GDA タグ集合の設計と応用

橋田 浩一
電子技術総合研究所

長尾 確
ソニーコンピュータサイエンス研究所

内山 将夫
信州大学

Christoph J. Neumann
東京工業大学

高橋 直人
電子技術総合研究所

1 はじめに

大域文書修飾 (Global Document Annotation; GDA) は、文書の意味構造 (語用論的側面を含む) を自動認識可能にするための、言語学的なタグ集合を策定し普及させることにより、自然言語処理技術の大規模な応用と統合、および研究用データの確保を目指すプロジェクトである。GDA タグによって構造化された文書は、翻訳、検索、情報抽出などによるさまざまな加工および提示が可能であるという意味で、**多用途の知的コンテンツ** (versatile intelligent contents) (橋田, 1998) である。現在、GDA タグ集合[1]の設計を進めながら、その実用性を評価しつつ、大量のデータの GDA タグによる構造化と GDA タグを用いた要約やプレゼンテーションなどの技術を開発しつつある。

GDA タグ集合は XML[2]のインスタンスなので用途に合わせてさまざまにカスタマイズできるが、その前提となる基本仕様は次のような方針の下に設計している。

- 広義の意味構造の明示を目的とする。
- 形態論的な分類の粒度をなるべく粗くする。
- 統語論に関するタギングの複雑さを抑制する。
- さまざまな詳細度のタギングを許容する。

以下では、これらの特徴と GDA の応用技術について述べる。

2 意味構造

GDA タグ集合は、語用論的な構造を含む広義の意味構造を計算機によって自動認識可能にすることを目的としている。意味構造を明示することにより、自然言語処理や人工知能のさまざまな (特に提示系の) 技術が高精度で適用可能になることが期待されるからである。GDA で考えている広義の意味構造は、以下の諸側面を含む。

- 主題役割 (2)
- 修辞関係 (2)
- 共参照 (2)
- 時制 (1)
- 相 (1)
- 対話機能
  - 働き掛け (1)
  - 応答 (2)
- 語義
- 様相演算子や量化子の作用域

働き掛け (forward-looking communicative function) は主張、命令、約束など、応答 (backward-looking communicative function) は了解、回答、受諾などである。こうした対話機能の分類は、DRI (discourse resource initiative)[3] (Carletta et al., 1997)、人工知能学会 SLUD 研究会の談話タグワーキンググループ (市川他, 1998)、DAMSL (Dialog Act Markup in Several Layers) (Allen & Core, 1996; Jurafsky et al., 1997) などを参考に設計しつつある。

このように意味構造の多くの側面に関するタグの仕様をなるべく単純化するため、一見異なるように思われる諸側面を統一的に捉えるように努めている。たとえば上記の意味構造の諸側面のうち、最後の作用域以外はすべて広い意味での語義に含めることができる。特に、(1) は 1 項述語であるような語義、(2) は 2 項関係であるような語義と考えられる。そこで、2 項関係であるような語義 (の識別子) を**関係項** (relational term) と呼び、関係項を値とする属性 **rel** によって上記の (2) の付いた側面を統一的にタギングすることにしている。GDA において語義を表わす属性は **sem** であり、関係項はその値にもなりうる。**sem** は (それが付随する) エレメント (開始タグから終了タグまでのテキスト) の意味クラスを示し、**rel** はエレメントと係り先との関係を示す。たとえば、下の 2 つの例はいずれも「健が」が係り先に対して **agt** (agent; 動作主) の関係に立つことを示し、2 番目の例はさらに **agt** の意味を持つのが「が」であることを示す。

[1] http://www.etl.go.jp/etl/nl/gda/tagman.html
[2] http://www.w3c.org/XML/
[3] http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html

- <adp rel="agt"> 健が </adp>
- <adp> 健 <ad sem="agt"> が </ad></adp>

また、関係項は **rel** および **sem** 属性の値になれるだけでなく、それ自身が属性名になることもできる。そのような属性を**関係属性** (relational attribute) と言う。関係属性の値は他のエレメントの **id** 属性である。エレメント A が関係属性 r によってエレメント B を指しているとき、これは、A が意味 (指示) する対象と B が意味する対象との間には関係 r が成立することを示す。これによってゼロ照応を含む共参照を明示することができる。たとえば下の例は、「健が健の母親を健の家に連れて行った」という解釈を示す。

```
<su>
  <adp><np id="K"> 健 </np> が </adp>
  <adp><np  eq="K"> 自分 </np> の母親を
</adp>
  <adp><np pos="K"> 家 </np> に </adp>
  連れて行った。
</su>
```

