

地形補完と方角・地形語ペア抽出を用いた文章からの地図自動生成

遠藤 純玲[†] 服部 峻^{††}

^{†,††}室蘭工業大学 ウェブ知能時空間研究室 〒050-8585 北海道室蘭市水元町 27-1
E-mail: [†]13024034@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

あらまし ファンタジー小説などで架空の世界について書く時、作者は読者に世界観を想像させるための描写をする必要がある。しかし「川は海へと流れ、その海の近くでは川によって運ばれた砂が浜を作る」などの複雑な地形関係を小説に反映させることは難しい。そこで本稿では、文書からの「地形語」と「方角語」の抽出とそれらをパラメータ化する（検出された地形オブジェクトに方角語をセットする）ことによって、文書から地図を自動生成するシステムを提案する。さらに、検出された地形オブジェクト間の関係の種類から推測される周辺地形オブジェクトを補完できるようにするために、本稿では実世界の地形同士の相互影響に基づいた地形補完アルゴリズムも提案する。

キーワード 地形, 地形補完, 地図生成, 自然言語処理, メディア変換

Automatic Map Generation from A Document Using Geographic Complement and Pair Extraction of Bearing-Words and Geo-Words

Sumire ENDO[†] and Shun HATTORI^{††}

^{†,††} Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27-1 Mizumoto-cho, Muroran, Hokkaido, 050-8585, Japan
E-mail: [†]13024034@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

Abstract When writing a fantasy novel about a fictitious world, the author needs to describe the geography to let readers imagine the world view. But it is difficult for an author to reflect complicated geographical relationships between terrains in her/his novel, e.g., “a river flows into a sea, and near the sea, a sandy beach is made of the sand carried by the river.” Therefore, this paper proposes a novel system that generates a graphical map from a textual document automatically by extracting Geo-words and Bearing-words from the document and parameterizing them (detecting geographic objects and setting their bearing). Furthermore, to enable the system to complement the peripheral geographic objects that can be inferred from the type of relationship between the detected geographic objects, this paper also proposes a geographic complementing algorithm based on actual mutual influence between terrains in the real world.

Key words Geography, Geographic Complementing, Automatic Map Generation, NLP, Media Conversion

1. ま え が き

小説などで、作者独自の架空の世界観を書く時、作者は読者に自分の物語を想像してもらうために、「西に山がある」「東に海がある」などの情報を文章中に組み込んで、作中の地形を描写する必要がある。物語中で特別な出来事が起こらない限り、物語の進行中でも、その世界の地形は一貫して同じになるはずであるが、自分が書いた地形の描写を忘れてしまい、地形の場所や方角などが以前の描写と食い違ってしまったたり、作者が地理に疎く、地形の繋がりを不自然に描写してしまい、読者に違和感を与えてしまうなど、地形の描写には気を払う必要があり

難しい。また、「川はいずれ海に流れ込み、その海の近くでは川によって運ばれた砂により、砂浜が出来る」など、近くにある地形が相互影響し合い作られる地形まで考えて世界観を作り込むのは、作者にとって非常に手間が掛かる。そこで、作者が意図した風景・地形描写に基づいた地図を文章から自動生成するシステムを開発することで、作者が描写をより豊かに行うことが出来る考えた。

本システムは、この問題を解決するために、入力した文章から地形と方角を表す単語を抽出して、地形の単語から地形オブジェクトを構成し、地形オブジェクトに抽出した方角を与えて配置することによって、文章に基づいた地図を自動生成する。

そのために、地形語と方角語のペアリングを行うアルゴリズム、および「同じ地形を示しているが何度も出現する地形語」を判断して重複した地形オブジェクトの作成を避けるために、文章中に出てくる同一名の地形を判定するアルゴリズムを導入し、その精度実験を行う。

さらに、作者の作る世界観の構成ルールが、現実世界とある程度一致しているという前提のもとで「地図に描画する時、文書中に描写があった川は自動的に海に繋げ、その川と繋がった海沿いには、文章中に描写が無くとも浜を作成する」など、検出した地形の種類より推測出来る周辺地形を補完する。そのために、地形と地形の間に繋がりがあると考えられる地形オブジェクト同士に関係性を結び、繋がった関係性のある地形同士から生成される地形がある場合に、地形を補完して地図上に表示を行う地形補完アルゴリズムも提案する。

