『現代日本語書き言葉均衡コーパス』に対する時間情報表現アノテーション

小西 光 †　　浅原 正幸 †　　前川 喜久雄 †‡
国立国語研究所　コーパス開発センター †/言語資源研究系 ‡
{hkonishi, masayu-a, kikuo} -at- ninjal.ac.jp

1　はじめに

情報検索や情報抽出において、テキスト中に示される事象を実時間軸上の時区間もしくは時点に関連づけることが求められている。Web 配信されるテキスト情報に関しては、文書作成日時 (Document Creation Time: DCT) が得られる場合、テキスト情報と文書作成日時とを関連づけることができる。しかしながら、文書作成日時が得られない場合や、文書に記述されている事象が起きる日時が文書作成日時と乖離する場合には他の方策が必要である。テキスト中に記述されている時間情報解析の精緻化が求められている。

我々は『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese; 以下 "BCCWJ") に対して時間情報表現の正規化情報アノテーションを行った。言語資源アノテーションの国際標準の一つである TimeML [1] の ⟨TIMEX3⟩ タグ相当の情報を日本語の言語資源に対して、レジスタ横断的に行った。以下ではアノテーション作業の概要について報告する。

2　国際標準 ISO-TimeML

国際標準化機構 (International Organization for Standardization: ISO) の標準化技術委員会 (Technical Committee) TC 37 は "Terminology and other language and content resources" と題し、言語資源に関するさまざまな標準化を提案している。そのなかに分科会 (Structure of the committee) が四つ設定されているが、TC 37/SC 4 [1]が言語資源管理 (Language resource management; LRM) に関する国際規格の規定を行っている。TC 37/SC 4 はさらに作業部会を六つ設定しており、さまざまな形式・出自の一次言語データに対するアノテーションや XML に代表される汎用マークアップ言語に基づくアノテーションの表現形式についての仕様記述言語を設計している。このうち TC 37/SC 4/WG 2 (Semantic annotation) がアノテーションと表現方法を議論する作業部会である。TimeML 開発者は、作業部会 TC 37/SC 4/WG 2 と連携を取りながら、時間情報表現と事象表現に関するアノテーション基準 Semantic Annotation Framework(SemAF)-Time (ISO-24617-1:2012) を 2012 年に正式に制定した。実際にアノテーションをしない作業部会がトップダウンに基準を決めることが多いなか、ISO-TimeML は数少ないコミュニティがボトムアップに策定した (Community driven) 基準であることが知られている。

この国際標準の中には、実世界上の特定のモノもしくはコトに関係づけられる言語横断的な標準化が有効なアノテーションと、言語の表現形態・表現機能のような言語横断的な標準化がそぐわないアノテーションとが混在している。我々は 2006 年より TimeML 開発者から TimeML 関連の情報を得ながら時間情報表現アノテーションと事象表現アノテーションに取り組んできた。標準化に適した時間情報表現アノテーションと、標準化に適さない事象表現アノテーションを切り分けたうえで、前者について ISO-TimeML に準拠する日本語版 ⟨TIMEX3⟩ アノテーション基準を検討し策定した。この部分が本研究の内容に相当する。一方、後者についてはモダリティが豊かな日本語の事象表現を国際標準に合わせてアノテーションすることが困難であり、別の方策 [4] でアノテーションする。

3　TimeML ⟨TIMEX3⟩ タグに基づいた日本語時間情報アノテーション

本節では日本語時間情報表現に対するアノテーション基準の概略を示す。アノテーション基準は、TimeML [1] ⟨TIMEX3⟩ タグの仕様に準拠している。以

[1] http://www.tc37sc4.org/

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

例 1 : ⟨TIMEX3⟩ タグに基づく日本語時間情報アノテーション (出典) PB59_00001

⟨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=”true”⟩ 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”⟩ 二時間半 ⟨/TIMEX3⟩ で収穫ゼロ。⟨/sentence⟩

下、⟨TIMEX3⟩ のタグの日本語適応について説明する。

アノテーション対象は日付表現・時刻表現・時間表現・頻度集合表現の 4 種類である。例 1 にアノテーション事例を示す。

日付表現は「二〇〇三年一〇月二〇日」「月曜日」のような日暦に焦点をあてた表現である。時刻表現は「午後五時三十分」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」という表現を時刻表現に分類する。時間表現は「二時間半」のような時間軸上の始点と終点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。頻度集合表現は上の事例（例 1）には出現しないが、例えば「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ある時間情報表現とする。

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。日付・時刻表現の線形結合はこれを一つの日付・時刻表現として切り出す。例えば「九日昼」のように日付表現と時刻表現が連接する場合には一つの時刻表現として切り出す。

⟨TIMEX3⟩ タグには正規化のための様々な情報が含まれている。ここで、属性のうち @tid, @type, @value, @valueFromSurface, @freq, @quant, @mod を概説する。また、作業・分析用に導入した @definite について説明する。

