International Journal of Geo-Information

MDPI

Review

Automatic Identification of Addresses: A Systematic Literature Review

Paula Cruz *, Leonardo Vanneschi, Marco Painho and Paulo Rita

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal; lvanneschi@novaims.unl.pt (L.V.); painho@novaims.unl.pt (M.P.); prita@novaims.unl.pt (P.R.)

* Correspondence: m2014175@novaims.unl.pt

Abstract: Address matching continues to play a central role at various levels, through geocoding and data integration from different sources, with a view to promote activities such as urban planning, location-based services, and the construction of databases like those used in census operations. However, the task of address matching continues to face several challenges, such as non-standard or incomplete address records or addresses written in more complex languages. In order to better understand how current limitations can be overcome, this paper conducted a systematic literature review focused on automated approaches to address matching and their evolution across time. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed, resulting in a final set of 41 papers published between 2002 and 2021, the great majority of which are after 2017, with Chinese authors leading the way. The main findings revealed a consistent move from more traditional approaches to deep learning methods based on semantics, encoder-decoder architectures, and attention mechanisms, as well as the very recent adoption of hybrid approaches making an increased use of spatial constraints and entities. The adoption of evolutionary-based approaches and privacy preserving methods stand as some of the research gaps to address in future studies.

Keywords: address matching; address parsing; machine learning; deep learning; natural language processing; address geocoding

Citation: Cruz, P.; Vanneschi, L.; Painho, M.; Rita, P. Automatic Identification of Addresses: A Systematic Literature Review. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 11. https://doi.org/10.3390/ijgi11010011

Academic Editor: Wolfgang Kainz

Received: 16 November 2021
Accepted: 26 December 2021
Published: 29 December 2021

1. Introduction

An address is a reference to a unique location on Earth and is usually expressed according to a certain addressing system (a combination of components such as street names, building numbers, units, levels, unit directions, postal codes, etc.), which can be distinguished from others based on its structure as well as on the types of used components [1]. Due to the hierarchical nature of the fields that compose an address, the association between addresses and address fields can be formally modelled, thereby taking into account the semantic characteristics of address fields [2].

In general terms, address matching consists of the process of identifying pairs of records through the comparison of full addresses or address fields, with the aim of obtaining the best matching result in relation to a searched address [3]. Address matching is also described as the process of relating the literal description of an address to its corresponding location on a map [4]. In this process, known as geocoding, addresses (up to the street name or street name and door number, combined with a postal code and/or an administrative division) are matched with a reference database in order to obtain the corresponding spatial geographic coordinates [5]. In the absence of a unique identifier (such as the social security number, for instance), addresses can also be used as quasi-identifiers in the linking of records related to the same entity in one or more data collections [6]. As such, address matching main areas of application include, among others, the enrichment of data quality [3], named entity recognition [6], and location-based analyses in general [7],

which are central to express delivery and take-out services, to disaster risk management and response, as well as in the construction of databases such as those used in census operations [3].

Closely associated to address matching is the task of address parsing or address segmentation, which consists of decomposing an address into its different components, such as a street name or a postal code. Basically, through parsing, it is possible to convert unstructured or semi-structured input addresses into structured ones, helping to overcome imprecise or vague addresses [5].

Regarding the matching of address fields or addresses, three types of similarities should be considered: string similarity, semantic similarity, and spatial similarity [2]. String similarity is mainly focused on finding common substrings or characters between address records or elements, whereas semantic similarity tries to capture the linguistic relations between words, such as synonyms (for instance, "street" and "road" both consist of street types but would be considered highly dissimilar based on a string similarity approach). Lastly, spatial similarity can be measured based on street numbers, when available. The combination of multiple similarities generally increases address matching accuracy, even though there is a tendency to overlook semantic characteristics and spatial proximities [2].

Traditional address matching methods are organized into two major categories: string similarity-based methods (calculation of the text-similarity between two addresses) and address element-based methods (comparison of the hierarchical results and the matching rate between each address element) [8]. However, these methods do not always manage to tackle non-standard address records, with redundant or missing address elements and few literal overlaps or written in more complex languages [9].

In order to better understand how these limitations can and have been overcome, this paper conducted a systematic literature review (SLR) focused on automated approaches to address matching and their evolution across time. Bibliometric analysis is a particularly useful method to discover hot topics, trends, research gaps, top authors and institutions [10] and, to the best of our knowledge, no similar studies have been previously published in the field under analysis. However, a brief reference should be made to related surveys, namely in the field of geographic information extraction from textual documents [11] and unstructured and diverse data, such as addresses on the Web [12]. The first study [11] mainly addresses the geographical coordinates' prediction of entire documents based on their textual contents, by conducting a survey on previous research in this field, with address geocoding being explicitly excluded from its scope, due to the significant difference in the used methods. As for the second of the mentioned studies [12], a review of different approaches to postal address extraction from the Web is performed with a twofold aim: firstly, to analyze the data quality of gazetteers (geographical dictionaries) like the Volunteered Geographical Information (VGI) based GeoNames (https://www.geonames.org/ (accessed on 9 November 2021)) and OpenStreetMap (https://www.openstreetmap.org (accessed on 9 November 2021)) [13]; secondly, to identify the factors that most hinder postal address extraction performance, including the diversity of styles and sources of addresses on the Web as well as their ambiguous and dynamic nature. The main conclusions point to the coverage of real Web pages and social networks with a view to obtain increased geographical knowledge on Points of Interest (restaurants, schools, hospitals, etc.), alongside the application of deep learning models in this area. It can thus be stated that, although related to address matching, the domain of postal address extraction differs from the former in the following main dimensions: the used methods (heavily dependent on gazetteers and involving the preprocessing of html documents containing very diverse entities), the source of the information (the Web, instead of semi-structured databases), and the components of addresses (which go as far as the street name and, in some cases, the street number). In relation to this last aspect, it should also be noticed that in address structures such as the one used in Portugal and several other countries [14], the street number is included after the street name, which can also contain numbers, such as streets which are named after important holidays, for instance. In the case of addresses related to apartment buildings

