文の過去/非過去分類と接続語に着目した 文学作品に関する時間情報解析ヒューリスティクス

田中 武揚
中央大学大学院 理工学研究科
taktanak@suzuki-lab.ise.chuo-u.ac.jp

鈴木 寿
中央大学 理工学部
suzuki@ise.chuo-u.ac.jp

1 はじめに

近年, TimeBank[5] など時間関係を付加したコーパスの整備などにより, 電子化されたテキストの時間情報抽出や時間的順序関係の推定などを扱う時間情報解析が盛んになりつつある.

本研究では時間情報解析の対象テキストとして, 特に散文で書かれた文学作品を扱う. 文学作品における時間表現は, “ある日”や “秋”など厳密に定義されない場合が多い. また文学作品は複数の場面から構成されるが, その時間的順序は新聞やニュース記事のように形式だっている場合は少なく, 過去/現在/未来などを複雑に行き来する. しかし, 我々人間は経験的に獲得した何らかの規則により, 様々な時間表現を取捨選択し, 作品の作者が意図した作品の時系列を適切に把握し理解することができる.

本研究では, 人間の読書の仕方からヒントを得た知見として, 主に文の過去/非過去分類, 接続語/指示語の活用という 2 つを用いて各文で述べられている話題の時間を判定し, 文の時間として付与することを試みる. これを人手により付与された時間と比較し, 知見の有用性について検討する.

2 時間情報解析システムの概要

対象作品は, 青空文庫 [2] より “駅伝馬車”, “クリスマス・イーブ”, “スリーピー・ホローの伝説”とする. 本研究で提案する文学作品の時間情報解析システムの流れについて, 図 1 に示す.

まず, ルビの削除など前処理を施したテキストを段落および文単位で分割する. 本研究において, 段落は青空文庫のファイル形式より改行で区切られたものとする. また, 文は “。”, “！”, “？”で区切られたもの, 会話文や引用文など特殊な文 (以下, 括弧文) は “「」” または “（）”それぞれの組で囲まれたもののみを扱うこととする.

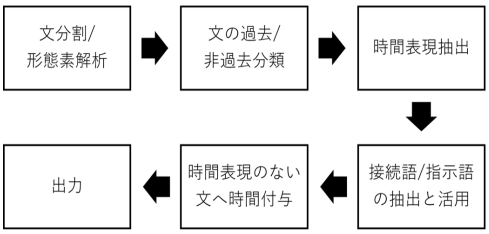

図 1: 時間情報解析システムの流れ

文分割されたテキストについて, テキストの先頭から各文ごとに形態素解析をおこない, 分かち書きされた文とと各単語の品詞情報を得る. 得られた品詞情報と人手によって作成された規則群を用いて, 文の過去/非過去分類, 時間表現の抽出をおこなう. そして, 文の直前および直後の文と時間区分が独立した 1 文について, 時間表現を持たず文頭に接続語および指示語がある場合, 直前の時間区分と一致させる. その後, 時間表現のない文に対し直近の同じ時間区分の時間表現を付与する. 最後に各文に割り当てられた時間表現を出力する.

3 文の過去/非過去分類

文学作品の各文には, 多くの時間区分が存在するだろう. 例えば, 文学作品の地の文において, 話の軸となる時系列で述べられる過去と現在が存在する. さらに, 作者が作品を俯瞰した立場から言及し, 話の軸となる時系列に配置できないメタ的な時間や, 未来や実際に起きていない事を述べる仮定的な時間が存在する. また括弧文について, 会話文と特に時間を指定することができない引用文の場合などがある. さらに, 会話文は発話された時間の話題である場合と, まったく別の時間を指定しそれについて述べる場合がある.

人間が読書をする際, 複雑な時間区分を文末の動詞や形容詞の時制などを基にしながら, 適切に感じ取ることができる. 例えば, 過去に属する話題が述べられ

ている最中に, 現在に属する文が得られた場合, その文はメタ的な時間を用いた補足としての役割などが考えられる. さらに, 過去に属する文が得られた場合, 元の話題に戻ったと考えて読むことが多いだろう. このように同一の時間区分に属する文において, 語られる話題は一致しており, 時間的にも連続している場合が多いと考えられる.

