Integration of automatic spatial annotations from different sources by means of semantic technologies

Javier Nogueras-Iso, Verónica Lázaro, Ludovic Moncla

Integration of automatic spatial annotations from different sources by means of semantic technologies

Javier NOGUERAS-ISO (1), Verónica LÁZARO (2), Ludovic MONCLA (3)

(1) Aragon Institute for Engineering Research (I3A), Universidad de Zaragoza, C/ María de Luna 1, 50018 Zaragoza, Spain; jnog@unizar.es (2) vlazgar@gmail.com (3) LIUPPA, Avenue de l'Université, BP 1155, 64013 Université Pau Cedex, France; ludovic.moncla@univ-pau.fr

Resumen

La extracción de información es una de las tareas principales de la minería de textos que resulta de gran utilidad para todo tipo de aplicaciones que exploten la información geográfica ya que hay gran cantidad de información geográfica que no se recopila directamente en formatos reconocibles por Sistemas de Información Geográfica, sino directamente como texto plano. Actualmente existen diversas soluciones informáticas para el procesamiento de textos y la anotación de entidades espaciales. Sin embargo, el problema que tienen estas herramientas es que producen como resultado de este procesamiento un texto anotado con lenguajes de marcado propio, que dificulta su integración en otros sistemas. El objetivo de este trabajo es proponer la conversión de la salida de estas herramientas a un lenguaje de anotación espacial común basado en tecnologías semánticas que facilite su integración e interoperabilidad. Como factor común de los lenguajes de marcado se propone una anotación de textos basada en RDFa y utilizando el vocabulario de la iniciativa internacional Schema.org. Para validar la utilidad de esta propuesta se ha creado la infraestructura necesaria para construir un repositorio semántico de documentos donde se integren y armonicen las anotaciones generadas por diversas herramientas de anotación existentes.

Palabras clave: Extracción de información. Integración de datos. Información geográfica. Web semántica. RDFa. Schema.org.

Abstract

Information extraction is one of the main tasks in text mining, which is essential for all types of applications exploiting geographic information because there is a big volume of geographic information not directly compiled in specific formats proposed by Geographic Information Systems, but just embedded in plain text sources. Currently, there are several software solutions for the processing of texts and the annotation of spatial named entities. However, the problem of these tools is that their output is based on heterogeneous annotation languages, which make it difficult their integration in other systems. The objective of this work is to propose the conversion of the output of these tools into a common spatial annotation language based on semantic technologies to facilitate their integration and interoperability. As a common annotation language we propose the use of a text annotation based on RDFa and using the vocabulary proposed by the international initiative Schema.org. In order to validate this proposal, we have created the necessary infrastructure to build a semantic repository of documents, where the annotations generated by different annotation tools can be integrated and harmonized.

Keywords: Information extraction. Data integration. Geographic information. Semantic web. RDFa. Schema.org.

1. Introduction

Information Extraction is one of the main tasks in Text Mining, which is essential for all types of applications exploiting geographic information because there is a big volume of geographic information not directly compiled in specific formats proposed by Geographic Information Systems, but just embedded in plain text sources. For instance, in this context we can find different types of narrative structures describing: itineraries in travelogues or travel novels (Leidner and Lieberman, 2011); geographic boundaries in legislative documents; locations through text transcriptions of calls to emergency services; or any type of official information associated to a geographic location and found in official bulletin documents (López-Pellicer et al, 2012).

For the processing of these information sources and building a structured and simplified view of their content, the recognition of named entities (Named Entity Recognition - NER) is an essential activity (Sekine and Ranchhod, 2009). Nowadays, there are several software solutions like OpenNLP, OpenCalais, CasEN or PERDIDO for the processing of texts and the annotation of spatial entities. However, the problem of these tools

is that their output is based on heterogeneous annotation languages, which makes it difficult to integrate them in more complex systems exploiting these annotations.

The objective of this work is to propose the conversion of the output of these tools into a common spatial annotation language based on semantic technologies to facilitate their integration and interoperability. As a common annotation language we propose the use of a text annotation based on RDFa and using the vocabulary proposed by the international initiative Schema.org. On the one hand, RDFa is a set of extensions to XHTML proposed by the World Wide Web Consortium (W3C) to embed semantics in documents through RDF triples. On the other hand, Schema.org is an initiative promoted by the main search engine companies (e.g. Google, Yahoo, Microsoft, Yandex) to define a common vocabulary of entities (Person, Place, Date, ...) to annotate the semantics of Web resources. The initial hypothesis is that an annotation approach based on standards accepted by the main stakeholders in Web development would increase the general use of these tools for the publication of enriched Web content.

