型理論的レキシコンに基づいたオントロジー抽出に向けて：例化談話の場合

緒方 典裕

1 序

自然言語は（形式言語同様）形式的体系において意味記述が可能である、という「Montague の提唱」[4] 以来、特に型理論に基づいた自然言語の意味モデルの研究がさかんになっている。[4] のモデルは、e(実体)、t(真理値)、s(インデックス = 可能世界 $\times$ 時間) を原始型とし、型構成子を $\rightarrow$ とする単純型付き λ 計算であり、「Montague 文法」と呼ばれている。多くの自然言語の形式意味論といわれている体系 ([3]) がこの体系かその拡張、多少の変更によるものである。さらに、Ahn[1] は Barendregt[2] の Pure Type System (PTT) 上の λP において、談話を含めた情報の形式化を提案し、Ranta[6, 5, 7] は Martin-Löf の ITT(直観主義型理論) をもちいて自然言語の意味論を提案している。これらの高階依存型理論 (Higher-order Dependent Type Theory) に基づいた自然言語の意味論では、人、車などの基本型が導入でき、「原始型 =（自然言語の）オントロジー」、「文脈 = 談話」等が実現されている。本稿では、Montague の一つの立場であった「自然言語 = 形式言語」という立場の一つの具現として「テキスト = 型理論的プログラム」を考察する。この立場では、レキシコンから談話構造までが型理論によって定義される。その応用例として、「例化談話」(exemplification discourse) を取り上げ、「オントロジー情報抽出 = 型推論」という図式を提案する。

2 例化談話

次の例はある株式ニュースの一パラグラフで、例化談話の典型例である。

(1) Shares of drug companies gave back their gains Wednesday.
Dow component Merck fell 1-1/2 to 114,
Pfizer finished down 1-13/16 at 79 and
Warner-Lambert lost 1-5/16 to 136-15/16.

(1) が示すように、例化談話は、第一行目は一般的な命題が表現され、その次にその例となる特定の命題がいくつか続くという形式をとる。第一行目の文を「一般文」、それに続く、例を表す文を「例文」と呼ぶことにする。この談話を論理形式で表現すると次のようになる。

$\forall x \in \mathtt{shares_of_drug_company.gave_back_their_gains}(x)$
$\mathtt{fell_to}(\mathtt{Dow_Component_Merch}, 1-1/2, 114)$
$\mathtt{finished_down_at}(\mathtt{Pfizer}, 1-13/16, 79)$
$\mathtt{lost_to}(\mathtt{Warner-Lambert}, 1-5/16, 136-15/16)$

ここで、ある種の帰納法 $\frac{\varphi(0)\ldots\varphi(n)\ldots}{\forall x \in \omega.\varphi(x)}$ との類推と、例化された各述語は一般命題の述語を含意する (上の例では fell_to, finished_down_at, lost_to が gave_back_their_gains を含意する) ことから、例化談話はさらに次のように抽象化できる。

$\forall x \in T.\varphi(x)$	
$\varphi_1(a_1)$	$a_1 \in T \ \forall x.\varphi_1(x) \rightarrow \varphi(x)$
$\ldots$	$\ldots$
$\varphi_n(a_n)$	$a_n \in T \ \forall x.\varphi_n(x) \rightarrow \varphi(x)$

ただし、右半分のボックスの情報は例化談話においては明確に表示されない、暗黙に前提される情報である。また下の左右の部分ボックスから演繹できる情報は $\varphi(a_1)\ldots\varphi(a_n)$ であるので、上の帰納法の前提に当たるものが導出できる。逆にいえば、この暗黙の情報は、各オブジェクトがある型を持つこと (例では Dow_Component_Merch, Pfizer, Warner-Lambert が shares_of_drug_company であること) と、各イベント型があるイベント型の部分型であること (例では fell_to, finished_down_at, lost_to が gave_back_their_gains の部分型であること) といった、特定領域で前提となっているオントロジー情報を表しており、これを形式化すれば、例化談話からそのようなオントロジー情報が導出できるといえる。

