文学作品における時間表現と規則を用いた時間情報解析ヒューリスティクス

Time Information Analysis Heuristics that Use Time-related Representations and Rules on Literary Writings

情報工学専攻 田中 武揚

Information and System Engineering TANAKA Takeaki

要約: 本研究は, 読書モデルを仮定し, 経験則に基づいた推論を用いて文学作品の各文の時間を取得する手法を探求する. 4 つの経験則に基づく推論を用いた読書モデルを実現し, 取得した時間と人手によって付与された時間を比較した. そして, 文学作品の時間情報解析における重要な課題や知見を得た.

キーワード: 自然言語処理, 時間情報解析, 時系列推定, 固有表現

1 はじめに

電子テキストにおける時間的順序関係や時間表現の抽出など, 自然言語の “時間” を扱う時間情報解析は, 因果関係の推定など高度な意味理解に深く関連するとされ, 重要な課題の一つである. しかし, これまでの計算機による “時間” の処理は, 新聞記事など事実を客観的に伝達することを目的としたテキストを対象とした研究や, 時間表現のみに着目した研究が多かった. 文学作品などの娯楽や芸術性を目的としたテキストは, 抽出の難しい時間表現や時間的順序がばらばらである場合が多く, テキストの性質や時間表現以外の構造に踏み込んだ時間解析手法が必要であると考えられる.

人間が読書する際, 経験則に基づく推論から, 必要な時間表現の取捨選択や作品の時系列の推定できる. 本研究は人間の読書から着想を得た “時間” の推論モデルを仮定し, 文学作品の時間情報解析を試みる. 一つの手法として, “時間表現”, “文の時間区分”, “接続語/指示語”, “文への時間付与” という 4 つの経験則を用いた簡易な推論モデルを提案する. 青空文庫 [2] より 21 作品を対象に, 人手によって付与された時間と比較することにより, 本研究で用いた経験則や推論手法の有用性を検証する.

2 準備

2.1 基本とする時間単位

本研究における基本とする時間の単位 (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.

2.2 時間表現の定義

本研究は, 時間表現として “朝” や “1993 年” など明示的な時間表現のみ扱う. 明示的な時間表現の内, 数値による時間表現を数値時間表現, 単語による時間表現を言語時間表現と呼称する. また, 明示的な時間表現の内, “朝” などその時間表現のみで時間を示すことができる時間表現を絶対時間表現と呼称する. これに対し, “明日” など他の時間表現から何らかの関係をもつことにより, 時間を示す時間表現を相対時間表現と呼称する.

2.3 文の時間区分の定義

地の文の場合, “過去”, “非過去”, “時間区分なし” の 3 つの時間区分に分類する. このほかに, “現在” や “未来” などの時間区分が考えられるが, それらは区別せず “非過去” として扱う. 会話文の場合, “発話時より過去”, “発話時”, “会話文でない” の 3 つの時間区分があるとする. “会話文でない” は, 作品の作者または登場人物によって引用された文であり, 発話という行為を伴わない文を指す.

2.4　時間独立文

作品内には, 時間表現を持たず, 直前と直後の文と時間区分も異なる文がある．これを本研究では, 時間独立文と呼称する．

3　時間情報解析システムの概要

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．

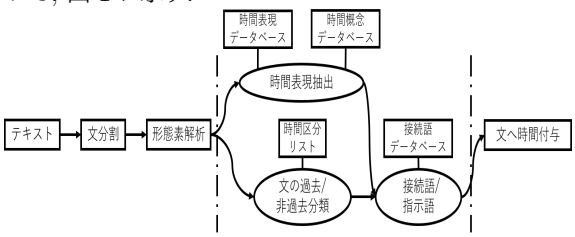

図 1　時間情報解析システムの流れ

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．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, 次に経験則に基づいた各推論をおこなっていく．

本モデルは, “時間表現抽出”, と, “時間区分の判断” という大きく 2 つの推論から成立する．“時間表現抽出” は, 時間表現を抽出し, 計算や比較できる形で保持することを試みる．分かち書きされた文から, 時間表現データベースと時間概念データベースという 2 つの知識と経験則に基づいて, 得られる時間表現を数値時間表現に変換し保持する．“時間区分の判断” は, 文の時間区分の取得を試みる．まず, “文の過去/非過去分類” として, 形態素解析の際に得られた品詞情報を用いて, 各文の時間区分を決定していく．そして, “接続語/指示語” を活用し, “時間表現抽出” および “文の過去/非過去分類” より得られた時間独立文を直前の文の時間区分に変換する．最後に “時間表現抽出” および “時間区分の判断” から得られた結果に基づいて, 各文が出現した場面の時間を推定し, 文の時間として付与する．

4　時間表現抽出

時間表現抽出は, 形態素解析を施した文へ, “朝” や “1993 年” など明示的な時間表現の抽出し, 場面の時間となり得る時間表現の候補の取得を目的とする．また時間表現の計算や比較をおこなうため, 時間表現を抽出した後, 言語時間表現を数値時間表現へ変換する．

4.1　時間表現抽出に用いるデータベースの作成

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 107 の時間概念をまとめたデータベースを作成する．時間概念データベースには, “朝” であれば “6∼9” のように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．次に, 時間概念データベースの見出し語の類義語, 計 321 語をまとめた時間表現の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．

