Artificial Intelligence Review (2022) 55:1085–1120
https://doi.org/10.1007/s10462-021-09983-1

Postal address extraction from the web: a comprehensive survey

Mohammed Kayed[1] · **Sara Dakrory**[2] · **A. A. Ali**[2]

Published online: 14 March 2021

Abstract
The Web is a source of information for Location-Based Service (LBS) applications. These applications lack postal addresses for the user's Point of Interests (POIs) such as schools, hospitals, restaurants, etc., as these locations are annotated manually by using the yellow pages or by the location owners (users/companies). Our study in this paper confirms that Google Maps, a common LBS application, only contains about 32.5% of the public schools that are registered officially in the documents provided by the Directorate of Education in Egypt. However, the remaining missed school addresses could be fished from the Web (e.g., social media). To the best of our knowledge, no prior survey has been published to compare the previous Web postal address extraction approaches. Additionally, all proposed approaches for address extraction are local (could be working in specific countries/locations with particular languages) and could not be used or even adapted to work in other countries/locations with other languages. Furthermore, the problem of Web postal address extraction is not addressed in many countries such as Arab countries (e.g. Egypt). This paper discusses the issue of address extraction, highlights and compares the recently used techniques in extracting addresses from Web pages. In addition, it investigates the discrepancy of knowledge among existing systems. Moreover, it provides a comprehensive review of the geographical Gazetteers used in the Web postal address approaches and compares their data quality dimensions.

Keywords Postal Address Extraction · Web Information Extraction · Gazetteers · Machine Learning

✉ Mohammed Kayed
mskayed@gmail.com

Sara Dakrory
sara.dakrory@gmail.com

A. A. Ali
abdelmgeid@yahoo.com

[1] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt

[2] CS Department, Faculty of Science, Minia University, Minia, Egypt

1 Introduction

Location-Based Service (LBS) applications have become indispensable as a result of the prolific spread of mobile communication. Further, the necessity of finding the geographical location of a mobile device and expanding services based on this location information becomes important. Updating and restoring location-based data on a large scale is considered a significant problem in these LBS applications. For example, people use map services to find Points of Interests (POIs) such as schools, hospitals, restaurants, gas stations, pharmacies, etc. Unfortunately, many of these POIs datasets are manually annotated by users/companies which is costly, error-prone and inefficient. There is no guarantee that all POIs are defined on the maps since none of the LBS applications provides any statistics about their coverage percentage or completeness. Several reasons cause the absence of information related to POIs such as lack of visitors knowledge, newly opened places, unattractive places that haven't had many visitors, and many other situations that shall be discussed later. Furthermore, map services mainly work to provide users with directions rather than addresses. Postal address extraction from the Web (surface or deep Web) has been suggested to augment such LBS applications with these lacked postal addresses data.

Search engines face many challenges in extracting addresses and location names from the Web. Traditional information extraction and natural language processing are not effective in the context of the Web because of the uncontrolled heterogeneous nature of the Web resources as well as the effects of HTML and other markup tags. Addresses on the Web could exist with different structures, formats or languages. Not all addresses were written with the same structure which makes many possibilities of the expected address structures. Cai et al. (2005a) have defined a postal address as "a knowledge structure which includes suite information, municipal location and regional position". So, a postal address has many entities that are combined to form the full address Borges et al. (2007). Some entities are compulsory to be shown in the address, while others are optional entities. Even, two addresses with the same entities could be provided in different orders. Generally, the task of postal address extraction includes three main steps: looking for geographical indicators, identifying the postal address boundaries and finally extracting the entities of the address. These three steps could be summarized as follows.

1. **Step 1 (Geographical Indicators):** Specific evidences/indicators can be used to indicate that a Web page contains one or more addresses Borges et al. (2007). Literatures usually use the place name as the key geospatial proof inside a page (i.e., a restaurant name like Pizza Hut). Postal code is also a strong proof of locations, as their identification enables the page to be specifically associated with a particular part of the world (i.e. 90209 is a postal code in Beverly Hills city in the state of California). In addition, ground line telephone numbers indirectly carry location information, since the numbering is structured according to geographical principles to ensure efficient cabling and distribution of equipment (i.e., +20-86-xxx-xxxx is the pattern for a phone number in Egypt while +20 is the country calling code and 86 is the code for Minia city).
2. **Step 2 (Address Boundaries Detection):** Web pages must be parsed carefully to locate the beginning and the ending of a postal address pattern using the address boundaries. Although each country has its standard address pattern, addresses on the Web pages may exist in different structures. In US, addresses start with a street number followed by street name, city, state name, optional ZIP code and country name Yu (2007). This means, a street number is the start boundary, while a county name is the end boundary. However,

Table 1 Results of a simple search about "public school names" in a particular city

Location	Minia City, Egypt
Dataset Size	560 records (School names/addresses provided by the formal office of Directorate of Education)
LBS	Google Maps
# of addresses found in the LBS	174
Missed address reasons:	– For some school names, Google Maps retrieves only one address to the school, while other branches of the school with different addresses are missed. – For other school names, the system has revealed no data. – Moreover, it has been found that even the formal organizations did not follow a strict address formula. The address pattern is a descriptive text rather than a formal structure. This problem leads to a variety of existing address patterns to be linked to the same location and subsequently affects the extraction of the addresses from the Web.

in other countries, those borders aren't the same. For example, Brazilian addresses start with the street type Borges et al. (2007), while in Egypt it has been observed that addresses might begin with a street name or some keywords such as "road", "street", or "avenue".

3. **Step 3 (Address Entities Extraction):** By studying the address patterns in most countries, an address may include a combination of the following entities Borges et al. (2007):

   (a) Basic entities: Entities such as a street name and a building number are required to identify the address.
   (b) Optional entities: These entities provide a complement to the basic entities in the address such as a neighborhood name.
   (c) Location entities: To recognize an address location, at least one out of four location identifiers must be presented. These identifiers are: postal address, phone number, building number and city/state.

Three main reasons motivate us to write this research paper. First, to the best of our knowledge, no prior work has surveyed, addressed or compared the previous approaches of address extraction from the Web. Second, all proposed approaches for address extraction are local and include several constraints of the address structure/format. Further, some of them use gazetteers that are available only at particular places. Consequently, there are no general Web address extraction systems, that can be used internationally for multi-languages or adapted to be used in various countries that have been proposed yet. For example, to the best of our knowledge, no prior approach has been proposed to solve the address extraction problem in Arab countries such as Egypt. Third, as mentioned above, address extraction from the Web is a hot research topic as it could be used to provide LBS applications with the postal addresses data that are annotated manually by users/companies. To confirm that LBS applications are lacked of annotated addresses and so address extraction is of great necessity, we practically conduct a searching experiment about school names at Minia city, Egypt. Table 1 summarizes the results of this searching. As shown in the table, 174 out of 560 ($\simeq$ 32.5%) school names (registered officially in the documents provided

Table 2 Examples of social media and other pages that containing the missed schools on Google Maps

School Name	Corresponding URL
Al-Gemhoria Primary School	https://bit.ly/2RM3szY
Tarek-Bin-Ziad Primary School	https://bit.ly/2Kig8dY
Omar Ibn Abd Elaziz Primary school	https://bit.ly/3aERq1Y
El Menia Preparatory Sport School for Boys	https://bit.ly/2VC9HaR
Kafr EL-Mansoura preparatory school for Girls	https://bit.ly/2XLFybO

by the Directorate of Education at this city) are identified on Google Maps services. Many school names are missed on Google Maps for the reasons that are addressed and summarized in the table. Moreover, Table 2 shows that most of these missed postal addresses could be seen and so extracted from the Web (Facebook, Twitter or other advertising pages). Therefore, this simple experiment confirms that postal address extraction from the Web is a hot topic and is still demanded many LBS applications.

This paper discusses the problem of address extraction, and highlights the recently used techniques in extracting addresses from Web pages as an alternative to the existing LBS data annotations. It also suggests criteria to evaluate and compare these discussed approaches. Moreover, this paper provides a comprehensive review of the geographical Gazetteers used in the Web postal address approaches and compares their data quality dimensions. In other words, this paper tries to answer the following questions: What are the existing approaches for address extraction? Which part of the address needs to be extracted? What are the datasets/Gazetteers available and used in these approaches? What are the most suitable criteria to evaluate these approaches?

The rest of the paper is organized as follows. Section 2 gives a background and discusses the related works. Section 3 presents the techniques used for address extraction, while Sect. 4 provides a comparative analysis of the gazetteers and the techniques discussed in the research literatures. Finally, Sect. 5 concludes our work and presents our future directions.

2 Background

This paper concerns the following topics: Named Entity Recognition and Classification (NERC), Geo-Parsing Tools, Gazetteers, and Segmentation in the problem of postal address extraction. In this section, we briefly discuss the details of these topics.

2.1 Named entity recognition and classification

Named Entity Recognition and Classification (NERC) is the process of identifying and classifying atomic elements in a sentence based on pre-defined named entities such as organizations, persons, locations, etc. Nadeau and Sekine (2007). The entity can be one of two categories: ENAMEX (which contains person, location, organization); or NUMEx (which contains time, currency and percentage expressions entities) Chinchor and Robinson (1997). The Named Entity Recognition (NER) systems expose persons' names, organizations, locations, dates and times, while the address extraction systems

consider classifying the detected location entities into names of cities, provinces/states and countries aiming to map them into physical locations. The postal address can be manipulated as a location entity with consolidated features. Further, it combines many entities to form the whole location pattern. Two of the most influential research groups that have given significant attention to the information extraction approach over the last decade are the Message Understanding Conference (MUC) and the Text Retrieval Conference (TREC). These two groups introduced several pieces of researches focusing on extracting street addresses information for the purpose of the vehicle launch event Chinchor and Robinson (1997). However, these researches were not proven to be successful in location-based data identification. In addition, it had difficulty in contending the diversity of address forms. On the other hand, a standard address database was exploited to perform geo-parsing of Web pages Morimoto et al. (2003) Sagara and Kitsuregawa (2001).

The NER was originally labeled as an essential subtask of the domain of Information Extraction (IE). NER is considered as one of the substantial Natural Language Processing (NLP) tools. It provides the Information Retrieval (IR) domain with recognized Named Entities (NEs) within the query and searched documents Benajiba et al. (2009), Freihat et al. (2018). Geographical Information Retrieval (GIR) Larson and Frontiera (1996) handles indexing, searching, retrieving and browsing georeferenced information sources. Compared to IR, GIR assumes that certain semantic data related to geography are presented either in the form of geographical metadata or by incorporating semantic ideas about spatial relationships and locations. It was found that (71%) of the queries in search engines contain NEs Guo et al. (2009).

NER has exploited both rule-based methods and statistical-based methods. However, the rule-based method is more reliable than others. In addition to that, it is more comprehensive since it is more relative to human reasoning. On the other side, this method has a lack of portability owing to the fact that it is dependent on the nature of the extracted language, area and text format. Furthermore, the coverage of the used rules cannot reach 100% for all named entities. In contrast, the statistical-based method has superior robustness and flexibility. It is objective and does not need too much manual intervention and domain knowledge. These models have been used in named entity recognition. Examples of these models are: Maximum Entropy Model (Borthwick et al. (1998), Chieu and Ng (2002) Hui et al. (2009)), Hidden Markov Model (HMM) (Bikel et al. (1997), Freitag and McCallum (1999), Borkar et al. (2001), Zhou and Su (2001), Zhao (2004)), Support Vector Machine (SVM) (Takeuchi and Collier (2002), Ekbal and Bandyopadhyay (2010)), C4.5 Decision Tree (Sekine et al. 1998) and Conditional Random Fields (CRFs) (Han et al. (2013), McCallum 2002).

Currently, several neural network architectures have been successfully implemented to NER instead of the traditional linear statistical models Tjong et al. (2003). Huang and Yu Huang et al. (2015) considered NER as a sequential token tagging task. Consequently, this limited the dependency on hand-crafted features extracted by NLP tools and external knowledge resources Zheng et al. (2017). For the purpose of eliminating the features engineering, Long Short-Term Memory (LSTM) based model on sequence tagging is proposed. LSTM is used with CRF model as well as Convolutional Neural Network (CNN) Chiu and Nichols (2016), Lample et al. (2016), Ma and Hovy (2016).