ここで **eq** は等値 (equality)、**pos** は所有者 (posessor) を表わす関係項である。

主題役割と修辞関係はいずれも基本的には意味的な二項関係であり、たとえば **cnc** (concession; 逆接) など、主題役割 (たとえば「失敗にも関わらず～」) でも修辞関係 (たとえば「失敗した。しかし～」) でもあると考えられる二項関係も多いので、これらは区別せずに扱っている。また、応答は発語内行為であり、厳密には意味的関係ではなく語用論的な関係だが、これを **rel** 属性の値としている。それでも、**rel="stt tim"** のように **rel** の値として複数の関係項を書くことができ、また関係項は属性名にもなるので、同じ 2 つのエレメントの組の間に意味的・語用論的な関係が複数通り成立する場合にもそれを明示することができる。**sbj** (主語) や **obj** (目的語) などの文法機能も関係項としているが、これはタギング作業者の判断を簡単にするためである。タギング作業者にとって文法機能の方が主題役割よりもわかりやすいことが多く、適当な辞書があれば、各単語 (特に動詞) について文法機能から対応する主題役割を求めることができる。

3 形態論

GDA では、形態論に関するタギングはかなり大雑把であり、EAGLES[4]や CES[5]で提案されているよりもはるかに分類が粗い。これは、GDA タグ集合が意味構造の明示を目的としているからである。言語間の差違はほとんど形態論と統語論に関するものであり、とりわけ形態論的な違いが大きい。これに対し、意味構造の表示に必要なタグや属性は言語に依存する度合が小さいと考えられる。したがって、形態論と統語論のタギングを大雑把にすることにより、多くの言語を同一のタグ集合によって扱うことを意図している。

GDA タグ集合では品詞はタグ名によって表わされる。そのようなタグ名は基本的に、**<n>** (名詞)、**<v>** (動詞と助動詞と終助詞)、**<aj>** (形容詞)、**<ad>** (後置詞、前置詞、副詞、連体詞、接続詞、補文化辞など)、**<ij>** (感動詞) しかない。接頭辞、接尾辞、接辞 (clitic) などは **<n>** や **<v>** によってタギングする。たとえば「っぽい」という形容詞化接尾辞は **<aj>** エレメントとする。フランス語の接辞 *en* は **<adp>** エレメントとする (後述のように **<adp>** は **¡ad>** の最大投射である)。

これに対し、エレメントの意味的な分類は形態論的・統語論的分類よりも詳細である。特に名詞は、**<date>** (日付)、**<time>** (時刻)、**<persname>** (人名)、**<num>** (数値) などに細分類してある。こうした分類は TEI[6]から取り入れたものである。

4 統語論

形態論と同じく統語構造のタギングも意味構造の明示に必要な最少限にとどめている。たとえば、通常の空所 (gap) と寄生空所 (parasitic gap) は区別しない。しかし、依存構造や等位構造を明示することは意味構造の明示に有効なので、GDA では、依存 (dependency)、並列 (parallel; または等位 coordination)、同格 (apposition)、修正 (repair) という 4 種に統語構造を分類し、各エレメントの子エレメントの間に成り立つ統語的関係を **syn** という属性の値によって示す (ただし **syn** 属性は文内の統語構造だけでなく文間の構造を示すのにも用いる)。

最も頻繁に現われる統語構造は依存構造だから、特に依存構造のタギングが簡単になるように工夫している。その簡単化は、**syn** (synthesis) 属性と**句タグ** (phrasal tag) による。**syn** 属性の値で依存関係に相当するものは 4 通りあり、そのうち日本語で主に用いるのは **f** (forward dependency; 前向き依存関係) と **fc** (forward chain; 前向き連鎖) である。これらはいずれも、子エレメントがそれぞれ原則として前方 (右側) にある他の子エレメントに係ることを意味する。また、句タグとは、**<p>** 以外のタグでタグ名が p で終わるものである。句タグを持つエレメントを**句エレメント** (phrasal element) と言い、それ以外の文内のエレメントを**主辞エレメント** (head element) と言う。句エレメントは最大投射を表わし、主辞にならない (つまり何も受けない)。日本語の場合、**syn** のデフォトル値は **f** なので、たとえば下の例は「健が」と「学校に」が「行っ」または「た」に係ることを示す (いずれに係るかは特定されない)。

```
<su>
```

