Dynamic Semantics Meets XML and IE

緒方 典裕

大阪大学 言語文化部
ogata@lang.osaka-u.ac.jp
http://lc301.lang.osaka-u.ac.jp

1 序：IE = ダイナミック・セマンティクス+XML+XSLT プログラミング

ダイナミック・セマティクスとは言語表現の意味を情報状態の推移関係とみなす形式意味論で、Hans Kamp [6] や Irene Heim [5] によって基本的なアイデアが出され、Jon Barwise [1]、Groenendijk & Stokhof [4] らによってほぼ原型が完成された。(詳しくは [11,8] を参照のこと。) 一階論理のダイナミック・セマンティクス (たとえば *DPL*[4]) では、この情報状態は変数への値割り当てとみなされる。Heim のファイル変化意味論 (File Change Semantics) では、変数という ID 毎に設けられたカードのファイルを更新するという比喩が用いられる。たとえば、文 (1) は図 (1) のようなファイル更新プロセスという意味を与えられる。[1]

(1) John loves a girl. He is a teacher. The girl is an artist. She is Mary. But she hates him.

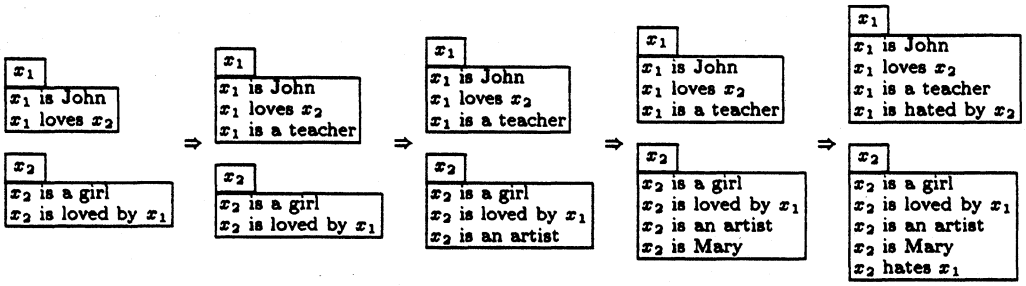

Fig. 1. 文 (1) のファイル変更過程

図 1 中の変数 x_1, x_2 は多くのダイナミック・セマンティクス (特に *DRT* [7]) では談話指示対象 (discourse referent) もしくはディスコース・エンティティといわれる。このファイル更新過程はディスコース・エンティティの観点からいえば、各カードの情報を収集する活動ともみなすことができる。

一方、IE(情報抽出 [3,9,10]) ではネイムド・エンティティ・タスク、シナリオ・テンプレイト・タスク、コレファレンス・タスク、推論等の各課題をこなして、例えば文 (1) から図 2 のようなテンプレイトを得る。ディスコース・エンティティ

[1] Heim のファイル変化意味論では、GB 理論の LF という表示から派生するために厳密には図 (1) とは違う。

EVENTTYPE	loving
PERSON1	John
PERSON2	Mary

Fig. 2. 文 (1) から抽出されたテンプレイト

は、イベントや状況なども含むことができるので、まさにIEとダイナミック・セマンティクスは同じタスクを違う観点から具現化したものであるとも考えられる。

また、入力文からの各カードの構成過程は、図3のような入力文の解析木の部分木の操作ともみなすことができる。XML 自体が木構造であるため、このよう

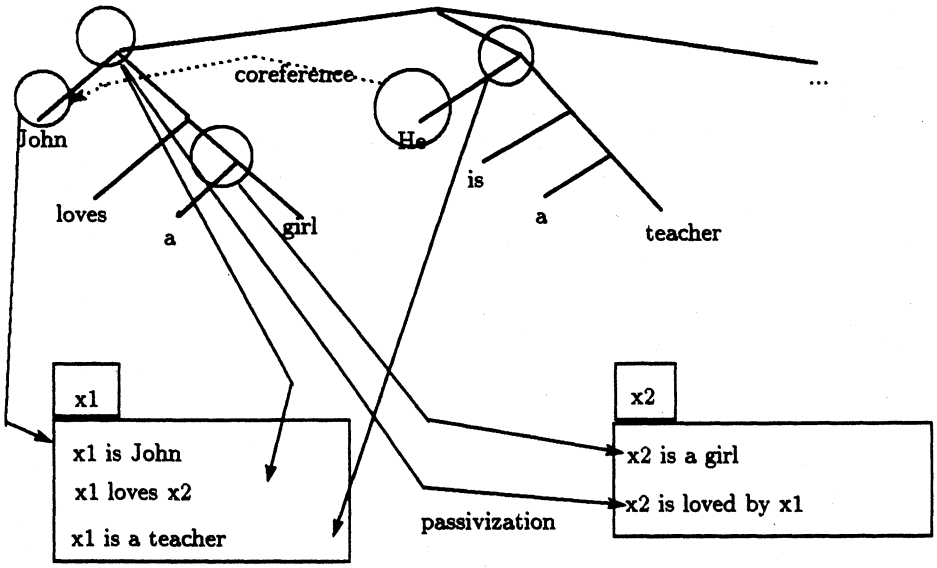

Fig. 3. 文 (1) の部分木操作としてのカード構成

な木構造の取り扱いはXML 関連分野でも議論されており、XSL Transformation [2](XSLT) という規格で扱われている。XML で自然言語の解析木・コレファレンス用のマークカップ言語を定義するのは自然であり、ここにもダイナミック・セマンティクスと XML の密接な関係を見出すことができる。

