Automatic Extraction of Geographic Locations on Articles of Digital Newspapers

Cesar García Gómez, Ana Flores Cuadrado, Jorge Díez Mínguez,
and Eduardo Villoslada de la Torre

Abstract. On this article, we present a model to make easier the reading of digital newspapers extracting the location of the news from the articles and showing the places associated with the news on a map. A module of supervised keyword-based extraction recognizes and classifies the geographical locations like named entities. The extraction results are improved using dictionaries or gazetteers (a list of named entities of the geographic area where the news are located). Thesauri are also used to check and complete the results, and for the named entities disambiguation. Finally, the model has been applied to "*El Norte de Castilla*", a digital publication of Vallladolid, to validate and identify the tools and techniques with the best results.

1 Introduction

In the last years, the content of more and more digital publications (magazines, newspapers) has been published on the Internet. There are even publications which can be exclusively read on the Web. To compete on this area, digital publications, besides taking into account the news quality and its importance, must offer value added services to attract a greater number of readers, such as displaying the news on maps based on its geographic location.

Currently, there are applications that allow searches on a group of fixed newspapers based on keywords. E.g. users can search for news, activities or services related to a location of United Kingdom using UpMyStreet [24].

Cesar García Gómez · Ana Flores Cuadrado · Eduardo Villoslada de la Torre
Telefónica Investigación y Desarrollo,
Parque Tecnológico de Boecillo, Parcela 120, 47151, Boecillo, Valladolid, España
e-mail: {cesargg,anafc,evdlt}@tid.es

Jorge Díez Mínguez
Universidad de Valladolid, Edificio TIT,
Campus "Miguel Delibes" s/n, 47011, Valladolid, España
e-mail: jdiez25@yahoo.com

J.M.C. Rodríguez et al. (Eds.): Trends in PAAMS, AISC 157, pp. 101–108.
springerlink.com

However, our goal is to create a more intuitive and visual way of reading digital newspapers, through a model which is able to extract the locations where the news happens or those ones whom the news refers to, in order that the news can be located on a map. The application of Geographical Location Extraction techniques [19, 18] on natural language documents has been used to built this model.

The geographical locations extraction is a subtask (called *Named Entities Recognition and Classification (NERC)* [18, 20, 25, 17, 16]) of a use case of the techniques to extract keywords and phrases (Keyword Extraction / Keyphrase Extraction) [13]. The named entities to be extracted, are usually classified in the following categories; Location names [LOC], Person names [PER], Organizations [ORG] and others [MISC].

The recognition and classification of named entities is based on the use of techniques of Machine Learning, Pattern Matching and Natural Language Processing. Many recognition and classification tools must be trained using a set of training documents manually classified. When the training is over, another set of articles manually classified is used to check the proposed tool, and the results are contrasted with the classification made by human raters. The success degree of these tools can be increased providing them a limited vocabulary (a keywords set (or its synonyms) extracted from a document). In addition, a model based on semantic techniques [26, 2] can be used to improve the process of geographical location, such as thesauri to confirm and complete the results and to resolve ambiguities. The semantic models set up the relation-ship between the news terms that are not addresses, (such as buildings, institutions,...) and the geographical units that represent them. The model is populated with characteristic keywords of an area and it forms part of the dictionary of streets, towns, organizations, institutions,..., which are readily available on Internet (street maps, *GIS (geographic information system)* systems).

Finally, this system has been implemented developing a Web application where the results of the extraction of locations of some sections of "*El Norte de Castilla*", a real digital publication, can be viewed.

The article contains the following parts: Section 2 describes the system architecture and its components, Section 3 presents and compares the results obtained of the improvements application (NERC tools, thesauri...) on the system. Finally Section 4 presents the conclusions and the future lines.