2. 関連研究

本研究に当たって永澤勇樹らの「モバイル端末における旅行記の理解支援のための行程抽出と地図化 [1]」を参考にした。これは旅行記を対象として、駅に着目した旅行の行程を抽出し、地図にマッピングを行うという研究である。この研究では、パターンマッチを用いて駅名や地名を旅行記から取得し、旅行記に出現した順番を考慮して旅行行程を構築している。本研究では、同じくパターンマッチを用いて、駅名や地名の代わりに地形語を取得することで、地図の自動生成を行う手法を提案する。

3. 提案手法

3.1 システムの概要

本章では、ユーザから入力された文書に基づいてその地形描写を地図として自動生成する提案システムについて詳述する。まず、提案システムが行う処理の流れを図1に示す。

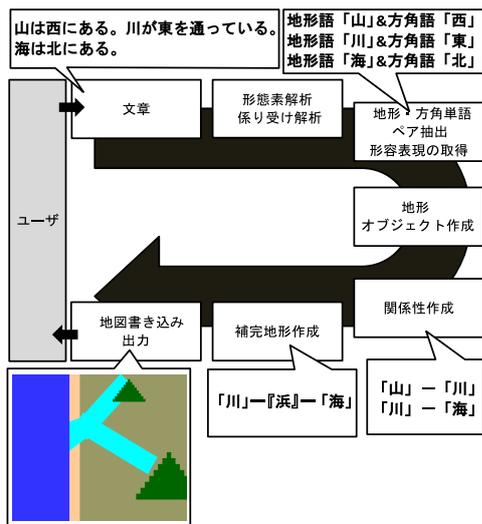


図1 提案システムの処理の流れ

提案システムは図1のように、入力された文書を形態素解析し、地形語と方角語を抽出して、それらのペアリングを行い地

形オブジェクトを生成する。この時、地形オブジェクトに関する特定の形容詞があれば文書から取得する。その後、生成した地形オブジェクト間に関係性を結び、オブジェクト間に補完すべき地形があれば、補完地形を作成した後に地図を出力する。関係性については3.6節にて後述する。

3.2 用語の定義

本節では、本研究で用いる以下の用語について定義する。

(1) 地形語：

システムが文書から抽出する地形を示す用語である。本稿では描画まで行う地形語の対象は「海」「山」「川」とする。また、現行では描画は行えないが地形語として認識する対象として、「砂漠」「街」「町」「森」「平野」「山脈」「王国」「国」「洞窟」を追加する。

(2) 補完地形語：

文書中に明確に出現していなくとも、他に出現している地形語との関わりによって地図上に描画すべき対象となる地形を指す用語である。本稿では「浜」を対象とする。

(3) 方角語：

システムが文書から抽出する方角を示す用語である。本稿では「北」「東」「南」「西」「北東」「南東」「北西」「南西」の8方向と、文書中の語からは抽出されない「方角なし」を対象とする。

3.3 方角探索アルゴリズム

地図上に地形オブジェクトを作成するためには、「どの種類の」地形を「地図上のどこに」配置するのかが示す、方角語と地形語のペアが必要である。本システムでは、地形語を起点として、前後5単語の範囲の中から方角語を探索してペアリングするアルゴリズムを実装した。ここでいう1単語とは形態素解析エンジン「MeCab」を用いて得られた分かち書き解析結果の中の一区切り分を指す(図2)。

北 の 海 は 深 く 蒼 い 色 を た た え て い る 。

図2 形態素解析による入力文書の単語への分割

探索アルゴリズムは、文書の最初から探索を始める。探索範囲内に方角語があれば、それを起点の地形語とのペアリング相手として採用する。方角語は自分が使われたことを示す使用フラグを所持しており、方角語が地形語のペアリング相手として使われるとフラグが立ち、以降再使用されない。探索範囲内に使用可能な方角語が複数ある場合は一番距離が近いものをペアリング相手として採用する(図3)。また、探索範囲内に使用可能な方角語が無い場合は、「方角なし」をパラメータとして地形語とペアリングする(図4)。



