Web 上のテキストを対象としたコーパス作成

内海 慶　　颯々野 学　　山田 薫　　谷尾 香里　　前澤 敏之
ヤフー株式会社
{kuchiumi,msassano,kayamada,katanio,tmaezawa}@yahoo-corp.jp

1　はじめに

我々は，Web 上のテキストを利用したサービスの開発を検討している．Web 上のテキストに自然言語処理技術を適用するときに問題となることの一つは，一般出版物には見られない表現が見られ，新聞などのテキストを対象とした従来の自然言語処理システムでは十分に対応できないことである．これに対処するため，我々は Web 上のテキストを対象としたコーパスの作成を進め，それと並行して，Web 上のテキストに現れる現象の整理や係り受け解析などのシステムの開発を進めていくこととした．本稿では，この取り組みの概要とコーパス作成過程で見つけた興味深い現象を報告する．

2　関連研究

ここでは，過去の日本語のコーパスや，Web 上のテキストの利用について，従来の研究を概観する．日本語の代表的なコーパスとして，京大コーパス (京都テキストコーパス) がある [3]．また，最近の注目すべき日本語コーパスのプロジェクトとして，『日本語話し言葉コーパス』[5] や，KOTONOHA プロジェクト [6, 10] がある．

Web を対象にしたコーパスの利用の研究も進みつつある．Liu らは Web 上のテキストをコーパスとして利用しようと，Web テキスト上に見られる綴りのミスなどを調べ，クリーニングを行なった後，言語処理で活用することを考えた [4]．Keller らは，検索エンジンでのヒット件数から未知の bigram の頻度数を推定しようとした [1].

これに対し，我々はもっぱら Web 上のテキストを対象として，人手で情報を付与したコーパスの作成を進める．

3　コーパス作成プロジェクトの概要

中長期的な我々の目的は，Web 上のさまざまな分野，形態のテキストのアノテーション付きコーパスを数万文用意することである．その本格的な遂行の前に，基本的な問題の洗い出し，アノテーションツールの作成，アノテーションのガイドラインの作成を進めることとした．

3.1　付与する情報

現在の我々のコーパスでは，形態素情報，文節区切り情報，係り受け情報が付与されている．形態素情報は，我々が独自に開発した形態素解析ツールにより付与し，人手の修正は行なっていない．この理由は，形態素解析の辞書や品詞体系の充実，チューニングは，本コーパス作成とは独立に行なわれているためである．将来，品詞体系の変更や辞書のエントリの変更があった場合には，形態素情報のみ入れ替える予定である．一方，文節区切り情報と係り受け情報は，ツールで付与した後，人手で修正して付与している．

文節区切り情報は，形態素ごとに {B, I} の二種類のタグを付ける．B は文節の先頭を意味し，I は文節の途中を表す [8].

係り受け情報は，文節ごとに，係り先と，通常の係り受け関係，並列関係，同格関係などの情報を区別して付与している．今回のアノテーションでは，{D，P，A} のタグを，それぞれ通常の係り受け関係，並列関係，同格関係を表すタグとして使用した．

アノテーションのガイドラインは，文献 [2, 9, 7] などを参考に設定した．図 1 に情報を付与した文の例を示す．

```
# written by 128.0.0.1 date Fri Jan 26 18:25:26 2007
# ID0000
0 5D
約 やく 約 接頭辞 冠数 * B
1 1 1 名詞 数詞 * I
年 ねん 年 接尾辞 助数 * I
前 まえ 前 名詞 名詞 * I
、 、 、 特殊 読点 * I
1 3D*2P
不安 ふあん 不安 名詞 名形 * B
と と と 助詞 助詞副詞化 * I
2 3D
期待 きたい 期待 名詞 名サ他 * B
を を を 助詞 格助詞 * I
3 5D
抱え かかえ 抱える 動詞 一段 連用形 B
ながら ながら ながら 助詞 接続助詞 * I
4 5D
ここ ここ ここ 名詞 名詞場所 * B
ハリウッド はりうっど ハリウッド 名詞 地名 * I
に に に 助詞 格助詞 * I
5 -1D
やって来 やってき やって来る 動詞 カ変 連用形 B
まし まし ます 助動詞 助動詞ます 連用形 I
た た た 助動詞 助動詞た 基本形 I
。 。 。 特殊 句点 * I
EOS
```

図 1　コーパスの例

3.2　スケジュール

このコーパス作成プロジェクトは，2006 年 7 月から始め，2007 年 1 月末まで初期の試行錯誤を行なった．2 月以降，対象テキストを増やすなどして本格的に取り組む予定である．まず，800 文程度を Web 上から選び，文節区切りのアノテーションを行ないながら，ガイドライン，文節区切りタグ付与プログラムの整備を行なった．次に，係り受け解析ツール，アノテーションツールの整備と並行して， 2,500 文程度の文に係り受け情報を付与した．現在までおよそ 3,000 文のコーパスが作成できた．2007 年中には 10,000 文規模のコーパスを作成する予定である．

