KEYSTONE COST ACTION IC1302, SHORT TERM SCIENTIFIC MISSIONS
-
SCIENTIFIC REPORT

STSM Topic: Extraction and representation of place names for the reconstruction of itineraries from texts
STSM Applicant: Ludovic Moncla
Applicant's affiliation: LIUPPA laboratory, Université de Pau et des Pays de l'Adour
Applicant's address: LIUPPA - UFR S&T de PAU (UPPA),
BP 1155, 64013 Pau Université Cedex, Pau, France
e-mail: ludovic.moncla@univ-pau.fr

1. Purpose of the STSM

In November 2012 three research groups on both sides of the French-Spanish border on the Pyrenees started collaboration for the development of a research project called PERDIDO (Project for Extracting and Retrieving DIsplacements from textual DOcuments). This project deals with the extraction and representation of place names for the reconstruction of itineraries from textual sources. This process can be divided in three phases: recognition of toponyms in the text, toponym resolution (also known as geocoding), and reconstruction of routes from detected toponyms. Our goal is to provide specialized mechanisms for toponym resolution.
The objective of this research stay has been to analyse the problem of toponym resolution in the case of processing texts that describe trip itineraries (e.g. hiking descriptions) in small areas where fine-grain toponyms are not usually found in well-known toponym databases such as Geonames or Open Street Maps.

2. Description of the work carried out during the STSM

Toponym resolution involves two important and related issues. One issue is toponym disambiguation, i.e. the task of assigning to a place name its correct reference in the world. Another issue is to assign an explicit georeference on the earth to a disambiguated toponym. In general, these two activities/issues are performed jointly thanks to the use of a geographic database, gazetteer or similar type of database that provides explicit georeferences to the possible locations of an input toponym. After this initial step, the possible locations are disambiguated using alternative techniques, which have been deeply studied in the literature. However, what happens if we cannot find the toponyms in a geographic database?
During this STSM we made a proposal of a map-based algorithm for toponyms disambiguation based on clustering techniques, taking profit of the experience on the use of these techniques at IAAA group. The technique is also able to infer the location of those toponyms not found in a geographic database thanks to the previously disambiguated toponyms. Finally, during the STSM we also prepared a paper for a submission to the ACM SIG Spatial conference.

3. Description of the main results obtained

The main result obtained during this research stay has been the proposal of a map-based algorithm for toponyms disambiguation based on clustering techniques. With this method we are able to define a geographic area where toponyms that are not referenced in gazetteer are located.
Our proposal is a hybrid solution that combines map-based disambiguation with the assignment of georeferences for new toponyms. The idea is to infer locations from the locations of previously disambiguated toponyms. The spatial inferences are represented by a geographical area which can be refined depending on various spatial information contained in the textual descriptions.
This approach has been implemented and we made experiments using a corpus of hiking descriptions in French, Spanish and Italian. Our proposal obtains good results. Table 1 shows the accuracy of our method. The difference between the accuracy with spatial inference and without represents the number of unreferenced toponyms that are well located by our approach.

	French	Spanish	Italian
Total # of toponyms	595	376	409
Accuracy without SI	57.14%	61.70%	45.96%
Accuracy with SI	87.39%	88.82%	75.30%

Table 1: Accuracy without and with spatial inference (SI)

4. Future collaboration with the host institution

We have initiated a new joint research line about the formal definition of a language for the spatial annotation of texts. For the definition of this language and the development of applications based on this language we are planning to adopt a Model Driven Engineering (MDE) approach. IAAA group has experience in the use of MDE for the definition of geographic metadata standards and deriving automatically metadata editors. A similar approach could be used for the definition of a spatial annotation language and the development of applications for spatial annotation, or more complex applications such as the reconstruction of routes from textual descriptions or building spatial maps.

5. Publications/articles resulting from the STSM

We submitted the paper: *"Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus"* to the ACM SIG Spatial conference.

6. Confirmation by the host institution of the successful execution of the STSM

See attached certificate signed by Javier Nogueras Iso, associate profesor of the University of Zaragoza, senior member of the IAAA group, and responsible of STSM at the host institution.