@tid 属性は一文書中の各時間情報表現に付与される識別子である。各時間情報表現を一意に同定するために用い、同一指示、参照、事象表現との時間的順序を表す際に用いる。

@type 属性は DATE, TIME, DURATION, SET の四つの値を持つ。それぞれ日付表現・時刻表現・時間表現・頻度集合表現を意味する。

@value 及び @valueFromSurface 属性は時間情報表現が含意する日付・時刻・時間の値を表す。値として ISO-8601 形式を自然言語表現向けに拡張したものを用いる。このうち @value は文脈情報を用いて正規化を行った値を付与し、@valueFromSurface 属性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。

ここで @value と@valueFromSurface 属性の違いについて例 1 を用いて説明する。「二〇〇三年十月二十日」という定時間表現は、文脈を用いなくても表層の文字列から時間軸上に一意に曖昧性解消ができ、@value と @valueFromSurface ともに “2003-10-20” と正規化する。「月曜日」という表現に対して、文脈情報からそれが 2003 年 10 月の第三月曜日であるとわかる場合には @value として “2003-10-W3-1”、@valueFromSurface として “XXXX-XX-WX-1” を記述する （ここで属性にわりあてる値の詳細については文献 [3] を参照すること）。定時間情報表現は @value と @valueFromSurface の値は同じになるが、不定時間情報表現は同じになるとは限らない。

@freq, @quant 属性は頻度集合表現に付与される頻度情報及び量化子情報である。@mod 属性は時間情報表現のモダリティを表す。例えば 「2000 年以前」をアノテーションするために@mod 属性に ON_OR_BEFORE という値をわりあてることにより「以前」というモダリティを表現する。それぞれのラベルの詳細については文献 [3] を参照すること。

@definite 属性は “true”, “false” のいずれかの値を持ち、@value 属性が、文脈情報により定時間情報が得られる時間情報表現は “true” の値を持ち、その他の時間情報表現は “false” の値を持つ。言い換えると、日付・時刻表現が時間軸上の特定時区間に写像できる場合と時間・頻度集合表現の時間幅が特定できる場合に “true” の値を持ち、そうでない場合に “false” の値を持つ。例 2 はともに @definite が “false” の

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

例 2：@value と @valueFromSurface が異なるが定時間情報が復元できない例

〈TIMEX3 @type="DATE" @value="XXXX-04" @valueFromSurface="XXXX-04" @definite="false"〉４月〈/TIMEX3〉の予定ですが
〈TIMEX3 @type="DATE" @value="XXXX-04-10" @valueFromSurface="XXXX-XX-10" @definite="false"〉１０日〈/TIMEX3〉は...

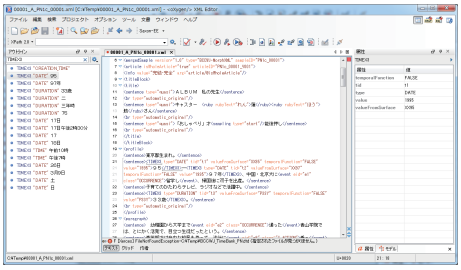

図 1: XML Editor oXygen によるアノテーション

例である。「１０日」という表現は、文脈から「４月」ということがわかるが、何年かまではわからないために定時間情報が得られない。尚、@definite 属性は作業・分析の便宜上導入したもので、元の ISO-TimeML の〈TIMEX3〉には規定されていない。

4　作業環境と作業対象

4.1　作業環境

アノテーション作業には XML Editor oXygen [2] を利用した。DTD や XML Schema を記述することにより時間表現の切り出し部分や属性に割り当てる値などを統制することができる。時間表現の切り出しはマウスもしくはキーボードで対象となる文字列を選択したうえで Ctrl-e とタイプし、タグを選択することで、XML タグ（閉じタグも含む）が挿入される。この状態で画面右の属性項目を記述することにより、XML ファイルを編集することができる。

言語資源に対するアノテーションにおいて、ある一定の基準を守ったうえで複数の作業者の主観を尊重してそれぞれの作業者間の判断の揺れを許す場合と、基準を厳格化し作業者間の判断の揺れを許さない場合の二通りの統制手法がある。本論文の時間情報表現の特定の時間軸上への写像作業は後者の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取った。1 台の PC に、キーボード二つ・マウス二つ・ディスプレイ二つを接続し、ディスプレイはミラーリングを行う。一つの PC を共有したうえで、作業者がアノテーション作業を行い、作業監督者がアノテーション仕様の改訂を行う。

