4R-4

要約のための物語文章の時系列推定

瀧本 洋喜 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

本研究は，著者による意図的な時系列の操作に着目し，時間的因果性を損なわない物語要約システムを提案する．

2 物語時系列推定システム

時語を手掛かりとした自動要約システムの一部として,「物語時系列推定」を構築している．このシステムは，以下に示す三つの仕様を満たすものとする．

1. 物語文章を入力として与えると，その文の時系列情報を推定し出力する．
2. 時系列情報は，一文に対して，時間帯，年月日，季節，人生の四つの異なる単位の情報を推測する．
3. 時系列情報を推測する手掛かりとして，時語データベースに登録された時語を用いる．

以上の仕様を考慮し，図 1 に示すシステムを提案する．

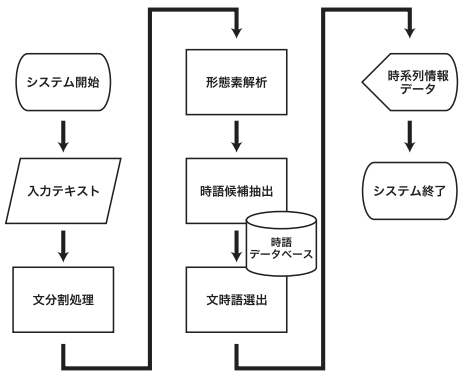

図 1: システムのフローチャート

なお，本稿では，時系列が操作された，或いは進行したことを時系列情報と定義する．また，時系列情報を表現する語句を，先行研究 [1] に従い時語と呼ぶことにする．

A Time-Line Inference for Summarizing Story Sentences
† Hiroki TAKIMOTO
† Noriyuki OKUMURA
Nagano National College of Technology, Department Electronics and Computer Science, noriyuki_okumura@ei.nagano-nct.ac.jp (†)

2.1 時語データベース

時間帯単位，年月日単位，季節単位，人生単位の四種類に分類された代表的な時語，511 語を使用する．

2.1.1 時間帯単位時語

一日の中のある時間帯を表す時語であり，二種類，計 110 語を登録している．

（例：「基準時語」: 朝，夜,「連想時語」: 始発，下校）

2.1.2 年月日単位時語

一年や一日単位の移り変わりを表す時語であり，三種類，計 60 語を登録している．

（例：「基準時語」: 日，年,「相対時語」: 今日，再来月,「週時語」: 火曜日，木曜日）

2.1.3 季節単位時語

一年の季節のどこかを表す時語であり，三種類，計 257 語を登録している．

（例：「基準時語」: 春，夏,「指定時語」: 春分，七夕,「連想時語」: 海，コスモス）

2.1.4 人生単位時語

人の一生のある期間を表す語句であり，84 語を登録している．

（例：高校生，青年）

3 評価実験

今回作成したシステムの評価実験として，青空文庫 [2] から入手できる物語 6 作品を使用し，2 節で提案した手法を用いた物語時系列推定システムについて評価した．

3.1 実験データ

評価実験のために，六つの小説から以下の二つの種類の実験データを用意する．

A 群： 時系列情報が含まれている文（167 文）
六つの小説の文章を対象に，人出で抜き出した，いずれかの時系列情報を持った文．

B 群： 時系列情報が含まれていない文（159 文）
つの小説の文章を対象に，人出で抜き出した，いずれの時系列情報も持たない文．

3.2 評価方法

システムの評価実験として，3.1 節に示したデータを使用し，以下の 3 種類の実験を行う．

1. 四つの時系列単位の推定について評価する．
2. 人生時系列推定の有意性を調査する．
3. 物語に対する提案手法の有効性を調査する．

3番目の実験のみ，青空文庫で入手できる「麦藁帽子」を実験データとして使用する．正しい時系列を推定できたものを「正答」，推定できなかったものを「誤答」とする．

3.3 結果

1番の実験の結果として，A群に対して4つの時系列単位それぞれの推定を行った結果を表1，B群に対して推定した結果を表2に示す．2番の実験の結果として，3単位の時系列推定を行った結果を表3に，4単位の時系列推定を行った結果を表4に示す．3番の実験の結果として,「麦藁帽子」に対して時系列推定を行った結果を表5に示す．

表 1: A群に対する4つの時系列単位の推定結果

時系列単位	正答	誤答
時間帯単位	0.787	0.213
年月日単位	0.552	0.448
季節単位	0.717	0.283
人生単位	0.754	0.246

表 2: B群に対する4つの時系列単位の推定結果

時系列単位	正答	誤答
時間帯単位	0.925	0.075
年月日単位	0.851	0.149
季節単位	0.842	0.158
人生単位	0.925	0.075

表 3: 3単位時系列推定の物語文章に対する精度

時系列情報の有無	文数	評価	評価数	割合
情報有り	167	正答	46	0.275
		誤答	121	0.725
情報無し	159	正答	120	0.755
		誤答	39	0.245

4 考察

表1の結果から，年月日単位以外の3単位については70% 以上の正答率が得られた．先行研究 [1] の正答率が69.4% 程度であったことから，十分有意な結果が得られたと考えられる．しかし，年月日単位の正答率は，ほかの3つの単位を比べて明らかに低い．原因は語句の多義性にあると考えられ，多義性をもった語句への対応が，今後の課題である．表2の年月日単位と季節単位の正答率についても，同様の課題がある．

表 4: 4単位時系列推定の物語文章に対する精度

時系列情報の有無	文数	評価	評価数	割合
情報有り	167	正答	76	0.455
		誤答	91	0.545
情報無し	159	正答	111	0.698
		誤答	48	0.301

表 5: 「麦藁帽子」に対する時系列推定精度

時系列情報の有無	文数	評価	評価数	割合
情報有り	123	正答	80	0.650
		誤答	43	0.350
情報無し	334	正答	294	0.880
		誤答	40	0.120

人生単位の物語文章への有意性は，表3及び表4から明らかである．3単位で推定を行った場合に比べ，4単位で推定を行った場合，B群に対する正答率がやや落ち込むが，A群に対しては，3単位時系列推定の正答率の倍近い正答率が得られた．

物語文章に対する提案手法の有意性は，表5から読み取ることが出来る．時系列情報を持ったA群相当の文に対しては65.0% ，時系列情報を持たないB群相当の文に対しては88.0% の正答率が得られた．どちらの数値も，表1及び表2に示された正答率より多少の落ち込みがみられるが，ほぼ同程度の正答率を得ることが出来た．

5 まとめ

本研究では，著者による意図的な時系列操作に着目し，時間的因果性を損なわないような提案手法を提案，その一部をシステムとして構成し，評価を行った．

年月日単位と季節単位の時系列推定に課題は残るものの,「麦藁帽子」に対する時系列推定の評価結果を示した表5にもあるとおり，一つの物語に含まれる時系列情報の65.0% を抽出することができた．また，時系列情報を含まない文は88.0% の正答率で時系列情報を含まないと判別することができた．これは，物語の76.5% の時系列情報を正しく判別できたことになる．要約率にもよるが，文章要約に使用されるのは文章の一部分であることから，要約に使用するには十分な精度で，物語時系列情報の推定ができたと言える．

参考文献

[1] 「連想メカニズムを用いた時間判断手法の有効性の検証」土屋 誠司, 渡部 広一, 川岡 司 : 自然言語処理研究会報告 No.73 pp.113-118, 2005-07-22

[2] 「青空文庫」http://www.aozora.gr.jp/index.html (閲覧日:2011/1/13)