Geographical Localization of Web Domains and Organization Addresses Recognition by Employing Natural Language Processing, Pattern Matching and Clustering

Paolo Nesi, Gianni Pantaleo and Marco Tenti

DISIT Lab, Department of Information Engineering (DINFO), University of Florence, Via S. Marta 3, 50139 Firenze, Italy.

Abstract

Nowadays, the World Wide Web is growing at increasing rate and speed, and consequently the online available resources populating Internet represent a large source of knowledge for various business and research interests. For instance, over the past years, increasing attention has been focused on retrieving information related to geographical location of places and entities, which is largely contained in web pages and documents. However, such resources are represented in a wide variety of generally unstructured formats, and this actually does not help final users to find desired information items. The automatic annotation and comprehension of toponyms, location names and addresses (at different resolution and granularity levels) can deliver significant benefits for the whole web community by improving search engines filtering capabilities and intelligent data mining systems. The present paper addresses the problem of gathering geographical information from unstructured text in web pages and documents. In the specific, the proposed method aims at extracting geographical location (at street number resolution) of commercial companies and services, by annotating geo-related information from their web domains. The annotation process is based on Natural Language Processing (NLP) techniques for text comprehension, and relies on Pattern

Email address: paolo.nesi@unifi.it, gianni.pantaleo@unifi.it (Paolo Nesi, Gianni Pantaleo and Marco Tenti)

URL: http://www.disit.dinfo.unifi.it (Paolo Nesi, Gianni Pantaleo and Marco Tenti)

Preprint submitted to Engineering Applications of Artificial IntelligenceNovember 20, 2015

Matching and Hierarchical Cluster Analysis for recognizing and disambiguating geographical entities. Geotagging performances have been assessed by evaluating Precision, Recall and F-Measure of the proposed system output (represented in form of semantic RDF triples) against both a geo-annotated reference database and a semantic Smart City repository.

Keywords:
Geographic Information Retrieval, Geoparsing, Geocoding, Web crawling, Data Mining, Semantic Web, Natural Language Processing, Pattern Matching, Hierarchical Clustering.

---

1. Introduction

In March 2015 the number of online active Web sites had been estimated in about 900 million (878 million, as reported by Netcraft[1], 903 million according to the CIA World Factbook[2], 930 million as stated by Internet Live Stats[3]). As a global leader in domain names and internet security, Verisign[4] periodically reviews the state of web domain name industry, reporting that the fourth quarter of 2014 ended with a total of 288 million Top-Level domain name registrations (TPDs), with an increase of 1.3 percent (about 4 million domain names) over the third quarter of 2014.

Such a huge amount of web resources represents an extremely vast source of knowledge, for the most part embedded in the textual content of web pages and documents (in the following, the term Web page will be used to describe every web resource identified by its own unique URL, containing a textual content that can be navigated and parsed, while the term Web document will be used to identify a larger variety of web available text files - such as .doc, .rtf, .pdf etc. - whose textual content can be downloaded and parsed). However, it is becoming increasingly difficult, for final users, to extract specific information items of interest, since web resources are not yet fully structured in a machine-readable format. The majority of web pages and documents are still a collection of unstructured text formats without any explicit meaning automatically inferable by machines (Schmidt et al., 2013a). An early

[1] http://news.netcraft.com/archives/category/web-server-survey/
[2] https://www.cia.gov/library/publications/the-world-factbook/
[3] http://www.internetlivestats.com/total-number-of-websites/
[4] http://www.verisigninc.com/en_US/innovation/dnib/index.xhtml

attempt for adding structure to HTML pages were Microformats[5]. Microformats, developed as part of the HTML5 standardization efforts, defines fixed vocabularies to annotate specific entities such as people, relationships, organizations and places, calendar entries, products, cooking recipes etc. within HTML pages (Loglisci et al., 2012). A more recent approach is represented by the Semantic Web applications: an increasing number of HTML pages embed structured XML/RDF data and schemas, according to the RDFa format (Bizer et al., 2008); besides, dedicated Ontologies and Taxonomies have been developed, such as the GoodRelations Ontology, which semantically models and describes details of business services and products in a fully integrated way with the schema.org markup vocabulary, used by the most important search engines such as Google, Yahoo!, Bing/Microsoft, and Yandex mobile applications (Hepp, 2013). Recently, a growing interest has arisen in the field of semantic data enrichment, since it covers the representational needs and interoperability requirements within the expanding e-commerce scenarios. However, since all these represent emerging standards, the vast majority of actual available online resources do not support yet such new reference benchmarks. Therefore, strong interests and needs are perceived to extract structured information in web pages, in a large variety of fields and application areas. Until the semantic enrichment of web content will not reach a significant degree of penetration, automatic annotation of information items represents an alternative to manual annotation, which is an extremely inefficient and time-consuming process.

This paper focuses on the annotation of geographic information from unstructured text in web pages and documents, with the aim of extracting geographical location of commercial entities. This process is defined as Geographic Information Retrieval (GIR), as it concerns the extraction of information involving some kind of geo-spatial reference (Mandl et al., 2008). Understanding place names mentioned in textual data can provide great benefits for data mining systems and search engines, enhancing the capabilities of geographic-based queries and filtering. From past studies emerged that 15% of the queries submitted to search engines contain references to geographic names (Anderson and Kohler, 2004). Analyzing some quite recent reports, in May 2011, 23% of USA citizens used location-based services. This number increased to 41% in February 2012, and it can be assumed that such

[5]http://microformats.org/

a trend is still growing (Zickuhur, 2012). Automatically retrieving address information about business entities, associations and organizations is an aspect that attracts a lot of commercial and research interest, ranging from geo-marketing to criminal investigation and fraud detection. Location-based applications can take advantage from automated harvesting of address data from Web sites, for instances recommendation-based systems can provide spatial-based suggestions on the surrounding of a user. An important application area for geographical annotation technology is found in the recently developing Smart City frameworks, aiming at helping citizens by providing different services and useful information on public available Open Data (OD), including geographical information and spatial location of places of interest, real-time traffic and parking structures, as well as any other kind of municipality resource which can be geolocated. This kind of services requires a very high spatial resolution (generally at street number level), although this is still far to be achieved by present GIR systems with a sufficient degree of confidence and reliability. The solution to this issue is not trivial, and it concerns not only with the improvement of the current GIR tools and services, but also with a desirable standardization of OD formats, descriptors and languages. Smart City services are often based upon unstructured Open Data representing public administration services and utilities which are not always geographically referenced. For this purpose, and moreover to cover commercial stakeholders which are not included in public OD, additional information is needed. Such a requirement is met by the presence of many online web domains that are usually representative of activities undertaken by business organizations. Consequently, the necessity arises to associate commercial web domains to geographic information related to their corresponding physical entities.