2 System Architecture

The architecture and the workflow of system operations are shown below (Figure 1):

1. The system connects to the *URL (Uniform Resource Locator)* where there are various RSS feeds *(Really Simple Syndication)* with a summary of the news of the digital newspaper. The system reads the contents of the RSS feeds and extracts the complete URLs of the news.
2. The news are read one by one and written into a file, only containing the title and the news text. The HTML (*HyperText Markup Language*) tags and other unnecessary elements are removed.

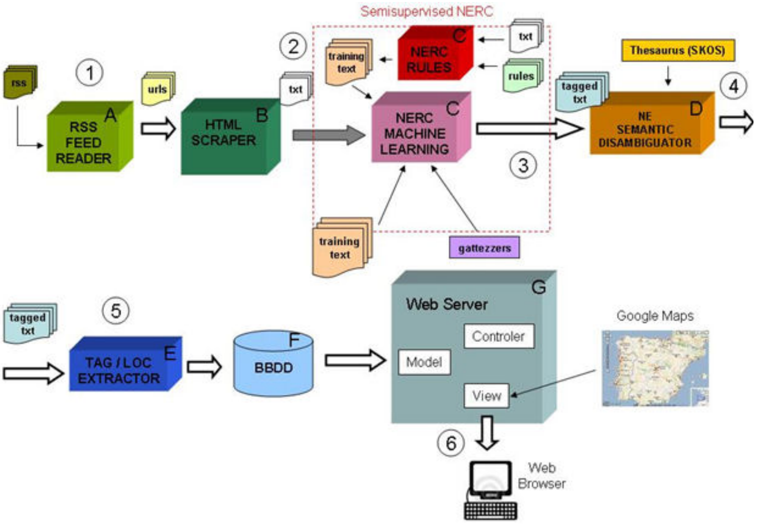


Fig. 1 High level architecture of the locations extraction system

3. Named entities of the news are labeled on the files by a semi-supervised NERC. The semi-supervised NERC is formed by the combination of a machine learning module, previously trained, and a rule-based NERC module
4. The disambiguation of named entities using thesauri and semantic techniques facilitates their identification on the files. The output files will contain the final labeled named entities.
5. In the next step, the named entities labels are extracted and stored on a database to facilitate subsequent consumption.
6. The Web application is in charge of showing a map with the possible locations of news where the users can interact with them.

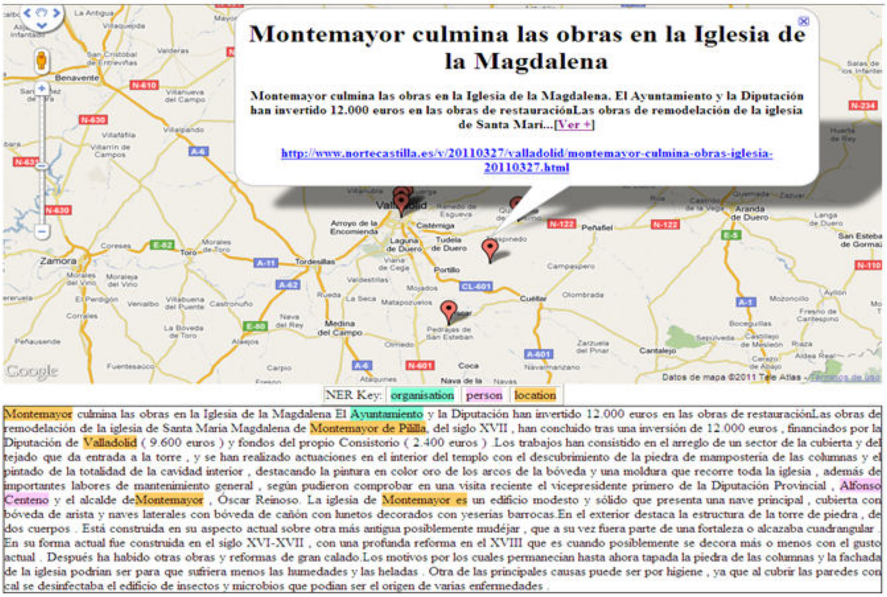
2.1 Components

Each of the architecture components are briefly described on the next paragraphs.

