2S-06

時空間情報を含むテキストデータを対象とした軌跡地図要約方式

石井　雄大†　　岡田　龍太郎†　　中西　崇文†
武蔵野大学データサイエンス学部データサイエンス学科†

1. はじめに

物語の概要を示すためには，一般的に，要点をまとめた文章による要約が主に用いられる．しかしながら，文章読解を苦手とするユーザにとっては，文章による要約だけでは，内容を理解することが難しいことがある．そのため，物語の展開を直感的に図示する可視化ツールの実現が重要となる．

本稿では，物語の概要を可視化するために必要な基本要素として，時空間情報，人物情報，感情極性情報に着目する．これらの情報をデジタルテキストデータで表されている物語から自動的に抽出し，整理，統合，可視化するシステムを実現することにより，文章読解を苦手とするユーザを対象とした新たな要約を表現するメディアを創成することができると考えられる．

本稿では，時空間情報を含むテキストデータを対象とした軌跡地図要約方式について示す．軌跡地図とは地図上において，順番に地点を移動する様を可視化したものである．本方式では，テキストデータに含まれる空間情報を出現順に可視化しつつ，同時に出現する時間情報，人物情報，感情極性情報をテキストデータから抽出し，軌跡地図に表現することで文章読解を苦手とするユーザの内容理解を補助することが可能である．

2. 関連研究

奥村ら[1]は，オフィスでの事象を表すイベントという概念を 5W1H で定義し，それらをテキストデータから抽出することで，鳥瞰要約と呼ばれるそのイベントに関する 5W1H で整理された経緯の抄録の生成を実現している．

山崎ら[2]は，テキストメールとして発信される犯罪発生情報から時間情報および場所情報を抽出し，地図と合成して表示するシステムを実現している．

本方式は，物語の概要を可視化するために必要な基本要素として，時空間情報，人物情報，感情極性情報に着目し，テキストデータに含まれる空間情報を出現順に可視化することで，時空間情報，人物情報，感情極性情報を直感的に可視化することで，軌跡地図として表現することを実現している．

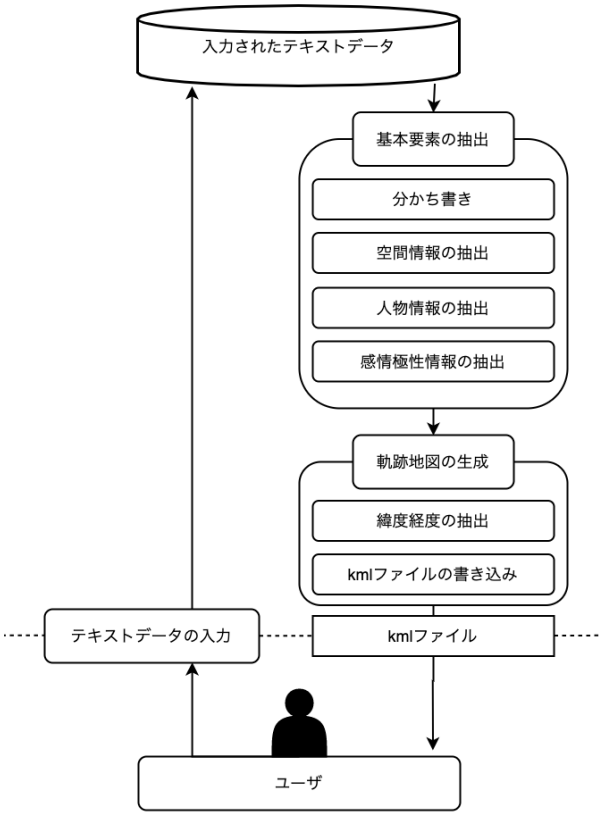

図 1: 軌跡地図要約方式のシステム全体図

3. 提案方式

3.1 システム全体像

図 1 に軌跡地図要約方式のシステム全体図を示す．本方式は主に，基本要素の抽出，軌跡地図の生成からなる．

3.2 基本要素の抽出

入力されたテキストデータから基本要素である時空間情報，人物情報，感情極性情報を抽出する．これは主に前処理と基本要素の抽出の二段階に分けられ、前処理として形態素解析器である MeCab[4]を用いて分かち書きをする．空間情報

Trajectory Map Summarization Method for Spatio-temporal Text Data
Yuta Ishii†, Ryotaro Okada†, Takafumi Nakanishi†
† Department of Data Science, Musashino University

