Abstract

This paper describes a way in which Natural Language Processing (NLP) can be employed to provide support for real-life geographical navigation guided by referenced locations and other entities found in texts. It is shown how NLP techniques, such as name entity recognition, anaphora resolution and entity linking, supported by digital geographical resources, can provide localization and cross-references of places, trajectories and other types of relations detected in texts. The architecture and functionality of a prototype application is briefly presented.

Key words: data mining, geographical references, mobile application

1. Introduction

In this paper we describe the MappingBooks project, intended to build a technology that complements a book content with information acquired from external sources. The intention is to develop a new type of electronic product with a high impact in education and tourism. The main users envisioned are high school pupils learning the geography of Romania and tourists visiting Romania. The technology combines text analytics, web cartography and mapping, augmented reality techniques and ambient intelligence/ubiquitous computing. Name entities mentioned in school manuals and touristic guides are localised on hypermaps and put in correlation with the reader’s location and related data. If the name entities refer to toponyms, they are supplemented with diagrams or graphical materials. For example, if the school book contains a mention of the Ceahlău Mountain and the young reader happens to be in its proximity, not only that a localisation of her/himself will be signalled on an electronic map that includes the mountain, but information fetched from external sources, such as background layers, as discrete raster and vector data, and additional multimedia information, mainly represented as pop-ups and hyperlinks, all depending on the textual context in which the toponym occurs, will be displayed on the reader’s mobile screen.

As for the tourist guide, up-to-date information regarding the planning of a trip, accommodation possibilities on the mountain, etc., in general data not directly available in the guide, will be displayed near the map.

---

1 First developed as term projects by students in the Artificial Intelligence course at the Faculty of Computer Science of the „Alexandru Ioan Cuza” University of Iași (UAIC).
As such, a *MappedBook* is a book connected with locations in the real and virtual world, which, to the discretion of the user, could be sensitive to the instantaneous location of the reader, as seized by her/his mobile, and that signals, at appropriate moments, events in the real or virtual world related to the location of the toponyms and other entity names the book contains.

After a short presentation of the current related technologies, we present a general overview of the system and some conclusions.

2. **State of the art and beyond**

*MappingBooks* will apply, enhance and invent new e-content and language processing technologies, capable to recognise and annotate different types of name entities, to link these mentions onto the real and virtual world, and to synchronise, by applying ambient intelligence and ubiquitous computing techniques, the mentions in the book with the actual position of the user and the images captured by the mobile user’s device.

In order to do that, basic text analytics include, minimally: segmentation at sentence level, tokenization, part-of-speech tagging, lemmatization, noun-phrase chunking, name entity recognition and anaphora resolution. These are well known techniques that have reached technological maturity for the language under investigation, which is Romanian. In past projects like CLARIN\(^2\), METANET4U\(^3\) and ATLAS\(^4\) modules and processing chains operating on the lines described have been realised. In ATLAS, for instance, individual tools have been combined (Anechitei et al., 2013) in processing chains by using CAS objects of U-Compare, a UIMA dialect (Ferruci and Lally, 2004).

A previous research developed at UAIC (Cărașu, 2011; Ciucanu, 2011) in correlation with a post-doc project (Dumbravă, 2010-2013) used patterns of pragmatic knowledge inferred from a manually annotated Romanian text\(^5\) to transform a textual description of a journey onto trajectories on Google maps.

Linking book mentions onto the real and the virtual world involves a very good name entity recogniser for Romanian. Some progress has been made in this respect, as for instance the ANNIE-GATE module\(^6\), the resources reported by Dumitrescu and Barbu-Mititelu (2013) and the recent work of Potolincă (2014) who acknowledges the acquisition of a gazetteer containing 103,601 names and a collection of 400 hand-written patterns. For English, as part of the MUC-7 competition, the system LTG (Mikheev et al., 1998) reports an F-measure of 93.4%, when human performance do not overpass 97%.

Important in *MappingBooks* is the discovery on the web of data linked to books’ mentions. For this, entity linking techniques are applied. A rich collection of methods is summarised in the exceptional survey of Năstase et al. (2013). The project intends to develop ad-hoc one-document and cross-document linking techniques, by exploiting the

---

\(^{2}\) http://www.clarin.eu/external/
\(^{3}\) http://metanet4u.eu/
\(^{4}\) http://www.atlasproject.eu/
\(^{5}\) “Iter in Chinam” by Milescu Spătarul – about 1675-1678.
\(^{6}\) A Nearly-New Information Extraction system, at: http://gate.ac.uk/sale/tao/splitch6.html
state-of-the art anaphora resolution techniques (Bagga and Balwdin, 1999; Saggion, 2007; Singh et al., 2011), mainly spotting persons and locations. The UAIC’s RARE system, using a mixed approach which combines symbolic rules with learning techniques, has been recently proved to work well for Romanian and other languages (Anechitei et al., 2013). Moreover, dynamic links will be created between mentions in the processed books and their reflexions in the virtual space. The project will develop an integrated technology which will put in the same melting pot known and new techniques to create multi-dimensional mash-ups that combine textual, geographical and, to some extent, temporal information in order to present them adequately to a reader, intermediated by a specially-designed human-computer interface (HCI). The links should be sensible to the context of the book, the moment of time the access is initiated by the reader, the momentary location of the reader and her/his orientation in space.

