Reconstruction of itineraries from annotated text with an informed spanning tree algorithm

Ludovic Moncla, Mauro Gaio, Javier Nogueras-Iso, Sébastien Mustière

International Journal of Geographical Information Science
Vol. 00, No. 00, Month 2015, 1–22

Reconstruction of itineraries from annotated text with an informed spanning tree algorithm

Ludovic Moncla[a,b,c*], Mauro Gaio[a], Javier Nogueras-Iso[b] and Sébastien Mustière[c]

[a]*LIUPPA, Université de Pau, Pau, France*; [b]*Computer Science and Systems Engineering Department, Universidad de Zaragoza, Zaragoza, Spain*; [c]*Université Paris-Est, IGN, Laboratoire COGIT, Saint-Mandé, France*

(*Received 00 Month 200x; final version received 00 Month 200x*)

Considerable amounts of geographical data are still collected not in form of GIS data but just as natural language texts. This paper proposes an approach for the automatic geocoding of itineraries described in natural language. This approach needs as an input a text annotated with part-of-speech and geo-semantic tags. The proposed method is divided into three main steps. Firstly we build a complete graph where vertices represent locations, all vertices are connected to each other by undirected edges. We assign a weight to all the edges of the complete graph using a multi-criteria analysis approach. Then we compute a minimum spanning tree to obtain an undirected acyclic graph connecting all vertices. And finally, we transform this graph into a partially directed acyclic graph in order to identify the sequence of waypoints and build an approximation of a plausible footprint of the itinerary described. Additionally, the rationale of the proposed approach has been verified with a set of experiments on a corpus of hiking descriptions.

Keywords: information extraction, itinerary reconstruction, multi-criteria decision, minimum spanning tree

1. Introduction

If the global understanding of a text is still considered an unattainable task to the current capabilities of computer systems, partial understanding with a predefined view has recently become a feasible task. Generally this task is called text mining (Kao and Poteet 2006). The objective is not to do extensive analysis of the textual contents of documents but tracking through indices for certain informational patterns. The interpretive process is not only led by the text, but it is also guided by a priori knowledge of the sought

*Corresponding author. Email: ludovic.moncla@univ-pau.fr

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/13658816.2015.1108422
http://www.informaworld.com

information. Linguistics analysis can be very accurate but still remains local. This allows both to master the complexity, and facilitates the portability of these systems on different natural languages. A first level of analysis, always triggered by the presence of specific keywords, builds a first interpretative structure. The first structure is integrated in a process to build on it more complex and richer structures, but the principle remains local. This focus and specialization allow building high-performance extraction systems.

Considerable amounts of geographical data are still collected not in the form of GIS data but in natural language texts form. The core of geographical information text mining lies in extracting place names as mentioned in a first notable work that can be attributed to Woodruff and Plaunt (1994). Nowadays, most approaches focus on extracting explicit geographic data from text and associating extracted location references with other information resources (Jones *et al.* 2008). For example in the geographical information area there are several types of narrative structures describing itineraries or displacements such as narrative descriptions of real journeys (travelogues) or travel novels like Gulliver's Travels. In this kind of texts the description of itineraries is just a piece of information in a story with lots of descriptions involving persons, events and places not always related to the itinerary. The description of a displacement is usually scattered in the text along many other things. Another kind of description of displacements are the ones provided in emergency calls, recorded by emergency services. These calls are made by citizens who require in-situ assistance for any kind of incident and try to describe their location using place names and motion expressions (from where they come or where they go) and using landmarks or landscapes features and perception expressions (what they are able to see around them). Hiking guides belong to another category of narrative text describing itineraries. In this kind of descriptions all the information is related to the itinerary. Although itineraries are described using similar elements as for the other narrative categories, they are structured as instructions. The reader must follow these instructions in order to take the same directions as expressed in the described route.

Our proposal is to use information extracted from text using natural language processing and information found in external geographical resources to build a geocoded representation of an itinerary. Vasardani *et al.* (2013) propose an approach for the reconstruction of the environment from a verbal description (translating spatial information into sketch maps). Although our work has some similarities with their proposal, our goal is different. Our approach is focused on the automatic reconstruction of routes and transcribed them in their geographical area of achievement (identifying waypoints and paths by interpreting spatial information in geographical context). The obtained geocoded representation of the itinerary may be used for geoindexing documents like old travelogues (Lesbegueries *et al.* 2006) or legal texts (Yahiaoui *et al.* 2014). It also paves the way to a further analysis of other information contained in the text like geocoding related events or landmarks (Li *et al.* 2014), or to an analysis of itineraries for searching spatial patterns (Laube *et al.* 2005). This kind of work may also have a great interest in digital humanities and in particular in spatial humanities (Gregory *et al.* 2015) such as displaying cultural heritage using textual analysis.

We divide the problem of the automatic reconstruction of itineraries from texts into three sub-problems. The first one is to obtain an annotated text described with a formal markup language where tags could be a combination of part-of-speech and geo-semantic information. We have characterized linguistic structures in terms of lexicosyntactic constraints. Especially we target, spatial named entities, information of motion, perception and spatio-temporal relations (see Figure 1). In the case of an automatic process this sub-problem involves the annotation of spatial information from natural language de-

scriptions. We introduce this process as a prerequisite for the second sub-problem and make a short description of the annotated information. The second sub-problem, which is the main focus of this paper, is to find the sequence of waypoints that provides the order in which the waypoints are visited during the displacement and build a first approximation of the representation of the itinerary. Then the third sub-problem is to propose a better approximation of the representation (not just straight lines between waypoints) taking into account the availability of route networks in urban areas and geographical obstacles in rural areas such as rivers and mountain peaks. This last step will be addressed in future works.

The remainder of this paper is structured as follows. Section 2 describes an overview of related work and its connection with our approach. Section 3 describes our proposed method for the automatic reconstruction of itineraries from information extracted from texts and the formalism used for the input of our proposal. Section 4 describes the experimental evaluation of the proposed method, presenting a specific corpus of hiking descriptions and the obtained results. Finally, Section 5 summarizes our conclusions and future work.

2. Related work

This paper involves connections between natural language processing and various other domains: spatial analysis, spatial cognition, and discourse analysis. Since the early 1990's few significant research, in the field of description of routes, has clarified the relationships between these domains. Appropriate use of "spatial language" thus depends on the addressee's capacity to translate linear linguistic information into multi-dimensional internal representations which incorporate realistic topological relations between the described objects (Denis 1997). But many aspects still need clarifications, in particular dependencies between linguistic expressions, reasoning and visual information (Landau and Jackendoff 1993, Jackendoff 2012).

More generally and always according to Jackendoff, a satisfactory computational theory is an essential factor in developing suitable general algorithmic and implementation theories. Unfortunately the mainstream models of language (such as the proposals arising from Chomsky's approach (Chomsky 1965)) do not lend themselves to being easily adapted to textual information processing. There are at least three reasons for this. One is that these models dot not include the component "meaning" of the natural language. A second reason is that a large part of these models are out of touch with other domains because they focuses almost exclusively on the computational theory. A third problem is that these frameworks do not really distinguish what has to be considered as a lexicon and could be therefore stored (in memory) from what is designed in runtime (when creating a sentence). Several frameworks such as the one proposed by Jackendoff characterized linguistic structures in terms of constraints rather than in terms of algorithms that generate sentences. The view of the lexicon shared in many of these constraint-based approaches is that there is no principled formal distinction between words, rules, and other larger constructs.