[4]http://www.ilc.pi.cnr.it/EAGLES/home.html
[5]http://www.cs.vassar.edu/CES/
[6]http://etext.virginia.edu/TEI.html

```
<adp> 健が </adp>
<adp> 学校に </adp>
行った。
</su>
```

syn="fc"は、 **syn="f"**の意味に加え、エレメントの中味に関して以下のことを意味する。

- プレインテキストの部分は、形態素を表わす主辞エレメントの列と見なす。それ以外のエレメントは明示されているものだけとする。
- 各エレメントは、なるべく近く (可能なら前方) に係る。

こうして、 **syn="fc"** を用いることによりタギングを大幅に簡略化できる。たとえば、下の例は「健が」と「学校に」が「行っ」に係ることを示す。

```
<su syn="fc">
  <adp> 健が </adp>
  <adp> 学校に </adp>
  行った。
</su>
```

また、「検討を始めたばかりのころは」の形態素への分割が「検討 + を + 始め + た + ばかり + の + ころ + は」という一通りに決まることを前提すれば、

```
<adp syn="fc"> 検討を始めたばかりのころは </adp>
```

は図 1 と等価である。 **syn="fc"** はさらに、明示されていない **dep**、 **pel**、 **phd**、 **grel** 属性を子エレメントが含まないことを意味する。これらの属性は開始タグと終了タグを越える依存関係を示すものなので、**syn="fc"** を持つエレメントの子エレメントの間の依存関係は唯一に定まる。

このようにエレメントの個数と入れ子の深さを抑制することにより、タグの構造が人間にとってわかりやすくなるので、タギング作業の負荷を軽減し、タギングの精度を向上させることができるだろう。特に日本語や韓国語のように依存関係がほとんど一方向に決まっているような言語の場合には、依存関係のタギングに必要なエレメントの入れ子の深さは中央埋め込み (center embedding) の深さの 2 倍を越えないようにできる。中央埋め込みの深さは高々 4 程度だから、それによってエレメントの入れ子の深さを 8 程度に抑制できるだろう。前後両方向の依存関係をいずれも多く含む言語の場合も各エレメントが前後どちら側に係るかを示す方法があれば同様にエレメントの入れ子の深さを抑制できるが、そのためのタグ集合の複雑化とのトレードオフを考慮する必要がある。

```
<adp>
  <seg>
    <seg>
      <seg>
        <seg>
          <seg>
            <seg>
              <seg> 検討 </seg>
              <seg> を </seg>
            </seg>
            <seg> 始め </seg>
          </seg>
          <seg> た </seg>
        </seg>
        <seg> ばかり </seg>
      </seg>
      <seg> の </seg>
    </seg>
    <seg> ころ </seg>
  </seg>
  <seg> は </seg>
</adp>
```

図 1: 不必要に複雑なタギング

5 詳細さの自由度

GDA ではタギングの詳細度をかなり自由に調整できる。これは、タグや属性の不在が原則として何も意味しないということである。これにより、アノテータが判断できない場合にはタギングを省略することができ、また使用目的に応じたさまざまな詳細度のタギングが可能になる。特に **syn="f"** (または **syn="b"**) であるようなエレメントの中では、依存関係を大雑把に指定することができる。たとえば前掲の例

```
<su>
  <adp> 健が </adp>
  <adp> 学校に </adp>
  行った。
</su>
```

では「健が」と「学校に」が「行った」の中のどこかに係ることは示されるが、「行っ」に係るか「た」に係るかは明示されない。(「健」が「が」に係ることや「学校」が「に」に係ることも明示されない。) ちなみに、このようにすれば、文節係り受け方式の統語解析結果を、曖昧性を保存しつつ GDA タグに自動変換できる。また、下のようにすれば文の構造に関して何も明示しないことになる。

```
<su> 健が学校に行った。 </su>
```

このようなタギングは情報量は少ないが誤りではない。もちろん、構造を詳しく確定することが可能ならば **syn="fc"** によってそうすることが望ましい。

関係項や作用域も、明示しない場合にはデフォルトの解釈を特定しないことにしている。作用域は属性 **sce** (scoping element) によってタギングする。 **sce** はエレメントの指示対象が属する作用域を導入する演算子 (下の例では量化子「どの男も」) を指す。 **top** は談話全体を意味し、 **sce="top"** はそのエレメントの指示対象がどの演算子の作用域にも含まれないということである。下の第 1 の例は男ごとに愛する女が異なるという解釈、第 2 の例はどの男も同一の女を愛するという解釈を表わすが、第 3 の例は **sce** を含まないので、そのいずれであるかを特定しない。