図3 方角探索アルゴリズム

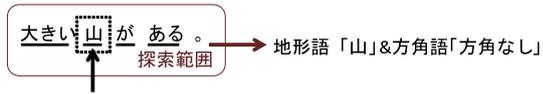


図4 地形語と方角なしがペアリングされる方角探索アルゴリズム

3.4 同一地形オブジェクト判別

「東にヒガシ山がある。その山ではリンゴが採れる」など、文書中に同一であると見なせる地形が2回以上出現することがある(図5)。文書を解析した時、人間が読めば同じであると判断できる地形にそれぞれ別の地形オブジェクトを作成してしまうと、地図として正確ではない。そこで、以前に出現した地形オブジェクトと同一のものを判別し、同一であると判断した場合は新たな地形オブジェクトを作成しないようにする必要がある。

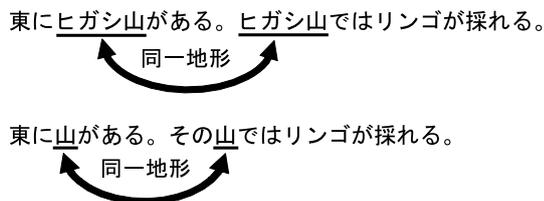


図5 同一地形と見なせる地形語が含まれた文書

本稿では、まず、地形が名前で呼び分けられている場合と、名前が無い場合の2種類を考える。図5にある「ヒガシ山」であれば、「ヒガシ」が名前である。ここで定義する「名前」とは、地形語の前に付く名詞もしくは未知語とする。ここでの名詞や未知語の定義はMeCabに依存するものとする。

まず、名前が付いている場合、名前用のデータベースを用意し、名前と地形語のペアリングを行い、ペアを登録する。以降同じペアが出現した場合は作成しない。名前を用いた同一地形オブジェクト判別の処理を次の図6に示す。

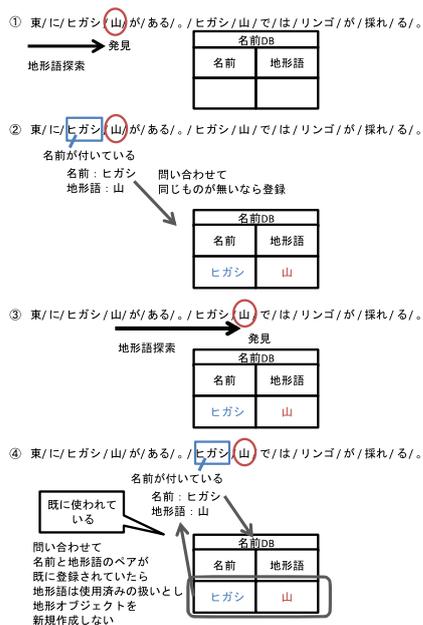


図6 名前を用いた同一地形オブジェクト判別処理

また、名前が無い地形語の場合、地形・方角ペアリングを利用する。文書から検出された順に地形語とペアの方角語をデータベースに登録し、2回目以降に同一の地形語が検出された場合、方角語と地形語のペアリングが既に登録されている状態ならば、その地形語に関してはオブジェクト化を行わない。また、地形語が同一であり、かつ、ペアリング対象の方角語が「方角なし」であった場合も、同一と見なしオブジェクト化を行わない。これは、同じ地形を示す地形語が2回目以降に出現した場合、方角語とともに出現する可能性が低いと考えられるからである。方角ペアリングを用いた同一地形オブジェクト判別の処理を次の図7に示す。