and other types of multifamily houses, the street number will be further followed by the identification of address elements such as the unit, level, and unit direction, which are hard to tackle through the use of dictionaries or even more traditional machine learning approaches, due to writing variations and the use of non-standard abbreviations, combined with missing elements.

To perform the proposed SLR, the paper is structured as follows: Section 2 presents the main data sources, search strategies, screening procedures and tools; Section 3 contains the main results, their discussion and identified research gaps; and, finally, in Section 4, we present our main conclusions, including some recommendations for future research.

2. Materials and Methods

2.1. Data Sources and Search Strategies

In order to select the most relevant set of articles, the following query was executed in April 2021 and December 2021 (in order to retrieve more recent papers) in the electronic repositories Scopus and Web of Science:

("address matching" OR "toponym matching" OR "address pars*" OR "address standardization" OR "address database*" OR "address string*" OR "postal address data" OR "non-standard addresses" OR "address element segmentation" OR "name and address data" OR "address geocoding" OR "geocoding addresses" OR "geocoded address*") **AND** ("machine learning" OR "deep learning" OR "neural network*" OR "vector representation" OR semantic* OR probabilistic OR automat* OR "similarity measure*") **AND NOT** ("IP address*" OR "mac address*" OR URL OR email*).

The Boolean expressions OR/AND imply that any article should contain at least one keyword from the first subquery inside curved brackets and one keyword from the second one. The Boolean expression "AND NOT" aims at further excluding any article containing one of the keywords inside the last subquery. The keywords included in the final query resulted from a fine-tuning process of alternative keywords' combination based on recently published papers in this field, such as the ones by Comber and Arribas-Bel [3] and Lin et al. [9], as well as on some of the earlier works, namely by Churches et al. [6]. The keywords "address data" were explicitly excluded from the query due to the ambiguous use of address as a noun and as a verb and to the presence of a significant number of papers related to the study of data imbalance issues, among others. The keyword "geocoding", although relevant, was also excluded due to its more general nature and conceptual "fuzziness" (in the sense that it can have different meanings depending on the user's perception and experience), with postal addresses consisting of one of the possible inputs that can be assigned a geographic code [15]. A combination of the keywords "geocoding" or "geocoded" and "addresses" was considered instead. Lastly, a time period of 20 years was considered, since seminal work about the researched topic based on machine learning approaches was first published in 2002 [6].

2.2. Screening Procedures

A total of 122 distinct documents were retrieved from the previously mentioned databases, after deduplication of common articles. For the selection of relevant papers, the following exclusion criteria were considered:

- Exclusion of reviews, book chapters, reports, and other duplicates (e.g.,: articles published as book chapters in the Springer series "Studies in Computational Intelligence");
- Exclusion of conferences not ranked as "A" (as of April 2021), according to the conference ranking provided in http://www.conferenceranks.com/ (accessed on 9 November 2021) (e.g.,: International Conference on Natural Computation);
- Exclusion of journals not ranked as Q1 or Q2 (as of April 2021), according to the SCImago Journal Rank indicator (https://www.scimagojr.com/ (accessed on 9 November 2021)) (e.g.,: Russian Journal of Forest Science);
- Exclusion of articles which were not in the scope of the research (e.g.,: articles dealing with inputs not related to addresses).

Regarding the screening procedures, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [16] were followed, resulting in the final inclusion of 41 articles as shown in Figure 1.

