大域文書修飾 (GDA) の進捗と展望

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

高橋 直人
電子技術総合研究所

内山 将夫
信州大学工学部

Christoph J. Neumann
東京工業大学

1 はじめに

機械が人間なみに言語の「意味」を理解するという自然言語処理の主要な目標を達成するにはあと1世紀以上かかるだろう。この長期目標に関する基礎研究への投資に根拠を与えるためには、未熟な技術を有用なものにするような新たな応用を見出す必要がある。それらの応用は、同時に、こうした基礎研究の推進に貢献するようなものであることが望ましい。特に、人間の知識の使用に関する巨大かつ良質のデータを集成することが上記の目標の達成に不可欠であるが、それを研究目的のためだけに行なうことは経済的に困難と考えられるので、そうしたデータベースが応用の副産物として自動的に生成されるような枠組を考えたい。

大域文書修飾 (**GDA**) (橋田他, 1997; 橋田, 1997) は、文書の意味的・語用論的な構造を明示するSGML (XML) のタグ集合を策定、公開し、そのタグを含むテキストを入出力するツールや応用プログラムの開発と普及を推進することにより、インターネット等でこのタグ集合を広めることを目指すプロジェクトであり、未熟な技術の実用化と基礎研究用のデータという、上記の二重の目的を持っている。GDAタグは現在の技術で機械が文書の意味的・語用論的な構造を理解することを可能にするので、タグの情報を利用することにより、機械翻訳、情報検索、情報抽出、要約、質問応答、事例に基づく推論、データマイニングなど、自然言語処理や人工知能のさまざまな応用の品質が飛躍的に向上する。したがって、こうした技術が安価に利用できるとすれば、自分の文書にタグ付けして公開することを多くのユーザに動機付けることになり、巨大なタグ付きコーパスがインターネットを中心として自動的に生成・成長するだろう。このコーパスは、意味やコミュニケーションの構造を明示しているので、事例に基づく知識ベースとなる。

2 タグの普及

言うまでもなく、GDAの最大の課題は、いかにしてタグを普及させるかという点にある。タグの普及は、タグ付けのメリットとそのコストとの関係にかかっている。GDAタグの最大のメリットは、タグ付けされた文書が機械翻訳(特に多言語翻訳)や情報検索や質問応答などの多くの用途に再利用できることである。こうした多種の応用技術を統合する文書処理環境は、特に、多量の文書処理の需要をかかえ、また専門のタグ付け作業者を養成したりタグ付けを外注したりできる大手の機関ユーザにとってきわめて有用かつ現実的な技術であろう。

しかし、教育を含めてこうしたインフラを整えるには手間と時間がかかる上に、オペレーティングシステム等の場合と違って、タグ付けのメリットはタグの仕様や技術内容の公開と共有を前提としており、また、タグの応用に関してはいわゆるキラーアプリケーションのようなものもさしあたりは存在しない。したがって、特定の企業がデファクト標準を握って独占的な利益を上げるなどということは考えにくいので、企業が独自にタグの普及を推進することはないだろう。しかし、タグ付けが普及すればユーザはより高度なサービスを享受でき、情報処理産業全体の市場規模も拡大するはずである。インターネットが研究者のツールとして始まったのと同様に、GDAタグも研究コミュニティを出発点として普及させるのがよいだろう。さしあたっては、タグに関連するさまざまなツールや応用技術を開発しながら、タグの普及を図る必要がある。

3 タグ集合とタグ付け

GDAタグ集合は、章、節、段落、タイトルなどの文章の構成を表わすタグ、統語範疇を表わすタグ、曖昧性を表わす選言的タグ (alternate tag) に加えて、主題役割 (thematic role) や修辞関係 (rhetorical relation) などの意味関係、照応、言語行為、テンス、アスペクト、量化やや否定や様相演算子の作用域、語義などを表わす属性からなる。タグの

