平成25年度　卒業研究論文

題目　地図化による旅行記の理解支援に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　永澤　勇樹

学籍番号　　10024113

提出年月日　　平成26年2月13日

目次

図目次

表目次

第 1 章

序論

本章では，研究背景と研究目的について述べる．

1.1 研究背景

旅行する前に行う手順として，まずは旅行先を決める．旅行先が決まったら次に行う事は，そこまでの移動手段を決める．そして，宿泊や交通手段を確保したら，旅行当日に出発するという事が多い．旅行先を決める際に，旅行先に関する情報を探すのに苦労する事が多いと考えられる．また，交通手段に関する情報を探して決定するのも苦労する事も多いだろう．ツアー会社が提供するパッケージ商品を用いないで長旅を行う場合，旅行先や交通手段，宿泊情報や観光情報等を自分で調べる必要がある．これらの情報は，公式サイトやパンフレット等を用いることで容易に得られるであろう．

だが，旅行先や交通手段を決めるための情報源は，公式サイトやパンフレット等の広く一般に向けた媒体からだけではなく，既にそこに行った事のある人たちの個人的な体験談からも情報源となり得る．具体的には，個人が発信する情報として，Web 上の旅行記を活用できないかと考えた．まとめサイトや評判を集めたサイト等もあるが，個人で公開されている旅行記に注目した．旅行記には，旅行で訪れた現地の写真や現地の様子を伝えている事が多いため，場所に関する情報が多く含まれている．また，作者がその土地で感じた主観的な事等も書かれており，旅行先についての情報を前もって調べる時に旅行記は参考になる．主観的な事についても書かれている点が，公式サイトやパンフレット等とは大きく異なる．

しかし，作者によっては，詳細を伝えようとした結果，文字数が非常に多くなってしまった旅行記もある．長い文章を何の抵抗も無く読むことはそう簡単なことではなく，有益な情報を見落とす事も考えられる．また，個人で Web 上に公開されているため，企業が運営しているホームページとは異なり商用ではなく，必ずしも読み手に理解させる必要が無い．その上，文章の書き方は基本的に自由であるため，同じ意味のことが書かれていても，作者が異なると文章も異なってしまう．書き方も上手いとは限らず，読み手に全然伝わらない事もあり得る．従って，Web 上の旅行記は，情報を見落としやすいという問題点があると考えられる．

Web 上の記事では，有益な情報が長い文章に埋もれてしまっていることがある．これでは

せっかくの情報も，活かせずに終わってしまう．このような長い文章から，有益な情報を得るためのハードルを下げられないかと考え，第一歩として旅行記を対象とした．

1.2　目的

一般的には，文章のみの記事だけではなく，図も用いる事でより理解を促すことが出来ると言われている．この事から，図を用いる事で旅行記の理解支援が行えるのではないかと考えている．具体的には，行程を地図に表す事によって，理解支援を行う．ここで言う行程とは，旅行記における作者の移動経路である．以上から，抽出した行程を地図に表す，地図化を行う．

旅行記は場所情報の他に，旅先で撮った写真や旅先の状況等の情報もある．まずはこれらの情報が混ざっている中から行程を抜き出す事で，作者の移動経路を見ながら読むことが出来て，その結果旅行記が読み易くなるのではないかと考えた．最終的には図 1.1 のように，抽出した行程がマッピングされた地図を思い浮かべながら読んでいる状態になれば良い．

図 1.1　最終出力のイメージ

本研究では，行程をより簡単に抜き出すことが出来るように，鉄道旅行に特化させる．主な理由は，移動地点を地名一般ではなく，鉄道駅に限定することが出来るからである．一方で，乗り換えなどによって目的地までまっすぐ行かないことも多い．従って，出発地と目的地だけでなく経由地に関しても書かれている事が多くなり，地名が多く出て来るため，他の交通機関を主に使用したときと比べて，地図に複雑な経路が描けるのではないかと考えられる．以上から，地図化による旅行支援がより有効ではないかと考えた．

第 2 章

関連研究

岡本健の「観光情報革命時代のツーリズム (その 4): 旅行情報化世代 [1]」によると，インターネットを利用して，Blog や Web の検索結果等から旅行先の情報を得ることが多くなってきている．そして今後も，Web コンテンツから旅行先の情報を得ることが多くなるだろう．また，このような状況に伴って，旅行を扱う Blog や Web の検索結果から情報を抽出する研究も行われている．

郡宏志らの「ブログからのビジターの代表的な行動経路とそのコンテキストの抽出 [2]」では，多数の京都観光に関するブログから，代表的な経路を抽出して，それを地図に表示するという研究である．

また，石野亜耶らの「旅行ブログエントリからの観光情報の自動抽出 [3]」では，Blog から，観光地に関する情報 (お土産や観光名所等) を抽出して，それらの外部リンクを表示するという研究である．

安村祥子らの「blog マッピングを用いたイベント情報抽出 [4]」では，Blog からイベントを，Blog 内に書かれている地名を利用することで，Blog からイベント情報を地図にマッピングできるようににするという研究が行われている．

これらの研究では，個人の Web ページから旅行に関する情報を抽出している点では，本研究と同じである．しかし，いづれも旅行先に関する情報を扱う一方で，旅行先までの過程には触れていない．文献 [2] のように行動経路を抽出している研究もあるが，旅行の目的地に関することであり，旅行の開始地点から目的地までの行程を扱っていない．また，本研究では旅行記を対象としているが，Blog も Web 上の旅行記も内容としてはあまり変わりがないと考える．これらの研究の対象は旅行 Blog であるが，旅行記に対しても同じ手法が使えるだろう．

第 3 章

提案手法

本章では，提案手法について述べる．

3.1 旅行記の構成

旅行記は，以下の 4 つの要素で構成されていると言える．

- 作者が旅行中に撮った写真
  旅行先で撮影された写真であり，旅行先の様子を表している事が多い．
- 旅行における作者の移動経路
  鉄道旅行の場合，「どの駅から」「いつ」「どの列車に」「どこまで」乗ったかが，全てではないものの旅行記中に書かれている．これらの文は，作者の移動情報である．また，これは鉄道に限らず，他の公共交通機関についても当てはまる．
- 現地の様子の説明
  作者が旅行先または途中で起きた出来事の説明である．
- 現地における作者の感想など主観的な事
  作者が旅行先または途中で，感じたことである．

写真と現地の様子，感想等はその場所の様子を知ることができる項目である．移動経路は，写真や現地の様子に場所情報を与える項目である．これらの情報を本文から抽出して，旅行記と併記した状態で見ることで，旅行記の理解支援が行えるのではないかと考えた．この 4 つの要素のうち，移動経路を抽出することが理解支援に最も効果的であると考えた．抽出した移動経路から，この旅行記の行程を表示する．ここで言う「行程」とは，旅行記における全移動経路である．また，駅から駅まで移動した情報を「経路」として使っている．「経路」を組み合わせたものが「行程」である．

3.2　支援の方法

旅行記に掲載されている写真と説明している現地の様子は，その場所がどのような場所なのかを知る手がかりとなる．但し，写真と説明文のみでは，場所情報が無い場合が多い．そのため，場所に関する情報が必要となるが，場所情報を得るためには旅行記を読み，それぞれの文がどの場所のことについて指しているのかを読み取らなければならない．場所情報とは，具体的には作者の移動経路であり，作者がどこにいるのかを本文を読みながら把握する必要がある．しかし地図等，一目で場所情報が読み取れる図を併記しておらず，文章のみに場所情報が記載されている旅行記が大多数を占める．

