3Q-09

多様な時間表現の解釈に基づく言明の抽出と整理

坂口 智洋　　黒橋 禎夫
京都大学 大学院情報学研究科

1　はじめに

人々の意見や価値観は時間とともに変化する。Web上には15年以上に渡る膨大な文書が蓄積されており、現在の視点からは古い情報も多く存在している。そのため異なる時期に書かれた文書を同じように扱うことによる弊害も生まれてきている。例えばWeb文書を分析した結果、反対意見と賛成意見が拮抗しているように見えても、本当は「昔は反対意見が多かったが、近年は事情が変わり多くの人が賛成している」という可能性もある。このような知見を得るためには、時間軸を用いたテキストの分析が必要である。

テキストの解析に影響を及ぼす時間情報には、その文書作成日時とテキスト中の時間表現がある。ニュースやブログ、SNSなどでは文書作成日時が記載されることが多いが、過去や未来の事を話題にすることも多く、テキストが書かれた日時と文書中で言及している日時が一致するとは限らない。本研究ではテキスト中から時間表現を抽出・正規化する手法を提案し、これを利用して言明の整理を行う。

2　時間表現の認識と正規化

2.1　背景

近年、時間表現と事象表現とを関連付ける国際的なタグ付け基準やコーパスが検討・整備され、時間表現の抽出・正規化などのタスクが設定された[1]。2007年以降にはこれらのタスクを対象とした評価型ワークショップが開催され、機械学習やルールベースの手法が研究されている。英語や仏語、中国語では早くからコーパスが整備されたのに対して、日本語はコーパスが無くほとんど研究が行われていなかった。しかし2013年に小西らによって時間情報タグ付きコーパス(BCCWJ-Timebank)が整備され日本語でも大規模に実験できる環境が整いつつある[2]。

2.2　問題設定

本研究では評価型ワークショップでも行われる、時間表現の1. 認識, 2. TYPE推定, 3. VALUE推定の3つのタスクを扱う(図1)。時間表現の認識とはテキストから時間表現を抽出するタスクである。TYPE推定とは時間表現を日付表現(例:1970年11月, 昨日)、時刻表現(例:昨晩, 18時半)、期間表現(例:半年, 1時間)、頻度集合表現(例:毎日, 1日おき)の4つの何れかに分類するタスクである。日本語の場合、特に日付表現と期間表現に曖昧性が多く存在する。例えば「95年の震災」と「95年の歴史に幕」では、前者の95年は1995年を表す日付表現であるのに対し、後者は期間を表す期間表現となる。

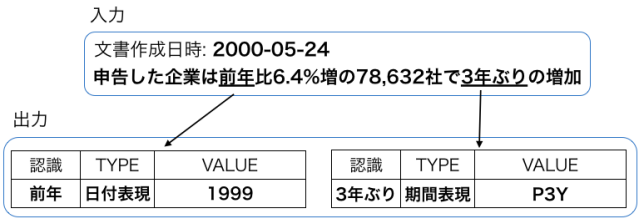

図1: 時間表現の認識と正規化

	Precision	Recall	F-score
認識・TYPE	0.86	0.82	0.84
VALUE	0.64	0.61	0.62

表1: 提案手法の評価

VALUE推定とは時間表現の正規化を行うタスクである(例:昨年6月 → 2014-06)。この時、時間表現のみから分かる情報だけでなく、文脈や文書作成日時から分かる情報も考慮して補完する必要がある。例えば2014年11月26日の記事に書かれた「昨日」という時間表現は2014-11-25と正規化する。

2.3　提案手法

本研究では分類器を作成して時間表現の認識とTYPEの推定を同時に行い、VALUEの推定はルールベースで行った。

TYPEはテキスト中の各文字に4種類のTYPEラベルかそれ以外の合計5つのラベルを貼る系列ラベリングの問題として解いた。TYPEラベルが貼られた部分が時間表現であるため、TYPEを推定することで結果的に時間表現の認識も行える。モデルにはConditional Random Field(CRF)を用い、CRF++[*]を利用して実装した。利用した素性は文字、品詞、文字が数字かどうか、VALUE推定で用いる時間表現パターンにマッチするかどうか、である。