多くと属性の一部はTEI[1]やEAGLES[2]などの既存のガイドラインに基づいている。GDAでは多くののタグは任意的である。また、意味関係についてはEDRコーパスやGeneralized Upper Model[3]など、照応についてはMUC[4]のコーパスやLancasterコーパス(Garside et al., 1997)の仕様も参考にして設計している。GDAタグ集合の仕様書のドラフトを昨年の夏から公開している[5]。コメントをいただければ幸いである。

GDAタグ付きテキストの例を図1に示す。<v>エレメントは動詞または動詞句、<n>エレメントは名詞または名詞

```
<v ctyp=fx><ad><n id=bk> 本 </n> を </ad>
          <v><ad> お母さんに </ad>
             <v><v obj=bk> 読んで </v><v> もらう </v></v>
          </v>
</v>
```

図 1: GDA タグ付きテキスト

句、<ad>エレメントは副詞や後置詞句や連体詞などある。ctyp属性は統語的構造、obj属性は目的語を表わす。fxは交差を含む前向きの依存関係である。日本語では、省略されたctypの値はfd(前向きの依存関係)と解釈される。

GDAタグ付きコーパスを作成中だが、そこではGDAタグの仕様全体を使うのではなく、さしあたりは統語構造と意味関係と照応だけに関するタグ付けを行なっている。後述のように、そのような単純な部分集合でも自然言語処理の応用の精度を高めることができる。

人手によるGDAのタグ付け作業を支援するソフトウェア・ツールを**タギングエディタ**と呼ぶ。現在のタギングエィタはGNU Emacsのモードのひとつとして実現されている。そのウィンドウ表示の例を図2に示す。ウインドウは左

```
.v_ad_n.....$| 本
  |   |______$| を
  |ad.......$| お母さんに
  |v_v......$| 読んで
    |v......$| もらう
```

図 2: タギングエディタのインタフェース

右に分かれ、左側にタグの入れ子構造を、右側にその構造の各部に対応する文字列を表示している。実線の枝はエレメントでない文字列を、点線の枝はエレメントになっている文字列を示す。タギングエディタは、タグの範囲の指定、タグと属性の編集、id属性の自動設定などの機能を持つ。タギングエディタも公開中[6]である。GDAタグ集合以外でもSGMLであれば比較的簡単にカスタマイズできる。

4 自然言語処理ツールの入出力の標準化と統合

GDAは自然言語処理のさまざまなツールの入出力形式をタグ付きテキストとして標準化することを含意する。これは、図3に示したような統合的なアーキテクチャを意味する。このようなツールは、実際にはゼロから作るのではなく、たいていは既存のツールにwrapperをかけるだけで安価に作ることができる。このアーキテクチャにより、(データへのアクセスを含む)多様なソフトウェアツールを再利用し、統合(プラグアンドプレイ)することが可能となり、自然言語処理システムの開発と管理が簡単になる。この方式によってタギングエディタにパーサや意味タガーをプラグインすれば、機械による解析の結果を用いて人間の負荷を軽減することができる。

同様の統合アーキテクチャはすでに最近いくつか提案されている。LT-NSL (McKelvie et al., 1997)はTEI (Sperberg-McQueen & Burnard, 1994)に基づくSGML形式の標準フォーマットを用いた統合アーキテクチャである。TIPSTERアーキテクチャ[7]では、タグを元のテキストに埋め込むのではなく、元のテキストとそれに対するタグを別のファイルとして扱うデータベース管理システムを核としてさまざまなツールを結合する。GATE[8] (Cunningham

[1]http://www.uic.edu:80/orgs/tei/
[2]http://www.ilc.pi.cnr.it/EAGLES/home.html
[3]http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html
[4]http://cs.nyu.edu/cs/faculty/grishman/muc6.html
[5]http://www.etl.go.jp/etl/nl/GDA/tagset.html
[6]http://www.etl.go.jp/etl/nl/gda/TE/
[7]http://www.tipster.org/arch.htm
[8]http://www.dcs.shef.ac.uk/research/groups/nlp/gate/