The main objective of this work is to present a system for extracting administrative information, including addresses and geographical coordinates, of web-visible human activities (intended, in the most generic meaning, all those human activities - ranging from commercial and research organizations, private companies and services, Public Administration services etc. - which are associated to a public available web domain) from unstructured text contents hosted in their web sites. The main areas and aspects of Engineering Applications of Artificial Intelligence addressed in the paper are Data Mining and NLP, in the specific: Text Mining and Annotation, Named Entity Recognition, Part-of-Speech (POS) tagging. Evaluation experiments have been conducted focusing on mining web domains and pages owned by

organizations located in the Tuscany region, Italy. Actually, according to the Italian National Institute of Statistics (ISTAT), Tuscany is one of the Italian regions with the highest number of firms and companies (reported as a total of 330 thousands registered companies in the last 2011 census, which is equivalent to nearly 80 firms per thousand citizens). However, only a small percentage of such commercial operators is estimated to be actively registered to the Open Data repositories provided by regional Public Administrations, as well as to external resources (e.g.: the GoodRelations users community). The proposed framework aims at filling this gap. The present paper is organized as follows: Section II illustrates the related work, in terms of state of the art and research issues for both commercial and research literature; in Section III, the functional architecture of the proposed system is presented; in Section IV, a two-steps validation of the system is reported (in order to assess both address information and geographic coordinates extraction); finally, Section V is left for conclusions and future perspectives.

2. Related Work

The processes of recognizing geographic context and assigning spatial coordinates are commonly referred to as *geoparsing* and *geocoding*, respectively (Scharl, 2007). Geoparsing deals with parsing unstructured text and extracting keywords and keyphrases describing geographical references, including the extraction of terms and/or metadata representing physical, natural and human-made features (e.g.: countries, cities, roads, addresses, postal codes, telephone numbers, buildings etc., but also forests, rivers, lakes, mountains etc.). In its most general meaning, geoparsing refers to the extraction of toponyms in texts (Pouliquen et al., 2004), and it is related to Natural Language Processing (NLP) and connected tasks such as Named Entity Recognition (NER, addressing the detection of general named entities) and Word Sense Disambiguation (WSD, aiming at resolving ambiguities occurring in presence of polysemous words and expressions). Geocoding, on the other hand, is defined as the process of mapping geographical annotations to their real-world counterparts by associating spatial coordinates (latitude, longitude and, in case, altitude).

2.1. State of the Art

Generally, geoparsing systems based on NLP and NER techniques can be classified into the two following main classes (depending on the processing

approach):

- Internal approaches. In this case, a text is analyzed and named entities are extracted without using external resources. The extraction methods can be based on the joint application of both Pattern Matching (in which explicit patterns are defined to extract specific information, relying also on the context) and Statistical based solutions (Schmidt et al., 2013b). A pattern-based approach for address extraction is presented by Asadi et al. (Asadi et al., 2008), in which manually chosen patterns are exploited for recognition of addresses, providing also different confidence scores. Both HTML and visual-based segmentations are used to increase the quality of address extraction. Other frequently used internal approaches are based on Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), and Maximum Entropy Models. NLP based solutions, such as POS tagging, commonly make use of internal approaches, and they usually achieve good results providing that they are properly optimized for the specific language (Tjong Kim Sang and De Meulder, 2003).

- External approaches. Here an external resource is used, usually in the form of an annotated list (called gazetteer) or database. Gazetteers can contain information on geographical references (e.g. toponyms, name variations, classes, sizes etc.), and have the advantage of being generally language-independent, provided that they contain also language-specific spellings, characters sets, as well as suitable alternatives for name variations. Other external sources employed are training data sets for GIR systems based on statistical techniques, which learn from manually annotated corpora. The efforts required to set up training data sets are balanced by a general improvement of extracting performances, especially in noisy environments.

Typically, both these solutions are often combined into hybrid approaches. Additional knowledge coming from external gazetteers is still considered as a significant help, as their availability is expanding since today we are surrounded by an Open Data populated environment and many global resources are present, such as *Geonames*, *OpenGeoData*, *OpenStreetMap* and several others (Moncla et al., 2014). Schmidt et al. (Schmidt et al., 2013a) have focused their work on extracting address information for German streets employing POS-tagging based techniques as a pre-processing phase for input

web pages. Then, single address attributes are extracted (at street number resolution) using a gazetteer of German cities and locations taken by querying the OpenStreetMap web service. Their solution achieves Precision = 0.656, Recall = 0.833 and F-Measure = 0.734 in complete address evaluation. The method proposed by Cai, Wang and Jiang (Cai et al., 2005) relies on patterns, but exploits also external databases, in order to determine the similarity between parsed text segments and reference patterns; if the similarity score exceeds a defined threshold, that specific text segment is considered as an address. Evaluation results show F-Measure = 0.734 (Precision = 0.745 and Recall = 0.724) for extracting addresses from Web sites of Yellowpages and Yahoo! Business finder. Yu (Yu, 2007) proposes a hybrid method combining pattern-based and machine learning techniques. The system segments web pages into tokens from which output addresses are extracted. Experimental evaluation on a dataset of 471 web pages containing 2257 labeled addresses reports F-score = 0.876 (Precision = 0.952 and Recall = 0.811). Also the work presented by Ahlers and Boll (Ahlers and Boll, 2008) relies on existing databases containing all possible combinations of street and city names, postal codes and other features for the identification of addresses. Loos and Biemann present a statistical approach using Conditional Random Fields (CRF) (Loos and Biemann, 2008); by applying this model, it is possible to perform tokens classification based not only on the direct textual context of processed item, but also on a wider, more general context obtained as a result for other, previously processed tokens. In their approach, they use both a small training set consisting of 400 Web sites manually annotated with address information, and a larger, not annotated data set. Evaluation results achieved are F-Measure = 0.83 (Precision = 0.918 and Recall = 0.73) on a dataset composed by 180.000 web pages and about 25.000 addresses matching their definition of full address. Chang and Li (Chang and Li, 2010) also jointly employ CRF to train models for geo-related information retrieval in their MapMarker tool, a NLP based framework for postal addresses extraction. Evaluation results yield F-Measure = 0.914 (Precision = 0.968 and Recall = 0.866) on a dataset containing 1205 pages with single address and 535 pages with multiple addresses (for a total 8519 postal addresses).

With the fast recent development of social media (blogs, forum, feddbacks and rating resources etc.) and social networks, the interest and need of geographical information extraction has been focused also in this area. Zhang and Gelernter (Zhang and Gelernter, 2014) proposed a system which annotates geographical locations in Twitter text messages, using supervised ma-

chine learning algorithm to weigh the different fields of Twitter messages and a dynamic gazetteer model. The system best performances report F-measure = 0.852 (Precision = 0.85 and Recall = 0.854) upon a set of 300 tweets for training and 100 for testing (on which toponyms and geographical coordinates were manually annotated for reference). Evaluation results and a short detail of the employed technologies for the above cited frameworks are shown in Table 1.

Table 1: Comparative Table of cited tools reviewed in literature (chronologically ordered starting from the most recent).