3　例化談話の型理論による形式化

前節で議論した例化談話の論理構造は、高階依存型理論の「formulas as types」、「proofs as objects」の概念を利用して、(2) のように型理論で仕様を与えることができる。ただし、「formulas as types、proofs as objects」とは、論理式を型とみなし、その型が割り当てられている λ 項をその論理式を導出する証明とみなすことである。例えば、論理式 $\forall x \in T.\varphi x$ は $\Pi x : T.\varphi x$ という型に当たり、ある λ 項 t に対して、$t : (\Pi x : T.\varphi x)$ が成り立つことがあり、この場合、t は、式 $\forall x \in T.\varphi x$ の証明を表していることになる。ただし、この型自体は $prop$(命題) という型をもつので、$(\Pi x : T.\varphi x) : prop$ も同じに成り立つ。

$$(2)\ \ \lambda T : type.\lambda\varphi : T \to prop.\overline{\lambda y : T.\lambda\psi : sub\varphi}.EXEM_n : \boxed{\begin{array}{c}\Pi x : T.\varphi x \\ \psi_1(y_1) \\ \cdots \\ \psi_n(y_n)\end{array}}$$

ただし、$\overline{\lambda y : T.\lambda\psi : sub\varphi} = \lambda y_1 : T.\lambda\psi_1 : sub\varphi \ldots \lambda y_n : T.\lambda\psi_n : sub\varphi$,
$sub\varphi = \{\varphi_i | \Sigma y : T.\Pi x : T.\varphi_i x \to \varphi x\}$,
$\boxed{\begin{array}{c}\varphi_1 \\ \varphi_2 \\ \cdots \\ \varphi_n\end{array}} \equiv (\Sigma x_1 : \varphi_1.\Sigma x_2 : \varphi_2.\ldots.\varphi_n)$ であるとする (つまり、$\varphi_1 \wedge \ldots \wedge \varphi_n$)。

(2) は例化談話の文法情報を利用してオントロジー抽出に関連した次の型推論を行う：

1. 例化談話の一般文の主語が S ならば、$S : type$ という情報を得る。
2. 1 の情報から一般文の述語 P に対して、$P : S \to prop$ という情報を得る。
3. 1 の情報から各例文の主語 S_i に対して、$S_i : S$ という情報を得る。
4. 2 の情報から各例文の述語 P_i に対して、$P_i : sub\ P$ という情報を得る。
5. 2 と 4 の情報から各例文の述語 P_i に対して、$P_i : S \to prop$, $\Pi z : T.P_i z \to Pz$ という情報を得る。

つまり、例化談話の論理形式は (2) に次のように談話の各項目を適用することで得られる。

(3) (((...((((2) 一般文主語) 一般文述語) 例文主語$_1$) 例文述語$_1$...) 例文主語$_n$) 例文述語$_n$)

そして、その過程で、領域特定オントロジーに関する情報が抽出される。

4　レキシコンの形式仕様の抽出に向けて

例化談話からは、少なくとも述語に関しては、単に型 (オントロジー) に関する情報だけでなく、近似的な語彙の意味内容も抽出可能である。直観的には、一般文のイベント型は、各例文のイベント型の上位型になっているからである。例えば、次のような一般文と例文の述語の型は同じであるが、意味内容が逆であるような例化談話はあり得ない。

(4) Shares of drug companies gave back their gains Wednesday.
Dow component Merck fell 1-1/2 to 114,
Pfizer finished down 1-13/16 at 79 and
Warner-Lambert climbed 1-5/16 to 136-15/16.

(4) の内容を表示するには次のように but などの接続詞を導入する必要がある。

(5) Shares of drug companies gave back their gains Wednesday.
Dow component Merck fell 1-1/2 to 114,
Pfizer finished down 1-13/16 at 79 but
Warner-Lambert climbed 1-5/16 to 136-15/16.

例化談話では、このような接続表現、副詞的表現が無い限り、一般文と例文の意味内容もおおよそ同じである。したがって、一般文の述語の意味内容がわかっていれば、例文の述語の意味内容つまりイベント型も見当をつけることができる。(また、(5) のような例化談話の拡張でも、述語の型は同じであり、また、述語の意味内容も逆であるという情報が得られる。)

このような、述語の意味内容の形式仕様の抽出は、(2) を利用して、次のような過程とみなすことができる。

1. 一般文の主語 S と述語 P から、各例文 $(S_i\ P_i)$ の述語 P_i の型は $sub\ P$ となり、$P_i : S \to prop$, $\Pi x : P_i x \to P$ が得られる。
2. P に対して $P \equiv Q$ が与えられていれば、1 から $\Pi x : P_i x \to Q$ が得られる。
3. さらに、例文 $(S_i\ P_i)$ と例文 $(S_{i+1}\ P_{i+1})$ が逆接接続表現で連結されていれば、$\Pi x : P_{i+1} x \to \neg Q$ を得る。

