言語処理システムの性能向上を目指した言語的注釈付けシステム

渡辺日出雄*、長尾確*、Michael C. McCord**、Arendse Bernth**
*日本IBM東京基礎研究所
** IBM Watson Research Center
{watanabe,nagao}@trl.ibm.co.jp

1 はじめに

近年の急速なインターネットの普及により、機械翻訳、情報検索、自動要約などの言語処理アプリケーションは、その主要な文書形式であるHTML(やXML)で書かれた文書を処理することが求められている。しかし、これらの文書を処理するにあたって幾つかの問題点があることを我々は[6]にて示した。これらの問題点の例としては、スタイル上の効果を得るために誤ってタグを使用したり(h3タグをボールド効果を得るために使用するなど)、テーブルのセル内で改行をもって文末とするなどがある。この様な問題以外に、言語処理自身がまだ一般ユーザーの期待する品質に到達していないという問題があることも事実である。これは、基本的には、言語が持つ曖昧さをうまく解消できていないということである。例えば、以下の二つの文は、それぞれ解釈が二通りあるが、これは、書いた本人でなければ解消できないであろう。

I saw a man with a telescope.
He likes accomodating people.

これらの問題点の解消、および、言語処理の品質向上のため、我々は言語処理を補助するようなタグセットを用いてあらかじめ文書に注釈付けをし、その情報を解釈できる言語処理システムを作成し、一般ユーザーに高度な言語処理システムを提供することを目標としている。

本論文では、我々が開発したLinguistic Annotation Language (LAL)という言語的注釈付けのためのXMLタグセットについて説明し、それを解釈可能な言語処理システムおよび、タグ付けを援助する環境(エディタ)について報告する。

2 Linguistic Annotation Language

LAL[8]は、XMLに準拠したタグセットであり、lalというprefixを用いたnamespaceを使用する[1]。以下に、LALの主要なタグについて説明する。

基本的には、文(s)と任意のレベルの句(seg)と単語(w)の範囲(スコープ)を指定するタグを用いる。

⟨lal:s⟩This is a sentence.⟨/lal:s⟩
This is ⟨lal:seg⟩a noun phrase⟨/lal:seg⟩.
This is a ⟨lal:w⟩word⟨/lal:w⟩.

これらのスコープ指定でもおおまかな依存関係は表現できる[2]が、より直接的にidとmod属性を用いて、依存関係を表現することも可能である。id属性でwあるいはsegにユニークなIDを付与し、mod属性にその設定されたID値を指定することにより依存関係を表現する。例えば、図1は"I saw a man with a telscope."という英文のLALによる注釈付けの例であるが、この例では"with"が"saw"に係っていることから、「望遠鏡で見る」という解釈であることを示している。

また、wタグは、idとmod以外に、品詞(pos)を表す属性なども記述可能である。

[1] これ以外に、既存のタグとの交差を防ぐため、Processing Instructionを用いて表現することも可能である。
[2] 完全に表現するにはヘッドがどの単語であるかが分からなければならない。

```
⟨lal:s⟩
⟨lal:w id="1" pos="pron" mod="2"⟩I⟨/lal:w⟩
⟨lal:w id="2" pos="v" mod="0"⟩saw⟨/lal:w⟩
⟨lal:w id="3" pos="det" mod="4"⟩a⟨/lal:w⟩
⟨lal:w id="4" pos="noun" mod="2"⟩man⟨/lal:w⟩
⟨lal:w id="5" pos="prep" mod="2"⟩with⟨/lal:w⟩
⟨lal:w id="6" pos="det" mod="7"⟩a⟨/lal:w⟩
⟨lal:w id="7" pos="noun" mod="5"⟩telescope⟨/lal:w⟩.
⟨/lal:s⟩
```

図 1: ESG の解析結果出力例

3 LAL-aware NLP Tools

我々は、LALにより注釈付けされた入力を受け付けることができる言語処理システムを幾つか開発した。

ESG[2] は、弊社ワトソン研究所で開発された英語の構文解析システムである。この ESG は、LAL で注釈付けされた英語文を入力として受け付け、かつ、解析結果を LAL で注釈付けされた英文として出力するように拡張されている。先に示した図 1 のタグ付け例は ESG による実際の出力例である。

PalmTree[3, 4, 5] は、英語から日本語への翻訳システムである。PalmTree も、LAL で注釈付けされた英語（プレーンテキストおよびＨＴＭＬ）を解釈可能であるように拡張された。PalmTree では、例えば、LAL タグのスコープ情報や依存関係情報は構文解析時の枝刈情報として使用 [7] されている。

LAL タグは入力文中の全ての単語について付与されている必要はなく、一部分にだけ付与されていても良いことにしている。これは、LAL タグ付きの文章を入力として受け付ける言語処理システムは、このように一部分にだけ注釈付けされている場合には、その部分情報だけを使用して、残りの部分には既存のデフォールトのロジックで対処するというシームレスな処理をすることを求めている。上記の二つのシステム (ESG と PalmTree) は、この制約に則って実現されている。

