EDR電子化辞書を利用した自然言語文から述語知識への自動変換法

4M−02

野中　昌行　グエン ベト ハー　石川　勉
拓殖大学工学部情報工学科

1　はじめに

我々は、不完全な知識を常識知識で補い、近似解を導く概略推論法 [1] の研究を行っている。この推論法では、大量の常識知識が必要となり、これを電子化文書から自動的に獲得することを想定している。

本論文では、自然言語文を自動的に一階述語論理に変換する手法を提案する。また、論理式の解釈が一意になるように、引数に8種類の格情報を持ったラベルを付ける手法を提案する。

2　知識の表現

知識の表現としては、概略推論で利用することを前提に、基本的には述語論理を用いることとする。しかし、従来の述語論理では、一つの論理式に複数の解釈が存在する。例えば、

歩く（私, 駅, 学校）

から、以下の二つの解釈が考えられる。

① 私は駅から学校まで歩く

② 私は学校から駅まで歩く

このため、異なった解釈で作成された知識を用いれば誤った結論を導く可能性がある。また、自然言語文から論理式を生成する場合には、何らかの順番で引数を配置する必要がある。これらを考慮して、本論文では自然言語文を一意に表現し、論理式を利用する人が異なっても解釈が一つに定まるよう、ラベルを付けることとした。例えば、主体を表すラベルを“Agent”とし、源泉を表すラベルを“Source”、目標を表すラベルを“Goal”とした場合、

歩く（Agent:私,Source:駅,Goal:学校）

は、上記の①の解釈のみを持つことになる。

ラベルとしては、基本的に語と語の間の意味関係を動詞中心に捉えたフィルモアの格文法 [2] の深層格を用いる。このための情報としては、日本語の主要動詞について、動詞の格に関連する情報を約13,000レコード記述した、EDR電子化辞書の日本語動詞共起パターン副辞書 [3](以後、共起辞書) があり、本論文では、この共起辞書を利用する。この辞書には各動詞の各概念について、表層格に対応する深層格の種類と、その深層格を取る場合の格助詞、意味情報構成要素 (深層格を満たす概念の範囲を示した概念識別子群) が記述されている。図1にEDR共起辞書の一部を示す。

	概念識別子	意味情報構成要素
<語1>		
が	agent	30f6b0:30f6bf
<語2>		
を	object	30f6e5:3f9639
食べる		

図 1:“食べる”の格情報

共起辞書の深層格のうち、表1に示す8種類を用いる。基本的には、これらの深層格を用いればある程度一般的な知識を表現できると考えた。

表 1: 深層格の種類

格	概念関係子	説明
主	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
材料	Material	材料
道具	Implement	動作における手段道具
条件	Condition	条件

3　述語知識への変換

3.1　変換の流れ

自然言語文を述語知識へ変換する手順を図2に示す。この方法では、文の動詞を述語部、名詞又は名詞句を引数とし、まず動詞の格情報を共起辞書から取り出す。取り出した格情報を利用して引数に深層格 (以後、ラベル) を付ける。次に、名詞がラベルの引数となるかEDR単語辞書 [4](以後単語辞書) を使いチェックをする。全ての名詞がラベルの引数となった場合、述語知識に変換する。

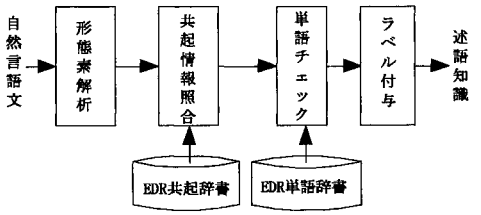

図 2: 述語知識変換の流れ

A Method to Automaticlly Transform Natural Language Sentences to Predicate Knowledge using EDR Caseframes
Masayuki Nonaka,Nguyen Viet Ha ,Tsutomu Ishikawa
Department of Computer Science , Takushoku University