3. General architecture of the system

3.1 Text pre-processing

Figure 1 displays a diagram of the interconnected modules of the system. The Text Analytics module (TA) receives in input the original text of the book in multiple document formats, including PDF and DOC. A process, known as boiler-plating, removes images, captions, footnotes, headers, footers and page numbers, i.e. any elements breaking the cursive unfolding of the text. However, the original layout of pages is also preserved and the page coordinates of all linguistic elements (words) are recorded in order to allow highlighting of specific word sequences that mark entities, spatial relations, or paths on the user interface, and to know where to place balloons or flying windows evidencing linking information.

Once cleaned from formatting, the content text is passed through some of the UAIC’s web services (Simionescu, 2011) that perform: sentence segmentation, tokenisation, POS-tagging, lemmatisation and noun phrase chunking, including also the identification of head words of noun phrases.

3.2 Entities and relations

The output, an XML file, is passed to the Name Entity Recognition module (NER), which is responsible for marking names and their types. Presently, we recognise 8 categories of names: people, institutions/buildings, streets/areas, cities, districts, countries/regions, mountains, waterways. If an ambiguity occurs (the same name identifies two cities in different countries or a city and a river, for example), the module asks help from the geographical information module (GEO), which returns the most probable category, usually that one which is placed on the map in a closer geographic vicinity to other entities recently mentioned in the same document. The NER module incorporates also an Anaphora Resolution algorithm, in order to link different mentions of the same entity (through pronouns or other nominal constructions). As such, all references to the same entity in the book are chained onto a unique semantic representation.

Marked entities are then crawled on the web by the Entity Crawler module (EC), in order to establish external links. Sources of external information are Wikipedia,
museum sites, whether reports, etc. The attached data varies according to the type of each entity and the context it is mentioned. For instance, for a country type entity, area, population, official language, national feats and other information could be fetched in. For an institution type as is a museum – visiting hours, entry fees, exhibitions currently hosted by the museum, etc.

The last step of the linguistic processing chain is realised by the Relations Detection module (RD), which is responsible for putting in evidence semantic relations mentioned in the text. To exemplify, the set of patterns (written as regular expressions, as used by the RD module) below detect built-by relations, that link institution/building type of entities with entities of type person.

@built-by
#Ent_[1]+{\s\d\W\p{L}}+a fost refacută{\s\d\W\p{L}}+#Ent_[0]+  
#Ent_[1]+{\s\d\W\p{L}}+a fost terminată{\s\d\W\p{L}}+#Ent_[0]+  
#Ent_[1]+{\s\d\W\p{L}}+este creația lui{\s\d\W\p{L}}+#Ent_[0]+  
#Ent_[1]+{\s\d\W\p{L}}+realizată de{\s\d\W\p{L}}+#Ent_[0]+  
#Ent_[1]+{\s\d\W\p{L}}+a fost construită{\s\d\W\p{L}}+#Ent_[0]+

Other relations are descriptions of trajectories. For instance, the ones below (written as XML patterns, as used by the geographical relations detection module) links two location entities as being near to each other.

@near
<w>in</w><w>in</w><w>apropiere</w><w>de</w><w>se</w><w>invecineaza</w><w>din</w><w>spre</w><w>pe</w><w>cu</w>

All patterns described above are built from extracted examples, by observing common types of relevant text fragments. However, in the project we envisage to build a corpus of annotated examples out of which an automatic learning system will infer more diversified patterns. At the end of the described chain, a heavily annotated XML file is accumulated. Then, the entity names, as identified by the NER component, are enriched with information of geographical nature.

3.3 Adding geographical data

The Geography module (GEO) uses free databases (Google GeoCoding API, GeoNames) to complement the information with geographical layers, as is the actual positioning of the entities found in the document on real maps (provided by GoogleMaps). Real-world geographical relations (such as a city being part of a country) are also determined by the GEO module. Let’s notice that, very often, this kind of
semantic relations are not directly specified in the original text. Sometimes GEO realises also a disambiguation step, as described before. Besides these functionalities, GEO also accumulates a JSON database of all entities found in the document (institutions, hotels, various landmarks, etc.). This database will be used to display important locations near the current position of the user, as can be seen in Figure 3.

