NLP のための Web ページ用標準フォーマット

橋本 力[*]　　河原 大輔[†]　　黒橋 禎夫[*†]
[*] 京都大学情報学研究科　　[†] 情報通信研究機構

平成 18 年 8 月 29 日

1　はじめに

近年、WWW のデータが自然言語処理の様々なタスクで活発に利用されている。それにともない、多くの研究機関で WWW データの蓄積が行われるようになった。

蓄積されたデータの再利用性を高めるためには、標準フォーマットを制定し、それに基づいてデータの蓄積を行うべきである。また、蓄積されるデータには、抽出された文集合だけでなく、その出処情報も明示されるべきである。これは、情報検索研究においてはもちろん、言語処理研究においても、抽出された文がどの Web ページのどの部分から得られたものなのかを知る必要がたびたび生じるためである。しかし、現在までに、このような標準フォーマットは提案されていない。

そこで本稿では、以上の用件を満たす、NLP のための Web ページ用標準フォーマットを提案する。我々のフォーマットは、元の Web ページとそこから抽出される文集合との間の中間形式として捉えることができる (図 1)。

図 1: 中間形式としての Web ページ用標準フォーマット

2　Web ページ用標準フォーマット

我々の Web ページ用標準フォーマット (以下、標準フォーマット) では、表 1 の情報が明示されている。「URL」と「開始位置」により、抽出された文がどの Web ページのどの位置から得られたものなのか分かるようになっている。「本文情報」の「種類」は、その本文がブログ (`blog`) か、ブログに対するコメントか (`comment`)、あるいは通常の Web ページの本文 (`default`) かを表す。「NLP ツールの解析結果」として、以下の例では KNP の解析結果を用いる。

標準フォーマットに基づくデータ (以下、標準フォーマットデータ) は XML で記述されるが、その DTD は図 2 のようになる。

また、標準フォーマットデータのエンコーディングは UTF-8 とする。

2.1　例 1

以下、§2.1 で、通常の Web ページからの標準フォーマットへの変換の例を挙げ、§2.2 では、ブログからの変換の例を挙げる。

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT StandardFormat (Text+)>
<!ATTLIST StandardFormat
    OriginalEncoding CDATA #REQUIRED
    Time CDATA #REQUIRED
    Url CDATA #REQUIRED
>
<!ELEMENT Text (S+)>
<!ATTLIST Text
    Author CDATA #IMPLIED
    Date CDATA #IMPLIED
    Title CDATA #IMPLIED
    Type (default|blog|comment) "default"
>
<!ELEMENT S (RawString,Annotation?)>
<!ATTLIST S
    Id CDATA #REQUIRED
    Length CDATA #REQUIRED
    Offset CDATA #REQUIRED>
<!ELEMENT RawString (#PCDATA)>
<!ELEMENT Annotation (#PCDATA)>
<!ATTLIST Annotation
    Scheme CDATA #REQUIRED>
```

図 2: **標準フォーマットの** DTD

表 1: 標準フォーマットに明示されている情報

種類	情報	XML タグ／属性	備考
出処情報	URL	Url	
	エンコーディング	OriginalEncoding	
	ページ取得日時	Time	「yyyy-mm-dd hh:mm:ss」形式。
本文情報	著者	Author	任意。
	更新日	Date	任意。
	タイトル	Title	任意。
	種類	Type	通常の Web ページかブログ、またはブログのコメント
文情報	文 ID	Id	
	バイト長	Length	
	開始位置	Offset	ファイル先頭からのバイトオフセット。
	文字列	RawString	文そのもの。
	NLP ツールの解析結果	Annotation	任意。
	NLP ツールの名称	Scheme	

まず、図 3 の HTML が標準フォーマットデータの派生元である。[1] そして、図 4 が標準フォーマットデータの例である。図 3 下部の 2 文が抽出されている。この例では「NLP ツールの解析結果」は付与されていない。図 5 は「NLP ツールの解析結果」が付与されている標準フォーマットデータの例である。

2.2 例 2

以下にブログから派生した標準フォーマットデータの例を挙げる。ブログは、通常、1 ファイルに複数の本文 (タイトルと記事のペア) が含まれるという独特の内部構造を持つ。そのため、ブログから派生した標準フォーマットデータの場合、通常、本文に相当する Text タグが複数回現れる。

ここで、タイトルは記事中の一文としても扱われることに注意されたい。そのため各記事のタイトルは、標準フォーマットにおいて、<Text>タグの Title 属性だけでなく、<RawString>タグの値としても現れる。

図 6 は標準フォーマットデータの派生元のブログである。[2] 図 7 がそのブログから得られた標準フォーマットデータである。この例では「NLP ツールの解析結果」は付与されていない。図 8 に「NLP ツールの解析結果」有りの例を挙げる。

3 HTML から標準フォーマットへの変換

我々は、標準フォーマットの制定とともに、HTML から標準フォーマットへの変換ツールを開発した。本節では、この変換で最も問題となる、Web ページからの日本語文の抽出方法について、その大枠を述べる。

我々のツールの日本語文抽出の大枠は以下の通りである。

[1] この例は http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html を加工して得たものである。
[2] この例は http://xtc.bz/を加工して得たものである。