- **RSS Feed Reader:** Reads the *XML (Extensible Markup Language)* / RSS files with the news summary whose URLs are stored on a configuration file, and extracts from them the full URL of each news.
- **HTML Scraper:** Performing HTML markup rules extracts the title and the body of the news from HTML files whose URLs have been obtained by the RSS Feed Reader. It removes the HTML marks and other unnecessary content (advertising, links to other sections of the newspaper, etc).
- **Semi-supervised NERC:** It has two modules: the **NERC module based on machine learning** and the **rule-based NERC**. The **rule-based NERC** module automatically tags the news not labeled manually. These tags are used to train the

NERC based on machine learning. The inputs of this module are the not labeled news and to label them uses the rules stored on a configuration file. The outputs are the *Name Entities* (NE) tagged on the articles depending on the type: [LOC] for locations, [PER] for people and [ORG] for organizations. The **NERC based on machine learning** is responsible of the final labeling of the named entities and it must be previously trained. This module, in addition to the output of the rule-based NERC, makes use of gazetteers of locations, people and organizations, providing a similar output to the one provided by the rule-based NERC, but improved.

- **NE Semantic Disambiguator:** Makes the disambiguation of named entities, classified by the semi-supervised NERC, that have several different meanings. It also checks if the named entities identified by the NERC module have been classified correctly. This module makes use of a thesaurus in SKOS [7] which has the NEs related to the newspaper section where the news appears. Initially, we've focused on the section of news of Valladolid, so we've used a thesaurus generated using the method [8] around the concept of the Wikipedia "*Valladolid*". The semantic disambiguation is done using a ranking of the N possible meanings that the named entity may have on the corresponding thesaurus and selecting the one that reaches the higher valuation as the meaning. To built the ranking, we've added a positively score when other words that maintains a semantic relationship (altLabel, broader and narrower) with the word we want to disambiguate appear on the same article. Unlike other solutions [3] that give the same score to all the words present in a term context, we give different weights to these words depending on their relation with the term to disambiguate
- **Tag and Localization Extractor:** It extracts the definitive named entities from the files with the news tagged by the previous modules. It stores the news on the database, linking each article with its set of named entities. It also makes use of simple heuristics to determine which is the location where the news are produced by applying precedence rules over the named entities of each article. These rules give priority to entities of LOC type over ORG and PER entities. Besides, if there are multiple LOC type entities, the less degree entity of the a political-administrative and geographic hierarchy is selected as location. (E.g. a street has preference over a city, if several entities are at the same level of the hierarchy, the one which appears more often, is chosen and in case of equality is selected the one which occurs above). A customized program has been created instead of using specific rule tools such as Drools [12], due to the rules simplicity.
- **Database:** Contains the tables where the news and the named entities are store in a structured way, facilitating the further consumption of the information by the Web tools.
- **Web Application:** Based on a *Model-View-Controller (MVC)* architecture, it is in charge of consulting the database and displaying the results properly. The user can view the located news of a date and for a particular newspaper section on a map and from there, he can access to the full news content on another section of the Website (figura 2).



Montemayor culmina las obras en la Iglesia de la Magdalena

Montemayor culmina las obras en la Iglesia de la Magdalena. El Ayuntamiento y la Diputación han invertido 12.000 euros en las obras de restauraciónLas obras de remodelación de la iglesia de Santa Mari...[Ver +]

http://www.nortecastilla.es/v/20110327/valladolid/montemayor-culmina-obras-iglesia-20110327.html