In many research studies on "spatial language" such as the ones of Miller and Johnson-Laird (1976), (Talmy 1985, p. 60), Vasardani *et al.* (2013) and many others, the linguistic representation of motion and locations requires two elements and some components: one object located or moving to the reference object ('Figure' and 'Ground', respectively in Talmy's typology). Components can be seen as sets of terms expressing spatial relations

between the Figure and the Ground. In other words the cognitive model of space frequently lies on the one proposed by Lynch (1960). It can be mentioned that Lynch put forward that landmarks (reference object in environmental context), nodes, and paths are the most important components in routes descriptions. In most research works about linguistics (Vandeloise 1986, Landau and Jackendoff 1993, Borillo 1998), the object to be located and the reference object are encoded as noun phrases; and the relationship is encoded as a spatial preposition that, with the reference object, defines an area in which the object to be located is situated. In addition to prepositions, there are many verbs that incorporate spatial relations. Talmy (1985, p. 131) distinguishes two typologies of languages describing how verbal phrases describe *path* and *manner* of motion: *verb-framed language* and *satellite-framed language*. In verb-framed languages such as French and Spanish, the *path* is expressed by verbs and the *manner* by adverb phrases. In satellite-framed languages such as English and German, the *path* is expressed by satellites (in, out, up, down) such as in the phrasal verb *branch off* (Fig.1) and verbs refer usually to the mode of travel such as walking, running, swimming, etc. The differences between languages are the main reason of our proposal to divide the problem of the automatic reconstruction of itineraries form texts into three independent sub-problems. Itineraries and displacements are described in texts using spatial named entities, spatial relations, perception expressions with description of landmarks, motion expressions and trajectories. An itinerary can be defined as a sequence of displacements between places called waypoints. Furthermore, route directions describe not only places and routes, but they also refer to landmarks located along the route (Michon and Denis 2001) that are supposed to be seen during the displacement such as buildings (such as church, school, shop, etc.) and natural landmarks (such as mountain peak, lake, river, etc.). Tom and Denis (2004) also show the important role of landmarks in the description of routes in comparison with the use of street names. A node refers to a place involved in the route where actions and decisions are taken, it is also known as "choice point". In this paper, we distinguish routes and paths according to the definition given by Montello (2005). A path refers to the physical feature (pathway) upon which travel occurs (streets, trails) and a route refers to a displacement occurring on a path or across areas that contains no paths. We propose a model for representing an itinerary as described in a text. Modelling and analysing itineraries lies in the general framework of Time-Geography (Winter and Raubal 2006) and received much attention in the literature. In particular, Spaccapietra *et al.* (2008) proposes a pattern for conceptually modelling itineraries and its implementation to store and query this model in a DBMS.

3. Proposal

In this paper we are mainly focused on the second sub-problem: finding the sequence of waypoints in order to build a geocoded representation of the route of the itinerary. We propose a generic method combining information extracted from texts and information obtained from geographical resources. Section 3.1 describes the formalism used as input of our processing chain.

3.1. *Input: Formalized Descriptions from Texts*

This section describes the different information extracted from the text with a specific formalisation used as input of our proposal. The following sentences are extracted from a

French hiking description and has been translated into English for the sake of clarifying the context of this paper:

> (1) This hike goes from Pralognan to the refuge of Leisse passing by the impressive Grande Casse, all in a wild and dotted with lakes. (2) In Pralognan, follow the road between Hotel de la Vanoise and Hotel du Petit Mont Blanc and go straight. (3) Further pass on Chanton bridge and cross the forest. Soon after, you will reach lake Des Vaches [...] (4) At a small crossroads, you can glimpse Pointe du Creux Noir then branch off south in the direction of lake Long which you will bypass from the right. [...] (5) You can see all the way down Croe-Vie bridge. (6) To reach it go all the way down, then cross it. [...] (7) At the crossroads, do not take south towards the refuge of Entre-Deux-Eaux, but go north and walk one hour. (8) Then follow Leisse torrent to achieve the day's stage. (9) This last part is done in a wild and beautiful steep-sided valley [...]

This typical example of route instructions illustrates the information used in itinerary descriptions to describe displacement between places. It shows that place names can have two roles: waypoint or visual cue. The places 'Pralognan', 'refuge of lake Long' and 'lake Des Vaches' are waypoints. On the opposite, the place 'Pointe du Creux Noir' is not considered as a waypoint because it is associated with the verb of perception 'glimpse', which means that this location is not reached during the displacement, but nevertheless useful because it acts as a visual landmark. And the place 'refuge of Entre-Deux-Eaux' is not considered as a waypoint because it is associated with a negation expression. Intuitively we might imagine that with this kind of association we can directly give the role of visual cue to the involved place name. But if we observe the two sentences (5) and (6) we can deduce that the place name 'Croe-Vie' has a double role. It has a visual cue role when mentioned in the sentence (5) in association with the verb 'to see'. But the sentence (6) changes his role into waypoint. Considering an automatic NLP process, the anaphorical form of its second evocation causes a problem not solved yet. Therefore the waypoint role will be given by default and visual cue retained as a possibility. The example also shows that the order of mention in the text does not always correspond

[1]Part-of-speech tags: CC = coordinating conjunction; PHV = phrasal verb; NN = common noun; IN = preposition or subordinating conjunction; DT = determiner; NP = proper noun

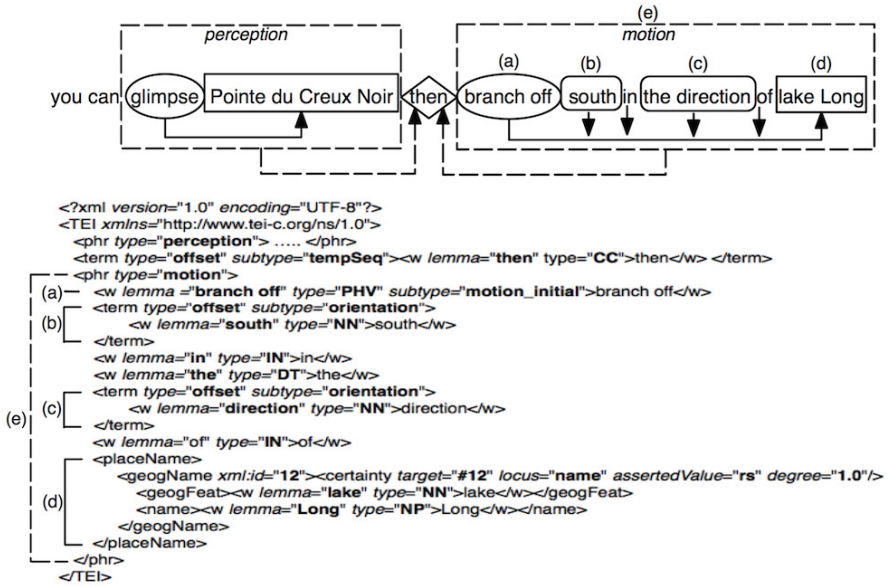


Figure 1. The input format: exemplified with part of sentence (4)[1]

to the order of achievement of the hike as in sentence (1), or in the case of perception or negation such as in sentences (4) and (7). Furthermore, motion or paths connecting waypoints are also used to describe itineraries. For instance, in sentence (4) 'branch off south in the direction of lake Long' means that arriving at a crossroads you have to take the junction going south. Finally, in texts describing travel stories as well as those describing hikes, starting and ending points are almost always given. But here again considering an automatic NLP process, different problems may occur. For instance, in the previous example the ending point 'refuge of Leisse' is given in the sentence (1) and not at the end of the description. How to get this information? Either make a complete semantic analysis of sentence (1) to identify that the three places mentioned represent the beginning of the hike, an intermediate waypoint and the end, or solve the anaphora: 'day's stage' in sentence (8) which refers to the ending point 'refuge of Leisse' mentioned in sentence (1). Thus, to process and identify these information automatically, it is necessary to use discourse analysis techniques, which still nowadays are complex and unreliable.

For the current work, we have used a previously annotated corpus of hiking descriptions following the method proposed by Moncla *et al.* (2014) which provides annotations (Figure 1) using a markup language based on TEI[2] standard. This method combines natural language processing based on a cascade of transducers with the use of gazetteers for the toponym resolution. For the problem of toponym disambiguation this method uses a clustering approach based on spatial density and a semantic matching of geographical feature types of places (Nguyen *et al.* 2013).

3.2. *Graph-Based Model*

From a spatial point of view, itineraries can be described by waypoints, routes connecting waypoints, visual cues and places not reached. Waypoints represent places reached during the displacement, and routes represent motion. We classify waypoints into three categories: starting point, ending point and intermediate points. In the case of loops, starting and ending points are the same. In addition to this, other spatial information is used to define an itinerary in a text, such as places not reached during the displacement called hereafter 'cues'. These places are not considered as waypoints because they are not directly involved in the route. They are used to describe landscape and can contribute to infer locations of unnamed waypoints located along the route between two other waypoints. For simplicity, we consider waypoints and visual cues as punctual objects and routes as linear geometries.

