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

橋田 浩一
電子技術総合研究所

杉村 領一
松下電器マルチメディア開発センター

柏岡 秀紀
ATR 音声翻訳通信研究所

内山 将夫
筑波大学工学研究科

Christoph J. Neumann
筑波大学応用言語学類

1 はじめに

自然言語処理における最大の課題は意味や常識や文脈の処理であり、その研究には機械に理解できる膨大な知識が必要である。CYC (Lenat, 1995) のように人手で知識ベースを構築する試みはあるが、これでは莫大なコストがかかるうえに、明示的な規則で書きやすい知識しかとらえられない。そこで、WWW 上の言語データをタグで修飾することによりWWW 全体を計算機で解釈可能な事例ベースに変えることを考える。ホームページにタグ付けすることによって情報発信が効率的に行なえるのであれば、一般ユーザが自分のホームページにタグ付けするようになり、巨大な知識ベースが自然発生するだろう。多くの一般ユーザがHTML のタグ付けをしている現状を見れば、もう少し詳しいタグ付けを普及させるに足るメリットを生むような機械翻訳や情報検索のサービスは現在の技術で提供可能と考えられる。以下ではそのために必要なタグの標準や意味タグ辞書について論じ、研究コミュニティが中心となってインターネット上の言語データの構造化を推進することを提案する。

2 大域文書修飾

大域文書修飾 (GDA: Global Document Annotation)[1]は、統語・意味等に関する多言語間に共通のSGML タグの標準を作って普及させようというプロジェクトである。タグは、係り受け、代名詞の指示対象、多義語の意味など、かなり細かい情報を含む。文書中でタグはたとえば以下のように用いられる。

```
<seg sem=time0>time</seg>
<seg><seg sem=fly1>flies</seg> <seg sem=like0>like</seg> an arrow</seg>
```

ただし、タグの標準は設計中であり、この例がその標準に合うものかどうかは未定である。

GDA は以下の3つのステップからなる。

(1) 文書の (統語的、意味的、語用論的、その他の) 構造を表示するSGML タグの標準を提案する。

(2) タグを用いた機械翻訳、情報検索、質問応答、知識発見など、情報発信・交換を支援する応用の開発・普及を振興する。

(3) それによってタグ付けのメリットを生じさせ、多くのユーザが自分のファイルにタグを付けるように動機付け、タグを普及させる。

適切なタグの標準を作って普及させれば、機械にも人間にも理解可能な知識ベースが世界規模で自己増殖し、自然言語処理や人工知能の技術が爆発的に実用化されて一般ユーザが恩恵を受けるのみならず、研究コミュニティにとっては基礎研究のための大量かつ良質のデータが手に入ることになる。意味や常識があと100年ぐらいは機械でまともに扱えないとすれば、そんな機械でもそれなりに活躍できる環境を整えてやる必要がある。その環境を整えることにより社会的ニーズに答えながら基礎研究をも進展させようというのがGDA である。

タグ付けには常識を含む知識を使った曖昧性の解消が必要だが、それが自動的にできれば苦労はないわけで、やはりタグ付けには人間が関与する必要がある。多くのユーザにタグを付けてもらうにはタグ付けのメリットが必要であり、それには、タグに基づく機械翻訳や情報検索のサービスが安価に利用可能になっていればよい。そうすれば、真剣に情報発信しようとしているユーザならタグを付けようという気になると考えられる。タグを付けておけば、さまざまな言

[1] http://www.etl.go.jp/etl/nl/gda

語で広く読んでもらえるし、また高精度の情報検索にかかるので適切な読者を得る可能性が高まる。ホームページを絵や音で装飾して目立たせようとするのが流行っているが、タグによる修飾は、内容そのものを構造化することによりホームページをもっと適確に目立たせる効果を持つ。インターネットにおいては、画像が大きな役割を演ずる場合も多いが、やはりほとんどの情報は自然言語で表現されているので、自然言語によるコミュニケーションを支援する技術がインターネットの発展に不可欠であることは論を待たない。