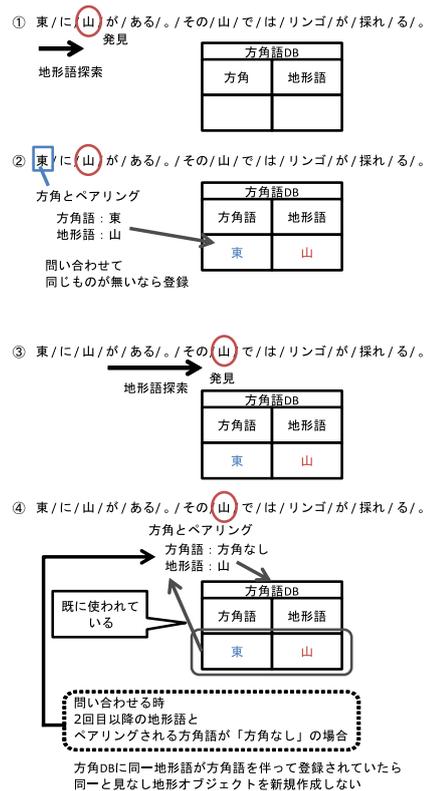


図7 方角ペアリングを用いた同一地形オブジェクト判別処理

3.5 形容表現の取得

本システムの目的は、作者が意図した風景・地形描写に基づいた地図を文書から生成することである。作者が世界観を描写する時に、地形に関する見た目について言及することが考えられる。地図に作者の意図をより反映させるために、地形に関する形容表現を地形オブジェクトの外見に反映させる。本稿では、パターンを用いて地形語の前にある「大小」の一部表現を取得している。具体的には、「大きい」「大きな」「小さい」「小さな」のパターンで大きさの表現を取得し、地図上の地形オブジェクトに反映させる。大きさは「大」「小」「デフォルト」の3段階である。色や形の表現への対応は今後の課題とする。

3.6 地形補完アルゴリズム

例えば、文書の中で「川があり、海がある」と書かれていた場合、おそらくその川は海へとつながっていると推測できる。このように、「一般的に関係があると考えられる地形語」が一緒

表 1 ペアリング精度実験に使用した文書例

文書例	アル山脈は東にあり、隣国との境界線になっている。ここは北をアイエス海、西をジャネットの森に囲まれた天然の要塞なのだ。		
正解定義ペア	東-山, 北-海, 西-森	この文書における再現率	0.333 (=1/3)
取得ペア	東-山, 西-海, 方角なし-森	この文書における適合率	0.333 (=1/3)

に出現した場合、本システム内部で地形語同士を関連付ける。地形オブジェクト同士の関連を「関係性」と呼び、これを本研究では「関係性を結ぶ」と呼称する。

本システムには、文書中に明記されていないが、地形と地形との繋がりを考えた時に「高確率で存在するはずの地形」を自動で地図上に補完する機能を搭載する。これを「地形補完機能」と呼び、地形補完機能に関するアルゴリズムを「地形補完アルゴリズム」と呼称する。本稿で補完される地形は「浜」である。「浜」は「海」と「川」の間に補完される地形である。

地形補完アルゴリズムの処理を図 8 に示す。

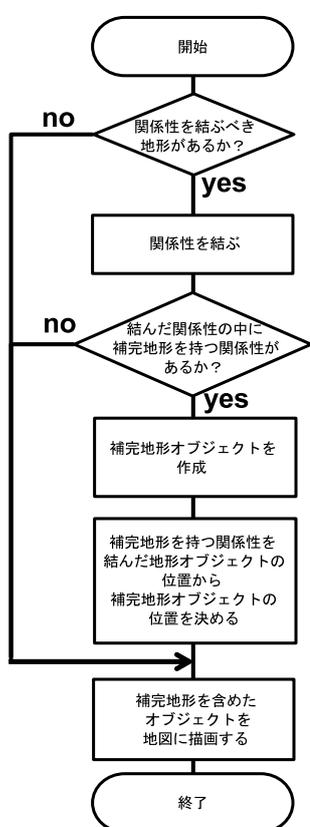


図 8 地形補完アルゴリズム

地形補完アルゴリズムは、文書中の地形語から地形オブジェクトを作成した後に行われる。まず、システムは地形オブジェクト間に関係性を結ぶかどうかを判断する。判断の基準はシステムに入っているデータベースに基づいて行われる。関係性の中には補完地形を生成する種類が存在する。関係性を結んだ後、補完地形を生成するかを判断する。補完地形を生成する関係性が結ばれた場合、システムは補完地形オブジェクトを生成する。その後、関係性が結ばれている地形オブジェクトの位置と種類を参照して、補完地形オブジェクトの位置を決める。最後に、補完地形を含めた地形オブジェクト群を地図上に描画する。

3.7 地形の描画

地形補完アルゴリズムの処理が終わり次第、地形オブジェクトの描画を開始する。ペアリング処理された方角語を用いて位置を決定し、生成された地形オブジェクトを決まった種類ごとに描画する。この時、同じ方角を指していても全く同じ場所に地形オブジェクトを配置するのを防ぐため、特定の範囲内でわずかにランダム性を持たせている。本稿では「大地（地形オブジェクトの無い場所に配置されるベースグラフィック）」「浜」「川」「海」「山」の順で描画を行う。また、山の大小はここでグラフィック処理に反映される。「森」や「平野」など、認識されてはいるが地図上への描画を行っていない地形語に関しては今後に対応する。