A displacement can be represented as a sequence of waypoints (locations). Each sequence has the form $(w_1, \cdots, w_n)$ where for each $i < j$, the w_i waypoint is reached before w_j. We define an itinerary as a Directed Acyclic Graph (DAG), $G = (V, E)$ comprising a set V of vertices and a set E of edges. The vertices of the graph represent locations mentioned in the text and the edges represent segments between two locations. Each vertex v of G is associated with its real-world location and each two consecutive vertices are connected by an edge. The leaves[3] of G represent the starting point and ending point and also *cues*. When a location is involved several times in the itinerary, for instance in the case of loops, we consider several vertices representing the same location in order

[2]TEI Consortium: `http://www.tei-c.org/Guidelines/P5/`, guidelines for electronic Text Encoding and Interchange

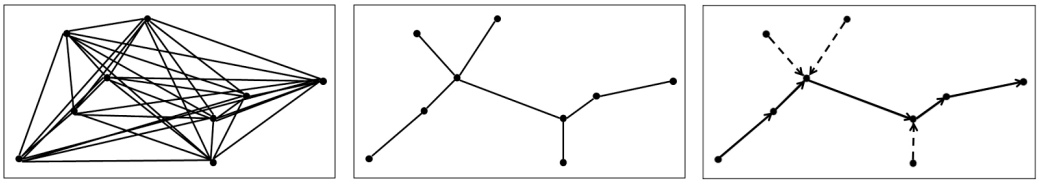
[3]vertices having only one incident edge (terminal vertices).

to avoid cycles. This graph contains 'main edges' (connecting waypoints) representing the displacement and 'secondary directed edges' representing the relations between waypoints and *cues* (places not reached during the displacement, such as places seen or described by the narrator).

From this definition, only vertices representing waypoints and the edges connecting them are needed to build a first approximation of the spatial footprint of the itinerary described. The other vertices (representing visual cues) can be used in another process to disambiguate some locations and infer locations of unnamed places in order to get a better approximation of the route of the itinerary.

We propose to consider a set of information needed for the automatic construction of a DAG that represents the described itinerary. Some of this information can be extracted from the textual description of the displacement: sequence of place names in the text, temporal relations ('after', '2 hours later'), spatial relations ('south of', '2 km', 'in the direction of'), polarity of the displacement ('to leave', 'to arrive'), and the use of a place name with a perception or negation expression ('to see', 'don't go to'). Other information can be obtained from external geographical resources: geographical distance or terrain profile between two places.

The purpose of the reconstruction of the itinerary is to interpret and link spatial information in order to reconstruct the route. Our proposal is to combine the use of all this information, when available, as criteria in order to find the most likely route linking each step of the displacement. Since the target is to provide a generic method that can deal with all types of narrative structure describing itineraries, we make the assumption that we do not know the starting and ending points of the itinerary and the sequence of waypoints either. Therefore, the challenge is to find the itinerary that is closer to the real route intended by the authors who wrote the text. For that purpose we need to identify the sequence of waypoints to build the graph representing the most likely route.



(a) Complete graph (b) Minimum Spanning Tree (c) DAG representing the itinerary

Figure 2. Example illustrated of the process: from a complete graph to a DAG.

We start by building a complete graph $K_n = (V, E)$ where all vertices v are connected, then we propose to use a multi-criteria analysis approach to compute and assign a weight to each edge of this graph (Fig. 2(a)). The weights represent the probability for an edge to be in the final path. Once we have a complete weighted graph, we compute a minimum spanning tree in order to get an undirected acyclic graph connecting all vertices (Fig. 2(b)). Then we transform this tree into a partially directed acyclic graph in order to identify the sequence of waypoints and build the DAG representing the itinerary (Fig. 2(c)).

3.3. *Multi-criteria analysis approach for itinerary reconstruction*

The first step of our approach is to build a weighted complete graph. Our proposal combines local information extracted from the text with physical features obtained from

external sources such as gazetteers or datasets providing digital elevation models. This combined spatial and textual analysis aims at resolving some ambiguities and reconstructing the geocoded representation of the route of the itinerary. The aim is to identify waypoints and find the most probable itinerary linking them with a minimal *length*. The term length is not referring only to geographical distance, but to an aggregated value that takes into account different criteria whose weight is going to be minimized. This length is a combination of contextual information extracted from the description and geometric information like terrain profile or geographical coordinates. Finding this optimal itinerary should help to remove ambiguities or places appearing in the text but not actually visited. This naturally leads to the notion of *minimal weighted spanning tree*. The minimum weighted spanning tree of a set of vertices is the tree connecting all the vertices together with the minimum weight, this weight being the sum of the weights of the edges linking vertices. As we are looking for the minimal spanning tree, all the criteria have to be minimized, that is to say, the lower the values are, the better it is. The criteria used in the proposed approach are described in Section 3.3.1 and the approach to combine these criteria is explained in Section 3.3.2.

3.3.1. Criteria

Sequence of the displacement (C_1)

The first information that can be easily extracted from the text is the sequence in which the places appear. However, the sequence of places in the text is not the same as the sequence of the itinerary (see sentence (1)). Indeed, ordering places as they occur in the text is not effective most of the time. In many cases, the discourse is not linear and the sequence of place names in the text can be totally different from the real sequence of displacements. In such cases, the order of place names should not be taken into account in our decision process or with a lower weight in comparison with the other criteria. Anyway, this can be an important information to help taking decision among several alternatives. For example, in the specific case of hiking descriptions, sequence of place names in the text is often close to the real one. In this case, the order of place names in text is an imprecise but useful information for the decision process. We use this information to define a criterion as the distance between two place names in the textual description, in other words it represents the number of place names appearing between those two place names. Each place name is associated with a number s_i equals to its order of apparition in the text, with $i \in [1, n]$ where n is equal to the total number of place names in the text. The value of the *text distance criterion* for an edge (i, j) is the distance between two place names in the text: $C_1 = |s_i - s_j|$.

Geographical distance (C_2)

Another important criterion is the geographical distance between each location (C_2). We compute the orthodromic distance, which is the shortest distance between two points on a sphere. We use the haversine formula (Sinnott 1984) to calulate this distance shown in equation (1): d is the distance between the two points A and B; r is the radius of the sphere; lat_A, lng_A and lat_B, lng_B are the latitude and longitude of points A and B respectively.

$$d_{AB} = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{lat_B - lat_A}{2}\right) + \cos(lat_A)\cos(lat_B)\sin^2\left(\frac{lng_B - lng_A}{2}\right)}\right) \quad (1)$$

This quantitative criterion based on geographical data and spatial analysis is important to fix errors introduced by the other criteria, and gives information even when other information are not available in the textual description such as spatial relations or expression of motion and perception.

Effort (C_3)

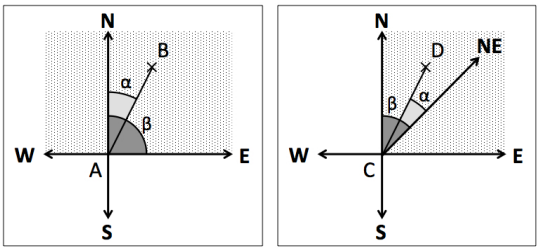
In addition to considering the geographical distance, we propose to consider the effort needed to go from one location to another one taking the slopes of the route into account. To obtain an approximation of the effort of the displacement made by a pedestrian during hikes and treks that are often occurring in mountains, we propose to take into account the elevation profile of the path. Indeed, hikes and treks occur most of the time in mountain areas where paths have ups-and-downs. Furthermore, the elevation gain is commonly used to describe the difficulty and estimate the duration of treks (Naismith's rule). This information is used to determine the steepness of a trail. It is an important factor to assign a difficulty rate. We compute the cumulative elevation gain and the cumulative elevation loss between two locations. For that purpose we compute the sum of elevation gain (pE) and the sum of elevation loss (nE) according to the terrain elevation profile. To determine the value of the effort criterion (ef), we compute the equation (2) widely recognized by experienced ramblers and hikers (equation (2)).

$$ef = (0.01 * pE) + (0.003 * nE) \tag{2}$$

Orientation (C_4)

Projective relations attempt to formalize relations expressed in natural language by orientation and cardinal relations (Clementini 2009) such as: north of, in the direction of, etc. For example, if it is written in the text that after one place we are going north, then we compare this information with geographical coordinates and assign a lower important weight to edges connected to the places that are north than places that are south. We also take into consideration binary relations expressing motion in the direction of a place. In this work we focus on directional and cardinal relations between two place names. We introduce the criterion (C_4) called *orientation criterion* and used to compare projective relations (north, south, in the direction of) extracted from the text and associated to a place with the locations of the other places. We use a projection-based calculus of directions known as *projection-based method* Frank and Mark (1991) or *cardinal algebra* Ligozat (1998) and defining nine basic cardinal relations (n, ne, e, se, s, sw, w, nw, eq). We calculate the angle α between the alternate locations and the azimut representing the orientation relation ($north = 0°$, $east = 90°$, etc). Then, to normalize this angle we divided α by β, where β is equal to 90° when the orientation is expressed by a cardinal direction (north, east, south, west, etc.) or 45° when it is expressed by an ordinal direction (northeast, southwest, etc), or a relative direction (in direction of a specific place). Indeed, we assume that the use of cardinal directions in natural language is fuzzier than the use of ordinal directions or azimuts.