VALUEは認識された時間表現とそのTYPEを入力として、人手で作成した時間表現パターンを利用して推定した。「昨日」や「来年」など書かれた日を基準とし

[*]http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar

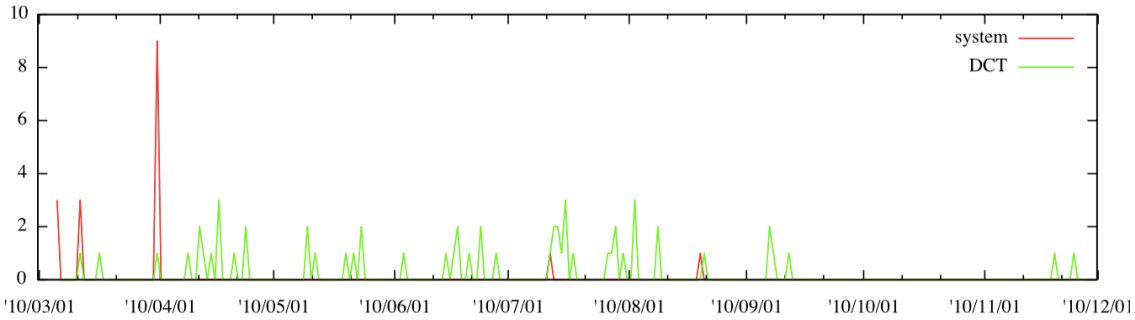

図 2: 文書作成日時 (DCT)、時間表現 (system) を用いた「Xperia, 発売」の日時推定

た相対的な表現は文書作成日時を用いた正規化を行った。一方「その前日」などは基準となる日があり、それが文書作成日時であるか文脈中の日付かが曖昧である。本手法では周辺の時間表現との類似性を考慮したルールを用いて基準となる日を決め正規化を行った。

2.4 実験

BCCWJ-Timebank を用いて提案手法の評価実験を行った。BCCWJ-Timebank は人手で作られた時間情報タグ付きコーパスであり、今回 6 つのジャンルからなる 950 文書 5482 個の時間表現を利用して実験した。本システムを実際に Web 文書に適用することを考えて、VALUE 推定の入力には正解の時間表現ではなく認識・TYPE での出力結果をそのまま用いた。実験結果は表 1 の通りである。2013 年に行われた TempEval-3[1] でも同様のタスクが行われた。今回の実験とは、対象とする言語やコーパスと各タスクで正解の入力を利用して評価している点が異なる。参加したシステムのおおよそ F 値が認識 0.7-0.8、TYPE 0.7-0.8、VALUE 0.6-0.8 であり、本システムもこの水準に達していると言える。

3 言明の抽出と整理

テキスト中の時間表現を利用することで、文書作成日時のみを利用した場合に比べてより正確な言明の抽出と整理が可能となる。このことを、2008 年から 2010 年に収集した Web 文書から「スマートフォン」に関連する文書 1000 件を取得し確認した。

まず「Xperia」「発売」の両単語を含む 160 文を抽出した。図 2 の緑線は該当する文の文書作成日時を、赤線は該当する文に時間表現が含まれる場合にその日時を解釈し図示したものである。実際に Xperia(Xperia X10) が発売されたのは 2010 年 4 月 1 日である。多くの記事は発売後に書かれているが、「1 日に発売された Xperia が〜」など文中で言及されている日時を解釈することで正確にこの出来事の日付を推定することができる。

次に時間表現が明示された言明を抽出し、時間軸上に整理した。具体的には時間表現を含む文を取り出し、

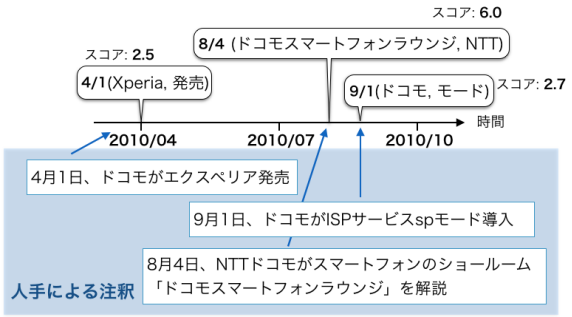

図 3: 「スマートフォン」に関する言明の整理

含まれる名詞句のペアを抽出する。その後各名詞句ペア p に対して下記のスコアを計算し、この値の上位のものを取り出す。

$$score(p,t) = \frac{etf(p,t)}{edf(p)}$$

ここで、$etf(p,t)$ はある日時 t に書かれた文のうち p を含むものの個数、$edf(p)$ は p を含む文が書かれた日が全部で何日あるかを表している。このスコアにより、ある日時 t に多く言及された p を取得することができる。図 3 に前述の文書 1000 件にこのアルゴリズムを適用した結果を示す。下部には実際の出来事を注釈として与えた。

4 まとめ

本稿では日本語テキストから時間表現の認識と正規化を行う手法を提案し、英語でのシステムに匹敵する精度を得た。また提案したシステムを利用して Web 文書中の出来事を時間軸上にまとめた。

参考文献

[1] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. pp. 1–9, Atlanta, Georgia, USA, June 2013.

[2] 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. 自然言語処理, Vol. 20, No. 2, pp. 201–221, 2013.