

Web 文書に対する言語処理を援助するタグセット*

6 R-4

渡辺 日出雄†
日本アイ・ビー・エム株式会社 東京基礎研究所‡

1 はじめに

近年のインターネットの普及により WEB 文書に対し自然言語処理アプリケーション（機械翻訳など）が使用される機会が増えてきている。しかし、言語処理技術の不完全さゆえ、一般ユーザーを完全に満足させる品質であるとは言えない状態である。この問題のかなりの部分は言語処理技術の不完全さから来ているが、WEB 文書を処理対象とした場合にはそれ以前の各種問題も存在する。本論文では、まず、WEB 文書に対する言語処理システムが遭遇する問題点について報告する。

これらの WEB 文書に対する言語処理の問題点を解消する一つの方策として、言語処理システムを援助するような情報を文書中に埋め込むということがある。このような観点から、言語リソースにタグ付けをして共有し、かつ、高度な言語処理を達成しようという動き [1, 2, 3, 5, 6] が多数出てきている。本論文では、基本的にこれらの動きに賛同しつつも、既存の WEB 文書にシームレスにタグ付けすることを目的とし、なおかつ、それほど複雑でないレベルのタグセットを提案する。

2 WEB 文書処理に関する問題点

ここでは、WEB 文書の翻訳と要約を例として、言語処理プログラムが遭遇する幾つかの問題点を紹介する。

2.1 視覚的効果のためのタグの誤用

ボールド書体で表示させたいなどの視覚的効果を得るために本来使うべきでないタグを使っている例がある。HTML 普及の初期の頃には、特にヘッダータグをそのような用途のために使っているケースがあった。機械翻訳にとってヘッダータグ内のテキストはタイトルであるとして特別な規則で解析をするという場合があり、このような場合に問題となる。

2.2 文スコープの認定

一般に、複数文からなる段落が与えられた場合、その中から文の単位を認定するのはそれほど簡単なタスクではない。これに加えて、HTML では、筆者が `
` タグで文の終わりの代用をしてしまっているケースがある。これは、特にテーブル内でよく見かける現象である。

<code><table> <tr> <td></code>
Internet Shops <code>
</code>
Cool Sites
<code></td> </tr> </table></code>

自然言語処理プログラムにとってテーブルのセル内が一行一文として訳すべきか、普通に句読点までを一文として訳すべきか知ることは出来ない。

* A Tag Set for Assisting NLP for Web Documents
† Hideo Watanabe (watanabe@trl.ibm.co.jp)

‡ IBM Research, Tokyo Research Laboratory

2.3 翻訳結果へのタグの挿入

機械翻訳プログラムでは、翻訳結果にソースと同じようにタグを挿入するという処理をする必要がある。デフォルトの動作としては、ソースの対応する単語に付いているタグをターゲット側にも付ければよいのであるが、そうでないケースもある。以下の例では、

```
<p>海の中に沢山の <a href="...">魚</a> がいます。
<br>
```

デフォルトの動作だけだと次のような出力になってしまうが、

```
There are many <a href="...">fishes</a> in the
<p>sea.<br>
```

本来は `<p>` は文頭になければならない。これは、`<p>` タグが文の先頭（性格には段落の先頭）に位置すべきものであるという位置に関する制約が分かって初めて可能となる。HTML を処理するプログラムではこれらの位置制約があらかじめプログラム中に埋め込まれているので正しい結果が得られるが、今後増えるであろう XML の翻訳を考えると、任意のタグを作成可能なので、何らかのタグの位置制約を記述する仕組みが必要となる。

2.4 文書内処理対象の認定

WEB 文書は一般に、関連リンク、タイトル、本文、著者名、著作権表示等、様々な要素から構成されている。しかし、現在の HTML ではこれらの領域及び意味を認識する仕組みはない。ここで、問題となるのは、テキストの自動要約処理である。現在実用化されている自動要約プログラムのほとんどのは、重要な文を取り出す文抽出型のものであり、それぞれの文に関して、含まれる重要なキーワードの数、文の文章中の位置¹等の表層の情報を基にして重要度を決定している。この様な処理の場合、要約処理に関する本文とタイトル部分だけを計算対象に含めないと、おかしな要約結果が出てしまうことになる。

3 Linguistic Annotation Language

ここでは、前節で挙げた各種の問題点を解消することを目指したタグセット (Linguistic Annotation Language or LAL) に関して説明する。

3.1 設計方針

LAL は以下のような方針に基づいてデザインした。

- 既存の文書へのタグ付け ... 新たな文書タイプを導入するのではなく、既存の任意の XML 文書にタグ

¹始めと終わりにある文は重要であることが多い

付け出来ることを目指す。このため、LAL のタグは処理命令 (Processing Instruction) を用いて実現している。

- 簡潔さ/効率の良さ … 人手であってもタグ付け可能な程度の簡単さを目指す。このため、主に構文的範囲指定のタグに限定している。²
- 自然言語処理プログラムの援助 … 自然言語処理プログラムの固有のアルゴリズムによる処理を援助するようなタグも導入する。
- 既存のタグセットとの連続性 … 従来から提案されているタグセットの中で、上記の基準に合うものは積極的に取りいれていくことにより、一種のサブセットの提案を目指す。