For example, Figure 3(a) can be the representation of the phrase 'Leave A and go to the north ...' and Figure 3(b) the representation of the phrase 'From C walk north-east to ...'. In these examples, B and D are two alternatives that can be reached from A and C, respectively. The questions are: 'how much' B is north of A ? And 'how much' D is north-east of C ? The value of the orientation criterion for the edge (A, B) (Fig. 3(a)) is $C_4 = \alpha/90$ and $C_4 = \alpha/45$ for the the edge (C, D) (Fig. 3(b)). C_4 is normalized between



(a) Cardinal relation (b) Ordinal relation

Figure 3. Illustration of the calculation of the orientation criterion

0 and 1, and the smaller the value is, the more consistent the alternate location is with the orientation relation expressed in the text.

Elevation (C_5)

We are also dealing with another kind of spatial relation called *elevation relation* (C_5), which can be denoted in the text by verbs (to climb, to come down). The *elevation relation* criterion is used to assign a specific weight to the edges connecting places associated with verbs that convey the sense of change of elevation. We use a trilean value for this criterion. If there are no such verbs expressed in the text, the value is equal to 0.5. When elevation relations are expressed in the text, we compare this information with the elevations of all the other places. If the elevation between two places is consistent according to the elevation relation expressed in the text the value is equal to 0 and 1 otherwise.

Temporality (C_6)

We define a criterion called *temporality criterion* (C_6) based on temporal relations automatically extracted from the text (Muller and Tannier 2004) such as temporal prepositions (before, after, then). Elements used to express motion in language are very important for the analysis and the reconstruction of an itinerary. Motion can be denoted by verbs (to go, to leave) and prepositions (from, to). If a temporal relation is expressed between two places, this helps us to determine that these two places are likely to be consecutive. We use this information to set a boolean value: 0 if two places are linked in the text by a temporal relation, and 1 otherwise.

Perception or negation (C_7 and C_8)

We propose to use the information that a place name is associated in the text with a perception or negation expression. The use of perception or negation expression with a place name implies that this place name is not reached during the displacement: it is only seen or used as a landmark to go somewhere else. Perception verbs are frequently used in itinerary descriptions to describe landscapes that we can see far away, such as mountains or lakes. This can be interpreted as a special kind of spatial relations. Information of perception can help to infer locations using the information that during the displacement between two places we are able to see a specific lake or mountain peak. However, in this current work we are not using perception information to infer new locations, but to decide whether a place name is not directly involved in the trajectory because it is not reached during the displacement. The value of the perception (C_7) and negation (C_8) criteria between two places is equal to 1, if at least one of the two places

is associated with a perception or negation expression, and 0 otherwise.

3.3.2. Weighted Sum Model

We have defined the different criteria that characterize an itinerary. All these criteria are defined using information extracted from the textual description of the itinerary or they can be computed using geographical data. We use these criteria to decide over a number of alternatives for the successive displacements in order to reconstruct the most likely route. Some criteria are quantitative, such as geographical distance or text distance, and the other are qualitative.

We propose to use the Weighted Sum Model (WSM), which is a well-known method for multi-criteria decision in decision theory (Triantaphyllou 2000). It combines criteria into a single criterion by multiplying each criterion with a weight and summing up the weighted criteria. The WSM method prioritises criteria by assigning weights and reduces the amount of information by summing the weighted standardized criteria.

$$a_i = \sum_{j=1}^{n} w_j a_{ij} \qquad \forall i \in [1, m] \tag{3}$$

The data input are a set of criteria $C = \{C_0 \cdots C_n\}$, a list of alternatives $A = \{A_0 \cdots A_m\}$, and a set of weights $W = \{w_0 \cdots w_n\}$, where n represents the number of criteria and m the number of alternatives. In our case, alternatives represent the location of place names, and a_i represents the cost to go from one place to another (A_i) and a_{ij} the cost of traversing edge a_i according to criterion j. We use the sum of the weighted criteria (equation (3)) to assign a weight a_i to each edge of the complete graph representing all the possible connections between places. The weights of criteria have been assigned according to the Analytic Hierarchy Process (AHP) indirect method (Saaty 1999) for deriving priorities of criteria. This method is based on the pairwise comparisons of criteria. Firstly, the criteria are compared, two by two, with respect to their importance to reaching the goal of establishing the right sequence of waypoints. This importance is assign using a fundamental AHP scale that ranges from 1 (both criteria have equal importance) to 9 (favoring one criterion over another is of the highest possible order of affirmation). Secondly, the results of these comparisons are entered into a matrix, whose principal right eigenvector will be used to derive the relative strengths of criteria.

Table 1. AHP priorities of criteria, and range of values for measuring each criterion

Description	Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	**AHP Priority**	Range of values
Text distance	C_1	1	3	4	4	4	2	1/3	1/3	**0.14**	$\mathbb{N}_{\geq 0}$
Geographical distance	C_2	1/3	1	2	2	2	1/2	1/5	1/5	**0.06**	$\mathbb{R}_{\geq 0}$
Effort	C_3	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**	$\mathbb{R}_{\geq 0}$
Orientation	C_4	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**	between 0 and 1
Elevation	C_5	1/4	1/2	1	1	1	1/3	1/6	1/6	**0.04**	0, 0.5 or 1
Temporality	C_6	1/2	2	3	3	3	1	1/4	1/4	**0.10**	0 or 1
Perception	C_7	3	5	6	6	6	4	1	1	**0.29**	0 or 1
Negation	C_8	3	5	6	6	6	4	1	1	**0.29**	0 or 1
								Sum of priorities		1	
								Inconsistency		0.022	

Taking into account the context of hiking descriptions, Table 1 shows the matrix with the pairwise comparisons of criteria that were used to derive the priorities. For the sake of paper size limits it is not possible to motivate all comparisons, but we manually choose the weights using the following principles: text distance (C_1) is important in the

specific context of hiking descriptions and route directions; geographical distance (C_2) is a key criterion for the reconstruction of itineraries; effort (C_3) reflects the difficulty of the trail but it is difficult to estimate; orientation expressions (C_4) are important in the description of displacements, but are not always available and become difficult to interpret due to the ambiguity of the language; expressions of change of elevation between two locations (C_5) are important clues, but sometimes human perception may differ from reality; temporality (C_6) is important when available as it informs about the sequence of waypoints reached, but it is hard to extract from natural language and interpret; perception (C_7) and negation (C_8) are very important clues to determine that these locations are not referring to waypoints. Anyway, as discussed later in section 5 the pairwise comparisons should be adjusted to the different types of texts to be processed, or alternative methods for deriving priorities could be considered.

Additionally, we normalize the criteria C_1 to C_4, whose values are beyond the range $[0-1]$ in order to make the criteria comparable with each other, using the formula in equation (4) with $k \in \{1-4\}$ (also known as 'fraction of the sum' normalization (Barba-Romero 2001)). And finally, we sum up the weighted criteria with equation (3) and assign the values to each edge of the complete graph.

$$a_{ik} = \frac{a_{ik}}{\max(a_{ik})}, \qquad \forall i \in [1, m] \tag{4}$$

3.4. *Minimum Spanning Tree (MST)*

We are working with a connected, weighted, complete undirected graph, where all the weights are positive numbers. We use Prim's algorithm (Prim 1957) to find a minimum spanning tree for a connected, weighted, undirected graph (see Algorithm 1). This algorithm builds the tree (T) one vertex at a time. It starts by adding randomly a vertex (v_0) to the set of nodes (Q) and removes it from the input weighted graph (G). Then, at each step it adds the edge with the minimum weight connecting a vertex v already inserted in Q with a new vertex w, not included in Q yet. To solve the problem of duplicate nodes, introduced to avoid cycles when the same location appear several times, we assign an infinite weight to the edges connecting two duplicate nodes.

The advantage of this approach is that we do not need a directed graph and we do not need to know which are the starting and ending points. Algorithm 1 shows a simple version of Prim's algorithm to facilitate the understanding of its applicability in this context.[4]

The result is a connected acyclic undirected graph, also called 'path graph', which means that this tree has no root. Furthermore, a unique simple path connects any two vertices in this tree. This tree represents the described itinerary, with vertices representing waypoints of the displacement and also vertices representing locations involved in the description of the itinerary but not reached during the displacement, such as visual cues (mountain peaks, lakes, etc.).

[4]if efficient structures such as heaps are used for the storage of weighted edges, the overall time complexity of this greedy algorithm is linearithmic $O(m \log n)$, where n is the number of vertices and m the number of edges.