4. これらの結果から、例文の述語の意味内容の近似を得る、つまり、$P_i \approx Q$ もしくは $P_{i+1} \approx \neg Q$ とする。

例えば、gave back their gains を次のように定義していたとする。

(6) Definition gave_back_their_gains(x:drug company,e:event)
≡ value(x,pre(t(e))) > value(x,post(t(e))).

ただし、value:drug company×time→real, t:event→time, pre(t:time) = t - 10 min, post(t:time) = t + 10 min であるとする。すると例化談話 (5) から次のような定義が得られることになる。[1]

(7) a. Definition fell(x:drug company,e:event)
≡ value(x,pre(t(e))) > value(x,post(t(e))).

b. Definition finished_down(x:drug company,e:event)
≡ value(x,pre(t(e))) > value(x,post(t(e))).

c. Definition climbed(x:drug company,e:event)
≡ value(x,pre(t(e))) < value(x,post(t(e))).

5 実装に向けて

実装には次の二つの手法が考えられる。

- ML, Coq, ALF, Twelf, λ-Prolog などの型理論的仕様や型推論を内在するプログラミング言語において、直接に仕様を与え、談話をプログラムとして扱う。
- いくつかの中間表示を経由して、上記の型理論的プログラミング言語のプログラムに変換する。

後者の一つの方法としては、(3) で示したように、一般文の主語・述語、各例文の主語・述語が抽出できればよいので、その情報を反映した、次のような例化談話のタグ付けを考えることができる。

[1] 厳密には、(7c) では $\leq$ の式が得られるので、$\forall x \forall e.value(x, pre(t(e))) \neq value(x, post(t(e)))$ のような一般知識が必要である。

(8)
```
<P TYPE=EXEM NUM=i> <!--paragaph i is an exemplification-->
  <SEN TYPE=GEN NUM=i.1> <!--i.1 is the general sentence of i-->
    <SUBJ TYPE=GEN NUM=i.1.1> <!--i.1.1 is the subject of i.1-->
      Shares of drug companies
    </SUBJ>
    <PRED TYPE=GEN NUM=i.1.2> <!--i.1.2 is the predicate of
i.1-->
      gave back their gains
    </PRED>
    <ADV NUM=i.1.3> <!--i.1.3 is an adverb of i.1-->
      Wednesday
    </ADV>.
  </SEN>
  <SEN TYPE=EX NUM=i.2> <!--i.2 is an examplified sentence of
i-->
    <SUBJ TYPE=EX NUM=i.2.1>
      Dow component Merck
    </SUBJ>
    <PRED TYPE=EX NUM=i.2.2>
      fell 1-1/2 to 114
    </PRED>,
  </SEN>
...
</P>
```

参考文献

[1] René Ahn. Communicating contexts: a pragmatic approach to information exchange. In Peter Dybjer, Bengt Nordström, and Jan Smith, editors, *Types for Proofs and Programs*, pp. 1–13. Springer, Berlin, 1995.

[2] H. P. Barendregt. Lambda calculi with types. In S. Ambramsky et al., editor, *Handbook of Logic in computer Science*, pp. 118–309. Clarendon Press, Oxford, 1992.

[3] David R. Dowty. *Word Meaning and Montague Grammar*. Reidel Publishing Company, Dordrecht, 1979.

[4] Richard Montague. *Formal Philosophy: Selected Papers of Richard Montague (ed. by Richmond Thomason)*. Yale University Press, New Haven, 1974.

[5] Aarne Ranta. *Type-Theoretical Grammar*. Oxford University Press, Oxford, 1994.

[6] Aarne Ranta. Type theory and the informal language of mathematics. In Henk Barendregt and Tobias Nipkow, editors, *Types for Proofs and Programs*, pp. 352–365. Springer, Berlin, 1994.

[7] Aarne Ranta. Syntactic categories in the language of mathematics. In Peter Bybjer, Bengt Nordstöm, and Jan Smith, editors, *Types for Proofs and Programs*, pp. 162–182. Springer, Berlin, 1995.