Many applications utilize the named entity task in their implementations. In this section, we will characterize these NER-based applications.

2.1.1 Auto query and answering (AQA)

These applications aim to produce answers to questions that were initially created in natural languages. An AQA system employs NERC to find the answers that respond to multiple fact-based questions. The detected entities by NERC system represent the answers to these questions. Therefore, by combining the NERC and AQA systems, the task of finding answers to some of the questions becomes more achievable and convenient Cavedon et al. (2006), Rodrigo et al. (2013).

2.1.2 Machine translation

These applications take text or speech as an input from a source language and then convert it into another (target) language automatically without any human interference Rodrigo et al. (2013). This process is not a trivial task. Many approaches and procedures are required to translate proper names rather than other word types Babych and Hartley (2003). Furthermore, the mistranslation of NEs such as generic nouns predominately leads to an incomprehensibility and the demand for a comprehensive editing.

2.1.3 Automatic text summarization and IR

Automatic text summarization is the process of exporting salient succinct points in a source document. In these applications, NEs are considered as important indications of the document topic. They are considered as useful key expressions for text summarization Nobata et al. (2002), Baralis et al. (2013). Another important domain that exploiting NEs is IR. The main function of an IR application is to respond to user queries by fetching the related information from a collection of resources Faloutsos and Oard (1998). The queries are represented by a collection of strings which include keywords or named entities. These keywords or entities are matched with the information content stored in large databases to facilitate the accessibility of the information system Betina and Mahalakshmi (2015). However, the existence and the number of named entities significantly influence the performance of IR systems Mandl and Womser-Hacker (2005).

2.1.4 Text clustering

Text clustering gathers text documents into groups (clusters) where texts in the same group have the same properties. This process is often exploited in knowledge discovery where words are clustered in groups such as persons, locations and organizations. The quality of the text clustering approaches revealed a prominent refinement through using NERs, especially in Suffix Tree Clustering (STC) Zhang et al. (2013).

2.1.5 Ontology

An ontology can be defined as a term of “a specification of a conceptualization”. It comprises three components: concepts, axioms and relationships. An ontology or a thesaurus can be utilized to merge synonyms and other related syntax. Designing an ontology includes the extraction of entities and concepts from data as well as learning the semantic and the conceptual relationship among them. NEs are considered as one of the

most effective methods in the ontology population. Designing the ontology in a manual manner ordinarily defines the concepts for the domain. However, individual occurrences of the concepts are frequently missing despite their importance in using ontology as a knowledge base. Moreover, it is costly to construct instances manually. To handle this issue, some endeavors were undertaken to construct an automatic system for ontology population. Such endeavors were proposed by Resnik (1995), Song et al. (2009) ('know-itall' is a type of these systems Etzioni et al. 2005). Moreover, the ontology of places defines concepts from a particular domain of urban geographical space. The hierarchy of territories, in which regions are subdivided into other regions, is explored by the ontology, as well as the concepts related to urban addresses and commonly used landmarks. An ontology also considers telephone numbers and postal codes, which can be used as indirect location identifiers. The ontology's spatial knowledge elements can be used to infer geographical relationships among objects, such as proximity, adjacency and containment. Besides, Geospatial ontologies can be employed in enabling semantic information extraction Kokla et al. (2018) specially in the geospatial semantics Kuhn (2005). Furthermore, the traditional vector cannot present the relationship among concepts. Literatures that have used this technique for address extraction will be discussed in Sect. 3.2.

2.1.6 Opinion mining

Opinion mining, or sentiment analysis, is the computational thinking about people's suppositions, states of mind and feelings towards substances such as products, administrations, organizations, people, occasions and their distinctive perspectives. Moreover, it has been a dynamic investigation zone in natural language processing and Web mining a long time. Researchers have considered opinion mining at the document, sentence or aspect levels. Aspect-level, known as aspect-based, opinion mining is frequently desired in practical applications. It provides the key conclusions or assumptions about different aspects of entities and entities themselves, which are ordinarily required for many activities Chu (2013). Moreover, people express their opinions openly on social Web pages on a variety of topics or products Popescu et al. (2005). These opinions influence decision-makers as it became a satisfaction measurement tool regarding their products or actions. Therefore, NERC represents an imperative role in opinion mining. The system OPINES Popescu et al. (2005) has been developed for the extraction of attributes of products and the analysis of related opinions.

2.2 Geo-parsing tools

Geo-tagging and Geo-parsing are two distinct processes that were defined in the context of address extraction from the Web Machado et al. (2010). Geo-tagging is the process of identifying geographical entities mentioned directly or indirectly in the text and creating tags that allow the document to be linked to a location or set of locations Teitler et al. (2008). On the other hand, geo-parsing encompasses location extraction and location disambiguation Jones and Purves (2008). The two processes require recognition of geographical references found in the text. If this task is fulfilled adequately, the geographical context of the document can be established. Many tools have been used in extracting the geographical information. Some of them were focused mainly on locations such as Metacarta and Digital Reasoning Geolocator 2.0, while others (e.g., NetOwl tool) encompassed other types

of entities such as people names, organizations, places (e.g., countries, cities), addresses, artifacts, phone numbers, titles, etc. In this section, we briefly introduce three Geo-parsing tools. For each one, the techniques used and the tool features are mentioned.

2.2.1 Metacarta

Metacarta[1] employs a text search algorithm which utilizing the geographical keywords to retrieve all the related contents. This framework can identify and detect geographical references, parse a document/a text query/a URL and return either an image location map or a JavaScript response for the locations. Moreover, this framework returns all possible location matches for a specific geographical query. This offers users a simple way to refine an initial area of focus. Moreover, it supports many languages such as Spanish, Russian, French, and Arabic.

2.2.2 Digital reasoning geolocator 2.0

The Digital Reasoning Geolocator 2.0[2] tool works on unstructured texts (i.e emails, instant messages and other documents) to extract the place's names and return the locations with their relevant geo-coordinates. It encompasses the following features: obtaining the current location of the device; the last known location; the continuous location updates; checking if location services are enabled on the device; translating an address to geo-coordinates and vice versa; and calculating the distance (in meters) between two geo-coordinates.

2.2.3 NetOwl

NetOwl[3] considers various entities that support many types of organizations (e.g., companies and governments), places (e.g., countries and cities), addresses, artifacts, phone numbers, titles, etc. Moreover, the entity extraction process can be extended to include relationship extraction and event extraction. NetOwl is a multilingual platform involving English, Arabic, Chinese (traditional and simplified), French, German, Korean, Persian (Farsi and Dari), Russian and Spanish. Furthermore, one of the features of NetOwl is Geotagging which enables disambiguation and normalization of place names. Name normalization is also used in cross-document name resolution for applications such as faceted search, geospatial analysis and link analysis. NetOwl can be integrated easily with many popular searches, geospatial and business intelligent tools such as Elasticsearch, Solr, MarkLogic, Esri ArcGIS, Tableau, Kibana, etc. Finally, it allows applying an English translation of named entities extracted from foreign language text. This translation uses different alphabets/scripts such as Arabic, Chinese (traditional and simplified), Korean, Persian (Farsi and Dari) and Russian.

[1] http://qbase.com/products/metacarta/
[2] https://digitalreasoning.com/blog/digital-reasoning-releases-geolocator-2-0/
[3] https://www.netowl.com/

2.3 Gazetteers

A Gazetteer can be defined as a geographical dictionary, also known as a toponymical dictionary, which is an important reference for information regarding locations and place names. Most digital gazetteers can be defined as geospatial dictionaries of geographical names which conventionally contain three core elements: a name (could also have variant names), a corresponding location (coordinates representing a point, a line or a real location) and a type (selected from a type scheme of categories for places/features) Hill (2000). Many studies are concerned with maintaining gazetteers or Point of Interest (POI) databases. The aim of constructing such a gazetteer or a database is to automatically provide an updated geographical source that simultaneously affects the user map searches. Here are some of the existing gazetteers that have been proposed by either researchers or commercial companies.

GeoNames[4] is an open-access geographical database which encompasses over 25 million geographical names such as place names, elevation and population. It consists of over 11 million unique features of 4.8 million populated places and 13 million alternate names. Despite this, it has been found that the existing data demonstrate a few random systematic errors which in a few cases do not differ significantly from the correct data. Moreover, there is a coordinate shortage as, in many cases, certain areas have an extent at the subminute scale that is not captured. Furthermore, there is an overlapping among places that results from the inaccurate topology Ahlers (2013).

OpenStreetMap (OSM)[5] depends on the Volunteered Geographical Information (VGI) Goodchild (2007), Gao et al. (2014) that are provided by the users which give it the capability to be editable and freely accessed. The database is authorized under copyright schemes. The uploaded data into OSM by volunteers are modeled and stored in 3 types of tagged geometric primitives: points, paths (polylines) and relationships (linking points and paths with tags). Recently, OSM applications have aimed to foster the mapping creativity of potential contributors to geographical data. Thus, there are various Web sites providing OSM data in a shape file format such as CloudMade or GeoFabrik.

Uryupina (2002): was one of the first approaches that combined a pattern based technique and Machine Learning (ML) techniques to extract gazetteers from Web pages. The aim of this work was to learn new gazetteers using a small set of reclassified examples. The author used a dataset collected randomly from the world atlases which resulted in a dataset of 1260 location names that are manually classified. Later on, Uryupina (2003) utilized the bootstrapping approach to efficiently combine a small portion of labeled (seed) examples with a much larger amount of unlabeled data. These approaches limited the need to encode knowledge manually and had the benefit to obtain new place names automatically. As mentioned by the authors, the use of classifiers and gazetteers would increase the efficiency of the extraction process.

Locus Souza et al. (2005): is a spatial locator system that has been utilized to extract spatial information on Web pages to build a gazetteer using ontology. Through ontology of places, Locus holds place names for entities such as cities and rivers. In addition, it handles intra-urban place names such as street names, urban landmarks and postal addresses, along with their spatial relationships. This work emphasizes the importance

[4] https://www.geonames.org/

[5] https://www.openstreetmap.org

of using ontology to raise gazetteers performance which in turn affects various fields such as the geocoding process. Additionally, it overrides the availability of spatial databases by proposing a semi-automatic technique to populate the Locus gazetteer with geographical content extracted directly from the Web.

Ontology-based Gazetteer: To provide a clear view of place's semantic, relations among the geographical names must be considered. An ontology-based gazetteer that has been proposed by Machado et al. (2010) achieved this by utilizing the connections among the places, and considered the terms and the expressions that characterize it. The introduced gazetteer added the benefits of recording concepts and terms belonging to a place in addition to identifying the name of this place. This initiative is anticipated to assist in solving challenging issues such as place name disambiguation, geographical text classification and geographical context recognition. The used dataset was comprised of news text. A collector was developed to extract and store its title and body text, using XPath. The sole focus was on "Minas Gerais", a state in the north of Southeastern Brazil. Regular expressions were designed to extract candidate names. The recognition of place names from the news documents was supported by the ontological gazetteer. From the 267 news documents containing place names, 2,244 relationships among places and documents were evaluated: 72% were considered as valid, while most of those relationships were considered as strong.

POI Gazetteer: Chuang et al. (2016) extended the work introduced in Chuang et al. (2014) by building a POI database using Apache Lucent's Solr4 in order to save the returned results from the map searches. This has been achieved by crawling the online Yellow Pages to train a linear-chained CRF model. For the purpose of collecting address pages, the system enabled two Web crawlers. The first one relied on the Yellow pages and the second utilized the store names as the query.

Although gazetteers have been used to improve the performance of postal address extraction approaches, they have some drawbacks that can be summarized as follows.

1. The coverage shortage: A gazetteer does not often include significant geographical data such as intra-urban place names, i.e. street names, neighborhoods, landmarks and tourist attractions.
2. The difficulty of determining reasons for the absence of the places: Many atlases do not list small islands, rivers and mountains. Such gazetteers contain only positive information: if X is not classified as an ISLAND, we cannot say whether there is really no island with the name X or the gazetteer is not complete.
3. The immutable classification of the gazetteer: In many cases, subdividing CITY might be changed into CAPITAL and NON-CAPITAL. In this case, it might be necessary to reclassify all (or a substantial part of) the items. Manual editing consumes time and effort.
4. The absence of spatial relationships that representing region hierarchies.
5. Most gazetteers don't implement generic relationships among object types, which limits the potential use of the gazetteer as a geographical ontology.
6. Gazetteers keep the names of well-defined footprints, thus imprecise locations might be lost.
7. Finally, the gazetteer languages can cause variance in geographical names. It might take a long time to adjust a French gazetteer to German. Moreover, such a resource is hardly effective for languages with non-Latin alphabets (e.g., Armenian or Japanese). Even collecting different proper names in one language is a non-trivial task.