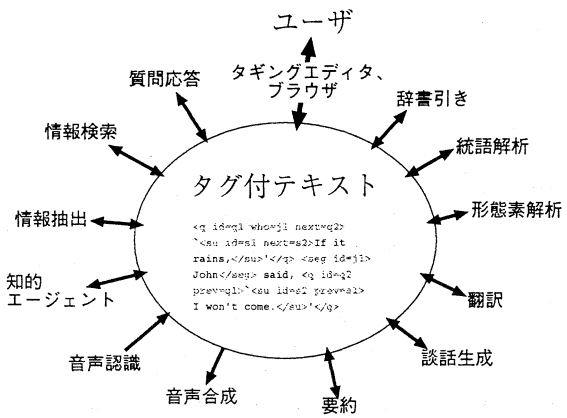

図 3: 標準インタフェースによる統合型自然言語処理環境

et al., 1997) および Corelli (Zajac, 1997) は TIPSTER アーキテクチャを実装・拡張したものである。TIPSTER アーキテクチャはもともと情報検索と情報抽出の統合開発環境として考えられたものだが、GATE と Corelly では機械翻訳などにも及ぶ。

GDA は、一般ユーザをタグ付けデータの提供者としてこうした統合環境に組み入れることにより、これを単なる開発環境ではなく利用環境へ広げようという試みである。TIPSTER などはインターネットにおける分散サービスも視野に置いているが、GDA の文脈ではこれはたとえば辞書の分散管理などに有用なので、これらのアーキテクチャとの連携を考えたい。

5 応用

GDA タグは自然言語処理のさまざまな応用の品質を劇的に向上させる。たとえば、タグの情報を用いることによって機械翻訳の精度の大幅な向上が期待される。ヨーロッパ語の間での機械翻訳は、文章の種類によってはかなり理解可能であり、ブラウジングのための翻訳としては何とか使えるレベルに達している。しかし、情報発信のための翻訳としてはまだ不十分であり、タグの利用によって情報発信のためのこなれた翻訳の可能性を視野に入れることができる。系統の異なる言語の間での翻訳はブラウジングに使えるレベルにも達していないが、タグの利用によって理解可能な翻訳にすることは可能であろう。

横山 (1995) は日本の主要メーカが開発した商用の 8 つの日英翻訳システムの性能を調査しているが、GDA タグの設計に当たっては、そこで指摘されているような問題を含むさまざまな問題の解決を意図した。たとえば、「何でも、ワームホールという現象を利用すれば、遠く離れた世界に瞬時に行けるのだそうだ」の英訳はかなり難しく、8 つのシステムによる翻訳のうちで最もまともなものは

> It is possible to go to the world away instantaneously far if the phenomenon such as anything and wormhole is used.

であった。ここでは「何でも」の係り先と意味が正しく解析できていないが、明らかにこの問題は統語構造と語義のタグによって解決できる。統語構造と言ってもさほど詳細なものである必要はない。「何でも」が文副詞であることがわかる程度の大雑把なもので十分である。また、語義タグを用いるには共有可能な語義の体系 (ontology) を整備する必要があるが、語義タグを用いなくても、パターンに基づく翻訳 (Maruyama, 1993) と統語構造のタグを組み合わせることによって多くの場合に対処可能である。たとえば下記のような翻訳パターンを用いることが考えられる。

`'何でも、 *X そうだ' → 'they say *Y'  where *X → *Y.`

`*X` と `*Y` は変項である。$A \rightarrow B$ は「A は B に翻訳できる」と読む。変項と → の両辺は SGML のエレメントと単一化可能とする。活用や語順に関する詳細は割愛した。

統語解析や照応解析などの言語的な解析は、情報検索や要約などではこれまではあまり用いられていなかったが、これはこうした解析に手間がかかる割には得られる結果の信頼性が低いためである。しかし、GDA のタグを用いれば、言語的な解析の手間を軽減し、かつその結果の信頼性を格段に高めることができるので、これまでそうした解析を使ってい

なかった応用においても高度な自然言語処理の手法が有用となる可能性が高い。情報検索は質問応答に近いものになるだろう。データマイニング、ネットワークエージェント、事例に基づく推論などに関しても同様の可能性が考えられる。