4. 評価実験

本章では、方角語・地形語ペアリングの精度、同一地形オブジェクト判別の精度について精度実験を行い、実際に生成した地図についてアンケート調査を行った。

4.1 方角語・地形語ペアリングの精度実験

本節では、方角語・地形語のペアを1つ以上含む文書50件を対象として、方角語・地形語ペアリングの精度について評価実験を行う。本評価実験では、文書中から方角語・地形語を取得してペアリングを行い、実験前に定義した正解ペアリングと合致していた場合、ペアリング成功であるとする。本実験による正解定義ペアの総数は79個、このアルゴリズムで取得したペアは81個であり、その内正解したペアは45個であった。例文を表1に、全50件の文書に対する平均ペアリング精度を表2に示す。

表 2 全 50 件の文書に対する平均ペアリング精度

正解定義ペア数	取得ペア数	取得正解数
79	81	45
再現率	0.570	
適合率	0.556	
F 値	0.563	

4.2 同一地形オブジェクト判別の精度実験

本節では、1つ以上の同じ地形語が含まれている文書20件を対象として、同一地形オブジェクト判別の精度について評価実験を行う。本評価実験では、文書中の同一地形1種類につき、文書全ての解析結果において1種であると認識された場合に成功とする。つまり、同一であるはずの地形語が別と判断された箇所が存在する、または地形語の種類としては同一であっても名前が違うなど別の地形オブジェクトとして認識されるべきオブジェクトが同一と判別された場合は失敗とする。地形語の取得漏れがあった場合も同様に失敗とする。本実験において文書

表 3 同一地形オブジェクト判別精度実験に使用した文書例

文書例	「西の森を抜ければお前さんの探す杖も見つかるじゃろう。だがな、間違っても東の森には行くでないぞ」		
正解オブジェクト数	森 2 種 (東の森, 西の森)	この文書における再現率	1.000
取得オブジェクト数	森 2 種 (東の森, 西の森)	この文書における適合率	1.000

表 4 全 20 件の文書に対する同一地形オブジェクト判別精度

正解定義種類	取得種類数	取得正解種類数
24	23	20
再現率	0.833	
適合率	0.870	
F 値	0.851	

20 件中に定義した正解地形オブジェクト種類は 24 個, この処理により取得した地形オブジェクトは 23 個で, その内正解数は 20 個であった. 例文を表 3 に, 全 20 件の文書に対する同一地形オブジェクト判別精度を表 4 に示す.

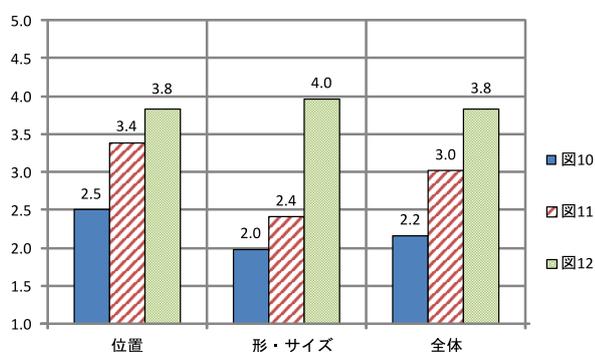
4.3 自動生成した地図についての評価

本節では, 実際に自動生成した地図の評価を行う. 本稿で説明した各処理の有効性を確かめるため, 1. 「方角語・地形語ペアリングのみ」(図 10), 2. 「方角語・地形語ペアリング&同一地形判別」(図 11), 3. 「方角語・地形語ペアリング&同一地形判別&形容詞処理」(図 12) のそれぞれの処理を行った図を用意し, 文書から各被験者が想像した図と自動生成で作られた地図がどれくらい似ているかについて, 『位置』『形・サイズ』『全体図』に関して 56 人に評価してもらった. 評価方法はそれぞれの地図に対して 1~5 点の評価点を整数で付けることで行い, 想像と合っているほど高得点とした. その評価結果(表 5, 図 9)と, 実験に使用した地図を図 10~図 12 に示す. また, ユーザ評価点において, 対応のある t 検定を用いた結果, 有意水準 1% で各処理ごとの全ての図の間に有意差が認められた.