Reference	**Technologies Used**	**Precision**	**Recall**	**F-measure**
(Zhang and Gelernter, 2014)	Machine Learning, Gazetteers	85.0%	85.4%	85.2%
(Schmidt et al., 2013a)	NLP (POS-tagging), Gazetteers	65.6%	83.3%	73.4%
(Chang and Li, 2010)	NLP, Condition Random Fields	96.8%	86.6%	91.4%
(Loos and Biemann, 2008)	Condition Random Fields	91.8%	73.0%	83.0%
(Yu, 2007)	Pattern Matching, Machine Learning	95.2%	81.1%	87.6%
(Cai et al., 2005)	Pattern Matching, External DB	74.5%	72.4%	73.4%

There are also several commercial products with geoparsing capabilities, for instance: MetaCarta[6] extracts information about place and time, while others like Digital Reasoning GeoLocator 2.0[7], AeroText[8] (produced by Rocket Software, formerly developed by Lockheed Martin) and NetOwl[9] (by SRA) extract places and annotate them, keeping relations with other entities such as persons, organizations, temporal information, etc. (Abascal-Mena and López-Ornelas, 2010). Yahoo! Placemaker[10] is a geotagging web service that offers the possibility to enrich web applications and sites with geographic information. The service is able to identify, disambiguate, and extract place names from unstructured and semi-structured documents, providing as output a unique *Where on Earth Identifiers* (WOEIDs) for each extracted toponym.

[6]http://www.metacarta.com/
[7]http://www.digitalreasoning.com/buzz/650647
[8]http://www.rocketsoftware.com/products/rocket-aerotext
[9]http://www.netowl.com/entity-extraction/
[10]http://developer.yahoo.com/geo/placemaker/

Many free available web services and commercial applications also provide IP geolocation features, such as IP Geolocator[11], Geo IP Address View[12], IP Location[13], GeoBytes IP Locator[14], SEO Mastering[15]. Also commercial applications have been developed for IP geolocation, such as IP2Location[16] and Maxmind GeoIP2[17]. However, tools based on this solution often don't provide an adequate level of accuracy, which is nowadays needed as a fundamental requirement in many application areas. Actually, they generally retrieves host IP locations, and this often represents not accurate or even erroneous information, since IP addresses does not correspond exactly to geographic locations: IP geocoding information is associated with the location of the Internet Service Provider, and not with the exact location of the legal entity that owns or is responsible for a certain web-domain or host. In addition, different web domains and their hosts are commonly assigned to a single, centralized Provider.

Regarding the conceptual difference between IP-based geocoding and geographical info based geocoding, the former basically refers to the *Source geography*, which represents the physical location of the host owner's server (i.e.: the origin of the page), while the latter is related to the *Target geography* is determined by the geographical information related to web page contents (geographical information on location of commercial utilities or services).

After this investigation conducted on the current state-of-the- art literature, some considerations can be made. The whole georelated information extraction process can be conceptually divided into different approaches, as depicted in Figure 1:

a) Approach based on IP host and whois-service geocoding applications, which directly extract geographic coordinates from web pages host IPs.
b) Approach based on systems employing NLP techniques (including internal, external and hybrid solutions based on NER, WSD, pattern-matching, statistical methods and external knowledge). In this context, we usually deal with two different phases:

[11]http://www.ipligence.com/geolocation
[12]http://www.geoipview.com.ipaddress.com/
[13]http://www.iplocation.net/
[14]http://www.geobytes.com/IpLocator.htm
[15]http://www.seomastering.com/ip-host-name-locator.php
[16]http://www.ip2location.com/
[17]http://www.geoiptool.com/

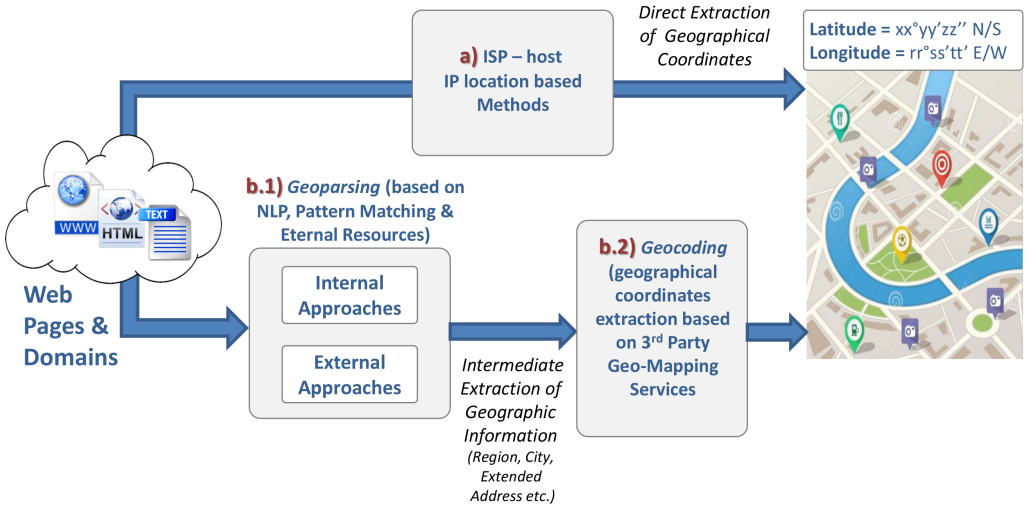

Figure 1: Overview of the different approaches and processes typically followed by a GIR system.

b.1) *Geoparsing*, leading to a first extraction of geolocation related information.

b.2) *Geocoding*, aiming at retrieving geographic coordinates by querying proprietary or third parties geomapping services.

2.2. Research Issues and Innovation

Present state-of-the-art systems for geographical information harvesting are still far from being perfect. Actually, identifying and disambiguating place names in text are difficult tasks, since they are language dependent and prone to errors, inconsistencies and ambiguities that have to be resolved (Zhang et al., 2012). Such ambiguities are commonly grouped into two categories: *geo/non-geo* and *geo/geo*.

A *geo/non-geo* inconsistency occurs when a toponym shares its lexical meaning also with non-geographical entities, such as person names and other common words (Amitay et al., 2004). Another example of geo/non-geo ambiguity is represented by those place names which are homonymic with names of persons or organizations in different languages: as an example, some of the most common words in the English, such as *And*, *To*, *Be* and *By*, represent also different locations in the world (located in Ireland, Ghana, India and Sweden, respectively) (Kimler, 2004).

On the other hand, *geo/geo* ambiguities arise when distinct places have the same name. As an example, there are over 2,100 toponym labels exactly matching the name "San Antonio" in the USA National Geospatial Intelligence Agency database (Loos and Biemann, 2008). Therefore, resolving geo/geo ambiguities requires toponym resolution, which deals with assigning names of places to the physical locations they actually refer to in the analyzed documents, as well as identifying the geographical scope resolution of the documents.

A lot of work has been made in the direction of toponyms disambiguation. According to Buscaldi (Buscaldi, 2011) different approaches can be identified, classified in three categories: map-based, knowledge-based, and supervised (or data-driven) approaches. Map-based methods use an explicit representation of toponyms, for instance a map or geographical coordinates list, in order to calculate the average distance of the unambiguous toponyms contained in the same document (context toponyms) from all the possible homologous candidates; the candidate with the shortest average distance among the contextual toponyms is selected. Kwoledge based methods exploit external sources such as gazetteers or semantic knowledge bases to find disambiguation clues. Data-driven methods apply machine learning techniques to extract and learn geographical references and relations on dataset by exploiting also non-geographical content. Recently, Zhao et al. (Zhao et al., 2014) have proposed a Geo-Rank algorithm inspired by Google Page Rank algorithm for disambiguating toponyms in Web resources. About these aspects, the goal of building smarter and more efficient machine readable models, in order to improve the quality of automatic extraction of high level knowledge and features, is becoming a matter of pivotal importance. Many efforts have been recently addressed to this direction, thanks to the development of Semantic technologies in the form of web Ontologies, Thesauri, general Knowledge Organization Systems (KOS), Taxonomies, Inference mechanisms and Reasoning engines (Bellandi et al., 2012).