Machado et al. (2010) have handled these difficulties by using geocoding services such as the ones available in Google Maps API, which do not make the gazetteer entries explicit, but are able to supply a pair of coordinates corresponding to a textual description.

2.4 Segmentation in postal address extraction

Web pages segmentation is a vital task in the field of postal address extraction and other related fields such as mobile Web Song et al. (2004), archiving Saad and Gançarski (2010), Web page phishing Cao et al. (2010), duplicate detection Chakrabarti et al. (2008), Information Retrieval (IR) Cai et al. (2004a), Information Extraction (IE) Chang et al. (2006), user interest detection Liu et al. (2004), and Web page clustering Kovacevic et al. (2002). The granularity of the term "segmentation" depends on the object to be segmented. Generally, the process of dividing an object into meaningful units is called "object segmentation". In the domain of postal address extraction, segmenting Web pages into smaller blocks (called address blocks) is called "Web pages segmentation", while segmenting an address block into smaller entities (attributes) is called "address block segmentation" (simply, we call it "block segmentation"). The aim of the Web pages segmentation is to identify the parts of the Web pages that have postal addresses (i.e., recognizing the beginning, the middle and the ending of the addresses in the Web pages). In the block segmentation problem, an address block is divided into entities such as street name, city, state name, ZIP code and country.

Addresses to be extracted are existed in a text format that are embedded as leaf nodes in a Web page DOM tree. Each node in the DOM tree is presented by the browser as an image which could be visually described. Therefore, segmentation approaches could be classified into three main categories: DOM Tree Based Segmentation, Vision-Based Segmentation and Text-Based Segmentation. In this subsection, we shall briefly discuss the aforementioned types of segmentation, respectively, while Section 3 will discuss the common segmentation algorithms that have been applied in different postal address extraction approaches.

2.4.1 DOM tree-based segmentation

A Web page DOM tree is used frequently in the domain of IE for page segmentation as it carries important structural information. The DOM tree based segmentation in the IE domain requires the input pages to be either generated from a database using a template Kayed and Chang (2010) or using ontology driven extraction techniques Gupta et al. (2003). Different from the page segmentation in IE, DOM tree based page segmentation in postal address extraction must consider that addresses in the Web pages don't have a common template; similar to the semantic un-supervised partitioner algorithm in Vadrevu et al. (2005). This partitioner algorithm can identify different segments in a Web page and their instances even in the presence of certain presentation irregularities. It aimed to find homogeneous segments, where the content is presented in a uniform way within each segment. It iteratively traversed the page DOM tree from the root in a top-down fashion Hattori et al. (2007). All uniform sets of nodes are treated as homogeneous segments, while non-uniform nodes are split into smaller segments until each segment is homogeneous. The principle of Entropy is used to evaluate if a section is homogenous.

Chang et al. (2016) proposed a DOM tree based segmentation algorithm based on the farthest distinguishable ancestor (FDA). They identified the subtrees that encompass

address landmarks (such as key suffixes for city/county/road/street names) for address block segmentation. Address blocks are identified based on three main assumptions: (1) all associated address information form a continuous block in the DOM tree, (2) address information blocks are mutually exclusive, and (3) an address information block is contained in a single subtree in the DOM tree, such that information blocks expand the region, provided that they are mutually exclusive. Therefore, address landmarks are used to identify candidate address blocks and the FDA algorithm is used to filter the non-mutually exclusive blocks.

2.4.2 Vision-based segmentation

As mentioned before, each node in the DOM tree is displayed as an image by the browser. Node images on the browser are nested, where the whole displayed image corresponding to the tag <body>, and the image corresponding to each child node in the DOM tree is presented totally inside the parent of this node. The dimension and position of a displayed image are provided by the browser via "offsetWidth and offsetHeight" and "offsetLeft and offsetTop", respectively.

The purpose of vision-based segmentation is to retrieve the visual block that occurred after detecting the separators such as white spaces, lines, photos, images and content. In addition, it uses these information to construct a content structure Cai et al. (2003, 2004b). Cai et al. (2005b) proposed a vision based text segmentation method for detecting postal addresses from webpages. Primarily, all the text snippets were obtained based on visible elements that contained both text content and layout information (i.e. font, border, color, position, size, etc.). Further, all blocks were merged into more integrated and meaning blocks based on their visual similarity and adjacency relationship. Each text block was categorized into cue blocks and body blocks. The first block is used for explanation purposes, while the second block contains the main text body content such as a postal address, telephone number, etc. The address is assumed to be located in the body block. This algorithm employs a top down approach, which is highly efficient.

Vision-based segmentation is computationally costly as it utilized external resources such as CSS files and images. Further, this method obviously has a higher complexity than other approaches because the layout must be rendered prior to the analysis, which may be too sluggish to be integrated into the Web crawling and indexing processes.

2.4.3 Text-based segmentation

In this context, we need to distinguish between two concepts: page segmentation based on text and text block segmentation. The first one refers to the process of extracting block segments from a web page based on low-level text properties rather than DOM-structural properties Kohlschütter and Nejdl (2008). The second concept refers to the process of dividing the text block into topically coherent segments Misra et al. (2011).

Topic modeling Du et al. (2015) has been used with Linear Dirichlet Allocation (LDA) algorithm Choi et al. (2001) to segment text and produce boundaries along with the topic distribution associated with each segment. This could resolve several IR applications such as segment retrieval, discourse analysis and dividing news broadcast transcription into stories. This segmentation technique is considered as a domain-independent and can effectively function if the framework is tested on documents using specific words as references

Misra et al. (2011). Further, the topics in the topic model are predefined, and each document's possible topics are given in the training dataset.

Page segmentation based on text declines the HTML's tree structure. Instead, it takes into account the web page's textual content and analyzes it for features such as link density or text density of sections of the web page. This technique is straightforward because there is no need to build DOM tree. Kohlschütter and Nejdl (2008) proposed Block Fusion algorithm that recognizes segments as an essential heuristic by using the text density metric to segment text documents. First, an HTML document is pre-processed into a list of atomic text blocks by removing the HTML tags. Afterwards the token density can be measured for each atomic block. If the gap between two neighboring blocks' token densities is below a certain threshold value, a merge strategy is used to combine blocks into increasingly larger blocks. The problem with this algorithm is during recursive application on sub-blocks, arbitrary adjustments to the text-density threshold are needed. Further, structural and visual cues are not regarded. On the other hand, block segmentation based on text usually uses alignment or sequence labeling algorithms such as CRFs to break down the block into entities. Word/token statistical information such as PunDensity (density of punctuation in the token), LetterDensity (density of letters in the token), DigitDensity (density of digits in the token) and CapitalStartTokenDensity (density of words that begins with an upper letter) could be used in segmentation. Also, gazateers and landmarks such as city/county/road/street names are used in segmentation as well.

3 The techniques used for postal address extraction

Nesi et al. (2014) have classified postal address extraction approaches into two categories: internal and external. The extraction process in the first (internal) category is maintained without using any external resources, which occurs in cases like extracting addresses based on patterns or statistical rules. The second category refers to the cases when an external source is used which may be a gazetteer or a trained dataset. The existing studies can be classified into four main groups according to how they extract postal addresses from the Web. The first group of studies applies the rule-based address extraction method, in which regular expressions and patterns are utilized for address extraction. In these approaches, patterns are identified and matched with an extracted token to determine whether it is an address or not. The second group uses ontology-based techniques that rely on describing the documents using a defined ontology to extract location-based information. The third category applies machine learning techniques. The fourth group includes hybrid techniques that combine aspects of the previous techniques. In this section, we shall briefly discuss the most recent approaches from the four categories.

3.1 Pattern-based address extraction approaches

Most pattern-based approaches employ two consecutive steps to extract addresses: Segmentation and Recognition. In the first step, unnecessary tags are removed from the Web pages and text segments that include the address blocks are obtained. While the second step aims to detect the different parts (entities) of the address pattern. However, these techniques mostly depend on gazetteers that are used to match a specific segment in the pattern. Gazetteers such as OpenStreetMap (OSM), Geonames and DBpedia may contain states, cities, streets or a subset from them. Regular expression techniques use predefined grammars for

almost all possible patterns of real postal addresses and compare the obtained tokens with lexicons. On the other hand, unstructured postal addresses cause a vital challenge in many countries such as India and Egypt. The problem arises from describing the geographical locations instead of using the formal address structure. Nagabhushan et al. (2006) used symbolic object properties to represent the variant information of the addresses through a knowledge base rather than gazetteers.

Many pattern-based postal address extraction approaches have been proposed. For example, Asadi et al. and can et al. Asadi et al. (2008) ; Cai et al. (2005b) have used pattern matching, while others have combined techniques that integrate patterns and gazetteers Schmidt et al. (2013). The rest of this section will introduce these attempts in a chronological order and the following sections will compare them. At the end of this subsection, a comparison of the discussed pattern-based approaches is shown in Table 3.

Can et al. was one of the primary studies that employed the Web page layout to identify the address on the structured English Web pages Cai et al. (2005b). They relied on segmenting the Web pages based on their visual similarity after converting the pages into DOM trees. In their work, the authors used regular grammars with confidence instead of rigid formats to detect the patterns of postal addresses. The content of a text block is considered as a postal address if its confidence value exceeded a defined threshold. They also utilized lexicons such as Nation lexicon, state lexicon, city lexicon, street suffix lexicon and organization suffix lexicon to improve the extraction process.

Similar to their previous work, Asadi et al. (2008) used vision segmentation with a pattern-based approach to extract addresses from Web pages. The system converts the HTML tags into XML, and then analyzes the leaf nodes in the XML format. Additionally, an XML parser was designed to classify the XML tags into categories such as: NUM, Trigger, NOTICE, Cased Word, Preposition, Period Ended, and quotes Ended. The used patterns were manually chosen for recognition of addresses, and different confidence scores have been given/assigned to them. This system uses several address patterns and a small table of geographical knowledge to find addresses and itemize them into smaller components. The author concluded that a pure pattern based address extraction model cannot extract and itemize addresses from Web pages properly. This is due to the large variation in address patterns on the Web. Finding all address patterns is almost impossible. However, by adding a small table of general or coarse location names (e.g. country and state names) and using some triggers and keywords, the system provides better results both in the extraction and itemization of addresses. This combined model is cheaper and more flexible than gazetteer-based extraction approaches.

Two other approaches are proposed in Yu (2007). The first one used regular expressions, and the second one used a gazetteer. Both approaches assumed a formal format for the US addresses that begin with a Street number, followed by Street Name, City, State Name, optional ZIP code and Country Name. Subsequently, other address patterns that might occur were discarded. In his research, the author used a defined address boundary to detect the address block in the web page. The author assumed that a number in the text could be considered as a street number that remarks the beginning of a potential address, while state name and ZIP code are indicators of the end of an address. This assumption can cause several errors since Web pages may contain numbers signifying things other than street numbers. The first technique formed the rules using regular expressions, generated by the lexical scanner generator Flex Nicol (1993). Moreover, the longest matching text was marked as address. In the second technique, a gazetteer is used which involves the entire US nation with street names, address ranges, geographical codes, demographics of each side of road/street segments and latitude-longitude of each intersection. The author

Table 3 Pattern-Based Address Extraction Approaches

Ref.	Tech.	Gazetteer	Segm.	Web Content	Address Lang.	Extracted Entities
Cai et al. (2005b)	Reg. Grammars	Lexicons	VIPS Cai et al. (2004b)	Structured	English	Org., Street, City, State, Nation
Yu (2007)	Reg. Exp. Using Flex	NA	Segment Web page into tokens	Un-Structured	English US	PO Box, Street, State, Country
Yu (2007)	Gazetteer	TIGER/ Line Digital mapping data	Segment Web page into tokens	Structured	English US	PO Box, Street, State, Country
Asadi et al. (2008)	Geographical Database	-	HTML and VIPS (Cai et al. (2004b) Yu et al. (2004))	Structured	English Australia	Street, State, City, Country
Ahlers and Boll (2008a) Dirk (2013)	Gazetteer	Postleitzahl[6]	-	Un-Structured	German	City, CityZip, Street

[6] https://public.opendatasoft.com/explore/dataset/postleitzahlen-deutschland/table/

Table 4 Ontology-Based Address Extraction approaches

Ref.	Dataset	Lang. / Country	Extracted Entities
Cai et al. (2004c) Cai et al. (2005a)	DMTI GIS Database	English / Canada	**Suite info:** Apt, Building, Room Number. **Municipal Location:** St. No, St. Name, St. Type, Direction. **Geographical Position:** City, Province, Country.
Borges et al. (2007)	WBR05	Brazilian	**Basic address:** St. Type, St. Name, Building No**Location Identifier:** Postal Code, Phone No, and City/State.

used this gazetteer to construct two tables. The first is the Street table which contains all the unique streets, cities, states and ZIP codes. A unique index was assigned to each row in the table. The second one is the Street Index table that is used as an index of the Street table to expedite the street name searching.