In order to validate this proposal, we have created the necessary infrastructure to build a semantic repository of documents, where the annotations generated by different annotation tools can be integrated and harmonized. Additionally, this repository will provide as well query functionalities for searching documents in terms of spatial filters that take profit of previously annotated spatial entities. Moreover, this repository can be seen as data integration solution (Doan et al, 2012) as annotations obtained from different tools are linked to the same original document.

The rest of the paper is structured as follows. Section 2 describes the annotation tools that have been reviewed for this work. Section 3 introduces our proposal for a common spatial annotation based on semantic technologies. Section 4 describes the design of the infrastructure for a document repository, whose validation is reported in section 5. Last, section 6 concludes the work and presents some issues for future work.

2. Tools for spatial named entity recognition

Among the systems for Named Entity Recognition and Classification (NERC) specifically focused on the detection of spatial named entities (place names or toponyms) we can find: approaches based on Natural Language Processing (NLP), in particular using syntactico-semantic rules (Moncla et al, 2014); approaches based on machine learning (Santos et al, 2014); or even hybrid solutions combining previous approaches (Leidner, 2008). In any case, the most typical feature of this specific NERC systems is the need to use external lexical resources, known as gazetteers, to facilitate the connection between place names and their explicit reference, in terms of geographic coordinates, to the Earth surface.

This section describes briefly the features of four tools for the annotation of spatial named entities and the annotation languages used in their output formats. These tools are OpenNLP, CasEN, PERDIDO and OpenCalais. The reason to study these tools is their availability to be freely accessed and their capability to process texts in English, French or Spanish.

OpenNLP (Apache, 2010) is an Apache Project, available as an Open Source library, which facilitates a set of tools based on machine learning for NLP tasks. In particular, NERC can be performed thanks to a *Tokenizer* tool, which divides sentences into *tokens*, and a tool known as *Name Finder,* which facilitates the named entity recognition. Currently, given the availability of dictionaries, the spatial named entity recognition is only possible for documents in Dutch, Spanish or English. The output generated by OpenNLP is just plain text. For instance, the result of processing the string *"I live in Zaragoza."* would be the following:

[3..4) location

where number 3 indicates the initial token of the place name in the text, number 4 indicates the position of the final token, and *location* indicates the type of the recognized spatial named entity.

CasEN (University of Tours, 2016) is a tool developed at University of Tours (France), also available as Open Source, which offers a solution based on syntactico-semantic rules for the recognition of named entities. In particular, the recognition of entities is based on the definition of cascades of transducers (a specific type of finite state automata) (Maurel and Friburger, 2004) working on top of the Unitex platform, a research project from the University of Paris-Est (France) to facilitate processing text tasks. As the dictionaries for spatial annotation are only available in French, CasEN can be used only for French texts. The annotation output generated by CasEN for a sentence containing the place name *Pau* is the following:

J'ai visité les villages au sud de {Pau,.entity+loc+adm+town+grftagNToponyme}

As it can be observed, the information of the recognized named entities is shown within curly brackets: *entity* for indicating the detection of an entity; *loc+adm+town* for indicating the type of entity;

and *grftagNToponyme* for indicating the graph (transducer cascade) that detected the entity.

PERDIDO (Moncla, 2015a) is the result of a research project made in collaboration between the LIUPPA laboratory at University of Pau (France), the Advanced Information Systems Laboratory at University of Zaragoza (Spain), and the COGIT laboratory at National Geographic Institute of France. This tool, accessible as a Web service, is specifically focused on the recognition of spatial named entities. As well as CasEN, NERC in PERDIDO is based on the application of syntactico-semantic rules on top of the Unitex platform, but also using some additional heuristics based on clustering techniques for the resolution of fine-grained toponyms. Currently, the lexical resources and rules used by this tool enable the processing of texts in French, Spanish and Italian, and it generates an XML based annotation based on the TEI (*Text Encoding Initiative*) markup language. Left side of figure 1 shows an example of the annotation produced for the string input *"Madrid",* which is recognized as a place name with the *placeName* tag. Additionally, the service provided by PERDIDO facilitates another XML file with additional information about the detected toponym (see right side of figure 1). This additional information is recovered from the gazetteer services used as external knowledge resources and consists of information items like the standard name of the toponym (*name* tag), entity type (*feature* tag), geographic coordinates (*coord* tag), country (*country* tag) or continent (*continent* tag).

