Automatic Extraction and Representation of Geographic Entities in eGovernment

Mário Rodrigues[1], Gonçalo Paiva Dias[2]
Polytech. Sch. Techn. & Manag.[1,2]/IEETA[1]/GOVCOPP[2]
University of Aveiro
Aveiro, Portugal
mjfr@ua.pt[1], gpd@ua.pt[2]

António Teixeira
Depart. Electronics, Telecom. & Informatics/IEETA
University of Aveiro
Aveiro, Portugal
ajst@ua.pt

Abstract— **in this paper we present a system that automatically extracts and geocodes named entities from unstructured, natural language textual documents. The system uses the Geo-Net-PT ontology and Google maps as auxiliary data sources. This type of system is particularly useful to automate the geocoding of existing information in e-government applications, which usually requires human intervention. Within the paper we introduce the relevant human language technologies, describe the system that was developed, present and discuss the preliminary results and draw the relevant conclusions and future work.**

Keywords: e-government; geographic information systems; human language technology;

I. INTRODUCTION

The Organization for Economic Co-operation and Development (OECD) defines e-government as "the use of Information and Communication Technologies (ICT) on government activities" [1]. This broad definition encompasses multiple dimensions of e-government, including, among others, electronic service delivery [2,3], one-stop government [4,5], interoperability [6,7], electronic procurement [8], transparency [9] and business process reengineering [10]. According to the World Bank, e-government benefits can be "less corruption, increased transparency, greater convenience, revenue growth, and/or cost reductions" [11].

Municipalities constitute one very relevant branch of government. Because of their land management responsibilities, they deal with geographical information and related textual information. One common problem is that the geocoding of existing textual information is not an easy task, since it usually requires human intervention. The availability of technology to automate this process can be an important breakthrough. It can lead to important cost reductions and, consequently, to stimulated e-government adoption at the local level.

When unstructured text documents are at stake, Human Language Technology (HLT) seems like a good alternative to geocode textual information [12]. In this paper we present a HLT based system that automatically geocodes information extracted from minutes of a Portuguese city council. The system uses Natural Language Processing (NLP) to find geographically related named entities in unstructured, natural language textual documents, structures and geocodes this information and displays it using spatial representation. Geo-Net-PT ontology [13] and Google Maps are used as additional data sources of the system.

The remaining of this paper is organized as follows. Section II includes the state-of-the art in terms of Named Entity (NE) recognition, the main HLT technology used in the system. Section III describes the developed work, including the overview of the system and the presentation of its components: the NE recognizer, the information manager, the database and the web server. Some preliminary results are presented in Section IV and the paper ends with future work and relevant conclusions in Section V.

II. NAMED ENTITY RECOGNITION

Named entity recognition is a subtask of the broader area of Information Extraction (IE), the area which studies how to produce unambiguous information from unseen, unstructured, natural language texts [14]. Named Entity (NE) recognition seeks to locate and classify atomic elements in the text into predefined categories such as the names of persons, organizations, locations and so on [15]. It involves processing a text and identifying certain occurrences of words or expressions as belonging to particular categories. NE recognition software serves as an important preprocessing tool for tasks such as information extraction, information retrieval and other text processing applications.

By assigning categories to entities, NE recognition allows to build systems to retrieve semantically relevant content. For example, NE annotation allows to search for texts about a person called "Gates", without receiving documents about things called gates. In a document collection annotated with NE information it becomes possible to find documents about Java the programming language without getting documents about Java the country or Java the coffee [16].

The approaches followed in NE recognition can be grammar based, statistical based, or a mix of both. Grammar based systems use hand-crafted rules – that usually reflect the linguistic rules of the natural language being analyzed – to find word patterns corresponding to the NEs. Statistical based systems use large amounts of manually annotated data to train a statistical model that is later used to decide which groups of

words form NEs. Early approaches to NE recognition needed extensive gazetteers – lists of names of people, organizations, locations, and other NE. The compilation of such gazetteers is considered a bottleneck in the design of NE recognition systems because it requires a big effort to create and update such lists, so modern approaches avoid using gazetteers [16]. For English and Portuguese languages, state-of-the-art systems are performing NE recognition over the web contents with good results [17, 18].