Another study was introduced based on the idea of increasing the granularity level. The researchers used a full address information database instead of the normal gazetteers since the latter does not describe the address at the street level Ahlers and Boll (2008a, 2008b, 2007). The used database contains postal codes, city names, street names and every city-postal codes combination for each street in the target area. The designed geo-parser categorizes the full German addresses using a combined extraction and verification process on unstructured Web pages through the geographical database. The authors presented 16 combinations of the address elements that might appear in the Web pages. Meanwhile, the authors concluded that some combinations don't make any sense and can't be deduced, which minimizes the combinations to 5 valid patterns. This dataset consists of one city, nine postal codes, approximately 1,364 streets and 1,440 Postal code-street combinations. Furthermore, Dirk (2013) added a new component for a comprehensive analysis of the data in the index. In turn, this maintains the enrichment, duplicate detection, merging and aggregation of entities based on similarity analysis. One of the drawbacks of this method is that it requires high experience to construct rules. Moreover, this technique is mainly associated with a specific domain, which results in many problems when it is applied with a new domain. Otherwise, rules designed manually hold the benefits of detecting patterns that statistical algorithms cannot learn from given features.

3.2 Ontology-based address extraction approaches

Many researchers have focused on the conceptual analysis of unstructured text on the Web. On the other hand, ontologies (defined in Sect. 2.1.5) are considered as semantically rich as the conceptual schemas; therefore they are more relative to the user's intellectual model. Indeed, there are a few pieces of literature that utilize the ontology approach to extract the addresses from the Web pages. In this subsection, we shall both briefly address the proposed two ontology-based postal address extraction approaches and summarize these approaches in Table 4. The graph structures are used efficiently in representing human knowledge. Cai et al. (2004c, 2005a) utilized a predefined ontology combined with graph matching techniques to describe documents in order to extract location- based information. The proposed technique in Cai et al. (2005a) has three steps: (1) Ontology Construction,

(2) Identify Concepts and (3) Graph Matching. The first step defined the ontology as a set of concepts. This concept set is a gazetteer, in which each concept can be defined as a double/couple c = (inscription, meaning), where the inscription is corresponding to the lexical inscription form of the concept and meaning is the syntactic sense of the concept. The authors related the concepts together by means of their semantic relations. These relations are vertical and horizontal: (1) The PartOf relation, (2) The InstanceOf relation, (3) The Similar relation, (4) The SyntacticNeighbor relation. The similarity between two graphs is obtained through the function shown in Equation 1.

$$Sim(G, G_T) = \frac{N_C(G, G_T) + E_C(G, G_T)}{N(G_T) + E(G_T)} \tag{1}$$

Where $N_C(G, G_T)$ is the number of nodes shared by the text segment's abstract descriptor sub-graph G and the template graph G_T, $E_C(G, G_T)$ is the number of edges common in G and G_T. $N(G_T)$ is the number of nodes in the graph G_T and $E(G_T)$ is the number of edges in G_T.

Borges et al. (2007) ; Cai et al. (2005a) proposed an ontology-based approach named OnLocus that recognized and extracted geospatial evidence with local characteristics, such as street names and area codes. This work has been initiated based on the results obtained from Locus Souza et al. (2005) that are mainly constructed depending on the conceptual schema which is specified using the OMT-G model Borges et al. (2001). However, OnLocus was intended to elicit geographical knowledge from Web pages. Therefore, it concentrated on indirect references to places such as postal codes and telephone area codes Borges et al. (2007). The convincing performance of Locus as an indirect reference in geographical information retrieval tasks suggested that a gazetteer might be much more helpful if it could record the various types of relationships among places. Furthermore, using relationships recorded in the ontological gazetteer will enable the detection of the connection between many documents and places that are not explicitly mentioned in their text.

3.3 Probabilistic modeling and ML address extraction approaches

Supervised learning techniques have been utilized for classification and labeling purposes. These approaches are focused on labeling training data, designing features and selecting a suitable learning algorithm that distinguishes positive from negative entities by consuming these features. Utilizing these algorithms requires a training dataset employed to recognize and classify named entities, which should be previously annotated. The manual annotation of these data causes time-consuming and intensive labor. Further, selecting convenient features with good representations is a crucial task that influences the performance of a machine learning task. Several classifiers such as Naive Bayes, C4.5 Decision Trees and Support Vector Machine are widely exploited to give a defined label to an unknown input. Moreover, the Hidden Markov Model, Maximum-Entropy Markov Model (MEMM) and Conditional Random Fields are used to label a sequence of input segments. In this subsection, as well as we briefly discuss these ML approaches, we shall summarize them in Table 5.

Yu (2007) used the C4.5 classifier which is trained with a dataset that utilized features of the n-gram. It has four main steps: tokenization, feature extraction, classification and post-processing. The first step is introduced previously in Section 3.1. Further, in the second step, the system used five categories of the address features: Word Level, Geographical, Part-of-Speech Tagger, Punctuation and Layout features. For the geographical features,

Table 5 Probabilistic Modeling and ML-Based Address Extraction Approaches

	Token. Tool	Segm. Label	Web Content	Lang.	Learn. Model	Features
Yu (2007)	Lexical Scanner, Flex	Start, Middle, End, Other	Un-Structured	English	C4.5	Geographical, Word Level, POS Tagger, Punctuation, Layout features.
Chang and Li (2010)	ANNIE Annotation	Street suffix, State name, Direction, Country	Structured	English	SVM, CRF	All-caps, Allower, Initialcaps, Alldigits, Punctuation, Street, State, Direction, Phone, Zipcode, Annielocation
Chang et al. (2012)	Yahoo Chinese Word, Individual Chinese Character	Admin. division, Street, House No	Structured	Chinese	CRF	Country, City, Township, village, StreetRoad, House No, Building, Contacts Punctuation, ChineseNo, AllDigits.
Liu (2016)	NA	NA	Structured	English	GB, CRF	NA
Chang et al. (2016)	Yahoo Chinese Word, Individual Chinese Character	BIEO	Structured	Chinese	CRF	County, City, Township, Village, StreetRoad, HouseNo, Building, Contacts, Punctuation, ChineseNo, AllDigits.
Efremova et al. (2018)	Predefined splitting criteria	State, Postal Code,City, etc.	Structured	English	SVM	Geographical, Word Level, POS Tagger, Punctuation, Layout features.

a small size dictionary is used to define the following address components: (1) US State Name; (2) Street Direction; (3) Street Suffix; (4) Secondary Unit Designator; (5) USZIP and (6) POBOX. After the system processed the feature values for each n-gram, the C4.5 classifier (in the classification step) would label the tokens as: START, MIDDLE, END or OTHER. A sequence of these labeled tokens is identified as an address only if its first token was predicted as START by the classifier, while the last token was classified as END. In the final step, the post-processing step was applied to the labeled tokens to extract and output the addresses. At least one token in the middle of the block was predicted as MIDDLE. Furthermore, each block must include less than 20 tokens.

Chang and Li (2010) have introduced a Web service (Map-Maker) which accepts inputs (as Web pages) and extracts the existing postal addresses with their associated information to be ready for marking on a map. Two different machine learning models (SVM and CRFs) were developed and trained by the dataset introduced in Yu (2007). The motivation to use these approaches is that SVM and CRFs are the best known discriminative models for structured and unstructured learning, respectively. The basic concept of SVM is to find a hyperplane to separate two sets of data apart with maximum margin and minimum displacement. The SVM classifier was developed using libsvm with the polynomial kernel to decide the label of each token. Given a sequence of word/character tokens, the beginning/inside/end of an address is labeled by the following labels B/I/E and others by O. The SVM classifier is utilized here to predict the label for each word/token based on its context information. As the feature selection process has a vital influence on the address extraction, the authors applied 14 features to indicate the address block. In addition to this, a BIEO tagging method was exploited to label the tokens: where B stands for the beginning position of a postal address, I stands for the inside position of a postal address, E stands for the ending position of a postal address and O stands for outside a postal address. In addition, the authors extracted the associated information of the extracted address for the user's easier comprehension.

Chang et al. (2012) have developed the idea that is introduced in Chang and Li (2010), applied it to Chinese postal addresses extraction and improved the extraction of associated information. The DOM tree has been utilized to distinguish the address layout in a Web page. Three suffix types were used as indicators of the address segments (e.g., Administrative Division, Street and House Number). The used features were increased to be 17 features instead of 14 features which have been employed as indicators of address segments. The IO tagging method was added to the BIEO method to mark the address segment. In the IO tagging method, "I" stands for inside and "O" stands for outside the required address.

In Chang et al. (2016), the authors have tried to use the Web pages to extract the Chinese addresses which don't exist on the maps. This technique utilized the linear chained CRF algorithm. In addition to that, the authors examined the proposed technique with word segmentation and without word segmentation. The results showed that the segmentation doesn't improve the performance. Moreover, an unsupervised algorithm for associated information segmentation was presented by making use of a DOM tree structure based on the farthest distinguishable ancestor (FDA) of each address. The FDA algorithm could successfully detect the associated information for each Chinese address.

By considering another source of addresses, Liu (2016) used news reports instead of Web pages for address extraction. The problem with extracting addresses from the news reports is that no special formatting exists; so, the address or addresses may be anywhere in the text. In addition to that, there are no labeled news report datasets for training and testing. Creating these datasets requires manual labeling. Eventually, a single address may have various forms in the news reports. Furthermore, Liu (2016) introduced a comparison

between three supervised learning approaches which differ only in the classification algorithms. The used algorithms are Gradient Boosting (GB), Gradient Boosting with Principal Component Analysis (GB+PCA) and Conditional Random Fields. The author used two different sources to collect training data from the Washington Post and the "Cavalier Daily reports" in the period between 1990 and 2015. These data are combined with an unlabeled dataset of 998 articles. Nonetheless, the comparisons show that CRFs has the most stable performance on news reports.

Microdata has offered a great aid to understand the information on Web pages and provide more relevant results to users. It is considered as a collection of attributes that are embedded into standard HTML tags to clarify the data context. Efremova et al. (2018) benefited from Web pages on the Common Crawl dataset. That was available in a microdata format to train the SVM classifier in order to develop a geo-tagging framework that elicits addresses from Web pages. Microdata is a collection of attributes that are embedded into existing content on Web pages. The detected addresses ara composed of names and phones, as well as micro-contents such as streets, regions, postal codes, etc.

3.4 Hybrid postal address extraction approaches

Hybrid approaches use a combination of different techniques to extract postal addresses. Also, in this subsection, a brief discussion of the proposed hybrid approaches is presented and a summary of these approaches is shown in Table 6 as well. Yu (2007) introduced a hybrid method which combined the pattern-based and machine learning approaches. The system followed the same preprocessing techniques as in the previous methods. The combined approach profited from the handcrafted patterns that couldn't be created by the statistical techniques. Moreover, machine learning is used to learn complicated rules rather than developing them manually.