一般的には，図を用いることで理解することが容易になると言われている．この事から，旅行記と図を用いることで，理解支援が出来るのではないかと考えた．旅行記から図示できる要素として移動経路がある．そして，それを地図に表示して図示するという方法がある．以上から，作者の通った駅を地図にプロットする方法で，理解支援を行うことを考えた．行程を図示するためには，まず初めに旅行記から，どのような経路を通ったかを読み取る必要がある．具体的には，どの駅からどの駅まで移動したという経路情報を抽出して，乗車駅または降車駅，乗換駅を順に繋げて行程を記す．図 3.5 が，行程を図示した例である．

3.3　行程を抽出

行程を地図化する前に，まずは旅行記の本文から行程を抜き出す必要がある．具体的には，旅行記から作者が移動した経路を抽出する．各経路の起点または終点には，鉄道駅を利用する．ここで言う鉄道駅は，Wikipedia の「日本の鉄道駅一覧」[5] の項目に書かれている駅名である．この鉄道駅から，駅名リストを作成する．

旅行記から行程を抽出するためには，旅行記を構文解析する必要がある．構文解析を行う方法として，形態素解析を用いる方法と，パターンマッチングを用いる方法の二つがあるが，本研究ではパターンマッチングを用いた．理由は構造が簡単であるからであり，地名や駅名等の固有名詞を容易に追加できると考えたからである．旅行記の文体は作者の自由であり，作者ごとに異なる事がある．また，同じ意味の文章でも表現方法は多様である．しかし，Web を通して不特定多数に公開する以上，相手に伝えようとする意図はあるから，多少の規則はあると考えた．以上の事から，作者が移動したと読み取れる文章をパターン化して，それをマッチングする方法を採用した．マッチングによって得られた文章中の駅名を，出発した駅または到着した駅とする．

旅行記から駅名を抽出するために，動詞に着目した．具体的には「乗車」「到着」「乗換」「出発」「移動」これらの 5 つを表す動詞ごとに異なるパターンを用意して，通った駅かどうかを判断する．パターンには駅名の他に，路線名や列車名，時刻表現も利用する．これらを用いて，旅行記から通ったと考えられる駅を抽出する．用いたパターンの一例を以下に記す．また，正規表現で表した全てのパターンを付録 A に添付する．

- 「時刻」発の「列車」に (乗車)
- 「時刻」に「駅」に (到着)
- 「駅」に「時刻」に (到着)
- 「駅」に (到着)
- 「駅」で「列車」に (乗り換え)
- 「時刻」に「駅」を (出発)
- 「駅」を「時刻」に (出発)
- 「駅」を (出発)
- 「列車」で「駅」に (移動)

ここで言う「時刻」は，時刻表記を正規表現にしたものの他，「定刻」や「–分遅れ」等も含まれる．「駅」は駅名リストの駅であり，この情報から乗降駅を求める．「列車」は例えば，「函館本線札幌行き最終特急列車」や，「特急すずらん 2 号」等，列車に関する情報である．

これらのパターンを用いずに，旅行記から単純に駅名を抜き出すことを行うと，実際に通っていない駅が多く抽出されてしまう．その理由は，旅行記に書かれている全ての駅を通ったとは限らないからである．例えば，旅行先の様子を説明するために駅名を用いている場合が挙げられる．または，一般名詞と紛らわしい駅名があり，それを抽出してしまうからである．

以下に上記のパターンを用いずに行程を抽出する手法の一例として，

- 単純に駅名リストの駅名を抜き出して繋げた結果
- 動詞を用いずに駅名に「○○駅」と続く単語を抽出した結果
- 移動または到着を表す動詞よりも，2 文字より前にある駅名を抽出した結果
  この 2 文字は助詞を想定している．正規表現で表すと「(駅名).{0,2}(動詞)」である．

と，パターンを用いた結果と実際の経路を比較した表 3.1 を記す．例文は，付録 B に添付する．

表 3.1 から，単純に旅行記本文の駅名のみを抽出した結果は，実際の経路とは全く関係のない駅が多く抽出された．「○○駅」を抽出した結果は，実際に通った駅を含んでいる．しかし，他の旅行記では，通った駅全てに「駅」と付いていない旅行記が多い．そのため，この旅行記では上手く抽出できたが，他の旅行記では上手く抽出できない可能性がある．駅名と動詞の組み合わせでは，実際の経路とほぼ同じではあるが，乗り換えた列車について言及している「東室蘭」は抽出できなかった．駅名と動詞の間に 2 文字よりも多い列車情報が挟まれていたためである．以上から，乗り換えた列車や，到着時刻についても言及していることを考慮して，複雑なパターンを用いた．

通った駅の順序については，例外があるが実際に通った駅名の順序と，旅行記に書かれている駅名の順序は一致することが多い．そのため，通った駅の順序については，旅行記に書かれている駅の順序と同じと仮定する．例えば，表 3.1 から，鷲別，東室蘭の順に駅が抽出された場合は，鷲別→東室蘭という経路を抽出したとする．

表 3.1　比較実験の結果

実際の経路	駅名のみ	○○駅	駅名 + 動詞	複雑なパターン (本研究で用いた)
鷲別	中間	鷲別	鷲別駅へと	鷲別駅到着
	鷲別駅		鷲別駅	
東室蘭	室蘭	東室蘭		東室蘭駅で特急 S 北斗 12 号に乗り換え
函館	千歳	函館	函館駅で	函館駅で乗り換え
	東京	知内		
青森	苫小牧	青森	青森駅	青森駅到着
	東室蘭駅		青森駅で	
	函館		青森駅	
秋田	札幌	秋田	秋田駅	秋田駅到着
	長万部			
	函館駅			
	青森駅			
	青森			
	青森駅			
	鷹ノ巣			
	大館			
	東能代			
	秋田駅			

3.4　地図にプロット

行程を地図に表示する手順について説明する．行程を抽出した後は行程を地図にプロットする．そのためにまずは，抽出した駅から位置情報を取得する必要がある．駅名から位置情報を取得するのに，HeartRailsExpressAPI[6] を利用した．この API は，路線や線路データ等の地理情報を，XML や JSON 形式により無料で取得できるサービスである．地理情報は具体的に言うと駅名の他に，乗り入れ路線名，緯度と経度，隣の駅，所在県である．このサービスを利用して，駅名を問い合わせて緯度と経度を取得する．例として，駅名「東室蘭」を問い合わせた結果を図 3.1 に記す．

```
<response>
  <station>
    <x>141.026212</x>
    <next>輪西</next>
    <prev nil="true"/>
    <y>42.349291</y>
    <line>JR室蘭支線</line>
    <postal>0500083</postal>
    <name>東室蘭</name>
    <prefecture>北海道</prefecture>
  </station>
  <station>
    <x>141.026212</x>
    <next>登別</next>
    <prev>本輪西</prev>
    <y>42.349291</y>
    <line>JR室蘭本線</line>
    <postal>0500083</postal>
    <name>東室蘭</name>
    <prefecture>北海道</prefecture>
  </station>
</response>
```

図 3.1　HeartRailsExpressAPI の問い合わせ結果の一例

図 3.2　日本中の市役所前駅

緯度と経度を取得した後は，それを利用して抽出した駅を地図にプロットする．しかし，プロットする際に問題がある．駅名が同じでも，場所が異なる駅が存在する場合は，どちらの駅かを判断しなければならない．例えば「市役所前駅」は日本国内において 9 駅存在する．図 3.2 はそれぞれの「市役所前」の位置である．なお，駅名が一文字でも異なる場合は，同名とは扱わない．例えば「市役所前駅」の他に「市役所駅」や「(自治体名) ＋市役所前駅」があるが，これらは異なる駅名として扱う．しかし，どちらの駅なのかについての情報が文章中に無い旅行記が大半を占める．それでも，判断の基準として文章中の情報で判断することは可能であるが，実際に旅行記を読む場合において，前後に通る駅の位置からどちらの駅かを判断している事が多い．そのため，前後に通る駅の位置関係を用いて判断する事とした．同名の駅が複数存在する場合は，基本的には直前に通った駅から最も近い駅を次に通る駅とする．各同名駅はお互いに遠く離れている，それに加えて，一度の乗車で一気に遠くへは行かないという前提である．