この知見の実現として, 単語の文の過去/非過去分類の段階では, 各文の文末の品詞情報などを利用し, 地の文や会話文の時間区分を明確にすることを目指す.

3.1 文の過去/非過去分類の手法

地の文は, “過去”, “非過去”, “時間区分なし”の 3 つに分類する. 会話文は, “発話時より過去”, “発話時”の 2 つに分類する. なお, 会話文とは括弧文の内 “「」” で括られたものとする.

まず, 文末に出現しうる単語の品詞情報を “過去”, “非過去”, “時間区分なし”, “判断対象とせず”の 4 つに分類し, 各時間区分の品詞情報のリストを作成する. ここで, ある品詞情報はただ一つのリストに登録されるものとする. “過去”には主にタ形, “非過去”には主に形容詞や動詞の基本形を含む. “時間区分なし”は体言で終わる文など, 時間区分を判断できない文末表現などを含む. “判断対象とせず”は “〜である”など, 文の時間区分が文末より前の表現によって判断される表現などを含む.

次に形態素解析を施したテキストの各文に対して, 地の文か会話文かを “「」”の有無により判定する. そして, 記号を除いた文末の単語の品詞情報と作成したリストを照合し, 該当するリストの時間区分を文末の時間区分とする. もし, “判断対象とせず”に該当した場合文の時間区分が確定するまで, 一つずつ前の単語の品詞情報とリストの照合を繰り返す. 最後に文が, *a*) 地の文の場合, 文末の時間区分をそのまま文の時間区分とする; *b*) 会話文で, 文末の時間区分が “過去”に該当する場合, 時間区分を “発話時より過去”と判断し, それ以外の場合の時間区分は “発話時”と判断する.

3.2 文の過去/非過去分類の評価と課題

本手法の評価のため, 青空文庫の 21 作品の各文に対して本分類手法を適用し得られた時間区分を, 人手によって付与された各文の時間区分と比較した. 人手による評価の際は, 会話文の分類を “発話時より過去”,“発話時”, “会話文でない”の 3 つとした. 人手によって “会話文でない”と判断された括弧文について, 提案手法では分類できなかったためすべて誤りとする. 判定対象の総文数は 2949 文であり, 本手法による分類と人手による分類が異なる文の総数は 90 文であった.

分類が異なった理由として, 会話文と引用文などの括弧文の区別を, 文末の表現のみでは判断できなかったことが挙げられる. また本研究では, より文末に近い単語の品詞情報をリストと照合し, 該当するリストの時間区分を文の時間区分とした. これにより, 文の中で倒置や省略が生じている場合は, 本手法では適切に分類することができなかった.

4 時間表現抽出

本研究では, 時間表現として “朝”や “1993 年”など明示的な時間表現のみを扱う. 明示的な時間表現の内, “1993 年”など数値によって時間を表した表現を数値時間表現, それ以外の表現を言語時間表現と呼称する.

この他の時間表現として, “スキー”といった連想により時間を示す暗示的な時間表現や, 比喩による時間表現が存在する. しかし, 暗示的な時間表現や比喩は文脈に強く依存する. 例えば, “起床”という言葉は時間帯として “朝”を連想させるが, “午後 3 時に起床した”という文は不自然ではない. このように暗示的, 比喩的な時間表現は文脈・意味解析を含むため, 本研究では扱わない.

また, 時間情報解析の分野においては, 文書作成時刻がよく用いられる. しかし, 文学作品では, 現実と異なる時間を軸として話が展開していく場合が多い. よって, 本研究では文書作成時刻を取得しない.

本研究における基本となる時間の単位 (以下, 基本時間単位) を大きいものから, “世紀”, “年”, “月”, “週”, “日”, “曜日”, “時”, “分”の 8 つとする. なお, “日”および “曜日”は同じ大きさであるとする.