4 タグ付けエディタ

上記のようなタグを実際にユーザーが人手で付与することは非常に困難であるので、タグ付けを容易に行うことができるような環境が必要となる。そこで、我々は言語的タグ付けのためのエディタを開発した。

このエディタは、文の係り受け関係を分かりやすくユーザーに提示し、ユーザーが簡単に係り受け関係を修正できることを目的としたものである[3]。このエディタを開発する上で考慮した点は以下の事柄である。

- 最初は文の全体構造を大まかに俯瞰できるようにする。このレベルでは、文の中心述語とその格要素等が分かるようにし、個々の格要素内の構造は気にしなくてもよくする。そして、次に、必要であれば、詳細なレベルの構造も見ることができるようにする。
- 構造表示上では入力文の単語の出現順に読んでいくことができるようにする。

最初の考慮点は、一度に詳細な構造にまで立ち入らずに全体を俯瞰することにより、ユーザーにそれ以上の詳細なレベルの変更をするかどうかのオプションを与えるためである。逆に言うと、最低限このレベルの構造上の修正は行ってほしいとの意図の裏返しでもある。また、二つ目の考慮点は、入力文と単語の出現順が変わってしまうと、元の入力文を思い出すのが大変となり結果的に正しい構造への修正が困難になってしまうためである。

文の構造表示は、基本的に係り受け関係を提示するものであり、表示のロジックは以下のようになる。

- 全体構造表示では、主動詞とそれを修飾する要素をルートとする部分をそれぞれ出現順に別の行に表示する。
- 詳細構造表示では、以下のように提示する。

[3] もちろん、詳細な属性の修正をする機能も持っているが、これはエンドユーザー向けではなく、専門家向けであると位置づけている。

- ある語 x の前方から修飾する語が $y_1, ..., y_n$ とあった場合、y_1 から y_n までの y_i はこの順に、x の上に行を挿入し、その行の x の直後のカラム位置に配置する。
- ある語 x の後方から修飾する語が $y_1, ..., y_n$ とあった場合、y_1 が x の隣接する直後の語であれば、x と同じ行の x の直後のカラム位置に配置する。y_n から y_2（あるいは、y_1 が直後の語でなければ y_1）までの y_i はこの順に、x の下に行を挿入し、その行の x の直後のカラム位置に配置する。

これにより、上から下、また左から右へと単語を読んでいけば、入力文の表層の出現順に読むことができる。

図 2 は、以下の英文に対する初期画面である全体構造表示を示したものである。

IBM announced a new computer for children with voice recognition function.

このモードでは、主動詞とその修飾要素（をルートとする部分）がそれぞれ別の行として表示されている。また、主動詞は他とは異なる色で表示されているので、一目でどれが主動詞か分かるようになっている。これにより、この文の主要な骨格が正しいかどうか一目で判断できる。図 3 は、システムのデフォールトの解釈を詳細構造表示で提示したものであり、幾つか係り受けが間違っている。そして、図 4 は、それらの係り受けの誤りを修正したものである。

この構造の修正結果の LAL タグ付き解析結果を図 5 に示す。

また、図 6 に、この LAL タグを付与した解析結果を使用した場合としない場合の PalmTree による翻訳結果を示す。LAL タグを付与したことにより正しく翻訳されていることが分かる。

このタグ付けエディタは、日本語文に対しても同様に使用することができる。

更に、本タグ付けエディタは筆者の一人が参画している GDA[1] というタグセットにも対応している。この研究を通じて、LAL と GDA のあるサブセット間では、データを相互に変換可能であることが分かった。

5 おわりに

我々は、言語処理を援助するための LAL という XML ベースのタグセットを構築し、この LAL により注釈付けされた文を解釈可能であるように英語の構文解析と英日の翻訳システムを拡張した。また、この LAL によるタグ付けを容易にするエディタを開発し、実際に翻訳の精度が向上することを示した。

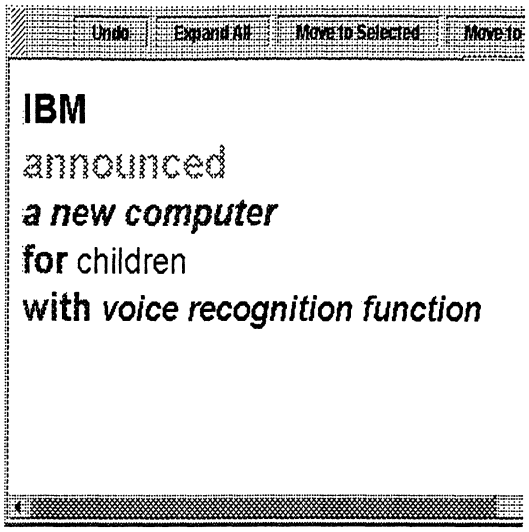

図 2: 全体構造表示画面

今後は、更にタグ付けエディタのユーザーインターフェースを改良し使いやすくするとともに、LAL タグをサポートした言語処理システムを増やしていく予定である。

参考文献

[1] Koichi Hashida, Katashi Nagao, et. al., "Progress and Prospect of Global Document Annotation," (in Japanese) Proc. of 4th Annual Meeting of the Association of Natural Language Processing, pp. 618–621, 1998.