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <s>
  <placeName>
   <name xml:id="2">
    <w lemma="Madrid" type="NPr">Madrid</w>
   </name>
  </placeName>
 </s>
</TEI>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<toponymes>
 <toponym>
  <id>2</id>
  <name>madrid</name>
  <feature>administrativeUnit</feature>
  <coord>0.16687,41.178711</coord>
  <country>SP</country>
  <continent>EU</continent>
  ...
 </toponym>
</toponymes>
```

Figure 1. Example of the annotation generated by PERDIDO service

OpenCalais (Thomson Reuters, 2016) is a Web service provided by Thomson Reuters for NERC, which adopts a hybrid approach based on a combination of statistics techniques, machine learning and rule-based methods. Within the scope of spatial annotation, OpenCalais is able to process and annotate texts in English, French and Spanish for the recognition of city and country names. The output format generated by OpenCalais is RDF serialized as XML (see discussion on RDF in next section). Figure 2 shows an example of this annotation format obtained as the result of processing the input string *"London"*. In this example it can be observed that tags *c:shortname*, *c:latitude*, *c:longitude* and *c:containedbycountry* inform about the name, the latitude, the longitude and the country to which the detected city belongs respectively.

```
<rdf:type rdf:resource="http://s.opencalais.com/1/type/er/Geo/City" />
 <c:docId
  rdf:resource="http://d.opencalais.com/dochash-1/02d32132-5c1d-38a3-8148-2977e6f9a0ef" />
 <c:name>London,Greater London,United Kingdom</c:name>
 <c:shortname>London</c:shortname>
 <c:latitude>51.517124</c:latitude>
 <c:longitude>-0.106196</c:longitude>
 <c:containedbystate>Greater London</c:containedbystate>
 <c:containedbycountry>United Kingdom</c:containedbycountry><!--London -->
 <c:subject
  rdf:resource="http://d.opencalais.com/genericHasher-1/6fda72fd-105c-39ba-bb79-da95785a249f" />
 <c:rcscode>G:21P</c:rcscode>
</rdf:Description>
<rdf:Description
 rdf:about="http://d.opencalais.com/genericHasher-1/6fda72fd-105c-39ba-bb79-da95785a249f">
 <rdf:type rdf:resource="http://s.opencalais.com/1/type/em/e/City" />
 <c:forenduserdisplay>true</c:forenduserdisplay>
 <c:name>London</c:name>
</rdf:Description>
```

Figure 2. Example of the annotation generated by OpenCalais service

In summary, after the description of these NERC tools, some of them widely used for English, French and Spanish texts, one can check that there is great heterogeneity with respect to the granularity of the toponyms detected and, overall, the output formats used for the annotation.

3. Annotation proposal based on RDFa and Schema.org

RDF (*Resource Description Framework*) is a family of specifications proposed by W3C, which was originally designed as a data model for metadata, but currently it is extensively used to describe semantics of information in a machine-readable way (Cyganiak et al, 2014). The representation model of RDF is based on the use of information triples *subject-predicate-object*: a resource *x* (subject) has a property *y* (*predicate*) with a value *z* (*object*).

Apart from expressing RDF semantics in a graphical way thanks to a graph formalism (e.g., see later discussion on figure 5), this semantics can be serialized as plain text (Turtle/N3 format), XML (RDF/XML format), or embedding the triples within XHTML documents using an extension of XHTML specification, which is known as RDFa specification (Herman et al, 2015).

Our proposal for a common language for spatial annotation of texts is to use XHTML with the RDFa extension and using the vocabulary proposed by Schema.org. On the one hand, XHTML is an XML version of the HTML for the construction of web pages that allows to maintain the original text processed by the spatial annotation tools. On the other hand, thanks to the use of RDFa we can benefit from *meta* and *link* XHTML attributes to introduce the semantics defined by the Schema.org vocabulary.