III. DEVELOPED WORK

The developed system is composed by four modules: a NE recognizer, an Information Manager, a Database and a Web Server (see Fig. 1). The NE recognition module parses the natural language unstructured documents to find NEs that have a physical location as streets, companies, markets, etc. The collected NE list, and respective positions in the document text, is then passed to the information manager which is responsible to query Google Maps to retrieve its location. If no information is retrieved, the NE is added to the database and marked as unresolved in order to be queried again in the future. When a location is retrieved, the information manager stores the new NE in the database, together with its location and accuracy. The first occurrence of the NE is then stored in the database.

Finally, when a user accesses the webpage, the web server queries the database and, using Google Maps for information rendering, composes a webpage that renders the map and the available locations. By selecting a location, the corresponding reference to the document is displayed, together with the context where that location occurred.

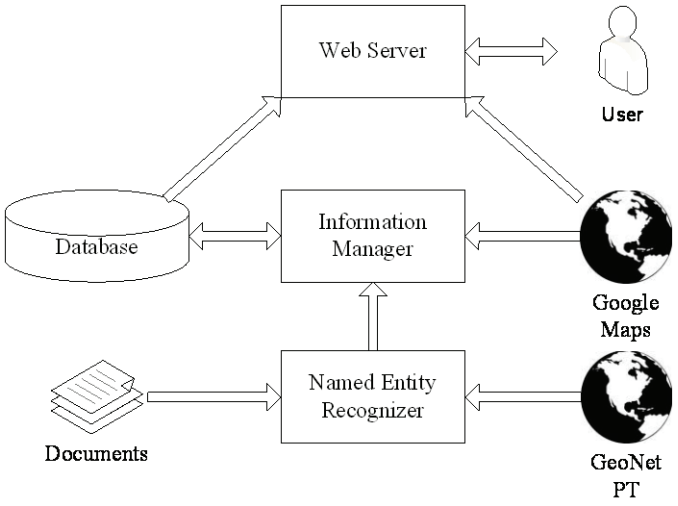

Figure 1 – Schematic representation of how the system modules interoperate: Named Entity Recognizer, Information Manager; Database and Web Server.

A. Named Entity Recognizer

Named entities include people, organizations, locations, and so on. For entities such as people, which are able to move, it does not make sense to attribute them a fixed mark in a map. As such, the NE recognizer was build to just detect NEs that have fixed physical location, as organizations or specific buildings.

The geographical ontology Geo-Net-PT was used to search the documents for entities containing some kind of address clues, as for example "… a associação de amigos dos animais de Sever do Vouga …", where "Sever do Vouga" is a location. The NE identification process starts by determining candidate words to look up in the ontology. The selected candidate words include: all capitalized word forms (e. g. company and/or association names); and the words that are included in the feature set of the ontology, for example *rua* (street), *estrada* (road), *rio* (river), etc. Then, for each candidate word, the ontology is searched to select every record that contains the candidate word. The recognized NE is the biggest neighborhood of the candidate word in the document text, and candidate words that are already part of a NE are not looked up in the ontology. A candidate word is discarded if it was not found in the ontology and, at the same time, does not include any word of the feature set (*rua*, *estrada*, …).

In the example "… a associação de amigos dos animais de Sever do Vouga …", there are two candidate words: Sever; Vouga. When searching the ontology by "Sever", three results are found: "Sever"; "Sever do Vouga"; and "rio Sever". The record "Sever" matches 1 word of the document, the record "Server do Vouga" spans across 3 words of the document and "rio Sever" does not fully match the document text. Thus "Sever do Vouga" is the biggest neighborhood, in the document, of the candidate word and is marked as a NE. After this, the word "Vouga" is part of a NE and, consequently, it is not looked up the ontology. Despite being a great source of structured information, the ontology does not have latitude and longitude values for the vast majority of entities, making it unusable to geocode them.