Algorithm 1: Minimum Spanning Tree

Input: undirected connected weighted graph $G = (V, E)$
being V a list of vertices $V = \{v_0 \cdots v_n\}$
Output: tree T representing the set of edges composing an MST of G

```
1  Q ← empty list;
2  Insert(Q,v0) ;                                   // with v0 chosen randomly
3  Remove(V,v0);
4  while V != empty do
5  |   minWeight ← ∞;
6  |   foreach vertex v ∈ Q do
7  |   |   foreach vertex w ∈ V do
8  |   |   |   if weight(ev,w)< minWeight then
9  |   |   |   |   bestEdge ← ev,w ;                // weight according to equation 3
10 |   |   |   |   w' ← w;
11 |   |   |   end if
12 |   |   end foreach
13 |   end foreach
14 |   Insert(T,bestEdge);
15 |   Insert(Q,w');
16 |   Remove(V,w');
17 end while
18 return(T);
```

3.5. *Building a DAG from the minimum spanning tree*

The last step of the process is to build a DAG representation of the described itinerary. We propose to find the longest path on the minimum spanning tree (maximum number of vertices between two leaves) in order to identify which leaves are the starting and ending points, and remove vertices that are not part of the displacement. This problem is the equivalent of finding the largest sub-graph having a Hamiltonian path or finding the topological order of a DAG.

We transform the tree into a Partially Directed Acyclic Graph (PDAG) also called Chain Graph. The class of chain graphs was introduced by Lauritzen and Wermuth (1989) as a generalization of both undirected graphs and acyclic directed graphs and admits both directed and undirected edges. Let $G = (V, E)$ be a chain graph with a finite vertex set V and an edge set $E \subseteq V \times V$. An edge $(v, w) \in E$ is directed if $(w, v) \notin E$ and undirected if $(w, v) \in E$. We denote a directed edge (v, w) by $v \rightarrow w$ and an undirected edge (v, w) by $v - w$. If $(v, w) \in E$, then v and w are adjacent.

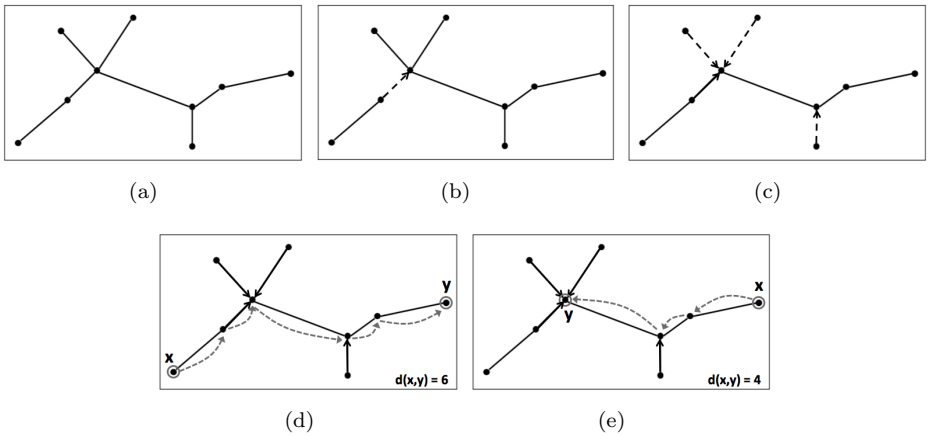


Figure 4. From an undirected acyclic graph to a partially directed acyclic graph (a-c) and illustration of the method to find the longest path (d-e)

To transform undirected edges into directed edges we use spatio-temporal relations that are expressing motion such as "goes to", "reach", etc. Indeed, motion can be denoted by verbs and prepositions and play an important role in the description of itineraries. It gives information concerning the polarity of the displacement (Slobin 1996, Aurnague 2011). When such relations are available, we use it to transform undirected edges into directed edges (Fig. 4(b)). We also transform undirected edges connecting locations and visual cues into directed edges (Fig. 4(c)). Considering a location represented by a vertex v and a visual cue represented by a vertex w, we transform the undirected edge $v - w$ into a directed edge $w \rightarrow v$. Then to find the longest path in the chain graph and assign a direction to all edges, we use a Depth First Search (DFS) algorithm (Tarjan 1972, Karger *et al.* 1997). We apply a DFS starting from every leaf (vertices having only one connected edge) except those which represent locations associated with perception or negation expressions. We compare the resulted distance of each DFS to identify the longest one. For example, Figure 4(d) and 4(e) shows the two possible paths, and in this example Figure 4(d) shows the longest path of the chain graph. The leaves x and y (Figure 4(d)) are considered as being the starting and ending points respectively. All other leaves are not considered as waypoints in the DAG representation of the displacement. When there are no spatial relations available to transform undirected edges, we are still able to find the longest path but we cannot distinguish starting and ending points.

4. Experiments and evaluation

4.1. *Description of the corpus*

The annotated corpus contains 90 documents divided into three sets of 30 documents extracted from specialized websites in French[1], Spanish[2], and Italian[3]. Each document describes one trail and is associated with the real trajectory (GPS) of the route. Each trail is only described by one document. Real trajectories are only used for the evaluation of the results of our automatic process. Although the main focus of this paper is not directly linked with Natural Language Processing for text mining, 5 out of 7 criteria are based on the ability of detecting language expressions about orientation, change of elevation, temporality, perception and negation. The precision to identify correctly these expressions varies from one language to another. The results of our proposal for the automatic reconstruction of itineraries depend on the results of the annotation process. In order to evaluate the proposed method of itinerary reconstruction without introducing errors as input, we manually corrected the results of the toponym resolution, correcting wrong locations and we assume that inputs of our proposed method are 100% correct.

4.2. *Implementation of the proposed method*

This section describes the customization of the method proposed in Section 3 for the adequate running of the experiments and taking into account the corpus of itineraries that has been selected. Firstly, it must be noted that the text mining method proposed by Moncla *et al.* (2014) for the generation of the geographically annotated corpus extracts different kinds of information such as motion expressions, spatial relations, perception

[1] `http://www.visorando.com` (fr)
[2] `http://senderos.turismodearagon.com` (es)
[3] `http://www.parks.it/parco.alpi.marittime/` (it)

expression or negation and not only spatial named entities. As mentioned in section 3.3.1, we use this information as criteria to weight the edges of the graph connecting each two places in order to take decisions and find the best route between places. Additionally, apart from the criteria extracted from texts, there are also some criteria that are described using information coming from digital elevation datasets.

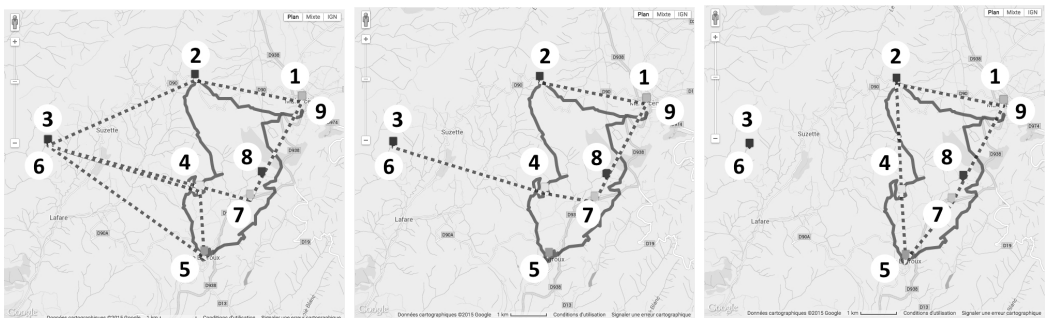
We propose to detail a typical case of hiking description that shows the strength of the proposed approach. The following phrases summarize the textual description of the hike:

> (10) From ***Malaucène*** to ***Col de la Chaîne*** northwest [...] (11) where you can admire a beautiful view of the ***Dentelles de Montmirail*** [...] (12) go south in the direction of ***Sainte-Madeleine Abbey*** [...] (13) we head straight to the ***castle of Barroux*** [...] (14) the view extends on the ***Dentelles de Montmirail*** [...] (15) passing near the old ***chapel Saint Jean*** and the ***Abbey of N.-D. de l'Annonciation*** [...] (16) return to ***Malaucne***.