表 1: 抽出した基本要素

	本文	空間情報	人物情報	感情極性情報
0	喜びというのは、今から十年以前に家出をした、長男の壮一君が、南洋ボルネオ島から、おとうさんに...	ボルネオ	A	pozi
1	喜びというのは、今から十年以前に家出をした、長男の壮一君が、南洋ボルネオ島から、おとうさんに...	日本	A	pozi
2	それから十年間、壮一君からはまったくなんのたよりもなく、ゆくえさえわからなかったのですが、つ...	ボルネオ	A	pozi
3	」壮二君が思いついたわな［#「わな」に傍点］というのは、去年でしたか、おとうさまのお友だちで...	アメリカ	A	pozi
4	建物のほうは、洋館はもちろん、日本座敷も雨戸がひらかれ、家中の電燈があかあかと庭を照らしてい...	日本	A	pozi
5	名探偵明智小五郎［#「名探偵明智小五郎」は中見出し］ネズミ色のトンビに身をつつんだ、小がらの...	左門	C	pozi
6	不安の一夜［#「不安の一夜」は中見出し］日下部左門老人が、修善寺でやとった自動車をとばして、...	修善寺	C	nega
7	不安の一夜［#「不安の一夜」は中見出し］日下部左門老人が、修善寺でやとった自動車をとばして、...	谷口	C	nega
8	小林少年が東京駅にやってきたのは、先生の明智小五郎を出むかえるためでした	東京	C	pozi
9	怪盗の巣くつ［#「怪盗の巣くつ」は中見出し］賊の手下の美しい婦人と、乞食と、赤井寅三と、気を...	代々木	C	nega
10	」そして、ふたりは、明智夫人のゆるしをえたうえ、ポーチのところへ出たのですが、壮二君はいきな...	門内	A	pozi

と人物情報としては MeCab の品詞細分類 2 において地域を表す単語と品詞細分類 3 において名を表す単語を抽出した．人物情報に関しては主要人物のみに限定するため出現頻度の高い 7 名の名前を抽出した．感情極性情報の抽出は，日本語評価極性辞書を利用した Python 用 Sentiment Analysis ライブラリである oseti を用いておこなった．極性値は文を単位として抽出した．値が 0 未満の場合をネガティブ，0 より大きい場合をポジティブとした．

3.3 軌跡地図の生成

抽出した基本要素を基に軌跡地図を生成し，google earth に反映させる．これは必要情報の取得と kml ファイルの書き込みの二段階に分けられる．必要な情報は各空間情報の緯度経度と画面の拡大率である．緯度経度の取得は Geocoder[5]と呼ばれるライブラリを使用した．画面の拡大率に関してはその地名が国名だった場合とそれ以外に分類し，国名の場合はズームアウト，それ以外の場合はズームインになるようにそれぞれに設定した値を割り当てた．必要情報を取得後，kml ファイルの書き込みを実行する．空間情報の出現順に沿って，地図の中心点を空間情報に対応する地点に移動させる．小説の出現順に空間情報を軌跡地図上に表現することで時間情報を反映させている．人物情報は人物名の出現頻度順に上位 7 名を A から G のアイコンに割り当て，地図上に表示する．感情極性情報はポジティブの場合は赤，ネガティブの場合は青，と色を割り振り，アイコンの色に反映させる．画面右側には抽出した各空間情報の抽出元となる本文を表示させる．これにより文章に限らない時空間情報，人物情報，感情極性情報の基本情報からなる可視化による要約が可能となる．

4. 実験

実験として入力するテキストデータとしてインターネット上に公開されている青空文庫[3]の作品である江戸川乱歩の「怪人二十面相」を用いた．この作品から抽出した基本要素の中からいくつかの例を表 1 に示す．表 1 より小説の重要な

図 2: 軌跡地図による可視化の例

基本情報が抽出されることが確認できる．しかしながら，地名である「麻布」，「サンダカン」「羽田空港」，「警視庁」「戸山ヶ原」，「伊豆半島」，「下田街道」，「外務省」，「国立博物館」，「早稲田大学」，「明治神宮」が抽出されなかった．これは MeCab の辞書の充足をおこなうことで改善できると考えられる．

表 1 より，人物情報を確認すると，A から G までに割り振ったにも関わらず，A と C のみしか確認できなかった．これは，本文中には多く登場した人物名が空間情報を含む文章には登場しなかったからだと考えられる．感情極性情報に関しては，感性に合致した値が抽出された．

これらの基本要素を反映させた軌跡地図を図 2 に示す．本地図上では画面左下のボタンを操作することで各情報を出現順に確認することができる．怪人二十面相は日本を舞台とした物語だが，軌跡地図を見ると世界的に地点が広がっていることが分かった．

5. おわりに

本稿では時空間情報を含むテキストデータを対象とした軌跡地図要約方式を示し，文章のみに限らない新たな要約方法を示した．本方式により，文章読解を苦手とするユーザの内容理解を補助することが可能となる．

今後の課題としては文章の重要度による可視化情報の絞り込みが挙げられる．

参考文献

[1] 奥村 明俊, 池田 崇博, 村木 一至, 5W1H 情報抽出・分類によるテキスト要約, 自然言語処理, 6(6), pp. 27-44, 1999.
[2] 山崎 竜平, 松田 純一, 水上 嘉樹, 多田村 克己, テキストからの情報抽出及びその可視化表現手法の開発, 画像電子学会研究会講演予稿, 2009, 09-02, pp. 69-74, 2010.
[3] 青空文庫, https://www.aozora.gr.jp/.
[4] MeCab, http://taku910.github.io/mecab/.
[5] Geocoder, https://geocoder.readthedocs.io/index.html.