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 碧
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 XML2のインスタンスなので用途に合わせてさまざまな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。
- 統語論に関するタグの多様さを抑制する。
- さまざまな詳細度のタグを許容する。

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、以下の側面を含む。

- 主題役割 (2)
- 修辞関係 (2)

http://www.etl.go.jp/etl/nl/gda/tagman.html
http://www.w3c.org/XML/

- 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）（Carletta et al., 1997）、人工知能学会 SLUD 研究会の論文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）なども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のは個々の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（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 において語義を表す属性は sem であり、関係項はその値に値が入る。sem が（それ自身を付随する）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、次の 2 の例はいずれも「誰か」が係り先に対して agent である関係に立つことを示し、2 番目の例はさらに agent の意味を持つのが「が」であることを示す。
<adp rel="agt">健が</adp>
<adp健</adp>
<adpsem="agt">が</adp></adp>

また、関係項は rel および sem 属性の値になるだけでなく、それ自身が属性名になることもできる。そのような属性は関係属性 (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 (指事) する対象と B
が意味する対象の関係 r で成立することを示す。
この意味において最も重要な関係 r が成立することを能
できる。たとえば下の例は、「健が彼の母を彼の家
に連れて行った」という解釈を示す。

ここで eq は等価 (equality), pos は所有者 (possession)
を表す接関係である。
主題役名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 con (concession; 逆接) な
ど、主題役名 (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
あると考えられる二項関係もないので、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
意味関係ではなく語用論的な関係だが、これも rel 属
性の値としている。それそれ、rel="attim"
のよう
に rel の値として複数の関係を表すことができ、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
の関係に意味的・語用論的な関係が複数に成立する
場合にもそれを示すことができる。

タグ名自体が文法的機能を関係項として、それらは
タグ名自体を文法的関係を示すものとする。さらに、
タグ名と文法的関係を示すものとして、それらは
タグ名を文法的関係を示すものとする。さらに、
タグ名（特に動詞）について文法機能から対応する主
題役名を求めることができる。

3 形態論

GDA では、形態論に関するタグ名はかなり大雑把
であり、EAGLES4 や CES5 で提案されているよりも
はるかに分類されている。これには、GDA タグ集が
意味構造の示す行為を目的としているからである。言語間の差違は
ほとんど形態論と形態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
に必要なタグ名が文法に依存する度合が小さいと考え
えられる。したがって、形態論と形態論のタグ名を大
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 タグ集では品詞はタグ名によって変わられ
る。そのようなタグ名は基本的には、<nn>（名詞）、<vr>
（動詞）と助動詞と助動詞）, <ap>（形容詞）, <ad>
（前置詞、助詞、助詞と助詞）, <aj>（動詞）
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と
（形態）と、<ia>（動詞）と、<pa>（動詞）と

これに対し、修辞名の意味的関数は形態論
の形態論的関数よりも詳細である。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 TEI6 から取り入れたものである。

4 統語論

形態論と同じく修辞構造のタグ名も意味構造
の示す必要に最少限とどまっている。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
ことは意味構造の示す行為がそれである。GDA で
は、依存（dependency），並列（parallel；または等位
構造），同格（apposition），修正（repair）とい
う 4 種に修辞構造を明示し、各エレメントの子エレメン
トの間に成立立つ修辞的関係を syn という属性の値に
よって示す（ただし syn 属性は文内修辞構造だけでなく
文間の構造を示すのにも用いる）。

最も頻繁に現われる修辞構造は依存構造だから、
特に依存構造のタグ名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 属性の値は依存に関係
するものは 4 通りであり、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
f（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
グには、pp 以外のタグ名で p で終えるもので
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
最大投射を表し、主辞にならない（つまり何も受けな
い）。日本語の場合、syn がデフォルト値なので、
たとえば下の例は「健が」と「学校が」が「行っ」
または「た」に係ることを示す（いずれに対しても特定され
ない）。

http://www.tnt.ric.cnr.it/EAGLES/home.html
http://www.cs.vassar.edu/CES/
健が
学校に
行った。

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、形態素を表す主辞エレメント以外のものである。
- 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」と「学校に」が「行って」に係ることを示す。

検討を始めたばかりのころは「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、

検討を始めたばかりのころは

は図1と等価である。syn="fc"はさらに、明示されている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し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的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倍を超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。

前後関係の依存関係を示す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。

詳細さの自由度

GDAではターゲットの詳細度をかなり自由に調整できる。これは、タグや属性の不必要性原則として何を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る、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。たとえば前掲の例は

健が
学校に
行った。

健が学校に行った。
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 scc (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」) を指す。top は語話全体を意味し、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 id="X">どの男</su>
  <np sce="X">ある女</np> を愛する。 </su>

- <su>どの男</su>
  <np sce="top">ある女</np> を愛する。 </su>

- <su>どの男</su>
  <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ても、「愛する」はそれに係る「どの男」の作用域に入る。

6 利用

GDA タグに基づく文書データベースの構造化は、「認知科学」の解釈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化ユーザーに与えると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具体的構造が利用可能なので、そうした文書の集合からは一定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化ユーザー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、きまりは検索要求を思い付かないことが多いので、こうしたインタラ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 (黒橋俊夫・満井康行, 1999) に加えて、意味からの情報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


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