この API は，駅に乗り入れる鉄道路線ごとに異なる駅扱いである上，たとえ同駅でもわずかに緯度と経度の値が異なる場合がある．そのためまずは，緯度経度の値が近い駅を同じ駅と見なす．ここでは，半径 0.005 度以内の駅は同一駅扱いとした．また，鉄道会社が別々であっても，同一駅として扱う．この操作により，複数の同名駅が無い場合は，この駅を通った駅とする．一方で複数の駅に分けられた場合は，直前に通った駅との距離を比較して近い方を通った駅とする．1 番最初に抽出された駅で候補が複数ある場合は，一旦その駅を断定する処理を飛ばして，2 番目以降の「複数の同名駅が無い駅」の緯度経度の値を取得して，そこから最後に抽出された駅の緯度経度の値を取得する．その後に，その「複数の同名駅が無い駅」から近い駅を通った駅とする．具体例として図 3.3 に郡山→福島→白石→鹿児島，図 3.4 に郡山→福島→白石→仙台を記す．「郡山駅」は福島県と奈良県の 2 駅，「福島駅」は福島県と大阪府の 2 駅，「白石駅」は北海道 (2 駅) と宮城県と熊本県の 4 駅が存在する．図 3.3 と図 3.4 の例では，最後の「仙台駅」と「鹿児島駅」でどちらの駅かを判断している．

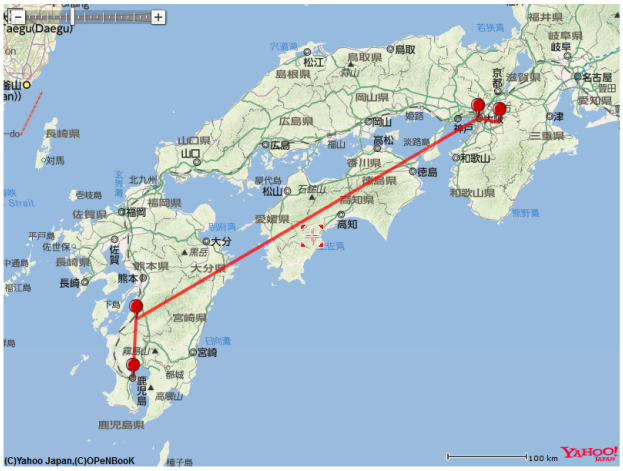

図 3.3　郡山→福島→白石→鹿児島の例

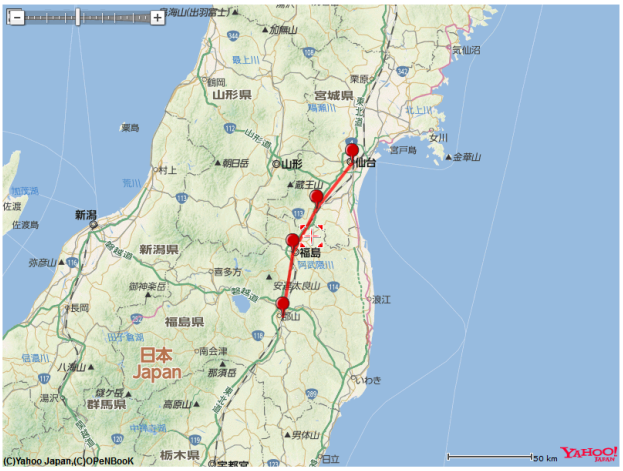

図 3.4　郡山→福島→白石→仙台の例

行程抽出した駅の緯度経度の値を取得した後は，Yahoo!地図 API[7] を利用して，駅を地図にプロットする．これは，JavaScript で Yahoo!地図を表示して，地図上にマーカー等を載せることが出来る API である．HeartRailsExpressAPI によって得られた駅の緯度と経度を指定して，地図上にマーカーをプロットする．その後，それぞれのマーカーを順番に線で結ぶことで，行程を地図に表す．本来はより分かりやすい UI の用意する，あるいはより理解し易い地図の表示をする予定であったが，ここまで手が回らなかった．最終的には，経由駅や列車に乗車した情報の抽出結果 (TXT ファイル) と，地図に行程がプロットされた HTML ファイルが出力される．出力結果を図 3.5 に記す．

==この旅行記の要素()==

宇都宮行きに乗車
大船行きに乗車
磯子で乗り換え
大船到着
横浜駅で乗り換え
小山駅で到着
黒磯駅到着
新白河到着
宇都宮行きに乗車
新白河で在来線に乗り換え
3分遅れで郡山到着
福島到着
普通列車に乗車
17:09に仙台駅到着
17:42発の一ノ関行きに乗車
一ノ関に到着
21:02に盛岡駅到着
新青森到着
青森駅到着
東室蘭駅に到着
一ノ関到着

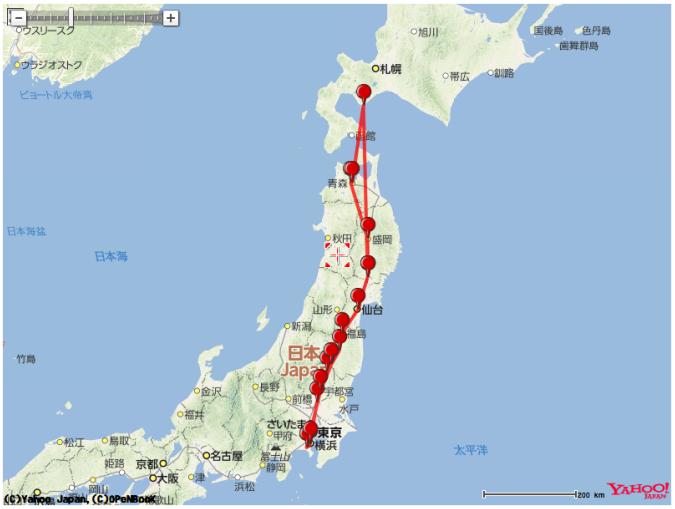

図 3.5　出力結果の行程表と地図

第 4 章

評価実験

本章では，評価実験について述べる．

4.1　評価方法

評価方法については，抽出した行程と実際の行程がどれだけ合っているかを評価する．そして，再現率と適合率を評価する．再現率は抽出した行程がどれだけ実際の行程を再現できているかであり，適合率は抽出した行程がどれだけ正確かを表す．図 4.1 に，再現率と適合率の関係を示す．

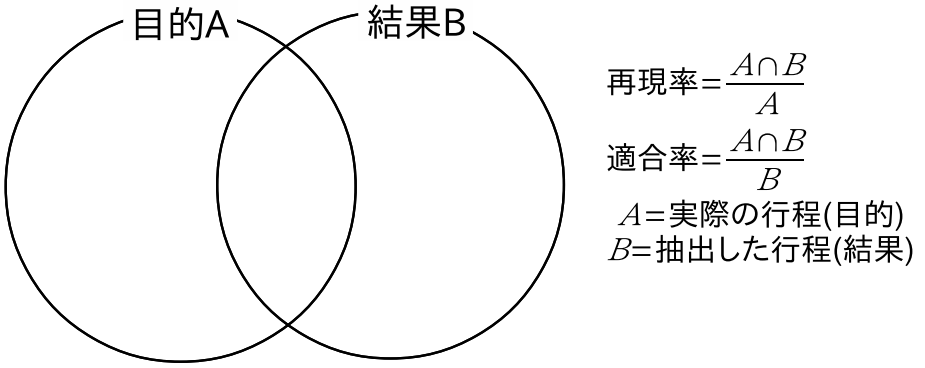

図 4.1　再現率と適合率の関係

A は実際の行程であり，B は抽出した行程である．また，再現率は $\frac{A \cap B}{A}$，適合率は $\frac{A \cap B}{B}$ と表すことができる．