- **<su><adp id="X">** どの男も **</adp>**
  <np sce="X"> ある女 **</np>** を愛する。 **</su>**
- **<su><adp>** どの男も **</adp>**
  <np sce="top"> ある女 **</np>** を愛する。 **</su>**
- **<su><adp>** どの男も **</adp>**
  <np> ある女 **</np>** を愛する。 **</su>**

ただし、タギングによって明示しなくても、「愛する」はそれに係る「どの男も」の作用域に入る。

6 利用

GDA タグに基づく文書データの構造化は、「認知科学」の解説論文などに対して行なわれているほか、RWC のテキストデータベース (Hasida et al., 1998) においても進行中であり、いずれも近い将来に公開する予定である。 RWC のテキストデータベースグループでは、毎日新聞の約 3,000 記事と岩波国語辞典第 5 版に対するタギングが進行中である。いずれも、岩波国語辞典に基づく語義のタギングを含む。今後は、これまで行なってきた要約 (Nagao & Hasida, 1998) やプレゼンテーション (内山・橋田, 1999) に加えて、翻訳や検索や辞書の利用などへの GDA の応用技術を開発する予定である。

従来の情報検索は正確には情報の検索ではなく文書の検索であり、検索要求を精密化するための情報をシステムがユーザに与えることもほとんど不可能だったが、GDA でタギングされた文書においてはキメ細かい構造が利用可能なので、そうした文書の集合からはピンポイントの情報検索が可能であり、検索要求を精密化するための情報をシステムがユーザに与えながらインタラクティブに検索を進めることもできる。たとえば、単語による検索要求に対し、文書の集合の中でその単語と直接の意味的関係を持つ他の単語の集合を提示することができる。ユーザはそれらのうちのいくつかを選択することによって、もとの単語キーワードと新たな単語の間の意味関係を含む文書の集合に検索対象を絞り込む。検索要求のこのような精密化はさらに何段階も続けることができ、インタラクティブな検索が可能となる。ユーザは、検索対象である文書の集合の内容を知らず、したがって、いきなり最適な検索要求を思い付かないことが多いので、こうしたインタクティブな検索が必要である。また、タギングされた辞書の利用に関しても、基礎的な研究課題 (黒橋禎夫・酒井康行, 1999) に加えて、意味からの単語のインタラクティブな検索など、実用的な可能性も多い。

参考文献

Allen, J. & Core, M. (1996). Draft of DAMSL: Dialog Act Markup in Several Layers.. ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz.

Carletta, J., Dahlback, N., Reitinger, N., & Walker, M. A. (1997). Standards for Dialogue COding in Natural Language Processing.. Dagstuhl-Seminar Report: 167. ftp://ftp.cs.uni-sb.de/pub/dagstuhl/reporte/97/9706.ps.gz.

橋田 浩一 (1998). GDA: 意味的修飾に基づく多用途の知的コンテンツ. 『人工知能学会誌』, **13** (4).

Hasida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S., Toyoura, W. K. J., & Takahashi, H. (1998). The RWC Text Databases. *Proceedings of The First International Conference on Language Resource and Evaluation*, pp. 457–461. Granada:

市川 熹・荒木 雅弘・石崎 雅人・板橋 秀一・伊藤 敏彦・柏岡 秀紀・加藤 佳司・熊谷 智子・樽松 明・小磯 花絵・田本 真詞・土屋 俊・中里 収・堀内 靖雄・前川 喜久雄・山下 洋一・吉村 隆 (1998). 談話タグ標準化の現状. 『人工知能学会研究会資料 SIG-SLUD-9703-6 (2/27)』, pp. 41–48.

Jurafsky, D., Schriberg, E., & Biasca, D. (1997). SWBD-DAMSL: Shallow-Discourse-Function Annotation Coders Manual.. http://stripe.colorado.edu/jurafsky/manual.august1.html.

黒橋禎夫 ・酒井康行 (1999). 国語辞典を用いた名詞句「A の B」の意味解析. 『情報処理学会自然言語処理研究会』.

Nagao, K. & Hasida, K. (1998). Automatic Text Summarization Based on the Global Document Annotation. *Proceedings of the 17th International Conference on Computational Linguistics.*

内山 将夫 ・橋田 浩一 (1999). GDA 文書からのスライド生成. 『「言語資源の共有と再利用」シンポジウム』. http://www.etl.go.jp/etl/nl/sympo99/.