3. Architecture of the Proposed System

The system proposed in this article is an extension of the one presented in (Nesi et al., 2014). New parts and functionalities have been added, such as the RDF Mapping Module (described in Section 3.4). Moreover, the Geoparsing module (described in Section 3.2) has been extended by implementing a Hierarchical Clustering technique in order to support a more robust deci-

sion process for the Entyity Info Arrays (EIAs) at domain level. The proposed system has been designed and developed with the aim of extracting geographical information related to location of web domains associated to entities such as commercial companies, business services, research institutes, and generally web-visible human activities located in the Tuscany region, in Italy.

The extraction of geographical information is based on a hybrid approach: both internal (NLP solutions, linguistic parsing, POS-tagging, pattern-matching based annotations) and external (use of external annotated gazetteers containing names of Italian regions, provinces and cities) approaches are adopted. The architecture of the system, as shown in Figure 2, is modular, and it is composed of four main functional blocks:

- A *Distributed Web crawler*, which represents a scalable solution for mining, parsing and fetching Big Textual Data and huge amounts of documents. Its architecture is illustrated in Section 3.1.

- A *Geoparsing Module*, based on a NLP-based linguistic analyzer which annotates input texts using custom syntactical and grammatical rules, as well as external gazetteers and pattern matching techniques. For each processed web page, one or more sparsely populated records, containing administrative and geographical information of the entity addressed in the same domain, are extracted. All these records will contribute to form the final outcome of the Geoparsing Module, which will be referred as *Entity Info Arrays* (EIA) in the following. A dedicated *Entity Information Array Estimation* module is in charge of estimating final output by applying a hierarchical clustering algorithm among the EIA candidates, based on the Levenshtein string distance. The Geoparsing Module will be described in deeper detail in section 3.2.

- A *Geocoding Module* retrieves the coordinates of extracted EIAs by querying a semantic Smart City repository, built upon the *Km4City* ontology[18] (Bellini et al., 2014a) created at DISIT Lab within the Sii-Mobility Project[19], and powered by the *Linked Open Graph*[20] visualization and browsing tool for multiple SPARQL end-points and Linked

[18]http://www.disit.org/km4city/schema
[19]http://www.disit.org/5478
[20]http://log.disit.org/service/

Open Data (Bellini et al., 2014b). The *Km4City Ontology* combines Linked Open Data, street graph data and Public Administration services, for instance real-time traffic and parking information provided by the Municipalities in the Tuscany region, tourism facilities etc. The geocoding module is described in Section 3.3.

- A *RDF Mapping Module* finally converts the EIAs information gathered from the geoparsing process and the geographical coordinates collected during the geocoding phase, into RDF triples describing the analyzed entities in terms of administrative and geographical elements. The *Km4City* Ontology, as well as external ontologies such as GoodRelations, Schema.org and Basic Geo WGS84, are used for semantic mapping. In addition, output RDF triples are later stored and indexed in the *Km4City* Ontology each time the whole extraction process is launched and correctly finalized. The functionalities of the RDF Mapping module are outlined in Section 3.4.

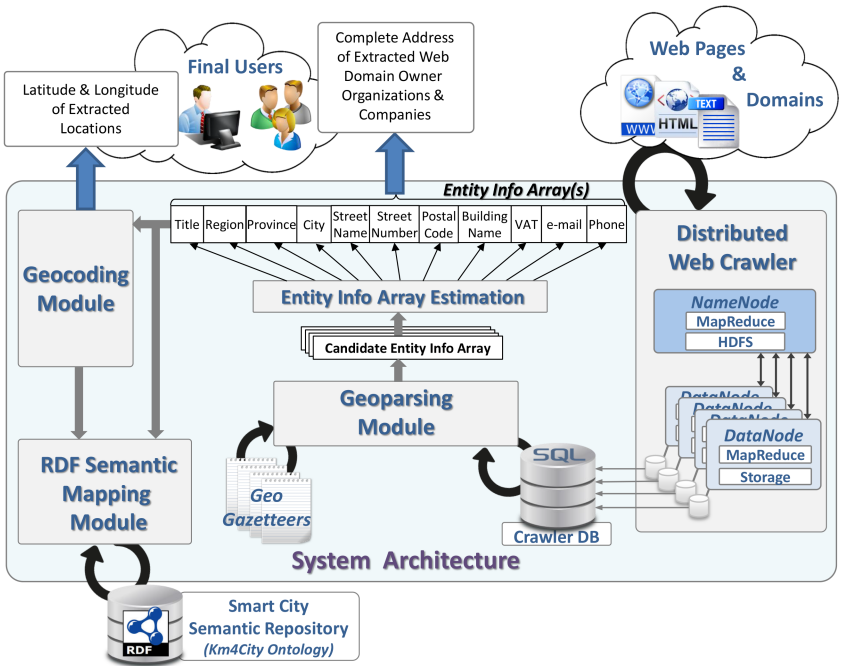

Figure 2: Functional block-Architecture of the proposed system.

3.1. Distributed Web Crawler

The crawling engine of the proposed system is based on the open source Apache Nutch[21] crawling tool, which has been integrated with Apache Solr[22] for document indexing, and initialized with a set of seed URLs of commercial companies and services operating in the Tuscany area. The URLs used to build the seed list have been taken from public available database providing administrative information about business organizations, indexed also by administrative regions. Since the proposed system is designed to process textual content of web pages and documents, the URLs to be processed are selected among all the crawled URLs by a simple filtering procedure, which discards the ones pointing to certain kinds of file formats, such as multimedia, executable files, archives and databases. As a qualitative evaluation, we found that the majority of crawled web URLs are associated to domains registered to private or public entities which can be geolocated. The remaining cases are eventually included in the handling of False Negatives count during the evaluation phase. The big amount of web resource to be crawled led to the implementing an efficient and scalable solution, in order to improve performances, as well as data integrity and failures handling. Apache Hadoop[23] has been chosen to distribute and parallelize the crawler architecture. Data access and management is based on the Hadoop Distributed File System (HDFS) mounted on a commodity hardware cluster. The cluster is composed by a master NameNode, which assigns scheduled crawling tasks to the different clients (DataNodes), which are in charge also of data storage. The MapReduce programming paradigm is used to parallelize the crawling work and database handling operations submitted by a node. The map function segments the work into smaller jobs or file blocks, which are subsequently mapped among the cluster DataNodes. Specific Map functions are defined to generate key/value pairs representing logical records from the input data source. Subsequently, Reduce tasks merge all intermediate values associated with the same key. At the end of the distributed crawling process, the collected URLs are stored in the HDFS filesystem. Later, their online content is parsed through scheduled processes. The gathered documents are then indexed by Solr and fetched into a unified SQL DataBase.