要約もそのような応用のひとつである。英語と日本語の文書に関して実験したところ、統語構造、意味関係および照応に関するタグがあれば、言語に依存しない非常に簡単なプログラムでも要約が十分に可能である (長尾, 橋田, ・宮田, 1997)。意味関係のタグ付けにはかなりの揺れがあるが、この揺れは要約の結果にはほとんど影響しない。

この要約プログラムは、意味的・語用論的な依存関係のネットワークをタグ付き文書から構成し、このネットワークの上で活性拡散 (Hasida, Ishizaki, & Isahara, 1987) によって各語句の重要度を評価し、この重要度と照応や統語論にまつわる制約に従って (必ずしも文全体ではない) 語句を抽出する。また、意味表現から文章を生成する一般的な技術を用いれば、単なる語句の抽出にとどまらず、さらに柔軟な要約が可能だろう。

6 おわりに

現在、GDA タグ集合仕様書の英語版と日本語用のタギングマニュアルを作成中であるが、タギングマニュアルは各言語用に必要である。フランス語、中国語、ドイツ語、インドネシア語、およびタイ語のタギングマニュアルの作成は今年度中に着手したい。

Section 4で述べたように、GDA は他のいくつかのプロジェクトと部分的に重なっているので、標準化と共有を目指す (GDA を含む) これらのプロジェクトの性質により、プロジェクトの間での連携が重要である。特に、UNL はタグ付きテキストではなく中間言語に基づく統合的自然言語処理のプロジェクトであるが、GDA のタグ付きテキストと UNL の中間言語との間での自動変換を可能にすれば、両プロジェクトの成果の普及する範囲を足し合わせて拡大することができる。タグ付きデータによる統合という構想は、自然言語処理のみならずさまざまな領域で実現が試みられている。たとえば音声認識と音声合成のための標準タグを策定しようという動き[9]もある。動画像などにも及ぶマルチモーダルなデータの統合的な処理についてもタグ付きデータによるアプローチが有効であろう。

参考文献

Cunningham, H., Humphreys, K., Gaizauskas, R., & Wilks, Y. (1997). Software Infrastructure for Natural Language Processing. In *Proceesings of the Fifth Conference on Applied Natural Language Processing.*

Garside, R., Fligelstone, S., & Botley, S. (1997). Discourse Annotation: Anaphoric Relations in Corpora. In Garside, R., Leech, G., & McEnery, A. (Eds.), *Corpus Annotation — Linguistic Information from Computer Text Corpora*, pp. 66–84. Longman.

橋田浩一 (1997). 大域文書修飾. 『人工知能学会全国大会 (第 11 回) 論文集』, pp. 62–63.

橋田浩一, 杉村領一, 柏岡秀紀, 内山将夫, ・ Neumann, C. J. (1997). 大域文書修飾: 標準タグによる言語データの大規模な構造化と再利用. 『言語処理学会第 3 回年次大会発表論文集』, pp. 135–138.

Hasida, K., Ishizaki, S., & Isahara, H. (1987). A Connectionist Approach to the Generation of Abstracts. In Kempen, G. (Ed.), *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pp. 149–156. Martinus Nijhoff.

Maruyama, H. (1993). Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP. In *Proceedings of Natural Language Processing Pacific Rim Symposium '93* Fukuoka.

McKelvie, D., Brew, C., & Thompson, H. (1997). Using SGML as a Basis for Data-Intensive NLP. In *Proceesings of the Fifth Conference on Applied Natural Language Processing.*

長尾確, 橋田浩一, ・宮田高志 (1997). GDA タグを用いた文書の要約に関する一考察. 『シンポジウム「実用的な自然言語処理に向けて」』.

Sperberg-McQueen, C. M. & Burnard, L. (1994). *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* ACH, ACL, ALLC.

Zajac, R. (1997). An Open Distributed Architecture for Reuse and Integration of Heterogeneous NLP Components. In *Proceesings of the Fifth Conference on Applied Natural Language Processing.*

横山晶一 (1995). 翻訳困難な例文の翻訳とネイティブチェック. *AAMT Journal*, **11**, 38–59.

[9] http://www.cstr.ed.ac.uk/projects/ssml.html