B. Information Manager

This module has two main tasks: to verify if it is necessary to query Google Maps; and to process information returned by Google Maps. To accomplish the first task, after receiving a list of NE (and respective positions in the document text) from the NE recognizer, the information manager queries the database to know if a given NE was already resolved. If it was resolved previously, the information manager just adds the NE occurrence to the database (the full filesystem path of the document and the NE position in the document). Otherwise it will query Google Maps for latitude and longitude, in order to know the location of the NE. The NE location and its occurrence in the document are stored in the database. If Google Maps does not give a valid answer, just the occurrence is added. Periodically, the module queries the database for unresolved NEs and then queries Google Maps for their locations. If information is available, the corresponding records are updated. Also it is possible to query periodically the location of known entities to update the database information.

The second task is very important because, when referring to locations, places are usually mentioned by an accuracy level that does not fully specify the place. For example, when speaking about "Valencia", people in Spain can refer to the city as "Valencia" instead of "Valencia in Spain", unless they have

asked if it is "Valencia in Venezuela". When locating cities the problem is minimal because just a few cities have the same name. When locating streets the same is not true. More than 10 Portuguese cities have an avenue called "Avenida da Liberdade", for example. Also, to be tolerant to typing errors, it's not uncommon for Google Maps to give information of a place that spells similarly to the one that is queried. The combination of these two characteristics leads to fact that, frequently, a query to Google Maps returns more than one location. In these cases, all places that do not contain the exact NE in the name are discarded. Then, if the result still has more than one location, the NE is marked as ambiguous and no location is assigned.

C. Database

For this first prototype, the information is stored in a relational database using MySQL as the database management system. The corresponding unified modeling language (UML) class diagram is presented in Fig. 2. The attributes of a document are its title and its filesystem full path. An occurrence is characterized by the page number where the named entity was found, and the start and end positions of the named entity text in the document. The attributes of a named entity are its name and if it was resolved or not. A named entity is resolved if a unique location was found for it, and a location is defined by its latitude, longitude, and accuracy.

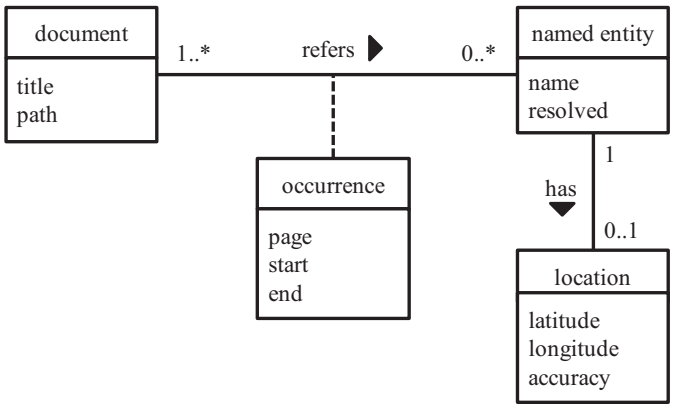

Figure 2 – UML class diagram. Documents have references to named entities and each reference is registered as an occurrence. If a named entity is resolved, then it has a known location.

D. Web Server

The web server used is Apache, and the access to the database is made using PHP language. All locations contained in the database are marked in the map. When a mark is selected, a request is made to the server and a PHP script selects and renders a web page containing a list of documents where the corresponding locations (and thus the entities) were found, and also the list of sentences containing the respective NE (see Fig. 3). The names of the documents are presented with a colored background and, below each name, the list of the NEs and the respective contexts are presented.

Javascript functions control the behavior of the client browser. Every time a mark is clicked, Javascript functions are responsible to make a new request to the web server, and also to automatically center the map in the area to be visualized, while maximizing the number of visible marks.

IV. Preliminary Results

The prototype was used to extract and display information contained in a set of minutes (28) of Águeda Council meetings. In Fig. 3, a part of the web page showing results for "Óis da Ribeira" is presented. The marks in the map show the locations where information is available. By selecting one location, a list of documents containing that geographic reference is displayed. The results represented in Fig 3 were obtained by analyzing documents named with numbers – the colored background number below the map. Below each document name is the list of the NEs that occur in that document – in boldface – and the respective context. From the displayed NE list it is visible that 3 slightly different (in meaning) NEs refer to the same place: "Óis da Ribeira", "Freguesia de Óis da Ribeira", and "Junta de Freguesia de Óis da Ribeira".