経路の抽出については，列車に乗った移動記録のみを対象とする．具体的には旅行記において，列車に乗車した駅から降車した駅までの経路を正解の経路とする．正解の経路は，実際に読んで乗車したと読み取った経路を正解の経路とする．基本的には，一列車ごとの移動記録になるが，実際には乗り換えが必要な経路でも，乗り換え無しで移動したように書かれている場合は，それで一つの経路とする．また旅行記の本文に，作者が乗った列車情報がまとめられた表やテキスト等がある場合は，その列車情報を正解の経路とする．原則作者が乗降または乗り換えした駅のみを扱うが，旅行記の最初または最後で列車に乗っている状態においては，乗車

している列車の確認できる途中駅を経路と認める．正解の経路と，抽出した経路が一致している場合のみを，正しい結果とする．仮に途中に通ったとしても，間に誤った駅が抽出された場合は正解としない．この理由は，途中駅を入れることで，正解の経路がいくつも出来てしまうからである．正解の経路を一つに絞るために，列車に乗車した駅から降車した駅までの経路のみを正解の経路とする．

4.2　行程抽出の結果

無作為に選んだ 85 件の旅行記に対して，行程の抽出実験を行った．それぞれの再現率と適合率をまとめたグラフを以下の図 4.2 に記す．また，平均値と中央値を以下の表 4.1 に記す．なお，85 件のうち 11 件の旅行記は，経路が一つも抽出できなかったため適合率の値が異常値となったので，図 4.2 と表 4.1 から除外した．

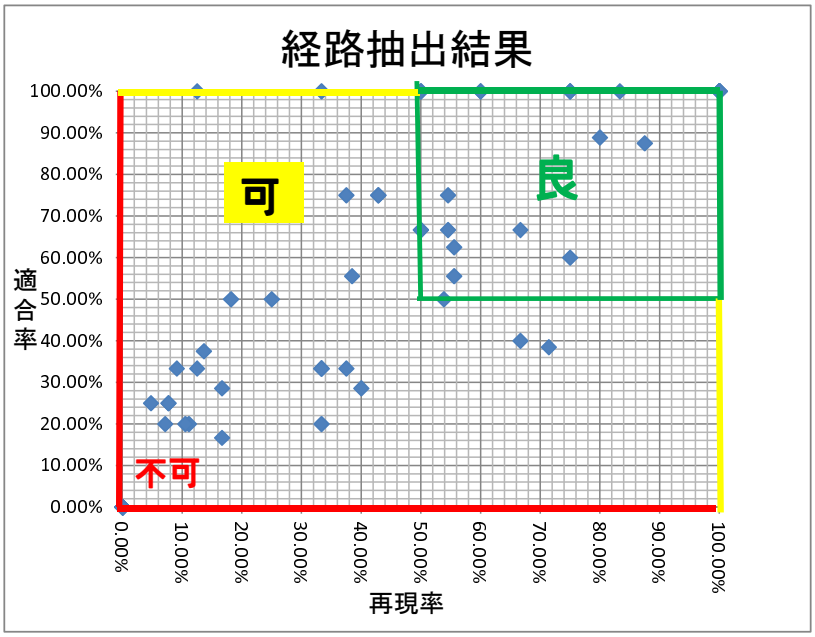

図 4.2　再現率と適合率の分布

表 4.1　再現率と適合率

	再現率	適合率
平均値	35.52%	45.56%
中央値	33.33%	35.42%
分散	0.115	0.151

図 4.2 と表 4.1 が表している事は，それぞれの平均値は低いものの，結果が良い場合と悪い場合で分かれているという事である．旅行記の実経路数ごとに結果が大きく異なるのではないかと考え，旅行記の実経路数ごとに再現率と適合率の集計を行った．実経路数ごとの平均値のグラフを図 4.3 に記す．また，図 4.4 は実経路数ごとの平均値から，実経路数当たりの抽出経路数と正解経路数を 0.1 単位で表している．図 4.1 から，正解経路数 $(A \cap B)$=実経路数 $(A)\times$ 再現率，抽出経路数 $(B)=\frac{\text{正解経路数 } (A \cap B)}{\text{適合率}}$ となる．

実経路数ごとに再現率と適合率を求めた結果，図 4.3 からは実経路数が 1 の旅行記は結果が悪く，実経路数 2 から 6 では揺れがある．実経路数が 7 から 10 では再現率は約 50%，適合率は約 60% で安定しているが，実経路数 11 から再現率と適合率ともに下がり，実経路数 12 以降は低い水準となっている．実経路数が 14 以降はサンプル数が少ないものの，再現率と適合率は共に低い．

また，図 4.4 からは，実経路数 9 までは抽出経路数，正解経路数共に上昇しているが，それより先は抽出経路数が下がり始めており，実経路数が 13 までは正解経路数 5 より多く抽出できない．実経路数が 14 以降はサンプル数が少なく，変化が激しいため正確には分からないものの，正解経路数は少ないと言える．このことから，提案した手法は実経路数が 2 以上 10 以下，特に実経路数 7 から 10 の旅行記に有効だと言える．