4.1 時間表現抽出の手法

まず言語時間表現を 24 時制の数値時間表現に変換するため, 計 106 の時間概念をまとめたデータベース (以下, 時間概念データベース) を作成する. 時間概念データベースには, 概念の見出し語と対応する数値時間表現が登録されている. 例えば, “朝”という概念には, 数値時間表現として “6〜9”時とすることが登録さ

れ, “明日”という概念には, 作品内の時間を “+1”日することが登録されている. 登録基準として, 主に気象庁 [3] が公表している用語を利用する. 深夜については記載がないので, 青少年育成条例などで用いられる 23 時から 4 時とした.

次に, 時間概念データベースの見出し語の類義語, 計 318 語をまとめた時間表現のデータベース (以下, 時間表現データベース) を作成する. 時間表現データベースには, 時間表現の見出し語と対応する概念が登録されている. 各見出し語に対応する概念はただ一つとする. 類義語は, 文学作品や辞書, 日本語 WordNet[4] から収集した.

次に, 形態素解析を施した各文に対して, 時間表現データベースに登録されている見出し語に合致する表現が存在するか否かを検索する. そして, 合致する表現が存在した場合, 時間概念データベースに基づき言語時間表現を数値時間表現へと変換し, 文の時間表現として保持する. 1 文に複数の時間表現が存在する場合は, すべて保持する.

5 接続語/指示語の抽出と活用

時間表現を持たず, 時間区分が直前の文と直後の文とは独立している場合がある. このような文の文頭が接続語または指示語である場合, 直前の文と同様の時間についての話題を述べていることが多いと考えられる. そこで, 文頭の接続語および指示詞から時間表現を持たず, 時間区分が文の前後とは独立した 1 文を直前の時間区分と一致させる.

5.1 接続語/指示語の抽出と活用の手法

まず, 接続語のデータベース (以下, 接続語データベース) を作成する. 作成にあたって主に 分類語彙表増補改訂版データベース [1] の接続に属する語のうち, 文頭に置くことができる語を登録した. ここで接続語の内, “さて”や “ところで”などの転換としての役割を持つ接続語は用いない. これは, 話題の転換を目的に使用される語であり, 直後に語られる時間が異なる場合が多いからである.

指示語については, 形態素解析の段階で指示詞として判定されたものを用いる. ここで, “どれ”などド系に属する指示詞は, 不定称なので用いない.

次に, 時間区分が直前の文と直後の文から独立している文について, 文の文頭の表現が指示詞または接続語データベースに登録された表現か判断する. 登録された表現の場合, 直前の文と同じ話題について述べている文として判断し, 文の時間区分を直前の文の時間区分と一致させる.

6 時間表現のない文へ時間付与

時間区分の分類および時間表現の抽出が終わったテキストに対して, 時間表現のない文へ同じ時間区分の直近の時間表現を付与する. このために, 各時間区分ごとに時間表現を取得と更新をおこなう.

文学作品において話題となる時間が変化する際, まったく別の時間へと変化することは少ない. つまり, 新たに時間表現を得た場合, 新たに取得された時間表現の最大の基本時間単位を求め, それ以下の基本時間単位の値を更新する. 例えば “2018 年 12 月 30 日朝”の話題について述べているとき, 新たに “夜”という時間表現を得たとする. その際, “朝”のみを更新し “2018 年 12 月 30 日夜”の時間の話題へと遷移したとする.

6.1 時間表現のない文への時間付与手法

まず, 各時間区分ごとに最新の時間表現を保持するため, 各基本時間単位の数値時間表現を要素として持つ配列を各時間区分ごとに用意する. 配列の初期値として空であることを示す値 null を与える.

次に, テキストの先頭から各文について,

- 時間表現を持つ文の場合
  1. 文の時間区分に対応する配列を取得する
  2. 新たに取得される時間表現の最大の基本時間単位を求め, それ以下の基本時間単位に属する値をすべて null とする
  3. 新たな時間表現を配列に与える
- 時間表現を持たない文の場合, 文の時間区分に対応する配列を取得し, 配列に保持されている時間表現を文の時間表現とする.