[21] http://nutch.apache.org/
[22] http://lucene.apache.org/solr/
[23] http://hadoop.apache.org/

3.2. Geoparsing Module

The Geoparsing Module relies on a linguistic parser which takes as input the content extracted from documents and pages previously collected by the Distributed Web Crawler Module. Input text is analyzed by means of NLP-based techniques, linguistic rules, POS-tagging, as well as through the use of external gazetteers providing names of Italian cities, regions, companies, abbreviations for address items and identifiers. The present module implementation is based on the open source GATE framework (A General Architecture for Text Engineering (Cunningham et al., 2002)), particularly exploiting some of its plugins for Information Extraction (ANNIE) and pattern recognition extraction (JAPE). After the input text has been segmented into morphologically and syntactically annotated tokens (annotating also the HTML markups), the extraction algorithm basically starts searching geographical information (potentially related to the analyzed domain) in the HTML footer or head of each processed web page. Actually, these HTML tags are often used to contain and display administrative and location information of the legal entity represented in the same domain. If no useful information is retrieved neither in the footer nor in the head, the system continues searching in the rest of the page. Geographical entities such as Region, Province and City are annotated relying on external Gazetteers. The main difficulties in the extraction of complete geographical addresses (including building names and internal block information) arise for handling the variety of formats in which they can be represented. For this reason, in order to deal with such heterogeneity of formats, the following JAPE patterns have been defined to be matched in the unstructured text:

- General (high level detail) address pattern:

$$[REGION] + [PROVINCE] + [POSTAL\_CODE] + [CITY].$$

- Specific (low level detail) address pattern:

$$[STREET\_IDENTIFIER] + [STREET\_NAME] +$$
$$+[STREET\_NUMBER].$$

- Pattern for retrieving address attributes in the form of internal block locations (in the common form usually encountered in Italian address forms, e.g.: *Scala A, Interno 4*):

$$[INNER\_BLOCK\_ID1] + [ID1\_VALUE] +$$

$$+[INNER\_BLOCK\_ID2] + [ID2\_VALUE].$$

- Pattern for extracting geographical coordinates, if they are present (in any case, this information is not inserted directly in the EIAs, but is stored separately, as an additional potential gold reference for geocoding evaluation):

$$[LATITUDE\_IDENTIFIER] + [LATITUDE\_VALUE] +$$

$$+[LONGITUDE\_IDENTIFIER] + [LONGITUDE\_VALUE].$$

  According to our training sets and evaluations, geographical coordinates are not so often provided by

Each field within square brackets in the above mentioned patterns represents further logic, rules, macros, gazetteer entries, as well as external lookup conditions, defined to optimize the overall quality of the extraction process. All the text segments of a same web page matching one or more of these patterns are annotated and later combined, with different priority levels. For each successfully parsed web page, a *Entity Info Array* (EIA) is created and populated with the extracted administrative and geographical annotations retrieved at different granularities: legal title and name, region, province, city, street or place name and number, postal code, building name/label, VAT, email and phone contacts.

The last issue in charge of the Geoparsing Module is to resolve which of the different extracted records have to be chosen as final EIAs at domain level. Actually, it is possible that a single web domain, representing a commercial company or service, may contain different locations, depending on the number of offices, headquarters, administrative/legal locations etc. For this purpose, a hierarchical agglomerative clustering procedure is performed among the different EIAs. The metric adopted as a criterion for combining clusters is the Levenshtein distance calculated on the strings obtained by merging city, street name, street number and postal code fields for each array, since these are considered mandatory items to retrieve geographical information. In fact, as a preliminary step, EIAs presenting null or empty values concurrently for city, street name, street number and postal code fields are discarded (in order to minimize false negative outcomes). A weighted linkage function has been used, based on the distance information previously generated, to determine the proximity among object pairs within the whole

set. The procedure is iterative: as EIAs are paired into binary clusters, the newly formed clusters are grouped into larger ones, until a hierarchical tree is built, usually represented as a *dendrogram*. Finally, in order to create the final data partition, a clustering function prunes all the tree branches whose mutual normalized distance is less than a defined threshold value; for our tests we have empirically chosen a normalized distance threshold equal to 60%, having reported the bets results in a preliminary training phase. Hierarchical clustering has been considered as a proper choice, since it does not require to know a priori the number of clusters. Within the obtained partition, clusters populated by a number of elements below a defined percentage of the whole set numerosity are pruned, in order to minimize false positive outcomes. Finally, for each partition, the estimated output EIA is obtained by selecting and assigning to each array field the most frequent value assumed among all the EIA candidates belonging to the same cluster. This is performed in order to improve the output records completeness. An example of EIAs estimation involving several web pages of a single web domain is illustrated as a dendrogram representation in Figure 3, based on sample EIA candidates shown in Table 2. As a remark, the numbering of EIA candidates shown in the example is ordered only for practicality's sake.

Table 2: Sample excerpt of Entity Information Array (EIA) candidates, extracted for a single domain and grouped by overlapping city, street name, street number and postal code fields.

EIA Candidates	**Street Name**	**Street Number**	**Postal Code**	**City**
EIA_1 ... EIA_5	Via Mascagni	17	-	Calenzano
EIA_6, EIA_7	Via Mascagni	17	-	-
EIA_8 ... EIA_9	Via Mascagni	-	50041	-
EIA_{10} ... EIA_{19}	Via del Saccardo	28	-	Calenzano
EIA_{20}, EIA_{21}	Via del Saccardo	28	-	-
EIA_{22}, EIA_{23}	strada	2520	-	-
EIA_{24}	Via dell'Arcovada	4	50100	Firenze

The present system provides multi-language capabilities, supporting Italian and English, and offers two different operating modalities: if the Language-dependent mode is activated, the system can be initialized with two different grammars (Italian and English) in order to exploit language-dependent syntactical and morphological rules for parsing and annotating Italian and En-

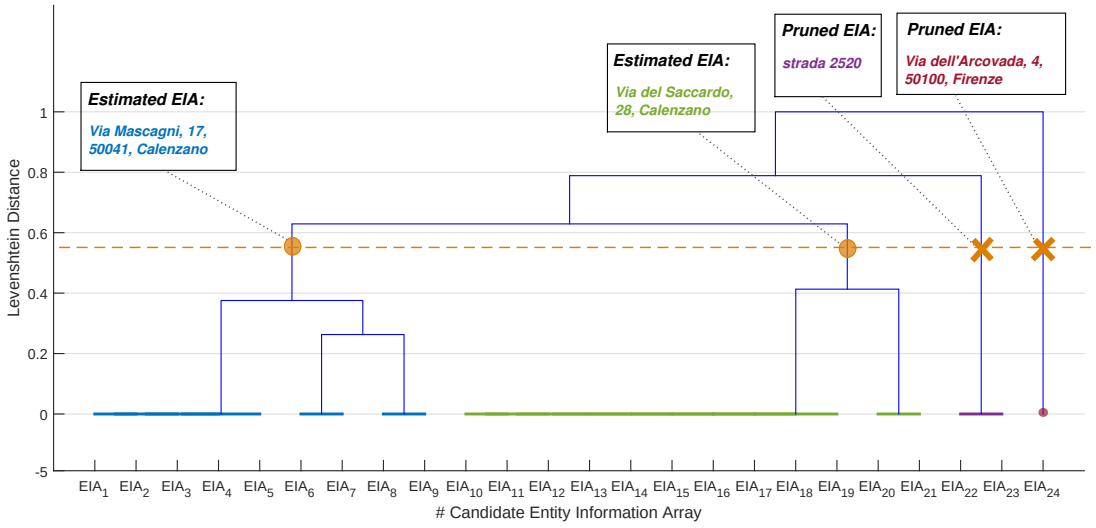

Figure 3: Example of Hierarchical Clustering, based on Levenshtein distance, for Entity Info Array (EIA) estimation related to EIA candidates presented in Table 2. The normalized threshold value discriminates the identification of three distinct clusters; two of them are pruned due to low numerosity of populating items (revealing actually as false positives).