(2) と (3) の間には正帰還の関係がある。つまり、タグを前提にしたアプリケーションが出回れば多くのユーザがタグ付けする気になり、そうして多くのファイルがタグ付けされればタグを前提にしたアプリケーションがますます出回るようになる。問題は、このサイクルをいかにして回し始めるかである。すでに翻訳システムを開発しているメーカーなら、タグを前提にした翻訳システムを1人月ぐらいで開発できるだろう。タグを利用した高精度の検索技術の開発もそれほど難しくないと考えられる。研究・開発に携わる十分多くの人々がその気になれば、半年ぐらいのうちにサイクルを回し始めることができるのではないだろうか。

多くの人々を動機付けるために、GDA の考え方のメリットを宣伝する必要がある。研究者にとっては、タグ付けによって自然言語処理や人工知能の技術が実用化されるだけでなく、タグ付けされたデータを使って研究できるというメリットがあるので、研究者のコミュニティを説得するのはあまり難しくないだろう。実際、(GDA と違って自動的な方法によってだが) WWW を知識ベースにしようという提案はすでになされている (Selman et al., 1996)。また、UNL (Universal Networking Language) プロジェクト (UNL Center, 1996)[2] では GDA と同様のタグ付けを考慮している。しかし、メーカーの経営者や一般ユーザを説得する材料としては、上記 (2) によって生ずるユーザのメリットしかない。どういうタグを普及させるためにどういうメリットを一般ユーザに提供できるかは、研究コミュニティの知恵の出しどころである。

3 タグの標準化

ユーザのメリットが大きい応用としては、翻訳、検索、要約などが考えられる。機械翻訳できる程度にタグ付けするということは、翻訳の際に問題になる曖昧性のうちで機械で自動的に解消するのが難しいものを解消しておいてやるということである。そのためのタグとしては少なくとも以下のようなものが必要だろう。

- 意味タグ: 多義語 (句) の意味
- 統語タグ: 構文木の構造、統語範疇、主辞の情報など
- 指示タグ: 代用表現の指示物、時間、場所など
- 語用論タグ: 文章の種類 (広告、論説、手紙など)、談話参加者の社会的関係など

検索の高度化や質問応答は翻訳用のタグがあれば大体できると考えられる。要約には別種のタグ集合が必要かも知れないが、いずれにせよ、各タグ集合のメリットが明確になるように設計することが望ましい。

タグの仕様はいきなり完成するのではなく、いろいろな要因を考慮に入れながら次第に練れて行くと考えられる。しかし仕様の変更によって以前のタグが無意味になっては困るので、変更はインクリメンタルなものとし、旧版に基づくタグでもそれなりに翻訳や検索ができることを保証する必要がある。もちろん、新しい仕様に基づいて精密なタグを付ければ翻訳や検索の精度が高まるわけである。また、知識表現言語を作ろうというわけではないので、機械的に処理できる程度ならタグには文脈依存性があってもよいだろう。たとえば、「健が来た。奈緒美も来た。」という場合に、第2文に時刻に関するタグが付いてなければ奈緒美が来たのは健が来た後と解釈する、などというわけである。

当然、タグ付けの間違いがなるべく起こらないようにして、結果の品質を保証する必要がある。ところが、人間がタグ付けするのだから、あまりうるさい制限を加えるのも現実的ではない。また、専門家によるタグ付けの結果も人と場合によって揺らぐのが当然であり、一般ユーザによるタグ付けではさらに揺らぎが大きいに違いない。しかし、タグが少々揺れても意味さえわかればよい。タギングの揺らぎから生ずる曖昧性は、タグなし入力に含まれる曖昧性に比べればずっと簡単に解消できるだろう。

UNL プロジェクトでは UNL という中間言語を考えているが、GDA のタグは UNL ほど詳細な情報を持つものではない。中間言語を用いればアプリケーションの開発が個別言語に依存しないので非常に効率的であるが、十分な普遍性を持つ中間言語の設計はきわめて難しい。UNL プロジェクトではその困難を見越して中間言語だけでなくタギングも考えており、この点で GDA に通ずる。UNL と GDA の間でタグの互換性を保ちながら協調して進める予定であ

