構文・照応・評判情報つきブログコーパスの構築

橋本 力* 河原 大輔† 黒橋 禎夫‡ 新里 圭司§ 永田 昌明¶
*†‡ 情報通信研究機構 ‡§ 京都大学 ¶ NTT コミュニケーション科学基礎研究所

1 はじめに

近年、ブログを対象とした情報アクセス・情報分析技術が盛んに研究されている。我々は、この種の研究の基礎データの提供を目的とし、249 記事、4,206 文からなる、解析済みブログコーパス（以下、KNB コーパス[1]）を構築した。主な特長は次の 4 点である。

(1) a. 自動＋人手チェックによる文区切り
b. 京大コーパス互換の、形態素、係り受け、格・省略・照応、固有表現アノテーション
c. 人手による評判表現アノテーション
d. アノテーションを可視化した HTML ファイルの提供（図 1）

記事は、大学生 81 名に「京都観光」「携帯電話」「スポーツ」「グルメ」のいずれかのテーマで執筆してもらうことで収集した。

解析済みブログコーパスを構築する際の課題は次の通りである。

(2) a. 不明瞭な文区切りへの対応
b. 括弧表現への対応
c. 誤字、方言、顔文字等、多様な形態素への対応

本稿では、KNB コーパスの全容とともに、いかに上記の課題に対応しつつコーパスを構築したかについて述べる。

2 関連研究

日本語の代表的な解析済みコーパスとして、京都大学テキストコーパス [1][2]（以下、京大コーパス）と NAIST テキストコーパス [3] がある。前者は、新聞記事を対象に、4 万文に対して形態素・構文情報を、5,000 文に対して格関係、照応・省略関係、共参照の情報を付与したものである。[3] 後者は、京大コーパスの 4 万文に対して、述語と表層格の関係、事態性名詞と表層格の関係、名詞間の共参照関係の情報を付与したものである。

一方、KNB コーパスは、ブログ記事を対象に、京大コーパスと同様のフォーマットで同様の言語情報（と、それらに加えて固有表現情報と評判情報）を付与している。ただし、後述するように、ブログの文章は話し言葉の特徴を有しているため、一部 KNB コーパス独自の仕様を新たに策定した。これらの独自仕様には、話し言葉を対象とした代表的な解析済みコーパスである日本語話し言葉コーパス [2][4]（以下、CSJ）の仕様と類似したものが含まれている。

ブログを対象とした既存のコーパスとして、ICWSM 2009 Spinn3r Blog Dataset[5]がある。これは 44,000,000 記事からなる大規模なものだが、係り受けや照応等の言語情報は付与されていない。

表 1: テーマごとの記事数と文数

テーマ	記事数	文数
京都観光	91	1,518
携帯電話	79	1,278
グルメ	57	888
スポーツ	22	522

3 コーパスの仕様

3.1 規模

コーパスは 249 記事、4,206 文、約 68,000 形態素から成る。記事のテーマは「京都観光」「携帯電話」「スポーツ」「グルメ」のいずれかである。テーマごとの記事数、文数は表 1 の通りである。

これらの記事は大学生 81 名によって書かれた。

3.2 アノテーション

KNB コーパスには、文区切り、括弧表現、形態素、係り受け、格・省略・照応、固有表現、評判表現に関

[1] **K**yoto University and **N**TT **B**log コーパス
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
[3] IREX'99 において、京大コーパス中の 1 万文に対する固有表現アノテーションが配布されている。
[4] http://www.kokken.go.jp/katsudo/seika/corpus/
[5] http://www.icwsm.org/2009/data/

KN002_Keitai_1-1-3-01

係り受け	格・省略・照応、固有表現	評判表現
悩んだ┐	一人称:ガ:文外, 機種:ニヲイテ:1文前	
末、──────┐		
カシオの──────┐ │	カシオ:ORGA	
G'z　oneと┐ │ │	G'z　one:ARTI	
いう、──────┤ │		
衝撃や─P │ │		衝撃や水濡れに強いのがウリの機種:採否＋
水┤ │		
濡れに┐ │ │	水:ニ, 電話:ガ:1文前	
強い┐ │ │	衝撃:ニ, 濡れ:ニ, 電話:ガ:1文前	
のが┐ │		
ウリの┤ │	の:ガ, 電話:ノ:1文前	
機種に┤	カシオ:修飾, G'z　one:トイウ, ウリ:修飾, 機種:=ヲ:1文前	
決めました。	機種:ニ, 一人称:ガ:文外, 末:時間	

表出形	読み	原形	品詞	活用
悩んだ	なやんだ	悩む	動詞	子音動詞マ行 タ形
末	すえ	末	名詞 時相名詞	
、	、	、	特殊 読点	

図 1: アノテーション可視化 HTML ファイル

するアノテーションを付与した。このうち、形態素、係り受け、格・省略・照応、固有表現のアノテーションは京都テキストコーパスに準拠したものである。

3.2.1 文区切り

ブログ記事では、(3) にあるように、文の終わりが不明確な場合がある。[6]