NER Key: organisation | person | location

Montemayor culmina las obras en la Iglesia de la Magdalena El Ayuntamiento y la Diputación han invertido 12.000 euros en las obras de restauraciónLas obras de remodelación de la iglesia de Santa María Magdalena de Montemayor de Pililla, del siglo XVII , han concluido tras una inversión de 12.000 euros , financiados por la Diputación de Valladolid (9.600 euros) y fondos del propio Consistorio (2.400 euros) .Los trabajos han consistido en el arreglo de un sector de la cubierta y del tejado que da entrada a la torre , y se han realizado actuaciones en el interior del templo con el descubrimiento de la piedra de mampostería de las columnas y el pintado de la totalidad de la cavidad interior , destacando la pintura en color oro de los arcos de la bóveda y una moldura que recorre toda la iglesia , además de importantes labores de mantenimiento general , según pudieron comprobar en una visita reciente el vicepresidente primero de la Diputación Provincial , Alfonso Centeno y el alcalde deMontemayor , Óscar Reinoso. La iglesia de Montemayor es un edificio modesto y sólido que presenta una nave principal , cubierta con bóveda de arista y naves laterales con bóveda de cañón con lunetos decorados con yeserías barrocas.En el exterior destaca la estructura de la torre de piedra , de dos cuerpos . Está construida en su aspecto actual sobre otra más antigua posiblemente mudéjar , que a su vez fuera parte de una fortaleza o alcazaba cuadrangular . En su forma actual fue construida en el siglo XVI-XVII , con una profunda reforma en el XVIII que es cuando posiblemente se decora más o menos con el gusto actual . Después ha habido otras obras y reformas de gran calado.Los motivos por los cuales permanecían hasta ahora tapada la piedra de las columnas y la fachada de la iglesia podrían ser para que sufriera menos las humedades y las heladas . Otra de las principales causas puede ser por higiene , ya que al cubrir las paredes con cal se desinfectaba el edificio de insectos y microbios que podían ser el origen de varias enfermedades .

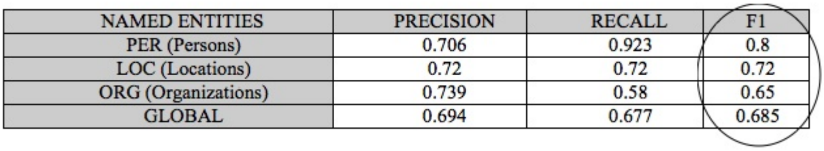
Fig. 2 Web application screenshot over an article of a digital publication.

3 Experimentation and Validation

The evaluation purpose is to quantify the improvements introduced in the system with the use of the techniques presented in this paper. The technique defined at the IREX [11] and CoNLL [6] conferences has been used for the evaluation. This technique is based on the comparison of MAF *(micro-averaged f-measure)* (MAF or F1 = 2 * P * R / (P + R)). Where *precision (P)* is the percentage of correct named entities found and the *coverage or recall (R)* is the percentage of named entities present on the articles found.

For the NERC tools selection, it has been made a study of several of them, which have selected by their good results at conferences and congresses [18, 10, 5] and by holding a steep learning curve. The tools selected were LBJ NER Tagger 1.2 [14, 21], Stranford NER [22], Lingpipe [1], CAGEclass [4], DRAMNERI [23], LT-TTT2 [15], Freeling [9]. Several tests using different tools over the same set of test articles of "*El Norte de Castilla*"have been performed. For all the tools, we've performed some tests using local and general gazetteers and some tests without them. As a result of the experiment, we can conclude that the best results have been obtained using the LBJ NER Tagger 1.2 tool trained with the CoNLL2002 dataset and using local gazetteers (Figure 3).

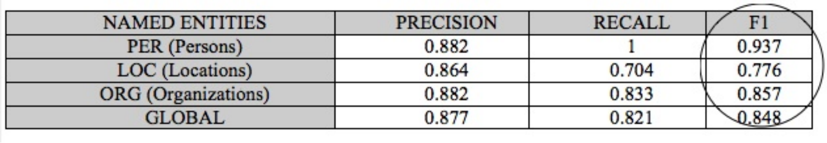
Two techniques have been used to improve the results: The first one consists on including new locations (popular names of areas, neighborhoods or buildings...), local business organizations (companies on the region, hotels and restaurants, churches, museums, ..) on the used dictionaries, and increasing the number