[2] Michael C. McCord, "Slot Grammars," Computational Linguistics, Vol. 6, pp.31-43, 1980.

[3] Takeda, K., "Pattern-Based Context-Free Grammars for Machine Translation," Proc. of 34th ACL, pp. 144–151, June 1996.

[4] Takeda, K., "Pattern-Based Machine Translation," Proc. of 16th COLING, Vol. 2, pp. 1155–1158, August 1996.

[5] H. Watanabe and K. Takeda. A pattern-based machine translation system extended by example-based processing. In *Proc. of 17th Coling (Coling-ACL'98)*, volume 2, pages 1369–1373, 1998.

[6] 渡辺日出雄,「Web 文書に対する言語処理の問題点と言語処理を援助するタグセットについて」, 情報処理学会自然言語処理研究会 98-NL-127, pp. 95–100, 1998.

```
⟨lal:s id="id1-0"⟩
⟨lal:w id="id1-0-1" pos="propn" mod="id1-0-2"⟩IBM ⟨/lal:w⟩
⟨lal:w id="id1-0-2" pos="v" mod="0"⟩announced ⟨/lal:w⟩
⟨lal:w id="id1-0-3" pos="det" mod="id1-0-5"⟩a ⟨/lal:w⟩
⟨lal:w id="id1-0-4" pos="adj" mod="id1-0-5"⟩new ⟨/lal:w⟩
⟨lal:w id="id1-0-5" pos="noun" mod="id1-0-2"⟩computer ⟨/lal:w⟩
⟨lal:w id="id1-0-6" pos="prep" mod="id1-0-5"⟩for ⟨/lal:w⟩
⟨lal:w id="id1-0-7" pos="noun" mod="id1-0-6"⟩children ⟨/lal:w⟩
⟨lal:w id="id1-0-8" pos="prep" mod="id1-0-5"⟩with ⟨/lal:w⟩
⟨lal:w id="id1-0-9" pos="noun" mod="id1-0-10"⟩voice ⟨/lal:w⟩
⟨lal:w id="id1-0-10" pos="noun" mod="id1-0-11"⟩recognition ⟨/lal:w⟩
⟨lal:w id="id1-0-11" pos="noun" punct="period" mod="id1-0-8"⟩function⟨/lal:w⟩.
⟨/lal:s⟩
```

図 5: 修正後の LAL タグ付きの解析結果

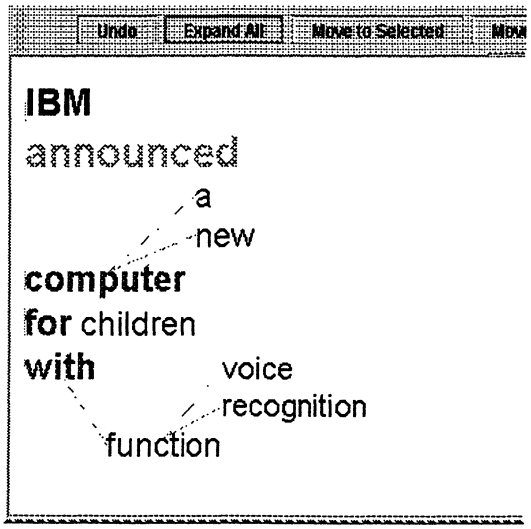

図 3: 詳細構造表示画面（修正前）

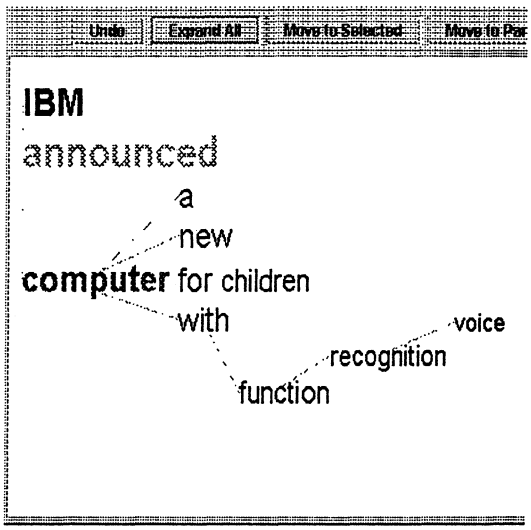

図 4: 詳細構造表示画面（修正後）

[7] 渡辺日出雄，「係り受け関係を用いたＣＦＧ構文解析の枝刈手法」，情報処理学会第５９回全国大会予稿集，Vol.2, pp.345–346, 1999.

[8] Hideo Watanabe, "Linguistic Annotation Language - The Markup Langauge for Assisting NLP Programs -", IBM Research Report RT0334, Nov. 16, 1999.

LAL タグ使用
IBM は、音声認識機能を持つこどものための新しいコンピュータをアナウンスした。
LAL タグ不使用
IBM は、音声認識機能でこどものための新しいコンピュータをアナウンスした。

図 6: LAL タグの有無での翻訳結果の相違