Table 2. Geographical coordinates and elevation of place names

	Place name	Latitude	Longitude	Elevation
1	Malaucène	44.1741	5.1322	331
2	col de la Chaîne	44.1793	5.0966	466
3	les Dentelles de Montmirail	44.1638	5.0478	347
4	Abbaye Sainte-Madeleine	44.1529	5.0983	364
5	le château du Barroux	44.1373	5.0996	300
6	les Dentelles de Montmirail	44.1638	5.0478	347
7	chapelle Saint-Jean	44.1508	5.1150	334
8	Abbaye N.-D. de l'Annonciation	44.1562	5.1187	364
9	Malaucène	44.1741	5.1322	331

This hiking trail is a loop, where the place name *Malaucène* is both the starting and the ending point. Table 2 shows the list of place names extracted from this hiking description. Place names are ordered as they appear in the text. They are associated with geographical coordinates (latitude, longitude) and elevation. The place names *Malaucène* and *Dentelles de Montmirail* appear twice in the description, and the place *Dentelles de Montmirail* is associated with expressions of perception (phrases (11) and (14))



(a) Text distance (C_1) (b) Geographical distance (C_2) (c) Multi-criteria

Figure 5. Results of automatic itinerary reconstruction using different criteria

Figure 5 shows the result of the itinerary reconstruction of this trail, using the *text distance* and the *geographical distance* criteria independently and using the proposed multi-criteria approach (Fig. 5(c)). The solid grey line represents the real GPS trajectory of the displacement, and the dashed lines represent the approximation of the route

computed automatically. In this example, we can notice that none of these two criteria taken independently can solve the problem of itinerary reconstruction, neither *text distance* (Fig. 5(a)) nor the geographical distance between places (Fig. 5(b)). The multi-criteria approach (Fig. 5(c)) taking into account all the criteria give better results than criteria taken independently.

To illustrate the multi-criteria approach used to weight the complete graph, Table 3 shows the value of the weights for each criterion and for all edges connected to the vertex 2 (*col de la Chaîne*). The multi-criteria column shows the weights assigned to each edge using the formula described in equation (3) of section 3.3.2.

Table 3. Weight values for all edges connected to the vertice 2

Edge	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Multi-criteria
2-1	**0.14**	**0.42**	**0.37**	**0.81**	**0.5**	**0**	**0**	**0**	**0.11**
2-3	0.14	0.62	0.48	0.80	0.5	1	1	0	0.52
2-4	**0.29**	**0.43**	**0.21**	**0.04**	**0.5**	**1**	**0**	**0**	**0.19**
2-5	0.43	0.68	0.24	0.05	0.5	1	0	0	0.22
2-6	0.57	0.62	0.48	0.80	0.5	1	1	0	0.58
2-7	0.71	0.51	0.26	0.36	0.5	1	0	0	0.27
2-8	0.86	0.45	0.17	0.48	0.5	1	0	0	0.29
2-9	1.00	0.42	0.22	0.91	0.5	1	0	0	0.33

4.3. *Evaluation of our approach*

For the evaluation of the proposed approach we only consider the spatial component of an itinerary and we consider displacements as geometric lines. We propose to use two different methods to evaluate the proposed approach. The first one (e_1) makes the comparison of the edges of the DAG obtained automatically with the edges manually built (Section 4.3.1). The second approach (e_2), fully automatic, makes the comparison of the real trajectory (GPS), associated with each description of the corpus, with the DAG built automatically (Section 4.3.2).

4.3.1. Comparison with manually produced trees (e_1)

We propose to use precision and recall measures to evaluate our approach and compare it to gold standard itineraries generated manually. Edges were manually built according to the textual description and taking into account the comparison with the real trajectory of the path from GPS data. The precision represents the length of the relevant edges obtained automatically over the length of all the edges automatically built. And the recall represents the length of the relevant edges obtained automatically over the length of the edges manually built. Table 4 shows the global precision, recall and F1-measure of 7 experiments on the corpus, where each experiment tests a combination of criteria.

Table 4. Evaluation of the precision and recall of edges obtained of the corpus of experiment

	Experiments (combination of criteria)	Precision	Recall	F1-Measure
(1)	C_1	89.5%	73.8%	80.9%
(2)	C_2	71.2%	51.2%	59.6%
(3)	$C_1 + C_2$	88.3%	80.5%	84.2%
(4)	$C_1 + C_2 + C_3$	89.1%	82.9%	85.9%
(5)	$C_1 + C_2 + C_3 + C_4 + C_5$	90.2%	82.8%	86.3%
(6)	$C_1 + C_2 + C_3 + C_4 + C_5 + C_6$	93.0%	86.0%	89.4%
(7)	$C_1 + C_2 + C_3 + C_4 + C_5 + C_6 + C_7 + C_8$	96.3%	95.8%	96.1%

Table 4 highlights the fact that each new criterion improves the accuracy of the automatic reconstruction. We can notice that, even if the qualitative information is not

always expressed in the text, they improve significantly the accuracy. Indeed, the overall accuracy of the method (line 7) is equal to 96.1% against 84.2% for the combination of the two quantitative criteria *text distance* and *geographical distance* (line 3). Line 4 shows that taking into account the effort (C_3) improves the accuracy of the automatic reconstruction. Line 5 shows the contribution of the spatial and elevation criteria. Although the score 86.3% is not significantly higher with respect to previous 85.9%, comparing the information expressed in the text such as spatial relations (north of, in the direction of, etc) or expressions referring to a change of elevation (to climb, to go down) with the geographical information found in gazetteers, improves the accuracy of the automatic reconstruction. Line 7 of Table 4 shows that the perception and negation expressions ($C_7 + C_8$) are very useful to identify places that are not waypoints.

4.3.2. Comparison with real GPS trajectories (e_2)

The evaluation corpus provides a ground truth of the route of the described displacement (GPS) associated with each hiking description. To evaluate the proposed method of automatic reconstruction of itinerary, we propose to compare the automatically computed route with the real trajectory (GPS) available with each document of the corpus, in order to evaluate the overall adequacy of the proposed reconstruction. The proposed method for the automatic reconstruction of itinerary builds an approximation of the route using straight lines and without taking into account road networks or geographical obstacles (rivers, mountains,...). The shape of the resulting route is obviously different from the real one. For instance, Llobera and Sluckin (2007) show the emergence of switchback patterns ("zigzags") in the case of steep slopes and propose a semi-quantitative theoretical model of the behaviour of humans moving on a terrain with relief. Therefore, the e_2 method aims at measuring how well straight routes are an approximation to the actual routes. This method takes into account an error margin to compare the similarity between the generated straight lines and the real trajectory. We create a buffer around the proposed route and we calculate the ratio of the real trajectory that is included in this buffer. If two waypoints are near, it is unlikely that intermediate points are missing. On the opposite, if two waypoints are far from each other, there might be some missing waypoints or some missing information needed for the reconstruction of the itinerary. The radius of the buffer is thus proportional to the length of each segment of the proposed route. Experimentally, we set the value of the radius buffer to 15% of the distance between two waypoints. The nearer two places are, the thinner the buffer is; and the farther two places are, the larger the buffer is.

We use this buffer to calculate the ratio of the length of the real trajectory that is included in the buffer of the proposed route. The average ratio of real trajectories included in buffers for all the documents of the corpus is equal to 71.4%. Figure 6 shows the accuracy of the two methods of evaluation (e_1 and $e2$), and Figure 7 shows some visual examples of results and makes the comparison between the automatically reconstructed route and the real trajectory. Each sub-figure of Figure 7 is associated with the score of the two methods of evaluation. As expected, we can notice that the scores obtained with the second evaluation approach are not as good as the scores obtained with the first evaluation approach (Fig. 6). For instance, considering the combinaison of all criteria (experiment 7), evaluation e_2 obtains 71.4% against 96.1% with the evaluation e_1. Indeed, as we are proposing an approximation of the route, even if the reconstruction is correct in comparison with the manually built sequence of waypoints, it may be not correct in comparison with the real route of the displacement. For instance, the reconstructions of itineraries shown in figures 7(a) and 7(f), are correct with the evaluation e_1 (100%) but