4.2　時間表現抽出の手法

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 文中に複数の時間表現が得られた場合, 相対時間表現が得られ, 既に取得された時間表現と共通の基本時間単位を持ち, 計算ができる場合, 計算する．相対時間表現以外の時間表現であるなら, 新たに取得された時間表現の最大の基本時間単位以下の部分のみを更新する．最後に, 取得された時間表現が整合性があるかどうか確認する．“1 月 32 日” など整合性がない時間表現である場合, 時間のくり上がりやくり下がりを用いて整合性のある時間表現へと変換する．

5　文の過去/非過去分類

同一の時間区分に属する文において, 語られる場面は一致しており, 時間的にも連続している場合が多い．そこで, 時間区分における経験則を活用することで, 本研究では時間表現を持たない文に対して時間を付与し, また, メタ的な時間で述べられている場面に関係のない時間表現の影響を, 特定の時間区分のみに抑えられる、と

本研究では仮定する．この仮定に基づき，文末の情報から文の時間区分を推定し，活用することを試みる．

提案手法では，地の文は，“過去”，“非過去”，“時間区分なし”の 3 つに分類する．また，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．

5.1　品詞情報リストの作成

主に文末の単語の品詞情報を “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．ある品詞情報はただ一つのリストに登録されるとする．

5.2　文の過去/非過去分類の手法

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，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．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．

5.3　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する．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 つとしている．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ていないので，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，会話文と引用文の区別を，文末の表現のみでは判断できなかったことがあげられる．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ている．これにより，文の中で倒置や省略が生じている場合は，適切に分類することができていない．

6　接続語/指示語の抽出と活用

時間独立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の話題を述べることが多くみられる．そこで，文頭の接続語および指示語を持つ際，時間的独立文の時間区分を直前の時間区分と一致させる．

6.1　接続語データベースの作成

接続語の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．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．

6.2　接続語/指示語の抽出と活用の手法

時間独立文について，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．

7　文への時間付与

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する．文学作品において話題におけ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．そこで，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”について述べる場面で，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．

7.1　時間表現のない文への時間付与手法

まず，各時間区分ごとに作品内で最新の時間表現を保持するため，各時間区分の直前に出現した時間表現における各基本時間単位の数値時間表現を要素としてもつ配列 (以下，場面時間配列) を各時間区分ごとに用意する．

作品内の各文について，時間表現をもつ文の場合，文の時間区分に対応する場面時間配列を取得し，時間表現の計算ができる場合は計算し，それ以外の場合は場面時

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	**a+d**	**a+b+d**	**a+b+c+d**
世紀	70.3	70.3	70.3
年	66.1	70.3	70.3
月	30.5	52.5	51.7
週	70.3	70.3	70.3
日	1.7	1.7	1.7
曜日	70.3	70.3	70.3
時	47.5	44.1	44.1
分	65.3	62.7	62.7

間配列の新規の時間表現の最大の基本時間単位以下の値を, 新規の時間表現へ更新する．時間表現を持たない文の場合, 文の時間区分に対応する場面時間配列を取得し, それを文の時間表現として保持する．

8　時間情報解析システムの評価

8.1　評価手法

評価のため対象の 21 作品について, 人手により各文に時間表現を付与する．メタ的な時間や引用文などは, 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**： 時間表現抽出”, “**b**： 文の過去/非過去分類”, “**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, “**a+d**”, “**a+b+d**”, “**a+b+c+d**” を組み合わせたシステムを構築し, 対象作品の各文に時間を付与する．そして, 人手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, 正誤によって評価する．

8.2　評価結果

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．

8.3　時間表現に関する検証

まず, 課題として, 時間表現の記述方法があげられる．“〜したとき” のような出来事や登場人物の行動により表される時間表現, “間” のような時間の区間を明確に指定する時間表現など本研究における記述方法だけでは表現しきれない時間表現がみられた．また, すべての時間表現を場面の時間として採用するのではなく, 場面の時間となりうる時間表現となりえない時間表現を, 抽出や更新に際に分類することは, 重要な課題である．例えば, 駅伝馬車はクリスマスイブの出来事について述べた話であるが, そのことを示す語は作品の冒頭で一度現れるだけである．その後は “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, それらを場面の時間として捉えている．結果, 基本時間単位 “日” は 1.7% という低い正答率となっている．

8.4　文の過去/非過去分類の検証

駅伝馬車の基本時間単位 “月” は約 22 ポイント向上している．この作品は主に, 時間区分 “過去” の文に作品の中心的な話が書かれ, 基本時間単位 “月” の場面に関わる時間表現も “過去” の文に多く出現する．一方で, 時間区分 “非過去” の文には筆者による補足の話が含まれ, 場面に関わらない時間表現が用いられている．本手法は, 時間区分を独立させ, それぞれで時間表現を保持されることで, “非過去” の文で筆者による補足の際に使用されている場面にそぐわない時間表現の影響を限定し, 結果, 正答率が大きく向上していると考えられる．

8.5　接続語/指示語の活用手法の検証

接続語や指示語で時間区分が変更される文は, 時間表現を持たない文のみであり, それぞれの時間区分で保持される時間が大きく変化することはない．結果として, 全体の作品で見た場合, 文の過去/非過去分類のみを用いた場合と比較して, 正答率が大きく変化していない．

9　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, 読書モデルを仮定し, 文学作品を対象に時間情報解析をおこなった．本研究を通して, 出来事による時間表現などを扱うことのできる記述方法や, 時間表現の取捨選択など, 多くの課題を明確にした．また, 言語時間表現を数値時間表現に変換することにより, 時間表現の計算できる, など有用な知見を得た．今後は, さらに人間の読書の仕方を観察し, 別の推論手法を模索する必要がある．

関連図書

[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”, Internet: ⟨https://www.aozora.gr.jp/⟩, [2019/2/10 アクセス].