本発表では、ダイナミック・セマンティクスをIEの観点からとらえなおし、それを XML 関連技術により定式化・実装可能であることを示す。

2 基本的文書構造の扱い

まず、浅い構文解析とコレファレンス処理がなされた文書のためのマーカップ言語 SNLPML(Shallow Natural Language Processing Markup Language) を次のように定義する。

```
<! DOCTYPE snlpml [
<! ELEMENT document (date, title, source, s+)>   <! ELEMENT date (day, time)>
<! ELEMENT source (organization, author)>
...
<! ELEMENT s (np, vp, period)>              <! ELEMENT np (det, adj*, cn, pp*)>
<! ELEMENT vp (tv, adv*, np, adv*)>         <! ATTLIST s id (#PCDATA) #REQUIRED>
<! ATTLIST np dref (#PCDATA) #IMPLIED
              coref (#PCDATA) #IMPLIED
...
]>
<! ATTLIST vp dref (#PCDATA) #IMPLIED      coref (#PCDATA) #IMPLIED
              icoref (#PCDATA) #IMPLIED    roletype (#PCDATA) #IMPLIED
              tense (past|pres) #REQUIRED
]>
<! ATTLIST adv anch (xpath) #IMPLIED       roletype (#PCDATA) #IMPLIED>
]>
<! ATTLIST tv nform (#PCDATA) #IMPLIED     pp (#PCDATA) #IMPLIED>
]>
```

重要な点は、(i) 文書全体あらわすものがトップノード (document) であること、(ii) その下には文書の作成時のデータ (ここでは date) と各文のノード (s) があること、(iii) 名詞句のノード (np) と動詞句のノード (vp) に属性としてディスコース・エンティティ (dref)、参照先 (coref) があること、(iv) 動詞のノード (iv, tv) の属性には不定形 (nform) や過去分詞形 (pp) があること、(v) 副詞句のノード (adv) の属性には指標詞のアンカー先の属性 (anch) があること、である。

3 ファイルの扱い

抽出されるべきファイルもまたマーカップ言語 FML(File Markup Language) として次のように定義される。

```
<! DOCTYPE fml [
<! ELEMENT file (card+)>        <! ELEMENT card (condition+)>
<! ATTLIST card token (#PCDATA) #REQUIRED>
<! ATTLIST condition type (#PCDATA) #REQUIRED>
]>
```

card の属性 token はある種の ID 番号をあらわし、文書中ではディスコース・エンティティを、知識ベース中ではその中での ID をあらわす。condition の属性 type は attribute-value 構造の attribute に該当し、condition でマーカップされたテキスト部分が value に該当する。

4 ファイル構成アルゴリズム

ファイルを構成するアルゴリズムは、基本的に n までの各 dref の値に関する XSLT のテンプレートの再帰的適用である。普通 for ループや while ループを使うが XSLT にはない。またパラメータの値更新もできない。そのため、パラメータの間接的な更新とテンプレートの再帰呼び出しにより間接的に実装できる。

5 指標詞の扱い

指標詞の処理は属性 anch によって行われる。例えば、時間指標詞の場合、anch は XPath によって/document/date/day などのように指定される。

6 固有名の扱い

固有名は知識ベース内で参照できるか否かによって処理する。これは、知識ベースもまたFMLで記述されているので、次のようなXPathを用いたXSLTのテンプレートによる「テスト」としてプログラムできる。

```
<xsl:when test="//np[@dref=$i]/pn/text()
          =document('kb.xml')//condition[@type='name']/text()">
...
</xsl:when>
```

ただし、kb.xmlは知識ベースをあらわす。もしテストが成功すれば、being Xというリンク付きのテキストを出力し、失敗すればnamed Xというリンクなしのテキストを出力する。その際、XLinkやXPointerを使うことが望ましい。

7 述語の扱い

述語は属性nformやppを利用して適宜、能動態や受動態に変形する。

References

1. Jon Barwise. Noun phrases, generalized quantifiers and anaphora. In Peter Gärdenfors, editor, *Generalized Quantifiers: linguistic and logical approaches*, pages 1–30. D. Reidel Publishing Company, Dordrecht, 1987.
2. W3C Consortium. *XSL Transformations (XSLT) Version 1.0.* http://www.w3.org/TR/xslt, 1999.
3. Ralph Grishman. Information extraction: Techniques and challenges. In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to and Emerging Information Technology*, pages 10–27. Springer Verlag, Berlin, 1997.
4. Jeroen Groenendijk and Martin Stokhof. Dynamic montague grammar. In L. Kalman and L. Polos, editors, *Papers from the Second Symposium of Logic and Language*. Akadémiai Kaidó, Budapest, 1990.
5. Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, 1982. published by Garland Publishing, New York, 1989.
6. Hans Kamp. Events, instants and temporal reference. In R. Bäuerle et al., editors, *Semantics from Different Points of View*, pages 376–417. de Gruyter, Berlin, 1979.
7. Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht, 1993.
8. R. Muskens, J. van Benthem, and A. Visser. Dynamics. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 587–648. Elsevier Science B.V., Amsterdam, 1997.
9. Maria Teresa Pazienza. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Springer Verlag, Berlin, 1997.
10. Maria Teresa Pazienza. *Information Extraction: Towards Scalable, Adaptable Systems*. Springer Verlag, Berlin, 1999.
11. Johan van Benthem. *Exploring Logical Dynamics*. CLSI Publications, Stanford, 1996.