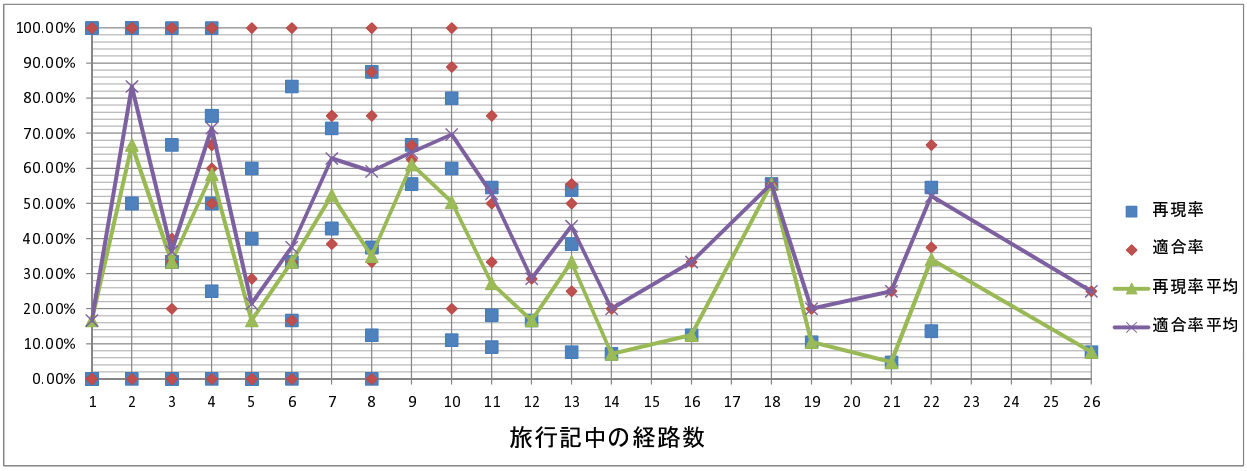

図 4.3　経路数ごとの再現率と適合率とそれぞれの平均値

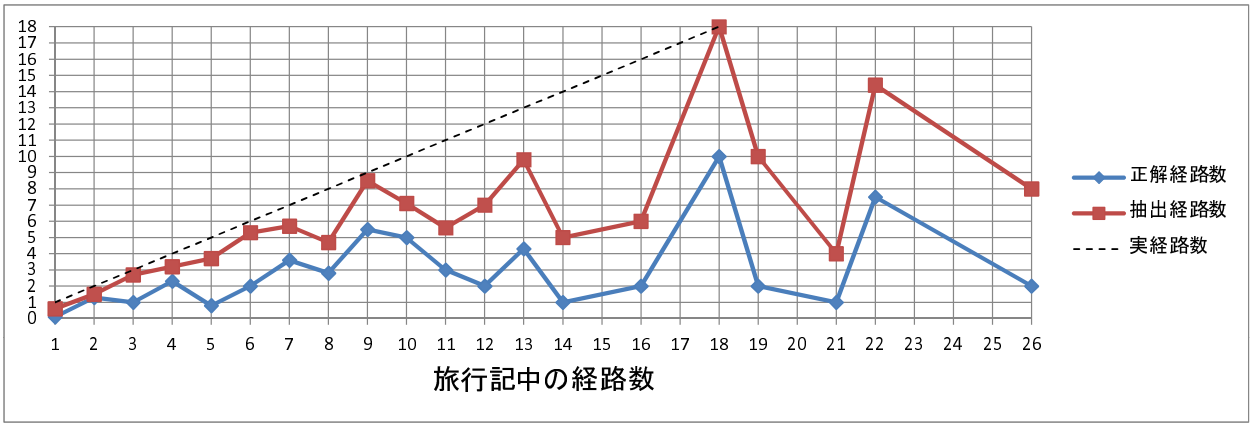

図 4.4　経路数ごとの平均抽出経路数と平均正解経路数

4.3　抽出結果の詳細

再現率と適合率を元に，各旅行記の抽出結果を以下のように 3 段階に分ける．上記の図 4.2 および下記の図 4.5 において，赤は再現率と適合率のどちらか一方が 0% または抽出できなかった旅行記で，「不可」とする．青は再現率 50% 以上かつ適合率 50% 以上であり，「良」とする．黄色はそれ以外であり，「可」とする．それぞれの範囲は，図 4.2 に記している．以下の図 4.5 に，経路数別に各評価の旅行記が含まれる割合を表したグラフを示す．

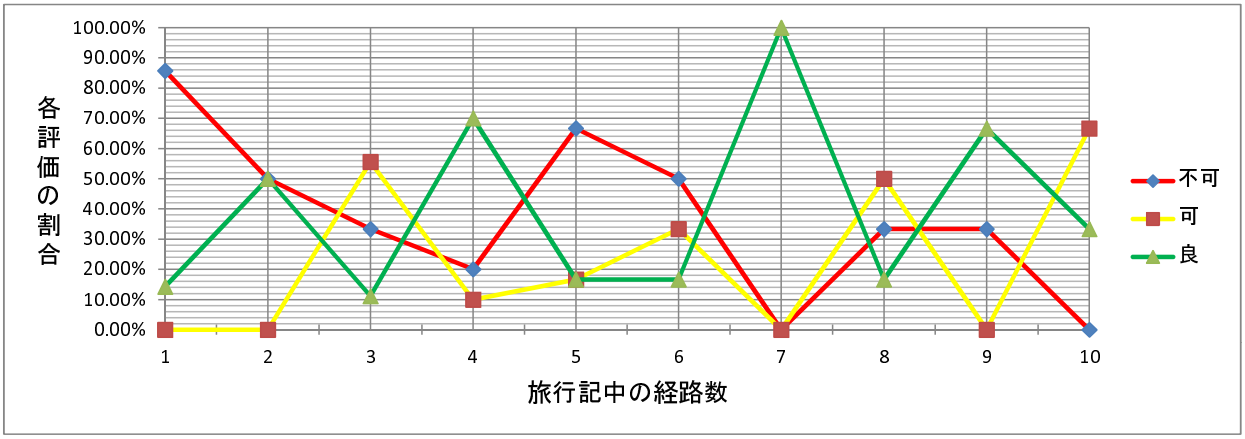

図 4.5　経路数ごとのクラス分け結果

傾向としては，経路数が少ないと不可の割合が高いが，経路数が多くなってくると，可と良の割合が高くなってくる．但し，これだけでは可と良の経路数との関係の違いは不明である．

また，図 4.2 から抽出結果について，無効の旅行記を含めて，5 つのエリアに分けることができる．これらのエリアと，経路数に関する特徴は以下の通りである．

1. 適合率が 80% 以上となる旅行記
   全経路数が少ない旅行記が大半を占める．また，経路数が 11 以上で，適合率が 80% 以上となる旅行記はない．
2. 評価が「良」であり，適合率が 80% 未満となる旅行記
   経路数が 9 以上の旅行記が 60% を占めている．
3. 評価が「可」の旅行記
   どの経路数においても分布しており，経路数はあまり関係がない．
4. 再現率と適合率のどちらかが 0% の旅行記
   経路数が 1 つの旅行記が多く，70% 以上を占める．二番目に多いのが経路数 5 つの旅行記である．このエリアの旅行記の経路数は全て 8 つ以下である．
5. 抽出結果が無い旅行記
   経路数が 2 以下の旅行記が半分以上を占める．また，経路数が 10 以上の旅行記は無い．経路数との関係については，「両方 0% の旅行記」とあまり変わらない．

第 5 章の考察では，これらの特徴を踏まえて，改善案を提案する．

4.4　地図化の結果と詳細

地図化の評価については，同名駅のうち正しい駅を通っているかどうかを評価する．実験の結果，85 件の旅行記中，誤った同名駅を表示した旅行記は 1 件である．なお，85 件の抽出結果中，同名駅が存在する駅を通る旅行記の総数は不明である．

結果の詳細は，草津駅 (滋賀県) から相生駅までの経路をプロットする際に問題が起こった．相生駅は滋賀県と岐阜県にあり，兵庫県の相生駅が正しい駅であるが，誤った駅である岐阜県の相生駅を表示した．草津駅 (滋賀県) と，兵庫県の相生駅と岐阜県の相生駅の直線距離は，岐阜県の相生駅の方が短いということである．図 4.6 に，草津駅 (滋賀県) とそれぞれの相生駅の位置関係を記す．また，草津駅は滋賀県の他に広島県にもあるが，こちらは何の問題もなく，実際に通った方の草津駅が表示された．

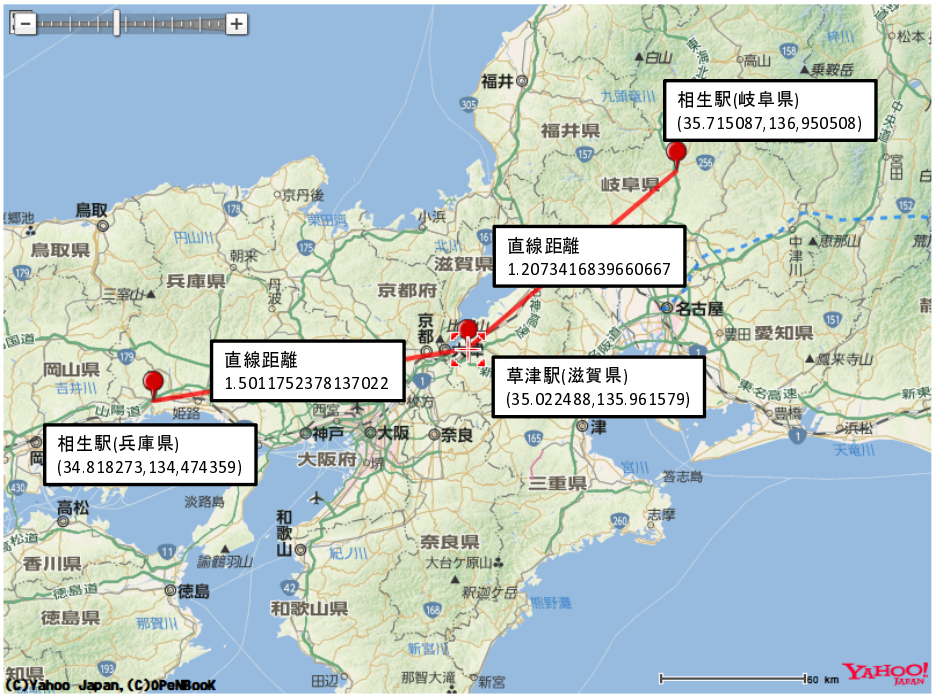

図 4.6　それぞれの相生駅と，草津駅 (滋賀県) から各相生駅の距離

この旅行記では，草津駅 (滋賀県) から相生駅 (兵庫県) の間で乗降しておらず，一本の列車で移動している．実際に 2014 年 2 月の時点で，草津駅 (滋賀県) から相生駅 (兵庫県) までは 1 本の列車で行く事が可能である．一方で，草津駅 (滋賀県) から岐阜県の相生駅へは 1 本の列車で行く事が出来ない．このことから，同名駅問題を解決するために，列車情報も利用できるだろうと考えた．