A combined approach was introduced by Borges et al. (2011) which implied ontology with urban gazetteers. The extended work included 17 different address patterns to verify which ones would be more useful for retrieval. Each pattern corresponds to a possible combination of address components in which addresses are usually found in the text. Moreover, in this work, they focused on extracting geographical knowledge from local Web business or service pages with the aim of providing a support to location-based services and integrating Web pages with urban locations. This meets the users' growing demand for such services, and has vast commercial, economic and social applications.

Schmidt et al. (2013) have extracted business German addresses from the Web. It combined patterns and gazetteers acquired from OpenStreetMap. The structure of the aimed address contains the name of a company to which the address belongs, the street name, the street number, the postal code and the city that the company is situated in. The proposed system contained three modules: Preprocessing, Identification of the address parts and Aggregation. The authors highlighted two significant issues that affect the system performance. Firstly, the uncommon structure of the company name can cause the detection of only a part of the name which will definitely influence the precision and the recall. Secondly, the improper assigning of the company name and the address. In many cases, a name is wrongly extracted to describe a company. This in turn has a negative effect on the precision.

Another hybrid approach is proposed in Nesi et al. (2014). It combined among the linguistic parsing, POS-tagging, pattern-based annotations Stab Christian (2017) and gazetteers that contain names of provinces and cities of the Tuscany region, as well as some of

Table 6 Hybrid Based Address Extraction approaches

Ref.	Used Tech.	Segm. Tech.	Token. Tool	Gazetteer	Web Content	Lang.
Yu (2007)	Pattern-based, ML	-	-	-	Structured	English
Borges et al. (2011)	Ontology-driven, gazetteers	NA	NA	WBR05	Semi-structured	Brazilian
Schmidt et al. (2013)	Patterns, gazetteers	Beautiful Soap[1]	Apache OpenNLP [2]	OpenStreetMap	Structured	German
Nesi et al. (2014)	Linguistic parsing, POS-tagging, pattern based annotations	ANNIE	JAPE	Annotated Gazetteers	Un-Structured	Italy

the POI places. The system architecture consisted of three modules: (1) A distributed Web crawler which aimed to fetch and mine the big textual data and the huge amounts of documents by using the Apache Nutch crawling tool, which has been integrated with Apache Solr for document indexing; (2) An Address Extractor which used a linguistic parser. This parser takes the documents and the pages retrieved from the previous module and analyzes it using the linguistic rules, POS-tagging, as well as through the use of external gazetteers. These gazetteers contain names of Italian cities, regions, companies, abbreviations for address items and identifiers using ANNIE Cunningham et al. (2002) (A Nearly-New Information Extraction system) and the pattern recognition algorithm JAPE (Java Annotation Patterns Engine). The best-case scenario for this module is to find the geographical information either in the HTML footer or head of each Web page. The worst case is to have to search through the rest of the page in case of not finding the required information neither in the footer nor the head; (3) A geocoding module which finally retrieved the coordinates of the extracted addresses or geographical information by querying a semantic Smart City repository created at DISIT Lab for the Sii-Mobility Project (proposed by Bellini et al. 2014).

In addition to the techniques mentioned earlier, a new branch of machine learning called Deep Learning (DL) has been broadly applied in geospatial studies since 2016. Unfortunately, deep learning is still not applied for predicting the label for each word of a postal address. None of the proposed deep learning based works aimed to identify the full address pattern. Deep learning has been employed to support the understanding of urban geography He et al. (2018), and processing the remote sensing images and street-view pictures Li and Hsu (2020). Further, Xu et al. (2020) have constructed a geospatial semantic address model based on bidirectional encoder representations from Transformers (BERT) to extract Chinese addresses' computational representations. Moreover, it predicts the location of the address characteristic expression of geographical address locations. They used the address coordinate prediction task as an example to show the workflow of the standard downstream task. This research's target was "building a transfer learning model based on the fine-tuning method". Moreover, Lin et al. (2020) proposed an address matching technique which is based on deep learning to identify the semantic similarity between address records. The used dataset was formed by 84,474 address pairs and the corresponding labels. The word-2vec model was trained to transform the address records into their corresponding vector representations. Further, the enhanced sequential inference model (ESIM) Fan et al. (2017) was applied to compute the semantic similarity of the compared address records to determine if two addresses are matched. Considering the previous deep learning results, it would introduce promising outcomes if applied in the domain of postal address extraction from the Web.

4 A comparative analysis

4.1 Gazetteers comparison

Quality plays an important role in working with all types of geo-data, particularly in data extraction and evaluation Goodchild (1992). Further, it influences the performance of common tasks such as geo-parsing and geocoding Acheson et al. (2017). Much of the work on geospatial data inaccuracy and uncertainty Devillers et al. (2010) are concerned with a positional mistake during geocoding or with positional accuracy and

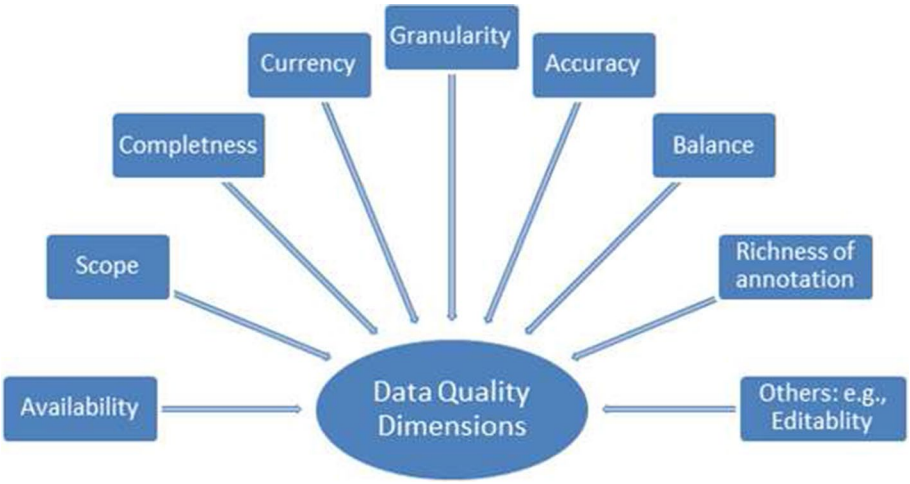

Fig. 1 Gazetteers Data Quality

prospective conflation methods of high granularity Ahlers and Boll (2009), Bakshi et al. (2005), which in turn influence the needs to verify the quality of data in the gazetteers. Data quality parameters have been varied in many researches. US Federal Geographical Data Committee listed the quality parameters to be: attribute accuracy, positional accuracy, logical consistency, completeness and lineage Guptill and Morrison (2013). Considering the context of the gazetteers, seven quality criteria have been defined by Leidner (2004) and extended later by Hill (2009). As shown in Figure 1, these criteria are defined as follows:

1. Availability: Degree to which a gazetteer is freely available and not limited by restrictive conditions of use.
2. Scope: Small communal database, regional/national coverage or worldwide coverage.
3. Completeness: Degree to which the scope of the gazetteer is covered completely.
4. Currency: Degree to which the gazetteer has incorporated changes.
5. Accuracy: Number of detectable errors in names, footprints and types.
6. Granularity: Includes large, well-known features only or features of all sizes and those that are less well known.
7. Balance: Uniform degree of detail, currency, accuracy and granularity across the scope of coverage.
8. Richness of annotation: Amount and detail of descriptive information, beyond the basics of name, footprint and type.
9. Moreover, one more criterion can be added since the essential dependence on gazetteers has shown the necessity of some features such as the editability; that can be defined as the degree to which the user can add/change attributes of the gazetteers.

In this subsection, a comparison among the existing gazetteers is established considering the granularity level of parameters, Editability, Accuracy, Availability, Scope and Currency. Table 7 shows the results of this comparison. Before that, quality of the gazetteers (presented before in Sect. 3) will be addressed as follows.

Table 7 Gazetteers Data Quality Dimensions

	Place	Neigh.	Land-mark	Island	River	Moun-tain	Lake	Co-ord.	Edit-able	Acc.	Avail-ability	Scope	Curr-ency
Geo Name	√	√	√	√	√	√	√	√	√	NA	Free	World wide	Daily
OpenStr-eetMap	√	√	√	√	√	√	√	√	√	NA	Free	World wide	NA
Uryupina (2002)	-	√	-	√	√	√	-	-	-	90%	NA	NA	Once
Souza et al. (2005)	√	√	√	√	√	√	√	-	-	77%	Braz. Cities	NA	NA
Machado et al. (2010)	-	-	-	-	-	-	-	-	-	NA	NA	Minas Gerais	NA
Chuang et al. (2016)	√	√	√	-	-	-	-	-	-	NA	NA	Taiwan	NA

GeoNames: It integrates data from multiple sources, while users can edit data in a wiki-like interface. This causes the variance in quality parameters such as scope, resolution or age. GeoNames data encompass the name, the coordinate (latitude and longitude), the parent administrative division and the country. Furthermore, the population data, height, alternative or translated names, or links to Wikipedia are organized in a hierarchy down from a country level. A proposed analysis by Ahlers (2013) indicates that there are different kinds of inaccuracies and partial error estimates. Moreover, it has been proved that there are shortened coordinates in very large numbers apart from other problems. In many cases, some areas have an extent that is not captured at the sub-minute scale. Giving a measurement degree to the accuracy parameter would be beneficial but it isn't applicable in the case of large gazetteers such as GeoNames.

OpenStreetMap (OSM): Numerous scientific studies were conducted to evaluate the data quality of OSM such as Haklay (2010), Helbich et al. (2012) Touya (2010), Zielstra and Zipf (2010). All these studies show the advantage of responsiveness and flexibility of OSM. Further, the results have been demonstrated high positional accuracy that highly depends on the data collection technique and an enormous amount of details that found around urban areas with a high number of contributors. Moreover, several factors such as GPS signal preciseness displaced aerial images or bulk movements proved impressive data quality. On the other hand, it showed also the problematic aspect of heterogeneity in OSM data, highly limiting the possible applications. This heterogeneity is particularly explained by the coexistence of different data sources Zhang et al. (2018), processes of capture and contributors' profiles, highlighting the importance of the followed accepted and well-defined specifications.

In Uryupina (2002), different entities such as City, Region, Country, Island, River and Mountain were supported. Furthermore, it didn't focus on a certain region. The system relied on running the queries for one time without any features to edit the obtained results such as the case of the previous gazetteers.

Locus: Souza et al. (2005) considered the place names and lakes entities in addition to the spatial relationships among the entities. The maximum accuracy of the two systems are 98% and 77% respectively. In addition to that, both researches didn't mention either the ability to update the dataset periodically or the ability to use the resulted gazetteer. Ontology-based Gazetteer: Machado et al. (2010) added the records concepts and terms related to a place to the structure used in the conventional gazetteers. Unfortunately, this work didn't provide any quality measurement to the proposed gazetteer. POI Gazetteer: Chuang et al. (2016) have evaluated the constructed POI database by comparing it with both Wikimapia and 'What's the Number?'. Wikimapia is an open-content collaborative mapping project that aims to mark and describe all geographical objects in the world, while 'What's the Number' is a popular app for smartphones that provides a telephone and address lookup service. The results show that the performance of the proposed technique exceeds the others in terms of the two kinds of queries (common keywords or POI names) and regardless of the area (whether urban and rural). This work treated with the address as a whole pattern (i.e. it didn't give details about the composition of the extracted addresses). Furthermore, the authors' evaluation discarded the data quality parameters such as Availability, Currency and Accuracy.

4.2 Technique-based comparison

In this subsection, we compare the performance of the address extraction approaches discussed in Section 3. The criteria used in the comparison rely on the obtained results

Table 8 Comparison of the results obtained from pattern-based approaches (with gazetteers)

Approach	Prec.	Recall	F-measure	#Web Pages	#Addresses	#Patterns
Cai et al. (2005b)	-	-	-	44	56	5
Yu (2007) Regular Expression	0.73	0.61	0.66	471	1370	6
Yu (2007) Gazetteers	0.83	0.69	0.75	471	1555	6
Asadi et al. (2008)	0.97	0.73	0.83	1,100	2,030	10
Ahlers and Boll (2008a)	-	-	-	180,000	25,000	5

and the volume of the used data. These criteria are Precision, Recall, F-measure, #Pages, #Addresses and #Patterns. The criteria are defined in the research literature as follows.