(3) a. なぜか清水寺に着きました笑 [EOS]
b. これに決めた !! ⬇ [EOS]
と思ったら… ⬇ [EOS]
なんと品切れ。 [EOS]
c. 京都のほうだったような… [EOS] まぁ観光スポット多いですもんね?京都は。 [EOS]

そこで他のアノテーションに先立ち、「文」の仕様を策定し、文区切りを明示する（一文一行にする）アノテーションを施した。(4) にその仕様の一部を挙げる。

(4) a. 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。
例）「散歩 ?? かな。」
b. 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。
例）「そんな日本語ないか。笑」

CSJ では、対象が話し言葉であり、文の終わりが不明確であるという特徴がより顕著である。そこで「節単位」という「文」に概ね相当する単位を規定し、我々と同様に文区切りアノテーションを施している。

[6] [EOS] はアノテートされた文区切りを、⬇は元の記事における改行を表す。

3.2.2 括弧表現

京大コーパスでは括弧表現は削除されていたが、ブログ記事の括弧表現は、新聞記事と比べると、本文と密接不可分な内容のものが多く無視できない。(5) は KNB コーパスの括弧表現の例である。

(5) a. 貴重な（まぁどのへんが貴重なのかはわからないけど）時間を無駄にしてしまう。
b. どでかい神楽松明（激しく燃えている！）を担いで、狭い鞍馬街道をどこからともなく練り歩き出す。

一方で、ブログ記事の括弧表現は多種多様で、文内に埋め込まれたままだと、係り受け等のアノテーションが困難になることが予想される。そこで、括弧表現を文中から取り出して、一つの独立した文とした。(7) は (6) から括弧表現を抽出したものである。

(6) # S-ID:KN012_Gourmet_6-1-23
ここもＦｏｏｄ（パンメインでカレーとか）の量は少なかったなー

(7) a. # S-ID:KN012_Gourmet_6-1-23-01
ここもＦｏｏｄの量は少なかったなー
b. # S-ID:KN012_Gourmet_6-1-23-02 括弧タイプ:例示 括弧 位置:7
パンメインでカレーとか

「#」の行は文 ID を表す。抽出され一文に昇格した括弧表現は、元の文の直後に置かれ、新たに文 ID が与えられる。なお、括弧表現の元の文における位置情報が記録されており、復元も可能である。

表 1 では、括弧表現も一文としてカウントしている。

3.2.3 形態素

括弧表現を考慮した文区切りアノテーションの後、形態素区切り、読み、原形、品詞、活用に関するアノテーションを付与した。これらは品詞・活用体系、フォーマットともに京大コーパスに準拠しているが、次のようなブログの特徴に対応するため仕様を拡張した。

(8) a. 誤変換、脱字、衍字
b. 口語的表現（方言、外国語、擬音／擬態語、意図的な言い淀み／言い直し）
c. 創造的表現（記号、Web で頻出のスラング）

a. と b. は話し言葉でも見られる特徴だが、c. はブログ（あるいは Consumer Generated Media）に特有のものと言える。

誤変換、脱字、衍字 これらは話し言葉の言い誤りに相当する。(9) に例を挙げる。

(9) a. 誤変換：「内蔵」が「内臓に」に
b. 脱字：「早めに行かないと」が「早めにかないと」に
c. 衍字：「高級な料亭」が「高級なり料亭」に

これらには次のように対応した。

(10) a. （元の表現を残しつつ）正式な書き方・表現に基づいてアノテーションする。例えば「早めにかないと」なら「早めに行かないと」としてアノテーションする。
b. メモ欄に誤りフラグと正しい書き方を記載。

CSJ においても、言い誤りは正しい形式に捉え直してアノテーションされている。

口語的表現 方言や外国語、擬音／擬態語、意図的な言い淀み／言い直し等がこれに該当する。(11) に例を挙げる。

(11) a. 方言：「面倒やん」
b. 外国語：「（祇園界隈へ）GO」
c. 擬音／擬態語：「ピリリリリリリ」
d. 意図的な言い淀み：「ぎゅうにゅ･･･にゅ･･･」

方言では活用をどう記述するかが特に問題となる。我々は、京大コーパスとの互換性と文法記述の正確性を最大限確保するため、形態素解析器 JUMAN[7] の活用記述法に準拠して、既存の活用に該当しない方言に対して、新たな活用を定義した。例えば「面倒やん」なら形容詞（面倒だ）ヤ列基本形となる。

外国語に対しては次のように対応した。

(12) a. 原則的にサ変名詞か形容詞としてとらえる。
b. それ以外は普通名詞か記号としてとらえる。

例えば「（祇園界隈へ）GO」はサ変名詞、「GREAT VIEW」なら「GREAT」がナ形容詞で「VIEW」がサ変名詞となる。

擬音／擬態語は全て副詞とした。

意図的な言い淀み／言い直しは、未定義語として、メモ欄に「言い淀み」あるいは「言い直し」と記載した。例えば (13) の下線部が言い淀みとなる。

(13) 牛乳を入れて･･･ぎゅうにゅ･･･にゅ･･･

結局、これらの表現は元の形のままでアノテーションされている。CSJ においても、融合、省略、フィラー、断片化といった口語表現特有の現象を、元の表現そのままでアノテーションしており、我々の方針と一致している。

