国語辞書からの常識知識の自動獲得法

佐々木智彦† 増子公徳† 野中昌行† 石川勉†
拓殖大学 工学部 情報工学科‡

1 はじめに

我々は、“ 言葉で考えるコンピュータ ”の実現を目指し研究している。このコンピュータでは、不完全な知識を類似知識で補い近似解を導く概略推論法 [1] と、常識知識ベースが核となる。常識知識の獲得については、事象間の因果知識の自動獲得法 [2] 等の研究がなされているが、本報告では国語辞書を利用した事実知識の自動獲得法について提案する。具体的には、辞書には少なくとも見出し語に関する事実知識が記載されており、かつその語義文には定型的な表現パターンが用いられていることに着目した獲得法を示す。さらに、この手法を人に関する知識を対象として評価する。

2 知識の表現形式と獲得の基本的な考え方

2.1 知識表現形式

知識の表現には述語知識を用いる。知識の意味を一意に特定可能とするため、以下のように引数はラベル (以下、`$_i$) 付きとする。

述語 (`$_1$:引数 1, `$_2$:引数 2, …)

なお、述語が動詞の場合には、語義文が否定や時制を伴って表現がされている場合があるため、それらを表すため述語の先頭に識別子を付加する。

2.2 知識獲得の基本的な考え方

語義文は、いくつかの定型的な表現パターンで記述されていることが多い。ここでは、そのうち、最も多用されていると考えられる“【見出し語】～を～する～。”というパターンを対象とする。例えば、“【医者】患者を治療する人。”等である。即ち、この表現パターンは一般的に

【A】B を C する D。

であり、A,B,D は名詞、C は動詞 (サ変名詞を含む) の語幹、また、D は A の上位概念となっている。従って、比較的簡単な構文解析を用いるだけで

C(`$_1$:A, `$_2$:B)

という述語形式に変換可能 (先の例では、“ 治療する (`$_1$:医者, `$_2$:患者) ”) である。

さらに、この知識を概略推論で利用可能にするため、以下のような含意を含んだ論理式 (見出し語が含意の右辺) に変換する。なお、この知識についてはその正当性が保証されないため、利用には注意を要する。

述語 (`$_1$:x,`$_2$:引数 2) →引数 1(`$_1$:x)

述語 (`$_1$:x,`$_2$:y) ∧引数 2(`$_1$:y) →引数 1(`$_1$:x)

3 知識獲得方法

知識獲得の基本的な流れとしては、まず獲得対象の概念に関する見出し語の語義文 (正確には“【見出し語】が ”という主部を補った文) を形態素解析し、多義の解消を行い、EDR 電子化辞書を用いてラベル付けを行う。

3.1 多義の解消

多義のある概念については、獲得の対象となる概念に属するか否かの判定が必要となる。例えば図 1 のように、「操り人形」は“ 芝居 ”と“ 人 ”の二つの意味を持つ。この場合、例えば人の概念についての知識を獲得するとしたときには、同図のように判定する。即ち、見出し語 (操り人形) と見出し語の上位概念 (芝居あるいは人) が人の概念に属しているか否かで判定する。

【操り人形】①糸をつけた人形を操ってさせる芝居。
②他人の命令のままに行動する人。

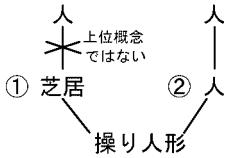

図 1: 多義の例

3.2 ラベル付け

ラベルとしては、フィルモアの格文法 [3] の深層格、具体的には EDR 電子化辞書の日本語動詞共起パターン副辞書 [4](以後、共起辞書) で利用されている概念関係子 (表 1) を用いる。

表 1: 深層格の種類

格	概念関係子	説明
主体	agent	動作を行う主体
対象	object	動作思考の対象
目標	goal	動作の終点
源泉	source	動作の起点
場所	place	事象の成立する場所
様態	manner	行為・現象における様態

ラベル付けは以下のように行う。語義文を形態素解析した結果を、それに含まれる最後部の動詞に関する共起辞書中の記述と照合する。この照合では、まず語義文と共起辞書に書かれている助詞が一致し、かつ引数となる概念が共起辞書の意味情報構成要素 (あるいはその上位概念、以後概念識別子) に属せばラベルを付与する。例えば、先の見出し語“ 医者 ”の例では、“ 治療する ”の共起辞書の内容 (図 2) を用い、まず語 1、語 2 の格助詞の照合を行い、次に“ 医者 ”と“ 患者 ”の概念が概念識別子に該当するか否かを判定する。

Acquiring Common Knowledge from Machine Readable Dictionary

†Tomohiko Sasaki, Masanori Mashiko, Masayuki Nonaka, Tsutomu Ishikawa
‡Department of Computer Science, Takushoku University

