

# 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 19 年 12 月 7 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、その出处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、以上の要件を満たす、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)。

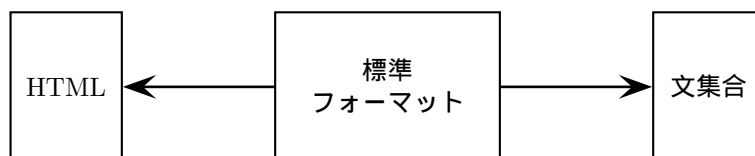


図 1: 中間形式としての Web ページ用標準フォーマット

## 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) 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
出处情報	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。<sup>1</sup> 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。<sup>2</sup> 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。

<sup>1</sup>この例は [http://www.kantei.go.jp/jp/koizumiprofile/1\\_sinnen.html](http://www.kantei.go.jp/jp/koizumiprofile/1_sinnen.html) を加工して得たものである。

<sup>2</sup>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">

</td>
<td valign="top" width="426"><b><font color="#0066ff">
      座右の銘</font></b>
<br>

<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、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、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（§3.1、§3.2 参照）。
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。

### 3.1 文分割

ブログ記事の文区切りは、新聞記事と比べると、ずっとはっきりしない。現在、次のような文区切り方針を設定している。

- 日付だけからなる文も一文とする
  - 「2006 年 10 月 09 日。」
  - <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30211804.html>
- URL だけからなる行も一文とする
  - 「<http://www.shigureden.com/>」
  - <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30211804.html>
- 箇条書きの一行一行をそれぞれ一文とする
  - 「・藤井大丸  
セレクトショップが多いです。大学生だと UNITEDARROWS や...

```

<?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、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イ</h4>
<p class="news">この前紹介したPOP2*0がヤバイのは、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zfy1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イ" 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イ</RawString>
    </S>
    <S Id="2" Length="148" Offset="308">
      <RawString>
        この前紹介した POP2*0 がヤバイのは、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イ" 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イ</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-
いい-いる-2-0-1-7-"代表表記:射る"><ALT-いいいい-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 がヤバイのは、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)

・紀伊國屋書店

新京極通三条下ルらしいです。前は河原町にあったブック...

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30105756.html>

- URL を含む文も一文とする

– 「URL 入力画面が出ますので, <http://url/> とでも入力すると, <http://url/> と出ます。」

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30142420.html>

- 途中に文末記号があっても、明らかな一文である場合は区切らない

– 「散歩?? かな。」

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30140027.html>

- 文開始直前、あるいは文終了直後の記号も、その文に含める

– 「そんな日本語ないか。笑」

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30140027.html>

– 「脱力。ORZ」

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/11/16145407.html>

– 「P.S. 数年前、電車の...」

– <http://nlp.kuee.kyoto-u.ac.jp/blog/KUNTTblog/2006/10/30003434.html>

## 3.2 括弧処理

括弧の使われ方も、新聞と比べるとずっと多様であり、構文情報付与の際の扱いが問題となる。

本研究では、原則として、括弧部分を元の文から取り除き、別の文として扱う。その際、元の文の ID 末尾に「-01」を付与し、取り除かれて別の文となった文の ID には「-02」「-03」「-04」等を付与する。