3.2 ラベルの付与

ラベルは、EDR 共起辞書の格パターンと形態素解析の結果とのパターン照合で決定する。

具体的には、動詞の全ての格情報を取り出す。さらに、入力された文の格助詞を全て含む格情報を特定する。格助詞が一致しても、ラベルが候補の単語を取らないことがある。そこで、このチェックを行い、入力された単語の概念識別子が、意味情報構成要素の下位概念と一致したとき、ラベルの引数とする。全ての格助詞についてこの処理を行う。格情報が複数ある場合、全てのラベルが候補の単語を引数とした格情報を使う。

しかし、共起辞書は人手で作成されている為、不完全な部分がある。例えば、ラベルが候補の単語を取り得る場合でも、概念識別子が意味情報構成要素と一致しない場合がある。そこでこの取りこぼしを少なくする為に、以下のルールを付けた。

ルール 1)

候補の単語の概念識別子が、意味情報構成要素と同じ上位概念を持つならば引数とする。また、特定の格助詞でラベルが候補の単語を取らないとき、以下のルールを適用する。

ルール 2)

i) 格助詞が“を”→ ラベルを“Object”とするただし、例外として動詞が“自動詞”かつ“移動動詞”、“Place”に変更する
ii) 格助詞が“が”→ ラベルを“Agent”とする
iii) 格助詞が“から”→ ラベルを“Source”とする

4 実験

4.1 対象としたデータ

以下の全ての条件を満たす文を変換対象として実験を行った。具体的にはこのような文を、ライトハウス英和辞典 [5] の例文から 100 文、共起辞書から 200 文無作為に取り出した。

- 単文
- 動詞を一つ含む文
- 文意が取れる文

4.2 実験と結果

4.1 で取り出した文を提案した手法により、述語知識に変換した。

英和辞典の例文と共起辞書の例文での結果を、それぞれ表 2 と表 3 に示す。表中の①は、全ての引数の概念識別子が、意味情報構成要素の下位概念と一致した場合である。②は、①に該当しなかった文に対して、ルール 1 を適用したものである。③は②にも該当しなかった文に対して、ルール 2 を適用したものである。

なお、このとき、(i) 共起辞書に動詞が登録されて無い、(ii) 単語辞書に単語が無い、(iii) 形態素解析が正確に行われなかったのような文を評価対象外とし、英和辞典と共起辞書の例文からそれぞれ 22 文と 28 文を削除した。

表 2: 英和辞典の例文の結果

	正解文の数 (%)	不正解文の数 (%)
① EDR	42(54)	10(13)
② ルール 1	4(5)	2(3)
③ ルール 2	16(20)	4(5)
合計	62(79)	16(21)

表 3: 共起辞書の例文の結果

	正解文の数 (%)	不正解文の数 (%)
① EDR	109(63)	19(11)
②ルール 1	19(11)	5(3)
③ルール 2	17(10)	3(2)
合計	145(84)	27(16)

これらの表からわかるように、約 80 %のかなり良い割合で正しく変換できた。

実際に変換した例を図 3 に示す。

自然言語文		述語知識
石が窓に命中した	→	命中した(Object:石, Goal:窓)
彼が私の提案に同意した	→	同意した(Agent:彼, Object:私・提案)
彼女が花を摘む	→	摘む(Agent:彼女, Object:花)

図 3: 実際の変換例

なお、共起辞書の例文を用い場合、本来①の段階で 100 %となるべきだが、そうなっていないのは、前述した EDR 共起辞書は人手で作成されたためゆらぎがあるためと考えられる。

5 まとめ

本論文では、EDR 共起辞書を利用して自然言語文を、引数に 8 種類の格情報を持った一階述語論理に自動的に変換する手法を提案した。また、英和辞典と EDR の例文を利用した実験により、約 80 %の正解率が得られ、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] 黒石禎夫他:自然言語処理:岩波講座ソフトウェア化学 15:pp200-230 (1996).

[3] 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.

[4] 日本電子化辞書研究所:EDR 電子化辞書, 日本語単語辞書,1996.

[5] 竹林滋, 小島義郎:ライトハウス英和辞典:研究社 (1993).