[2] http://unl.ias.unu.edu/~unl/

る。また、タグの標準化に関しては、TEI[3]、EAGLES[4]、CES[5]などの成果も取り入れることが望ましい。

4 意味タグ辞書

人間が多義語に意味タグを付ける作業のため、さまざまな自然言語による直観的な説明を各意味タグに付けておく必要がある。したがって、意味タグ辞書は多言語辞書となる。また、一般ユーザによるタグ付けのコストを抑えるためには、少なくとも一般ユーザには意味タグ辞書を無料で配布する必要があるだろう。

さしあたり英語と日本語の説明を含む辞書を作るため、WordNet (Miller, 1995) と EDR 辞書 (Yokoi, 1996)[6] とのアラインメントを行なっている。これに関しては、同様の作業をしている東大・東工大・EDR の合同プロジェクト (西野他, 1997; 出羽他, 1997) と協同で進めている。WordNet はフリーウェアだが、きわめて残念ながら EDR 辞書は国有財産であるため、このようにしてできた辞書を一般ユーザに無料で配布する際に工夫を要する可能性がある。無料配布可能な辞書の材料としては、EuroWordNet[7]などもある。また、日本の機械翻訳協会のプロジェクト (伊藤他, 1997) では機械翻訳用のユーザ辞書の共有化を進めており、その成果も利用できるだろう。

意味タグの体系は一種のオントロジーであるから、オントロジーの標準化に関するさまざまな試みとも協調することが望ましい。たとえば SHOE[8] (Luke et al., 1997) は、HTML をオントロジーに関するタグによって拡張し、これに基づいて情報検索などを高度化しようという、GDA と似た発想のプロジェクトである。Ontolingua[9]など、大規模なオントロジーを公開する動きもある。また、ANSI で検討されているオントロジーの標準化[10]では、WordNet と EDR 辞書に加えて CYC や Pangloss の辞書を用いてオントロジーのアラインメントのための手法などを探っており、その成果は GDA で利用できるだろう。

5 タグの応用と普及

前述のように、タグを普及させるための主な応用技術は、翻訳、検索、要約などであろう。タグを付けておけば、さまざまな言語で読まれ、また正確に検索されるので、多言語での効率的な情報発信が安価にでき、ホームページの著者にとって大きなメリットとなる。また、要約が簡単にできれば文書の概要を素早く把握できるので読者にとって便利なのは明らかであるが、読者に読みやすい文書が作れるということは著者にとってのメリットにもなり、タグの普及が促される。

もっと公共的な視点から見ると、インターネットや学問研究の場で英語しか通じないというのは不健全かつ非効率的であるが、タギングによって世界の言語的多様性を保持したまま情報の流通を促進できることも、社会的インパクトという意味では重要である。また、視覚障害や文盲のためにそもそも文字を読めないユーザに対する情報提示のための音声合成の品質を向上させるためにも、タグの情報が有用だろう。

一般ユーザにとってよりも自然言語処理や人工知能の研究者にとっての方がタグ付けのメリットが大きいので、いきなり一般ユーザにタギングを広めようとするのではなく、まず研究コミュニティが率先してタグを普及させる方が現実的かも知れない。たとえば、言語処理学会などが学会誌の論文にタグ付けして翻訳システム込みで公開するというのは荒唐無稽な話ではない。文学的情緒を保存するのは無理としても、技術的な文章なら原文の論理的意味を保存する翻訳ぐらいはできるだろう。また、研究所や学会がその活動内容などに関する文書情報にタグを付けて、WWW で自動的な質問応答が可能な形で公開するのも面白い。それに近い質問応答システムとして START[11]がある。START は、タグ付けされたデータに基づいて、英語の質問に対してマルチメディア情報を交じえながら英語で答えることができる。

翻訳とか検索以外に、データマイニング、CBR、ネットワークエージェントなどの技術もタグを普及させる原動的となるようなメリットを生み出す可能性があるが、本稿ではこれらに関して詳述するゆとりはない。

[3] http://www.uic.edu:80/orgs/tei/
[4] http://www.ilc.pi.cnr.it/EAGLES/home.html
[5] http://www.cs.vassar.edu/CES/
[6] http://www.iijnet.or.jp/edr/
[7] http://www.let.uva.nl/ ewn/
[8] http://www.cs.umd.edu/projects/plus/SHOE/
[9] http://www-ksl-svc.stanford.edu:5915/doc/ontology-server-projects.html
[10] http://www-ksl.stanford.edu/onto-std/index.html
[11] http://www.ai.mit.edu/projects/infolab/index.html