第 5 章

考察

本章では，実験を行った結果の考察と，改善手法について述べる．

5.1 行程抽出失敗の原因分析

行程抽出において，失敗する原因は主に二つある．

- 誤認…不正解の駅を誤認してしまうこと．適合率が低下する原因となる．
- 見逃し…実際の行程において，通った駅を認識しないこと．再現率低下の原因となる．

これらについては，以下でそれぞれ詳しく考察する．

5.1.1 誤認の主な原因

誤認した駅の大抵は，移動途中に通った駅である．通っているのだが，作者が列車を降りていない駅ついて述べていることが多い．この様な場合は，経路としては通っているのだが，本研究では，誤った経路と見なしている．しかし，文章による情報のみでは，乗降または乗り換えた駅との区別がつかない．

また，駅の様子や列車の様子などの説明や，次の予定を述べている場合も，誤認の原因である．例えば行程が A 駅→ B 駅の旅行記で，「これから B 駅に行く」という予定を先に述べてから「A 駅から B 駅へ行った」と事実を書いている場合では，抽出結果は B → A → B となる．このように，通った駅に問題はないが，駅同士の接続に問題がある結果が多い．特に，一度通った駅が抽出されてしまう事で再現率が低くなる結果が多いので，同じ駅を通らないようにすることで解決ができると考えた．このようにすることで，同じ駅を通る旅行記の結果が上手く行かない可能性があるが，調査対象の 85 件の旅行記中で，同じ駅を通る旅行記は 85 件中約 17 件 (約 20%) と多いとは言えないので，有用だと考えられる．

一般名詞を駅名と誤認した例もある．本実験の結果からは，「駅前」や「フェリーターミナル」「関」の 3 駅である．これらの駅は存在しているため，駅名リストにも含まれている．しかし，旅行記において，これらの具体的な駅を意味しないにも関わらず使われることも多いた

め，誤認の原因となる．これについては，これらの駅を駅名リストから除外するのが有効だと考えられる．なお，本実験において，一般名詞を駅名と誤認した例は，この 3 駅のみである．また，全 85 件の旅行記中で，これらの駅を通る旅行記は無い．

5.1.2　見逃しの主な原因

最も多い抽出できなかった原因は，動詞が無いため抽出できなかったことである．動詞を元に抽出しているので，動詞が無いと駅名を抽出できずに見逃してしまう．動詞を用いずに駅について説明している例として，例えば動詞の代わりに，駅の写真を載せて説明していることが多い．この様な場合は，駅名を画像のキャプションとして使い，写真で駅にいることを表している．動詞が省略されていることもある．例えば，「A 駅から B 駅へ」等，助詞のみで移動したことを表している文である．または，駅名すら出てこない場合がある．特に経路情報を本文中に載せている場合には，本文に駅名や移動したことを省略する傾向がある．しかし，駅名が出てこない一方で，列車名や路線名が書かれていることが多い．列車名や路線名から，どの駅から出発したか，またはどの駅まで行ったかを推測することは可能である．旅行記の経路情報を優先的に抜き出すという方法も有効ではないかと考えた．

パターンマッチングを用いている以上，用意したパターンと合わなかったために駅を見逃した例も多数ある．用いたパターンは，駅名の後に動詞が来る前提であるため，「次に降りた駅は，札幌駅です」のように，駅名の前に動詞がある文章は抽出できない．

また，駅名リストの駅と完全に一致しなければ，駅として抽出することが出来ない．ひらがなが混ざっている，もしくは誤字や括弧が含まれている場合は，駅名と認識されない．駅名にふりがなが振られている場合も見逃している．例えば「音威子府 (おといねっぷ) 駅に到着した」では，(おといねっぷ) があるため抽出できない．また，代名詞が用いられている場合も抽出できない．

「いよいよ」や「やっと」等，感嘆詞等が入っているために抽出できなくなる場合もある．例えば，「駅に到着」では抽出できるが，「駅にやっと到着」では抽出できない．また，日常であまり使わない動詞を用いている例も少数だがある．全てのパターンを適用するのは困難であるため，多く使われていると考えられる例のみを追加もしくは修正する．

経路数があまりにも多いと，一つの文に，複数の乗換駅を羅列する等，途中駅に関する説明が大幅に省略されており，その上，写真も列車情報も無いこともある．このような場合は，地図にプロットされても文章中には殆ど出てこないため，抽出されなくても理解支援に影響が無いと考えられる．

5.2　行程抽出の改善手法

誤認については，経路外の駅が抽出されることはあまりなく，駅同士の組み合わせに問題がある場合が多い．旅行先までの経路または，旅行先からの経路ならば，途中で引き返すことや，同じ駅を通ることはあまり無い．この様な旅行記ならば，旅行記において起点となる駅か

ら，同じ駅を通らない前提で，最も近い駅を辿る方法により，適合率を高めることが出来るであろう．しかし，行程によっては結果が悪くなる可能性もある．例えば図 5.1 のように，実際の行程が新大阪→東京→長野 (赤線) であり，文章抽出による結果も同じである旅行記では，起点の新大阪から最も近い駅を辿ると，新大阪→長野→東京 (青線) となってしまい，正しい結果が誤った結果となってしまう．この改善手法で良い結果となる旅行は，起点から終点までの行程が直線に近い形となり，且つ全ての経由駅が起点から終点の間にある旅行である．例えば図 5.2 のように，実際の行程が福島→郡山→宇都宮→大宮→東京 (緑線) が福島→東京→郡山→大宮→宇都宮の順序 (赤線) で抽出された場合である．

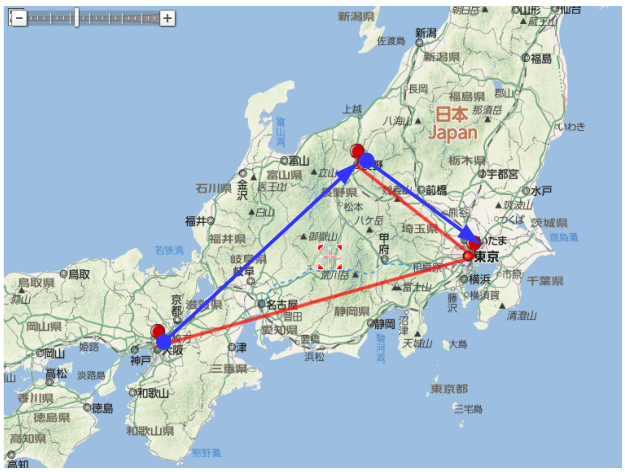

図 5.1　駅間距離を用いて補正して失敗する例 (新大阪→東京→長野)