NAMED ENTITIES	PRECISION	RECALL	F1
PER (Persons)	0.706	0.923	0.8
LOC (Locations)	0.72	0.72	0.72
ORG (Organizations)	0.739	0.58	0.65
GLOBAL	0.694	0.677	0.685

Fig. 3 LBJ NER. Training with CoNLL2002 and local gazetteers

of names of registered people. The other one is based on the thesauri use to make the disambiguation of the named entities and to confirm the classification made by the NERC tool. The thesauri are also used to help on the selection of the locations shown on the Web. The Figure 4 shows the improvement on the results made using these techniques.



NAMED ENTITIES	PRECISION	RECALL	F1
PER (Persons)	0.882	1	0.937
LOC (Locations)	0.864	0.704	0.776
ORG (Organizations)	0.882	0.833	0.857
GLOBAL	0.877	0.821	0.848

Fig. 4 LBJ NER. Training with CoNLL2002, improved with local gazetteers and thesauri use

The improvement on the results is evident. The global F1 parameter (related with all named entities) improves about 20% compared to the value obtained before introducing the improvements. Moreover, in the geographical locations (LOC entities) case, the F1 improvement is over 7%. The improvements got in other types of entities are even greater, due to the use of better local dictionaries. If we add to the dictionaries the records of most companies and organizations, which belong not only to the province but also to the autonomous region, the F1 of organizations will be increased around 31%. If we increase the number of names of people, the F1 will be improved on 17%.

Most failures happen on the entity classification, when a entity is classified like a wrong type. E.g. *José Zorrilla stadium* should have been labeled like a ORG type, but it was classified like a person [PER], due to this combination of words does not often appear on the training set and the Wikipedia reinforces this term like a person concept, since the principal meaning of *José Zorrilla* on the Wikipedia is related with the famous writer born in *Valladolid*.

In conclusion, the results of NERC tools can be improved by optimizing the local dictionaries, adding to them entities directly related to the articles analyzed scope and using one or more thesauri related to the geographic area (city, province, region) where the most articles of the digital publication are mainly located.

4 Conclusions and Future Lines

In this paper, we've presented a system and a Web application that offer a more intuitive and visual way to access and read of the news on digital newspapers, applying location extraction techniques. The location information extracted is used to facilitate the reading of online newspapers, identifying visually the news location extracted from the text and showing the places associated with the news on a map.

The use of NERC tools to extract locations on delimited domains provides significant improvements on the performance of some existing NERC systems. Several of these tools improve their performance completing their gazetteers with local named entities. So, our scientific contribution is developing a system to prove that the success rates of these tools can be substantially improved applying two techniques: adding to the dictionaries and gazetteers the named entities related to the area covered by the publication and using semantic techniques such as thesauri to check the results got by the NERC tool and to resolve the named entity ambiguities

There are several areas for future evolution of the system. For example, it could be incorporated new rules which do not limit the identification to independent entities to select the final location of the news and allow to identify the routes or neighborhoods. Likewise, it could be included features that allow users to search for news happened in a date range, or in a given location or even in a geographic area selected on the map.

Another possible future line would be the thesauri generation based on the different categories (sports, economy, society,...) and the sections of the publication to analyze. The thesauri could be selected automatically depending on the section of the newspaper containing the news that we want to locate geographically. Besides, to increase the thesaurus reliability, when it is generated, it would be interesting to integrate other data sources, not just the Wikipedia. Finally regular updates of gazetteer used by NERC tools should be considered to maintain a high success rate.