6 タギングエディタ

人間によるタグ付け作業の負荷を軽減するためのタギングエディタが必要である。これは、HTMLファイルのエディタのように、人間がタグをあからさまに意識することなくファイルを修飾できるようにするツールであるが、GDAでは現在のHTMLよりも複雑な修飾を行なうので、ユーザインタフェースの設計には工夫が必要だろう。おおよそ以下のような方針の下にタギングエディタを作成中である。

タグの仕様(DTDファイルの形式)に一定の制約を設けることによって、タグの記述力に関する一般性を失うことなく、タグ付け作業をたとえば次のような少数の種類に限定できる。

- テキスト中のある範囲についてその性質(品詞や意味など)を指定する。
- テキスト中の2つの範囲についてそれらの関係(一方が他方の補語や先行詞やせりふの続きや根拠などであること)を指定する。

このほか、タギングエディタには、タグ付け作業をしている人間にタグと属性(attribute)の直観的な意味を提示する機能が必要である。そこで提示すべき情報はDTDファイルとは別の入力ファイル(意味タグ辞書など)に書かれる。

タギングエディタは、最低限DTDファイルと意味タグ辞書さえあればタグ付け作業ができるように設計する。さらに、形態素解析や統語解析などのプログラムをプラグインすることによってユーザの作業をガイドする。解析プログラムとタギングエディタとの間の通信はやはりタグ付きテキストによって行なう。その際に曖昧性を含む解析結果を表現するためのタグを定義しておく必要がある。この形式で入出力を行なうようなフィルタをかけることにより、多くの解析プログラムを簡単にプラグインできるようする予定である。

7 おわりに

情報発信したいとか自分を表現したいとかいう、莫大なエネルギーを持つ需要が、特にインターネットに喚起されて世界中で醸成されつつある。その需要を満たす技術の供給を通じて、そのエネルギーを利用することにより、意味や文脈や常識の研究を大きく進展させるための環境を構築しようというのがGDAである。現在、タグ集合、意味タグ辞書、およびタギングエディタの第1版を1997年夏に公開すべく準備を進めている。これらに基づく応用プログラムをぜひ多くの方々に作っていただき、上のような応用・研究環境の実現を図りたい。

謝辞

多数にわたるので名前は挙げないが、GDAメイリングリスト(gda@etl.go.jp)での議論に加わって下さった方々に感謝する。

参考文献

出羽達也・西野文人・辻井潤一 (1997). 知識獲得に向けての品詞体系の考察. 『言語処理学会第3回年次大会論文集』.

伊藤悦雄・村木一至・桧山努・赤羽美樹子・斎藤由香梨・平井徳行・亀井真一郎 (1997). 機械翻訳ユーザ辞書の共通フォーマットの設定 — アジア太平洋機械翻訳協会 における活動中間報告 —. 『言語処理学会第3回年次大会論文集』.

Lenat, D. B. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, **38** (11), 33–38.

Luke, S., Spector, L., Rager, D., & Hendler, J. (1997). Ontology-based Web Agents. *Proceedings of First International Conference on Autonomous Agents.*

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, **38** (11), 39–41.

西野文人・杉山健司・辻井潤一 (1997). 知識ベース増殖のための中央データベース. 『言語処理学会第3回年次大会論文集』.

Selman, B., Brooks, R., Dean, T., Horvitz, E., Mitchell, T. M., & Nilsson, N. J. (1996). Challenge Problems for Artificial Intelligence. *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Conference on Innovative Applications of Artificial Intelligence*, pp. 1340–1345.

UNL Center, (1996). *UNL: Univeral Networking Language — An Electronic Language for Communication, Understanding, and Collaboration.* Tokyo: UNL Center, Institute of Advanced Studies, The United Nations University.

Yokoi, T. (1996). The EDR Electronic Dictionary. *Communications of the ACM*, **38** (11), 42–44.