〈論文紹介〉 小西光,浅原正幸,前川喜久雄「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」言語処理学会誌『自然言語処理』20(2): 201-222. (2013)

著者	小西 光
雑誌名	国語研プロジェクトレビュー
巻	5
号	1
ページ	54-56
発行年	2014-06
URL	http://doi.org/10.15084/00000771

国語研プロジェクトレビュー
NINJAL Project Review　Vol.5 No.1 pp.54–56（June 2014）

〈論文紹介〉

小西光，浅原正幸，前川喜久雄
「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」
言語処理学会誌『自然言語処理』20(2)：201–222.（2013）

小西 光

本稿は，テキスト中に記述される事象の生起時刻を推定するための重要な手がかりとなる時間情報表現を抽出し，その生起順序を解析することを目的に，アノテーション（タグづけ）基準の再定義と『現代日本語書き言葉均衡コーパス』（BCCWJ）コアデータの一部に付与した時間情報アノテーションの問題点と課題を考察した。時間情報表現を含む数値表現の抽出は，固有表現抽出の部分問題として解かれてきた。英語においては，評価型国際会議が開かれ，時間情報表現のテキストからの切り出しだけではなく，曖昧性解消・計算機可読な形式への正規化のための様々な手法が提案されている。さらに，時間情報と事象とを関連づけるアノテーション基準 TimeML の定義や新聞記事にアノテーションを行ったコーパス TimeBank の整備が進んでいる（James Pustejovsky et al. 2003）。一方，日本語では時間情報表現の抽出技術は未だ確立していない。したがって日本語の事象の生起順序関係解析のための基礎データは現在のところ存在していない。本稿では，TimeML の時間情報表現を表す <TIMEX3> タグに基づいたアノテーション基準を日本語向けに再定義し，BCCWJ の一部データにアノテーションを実施した。また今後事象の生起時刻を推定するために必要な課題を考察した。

日本の言語資源整備は，実データを作成せずに標準化活動を行うものと実データを作成するが標準化活動を無視して行うものとに二極化しているという問題点がある。本研究は，標準化に適した ISO-TimeML に準拠する日本語版 <TIMEX3> 時間情報表現アノテーション基準を検討・策定し，実データへのアノテーションを実施した点およびコーパスアノテーションにおける標準化活動の点から重要である。

アノテーション対象となる時間情報表現は，日付表現・時刻表現・時間表現・頻度集合表現の 4 種類である。図 1 にアノテーション事例を示す。日付表現は「一九二九年二月」「前日」のような日暦に焦点をあてた表現である。時刻表現は「午前十時ごろ」「午後六時ごろ」「昼」「九日昼」のような一日のうちのある時点に焦点をあてた表現である。日付表現と時刻表現の区別は時間軸上の粒度の区別でしかない。便宜上不定の現在を表す「今」という表現は時刻表現に分類する。時間表現は「その間」のような時間軸上の両端に焦点をあてておらず，期間を表すことに焦点をあてている表現である。頻度集合表現は「毎日」のような複数の日付・時刻・時間に焦点をあてた表現である。この分類は，解析の方便のために導入したものである。時間軸上一つもしくは複数の時点・時区間を表現するものをアノテーション対象で

（出典）書籍サンプル PB59_00001（優先順位 00003）

<sentence type="quasi"> <TIMEX3 @tid="t1" @type="DATE" @value="2003-10-20" @valueFromSurface="2003-10-20" @definite="true"> 二〇〇三年十月二十日 </TIMEX3> <TIMEX3 @tid="t2" @type="DATE" @value="2003-10-W3-1" @valueFromSurface="XXXX-WXX-1" @definite="true"> 月 曜 日 </TIMEX3> </sentence> <br type="automatic_original"/> <sentence type="quasi"> <TIMEX3 @tid="t3" @type="TIME" @value="2003-10-20T17:30:XX" @valueFromSurface="XXXX-XX-XXT17:30:XX" @definite="false"> 午後五時三十分 </TIMEX3> </sentence> <br type="automatic_original"/> <blockEnd/> <paragraph> <sentence> ステイシーはだらけた姿勢でモニターの前に陣取り、白黒の画像に見入っていた。</sentence> <sentence>彼女は伸びをし、腕時計に目をやった。</sentence> <sentence> <TIMEX3 @tid="t4" @type="DURATION" @value="PT2H30M" @valueFromSurface="PT2H30M" @definite="true"> 二時間半 </TIMEX3> で収穫ゼロ。</sentence>

図 1　アノテーション例

ある時間情報表現とする。日本語適応時に問題となった和暦や「上旬」「盆」など日本語特有の表現については，日本語に限定した形式での正規化を行った。

現在のアノテーション基準では <TIMEX3> タグの入れ子を許さない。また，次に挙げる属性によって情報を付与している（@tid, @type, @value, @valueFromSurface, @freq, @quant, @mod etc.）。また作業は統制手法を取るためにペアプログラミングのような手法を採用し，表 1 にある分量のデータにアノテーションを実施した。

アノテーションした情報について，時間情報表現の正規化の観点から分析を行った。

日付表現（“DATE”）・時刻表現（“TIME”）については，時区間特定可であるものの多くが，人手による曖昧性解消が行われていた。このことから本アノテーションの目的とする時間表現の正規化作業の重要性がうかがえる。白書や新聞など，出版年・発行年月日が明らかであるものほど曖昧性解消がよく行われる傾向があった。一方 Yahoo! 知恵袋や雑誌は，お店の営業時間など時間軸上の特定の時刻を表現しないものが多かった。

時間表現（“DURATION”）と頻度集合表現（“SET”）については，時間軸上の時区間を特

表 1　作業対象データ

レジスタ		ファイル数	うち時間表現あり	文数	うち時間表現あり	長単位形態素数	短単位形態素数
白書	(A)	17	16 (94%)	1,439	405 (28%)	40,690	58,336
書籍	(A)	25	25 (100%)	2,568	289 (11%)	50,257	57,929
新聞	(A, B)	110	110 (100%)	5,582	1,562 (28%)	88,733	116,834
Yahoo! 知恵袋	(A)	518	250 (48%)	3,479	488 (14%)	51,240	60,086
雑誌	(A)	23	23 (100%)	3,066	413 (13%)	49,715	59,372
Yahoo! ブログ	(A)	257	198 (77%)	3,986	765 (19%)	53,333	63,459

定することを目的とせず，実際に時間軸上の時区間に写像する際には，日付・時刻表現や事象表現との時間的順序関係（TimeML の <TLINK>）を定義することが必要になる（Masayuki Asahara et al. 2013）。

今後，TimeML で行われている事象表現と時間表現間の時間的順序関係（TimeML 中の <TLINK>）付与を進めていきたい。そのためには，対象となる事象表現の策定，事象表現に対する分類，テンス・アスペクト体系の整備，節間の関係定義など解決すべき問題は山積している。今後 TimeML に準じた事象表現に対するアノテーションを行い，最終目標である事象表現に対する時間情報付与の研究へと進んでいきたい。

●参照文献●

Asahara, Masayuki, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa（2013） BCCWJ-TimeBank: Temporal event information annotation on Japanese text. *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation*, 206−214.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo（2003） The TimeBank Corpus. *Proceedings of Corpus Linguistics 2003*, 647−656.

小西　光（こにし・ひかり）

国立国語研究所コーパス開発センター プロジェクト非常勤研究員。修士（文学）（上智大学）。2011 年 8 月より現職。『日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』『日本語超大規模コーパス』の整備に携わる。