References

1. Baldwin, B., Carpenter, B.: LingPipe, http://www.alias-i.com/lingpipe/
2. Blázquez, L.M.V., Pascual, A.F.R., Ángel, M., Poveda, B.: Ingeniería ontológica: El camino hacia la mejora del acceso a la información geográfica en el entorno web. In: Subdirección General de Aplicaciones Geográficas del Instituto Geográfico Nacional. Avances En Las Infraestructuras De Datos Espaciales, p. 95 (2006)
3. Brugmann, H., Malaisé, V., Gazendam, L.: Disambiguating automatic semantic annotation based on a thesaurus structure. In: Proc. 14e Conference Sur le Traitement Automatique des Langues Naturelles, TALN 2007 (2007)
4. CAGEclass, http://cageclass.sourceforge.net/ (last visit January 2011)
5. Chinchor, N.: Overview of MUC-7/MET-2. In: Proc. Message Understanding Conference, MUC-7 (1999)
6. CoNLL-2011, http://www.clips.ua.ac.be/conll/ (last visit March 2011)
7. Drools: The Business Object Integration Platform,
   http://www.jboss.org/drools

8. Flores Cuadrado, A., Villoslada de la Torre, E., Peláez Gutiérrez, A.: Generación de Tesauros basado en Media Wiki. Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos 3(6) (2009)
9. FreeLing Home Page, `http://nlp.lsi.upc.edu/freeling/` (last visit April 2011)
10. Grishman, R., Sundheim, B.: Message Understanding Conference-6: A Brief History. In: Proc. 16th Conference on Computational Linguistics, USA, vol. 1, pp. 466–471 (1996)
11. IREX: Information Retrieval and Extraction Exercise, `http://nlp.cs.nyu.edu/irex/`
12. Isaac, A., Summers, E.: SKOS: Simple Knowledge Organization System primer (2008), `http://www.w3.org/TR/skos-primer` (last visit March 2011)
13. Keyphrase Extraction Algorithm. Technical Report. Computer Science Department, University of Waikato. Hamilton, New Zealand, `http://www.nzdl.org/Kea/index.html`
14. Learning Based Java. Cognitive Computation Group. Universidad de Illinois, EEUU, `http://cogcomp.cs.illinois.edu/page/software_view/11` (last visit April 2011)
15. LT-TTT2. Language Technology-Text Tokenisation Tool, `http://www.ltg.ed.ac.uk/software/lt-ttt2` (last visit March 2011)
16. Mansouri, A., Affendey, L.S., Mamat, A.: Named Entity Recognition Approaches. International Journal of Computer Science and Network Security 8, 339–344 (2008)
17. Marrero, M., Sánchez-Cuadrado, S., Lara, J.M., Andreadakis, G.: Evaluation of named entity extraction systems. In: Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), pp. 47–58 (2009)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes 30, 3–26 (2007)
19. Ortega, J.M.P., Cumbreras, M.A.G., Vega, M.G., López, L.A.U.: Sistemas de Recuperación de Información Geográfica multilinges en CLEF. Procesamiento Del Lenguaje Natural 40, 129–136 (2008)
20. Ortega, J.M.P., Ráez, A.M., Santiago, F.M., López, L.A.U.: Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. Procesamiento Del Lenguaje Natural 43, 33–40 (2009)
21. Ratinov, L.: Design Challenges and Misconceptions in Named Entity Recognition, `http://cogcomp.cs.illinois.edu/page/publication_view/199` (last visit April 2011)
22. Stanford Named Entity Recognizer. The Stanford Natural Language Processing Group, `http://nlp.stanford.edu/software/CRF-NER.shtml` (last visit April 2011)
23. Toral, A.: DRAMNERI: a free knowledge based tool to named entity recognition. In: Proc. 1st Free Software Technologies Conference, La Coruña, España, pp. 27–31 (2005)
24. UpMyStreet, `http://www.upmystreet.com/` (last visit April 2011)
25. Vargas, J.D.: Reconocimiento de Entidades Nombradas en Textos no Estructurados. Technical Report. Universidad Nacional de Colombia (2008)
26. Zapater, S., Javier, J.: Ontologías para servicios web semánticos de información de tráfico. Revista digital Dialnet. Lectura en la Universitat de Valencia en 2006 (2006), `http://dialnet.unirioja.es/servlet/tesis?codigo=7157` (last visit March 2011)