3.2 LAL タグの構造

LAL タグは、XML の処理命令に基づき、以下のような形式とする。

```
開始タグ ::= <?lal タグ名 処理内容?>
終了タグ ::= <?lal_ タグ名 処理内容?>
空要素タグ ::= <?lal タグ名 処理内容_?>
```

3.3 LAL タグ例

構文タグとしては以下のように文 (s)、単語 (w)、任意の句 (seg) を指定できる。

```
<?lal s?> これは文です。<?lal_s?>
<?lal w?>New York<?lal_w?>
She saw <?lal seg cat="np"?>a man with
telescope<?lal_seg?>.
```

意味的な情報指定タグとしては固有名詞 (proper)、日付 (date)、時間 (time)、数値表現 (num) に限定する。これらは、基本的に句の境界を指定すると同時に意味的な属性も指定していると考える。

```
<?lal proper type="country"?>U.S.A.<?lal_proper?>
<?lal date?>Dec. 25, 1999<?lal_date?>
<?lal time?>6:00 PM EST<?lal_time?>
<?lal num?>three hundred and twenty-one<?lal_num?>
```

特定の自然言語処理アプリケーション用の処理依存タグということで、機械翻訳と自動要約用のタグを用意した。

翻訳プログラムに対して明示的に翻訳のスコープを指定するタグとして以下のものを用いる。

```
<?lal tranStop_?> … 翻訳処理の停止
<?lal tranStart_?> … 翻訳処理の開始
```

また、前述したように、翻訳結果へのタグ付けのため以下の例に示す様なタグ情報 (taginfo) タグを用いて直後の XML (又は HTML) タグの使われ方に関する情報を付加する。

²これは、ツールを使った場合でも、ユーザーへのフィードバックを考えた場合に必要であると思われる。また、言語処理プログラムの誤りの半分くらいは、並列句の範囲認定などの構文的な範囲認定が分かれば解消できると思われる。よって、少ない労力で大きな効果を見込むことが出来る。

```
<?lal taginfo type="single" loc="bos_?> <p> <?lal
taginfo type="open_?> <a href="...> IBM <?lal
taginfo type="close_?> </a>
```

ここで、loc の bos はそれが文頭に置かれること³を示す。要約に関しては、以下の様な要約処理対象を指定するタグを用いる。

```
<?lal smrycalcStart_?> … 要約処理の開始
<?lal smrycalcStop_?> … 要約処理の停止
```

3.4 LAL タグと構文解析

LAL タグの範囲指定タグを利用してすることで、構文解析の精度はかなり向上すると思われる。構文解析では、LAL タグで指定された構文的範囲と矛盾するような解釈を保持する必要はなく、処理の途中での曖昧さの減少を大いに見込むことが出来る。

4 おわりに

本論文では、WEB 文書に対しての自然言語処理の問題点について説明し、ユーザーからの補助によりもう少し良い処理結果を期待できることを示した。そして、ユーザーからの補助手段として、言語的情報をタグ付けるためのタグセットである LAL を提案した。これは、任意の XML 文書にシームレスに言語的情報を挿入できるものであり、主要なタグは単語や句などの様々なレベルでの境界を指定するものである。更に、機械翻訳や自動要約などの言語処理に依存した情報付加のためのタグも用意した。

従来から、言語リソースにタグ付けをして共有しようという試み [1, 2, 5, 6] が知られているが、これらは主に自然言語処理等の研究のために考えられたこともあり、かなり詳細で複雑な記述となっていて一般的に普及するに至っていない。最近の HTML の普及により一般ユーザーのある程度のタグ付けを期待した試みとして GDA [3, 4] があるが、これもかなり詳細な記述となっている。LAL で目指したのは、実際にはツールが必要になるにせよ、ユーザーが何とか人手でタグ付けが出来るレベルの簡易なタグセットの提案であり、先行研究のある意味でのサブセットレベルを提示することである。

今後、LAL によるタグ付けツールと LAL タグを認識する言語処理プログラムの開発などにより、言語的タグ付けがされた WEB 文書が増え、ユーザーが現在の言語処理プログラムを最大限に利用できるようになることを目指していきたい。

参考文献

- [1] Corpus Encoding Standard (CES) (<http://www.cs.vassar.edu/CES/>)
- [2] Expert Advisory Group on Language Engineering Standards (<http://www.ilc.pi.cnr.it/EAGLES/home.html>)
- [3] Global Document Annotation (<http://www.etc.go.jp/etc/nl/gda/>)
- [4] Koichi Hashida, Katachi 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.
- [5] A Standard Extraction/Abstraction Text Format for Translation and NLP Tools (<http://www.opentag.org/>)
- [6] Text Encoding Initiative (TEI) (<http://www.uic.edu:80/orgs/tei/>)

³ちなみに eos は文末に置かれることを示す。