3.3　アノテーションツール

アノテーション作業にあたり，アノテーションツール Y!CAT の開発を行った (図 2)．以下では，開発したツールについて説明する．

3.3.1　特徴

Y!CAT は，Web ベースのマルチユーザ型アノテーションツールである．ツールは，導入の迅速さを考慮して，Ajax による UI をベースとした実装にした．このようなブラウザ経由で利用できるツールの利点として，導入と開発・メンテナンスの容易さが挙げられる．

アノテーターは，各自の PC からブラウザを通じて Y!CAT にログインし，アノテーション作業を行う．作成するコーパスは，アノテーターごとに管理される．

次に，Y!CAT の機能について説明する．

3.3.2　機能

Y!CAT では，アノテーターの作業支援のために次のような機能を実装している．

1. パーサー出力結果のツリー表示
   パーサー出力を直感的に理解しやすくする．
2. 係り先情報の変更に応じた自動ツリー表示
   アノテーターの修正にリアルタイムで応答して，係り受け構造のツリーを表示する．
3. ラベル選択
   手入力によるタイプミスを防止する．
4. 文節ごとの形態素解析結果表示
   文節切りの誤りや形態素解析の誤りの確認を行う．
5. 文節に対するメモ
   文節切り誤りや形態素解析結果の誤りを記述する．複数の係り先候補がある場合にも記述する．
6. 文に対するメモ
   文全体で見られた特徴や，アノテーターが気づいた点を記述する．
7. 検索機能
   コーパスに対する文字列検索を行う．
8. ID 指定によるジャンプ
9. 係り受け関係の非交差条件を破る場合の自動修正
   アノテーターが非交差条件を破るような修正を行った場合に，係り先情報を自動修正する．

4　Web 上のテキストでの興味深い言語現象

我々が Web テキストを対象に係り受け関係を付与する際，新聞などの一般的なテキストには現れにくい文が見られた．本稿ではそれを，表記的特徴と文体の特徴という二つに分類し整理した．以下で，表記的特徴と文体の特徴について説明する．

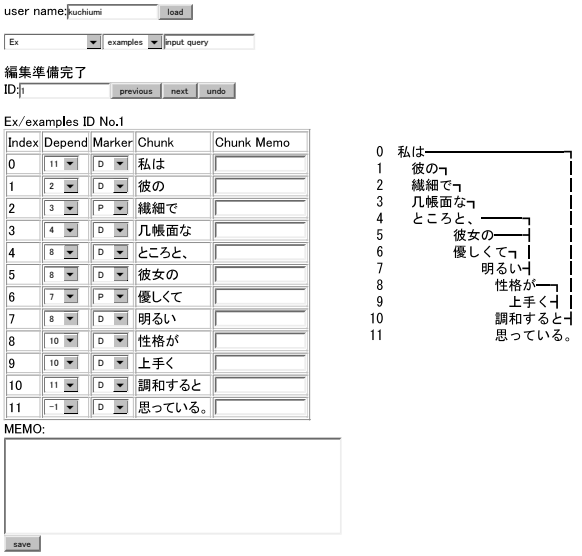

図 2　Y!CAT 画面

4.1　表記的特徴

Web 上のテキストで頻繁に観察される表記と，機械処理を行うにあたっての問題について説明する．

4.1.1　句読点などの記号の用法が異なるケース

句読点などの記号を，本来の句読点としての用法から外れて，言外の意味などの感情を表現するために用いている場合が見られた．以下に，コーパス作成時に見られた句読点記号の標準的ではない用法を挙げる．

(1)　日本 (アジア?) をモチーフにしているようです。

(2)　はずれ！ってことはありません。

上記のような文の場合，機械的な処理で正しく文切りを行うためには，単純な句読点記号によるパターンマッチだけでは上手く行かず，工夫が必要となる．

4.1.2　新聞などでは用いられない記号が用いられるケース

句読点以外の記号が，感情を表現する目的で，一般出版物の紙面とは異なる用法で使用されている文が見られた．以下に，感情を表現している特殊な記号用法を含む文例を示す．

(3)　値段も普通のより高め（´Д｀；）

(4)　まさに搾りたて♪の美味しさです。

(5)　アナログしかないので、更に画質悪いと思われ orz

(6)　すごく美味しいので、買う価値アリ㊤㊤

(7)　濃厚だけどパサついてて、自分的には★☆☆です。

例文 (7) は，形容詞相当として記号が使われている．(3)-(6) と異なり，記号を取り除くと文が成立しないため，注意が必要である．

4.1.3　文の中に括弧で文を挿入するケース

文中に単語の説明などが挿入される現象は新聞やニュース記事でも見ることが出来る．しかし，新聞やニュース記事では表記法が限定されており，その記法を考慮して機械処理を行うことが出来る．一方で，Web 上のテキストでは特定の表記法はないためため，予め挿入文をパターン化して判別することは困難である．以下に，コーパス作成時に見られた特徴的な挿入文を示す．