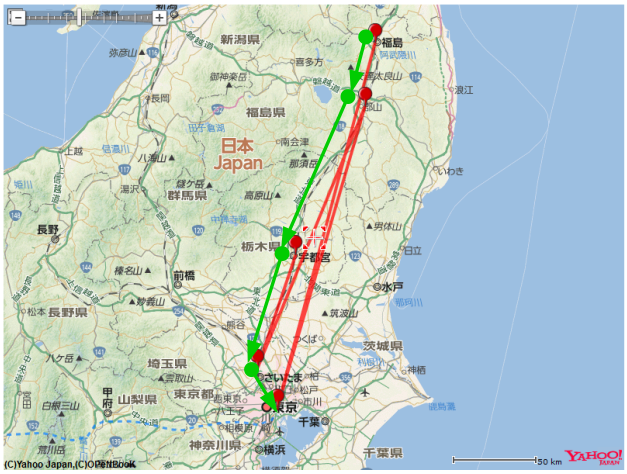

図 5.2　駅間距離を用いて補正して成功する例 (福島→郡山→宇都宮→大宮→東京)

見逃しについては，まずは動詞が無くても，通った駅を判別できるようにする必要がある．また，駅名が無くても，書かれている列車名や路線名から通った駅名を推測するのも有効だろう．これについては，列車ごとの駅リストや，路線ごとの駅リストを作成する必要がある．駅名や列車名，路線名も無い場合はどうしようもない．そして，見逃す駅が減る一方で，誤認する駅が増える可能性がある．旅行記の経路情報を優先的に抜き出すという方法も有効ではあるが，経路情報は作者ごとに形式が異なるので，効果は限定的だと思われる．

5.3　地図化の考察

地図化については，提案した同名駅問題の解決方法について説明する．

提案した手法について，実験結果には無かったものの，4.4 節で挙げた事例以外にも問題がある．同名駅は互いにかなり離れていて，都道府県あたり一駅という前提で 3.4 節で挙げた解決法を提案した．しかし，この前提に当てはまらない駅があり，例として，琴似駅 (JR 北海道)(札幌市地下鉄) と浅草駅 (東武鉄道と都営地下鉄と東京メトロ)(つくばエクスプレス) がある．それぞれの駅は互いに近くにあるとはいえ，歩いて乗り換えると 10 分以上掛かるうえ，互いに連絡されていないため，同一駅として扱うには無理がある．とはいえ，どちらの駅が地図にプロットされても概ね行程を理解できるが，行程抽出に路線名等，位置情報以外の駅情報

を利用する事も考慮に入れると，区別することが望ましい．しかし，互いに近くにあるため，位置情報のみでどちらの駅を通ったかを判断すると，図 5.3 のように，どちらの駅に関しても距離差が僅かとなり誤認する可能性が高い．

図 5.3　各浅草駅 (左の二駅) から成田駅 (右の駅) の距離比較

この解決方法として，路線情報を用いることで，どちらが通った駅かを判断する方法を考えた．前後の駅と同じ路線情報の駅を通った駅とすれば，判断は可能である．路線情報を活用してどちらかを判断することで，解決できるだろう．但し，抽出駅数が少ない場合も考慮に入れて，路線情報でも判断が出来ない場合には，代表駅として問答無用でどちらか一方に絞るということも必要だろう．

外部 API を利用している事に関する問題もある．HeartRailsExpressAPI[6] による問い合わせ結果は，旅客駅のみ有効であるため，位置情報が得られる駅は旅客駅のみ有効である．しかし，Wikipedia の「日本の鉄道駅一覧」[5] には貨物駅も含まれているため，用意した駅名リストにも貨物駅が含まれており，抽出結果に貨物駅が含まれる事もある．そのため，位置情報を取得することができない駅がある事も考慮する必要がある．本実験での例では「仙台港」があった．仙台港はフェリーが就航しており，旅行記に仙台港を通った事について書かれることもあり得る．この問題については，位置情報が取得できない駅をプロット対象から外すことで，解決している．

また，抽出結果を増やすために，よく使われる駅の通称を駅名リストに追加しても，位置情報が得られる駅は変わらない．経路抽出が出来ても，位置情報が得られないため，地図にプロットが出来ない．対策としては，API を用いずに位置情報を得られるようにする方法もあるが，追加した駅名を API に問い合わせる際に，位置情報が得られる駅名に変換してから問い合わせる方法が有効であると考えた．

第 6 章

結論

旅行を行う際には，情報収集が欠かせない．情報源は公式サイトやパンフレット等多数ある．一方で，近年は Web を用いた個人から発信される情報は増え続けており，それらの情報を利用する機会も増えつつある．この中には，既に訪れたことについて書かれている Web 上の旅行記もある．本研究では，Web 上の旅行記を情報源として活用できないかと考えた．旅行記には，写真や旅行先の出来事等があり，情報源として十分活用が出来る．しかし，個人で発信される以上，読み易いとは限らない．そのため，本研究では旅行記から行程を図示することによって，理解支援を行った．まずは旅行記から行程を抽出する必要がある．旅行記に書かれている駅が全て通った駅とは限らないため，行程の抽出は動詞に着目したパターンマッチングを用いた．また「駅名」の他に，「時刻」や「列車」の情報も，経路抽出に用いた．

経路抽出の評価は，抽出した行程が実際の行程とどれだけ正しいかを評価する．抽出の結果は，再現率は 35.52%，適合率は 45.56% と，あまり満足のいかない結果となった．しかし，全く通っていない駅が抽出される事はあまり無く，駅の組み合わせに問題がある場合が多い．このことから，同じ駅を通らない前提で経路をプロットする事と，位置情報による修正で改善できると考えた．

抽出した行程を地図に表す際に，同名駅に関する問題があるが，本研究では前に通った駅から近い駅を通った駅とすることで解決しようとした．この方法ではある程度は上手く行ったようではあるが，この方法では，移動距離が長い場合，または同名駅がそれぞれ近い所にある場合に，誤認する可能性がある．これらは，路線情報や列車情報を用いる，あるいは駅ごとに重みを用いることで解決が出来るであろう．

本研究から言えることは，旅行記から行程を抽出して，それを地図にプロットするには，本研究の方法では不十分となってしまったが，路線情報や位置情報等を併用することで，より精度を上げられるのではないかということである．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教を初めとして，服部研究室の皆様に感謝を致します．また，実験に使った各旅行記の作者の皆様にも，感謝致します．そして，本研究で用いたフリーの API を提供している二社に感謝致します．

参考文献

[1] 岡本 健：観光情報革命時代のツーリズム (その 4):旅行情報化世代，北海道大学文化資源マネジメント論集 Vol.006，pp.6 (2009).

[2] 郡 宏志，服部 峻，手塚 太郎，田島 敬史，田中 克己：ブログからのビジターの代表的な行動経路とそのコンテキストの抽出，ウェブ属性抽出, 夏のデータベースワークショップ DBWS 2006，pp.36–40 (2006).

[3] 石野 亜耶，難波 英嗣，竹澤 寿幸：旅行ブログエントリからの観光情報の自動抽出，2010 日本知能情報ファジィ学会，pp.669–672 (2011).

[4] 安村 祥子，池崎 正和，渡邉 豊英，牛尼 剛聡：blog マッピングを用いたイベント情報抽出，DEWS2007 D8-3，pp.2–6 (2007).

[5] Wikipedia 日本の鉄道駅一覧：http://ja.wikipedia.org/wiki/日本の鉄道駅一覧, 2013 年 7 月 25 日時点の駅データを使用.