7 時間情報解析システムの評価

7.1 評価手法

まず評価のため対象作品について人手によって, 各文に基本単位時間ごとに時間表現を付与する. メタ的

な時間や引用文などについては, それぞれに対応した値を付与する.

次に, 本研究で提案した a 文の過去/非過去分類; b 時間表現抽出; c 接続語/指示語の抽出と活用; d 時間表現のない文への時間付与 のうち, "b+d", "a+b+d", "a+b+c+d"を組み合わせたシステムをそれぞれ構築し, 対象作品の各文に時間を付与した. そして, 人手によって付与された時間とそれぞれの手法で付与された時間を, 各基本時間単位ごとに比較し, その正誤によって評価をおこなう.

7.2 評価結果および課題

評価結果の例として, "駅伝馬車"に各手法を適用し各基本時間単位ごとの正答率 (%) を求めたものを, 表 1 に示す. "駅伝馬車"の総文数は 118 文である. また, 各正答率は小数点第 2 位で四捨五入している.

表 1: 提案手法による時間情報解析の正答率 (%)

基本時間単位	b+d	a+b+d	a+b+c+d
世紀	64.4	65.3	64.4
年	60.2	65.3	64.4
月	40.7	50.8	50.0
週	64.4	65.3	64.4
日	1.7	1.7	1.7
曜日	64.4	65.3	64.4
時	44.1	43.2	43.2
分	63.6	64.4	63.6

"世紀", "週", "曜日", "分"についての時間表現は 3 作品中に出現することがほとんどなかった. これらの基本時間単位の正答率を下げた主な要因として, 提案手法では会話文と引用文などの区別をつけるに至らなかったことが挙げられる. また, メタ的な時間や仮定の時間などを分類できなかったため正答率が下がった.

"日"は 1.69%と低い正答率となった. "駅伝馬車"は, 主に "12 月 24 日"の出来事を述べたテキストだが, "クリスマスの準備"のように "クリスマス"という表現が話題の時間に関わることなく出現し, それらを文の時間表現として判定したことによる. また, 他作品では "アメリカの歴史がはじまったころ"など歴史的な出来事や知識を利用した明示的な時間表現, "一時"など多義性のある単語などによって正答率が下がった.

その他の基本時間単位について, "駅伝馬車"の場合, 手法 a+b+d の結果が正答率が他の手法と比べ良かった. 特に "月"に関しては, 手法 b+d よりも約 10%高くなった. 文を過去/非過去に分類するとことは, 一定の効果があると考えられる. 一方で, 他作品では手法 b+d が最も良い結果となる場合もあった. これは主に, 時間表現を伴って出現することが少ない時間区分に属する文が保持する時間表現が, 長く更新されなかったことによる.

最後に対象としたテキストにおいて, 提案した接続語/指示語の活用手法は効果がなかった. これは主に, 接続語/指示語の参照先を考慮していないこと, 直前の文が保持する時間表現が正しくなかったことによる.

8 おわりに

本研究は, 人間の読書の仕方から得た知見を基に, 文学作品を対象に時間情報解析をおこなった. 文の過去/非過去分類は作品により正答率を向上させたが, 更新の手法に課題があった. 接続語は提案手法では効果がなかった. 今後は, 検証対象の作品数を増やし詳細な評価をおこなうとともに, 得られた課題への解決手法や, 別の知見を用いた手法を模索していく.

参考文献

[1] "分類語彙表増補改訂版データベース - ver.1.0", 国立国語研究所, 2004.

[2] "青空文庫", ⟨https://www.aozora.gr.jp/⟩, 2018/12/28 アクセス.

[3] "気象庁｜予報用語　時に関する用語", ⟨https://www.jma.go.jp/jma/kishou/know/yougo_hp/toki.html⟩, 2018/12/29 アクセス.

[4] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto, "Japanese SemCor: A Sense-tagged Corpus of Japanese" in The 6th International Conference of the Global WordNet Association (GWC-2012), Matsue, 2012.

[5] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo, "The TimeBank corpus", *Proceedings of Corpus Linguistics*, pp. 647–656, 01 2003.