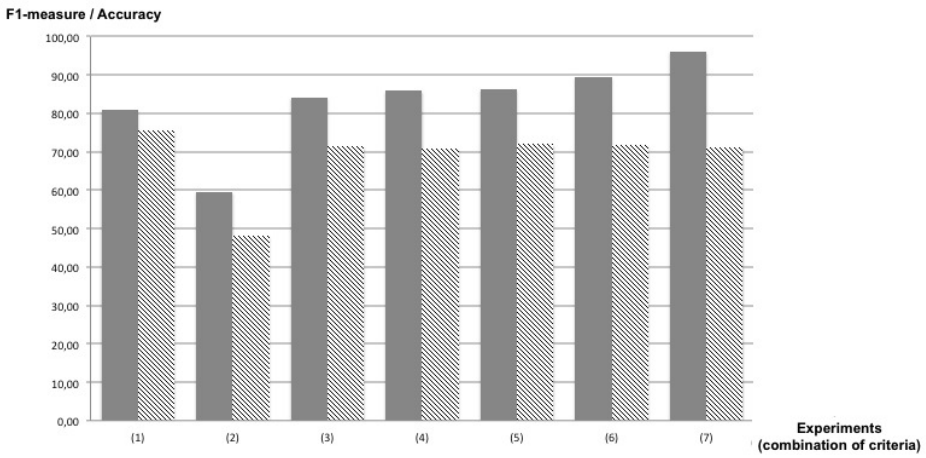
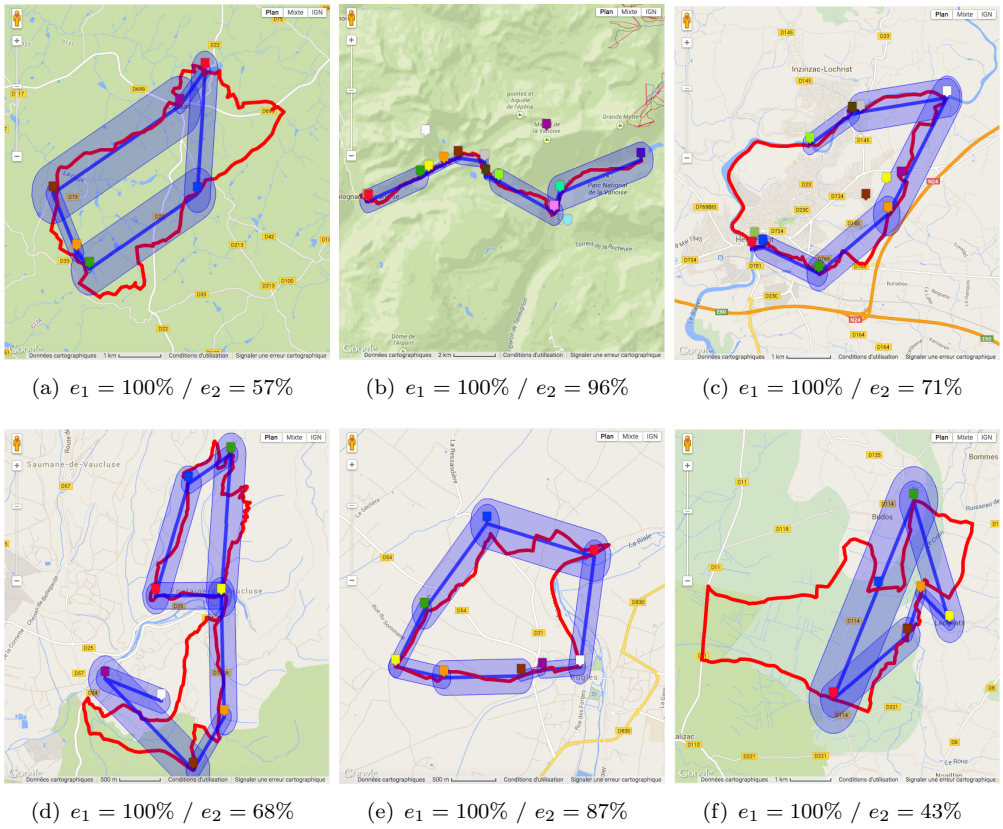


Figure 6. Comparison of measures obtained by the 2 evaluation methods on the 7 experiments described in Table 4: grey bars show the F1-measure (e_1); hatched bars show accuracy (e_2)

have lower scores with the evaluation e_2 (57% and 43%). These scores are explained by the fact that some locations are not named in the descriptions and that other information is expressed in terms of road names or relative directions ("turn left on road RN 12"). Another typical case of issue of reconstruction is shown in figures 7(c) and 7(d), the problem in these hikes, is that at the end of the descriptions the names of the ending locations are not mentioned, and are supposed to be the same as the starting points. Furthermore, at the end of the description of hike 7(c), it is written that the route follows the river bank to come back at the starting point. The lower score obtained with the method e_2 in figure 7(e) is explained by the fact that between two waypoints the route is following a river. In further work, we plan to consider natural obstacles (such as river, chasm, mountainous topography) and route network in order to create *intermediate virtual waypoints* and obtain a spatial representation with high precision.

5. Conclusions

This paper has proposed a formal model for representing an itinerary as described in a text. A Directed Acyclic Graph (DAG) is used foreseen, where the vertices of the graph represent locations and the edges represent segment between two locations. These elements are the core of most models representing routes or itineraries with a graph approach (Werner *et al.* 2000, Spaccapietra *et al.* 2008, Vasardani *et al.* 2013). The model is original in that in addition to taking into account the classic elements (routes and waypoints), it emphasizes other elements describing an itinerary: features seen or mentioned as landmarks. To go further, this model could be enriched with other elements for a more precise description. In particular, texts may describe the itinerary at several levels of granularity (some times a global description of the whole itinerary, and some times a precise description of particular pieces of the itinerary) and modelling multiple granularities would then be useful, as proposed by Hornsby and Egenhofer (2002). Other key elements could be to rely on linear referencing principles to model events appearing within a particular route (Güting *et al.* 2006), or allowing the modelling of fuzzy information like in sentences (3) and (8). Additional notions like the one of "entry, course and exit" emphasized by the Route Graph model would also be useful to represent orientation elements on how the itinerary enters or exits waypoints, and how it is related to other



(a) $e_1 = 100\%$ / $e_2 = 57\%$ (b) $e_1 = 100\%$ / $e_2 = 96\%$ (c) $e_1 = 100\%$ / $e_2 = 71\%$

(d) $e_1 = 100\%$ / $e_2 = 68\%$ (e) $e_1 = 100\%$ / $e_2 = 87\%$ (f) $e_1 = 100\%$ / $e_2 = 43\%$

Figure 7. Comparison of the automatic reconstruction (blue) with the real trajectory (red).

elements along routes (Werner *et al.* 2000, Krieg-Brückner and Shi 2006).

Moreover, a comprehensive approach is proposed for automatically identifying the sequence of waypoints from a geoparsed text and building an approximation of a plausible sequence of the described itinerary. The feasibility of this approach has been tested for the automatic approximate geocoding of itineraries described in a corpus of hiking descriptions. This feasibility study also allowed us to illustrate that combining quantitative and qualitative criteria, based on knowledge extracted from the text and knowledge extracted from geographic databases, improves the approximation of a described itinerary.

As ongoing work, we study how to improve the geoparsing by adding a deeper linguistic processing and a deeper spatial analysis to take into consideration new categories of spatial relations and to annotate unnamed locations. In conjunction, we are about to consider a method to integrate more information coming from geographic databases describing feature shapes. For example, if the itinerary description mentions "follow the road/river" (sentences (2) and (8)), "cross the forest" (sentence (3)), or "walk one hour" (sentence (7)), that information could be used to define some new criteria (some other works also propose to use this information (Richter and Klippel 2005)), if they are crossed with databases describing forests and rivers or digital elevation models, and if some spatial analysis tools are defined to approximate the notion of "cross", "follow" or to approximate distance from walking durations. Some other information in the text describe relations between parts of the itinerary, like "go straight" (sentences (2)). In order to handle that, we cannot directly define new criteria to weight each edge of the

graph, but we should extend the notion of minimal spanning tree and constrain the search so that those relations between parts are fulfilled.

Another forthcoming work concerns the setting of the multi-criteria approach. One key issue in our multi-criteria approach is how to define weights and the combination strategy. For such a problem of setting the weights of a multi-criteria combination, or for setting the suited model of combinations, machine learning is a widely used approach that we could follow. In particular, machine learning is used in the natural language processing domain for approaches to entity tagging that are based on probabilistic models such Hidden Markov Models (Rabiner 1989) or, approaches to extracting semantic relationships between entities (Béchet *et al.* 2014). However, whatever machine learning technique is used, a key issue is to get a sufficient number of examples and to precisely define the learning task (Mitchell 1997). Those examples cannot be direct examples for our task, as we have seen that the text alone is never sufficient to reconstruct the actual precise itinerary, and as the measure that we propose for comparing GPS tracks and reconstructed itinerary is only an approximate one (cf. Section 4.3.2). However, we may expect that a huge number of examples could overcome some of those difficulties, if one tries to learn weights that minimise the proposed evaluation distance. This could be faced as further work.