創造的表現 顔文字等の記号や、「サーバ」を意味する「鯖」等の Web 上で多用されるスラングは、ブログ特有の表現といえる。

(14) に記号に関する仕様の一部を挙げる。

(14) a. 顔文字：一語とする。品詞は「記号」。
b. 同じ記号の連続：一語とする。品詞は「記号」。
c. 「～」「ー」：直前の語の一部とする。

スラングも、他の表現と同様に京大コーパスとの互換性を最大限保つよう配慮した。例えば上記の「鯖」は、普通名詞として扱った。

3.2.4 係り受け

京大コーパスに準拠する形式で、文節の認識と文節間の係り受け関係のアノテーションを付与した。

CSJ では話し言葉特有の現象に対応すべく独自の仕様を設けている。(15) にその一部を挙げる。

(15) a. 「倒置」に対応すべく、右から左への係りを認める。
b. 「言いさし」「ねじれ」等の現象に対応すべく、係り先のない文節を認める。

KNB コーパスでは、京大コーパスとの互換性を重視し、上記のような特殊な仕様は避けた。つまり (15) に相当する現象に対しても、左から右へ、最も適切と思われる文節に係るようにアノテーションした。

[7] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

3.2.5 格・省略・照応

京大コーパスに準拠する形式で、格・省略・照応のアノテーションを付与した。(16)、(17)、(18) に、格、省略、照応の例を挙げる。

(16) Food は量が 少ない。(ガ格:量, 外の関係:Food)

(17) 4, 5 回しか 行ったことないけど。(ガ格:一人称)

(18) a. 父と野球。
b. 父 は野球が好きだった。(父 = 一文前)

3.2.6 固有表現

IREX'99 準拠の固有表現アノテーションを付与した。(19) に例を挙げる。

(19) a. 京都 を回ってみようと思います☆ (京都:LOCATION)
b. 近鉄 ファンだった。(近鉄:ORGANIZATION)

3.2.7 評判表現

評判アノテーションは、何らかの評判表現を含む文に対して次の情報を付与することで行った。

(20) a. 評判表現
b. 評判対象
c. 評判タイプ (当為、要望、感情 +/−、批評 +/−、メリット +/−、採否 +/−、出来事 +/−)
d. 評判保持者

(21)(21) に例を挙げる。

(21) おかきやせんべいの店なのだが、これがオイシイ。
a. 評判表現 : オイシイ
b. 評判対象 : おかきやせんべい
c. 評判タイプ : 批評 +
d. 評判保持者 : [著者]

(22) 貧乏人臭い、なんか怪しげな人っぽいといった類のものです。
a. 評判表現 : 貧乏人臭い、なんか怪しげな人っぽいといった類のものです。
b. 評判対象 : [プリペイドユーザ]
c. 評判タイプ : 批評 −
d. 評判保持者 : [不定]

文中に存在しない対象、保持者は [...] でマークされている。

結果として、約 48%の文に何らかの評判表現が含まれていた。

3.3 アノテーション可視化 HTML ファイル

以上のように KNB コーパスのアノテーションは多岐にわたり、そのままでは (人間にとって) 可読性が低い。そこで、アノテーションを可視化した HTML ファイルを別途用意した (図 1)。

4 構築手順

KNB コーパスの構築は、概略、次の手順で行った。

1. 記事収集
2. 文区切り／括弧抽出 (自動 + 人手)
3. 形態素／係り受け／格・省略・照応／固有表現 (自動 + 人手)
4. 評判表現 (人手)

記事収集は、独自にブログサーバを設置し、大学生にアルバイトとして記事を書いてもらうことで収集した。手順 2. と 3. は、自動でアノテーションした後、人手修正を施した。特に手順 3. は、京大コーパスと同様、JUMAN ／ KNP[8]を用いた。一方、手順 4. は全て人手で行った。手順 3. と 4. は、アノテーション基準の見直しと再アノテーションというサイクルを何度か繰り返して行った。

5 おわりに

本稿では、我々が構築した、他に類を見ない、解析済みブログコーパスについて報告した。特に、ブログ特有の現象とそれらの正確な記述、京大コーパスとの互換性の重視について述べた。

謝辞

評判情報のアノテーションについて御協力いただいた情報通信研究機構知識処理グループに感謝いたします。

参考文献

[1] 河原大輔, 黒橋禎夫, 橋田浩一. 「関係」タグ付きコーパスの作成. 言語処理学会 第 8 回年次大会, pp. 495–498, 2002.

[2] 前川喜久雄, 籠宮隆之, 小磯花絵, 小椋秀樹, 菊地英明. 日本語話し言葉コーパスの設計. 音声研究, Vol. 4, No. 2, pp. 51–61, 2000.

[3] 飯田龍, 小町守, 乾健太郎, 松本裕治. NAIST テキストコーパス : 述語項構造と共参照関係のアノテーション (解析・対話). 情報処理学会研究報告. 自然言語処理研究会報告, pp. 71–78. 社団法人情報処理学会, 2007.

[8] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html