表 5 アンケート評価の平均点

図番号	位置について	形・サイズについて	全体図について
図 10	2.5 点	2.0 点	2.2 点
図 11	3.4 点	2.4 点	3.0 点
図 12	3.8 点	4.0 点	3.8 点

※アンケート選択は 1~5 点の整数とする



※それぞれの項目の2点間において全て $p < 0.01$ で有意差あり

図 9 自動生成した地図についての評点

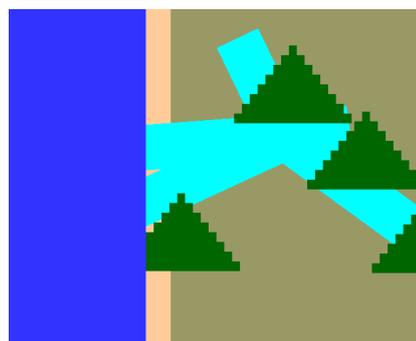


図 10 1. 「方角語・地形語ペアリングのみ」

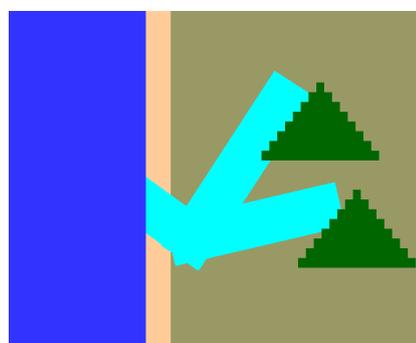


図 11 2. 「方角語・地形語ペアリング」「同一地形判別」

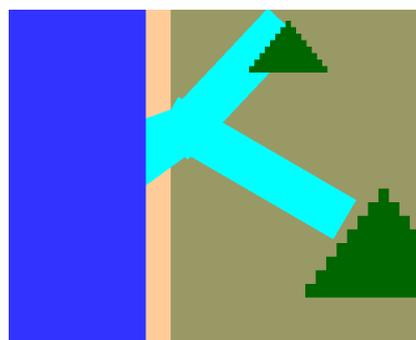


図 12 3. 「方角語・地形語ペアリング」「同一地形判別」「形容詞処理」

5. 考 察

前章の評価実験, 及びアンケート調査で得られた結果を詳しく考察すると以下ようになる.

5.1 方角語・地形語ペアリングに関する考察

本稿の実験においてペアリング精度が低い文書は, 「方角語が地形語から遠く探索範囲外だった」「文書内に多数の地形語と方角語がありペアリングを間違えている」, 特に方角語において「表現が想定と違ったために取得不能だった」のいずれか

に当てはまる。「文書内に多数の地形語と方角語がありペアリングを間違えている」場合の原因は、本稿で提案したアルゴリズムでは単純に単語と単語の距離によって判定しているため、表1のような「方角語+（名前、形容表現）+地形語」が複数回続くパターンの文書では、1回目の地形語と2回目の方角語の距離が近くなってしまい、取り違えが起きたと考えられる。これを防ぐためには、このパターンでは地形語の後に句読点が付く場合が大半であるため、句読点に距離に掛かる重みを付けて、句読点を跨いだ文書は単語間の距離を遠くするなどの処理が考えられる。

「表現が想定と違ったために取得不能だった」場合、対応表現を増やすことが解決策であるが、方角語は表現の幅が広がるのが考えられる。本稿の実験でも「北」「北方」「北側」「北部」などが出現したが、これらは後方の正規表現で全て対応できると考えられる。ただし、正規表現を用いるとノイズも含むことが予想されるため、ノイズの種類を検討し、ノイズが多い場合はノイズを除去する方法も検討する必要がある。また、川の方角指定については「西から東へ」「南北を流れる」などの表現が出現した。川に関しては山などのように一か所の指定だけでなく、始点と終点の描写がある場合には、それらの方角を取得する必要がある。

本稿の実験中に散見された表現として、「(地形語/ここ)から見て(方角語)にある～」という表現がある。今回の実験はペアリング精度の検証であるため描画はしていないが、描画を行う時、現在はウィンドウの中心から見て8方位で位置を決めているが、基準となる場所が地形語であり、かつ、先に場所が指定されている場合には、基準の場所から見て位置を決めなくてはならない。