[2]http://www.oxygenxml.com/

4.2　作業対象

作業対象である BCCWJ[2] のコアデータは、OW: 白書、PB: 書籍、PN:新聞、OC: Yahoo! 知恵袋、PM: 雑誌、OY: Yahoo! ブログの六つのレジスタからなり、それぞれ約 5 万語単位で、アノテーションすべき優先順位に基づいた部分集合が規定されている。全レジスタの部分集合 “A” と比較的時間表現が多いレジスタである PN の部分集合 “B” をアノテーション対象とした。表 1 にデータの概要を示す。表中「時間表現あり」は時間表現一つ以上含むもののファイル数・文数を表す。

5　タグの分析

本節では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う。表 2 に文書作成日時を示すタグを除いた@type ごとのタグの出現数を曖昧性解消の観点から二つの視点で四つに分割して示す。一つ目の視点は @definite が “true” か “false” かである。“true” の場合、時間軸上に時区間が特定可（“DURATION” と “SET” は時間幅が特定可）であることを意味し、“false” の場合、時間軸上に時区間が特定不可であることを意味する。二つ目の視点は @value と @valueFromSurface の値が一致する (“=” で表記) か、一致しない (“≠” で表記) かである。一致する場合人手による文脈を用いた正規化作業が行われていないことを意味し、一致しない場合人手による文脈を用いた正規化作業が行われたことを意味する。

文中に出現した 5297 件の時間情報表現のうち正規化作業が不要な表現が 1639 件 (約 30%) であった、残りの表現のうち 2023 件 (約 37%) がテキスト中の情報より正規化が可能である一方、1875 件 (約 34%) についてはテキスト中の情報のみでは正規化ができないことがわかった。

日付表現 (“DATE”) について、時区間特定可であるもの (@definite が “true”; 61%) の多くが、

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

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 OW	(A)	17	16 (94%)	1439	405 (28%)	40690	58336
書籍 PB	(A)	25	25 (100%)	2568	289 (11%)	50257	57929
新聞 PN	(A,B)	110	110 (100%)	5582	1562 (28%)	88733	116834
知恵袋 OC	(A)	518	250 (48%)	3479	488 (14%)	51240	60086
雑誌 PM	(A)	23	23 (100%)	3066	413 (13%)	49715	59372
ブログ OY	(A)	257	198 (77%)	3986	765 (19%)	53333	63459

表 2: `@type` 属性ごとの出現数と文脈による曖昧性解消可能性

`@definite`	`true` (特定可)			`false` (特定不可)		
`@value` と `@valueFromSurface`	all	=	≠	all	=	≠
DATE	2214(61%)	381(10%)	1833(50%)	1438(39%)	1275(35%)	163(4%)
TIME	188(37%)	1(0%)	187(37%)	315(63%)	239(48%)	76(15%)
DURATION	1129(92%)	1128(92%)	1(0%)	99(8%)	99(8%)	0
SET	131(85%)	129(84%)	2(1%)	23(15%)	22(14%)	1(1%)
ALL	3662(66%)	1639(30%)	2023(37%)	1875(34%)	1635(30%)	240(4%)

人手による曖昧性解消が行われている (`@value`≠ `@valueFromSurface`;50%) ことがわかる。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。日付表現の曖昧性解消は、和暦から西暦への換算や、西暦二ケタ表記から西暦四ケタ表記への換算、さらに年が省略されている表現の文脈や文書作成日時に基づく年の補完によるものがあり、白書の多くの事例がこの暦の換算作業であった。

時刻表現 (“TIME”) については、書籍の 1 件を除いて時区間特定可であるものの殆どが人手による曖昧性解消が行われている。時刻表現の曖昧性解消は、日付が省略されている場合の日付の補完のほか、午前と午後の曖昧性解消が含まれる。

時間表現 (“DURATION”) と頻度集合表現については、時間軸上の時区間を特定することを目的とせず、時間幅が特定できれば `@definite` が “`true`” になると定義している。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。

6 おわりに

本稿では BCCWJ に対する日本語時間情報アノテーションについて説明した。アノテーションは各国で進められている国際標準 ISO-TimeML に定義された ⟨`TIMEX3`⟩ タグに準拠している。他言語においては対象を新聞記事に限定しているのに対し、本研究は 6 種類のレジスタを対象にした。

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。

謝辞

本研究は、文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築 : 21 世紀の日本語研究の基盤整備」(平成 18〜22 年度、領域代表者 : 前川喜久雄) および国立国語研究所「超大規模コーパス構築プロジェクト」によるものです。

参考文献

[1] J. Pustejovsky et al. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, 2003.

[2] 国立国語研究所コーパス開発センター. 『現代日本語書き言葉均衡コーパス』利用の手引き, 第 1.0 版, 2011.

[3] 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. 第 2 回コーパス日本語学ワークショップ発表論文集.

[4] 保田祥, 小西光, 浅原正幸, 今田水穂, 前川喜久雄. 『現代日本語書き言葉均衡コーパス』に対する　時間表現・事象表現間の時間的順序関係アノテーション. 第 3 回コーパス日本語学ワークショップ発表論文集.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.