[6] HeartRails Express：http://express.heartrails.com

[7] Yahoo! Open Local Platform：http://olp.yahoo.co.jp

付録 A

行程抽出に用いたパターン

行程抽出において，用いたパターンの正規表現を記載する．パターンの末尾にある「乗車」「到着」「乗換」「発車」「移動」は動詞であり，それぞれの中身を表 A.1 で説明する．また，「駅名」「時刻」「列車」「路線」は，それぞれを表す単語である．

- 乗車…このパターンのみ駅名を抽出しない．
  (「時刻」発?車?の? 、?)?+「列車」に「乗車」
- 到着
  (「時刻」[にで、]?、?)?+「駅名」(((には?)|を)?、?「時刻」)?[にで]?「到着」
- 乗換
  「駅名」(で|にて)?、?(「路線」|「列車」)?に?+「乗換」
- 発車
  (「時刻」.{0,2})?+「駅名」を(「時刻」.{0,2})?「発車」
- 移動
  ((「路線名」.{0,2})?(「列車」.{0,2})?(「駅名」[にへ])「移動」

各動詞の一覧
それぞれの動詞の中身は以下の表 A.1 の通りである．活用語尾は一部省略している．

表 A.1　各動詞の説明

「乗車」	「到着」	「乗換」	「発車」	「移動」
乗車	着	乗り換え	発つ	向か
乗る	到着	乗換	出発	向い
乗った	下車	乗り継	発車	移動
乗って	途中下車	乗継	出る	
乗り				

各単語の説明

- 「駅名」
  Wikipedia の「日本の鉄道駅一覧」[5] に載っている駅名
- 「時刻」
  (午前|午後|AM|PM)?([0-9]{1,2}+(:|**時**))([0-9]{0,2}+ **分**?)
  上記のような時刻の正規表現ほか,「10 分遅れ」等，遅れている事を示す表現も含まれている.
- 「列車」 一例:室蘭本線室蘭行き最終普通列車
  「スーパー北斗」等の列車愛称の他,「快速」等の列車種別も含まれている．そのほか,「電車」「列車」「新幹線」等，列車を表す言葉も含まれている．列車愛称は，北海道内を走る列車は全て含まれている.
- 「路線」
  現在は JR の路線のみ対応．廃止路線は含まれていない．また，北海道内の路線のみ愛称名を記載.

付録 B

予備実験で用いた旅行記

3.3 節で，予備実験に用いた旅行記を添付する．内容は，Web ページの内容をテキストとして貼り付けた文である．句読点を変更して，画像のキャプションを箇条書きにして，誘導のリンクを削除して，見出しの文字サイズを大きくして，さらに太字にした他は，原文そのままである．また，画像は添付しない．出典は「はてなグループ」内の日記であり，同グループ内のみ閲覧可能である．

2012 年 11 月 15 日 (木)11:55–作戦開始

朝 2 の授業が終わった後，約 2 か月間かけて作った計画が実行された．本当は旅行の準備を 8 時頃までしたかったんだけど，中間試験なるものがあったので，さっと済ませた．

まずは，工大生協にて昼食を購入．学生食堂だと，間に合わない可能性もあるので．ついでにジュースを買う．あとは，歯ブラシ (用意するの忘れてた)．昼食を食べ終えた後は，鷲別駅へ．ただ，食後の運動は避けたいところなので，図書館で暇つぶし．2 階で新聞読んで面白いネタないかと．13 時前には鷲別駅へと向かう．

13 時頃に鷲別駅到着．室蘭行きの列車が来る前に，その辺をうろうろ．

来た列車はキハ 143 系 (右の写真がそうです．鷲別駅にて)．室蘭での PDC は，これが初乗車だったりする．もっとも，PDC に乗る機会自体がそんなに無かったのだが (千歳線沿線に住んでいたので．そして，いざ乗りに行ったらキハ 201 系が来るという…)．

乗った感想だけど，キハ 150 系やキハ 40 系とは全然違う感じがした．乗る前の印象が，2 編成のキハ 150 系と同じようなものだろ，と思っていたわけで．運賃表示機は液晶，側面の方向幕は LED，放送の声も違う (慣れていないだけだと思うけど，若干暗い感じがする)．共通点は整理券発行機と運賃箱 (この 2 つはキハ 150 系と同じ物だったと思う)，セミクロスシートというレイアウト，塗装くらいかな．

OutofBounds 関東 OB 会

あ，この旅行のタイトルだからね，コレ．写真の整理を容易にするために，旅行に (変な) 題名を付けている．

タイトルの意味だけど，OB と聞かれて真っ先に思いついたのが Out of Bounds(ゴルフで

言う OB) なので．北海道から東京へ行くのに，あけぼのっていう少し外したルートを使ったという意味もある．

2 編成の PDC は，ボックスシートがほとんど埋まるほどの乗車率であった．711 系だった頃とは違い，室蘭-東室蘭と東室蘭-苫小牧を直通する運用が増えた気がする．乗換の有無は乗客にとって重要な要素だと思う，たとえ所要時間が少し長くなっても．

東室蘭駅で特急 S 北斗 12 号に乗り換えて函館へ．久々に乗った気がするキハ 281 系．札幌に行くときは，いつも 14 系客車とか 785 系とかだからなあ．そのうえ，札幌に行く機会が少なくなったので．自由席の窓側，海側の座席を確保．乗車率はガラガラではないが，多いともいえない．海側に座ったのは，そっちの方が眺めが良いから．

- 14:09 頃 (洞爺→長万部) に撮った写真．奥に見える山は駒ヶ岳なのかな．

でも直射日光が当たる．眩しいので，カーテンを閉めていたことが多かった．気が付いたら長万部を過ぎたあたりで，太陽の光は山側に移った．やっと眩しいのから解放されると思ったら，反対側から直射日光が当たる．今度は，カーテンが閉められないっていう問題が発生．やっぱり眩しい．函館駅で乗り換え．次の列車がホームの向こうに待機している．でもすぐには乗らない．撮影したり歩き回ったり．

- 左が S 白鳥，右が S 北斗 (函館駅にて)

乗った車両だけなんだけど，S 北斗よりも S 白鳥の方が混んでいる気がする．立ち客は無かったけど．江差線内は，そんなに速くない．もともとローカル線が故の宿命なのだろうか．カーブも多い．以前，二度にわたって貨物列車が脱線したあの地点では特にゆっくりと．海峡線からは一気に速く．でも速度計とかないから，140km/h が出ているのかどうかが分からない．そういえばこの列車，知内駅に停まるんだった．検札でも，知内までという声があったし (実際，降りた人がいたようだ)．青函トンネルを抜けると，もう真っ暗．トンネルを抜けると (抜けても?) そこは真っ暗闇な世界．青森駅到着前，青森車両センターにあけぼのが待機していているのを見た．

青森駅で途中下車して，夕食を購入．入線していくところを見たかったが，何時ごろにあけぼのが入線してくるのかが分からないため，とりあえず急ぐ．まあ，間に合わなかったけどな (18 時頃に入線したようだ)．

- 青森駅で発車を待つ寝台特急あけぼの (青森駅にて)

乗る車両であるゴロンとシートは一番後ろ．跨線橋は前の方にしかないので遠い．下段のベッドを指定した．二段ベッドは上と決めているのだが，窓が無いので下にした．ただ，ベッドを座席として使った時には進行方向逆向きとなってしまう．JR 時刻表には座席配置図こそあるが，どちらがどっちへ行くのかが分からないからなあ．

青森駅発車時，ゴロンとシートにいた人は俺を入れて 3 人．そのあと，弘前，大館，鷹ノ巣，東能代でどんどん乗ってくる．俺の上は大館から，向かい側 2 席は東能代からの乗客だった．秋田駅到着時でも何人か乗車．この時点で満席となる (少なくとも，ゴロンとシートは)．

秋田駅発車後に寝ることにする．が，眠れない．ベッドが硬すぎるのか (よく考えると，座布団の上で寝ているようなものだもんな．よくロングシートをベッドにしている人がいるけど，寝心地は悪いんだろうな)．あるいはコーヒー (といってもミルクコーヒー．黒なんて飲めません．) を遅い時間に飲んだからか．窓のカーテンが完全に閉まっていない (街灯とか眩しくて) からなのか．または S 北斗の車内で寝た，朝 1 の授業で寝た (要は昼間いっぱい寝た) からか…とりあえず，なかなか眠れなかった．0 時くらいまで起きていたんじゃないかなあ．