```
<html><head><title>小泉総理プロフィール・信念</title>
<meta http-equiv="Content-Type"
   content="text/html;charset=Shift_JIS">
</head>
<body bgcolor="#ffffff" text="#000000">
<center>
<table cellpadding="0" cellspacing="0" width="610">
<tbody><tr>
<td valign="top" width="172">
<img src="sample_files/0_001.jpg"
   alt="無信不立" height="160" width="160">
</td>
<td valign="top" width="426"><b><font color="#0066ff">
 ▼座右の銘</font></b>
<br>
<img src="sample_files/1px_spacer.gif" alt="" height="5" width="426">
<br>
  小泉総理の好きな格言のひとつに「無信不立（信無くば立たず）」があります。論語の
下篇「顔淵」の言葉で、弟子の子貢（しこう）が政治について尋ねたところ、孔子は「食
料を十分にし軍備を十分にして、人民には信頼を持たせることだ」と答えました。<br>
</td>
</tr>
</tbody></table>
</center>
</body></html>
```

図 3: **派生元の** HTML **ファイル**

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat
  Url="http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html"
  OriginalEncoding="Shift_JIS" Time="2006-08-14 19:48:51">
  <Text Type="dafault">
    <S Id="1" Length="70" Offset="525">
      <RawString>小泉総理の好きな格言のひとつに「無信不立」があります。
</RawString>
    </S>
    <S Id="2" Length="160" Offset="595">
      <RawString>
        論語の下篇「顔淵」の言葉で、弟子の子貢（しこう）が政治について尋ねたところ、孔子は「食料を十分にし軍備を十分にして、人民には信頼を持たせることだ」と答えました。</RawString>
    </S>
  </Text>
</StandardFormat>
```

図 4: **標準フォーマットデータの例** (NLP **ツールの解析結果無し**)

1. **まず、日本語ページかどうか判定する。これには、文字コードや、日本語の助詞（「が」「を」「に」等）の有無などの情報を利用する。**
2. **次に、ページを文リストへ分割する。これには、**HTML **タグ（**
や<p>**など）と句点を利用する。**
3. **最後に、文リストから日本語文だけを抽出する。これは、日本語ページと判定されていても、文ごとに見ると英語の場合もあるためである。ひらがな、カタカナ、漢字のいずれかが** 60%**以上含まれる文のみを抽出する。**

4 おわりに

本稿では、NLP **のための** Web **ページ用標準フォーマットを提案した。標準フォーマットは、**Web **ページとそこから抽出される文集合との間の中間形式としての役割を果たすため、抽出された文だけでなく、その文の出処情報も含む。**

また本稿では、Web **ページからの日本語文の抽出方法についても述べた。**

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat
  Url="http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html"
  OriginalEncoding="Shift_JIS" Time="2006-08-14 19:48:51">
  <Text Type="default">
    <S Id="1" Length="70" Offset="525">
      <RawString>小泉総理の好きな格言のひとつに「無信不立（信無くば立たず）」が
ありま す。</RawString>
      <Annotation Scheme="KNP">
        <![CDATA[# S-ID:1 KNP:2006/08/10
* 1D <文頭><サ変><人名><助詞><連体修飾><体言><係:ノ格><区切:0-4><RID:1056>
小泉 こいずみ 小泉 名詞 6 人名 5 * 0 * 0 NIL <文頭><漢字><かな漢字><名詞相当
語><自立><タグ単位始><文節始><固有キー>
 ... 中略...
ます ます ます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 NIL <
表現文末><かな漢字><ひらがな><活用語><付属><非独立無意味接尾辞>
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS]]>
      </Annotation>
    </S>
    <S Id="2" Length="160" Offset="595">
      <RawString>
        論語の下篇「顔淵」の言葉で、弟子の子貢（しこう）が政治について尋ねたと
ころ、孔子は「食料を十分にし軍備を十分にして、人民には信頼を持たせることだ」と
答えました。</RawString>
      <Annotation Scheme="KNP">
        <![CDATA[# S-ID:1 KNP:2006/08/10
* 1D <文頭><助詞><連体修飾><体言><係:ノ格><区切:0-4><RID:1056>
論 ろん 論 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 代表表記:論" <漢字読み:音><
代表表記:論><文頭><漢字><かな漢字><名詞相当語><自立><タグ単位始><文節始>
 ... 中略...
ました ました ます 接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 タ形 5 NIL <
表現文末><かな漢字><ひらがな><活用語><付属><非独立無意味接尾辞>
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS]]>
      </Annotation>
    </S>
  </Text>
</StandardFormat>
```

図 5: **標準フォーマットデータの例** (NLP **ツールの解析結果有り**)

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="ja" lang="ja">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=EUC-JP">
<title>音楽配信メモ</title>
<div class="theday">
<h3 class="date5">2006 年 08 月 04 日（金）</h3>
<h4 class="title">はてなの音楽ブログが地味にヤバい</h4>
<p class="news">この前紹介した POP2*0 がヤバいのは、音楽フリークの中でもはや既
に常識であることは当然として、音楽業界に興味がある人も、この↓エントリは読んで
おくべき。</p>
<p class="newsfoot">| <a href="http://xtc.bz/index.php?cID=11">サイト紹
介</a> | <a href="http://xtc.bz/index.php?ID=366">この記事の URI</a> |
Posted at 23 時 17 分 |</p>
<p class="toplink">
<a href="#top">▲ページトップへ</a></p>
</div>
<div class="theday">
<h3 class="date1">2006 年 07 月 31 日（月）</h3>
<h4 class="title">第４回著作権分科会私的録音録画小委員会のまとめ</h4>
<p class="news">相変わらずお仕事が早い zfyl さんのところでまとめ記事が上がって
います。</p>
<p class="newsfoot">| <a href="http://xtc.bz/index.php?cID=3">著作権</a>
 | <a href="http://xtc.bz/index.php?ID=365">この記事の URI</a> | Posted
at 11 時 58 分 |</p>
<p class="toplink"><a href="#top">▲ページトップへ</a></p></div>
<div id="footer">
<p>作成者：津田大介<br>
デザイン：milkboy studio<br>
</p></div>
</body></html>
```