The mission of Schema.org is to create, maintain and promote schemas for structured data on the Internet. The vocabulary of Schema.org covers entities, relations between entities, and actions. Additionally, it can be easily extended thanks to a well-documented extension model.

In particular, taking into account the context of this work where we are going to process mainly texts describing itineraries, we will initially consider the annotation of the full input document as an instance of a *TravelAction*, i.e. an action involving a displacement. Within a document, we propose to annotate every recognized spatial entity using the following entities proposed by Schema.org: *Place* for the general annotation of place names; *City* for the specific annotation of city names; *Country* for country names; *Continent* for continent names; *BodyOfWater* for water bodies like lakes or rivers; and *Mountain* for geologic accidents. These entities cover the types of spatial named entities recognized by the tools reviewed in section 2, but in the future this set of entities could be extended to take into account other types covered by new annotation tools.

Schema.org	PERDIDO
Place	placeName
	toponym
name	name
description	feature
geo	coord
Country	country
name	
Continent	continent
name	

Schema.org	OpenCalais
Country	Country
geo.latitude	c:latitude
geo.longitude	c:longitude
name	c:shortname
City	City
name	c:shortname
geo.latitude	c:latitude
geo.longitude	c:longitude
Country	c:containedbycountry
name	

Schema.org	CasEn
Place	loc.adm.town /
name	loc.adm.reg / loc.adm.nat / loc.phys.sup
Mountain	loc.phys.geo
name	
BodyOfWater	loc.phys.hydro
name	

Schema.org	OpenNLP
Place	location
name	

Figure 3. Mapping between Schema.org and annotation languages of other tools

With respect to the information provided for each annotated entity, all types of entity share a property called *name* to store the recognized place name. Additionally, depending on the detailed information provided by each annotation tool, additional properties like *geo* or *description* can be used to annotate the associated geographic coordinates or the type of entity respectively. Figure 3 shows the mapping between the vocabulary proposed by Schema.org and the annotation languages used by the annotation tools reviewed in section 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<HTML xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:schema="http://schema.org/"
xmlns="http://www.w3.org/1999/xhtml"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<HEAD/>
<BODY typeof="schema:TravelAction" property="schema:url" content="Near_the_Puerta_del_Sol4">
<p>
    I live in
    <div property="schema:location" typeof="schema:Place">
        <h1 property="schema:name">Madrid</h1>
        <h1 property="rdfs:comment">OpenCalais</h1>
        <div property="schema:geo" typeof="schema:GeoCoordinates">
            <meta property="schema:latitude" content="40.4"/>
            <meta property="schema:longitude" content="-3.6833"/>
        </div>
        <div property="schema:location" typeof="schema:Country">
            <h1 property="schema:name">Spain</h1>
        </div>
    </div>
</p>
</BODY>
</HTML>
```

Figure 4. *Example of our annotation proposal (XHTML-RDFa + Schema.org)*

Figure 4 shows an example of the annotation for the text *"I live in Madrid"* in XHTML-RDFa format and using the Schema.org vocabulary. As it can be seen, the full document is a resource of *TravelAction* type and the recognized spatial named entity *Madrid* has been annotated as a *location* property of this resource.

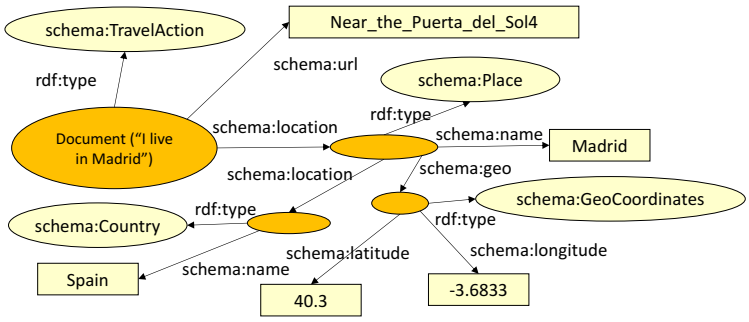

Figure 5. *RDF triples under the perspective of a graph model*

Additionally, in our annotation proposal we include complementary information about the annotation process. For instance, thanks to the

schema:url property we store the identifier of the file containing the original input text that has been processed. We also use the *rdfs:comment* property in each entity of *Place* type (or any of its derived types) to indicate the annotation tool that recognized the toponym, e.g. OpenCalais in this example.