Our approach aims to reconstruct the sequence of displacement taking account the geographical area of achievement. An ongoing work is to approximate the actual footprint of the displacement. To do that, we may extrapolate from a very small amount of information present in the text. Some external knowledge has to then be introduced, like displacement habits: for example, hikers do not cross rivers and may minimise they effort. Other information could come from other itinerary descriptions, in any format (text or geolocalised paths). This would require to introduce some mechanism for merging itinerary descriptions, as proposed by (Belouaer *et al.* 2013).

Finally, we would also like to extend our experiments with a larger corpus of texts describing itineraries, including not only hiking descriptions but also other types like travelogues or spatial orientation instructions.

Acknowledgements

This work has been partially supported by : the Communauté d'Agglomération Pau Pyrénées (CDAPP) and the Institut National de l'Information Géographique et Forestière (IGN) through the PERDIDO project ; the Spanish Government (project TIN2012-37826-C02-01); and the Aragon and Aquitaine Regional Governments through the transborder Aragon-Aquitaine cooperation programme 2014.

References

Aurnague, M., 2011. How motion verbs are spatial: The spatial foundations of intransitive motion verbs in French. *Lingvisticae Investigationes*, 34 (1), 1–34.

Barba-Romero, S., 2001. The Spanish Government Uses a Discrete Multicriteria DSS to Determine Data-Processing Acquisitions. *Interfaces*, 31 (4), 123–131.

Béchet, N., *et al.*, 2014. How to combine text-mining methods to validate induced Verb-Object relations?. *Comput. Sci. Inf. Syst.*, 11 (1), 133–155.

Belouaer, L., Brosset, D., and Claramunt, C., 2013. Modeling Spatial Knowledge from

Verbal Descriptions. *In*: *Proc. 11th Intl. conf on Spatial Information Theory, COSIT*, 338–357.

Borillo, A., 1998. *L'espace et son expression en français, L'essentiel*. Orphrys.

Chomsky, N., 1965. *Aspects of the theory of syntax*. Oxford, England: M.I.T. Press.

Clementini, E., 2009. A Conceptual Framework for Modelling Spatial Relations. Thesis (PhD). Institut National des Sciences Appliques de Lyon, Lyon.

Denis, M., 1997. The description of routes: a cognitive approach to the production of spatial discourse. *Cahiers de Psychologie Cognitive*, 16, 409–458.

Frank, A.U. and Mark, D.M., 1991. *Language Issues for Geographical Information Systems*. Maguire DJ et al.

Gregory, I., *et al.*, 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *IJHAC*, 9 (1), 1–14.

Güting, R.H., de Almeida, T., and Ding, Z., 2006. Modeling and Querying Moving Objects in Networks. *The VLDB Journal*, 15 (2), 165–190.

Hornsby, K. and Egenhofer, M.J., 2002. Modeling Moving Objects over Multiple Granularities. *Annals of Mathematics and Artificial Intelligence*, 36 (1-2), 177–194.

Jackendoff, R., 2012. Language as a source of evidence for theories of spatial representation. *Perception*, 41 (9), 1128–1152.

Jones, C.B., *et al.*, 2008. Modelling vague places with knowledge from the Web. *IJGIS*, 22 (10), 1045–1065.

Kao, A. and Poteet, S.R., 2006. *Natural Language Processing and Text Mining*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Karger, D., Motwani, R., and Ramkumar, G.D.S., 1997. On approximating the longest path in a graph. *Algorithmica*, 18 (1), 82–98.

Krieg-Brückner, B. and Shi, H., 2006. Orientation Calculi and Route Graphs: Towards Semantic Representations for Route Descriptions. *In*: *Proc. 4th Intl. conf on GIS*, GIScience'06, Berlin, Heidelberg, 234–250.

Landau, B. and Jackendoff, R., 1993. What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16 (02), 217–238.

Laube, P., Imfeld, S., and Weibel, R., 2005. Discovering relative motion patterns in groups of moving point objects. *IJGIS*, 19 (6), 639–668.

Lauritzen, S.L. and Wermuth, N., 1989. Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. *The Annals of Statistics*, 17 (1), 31–57.

Lesbegueries, J., Gaio, M., and Loustau, P., 2006. Geographical information access for non-structured data. *In*: *Proc. ACM Symposium on Applied Computing*, 83–89.

Li, R., Fuest, S., and Schwering, A., 2014. The effects of different verbal route instructions on spatial orientation. *In*: *the 17th AGILE conference on geographic information science, Castellon, Spain*.

Ligozat, G., 1998. Reasoning about Cardinal Directions. *J. Vis. Lang. Comput.*, 9 (1), 23–44.

Llobera, M. and Sluckin, T.J., 2007. Zigzagging: theoretical insights on climbing strategies. *Journal of Theoretical Biology*, 249 (2), 206–217.

Lynch, K., 1960. *The image of the city*. Vol. 11. MIT press.

Michon, P.E. and Denis, M., 2001. When and Why Are Visual Landmarks Used in Giving Directions?. 2205, *In*: D.R. Montello, ed. *Spatial Information Theory.*, 292–305.

Miller, G.A. and Johnson-Laird, P.N., 1976. *Language and perception*. Vol. viii. Cambridge, MA, England: Belknap Press.

Mitchell, T.M., 1997. *Machine learning*. McGraw-Hill, Inc.

Moncla, L., *et al.*, 2014. Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus. *In*: *Proc. 22nd Intl. Conf on Advances in Geographic Information Systems*, SIGSPATIAL '14 ACM, 183–192.

Montello, D.R., 2005. Navigation. *In*: *The Cambridge handbook of visuospatial thinking.*, 257–294 Cambridge: Cambridge University Press.

Muller, P. and Tannier, X., 2004. Annotating and measuring temporal relations in texts. *In*: *Proc. 20th intl. conf on Computational Linguistics* ACL, p. 50.

Nguyen, V.T., Gaio, M., and Moncla, L., 2013. Topographic Subtyping of Place Named Entities: a linguistic approach. *In*: *Proc. 15th AGILE Intl. conf on GIS*.

Prim, R.C., 1957. Shortest connection networks and some generalizations. *Bell system technical journal*, 36 (6), 1389–1401.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2), 257–286.

Richter, K.F. and Klippel, A., 2005. A Model for Context-specific Route Directions. *In*: *Proceedings of the 4th International Conference on Spatial Cognition: Reasoning, Action, Interaction*, SC'04 Berlin, Heidelberg: Springer-Verlag, 58–78.

Saaty, T.L., 1999. *Decision making for leaders: the analytic hierarchy process for decisions in a complex world*. Vol. 2. RWS publications.

Sinnott, R., 1984. Virtues of the haversine. *Sky and Telescope*, 68 (2), 159.

Slobin, D.I., 1996. Two ways to travel: Verbs of motion in English and Spanish. *Grammatical constructions: Their form and meaning*, 195–219.

Spaccapietra, S., *et al.*, 2008. A conceptual view on trajectories. *Data & Knowledge Engineering*, 65 (1), 126–146.

Talmy, L., 1985. Vol. 3, Lexicalization patterns: Semantic structure in lexical forms. *In*: *Language typology and syntactic description.*, 57–149 Cambridge University Press.

Tarjan, R., 1972. Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing*, 1 (2), 146–160.

Tom, A. and Denis, M., 2004. Language and spatial cognition: comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology*, 18 (9), 1213–1230.

Triantaphyllou, E., 2000. *Multi-criteria Decision Making Methods: A Comparative Study*. Applied Optimization Vol. 44. Boston, MA: Springer US.

Vandeloise, C., 1986. *L'Espace en français. Sémantique des prépositions spatiales*. Editions du Seuil, Paris.

Vasardani, M., *et al.*, 2013. From Descriptions to Depictions: A Conceptual Framework. *In*: *Proc. 11th Intl. conf on Spatial Information Theory, COSIT*, 299–319.

Werner, S., Krieg-Brückner, B., and Herrmann, T., 2000. Modelling Navigational Knowledge by Route Graphs. *In*: *Spatial Cognition II, Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications*, 295–316.

Winter, S. and Raubal, M., 2006. Time Geography for Ad-Hoc Shared-Ride Trip Planning. *In*: *Proc. 7th Intl. conf on Mobile Data Management, MDM*, May., p. 6.

Woodruff, A.G. and Plaunt, C., 1994. GIPSY: Automated Geographic Indexing of Text Documents. *J. Am. Soc. Inf. Sci.*, 45 (9), 645–655.

Yahiaoui, S., *et al.*, 2014. Vérification et (re)construction automatiques des limites des bureaux de vote par l'étude des textes juridiques. *In*: *SAGEO*, Grenoble, France.