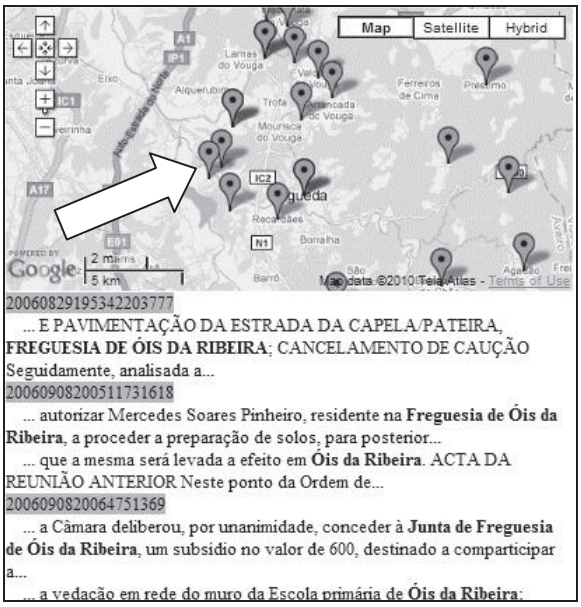

Figure 3 – Grayscale image of the web page. The results presented are a part of the ones shown when the mark of "Óis da Ribeira" is clicked (pointed by the arrow).

From the set of 28 documents containing an average of 6 pages each, the NE recognizer was able to detect 126 named entities with fixed location. The location of all NEs was queried to Google Maps which was able to locate 100 (79.4%) NEs, and unable to locate 26 (20.6%) NEs. A unique answer was obtained for 56 of the 100 located NEs, and more than one location was retrieved for the remaining 44 (see TABLE I).

The proportion of unknown locations should tend to decrease because maps are including more information every day. An approach to try to reduce the rate of unknown locations can be to use other information sources than Google Maps, for example Yahoo Maps or Nokia Maps. The high value of NEs with more than 1 possible location (about 35%)

makes this problem relevant and it should be tackled. A possible way to disambiguate the location is to choose the closest one to the locations mentioned in adjacent paragraphs, since it is likely that adjacent paragraphs refer to the same location. Another is to limit locations to the geographic area of influence of the City Council. Yet another approach can be to mark all locations and let users decide if the information is relevant or not. One example of a NE with more than one location is "Barrô", that is the name of neighborhoods of 3 cities: "Águeda"; "Mealhada"; and "Resende". Examples of unknown NEs are "Espinhel", a neighborhood of "Águeda" that apparently Google Maps does not know, and "Forno" (oven) which can be the name of a very small place.

TABLE I. DISTRIBUTION OF LOCATIONS BY OUTCOME

Results	*N. of Locations (%)*
Known	100 (79.4%)
unique	56 (44.4%)
> 1 location	44 (34.9%)
Unknown	26 (20.6%)

V. FUTURE WORK AND CONCLUSIONS

In this article we presented a system that, using a named entities extraction strategy, automates the process of geocoding entities that occur in unstructured, natural language textual documents. Results seem promising. Nevertheless, further measurements of the system performance are needed. The proportion of NEs detected (compared to the total amount of NEs in the documents), and the amount of correct NEs recognized, were just asserted as good by a close look to a sample of sentences and not thoroughly measured. To have accurate measures it is necessary to compare the system against an ideal system, being considered an "ideal system" a human proficient in the subject. In this early stage of development, our goal was to perform a proof of concept, and such measurements are planned for a near future.

Currently, the system only recognizes entities that have some kind of address clues in its name or that exist in Geo-Net-PT. A future development is to extend the system in order to recognize entities that have a fixed location but do not have any address clues in its name, for example the entity "Escola Secundária de Adolfo Portela" it's an high school – an entity with fixed location – and does not have a reference to an address in its name. The Geo-Net-PT includes some of these entities, but much more are needed and a better way to detect them would improve the system coverage.