5.2 同一地形オブジェクト判別に関する考察

本稿の実験で不正解となった要因は「ペアリングのミスのために別の地形オブジェクトとして認識された」「方角語表現にぶれがあり別の地形オブジェクトとして認識された」が挙げられる。このうち「方角語表現にぶれがあり別の地形オブジェクトとして認識された」場合に関しては、正規表現を用いて「西」「西方」「西側」を「西」に収束させるなどの処理を挟み、文書内の表現をシステム内では全て同一の「西」として処理することで表現のぶれを改善出来ると考えられる。また「ペアリングのミスのために別の地形オブジェクトとして認識された」場合に関しては方角語・地形語ペアリングの精度向上が同一地形オブジェクト判別の精度向上にも繋がると考えられる。

本稿の実験において不正解となった特殊なパターンとして、「(地形語)が3つある。それぞれ(名前)、(名前)、(名前)という」という文書があったが、自然における地形であれば、あまりこのパターンの文書は見付けられない。山であれば、多くの場合「(名前)山である」などのように名前の後には地形語が付く。しかし、国や都市などを地形オブジェクトとして扱う場合、「国」や「街」が地形オブジェクトの種類として設定されるが、国や都市は名前の後に「国」「街」が省略されることも多い。その場合の対策として、名前と地形オブジェクト種類の関連付けも重要になると考えられる。

5.3 自動生成された地図への評点に関する考察

本稿では自動生成した地図についてユーザ評価を行った。本ユーザ評価では各項目ごとの全ての2点間において、 $p < 0.01$ で有意差を確認した。全体的に、適用した処理が最も少ない図10は3項目ともに低い評点であるが、処理の数が増えていくごとに評点が上がっている。

まず『位置』についての評点について考察する。図10も方角語・地形語ペアリングは行っており、位置については図11、図12と同様の処理を適用しているが、同一地形オブジェクト判別が行われておらず地形オブジェクトが乱立しているため、図10では文書通りの位置に地形オブジェクトが存在したとしても、ノイズである本来あるべきではない地形オブジェクトの影響が大きく、想像と位置が異なっていると多くのユーザが評価したのではないかと考えている。これにより、同一地形オブジェクト判別はユーザ評価の結果、かなり重要であると考えられる。

『形・サイズ』についての項目も3つの処理を全て行った図12が最も高い結果となっている。図10より図11が0.4pt高い理由として、同一地形オブジェクト判別が適用されているため、余計な川部分の描画が減っているからではないかと推測される。しかし、図12に比べ、図10と図11の評点が低いことから、この項目においては図12のみに適用した形容詞処理が重視されていると考えられ、形容詞処理がユーザの想像に近い地図を生成することに有効であると言える。

また、『全体図』としての評点も図12が最も高い結果となった。これは上記2つの処理が加わったことにより、ユーザの想像と自動生成された地図が近づいたからであると言える。

6. む す び

本稿では、文書中の方角語・地形語ペアリングを中心とした地図の自動生成について提案した。方角語・地形語ペアリング処理精度としては、F値0.563の精度が得られ、同一地形オブジェクト判別精度としてはF値0.851の精度が得られた。また、自動生成した地図のユーザ評価を行い、同一地形オブジェクト判別や大小の形容詞処理が、ユーザの想像と自動生成された地図の図柄を近付けるために有効であることを確認した。

今後はペアリング処理の精度向上のため、正規表現による幅広い表現の方角語の取得や、句読点に関して重みパラメータの導入、距離算出以外の有効なペアリング手法の確立を行う。また、地形補完を行う地形種類の増加とその有効性の評価、「森」など現行では描画できていない地形の描画などのシステムの改良、色や形などの大小以外の形容詞処理の追加等を考えている。

謝辞 室蘭工業大学情報電子工学系学科の各研究室の皆様には、ユーザ評価のアンケート収集に協力して頂きました。感謝の意をここに記します。

文 献

- [1] 永澤 勇樹, 吉田 京平, 服部 峻, “モバイル端末における旅行記の理解支援のための行程抽出と地図化,” 電子情報通信学会 モバイルネットワークとアプリケーション研究会, 信学技報, Vol.114, No.31, MoNA2014-4, pp.19-24 (2014).