	概念関係子	意味情報構成要素
＜語 1 ＞が	agent	30f6b0
＜語 2 ＞を	object	30f6b0;30f60f
治療する		

図 2: “ 治療する ”の格情報

3.3 例外的表現への対応

前述の基本的な表現パターンのみを対象とすると、適合する見出し語が限定されるため、いくつかの変形に対処することとした。主な変形とそれへの対処を以下に示す。

- B の部分が名詞句の場合
  名詞句を構成する形容詞または名詞を“・”で連結する。
  例)【髪結】他人の髪を結う人。
  ⇒ 結う (agent:髪結, object:他人・髪)
- 並列助詞を含む文
  “ ～や～ ”、“ ～・～ ”など名詞が並列している場合、複数の独立した知識として作成する。
  例)【医員】病院・診療所などにつとめる医師。
  ⇒ つとめる (agent:医員, object:病院)
  つとめる (agent:医員, object:診療所)
- B の部分が指示語の場合
  その直前の文の最後部の名詞 (並列助詞で結ばれた名詞を含む) を B と仮定する。
  例)【小作農】小作による農作。それを営む人。
  ⇒ 営む (agent:小作農, object:農作)
- 動詞直前に副詞がある場合
  共起辞書には必須格のみが記述されており、副詞は任意格とされている。しかし、下記の例のように「餓鬼」に対して副詞“ がつがつ ”は必要不可欠である。そこで、動詞直前に副詞がある場合にはその副詞は重要な意味を持つと考え、ラベル“ manner ”を付与する。
  例)【餓鬼】食べ物をがつがつと食べる人。
  ⇒ 食べる (agent:餓鬼, object:食べ物, manner:がつがつ)

4 実験

上述の手法を用いて、学研国語大辞典 [5] 中の主要語に対し人に関する概念の知識獲得を行った。主要語としては 4 万語概念ベース [6] の概念を選んだ。人に関する見出し語は 3,031 個あり、多義を含めた語義文数は 5,229 個であった。そのうち、ここで対象とした表現パターンを持った語義文が 3,998 個 (約 76%) あった。さらに多義の解消を行うと、1,743 個が人に関する語義文となった。

この 1,743 個に対して、3.2 節で述べたラベル付けを行った。その結果を図 3 に示す。横軸の段数は 3.2 節で述べた上位概念の段数である。すなわち、0 は概念識別子そのものに含まれていた場合であり、1 はそれの 1 段上の概念あるいはその下の概念に含まれていた場合である。以下、2 以降も同様である。また、知識獲得率は語義文 (1,743 個) に対する獲得知識数の割合、総正解率は総獲得知識数に対する正解の割合、段正解率はその段で新たに獲得された知識の正解率である。

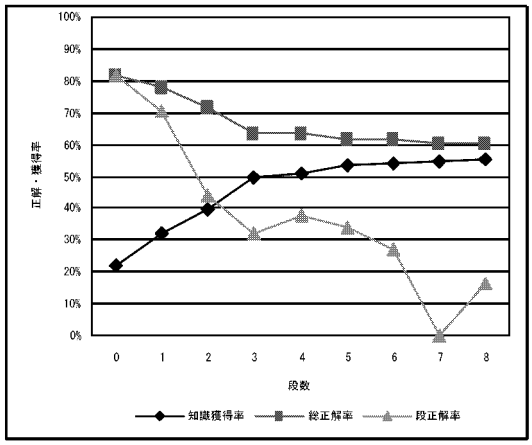

図 3: 範囲拡大時の獲得数と正解率

同図より、段正解率が 2 段で激減することがわかる。したがって、1 段上位の概念識別子を用いるが妥当であると考えられる。この場合の獲得知識数は 563 個であり、そのときの正解率は約 70%であった。なお、獲得された知識の例を以下に示す。

【園児】幼稚園などに通っている子供。
通う (agent:園児, goal:幼稚園)
通う (agent:x, goal:幼稚園) →園児 (inst:x)
通う (agent:x, goal:y) ∧幼稚園 (inst:y)
→園児 (inst:x)

なお、総獲得率は 8 段上位の概念識別子を対象としても 50%程度となっているが、この主な原因は、EDR 辞書における格パターンや概念識別子の不足、あるいは A から D の概念が単語辞書に登録されていない等による。また、知識が誤って獲得された主な原因は、文型が複雑なため引数部を正しく特定できなかったこと等による。

5 まとめ

本論文では、国語辞書の語義文に関する常識知識を、EDR 電子化辞書を利用して一階述語論理形式で自動獲得する手法を提案した。この手法を、国語辞書の人に関する見出し語 3,031 個に適用した結果、563 個の事実知識を獲得でき (正解率は約 70%)、その有効性を確認した。

参考文献

[1] Nguyen Viet Ha, 石川勉, 阿部明典:知識の類似性を利用した概略推論法, 電子情報通信学会論文誌 D I, Vol.J84-D-1, No. 4, pp.389-400(2001).
[2] 佐藤浩史, 笠原要, 松澤和光:表層的因果知識ベースによる事象推移予測方式, 情報処理学会第 56 回全国大会, Vol. 2, pp. 251-252(1998).
[3] 黒橋禎夫他:自然言語処理, 岩波講座ソフトウェア科学 15,pp200-230(1996).
[4] 日本電子化辞書研究所:EDR 電子化辞書, 日本語動詞共起パターン副辞書,(1996).
[5] 金田一春彦, 池田弥三郎:学研国語大辞典第二版, 学習研究社 (1988).
[6] Nguyen Viet Ha, 帆苅譲, 石川勉, 笠原要:単語の意味の類似性判別のための大規模概念ベース, 情報処理学会論文誌 Vol. 43, No. 10, pp. 3127-3136(2002).