glish text documents. The Language-independent mode, on the other hand, does not provide the use of such linguistic feature. A preliminary evaluation has been made, in order to assess if address extraction performances are significantly affected by switching between one of the two working modes. Results obtained against a small URL subset of the Crawler Database shows that little improvements are achieved in the Language-dependent mode, though they are not significant enough to justify the noticeable increasing of computational times due to running two different instances of the ANNIE pipeline for the linguistic parser, as well as several instances of the GATE LingPipe Language Identifier plugin.

3.3. The Geocoding Module

Once EIAs are obtained for each domain, the sparsely populated fields are used to build expanded queries submitted to the SPARQL endpoint of the *Km4City* repository, in order to extract geographic coordinates of retrieved locations, where it is possible. This choice was made in order not to

depend from third parties tools (e.g. OpenStreetMap, Google Maps, Google Geocoding APIs etc.), in case to allow performing comparative evaluations. Anyway, the proposed system can be easily adapted for working with any other geocoding repository or services, outside the considered domain.

Single queries are built providing different options and combinations (for instance, queries by company/service legal name, queries by province, city, address, which may include or not the building name, query by complete EIAs etc.) set to different priority levels. By this way, extracting also non-geographic features allows the system to potentially obtain correct coordinates even for domains with missing address or other fields. As an example, if we consider the web page at the following URL: http://www.apretoscana.org/portal/ (extracted from the URLs fetched by the crawler), we see that no typical geographic information is present in the page, neither in the footer nor in the head. However, in the footer it is possible to extract the name of the building (*Polo Scientifico di Sesto Fiorentino*) which hosts the organization, and actually it is correctly recognized by querying our *Km4City* repository, so that its geographical coordinates are properly returned. Finally, as an intermediate result, output latitude and longitude values are inserted into a SQL database, together with the corresponding web domain string and Entity Info Arrays.

3.4. The RDF Mappging Module

The RDF Mapping Module address another application of Artificial Intelligence which is Semantic Computing and Information Processing. This module takes as input the EIAs and their corresponding geographical coordinates from the Geoparsing and Geocoding Module, respectively, and maps all the collected administrative and geographical information into RDF triples that are subsequently stored into the semantic *Km4City* repository; they can also be visualized through a dedicated web-based application, our DISIT Service Map[24]. This is the final goal of the whole mining and annotation process described in the present paper, that is the improvement of georelated services, annotating also private business entities and commercial activities (which are usually not covered by Public Administrations Open Data), as well as providing more advanced geographic based search and visualization capabilities for end users. RDF mapping is carried out by means of a dedi-

[24]http://servicemap.disit.org/

cated Sql2RDF Tool, based on the Karma open source software (Gupta et al., 2015), which automatically performs data-integration (supporting different data formats, such as relational databases, XML, JSON, CSV). An input TTL model has been specifically designed to properly define entities, classes, attributes and relations, as well as specifying imported external ontologies and schemas. The mapping procedure takes as input, in addition to the dataset to be modeled, a OWL based ontology to which input dataset will be mapped, and a database of semantic types previously collected, used to infer, propose and assign semantic models for data mapping. The output is the model describing, for each source, the formal mapping between the source and the ontology, which can be then used to generate RDF. A secondary output is the enrichment and update of the database with the last learned semantic types.

In our case, the *Km4City* proprietary model, as well as external resources such as the GoodRelations and the Geo WGS84 ontologies, are used to model the extracted administrative and geographic information of services and companies. The input Entity Info Arrays and extracted coordinates need no further processing in order to populate required fields for RDF mapping. The RDF triples, generated in output, are then stored in the *Km4City* Ontology. A sample template of RDF triple output describing a single commercial entity is showed in the following:

```
@prefix gr:<http://purl.org/goodrelations/v1#> .
@prefix schema: <http://schema.org#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix km4city: <http://www.disit.org/km4city/schema#> .

gr:<URI_Identifier> rdf:type gr:BusinessEntity .
gr:<URI_Identifier> gr:legalName "<COMPANY_OR_SERVICE_LEGAL_NAME>" .
gr:<URI_Identifier> gr:vatID "<COMPANY_VAT_ID>" .
gr:<URI_Identifier> schema:url "<WEBSITE_URL>" .
gr:<URI_Identifier> schema:addressRegion "<REGION_NAME>" .
gr:<URI_Identifier> schema:addressLocality "<CITY_NAME>" .
gr:<URI_Identifier> schema:postalCode "<POSTAL_CODE_VALUE>" .
gr:<URI_Identifier> schema:streetAddress "<STRRET_NAME>" .
gr:<URI_Identifier> km4city:houseNumber "<STREET_NUMBER>" .
gr:<URI_Identifier> schema:telephone "<TELEPHONE_NUMBER>" .
gr:<URI_Identifier> schema:faxNumber "<FAX_NUMBER>" .
gr:<URI_Identifier> schema:email "<EMAIL_ID>" .
gr:Location_<URI_Identifier> rdf:type gr:Location .
gr:<URI_Identifier> grt:hasPOS gr:Location_<URI_Identifier> .
gr:Location_<URI_Identifier> geo:latitude "<LATITUDE_VALUE>" .
gr:Location_<URI_Identifier> geo:longitude "<LONGITUDE_VALUE>" .
```

As shown, syntax and fragments imported from external resources and

languages, such as the schema.org, RDF and XML Schema and our *Km4City* Ontology, have been used for mapping.

4. Evaluation

The crawler module, initialized with a predefined set of seed URLs, periodically runs scheduled crawling tasks, in order to collect the most up-to-date resources, since web sites maps can change frequently over time. More than 9 million URLs have been fetched starting from a list of seed URLs containing web domain references of business entities and research institutes operating in the Tuscany region, in Italy. The proposed system has been evaluated by assessing its geoparsing and geocoding performances on a subset of the above mentioned URL collection, containing about 160000 URLs. Since we are interested in extracting geographical locations at domain level, a ground-truth reference was built by automatically acquiring administrative and geographical information related to business entities from free available online resources, covering nearly 6 million Italian commercial entities, which will be referred in the following as our reference dataset for geoparsing evaluation.