- $Precision = \frac{N_{correct}}{N_{response}}$; where $N_{correct}$ is the number of the correct addresses in the system and $N_{response}$ is the total number of addresses candidates
- $Recall = \frac{N_{correct}}{N_{key}}$; where N_{key} is the total number of real addresses in the answer key.
- F-measure is the combination of the two metrics scores recall and precision.
- #Pages: the total number of used Web pages.
- #Addresses: the total number of the extracted addresses.
- #Patterns: the number of the designed address patterns.

4.2.1 Pattern-based approaches comparison

Table 8 shows the results of comparing pattern-based address extraction approaches using the above mentioned criteria. The details of this comparison are discussed as follows.

Results in Cai et al. (2005b) are obtained by submitting a query to the Google search engine with the subject "contact". The examined pages were only the first 50 returned pages, which were examined manually to check whether they contained addresses or not. Six of the pages were not allowed to be accessed, sixteen pages had addresses and the remainder contained no addresses. This approach was evaluated regarded to the precision of the segmentation process. It revealed that: among 56 addresses, one address was not segmented, 7 were segmented as larger blocks, 5 were segmented as smaller blocks and 43 were segmented well. Therefore, the total accuracy was 0.89 and the false alarm rate was 3.8%. Unfortunately, the test dataset (mentioned by the author at http://idke.ruc.edu.cn/wdml/address-truth.zip) could not be accessed. This approach didn't test using a large dataset. Moreover, some address parts were neglected such as: telephone number, product price and product description.

A dataset in Yu (2007) is constructed (using a regular expression) by querying Google with query sets of 3 main subjects: contacts, hotels and pizza restaurants. The first, "Contact Collection", was collected using the two queries: "contact us" and "contact information". The second, "Hotel Collection", was collected using the queries "Hotel Los Angeles", "Hotel San Francisco", "Hotel New York" and "Hotel Seattle". The last, "Pizza Collection", was collected using the queries "Pizza Los Angeles", "Pizza San Francisco", "Pizza New York" and "Pizza Seattle". The queries resulted in 2,375 Web pages and 12,895 US addresses. Only 20% of the total Web pages were chosen randomly and utilized as a testing set. The testing dataset consists of 471 Web pages with 2,257 labeled addresses. The regular expression system successfully detected 1,370 addresses. The F-Measure for exact

matching using the regular expression was 0.665 with a precision of 0.735 and a recall of 0.607. However, the results were improved (high recall and precision) by using a gazetteer. The system achieved precision of 83.1% and recall of 68.9%. The enhancement in precision was due to the validation of address elements with gazetteer which covered all US cities and street names. The system avoids many false positives; therefore, it improves the precision. However, the authors reasoning the low recall percentage due to the used assumption which considered the address starting with a street number or PO Box, and ending with a state name plus ZIP code. This hypothesis leads to the dismissal of addresses that do not precisely follow these defined rules. By comparing the two proposed approaches, it is evident that the performance of the regular-expression approach is lower than that of the gazetteer-based approach. On the other hand, the first approach did not rely on any database, while the second approach demanded a dataset with every street and city in a country.

A manually constructed dataset is used in Asadi et al. (2008). The designed patterns were based on the most repeated 10 Australian styles that are appeared in the collected Web pages. The testing scenarios run through many cases. The first scenario used only the designed patterns to extract addresses. The second one combined the designed addresses with a small geo-graphic table containing triggers, keywords, country names, Australian states and major cities. The results revealed that: in the first test case scenario, the recall was 65%, and was raised to 73% in the second scenario. This indicates that the pattern based approach can't extract all the patterns that exist in the Web pages. By doing so, the authors achieved better results than a pure pattern based approach. Furthermore, it was found that this result can be enhanced through adding the geographical table. Moreover,a recall was found to be 0.73, precision to be 0.97 and F-measure to be 0.83 for the complete address. This system didn't provide any usable dataset for other researchers.

Ahlers and Boll (2008a) have presented 16 combinations of the address elements that might appear in Web pages. Meanwhile, they concluded that some combinations didn't make any sense and cannot be detected. So, they minimized the combinations to 5 valid patterns. This dataset consists of one city, 9 postal codes, 1,364 streets and 1,440 Postal code-street combinations. The authors restored approximately 180,000 Web pages and about 25,000 addresses that are coincided with the definition of a complete address. This is equivalent to a result of approximately 13% of location-aware Web pages. The obtained results showed a very high precision. However, recall wasn't recorded due to the difficultly of measuring the numerous relevant documents.

4.2.2 Ontology-based approaches comparison

Two Web page representations were used in Cai et al. (2005a) Borges et al. (2007). The first described the Web pages as a subgraph of a predefined ontology while the second converted the HTML pages to plain text. Cai et al. (2005a) didn't show a high performance according to the difficulty of the graph matching. The experimental results revealed 74.5% for precision, 72.4% for recall and 73.34% for F-Measure at 0.29 for similarity threshold. Moreover, the problem of typing-errors in address identification, noise elimination and separation of joined address phrases still unsolved. On the other hand, Borges et al. (2007) evaluated the proposed work using the False Positive measurement while precision and recall were not measured. The authors also examined each pattern of the address separately. The system showed the highest performance in detecting postal codes, followed by the pattern: street type + street name + building number + postal code. This is because the more conventional syntax of the address patterns that appears in the text will increase the ability

Table 9 Comparison of the results obtained from the Ontology-Based Approaches

Approach	Prec.	Recall	F-measure	#Web Pages	#Addresses	#Patterns
Cai et al. (2005a)	0.74	0.72	0.73	11	105	NA
Borges et al. (2007)	-	-	-	43,121	893,260	11

to recognize the address pattern. As can be observed, the number of extracted addresses by Borges et al. (2007) is significantly greater than that of the other proposals. This is due to the calculation of each element of the address pattern separately. For instance, the extracted postal code was calculated as a pattern even if there is no other address information extracted. Table 9 shows the details of comparing Ontology-Based Approaches using the above mentioned criteria.

4.2.3 Probabilistic modeling and ML-based approaches comparison

The machine-learning-based system exceeded the precision and the recall of the two rule-based systems introduced previously in Yu (2007), while the regular expression technique was limited in its coverage of postal address formats. Meanwhile, the Gazetteer contains a large amount of information for matching and extracting. Furthermore, it's shown that the performance of the address extraction task was higher while using all features rather than each feature separately. The system also demonstrated that precision was inversely affected by the number of n-grams. The rising order of n-gram gradually decreased the precision. The peak F-measure was 0.843 while the highest order of n-gram was 8. Using the same dataset, the performance of the proposed approaches in Chang and Li (2010) has shown a rising trend as the F-score for SVM increased by 2.7%, compared to the F-score of the C4.5 Decision tree Yu (2007). Meanwhile, the F-score for CRF has the highest result of 0.914 with an increase of 3.8%. These results confirm that the SVM and CRF have a superior learning capability than C4.5. Further, associated information for each address is also identified based on the clustering of the addresses into address blocks. The accuracy of associated information extraction was measured with and without adjustment procedure. The experimental result shows that the accuracy is 0.851 without adjustment which is enhanced by 1.5%.

The sequences labeling technique was applied later rather than the classification technique to the Chinese postal address extraction using both BIEO and IO tagging methods in Chang et al. (2012). The performance was measured with and without Yahoo Chinese word segmentation and the results demonstrated that the best outcomes can be obtained using a conditional random field with BIEO tagging without word segmentation. The proposed work emphasized that improper segmentation can lead to worse labeling of address tokens. Moreover, the F-measure was increased from 0.90 to 0.92. Furthermore, the FDA algorithm reported a higher F-measure (from 0.811 to 0.96) when applied to associated information for Chinese addresses Chang et al. (2016).

The SVM classifier also presented high results in Efremova et al. (2018). The authors didn't mention the number of Web pages or the extracted addresses. Instead, they only mentioned the number of used annotated data partitions. This approach seems limited as it depends on the microdata tags, while according to the W3Techs survey[6], only around

[6] https://w3techs.com/

Table 10 Comparison of the results obtained from the ML-Based Approaches

Approach	Prec.	Recall	F-measure	#Web Pages	#Addresses
Yu (2007) C4.5 Decision Tree	0.94	0.72	0.81	471	2,257
Chang and Li (2010) SVM	0.96	0.85	0.90	1,740	8,519
Chang and Li (2010) CRF	0.97	0.86	0.91	1,740	8,519
Chang et al. (2012) BIEO + ICCS	0.97	0.97	0.97	549	3,896
Chang et al. (2012) BIEO + YCWS	0.97	0.96	0.96	549	3,896
Chang et al. (2012) IO + ICCS	0.95	0.95	0.95	549	3,896
Chang et al. (2012) IO + YCWS	0.94	0.93	0.94	549	3,896
Chang et al. (2012) Regular Expression	0.88	0.90	0.89	549	3,896
Chang et al. (2016) BIEO + ICCS	0.97	0.969	0.97	549	3,896
Chang et al. (2016) BIEO + YCWS	0.97	0.958	0.96	549	3,896
Chang et al. (2016) IO + ICCS	0.95	0.946	0.949	549	3,896
Chang et al. (2016) IO + YCWS	0.95	0.93	0.94	549	3,896
Chang et al. (2016) Regular Expression	0.88	0.90	0.89	549	3,896
Efremova et al. (2018) SVM	0.91	0.928	-	-	-

Table 11 Comparison of the results obtained from the Hybrid-Based Approaches

Approach	Prec.	Recall	F-measure	#Web Pages	#Addresses
Yu (2007) ML+Reg. Exp.	0.952	0.811	0.876	471	2,257
Yu (2007) ML+Gazetteer	0.943	0.784	0.856	471	2,257
Borges et al. (2011)	-	-	-	603,798	2,137,601
Schmidt et al. (2013)	0.61	0.80	0.69	1,576	4,449
Nesi et al. (2014)	0.90	0.93	0.92	100,000	-

12.7% Web pages have microdata tags compared to the entire Web. However, this represents a very small percentage since not all of these pages contain addresses. Table 10 shows the details of comparing ML-Based Approaches.

4.2.4 Hybrid approaches comparison

As can be observed, the first two hybrid systems proposed in Yu (2007), which integrated the machine-learning technique with the regular expression technique and the machine-learning technique with the gazetteer have exceeded the precision of the other three hybrid systems by achieving 0.95 and 0.94 respectively. While the other systems used patterns with the gazetteers technique and Part-Of-Speech-tagging with patterns, they have demonstrated a lower precision. On the other hand, Nesi et al. (2014) showed higher performance in both recall and F-measure by scoring 90.5% and 92.8%, respectively. Schmidt et al. (2013) identified 4,449 addresses (an average of 2.8 addresses per Web site) and achieved a moderate results in all the measurements.

Borges et al. (2011) have the highest number of extracted addresses because they considered the postal code and the city name as a separate address. The postal code pattern and city name pattern have been extracted 1,083,913 and 470,879 times,

respectively. However, the authors pretended that the geocoder's performance has a sustainable relevance with the quality of addresses data. The extracted patterns were geocoded using the techniques introduced in Davis and Fonseca (2007), while confirming the precision of location detection according to the designed patterns. Table 11 represents the obtained results from these proposals.

5 Conclusion and future work

This survey intended to review the different approaches introduced to extract postal addresses from the Web. Our review focused on two main directions. The first direction has investigated the data quality of the existing gazetteers, which has found that GeoNames and OSM had the highest affirmative data quality. The second direction has addressed the challenges facing the problem of postal addresses extraction. It has examined the techniques employed to handle these challenges. This direction has explored that many issues (such as the ambiguous and dynamic nature of a location name, the various styles of the address and the different sources of addresses on the Web) influence the postal address extraction process. Moreover, the semantic misunderstanding consequences of the synonyms (different words with the same meaning) and polysemy (the same word with different meanings) still present an obstacle in many used languages which consequently affects the performance of the postal address extraction process.