Graphically, the content annotated in figure 4 could be also seen under the perspective of an RDF graph as shown in figure 5. Moreover, this RDF content can be stored as well in a triple store enabling the construction of a document repository, which can be queried by query languages like SPARQL (SPARQL Protocol and RDF Query Language) (Harris and Seaborne, 2013). SPARQL is a standardized language proposed by W3C for the query of RDF graphs. For instance, figure 6 shows a query to retrieve documents containing the place name *Madrid* in a graph model like the one proposed in figure 5.

```
PREFIX schema:<http://schema.org/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?x ?name ?nameFile
WHERE {
  ?x rdf:Description ?id.
  ?x rdf:type <http://schema.org/TravelAction>.
  ?x schema:url ?nameFile.
  ?x schema:location ?lo.
  ?lo rdf:type <http://schema.org/Place>.
  ?lo schema:name ?name.
  {FILTER regex(?name, 'Madrid')}*
}
```

Figure 6. *Example of a SPARQL query for documents containing spatial named entities*

This duality of RDFa and RDF is useful for the work proposed in this paper as our purpose is twofold. On the one hand, XHTML-RDFa allows to maintain the original text with the in-line annotations. On the other hand, if more specific queries on documents are required, the storage of RDF triples on a triple-store enables the use of powerful query languages like SPARQL.

4. Infrastructure for a document repository

Figure 7 shows the multilayer architecture of the infrastructure we have designed to build a system able to convert the output of annotation tools into our annotation proposal and, at the same time, feed the contents of a document repository, which can be searched in terms of the spatial entities that have been previously recognized.

In the bottom layer of the architecture we can find the external annotation tools and the RDF triple store based on JENA technology, which gives support to the document repository providing search functionalities. Additionally, it must be noted that we also store the XHTML-RDFa documents to let users see the annotations produced by the annotation tools at the exact position of the original text.

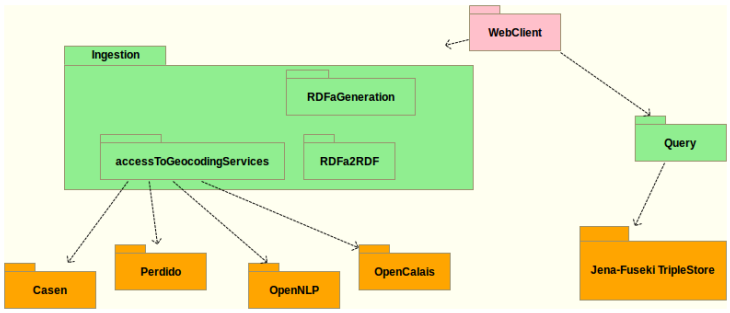

Figure 7. *Architecture of the document repository*

On top of this bottom layer, there is a business logic layer with the *Ingestion* and *Query* subsystems. The *Ingestion* subsystem enables the access to the external annotation tools, the conversion of their output to our annotation proposal, and the ingestion of documents in the repository. The *Query* subsystem facilitates the services that are necessary to facilitate searches in the document repository using SPARQL.

Last, the upper layer consists of the web application that allows the interaction with the final users for the setup and ingestion of documents in the repository, and the searching of documents according to different spatial filters. The following subsections describe in more detail the *Ingestion* subsystem and the Web application.

4.1. *Ingestion* subsystem

In order to facilitate the access to the different annotation tools, the *Ingestion* subsystem, developed with Java technology, uses an *AbstractFactory* design pattern to separate the problem of creating components with a standard annotation functionality from the problem of using different access protocols: whereas CasEN and OpenNLP can be integrated as libraries in the source code, PERDIDO and OpenCalais must be accessed as Web services.

Another important aspect in the *Ingestion* subsystem has been the conversion of the outputs of the different annotation tools into our common proposed language. In order to implement the mapping specified in figure 3, specific customizations were required for each tool. In the case of CasEN and OpenCalais, generating plain text as output, it is required to perform a pre-processing to generate

an intermediate XML markup language. Then, this intermediate markup language, as well as the annotation output of PERDIDO (directly available as XML), is transformed into our proposed XHTML-RDFa format using XSLT technology (eXtensible Stylesheet Language Transformations). On the opposite side, in the case of OpenCalais, as this tool generates an RDF annotation, the transformation can be directly programmed using the JENA API for accessing RDF content.