To conclude, the proof of concept system automates the process of extracting and geocoding entities included in unstructured, natural language textual documents. The system stores the acquired information in a structured way, and supports access to it through the web by displaying the locations for which there is information in the documents on top of a world map. Furthermore, due to the way information is structured, it is possible to access the original text by selecting that location on the map.

Preliminary tests and consultations show that the system may be useful to automate relevant back-office functions in Portuguese municipalities. Other applications are also envisioned, including geographic based searches of government public information, such as laws, regulations, guidelines, resolutions, studies or plans.

REFERENCES

[1] OECD, Public Management Service, E-Government: Analysis framework and methodology, December 2001.

[2] J. Chamberlain, T. Castleman, "e-Government business strategies and services to citizens", in Proc. of the 2nd Working Conference on e-Business, Seeking Success in e-Business: a Multidisciplinary Approach, International Federation for Information Processing, vol. 123, pp. 309-325, 2003.

[3] R. Leenes, J. Svensson, "Size matters - Electronic Service Delivery by municipalities?", in Proc. of the 1st International Conference on Electronic Government (EGOV 2002), Lecture Notes in Computer Science, Vol. 2456, pp. 150-156, 2002.

[4] G. P. Dias, J. A. Rafael, "A simple model and a distributed architecture for realizing one-stop e-government", Electronic Commerce Research and Applications, vol. 6, 81-90, Spring 2007.

[5] D. Gouscos, M. Kalikakis, M. Legal, S. Papadopoulou, "A general model of performance and quality for one-stop e-Government service offerings", Government Information Quarterly, vol. 24, pp. 860-885, October 2007.

[6] G. P. Dias, J. A. Rafael, "Proposal for a platform for the integration into public administration", in Proc. of the 1st Iberian Conference on Information Systems and Thechnologies, vol. 1, pp. 179-194, 2006.

[7] L. Guijarro, "Interoperability frameworks and enterprise architectures in e-government initiatives in Europe and the United States", Government Information Quarterly, vol. 24, pp. 898-101, Jan 2007.

[8] C. A. Hardy, S. P. Williams, "E-government policy and practice: A theoretical and empirical exploration of public e-procurement", Government Information Quarterly, vol. 25, pp. 155-180, April 2008.

[9] G. P. Dias, J. M. Moreira, "Transparency, corruption and ICT (illustrated with Portuguese cases)", in Transparency, information and communication technology: social responsibility and accountability in business and education, pp. 151-162, 2008.

[10] J. Y. L. Thong, C. S. Yap, K. L. Seah, "Business process reengineering in the public sector: The case of the housing development board in Singapore", Journal of Managemnt Information Systems, vol. 17, pp. 245-270, Summer 2000.

[11] World Bank, http://www.worldbank.org/egov, 2009.

[12] M. Rodrigues, G. P. Dias, A. Teixeira, "Human language technologies for e-gov", in Proc. of the 6th International Conference on Web Information Systems and Technologies, in press.

[13] M. Chaves, M. J. Silva, B. Martins, "A Geographic Knowledge Base for Semantic Web Applications". In Proc. of the 20th Brazilian Symposium on Databases - SBBD Uberlândia, Minas Gerais, Brazil, October, 2005.

[14] D. E. Appelt, "Introduction to Information Extraction", AI Communications, 12(3):161–172, 1999.

[15] K. Bontcheva, B. Davis, A. Funk, Y. Li, T. Wang, "Human Language Technologies", Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies, pp. 37-49, 2008.

[16] A. Mikheev, M. Moens, C. Grover, "Named Entity Recognition Without Gazetteers", in Proc. Of th 9th conference on European Chapter of the Association for Computational Linguistics, pp. 1–8, 1999.

[17] C. Whitelaw, A. Kehlenbeck, N. Petrovic, L. Ungar, "Web-scale named entity recognition", in Proc. Conference on Information and Knowledge Management, pp. 123–132, New York, NY, USA. ACM, 2008.

[18] C. Mota, D. Santos (eds.), "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas", Linguateca, 2008, ISBN: 978-989-20-1656-6.