We concluded that, using one or more specific Web sites as fixed data sources would limit the data diversification and affect the comprehensiveness of geographical data. Therefore, covering real Web pages instead of relying on Yellow Pages would increase the obtained geographical knowledge. Additionally, more attention should be directed towards social networks as they are vital dynamic sources of geographical data. Social networks are prolific sources to detect locations for geospatial applications. They represent a fertile source of real-world events, especially in times of mass emergencies. Also, there are still promising models such as "Deep Learning (DL)" that have not been strongly applied in this area, despite the impressive results of DL in other fields related to the domain of postal address extraction from the Web. Furthermore, it's noticed that the comparisons among the existing approaches don't exist due to a lack of public geographical datasets. Therefore, collecting a public dataset that can be used in exploring new hypotheses and validate methods is needed, which subsequently could increase the efficiency and quality of the research in this domain. Moreover, designing geographical gazetteers from scratch consumes time and effort. Most of the systems that depend on hand-designed gazetteers had a narrow coverage scope, which could be expanded by using the open-accessed geographical databases (e.g., GeoNames, OSM, etc.).

In the future, further research in this domain should propose a multilingual framework that aims to extract geographical data from social media networks to overcome the absence of many POIs on the current location-based services. On the other hand, location-based service providers (such as Google Maps, Wikimapia, Garmin and Yelp) should focus on tracking the percentages of uncovered POIs and at the same time providing some tools that seek to compensate this lack of POIs from the customers themselves.

References

Acheson E, De Sabbata S, Purves RS (2017) A quantitative analysis of global gazetteers: patterns of coverage for common feature types. Comput Environ Urban Syst 64:309–320. https://doi.org/10.1016/j.compenvurbsys.2017.03.007

Ahlers D (2013, November). Assessment of the accuracy of GeoNames gazetteer data. In Proceedings of the 7th workshop on geographic information retrieval (pp. 74-81). ACM

Dirk Ahlers (2013) Business entity retrieval and data provision for yellow pages by local search. In IRPS Workshop (ECIR2013)

Ahlers D, Boll S (2009), On the accuracy of online geocoders. Geoinformatik

Ahlers D, Boll S (2008), Retrieving address-based locations from the Web. In Proceeding of the 2nd international workshop on geographic information retrieval - GIR '08, 27. https://doi.org/10.1145/1460007.1460015

Ahlers D, Boll S (2008), Urban Web Crawling. First international workshop on location and the web (LocWeb 2008), 25–32. https://doi.org/10.1145/1367798.1367803

Ahlers D, Boll S (2007) Location-based web Search. The Geospatial Web. https://doi.org/10.1007/978-1-84628-827-2_6

Popescu AM, Nguyen B, Etzioni O (2005) OPINE: Extracting Product Features and Opinions from Reviews. Proc. of the HLT/EMNLP (2005) Human language technology conference and conference on empirical methods in natural language processing, 6–8 October 2005. Vancouver, British Columbia, Canada, pp 32–33

Asadi S, Yang G, Zhou X, Shi Y, Zhai B, Jiang WWR (2008) Pattern-based extraction of addresses from Web page content. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-540-78849-2_41

Babych B, Hartley A, (2003) Improving machine translation quality with automatic named entity recognition. In proceedings of the 7th international EAMT workshop on MT and other language technology tools, improving MT through other language technology tools resources and tools for building MT - EAMT '03 (pp. 1–8), Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1609822.1609823

Bakshi R, Knoblock CA, Thakkar S (2005) Exploiting online sources to accurately geocode addresses. Proceedings of the 12th annual ACM international workshop on geographic information systems, 194, https://doi.org/10.1145/1032222.1032251

Baralis E, Cagliero L, Jabeen S, Fiori A, Shah S (2013) Multi-document summarization based on the Yago ontology. Expert Syst Appl 40(17):6976–6984. https://doi.org/10.1016/j.eswa.2013.06.047

Bellini P, Benigni M, Billero R, Nesi P, Rauch N (2014) Ontology construction and knowledge base feeding and cleaning for smart-city services. IEEE 19 Int. Conf. on Engineering of complex computer systems (ICECCS 2014)

Benajiba Y, Rosso P, Diab M (2009) Arabic named entity recognition: a feature-driven study. IEEE Trans Audio Speech Language Process 17(5):926–934. https://doi.org/10.1109/TASL.2009.2019927

Betina Antony J, Mahalakshmi GS (2015) Content-based information retrieval by named entity recognition and verb semantic role labelling. J Univ Comput Sci 21(13):1830–1848

Bikel DM, Miller S, Schwartz R, Weischedel R (1997) Nymble: a high-performance learning name-finder. Proceedings of the fifth conference on applied natural language processing. https://doi.org/10.3115/974557.974586

Borges KAV, Davis CA, Laender AHF (2001) OMT-G: an object-oriented data model for geographic applications. GeoInformatica 5(3):221–260. https://doi.org/10.1023/A:1011482030093

Borges KAVV, Laender AHFF, Medeiros CBand Davis Jr., Ca (2007), Discovering geographic locations in Web pages using urban addresses. GIR '07 proceedings of the 4th ACM workshop on geographical information retrieval, 31–36. https://doi.org/10.1145/1316948.1316957

Borges KAV, Davis CA, Laender AHFand Medeiros CB, (2011) Ontology-driven discovery of geospatial evidence in web pages. GeoInformatica 15(4):609–631. https://doi.org/10.1007/s10707-010-0118-z

Borkar V, Deshmukh K, Sarawagi S (2001) Automatic segmentation of text into structured records. ACM SIGMOD Record 30(2):175–186. https://doi.org/10.1145/376284.375682

Borthwick A, Sterling J, Agichtein E, Grishman R (1998) Exploiting diverse knowledge sources via maximum entropy in named entity recognition. Proceedings of the 6th workshop on very large Corpora, 152–160. http://acl.ldc.upenn.edu/W/W98/W98-1118.pdf

Cai D, Yu S, Wen JR, Ma WY (2003) Vips: a vision-based page segmentation algorithm. Technical Report, MSR-TR-2003-79. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2003-79.pdf

Cai D, He X, Wen JR, Ma WY (2004, July). Block-level link analysis. In Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (pp. 440-447). https://doi.org/10.1145/1008992.1009068

Cai D, Yu S, Wen J-R, Ma W-Y (2004) Block-based Web search. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, 456–463. https://doi.org/10.1145/1008992.1009070

Cai WT, Wang SR, Jiang QS (2005) Address extraction: extraction of location-based information from the web. Web Technol Res Dev - Apweb 3399:925–937

Cai W, Wang S, Jiang Q (2004) Address extraction: a graph matching and ontology-based approach to conceptual information retrieval. Proceedings of international conference on machine learning and cybernetics. https://doi.org/10.1109/ICMLC.2004.1382024

Can L, Qian Z, Xiaofeng M, Wenyin L (2005) Postal address detection from web documents. International workshop on challenges in web information retrieval and integration, 40–45. http://dl.acm.org/citation.cfm?id=1105926.1106228

Cavedon IL, Zukerman I, Moll D, Zaanen M Van, Smith D, (2006) Named entity recognition for question answering. Proc. of the (2006) Australasian language technology workshop 2006, November 30-December 1, 2006. Sancta Sophia College. Sydney. Australasian Language Technology Association, Carlton, Vic, pp 51–58

Cao J, Mao B, Luo J (2010) A segmentation method for web page analysis using shrinking and dividing. Int J Parallel Emerg Distributed Syst 25(2):93–104. https://doi.org/10.1080/17445760802429585

Chakrabarti D, Kumar R, Punera K (2008, April) A graph-theoretic approach to webpage segmentation. In Proceedings of the 17th international conference on World Wide Web (pp. 377-386). https://doi.org/10.1145/1367497.1367549

Chang CH, Li SY (2010), MapMarker: Extraction of postal addresses and associated information for general Web pages. Proceedings - 2010 IEEE/WIC/ACM international conference on web intelligence, WI 2010, 1, 105–111. https://doi.org/10.1109/WI-IAT.2010.64

Chang C-H, Huang C-Y, Su Y-S (2012) On Chinese postal address and associated information extraction. The 26th annual conference of the Japanese society for artificial intelligence

Chieu HL, Ng HT (2002) Named entity recognition: a maximum entropy approach using global information. Coling '02, 1, 1–7. https://doi.org/10.3115/1072228.1072253

Chinchor N, Robinson P (1997, September). MUC-7 named entity task definition. In proceedings of the 7th conference on message understanding (Vol. 29, pp. 1-21)

Chiu JPC, Nichols E (2016) Named entity recognition with bidirectional LSTM-CNNs, transactions of the association for. Comput Linguist 4(2003):357–370. https://doi.org/10.3115/1119176.1119204

Choi FY, Wiemer-Hastings P, Moore JD (2001) Latent semantic analysis for text segmentation. In Proceedings of the 2001 conference on empirical methods in natural language processing

Chu WW (2013) Erratum: data mining and knowledge discovery for big data. Data mining and knowledge discovery for big data pp 305-308. https://doi.org/10.1007/978-3-642-40837-3_10

Chang C-H, Kayed M, Girgis MR, Shaalan KF (2006) A survey of web information extraction systems, IEEE transactions on knowledge and data engineering, 18(10): pp. 1411–1428. https://ieeexplore.ieee.org/document/1683775

Chang C-H, Chuang HM, Huang CY, Su YS, Li SY (2016) Enhancing POI search on maps via online address extraction and associated information segmentation. Appl Intell 44(3):539–556. https://doi.org/10.1007/s10489-015-0707-5

Chuang H-M, Chang C-H, Kao T-Y (2014) Effective web crawling for chinese addresses and associated information. Int Conf Electron Commerce Web Technol. https://doi.org/10.1007/978-3-319-10491-1_2

Chuang H, Chang C, Kao T, Cheng C, Cheong K (2016) Enabling maps/location searches on mobile devices- constructing a POI database via focused crawling and information extraction. Int J Geogr Inform Sci 30(7):1405–1425. https://doi.org/10.1080/13658816.2015.1133820

Cunningham H, Maynard D, Bontcheva K, ACL VT (2002) GATE: A framework and graphical development environment for robust NLP tools and applications. Proceedings of the 40th annual meeting of the association for computational linguistics, July 6-12, 2002, Philadelphia, PA, USA. http://www.aclweb.org/anthology/P/P02/P02-1022.pdf

Davis CA, Fonseca FT (2007) Assessing the certainty of locations produced by an address geocoding system. GeoInformatica 11(1):103–129. https://doi.org/10.1007/s10707-006-0015-7

Devillers R, Stein A, Bédard Y, Chrisman N, Fisher P, Shi W (2010) Thirty years of research on satial data quality achievements, failures, and opportunities. Trans GIS 14(4):387–400. https://doi.org/10.1111/j.1467-9671.2010.01212.x

Ding R, Chen Z (2018) RecNet: a deep neural network for personalized POI recommendation in location-based social networks. Int J Geogr Inform Sci 32(8):1631–48

Du L, Pate JK, Johnson M (2015, February). Topic segmentation with an ordering-based topic model. In 29th AAAI conference on artificial intelligence

Efremova J, Endres I, Vidas I, Melnik O (2018, July) A geo-tagging framework for address extraction from Web pages. In industrial conference on data mining (pp. 288-295)

Ekbal A, Bandyopadhyay S (2010) Named entity recognition using support vector machine a language independent approach. Int J Electr Comput Eng 4(3):155–170

Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Yates A (2005) Unsupervised named-entity extraction from the Web: an experimental study. Artif Intell 165(1):91–134. https://doi.org/10.1016/j.artint.2005.03.001

Faloutsos C, Oard DW (1998) A survey of information retrieval and filtering methods. A J Comp Educ, 1–24. http://drum.lib.umd.edu/handle/1903/436

Fan Y, Pang L, Hou J, Guo J, Lan Y, Cheng X. Matchzoo: A toolkit for deep text matching. arXiv preprint arXiv:1707.07270. 2017 Jul 23

Freihat AA, Bella G, Mubarak H, Giunchiglia F (2018) A single-model approach for Arabic segmentation, POS tagging, and named entity recognition. The 2nd International conference on natural language and speech processing. ICNLSP 2018:1–8. https://doi.org/10.1109/ICNLSP.2018.8374393

Freitag D, McCallum AK (1999) Information extraction using HMMs and shrinkage. AAAI99 workshop on machine learning for information extraction, 31–36. https://doi.org/10.1017/CBO9781107415324.004

Gao S, Li L, Li W, Janowicz K, Zhang Y (2014) Computers, environment and urban systems Cconstructing gazetteers from volunteered Big geo-data based on Hadoop. Comput Environ Urban Syst. https://doi.org/10.1016/j.compenvurbsys.2014.02.004