Finally, the *Ingestion* subsystem is also in charge of the transformation of XHTML-RDFa documents into RDF triples for its storage in a triple store (which can be later queried by the *Query* subsystem). This transformation is made thanks to the RDF Translator tool (Stolz, 2014), which is accessible online as a Web service. Figure 8 shows an example of the conversion to RDF of the annotated document in figure 4 (and its graph representation in figure 5).

At this point it must be noted that both XHTML-RDFa annotations and their conversion into RDF maintain a reference to the original document identifier using the *schema:url* property. This facilitates the integration of the input of different annotation tools at the moment of storing them in the same triple-store. Thanks to this, it is possible to query later for place names that have been identified by different tools on the same document.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:schema="http://schema.org/"
>
  <rdf:Description rdf:nodeID="N6d2e01594db44edf9751591df57c3939">
    <schema:location rdf:nodeID="N6c1e82d03f9d48c1b5cf0db3c918cfc2"/>
    <rdf:type rdf:resource="http://schema.org/TravelAction"/>
    <schema:url>Near_the_Puerta_del_Sol4</schema:url>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N6c1e82d03f9d48c1b5cf0db3c918cfc2">
    <rdf:type rdf:resource="http://schema.org/Place"/>
    <schema:location rdf:nodeID="N351fee839c1247b49be2dec603417deb"/>
    <schema:name>Madrid</schema:name>
    <schema:geo rdf:nodeID="N1bcc407fec644c2088ccd2610148ffe8"/>
    <rdfs:comment>OpenCalais</rdfs:comment>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N1bcc407fec644c2088ccd2610148ffe8">
    <schema:longitude>-3.6833</schema:longitude>
    <schema:latitude>40.4</schema:latitude>
    <rdf:type rdf:resource="http://schema.org/GeoCoordinates"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N351fee839c1247b49be2dec603417deb">
    <schema:name>Spain</schema:name>
    <rdf:type rdf:resource="http://schema.org/Country"/>
  </rdf:Description>
</rdf:RDF>
```

Figure 8. Example of RDF triples stored in the semantic document repository

Figure 9. User interface of the Web application

4.2. Web application

The Web application prototype for interacting with the document repository has been developed with AngularJS, Bootstrap and Spring MVC technologies. On the one hand, AngularJS (Google, 2016) is an open source project, based on Javascript, which contains a set of libraries to facilitate the development of front-ends in web applications. On the other hand, Bootstrap (2016) is an Open Source framework that provides Cascading Style Sheets (CSS) to define the layout of HTML or XML documents (including also XHTML documents). Last, Spring MVC (Pivotal Software, 2016) facilitates the connection between the web

application front-end and the business logic of the *Ingestion* and *Query* subsystems through HTTP GET and POST requests.

Figure 9, in the previous page, shows an overview of the screenshots provided by the web application: an interface for the ingestion of documents or direct typed text (left side of figure 9), and an interface for querying documents in the repository and browse the annotation and toponyms found in each document.

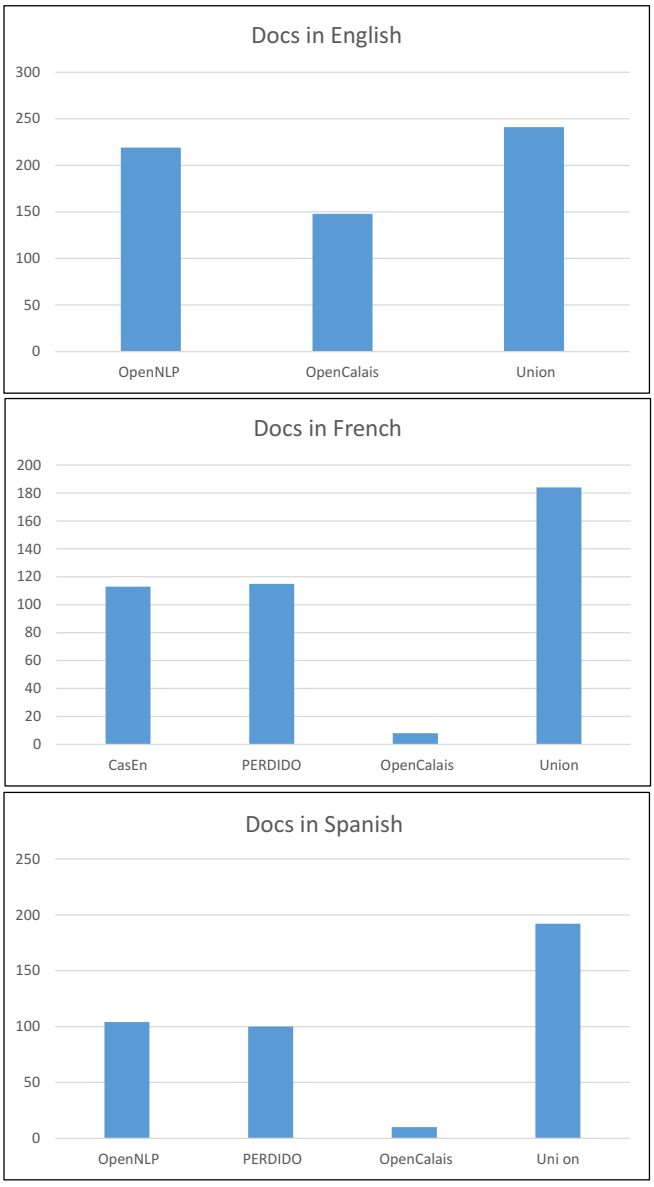

Figure 10. Number of toponyms detected as a result of the ingestion process

5. Validation of the approach

In order to validate the infrastructure proposed for a document repository we have uploaded a corpus of documents consisting of: 32 documents in English selected from a corpus created for the testing of spatial queries in an information retrieval conference (SemEval, 2015); 20 documents in French selected from a Web site of hiking descriptions (Visorando, 2016); and 28 documents in Spanish selected from another Web site with hiking descriptions (Gobierno de Aragón, 2016) (1). Figure 10 shows three bar charts with the results obtained by the ingestion process in each language. Each bar chart shows the number of spatial entities detected by each annotation tool and a final bar with the number of spatial entities detected by all tools excluding duplicates.