A two-steps validation has been performed: in the first step, the extracted Entity Info Arrays for each web domain are automatically compared with the reference dataset fields; for those cases in which inconsistencies and unexpected ambiguities are found during the automatic string matching procedure, a manual, supervised comparison is made against geolocation information perceived by a human evaluator navigating the same analyzed domain. In the second step, geographic web domain coordinates automatically extracted by the Geocoding Module are sent in input to a third party geovisualization service, so that the correctness of the resulting location can be manually assessed. To this aim, OpenStreetMap has been found to be useful tool for this purpose, providing street number resolution; we also exploited the *Street View* functionalities of Google Maps whenever OpenStreetMap has not provided sufficient spatial resolution. Human support is considered still necessary for an accurate evaluation; moreover, a manual validation is coherent with human expectations of web users searching for geographic information associated to a certain entity. Standard metrics *Precision*, *Recall* and *F-Measure* have been adopted: this approach relies on the observation of true/false positives and true/false negatives, usually performed in a typical binary classification test. Before describing the two evaluation scenarios, let

us to recall the definition of employed validation metrics:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN},$$

$$F\text{-}Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall},$$

being TP, FP and FN the number of true positives, false positives and false negatives, respectively. In the following, the details of the two-steps validation processed are reported.

4.1. Geoparsing Evaluation

For a general description of statistical metrics please refer to (Han et al., 2012), (FpFn, 2015) and (BinCla, 2015). In this context, the following matching criteria have been followed: a true positive (TP) is detected whenever the geographical information returned by our proposed system for a certain domain, matches the physical location of the legal entity described in the same analyzed web domain. This involves a string match between the extracted Entity Info Array and the previously collected ground-truth reference. As stated earlier, this is a semi-automatic procedure, allowing a manual supervised evaluation in those cases in which inconsistencies and unexpected ambiguities are found in the automatic matching with the reference dataset. A false positive (FP) is found whenever the information contained in a given EIA, extracted for a certain web domain, does not match with the corresponding entity in the ground truth, or whenever it differs from the one perceived by the human evaluator. A false negative (FN) is reported when the proposed system is not able to extract significant geographical information at required granularity level, while such information is instead present in the reference data, or in case when it is perceived in the corresponding web domain. Finally, a true negative (TN) is detected when neither the proposed system is able to extract spatial location, nor the analyzed web site contains such information. Table 3 shows the detailed results of the geoparsing evaluation.

The proposed system shows very good capabilities on retrieving relevant information on geographic location of analyzed web domains (with Precision and Recall rates both above the 90%), with only a 6.7% False Negative rate. Through a qualitative error analysis we found that some of the most commons sources of False Positives in the Geoparsing task (which we believe is the most interesting case) are represented by erroneous assignation

Table 3: Details of the *Geoparsing* evaluation results.

TP	**FP**	**FN**	**TN**	**Precision**	**Recall**	**F-measure**
83.5%	7.1%	6.7%	2.7%	92.1	92.6	92.3

of EIA fields by the Hierarchical Clustering algorithm. This may occur, for instance, when the text content of a web page or document contains many different toponyms and geographical names, not related to the organization administrative address, which can be collected as potential candidates for EIAs fields such as City, Region or Country. False Negatives mainly occur when the system does not find geographical information in the header nor in the footer of the web pages, so that it have to process the entire text content of the web page, dealing with larger and unstructured text fragments. Reported results show how the proposed system often outperforms reported values for GIR system reviewed in the State of the Art (see Section 2.1) and reported in Table 1, although this can be only a qualitative comparison since test conditions and data sets are different.

4.2. Geocoding Evaluation

As for the geocoding process, geographical coordinates are obtained by submitting SPARQL expanded queries (formed with combinations of EIAs fields) to our semantic *Km4City* repository. In this context, evaluation criteria are slightly different: a true positive (TP) is detected whenever both latitude and longitude values are correctly returned by our system for a certain web domain, that is when returned coordinates identify the same location (at street number definition) perceived by a human user/evaluator using the visualization tool chosen for the assessment (in this case, as previously stated, OpenStreetMap has been chosen). A false positive (FP) is detected when the system extracts coordinates which are different from the ones perceived by the evaluator (always considering a street number granularity). A false negative (FN) is found when the system is not able to extract geographical coordinates, while the human evaluator actually perceives them. Finally, a true negative (TN) is reported when highly sparsely populated, incomplete or not-well formed EIAs do not allow neither our system nor the human evaluator to extract coordinates.

Concurrently, a parallel evaluation has been conducted, on the same EIAs set produced in output by our Geoparsing Module, using an external geocod-

ing system: the Google Geocoding APIs. String combinations and patterns from EIAs fields have been defined in a similar way with respect to the approach adopted in designing our Geocoding Module (although in this case we do not deal with semantic queries), and they have been used to form evaluation requests submitted to Google Geocoding APIs, returning the estimated geographical coordinates which are evaluated against the same visualization tool used for assessing our framework. The results of both these geoparsing evaluations are reported in detail in Table 4.

Table 4: Details of *Geocoding* evaluation results against both our *Km4City* semantic repository and the Google Geocoding APIs.

Employed Framework	**TP**	**FP**	**FN**	**TN**	**Precision**	**Recall**	**F-measure**
Km4City Repository	73.4%	7.0%	14.1%	5.5%	91.3%	83.9%	87.4%
Google Geocoding APIs	62.7%	30.8%	2.9%	3.6%	67.0%	95.5%	78.6%

In this case, the higher Precision reached by the proposed system reflects the quality of the *Km4City* semantic repository for local resources. Actually, it is composed by Linked Open Data generated from local Public Administration OD, which have been carefully mapped and reconciliated regarding geographical coordinates. The higher recall rate achieved when using the Google Geocoding APIs can be explained by the fact that Google exploits improved search solutions, such as fuzzy technology, so that it is almost always able to retrieve a result for submitted queries (actually, it shows only a 2.9% of false negatives). Moreover, the Google Repository is by far larger than our Smart City data store, indexing a huge amount of online web resources. On the other hand, this significantly affect the precision rate, together with a not always aligned mapping between real locations and provided coordinates (considering the desired street number resolution), such cases occurring especially in sub-urban and rural areas.