Finally, the Maps and Trajectories module (M&T) uses Google Maps APIs to trace routes (geographical trajectories) out of the XML notations that complement the descriptions of trajectories in the text. When patterns containing sequences of location entities are fired, routes are generated. The coordinates determined by the GEO module are used to compute the distances between locations. Routes need to have at least three locations in order to be marked as such.

### 3.4 The client-server model

The system is developed as a Client-Server architecture in which various modules continuously exchange data. The modules described up to this point run on the Server. Also there resides the file marked in Figure 1 as the Portrait, which keeps all relevant information about the user.

Two other modules run on the Client – in our case, the user’s mobile device: the HCI Interface (INT), displaying appropriate maps, diagrams, associated information from web sources and highlights on the text, and the Augmented Reality (AR) module.

The client interface receives from the Server and displays the current section of the active document, which is fragmented in chunks smaller than page-size. The reason for the segmentation is to minimize the volume of circulated data. As seen in Figure 2, the interface displays the current text chunk and the location of a selected entity (in this case “Bucium”), as well as the additional information fetched by the EC module (when it exists). Figure 3 shows the current position of the device (determined by the mobile’s GPS) as well as minimum 3-4 locations nearby (sorted dynamically).

The Augmented Reality module (only sketched in the current prototype) is intended to augment the image captured by the mobile’s camera with fleshes pointing the direction of relevant objectives revealed by the current chunk of text (important buildings and streets in a town, mountains in the open, etc.).

### 3.5 Current development state and evaluation status

This architecture offers flexibility and reusability. The described design, as proved during the development of the prototype realised by students, has also the advantage of modularity, the individual modules being built, added and tested as unrelated small projects by groups of students. The prototype was coded in Java and uses JSON databases and Android APIs.

MappingBooks is designed as a mobile application, allowing users to access the content they have rights to (by entering an access code or by uploading a document themselves).

After the prototype system is closer to the final stage of development an evaluation stage is planned. Evaluation can be done both at module level and as a whole system. Since the individual modules are changeable with any equivalent tool providing the
same functionalities, and since the evaluation of those modules is a task optimally performed by their developers and to be discussed when describing those tools individually, we will focus mostly on evaluating the general performance of the whole system. Quantitative evaluations will be done by comparing the output of the system against annotations in a corpus containing relevant examples. But the nature of the system makes it more suited to a qualitative evaluation, to be performed by people in interested social categories: students, authors and editors, general consumers of geographical textbooks and travel guides. This type of evaluation requires a system near completion, so it will be carried out and it’s results described in a forthcoming paper.

4. **Conclusions**

The project advances the state-of-the-art in several directions. First, it invents a technology able to automatically annotate toponyms in Romanian texts and link them onto digital maps. Then, the hypermap structures are automatically updated based on user-dependent contexts. Starting from the book’s text, the user receives support which is presented in a graphical interactive manner. A mixed reality technology indicates on the screen of a graphic tablet or of a smart phone, superposed on an image captured by the device’s camera, the position of certain geographical guiding marks (mountains, hills, localities, roads, railway lines, outstanding buildings, etc.), simultaneously with highlighting them on the text of the book. Moreover, exploiting a collaboration with the Faculty of Geography at UAIC, the project creates an inventory of spatial and geocoded data, which can supplement the map with information related to the context of the identified toponyms. The project also intends to augment the linguistic resources for the Romanian language with a manually annotated corpus (a Geography manual or a touristic guide) at geo-content, out of which minimum 500 instances of geographical concepts will be generated and integrated as instances onto the Romanian WordNet (Tufiș and Cristea, 2002), as open data for research.

**Acknowledgements**

The authors thank to the students in Computer Science at UAIC, who, during the first term of the university year 2013-2014 realised a prototype of the MappingBooks system as their class project in the Artificial Intelligence course, under the supervision of the authors and our colleague Mădălina Răşchip. The research described in this paper constitutes preliminary work to the PN-II-PT-PCCA-2013-4-1878 Partnership PCCA 2013 grant, having as partners UAIC, SIVECO and „Ştefan Cel Mare” University of Suceava.
MAPPING BOOKS: LINGUISTIC SUPPORT FOR GEOGRAPHICAL NAVIGATION IN TEXT RESOURCES
Figure 1: The architecture of the MappingBooks system

Figure 2: Application tablet interface: text and geolocations viewer

Figure 3: Application tablet interface: important locations near user device
References