Goodchild MF (1992). Geographical data modeling. Computers Geosciences, 401–408. https://www.sciencedirect.com/science/article/pii/0098300492900694

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221. https://doi.org/10.1007/s10708-007-9111-y

Guo J, Xu G, Cheng X, Li H (2009) Named Entity Recognition in Query. Proceedings of the 32nd International ACM SIGIR conference on research and development in information retrieval - SIGIR '09, 267. https://doi.org/10.1145/1571941.1571989

Gupta S, Kaiser G, Neistadt D, Grimm P (2003, May) DOM-based content extraction of HTML documents. In proceedings of the 12th international conference on World Wide Web (pp. 207-214)

Guptill SC, Morrison JL (2013) Elements of spatial data quality

Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. Environ Plan B: Plan Des 37(4):682–703. https://doi.org/10.1068/b35097

Han AL, Wong DF, Chao LS (2013) Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. Lang Process Intell Inform Syst. https://doi.org/10.1007/978-3-642-38634-3_8

Hattori G, Hoashi K, Matsumoto K, Sugaya F (2007, May) Robust web page segmentation for mobile terminal using content-distances and page layout information. In Proceedings of the 16th international conference on World Wide Web (pp. 361-370)

He J, Li X, Yao Y, Hong Y, Jinbao Z (2018) Mining transition rules of cellular automata for simulating urban expansion by using the deep learning techniques. Int J Geogr Inform Sci 32(10):2076–97

Helbich M, Amelunxen C, Neis P, Zipf A, (2012) Comparative spatial analysis of positional accuracy of openStreetMap and proprietary geodata. Proceedings of GI_Forum, 24–33 http://gispoint.de/fileadmin/user_upload/paper_gis_open/537521013.pdf

Hill LL (2009) Georeferencing: The geographic associations of information. Mit Press

Hill LL, (2000) Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. The 4th European Conference, ECDL, (2000) Lisbon. Portugal. https://doi.org/10.1007/3-540-45268-0_26

Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv 2015. http://arxiv.org/abs/1508.01991

Hui N, Hua Y, Ya-zhou T, Hao W (2009) A method of Chinese named entity recognition based on maximum entropy model. Mechatronics and automation, 2009. IEEE conference on mechatronics and automation, 2472–2477, https://doi.org/10.1109/ICMA.2009.5246408

Jones CB, Purves RS (2008) Geographical information retrieval. Int J Geogr Inform Sci 22(3):219–228. https://doi.org/10.1080/13658810701626343

Kayed M, Chang C-H (2010) FiVaTech: Page-Level web data extraction from template pages, IEEE Transaction on knowledge and data Eng., vol. 22, no. 2, pp. 249-263, https://ieeexplore.ieee.org/document/4476640/

Kohlschütter C, Nejdl W (2008, October). A densitometric approach to web page segmentation. In Proceedings of the 17th ACM conference on Information and knowledge management, 1173-1182

Kokla M, Papadias V, Tomai E. Enrichment and population of a geospatial ontology for semantic information extraction. International archives of the photogrammetry, remote sensing and spatial information sciences. 2018 Sep 19;42(4)

Kovacevic M, Diligenti M, Gori M, Milutinovic V (2002, December). Recognition of common areas in a web page using visual information: a possible application in a page classification. In 2002 IEEE international conference on data mining, 2002. Proceedings. (pp. 250-257). IEEE

Kuhn W (2005) Geospatial semantics: why, of what, and how?. InJournal on data semantics III 2005 (pp. 1-24). Springer, Berlin, Heidelberg

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural Architectures for Named Entity Recognition. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies arxiv: abs/1603.01360

Larson RR, Frontiera P Geographic (1996) Information Retrieval and Spatial Browsing. 32nd Clinic on library applications of data processing, (January 1995), 81–124 https://doi.org/10.1145/100892.1009143

Leidner JL (2004) Towards a reference corpus for automatic toponym resolution evaluation. Workshop on geographic information retrieval, Sheffield, Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner 20(2):22–23

Li W, Hsu CY (2020) Automated terrain feature identification from remote sensing imagery: a deep learning approach. International Journal of Geographical Information Science (**2;34(4):637–60**)

Lin Y, Kang M, Wu Y, Du Q, Liu T (2020) A deep learning architecture for semantic address matching. Int J Geogr Inform Sci 34(3):559–76

Liu Y, Liu W, Jiang C (2004, July) User interest detection on web pages for building personalized information agent. In International conference on web-age information management (pp. 280-290). Springer, Berlin, Heidelberg

Liu X (2016) Extracting Addresses From News Reports Using Conditional Random Fields. 15th IEEE International conference on machine learning and applications (ICMLA) https://doi.org/10.1109/ICMLA.2016.94

Ma X, Hovy E, (2016) End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Proceedings of the 54th annual meeting of the association for computational linguistics arxiv: abs/1603.01354

Machado IMR, Alencar RO De, Oliveira R De, Junior C, Junior CAD (2010) An Ontological Gazetteer for geographic information retrieval. Proceedingd XI GEOINFO, Campos Do Jordao, Brazil, (Hill 2000), 21–32

Mandl T, Womser-Hacker C (2005) The effect of named entities on effectiveness in cross-language information retrieval evaluation. Proceedings of the 2005 ACM symposium on Applied computing https://doi.org/10.1145/1066677.1066919

McCallum A (2002) Efficiently inducing features of conditional random fields. Proceeding UAI'03 proceedings of the nineteenth conference on uncertainty in artificial intelligence, 19(July), 168–175, . https://dl.acm.org/citation.cfm?id=2100633

Misra H, Yvon F, Cappé O, Jose J (2011) Text segmentation: a topic modeling perspective. Inform Process Manag 47(4):528–544

Morimoto Y, Houle ME, Mccurley KS, Road H, Jose S, Extracting spatial knowledge from the web. In 2003 symposium on applications and the Internet, pp. 326-333 (2003). https://doi.org/10.1109/SAINT.2003.1183066

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26. https://doi.org/10.1075/li.30.1.03nad

Nagabhushan P, Angadi S, Anami B (2006) A fuzzy symbolic inference system for postal address component extraction and labelling. Fuzzy Syst Knowl Discov. https://doi.org/10.1007/11881599

Nesi P, Pantaleo G, Tenti M (2014) Ge(o)Lo(cator): Geographic Information Extraction from Unstructured Text Data and Web Documents. 9th International Workshop on Semantic and Social Media Adaptation and Personalization https://doi.org/10.1109/SMAP.2014.27

Nicol GT (1993) Flex: the lexical scanner generator. Free Software Foundation

Nobata C, Sekine S, Isahara H, Grishman R (2002) Summarization system integrated with named entity tdagging and IE pattern Discovery. Proceedings of the Third International conference on language resources and evaluation (LREC'02, 1, 1–4) http://pdfs.semanticscholar.org/c500/40ac812c3f3le0cf37802ff87de2dce87821.pdf

Resnik P (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceedings of the 14th international joint conference on Artificial intelligence vol. 1 http://arxiv.org/abs/cmp-lg/9511007

Rodrigo Á, Pérez-Iglesias J, Peñas A, Garrido G, Araujo L (2013) Answering questions about European legislation. Expert Syst Appl 40(15):5811–5816. https://doi.org/10.1016/j.eswa.2013.05.008

Saad MB, Gançarski S (2010, March) Using visual pages analysis for optimizing web archiving. In Proceedings of the 2010 EDBT/ICDT Workshops (pp. 1-7). https://doi.org/10.1145/1754239.1754287

Sagara T, Kitsuregawa M (2001) Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents. SIGIR Workshop on Geographical Information Retrieval (2004). http://www.geo.unizh.ch/~rsp/gir/

Schmidt S, Manschitz S, Rensing C, Steinmetz R (2013) Extraction of Address Data from Unstructured Text using Free Knowledge Resources. 13th International Conference on Knowledge Management and Knowledge Technologies, At Graz, Austria https://doi.org/10.1145/2494188.2494193

Sekine S, Grishman R, Shinnou H (1998) A Decision Tree Method for finding and classifying names in Japanese texts. Proceeding of the 6th workshop on Very Large Corpora, (May), 171–178

Song HJ, Park SB, Park SY (2009) An automatic ontology population with a machine learning technique from semi-structured documents. IEEE Int Conf Inform Auto ICIA 2009:534–539. https://doi.org/10.1109/ICINFA.2009.5204981

Song R, Liu H, Wen JR, Ma WY (2004, May) Learning block importance models for web pages. In Proceedings of the 13th international conference on World Wide Web (pp. 203-211). https://doi.org/10.1145/988672.988700

Souza LA, Davis CA, Borges KAV, Delboni TM (2005) Laender AHF (2005) The role of gazetteers in geographic knowledge discovery on the Web. Proceedings - Third Latin American Web Congress, LA-WEB 2005:157–165. https://doi.org/10.1109/LAWEB.2005.38

Stab Christian IG (2017) Parsing argumentation structures in persuasive essays christian. Jurnal Pengurusan 38(April):41–51. https://doi.org/10.1162/COLI

Takeuchi K, Collier N (2002) Use of support vector machines in extended named entity recognition. Proceedings of the 6th Conference on Natural Language Learning-Volume 20. Association for Computational Linguistics, 2002., 1–7 http://dl.acm.org/citation.cfm?id=1118882

Teitler BE, Lieberman MD, Panozzo D, Sankaranarayanan J, Samet H, Sperling J (2008, November). NewsStand: A new view on news. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (pp. 1-10)

Tjong EF, Sang K, Meulder F De., Introduction to the CoNLL Shared Task Language Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (2003)

Touya G (2010) Quality assessment of the French. Trans GIS 14(4):435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

Uryupina O (2002) Extracting geographical knowledge from the internet 2002.pdf. Proc. of the ICDM-AM International Workshop on Active Mining - Maebashi, 113–118

Uryupina O (2003) Semi-supervised learning of geographical gazetteers from the internet. Proceedings of the HLTNAACL 2003 Workshop on Analysis of Geographic References, 1, 18–25 https://doi.org/10.3115/1119394.1119397

Vadrevu S, Gelgi F, Davulcu H (2005, November). Semantic partitioning of web pages. In International Conference on Web Information Systems Engineering (pp. 107-118). Springer, Berlin, Heidelberg

Xu L, Du Z, Mao R, Zhang F, Liu R (2020) GSAM: A deep neural network model for extracting computational representations of Chinese addresses fused with geospatial feature. Comput Environ Urban Syst 1(81):101473

Yu S, Cai D, Wen J-R, Ma W-Y (2004) Improving pseudo-relevance feedback in Web information retrieval using Web page segmentation. Proceedings of the 12th international conference on World Wide Web, 11-18 https://doi.org/10.1145/775152.775155

Yu Z (March, 2007), High accuracy postal address extraction from web pages.In Masters Abstracts International (Vol. 45, No. 05)

Zhang J, Dang Q, Lu Y, Sun S (2013) Suffix tree clustering with named entity recognition. Proceedings - 2013 International Conference on Cloud Computing and Big Data, CLOUDCOM-ASIA 2013, 549–556, https://doi.org/10.1109/CLOUDCOM-ASIA.2013.102

Zhang Y, Gao M, Zhang X, Yang P, Ma Q, Wang C, Hu X (2018) An Automatic Approach to Extracting Geographic Information from Internet. IEEE Access, 3536(c), 1–1, (2018). https://doi.org/10.1109/ACCESS.2018.2844470

Zhao S, (2004) Named entity recognition in biomedical texts using an HMM model. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications - JNLPBA '04, (Grefenstette 1994), 84, https://doi.org/10.3115/1567594.1567613

Zheng S, Hao Y, Lu D, Bao H, Xu J, Hao H, Xu B (2017) Joint entity and relation extraction based on a hybrid neural network. Neurocomputing 257(2017):59–66. https://doi.org/10.1016/j.neucom.2016.12.075

Zhou G, Su J (2001) Named entity recognition using an HMM-based chunk tagger. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, (July), 473. https://doi.org/10.3115/1073083.1073163

Zielstra D, Zipf A (2010) A comparative study of proprietary geodata and volunteered geographic information for Germany. 13th AGILE International Conference on Geographic Information Science 2010 Guimarães, Portugal, 1, 1–15, https://doi.org/10.1119/1.1736005