Although the results obtained depend greatly on the features of the corpus and the granularity of place names recognized by each tool, it seems that a document repository able to integrate different existing annotation tools provides an exhaustive annotation. In the case of the English experiment, the documents describe cities and countries all over the world. Thus, OpenNLP and OpenCalais, the tools available for processing English texts and focused on the detection of administrative units like cities or countries, provide similar results. In the case of French and Spanish experiments, the documents contain fine-grain toponyms and small to medium size names of administrative units. In these cases, we can see that the annotation provided by PERDIDO annotation tool, and focused on fine-grained toponyms, is an excellent complement to the annotations provided by CasEN or OpenNLP, which are more focused on the detection of names of small/medium administrative units. Additionally, it can be observed that the use of OpenCalais for French and Spanish documents provides little help as there are a very small number of city and country names.

6. Conclusions and future work

This work has described an approach to transform the heterogeneous output of different NER tools into a common annotation framework based on the use of RDFa and the vocabulary of Schema.org. In principle, this annotation approach should increase the visibility and recoverability of spatial annotated texts since big web search engine companies have agreed in the use of Schema.org as a de-facto standard for semantic annotation on the Web. There are other more specific application level ontologies for defining spatial concepts such as the ones referred by Lopez-Pellicer et al (2012), but our purpose was to select a common core set of concepts accepted by different communities adopting Semantic Web technologies. Other well-known Web content authoring tools like RDFace (Khalili and Auer,

2015) have also adopted Schema.org for their annotations.

Additionally, this work has also presented the design of the infrastructure for a semantic document repository with a Web interface that facilitates the ingestion of documents and their spatial annotation with the different external NER tools, and the spatial searching of documents based on the annotated spatial entities.

As future lines of this work we plan to study and integrate other existing annotation tools, establishing the mapping to our annotation proposal. Additionally, we would like to improve our annotation approach to take into account other types of named entities like Persons, Organizations or Dates. Furthermore, instead of considering each annotated document as an instance of a *TravelAction*, we would like to provide a finer heuristics to annotate the general intent of the annotated document.

Finally, our goal is also to extend the web application to facilitate the interaction of final users in different aspects: facilitate the ingestion of documents in a wider range of input formats (e.g., PDF, Word, ODF); provide an analytical comparison of the results produced by the different annotation tools; and facilitate the possibility of making a manual revision of these annotations.

Acknowledgements

This work has been partially funded by Universidad de Zaragoza through Catedra Logisman on Technological Document Management and UZ2016-TEC-05 research project.