図 6: **派生元のブログ**

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat Url="http://xtc.bz/" OriginalEncoding="EUC-JP"
  Time="2006-08-14 19:48:51">
  <Text Author="津田大介" Date="2006-08-04"
    Title="はてなの音楽ブログが地味にヤバい" Type="blog">
    <S Id="1" Length="32" Offset="254">
      <RawString>はてなの音楽ブログが地味にヤバい</RawString>
    </S>
    <S Id="2" Length="148" Offset="308">
      <RawString>
        この前紹介した POP2*0 がヤバいのは、音楽フリークの中でもはや既に常識であることは当然として、音楽業界に興味がある人も、この↓エントリは読んでおくべき。
</RawString>
    </S>
  </Text>
  <Text Author="津田大介" Date="2006-07-31"
    Title="第４回著作権分科会私的録音録画小委員会のまとめ" Type="blog">
    <S Id="3" Length="46" Offset="777">
      <RawString>第４回著作権分科会私的録音録画小委員会のまとめ</RawString>
    </S>
    <S Id="4" Length="68" Offset="845">
      <RawString>相変わらずお仕事が早い zfyl さんのところでまとめ記事が上がっています。</RawString>
    </S>
  </Text>
</StandardFormat>
```

図 7: **ブログからの標準フォーマットデータの例** (NLP **ツールの解析結果無し**)

```
<?xml version="1.0" encoding="UTF-8"?>
<StandardFormat Url="http://xtc.bz/" OriginalEncoding="EUC-JP"
  Time="2006-08-14 19:48:51">
  <Text Author="津田大介" Date="2006-08-04"
    Title="はてなの音楽ブログが地味にヤバい" Type="blog">
    <S Id="1" Length="32" Offset="254">
      <RawString>はてなの音楽ブログが地味にヤバい</RawString>
      <Annotation Scheme="Knp">
        <![CDATA[# S-ID:1 KNP:2006/08/11
* 3D <文頭><形副名詞><体言><用言:判><タグ単位受:-1><係:未格><レベル:B><区
切:3-5><ID:~ の~ ><RID:346><連体並列条件><格要素><連用要素>
 ... 中略...
い い い 名詞 6 普通名詞 1 * 0 * 0 "代表表記:居" <代表表記:居><品曖><ALT-い-
い-いる-2-0-1-7-"代表表記:射る"><ALT-い-い-いる-2-0-1-7-"代表表記:鋳る"><品
曖-動詞><品曖-その他><文末><表現文末><かな漢 字><ひらがな><品詞変更:い-い-い
る-2-0-1-7><名詞相当語><自立><複合←><タグ単位始>
EOS]]>
      </Annotation>
    </S>
    <S Id="2" Length="148" Offset="308">
      <RawString>
        この前紹介した POP2*0 がヤバいのは、音楽フリークの中でもはや既に常識で
あることは当然として、音楽業界に興味がある人も、この↓エントリは読んでおくべき。
</RawString>
      <Annotation Scheme="Knp">
        <![CDATA[# S-ID:1 KNP:2006/08/18
* 1D <文頭><時間><強時間><外の関係><体言><係:無格><区切:0-0><RID:1316><格要
素><連用要素>
 ... 中略...
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS]]>
      </Annotation>
    </S>
  </Text>
  <Text Author="津田大介" Date="2006-07-31"
   Title="第４回著作権分科会私的録音録画小委員会のまとめ" Type="blog">
   ... 中略 ...
  </Text>
</StandardFormat>
```

図 8: **ブログからの標準フォーマットデータの例** (NLP **ツールの解析結果有り**)