2, pp.39-54, 2001.