Notes

(1) The documents in Spanish and French form part of the PERDIDO corpus: a collection of documents that are representative of the type of documents available on Web sites of hiking descriptions and have been already manually annotated for the purpose of testing geoparsing algorithms. More details about the corpus can be found at (Moncla, 2015b).

References

Apache (2010). The Apache Software Foundation. Apache OpenNLP web site. https://opennlp.apache.org/ (2016-04-28)

Bootstrap (2016). Bootstrap web site. http://getbootstrap.com/ (2016-04-28)

Cyganiak, R., Wood, D., Lanthaler, M. (eds.) (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (2016-06-14)

Doan, A., Halevy, A., Zachary, I. (2012). Principles of Data Integration (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 9780124160446

Gobierno de Aragón (2016). Turismo de Aragón. Senderos de Aragón. Excursiones y rutas para GPS. http://senderos.turismodearagon.com (2016-04-28)

Google (2016). AngularJS API Docs. https://docs.angularjs.org/api (2016-04-28)

Harris, S., Seaborne, A. (eds.) (2013). SPARQL 1.1 Query Language. W3C Recommendation 21 March 2013. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/ (2016-06-14)

Herman, I., Adida, B., Sporny, M., Birbeck, M. (eds.) (2015). RDFa 1.1 Primer - Third Edition. Rich Structured Data Markup for Web Documents. W3C Working Group Note 17 March 2015https://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/ (2016-06-14)

Khalili, A., Auer, S. (2015). WYSIWYM–Integrated visualization, exploration and authoring of semantically enriched un-structured content. Semantic Web, 6(3), 259-275.

Leidner, J. L. (2008). Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal-Publishers

Leidner, J. L.; Lieberman, M. D. (2011). Detecting geographical references in the form of placenames and associated spatial natural language. // SIGSPATIAL Special 3(2), 5-11.

López-Pellicer, F. J.; Lacasta, J.; Florczyk, A.; Nogueras-Iso, J.; Zarazaga-Soria, F.J. (2012). An Ontology for the representation of Spatio-Temporal Jurisdictional Domains in Information Retrieval Systems. // Int. Journal of Geographical Information Science. 26:4, 579-597.

Maurel, D.; Friburger, N. (2004). Finite-state transducer cascades to extract named entities in texts. Theoretical Computer Science. 313:1, 93-104.

Moncla, L. (2015a). PERDIDO REST API Specifications. http://erig.univ-pau.fr/PERDIDO/api.jsp (2016-04-28)

Moncla, L. (2015b). Automatic reconstruction of itineraries from descriptive texts. PhD Dissertation, Université de Pau et des Pays de l'Adour & Universidad de Zaragoza. https://zaguan.unizar.es/record/47425/ (2016-06-16)

Moncla, L.; Gaio, M.; Mustière, S. (2014). Automatic Itinerary Reconstruction from Texts. Proc. 8th International Conference, GIScience 2014, 253-267.

Pivotal Software (2016). Spring Framework. http://projects.spring.io/spring-framework/ (2016-04-28)

Santos, J.; Anastacio, I.; Martins, B (2014). Using machine learning methods for disambiguating place references in textual documents. // GeoJournal. 80:3, 375-392.

Sekine, S.; Ranchhod, E. (eds.) (2009). Named Entities: Recognition, Classification and Use. John Benjamins. ISBN 9789027289223.

SemEval (2015). International Workshop on Semantic Evaluation (SemEval-2015): Task 8-SpaceEval Test Data. http://alt.qcri.org/semeval2015/task8/index.php?id=data-and-tools (2016-04-28).

Stolz, A. (2014). RDF Translator. http://rdf-translator.appspot.com/ (2016-04-28).

Thomson Reuters (2016). Open Calais web site. http://www.opencalais.com/ (2016-04-28).

University of Tours (2016) CasEN web site. http://tln.li.univ-tours.fr/Tln_CasEN.html (2016-04-28).

Visorando (2016). Visorando web site. http://www.visorando.com (2016-04-28).

Enviado: 2016-0-28. Segunda versión: 2016-0-.
Aceptado: 2016-0-.