5. Conclusions

In this paper, a system for extracting geographical locations and coordinates of web-visible human activities from their web domains is presented. This work mainly addresses some aspects of Engineering Applications of Artificial Intelligence such as Data Mining, Semantic Information Processing and

NLP, in the specific: Text Mining and Annotation, Named Entity Recognition, Part-of-Speech (POS) tagging, RDF representation of extracted data for semantic ontology enrichment. The proposed system falls into the hybrid approaches category, actually it is based on NLP techniques jointly combined with external knowledge (in the form of annotated gazetteers and stop-word blacklists) and data clustering techniques to improve the extraction capabilities and performances. Geographic information retrieved is then semantically mapped into RDF triples (applying proprietary and external models, such as the GoodRelations and the Geo WGS84 ontologies) which are at the basis of our *Km4City* ontology, repository and services, designed for Smart City applications in the Florence metropolitan area and the Tuscany region, in Italy. Evaluation results show high values for F-Measure, Precision and Recall metrics. These seem to be really encouraging outcomes, taking into account the several difficulties encountered in geographical information retrieval from unstructured text. To solve these aspects, future perspective lead towards refinement of geoparsing methods, since this can in turn improve geocoding performances. Open issues involve further advancements in NLP-based techniques and hybrid approaches; exploiting external resources could be useful to improve identification and disambiguation of VIP names (often present in street labels). Besides, further advancements can be brought by achieving a deeper semantic integration with more expressive semantic models and linked data. For instance, this can foster machine learning processes, automatic reasoning and inference mechanisms which can contribute to improve results in presence of unstructured and incomplete data. For future work, our goal is also to extend the use case and application domain to larger geographical areas. Moreover, additional features can be planned, such as the annotation of the organizations' activities, which may require additional NLP-related techniques, such as text summarization and content extraction.

References

Abascal-Mena, R., López-Ornelas, E., 2010. Geo information extraction and processing from travel narratives. In: Proc. of the 14th International Conference on Electronic Publishing. Helsinki, Finland. pp. 363–373.

Ahlers, D., Boll, S., 2008. Retrieving address-based locations from the web. In: Proc. of the 2nd International Workshop on Geographic Information Retrieval, GIR '08. pp. 27–34.

Amitay, E., Har'El, N., Sivan, R., Soffer, A., 2004. Web-a-where: Geotagging web content. In: Proc. of the 27th Annual international ACM SIGIR conference on Research and development in Information Retrieval), Dallas, Texas, USA. pp. 273–280.

Anderson, M., Kohler, J., 2004. Analyzing geographic queries. In: Workshop on Geographic Information Retrieval (SIGIR).

Asadi, S., Yang, G., Zhou, X., Shi, Y., Zhai, B., Jiang, W. R., 2008. Pattern-based extraction of addresses from web page content. Progress in WWW Research and Development, Lecture Notes in Computer Science 4976, 407–418.

Bellandi, A., Bellini, P., Cappuccio, A., Nesi, P., Pantaleo, G., Rauch, N., 2012. Assisted knowledge base generation, management and competence retrieval. International Journal of Software Engineering and Knowledge Engineering (JSEKE) 32 (8), 1007–10038.

Bellini, P., Benigni, M., Billero, R., Nesi, P., Rauch, N., 2014a. Km4City Ontology bulding vs data harvesting and cleaning for Smart-city services. International Journal of Visual Language and Computing 25 (6), 827–839.

Bellini, P., Nesi, P., Venturi, A., 2014b. Linked Open Graph: browsing multiple SPARQL entry points to build your own LOD views. International Journal of Visual Language and Computing 25 (6), 703–716.

BinCla, 2015. Wikipedia Binary Classifications Metrics. `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`, [Online; accessed 18-November-2015].

Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J., 2008. Deployment of RDFa, Microdata, and Microformats on the web - a quantitative analysis. Knowledge Engineering: Practice and Patterns, Lecture Notes in Computer Science 5268, 329–346.

Buscaldi, D., 2011. Approaches to disambiguating toponyms. SIGSPATIAL Special 3 (2), 16–19.

Cai, W., Wang, S., Jiang, Q., 2005. Address extraction: Extraction of location-based information from the web. Web Technologies Research and

Development - APWeb 2005 - Lecture Notes in Computer Science 4976, 925–937.

Chang, C. H., Li, S. Y., 2010. Mapmarker: Extraction of postal addresses and associated information for general web pages. In: Proc. of the 2010 IEEE/WIC/ACM Int. Conference on Web Intelligence and Intelligent Agent Technology. pp. 105–111.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: A framework and graphical development enviroment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL '02.

FpFn, 2015. Wikipedia False Positives and False Negatives. `https://en.wikipedia.org/wiki/False_positives_and_false_negatives`, [Online; accessed 18-November-2015].

Gupta, S., Szekely, P., Knoblock, C. A., Goel, A., Taheriyan, M., Muslea, M., 2015. Karma: A system for mapping structured sources into the Semantic Web. The Semantic Web: ESWC 2012 Satellite Events, Lecture Notes in Computer Science 7540, 403–434.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, 3rd Edition. The Morgan Kaufmann Series in Data Management Systems, Elsevier.

Hepp, M., 2013. Goodrelations: an ontology for describing products and services offers on the Web. The Semantic Web, Lecture Notes in Computer Science 8219, 17–32.

Kimler, M., 2004. Geocoding: Recognition of Geographical References in Unstructured Text, and Their Visualisation.

Loglisci, C., Ienco, D., Roche, M., Teisseire, M., Malerba, D., 2012. Web data commons: extracting structured data from two large web corpora. Proc. of the 4^{th} Linked Data on the Web Workshop LDOW2012.

Loos, B., Biemann, C., 2008. Supporting web-based address extraction with unsupervised tagging. Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization, 577–584.

Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Sanderson, M., Santos, D., Womser-Hacker, C., 2008. An evaluation resource for geographic information retrieval. In: Proc. of the 6th International Conference on Language Resources and Evaluation (LREC).

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., Gaio, M., November 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In: Proc. of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014), Dallas, Texas, USA.

Nesi, P., Pantaleo, G., Tenti, M., Nov 2014. Ge(o)lo(cator): Geographic information extraction from unstructured text data and web documents. In: Proc. of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2014). pp. 60–65.

Pouliquen, B., Steinberger, R., Ignat, C., De Groeve, T., 2004. Geographical information recognition and visualisation in texts written in various languages. In: Proc. of the 19th Annual ACM Symposium on Applied Computing, Nicosia, Cyprus. pp. 1051–1058.

Scharl, A., 2007. Towards the Geospatial Web: Media platforms for managing geotagged knowledge repositories. The Geospatial Web, Advanced Information and Knowledge Processing, 3–14.

Schmidt, S., Manschitz, S., Rensing, C., 2013a. Extraction of address data from unstructured text using free knowledge resources. Proc. of the 13^{th} International Conference on Knowledge Management and Knowledge Technologies Graz, Austria.

Schmidt, S., Manschitz, S., Rensing, C., Steinmetz, R., 2013b. Extraction of address data from unstructured text using free knowledge resources. In: Proc. of the 13th International Conference on Knowledge Management and Knowledge Technologies. pp. 273–280.

Tjong Kim Sang, E. F., De Meulder, F., 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proc. of CoNLL-2003. pp. 142–147.

Yu, Z., 2007. High Accuracy Postal Address Extraction From Web Pages.

Zhang, Q., Jin, P., Lin, S., Yue, L., 2012. Extracting focused locations for web pages. Lecture Notes in Computer Science 7142, 76–89.

Zhang, W., Gelernter, J., 2014. Geocoding location expressions in Twitter messages: a preference learning method. Journal of Spatial Information Science, 37–70.

Zhao, J., Jin, P., Zhang, Q., Wen, R., 2014. Exploiting location information for web search. Computers in human behavior. SIGSPATIAL Special 30, 378–388.

Zickuhur, K., 2012. Three-quarters of smartphone owners use location-based services. Report of the Pew Internet And American Life Project.