(8)　しかし飲み物のメニュー（＝価格が明記してあるもの）がなかったのは、ヒヤヒヤものの経験でした。

(9)　赤坂氏（30）が結婚することが 12 日、事務所より発表された。

上記は名詞と助詞の間に文の挿入が行われたケースである．(8) の文では，助詞までを 1 つの文節とした場合，挿入文自体の係り受け関係が付与できない．(9) の文は，助詞までを 1 つの文節として問題がない挿入文の例である．挿入文に対して係り受け関係を付与する場合，適切な文節切りが難しく，どうアノテートするか問題となる．

4.1.4　仮名表記を変形させるケース

Web 上のテキストを見ると，形容詞などの仮名表記に数多くのバリエーションが見られた．以下に，コーパス作成時に現れた典型的な仮名表記の変形を示す．

(10)　前菜がさっぱりしていてオイシかった♪

(11)　しかし新作メニューは、ことごとくマズイ。

(12)　ホント綺麗でびつくり！！

(13)　そんなこと聞いてない YO!

上記のように，平仮名や漢字で書くべきところを片仮名や英字で書いた場合，通常の形態素解析では誤解析の可能性が高くなる．

表 1 表記の種類と検索ヒット数
[Yahoo!検索 (2007. 1. 25) 調べ]

Word	検索ヒット数
おいしい	47,800,000
美味しい	39,900,000
オイシイ	1,750,000
オイシい	26,200

標準的な表記に対して，標準的ではない表記がどの程度 Web 上で用いられるかを調べるため，「美味しい」について Web 検索を行い，そのヒット数を比較した．検索件数の比較を表 1 に示す．

表 1 を見ると，全表記の検索ヒット数に対して，標準的ではない表記の検索ヒット数が占める割合は約 2% であった．しかし，件数を見ると標準的ではない表記をする文書数も 180 万件程度あることが分かる．

標準的ではない表記に対する対処としては，

1. 全体の割合からすれば少ないと判断して捨てる
2. 変化のパターンを抽出して標準的な語へ変換する
3. 標準的な語とは区別して対処する

などが考えられるが，具体的にどの方向へ進むかは検討中である．

4.2 文体の特徴

口語体の文では，終止形・連体形の解釈に注意が必要な例があった．

(14) シャッフル・バッテリーの減りが遅くて、月 1 回充電するくらいだったけど、ナノ・ヤバイ減りが早すぎ。

この文の「ヤバイ」は，意味から考えると連体形ではなく，直後に句点は存在しないが，終止形であると考えられる．それ以外にも，「ヤバイ」が感動詞的に挿入されているとの解釈もありうる．

5 おわりに

本稿では，我々が進めているプロジェクトの概要と，アノテーション作業で見つけた Web 上のテキストに現れる言語現象を，表記的特徴と文体の特徴という二つの観点から整理，分類し，報告した．

本稿で述べた種々の問題に対して，適切な解決法を探ることが，我々の今後の課題である．

参考文献

[1] Frank Keller and Mirella. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, Vol. 29, No. 3, pp. 459 – 484, 2003.

[2] 黒橋禎夫, 居蔵由衣子, 坂口昌子. コーパス作成の作業基準 version 1.8., 2000.

[3] Sadao Kurohashi and Makoto Nagao. *Treebanks*, chapter 14, Building a Japanese parsed corpus while improving the parsing system, pp. 249 – 260. 2003.

[4] Vinci Liu and James R. Curran. Web text corpus for natural language processing. In *Proc. of EACL 2006*, pp. 233 – 240, 2006.

[5] 前川喜久雄. 『日本語話し言葉コーパス』の概観 version. 1.0, 2004.

[6] Kikuo Maekawa. Kotonoha, the corpus development project of the national institute for Japanese language. In *Proc. of the 13th NIJL International Symposium*, pp. 55 – 62, 2006.

[7] 西川賢哉, 小椋秀樹, 相馬さつき, 小磯花絵, 間淵洋子, 土屋菜穂子, 斉藤美紀. 文節の仕様について version 1.0., 2004. http://www2.kokken.go.jp/csj/public/manuals/bunsetsu.pdf.

[8] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Proc. of the Third Workshop on Very Large Corpora*, pp. 82 – 94, 1995.

[9] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision, 2nd printing), 1990.

[10] 山崎誠, 前川喜久雄, 田中牧郎, 小椋秀樹, 柏野和佳子, 小磯花絵, 間淵洋子, 丸山岳彦, 山口昌也, 秋元祐哉, 稲益佐知子, 吉田谷幸宏. 代表性を有する現代日本語書き言葉コーパスの設計. 言語処理学会第 12 回発表論文集, pp. 440 – 443, 2006.