括弧処理前： [001] 「大」の字はいくつかのブロック（松明を掲げるのかな？）から構成されていました。

括弧処理語： [001-01] 「大」の字はいくつかのブロックから構成されていました。  
[001-02] 松明を掲げるのかな？

さらに、括弧文には次のラベルを与える。

- 括弧タイプ: ... 年齢, 読み, 日付, 金額, 場所, 所属, 同義, 文, その他
- 括弧位置: 元の文における文字単位の位置
- 括弧始: (
- 括弧終: )

### 3.2.1 括弧タイプ「年齢」の例

括弧処理前

# S-ID:950101016-001

共産党は一日付の機関紙「赤旗」で、宮本顕治議長（ 8 6 ）のインタビューを掲載した。

括弧処理後

# S-ID:950101016-001-01

共産党は一日付の機関紙「赤旗」で、宮本顕治議長のインタビューを掲載した。

# S-ID:950101016-001-02 括弧タイプ:年齢 括弧位置:23 括弧始:( 括弧終:)  
8 6

### 3.2.2 括弧タイプ「読み」の例

括弧処理前

# S-ID:950101010-002

こんな単純な発想にあやうさ、脆（もろ）さを感じる人は多いでしょうが、混迷の転換期を乗り切るため「日本は変わった」ことの証（あかし）であり、メッセージになるはずです。

括弧処理後

# S-ID:950101010-002-01

こんな単純な発想にあやうさ、脆さを感じる人は多いでしょうが、混迷の転換期を乗り切るため「日本は変わった」ことの証であり、メッセージになるはずです。

# S-ID:950101010-002-02 括弧タイプ:読み 括弧位置:15 括弧始:( 括弧終:)  
もろ

# S-ID:950101010-002-03 括弧タイプ:読み 括弧位置:60 括弧始:( 括弧終:)  
あかし

### 3.2.3 括弧タイプ「日付」の例

括弧処理前

# S-ID:950101020-007

日本の選挙制度は一八八九（明治二十二）年、小選挙区制中心でスタートしたのだが、現行の制度は、この男子普選法が原型となっている。

括弧処理後

# S-ID:950101020-007-01

日本の選挙制度は一八八九年、小選挙区制中心でスタートしたのだが、現行の制度は、この男子普選法が原型となっている。

# S-ID:950101020-007-02 括弧タイプ:日付 括弧位置:12 括弧始:( 括弧終:)  
明治二十二

### 3.2.4 括弧タイプ「金額」の例

括弧処理前

# S-ID:950101071-031

昨年十月には総額六百万カナダドルを投じて、鉄筋三階建ての編集・印刷センターを建設した。

括弧処理後

# S-ID:950101071-031-01

昨年十月には総額六百万カナダドル（約四億四千百万円）を投じて、鉄筋三階建ての編集・印刷センターを建設した。

# S-ID:950101071-031-02 括弧タイプ:金額 括弧位置:16 括弧始:( 括弧終:)  
約四億四千百万円

### 3.2.5 括弧タイプ「場所」の例

括弧処理前

# S-ID:950101091-004

2年前の世界選手権（カナダ・ハミルトン）で、男子3、女子1の金メダルを獲得した開催国・日

本がお家芸の座を守れるか。

括弧処理後

# S-ID:950101091-004-01

2年前の世界選手権で、男子3、女子1の金メダルを獲得した開催国・日本がお家芸の座を守れるか。  
# S-ID:950101091-004-02 括弧タイプ:場所 括弧位置:9 括弧始:( 括弧終:)  
カナダ・ハミルトン

### 3.2.6 括弧タイプ「所属」の例

括弧処理前

# S-ID:950101091-009

女子では、48キログ級で61連勝中の田村亮子(帝京大)に連覇の期待がかかる。

括弧処理後

# S-ID:950101091-009-01

女子では、48キログ級で61連勝中の田村亮子に連覇の期待がかかる。

# S-ID:950101091-009-02 括弧タイプ:所属 括弧位置:21 括弧始:( 括弧終:)

帝京大

### 3.2.7 括弧タイプ「同義」の例

括弧処理前

# S-ID:950101018-001

アジア・太平洋経済協力会議(APEC)は、今年五月にもソウルで情報・通信担当閣僚を集めた「APEC情報サミット」を開催する。

括弧処理後

# S-ID:950101018-001-01

アジア・太平洋経済協力会議は、今年五月にもソウルで情報・通信担当閣僚を集めた「APEC情報サミット」を開催する。

# S-ID:950101018-001-02 括弧タイプ:同義 括弧位置:13 括弧始:( 括弧終:)

APEC

### 3.2.8 括弧タイプ「文」の例

括弧処理前

# S-ID:950101067-003

秋山和慶の音楽監督就任三十周年、そして第四百回定期という記念すべき機会に演奏された、シェーンベルクの未完のオペラ《モーゼとアロン》のサントリーホールにおける演奏会形式による(しかし作曲者の指示を守り、可能なかぎりの演出、装置や照明等を備えた)公演は、この作曲家が生涯の課題とした宗教的命題と芸術表現の問題を一気呵成(かせい)劇的に演じきるものであり、「旧約」の成立の場を感動的に表現するものであった。

括弧処理後

# S-ID:950101067-003-01

秋山和慶の音楽監督就任三十周年、そして第四百回定期という記念すべき機会に演奏された、シェーンベルクの未完のオペラ《モーゼとアロン》のサントリーホールにおける演奏会形式による公演は、この作曲家が生涯の課題とした宗教的命題と芸術表現の問題を一気呵成、劇的に演じきるものであり、「旧約」の成立の場を感動的に表現するものであった。

# S-ID:950101067-003-02 括弧タイプ:文 括弧位置:86 括弧始:( 括弧終:)  
しかし作曲者の指示を守り、可能なかぎりの演出、装置や照明等を備えた  
# S-ID:950101067-003-03 括弧タイプ:読み 括弧位置:157 括弧始:( 括弧終:)  
かせい

### 3.2.9 括弧タイプ「その他」の例

括弧処理前

# S-ID:950101071-029  
地元最大の中国語日刊新聞「星島日報」(本社・香港、四万八千部、古偉凱編集局長)が、その実態を語る。

括弧処理後

# S-ID:950101071-029-01  
地元最大の中国語日刊新聞「星島日報」が、その実態を語る。  
# S-ID:950101071-029-02 括弧タイプ:その他 括弧位置:18 括弧始:( 括弧終:)  
本社・香港、四万八千部、古偉凱編集局長

### 3.2.10 例外

以下のような列挙のための数字付きの括弧は例外とし、文中に括弧を入れたままとする。

# S-ID:950105011-003-01  
改革の柱として(1)財界・大企業本位の政治にメスを入れる(2)日米安保と在日米軍の横暴に取り組む の二点を挙げた。

## 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、その文の出处情報も含む。

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。