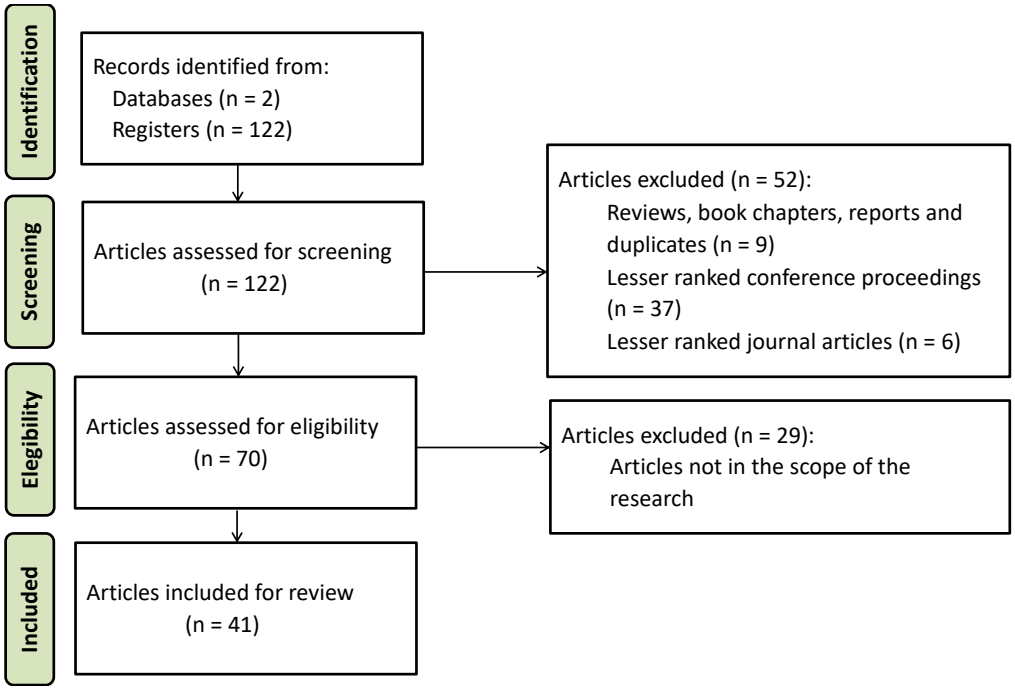


Figure 1. PRISMA flow diagram of the followed screening procedures.

2.3. Tools

Excel 2010 (Microsoft Corp), VOSviewer 1.6.16 (https://www.VOSviewer.com (accessed on 9 November 2021)) and Gephi 0.9.2 (https://gephi.org/ (accessed on 9 November 2021)) were used for the qualitative and quantitative analyses of keywords and author co-occurrences as well as publication trends, top countries, research gaps, application areas and methods. VOSviewer consists of a bibliometric analysis tool for network analysis based on clustering techniques and text mining [17]. It enables three types of visualizations: network visualization, overlay visualization, and density visualization. In this analysis, only the first two were used, for the sake of simplicity. Gephi consists of an open-source software for network analysis [18] in a broader range of subjects, enabling the extraction of graph theory metrics additional to the ones provided by VOSviewer. As such, in this paper, Gephi was mostly used in a subsidiary manner, whenever considered necessary.

3. Results and Discussion

3.1. Results

3.1.1. Publication Venues of the Selected Papers

Of the 41 articles that met the inclusion criteria, 17 were published in Q1 journals, 10 in Q2 journals, 6 in a Q2 book series, and the remaining 8 in conference proceedings, as illustrated in Figure 2:

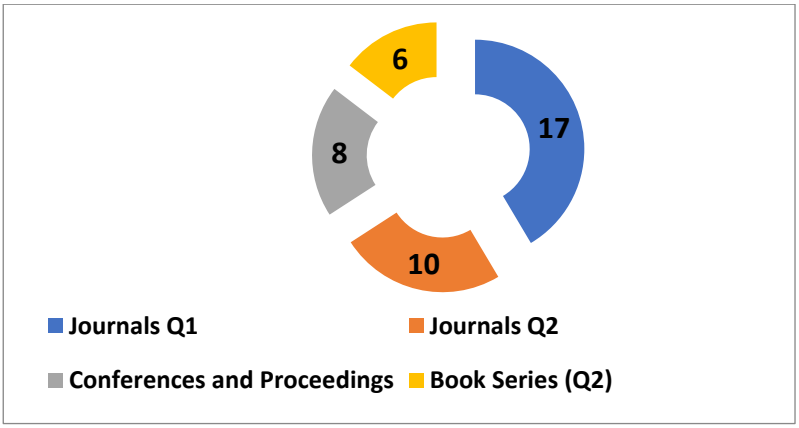


Figure 2. Type and ranking of the selected papers, according to SCImago Journal Rank indicator and http://www.conferenceranks.com (accessed on 9 November 2021).

In Tables 1 and 2 the main journals and conference proceedings are presented, respectively.

Table 1. Publication sources of the selected journal articles.

Journal	No.	Quartile	Publisher	Field(s)	Publisher Country
ISPRS International Journal of Geo-Information	4	Q1	MDPI AG	Earth and Planetary Sciences; Social Sciences	Switzerland
International Journal of Geographical Information Science	2	Q1	Taylor and Francis Ltd.	Computer Science; Social Sciences	United Kingdom
Applied Sciences (Switzerland)	2	Q1	MDPI Multidisciplinary Digital Publishing Institute	Chemical Engineering; Computer Science; Engineering; Materials Science; Physics and Astronomy	Switzerland
Transactions in GIS	2	Q1	Wiley-Blackwell Publishing Ltd.	Earth and Planetary Sciences	United Kingdom
Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University	2	Q2	Wuhan University	Computer Science; Earth and Planetary Sciences	China
Computers, Environment and Urban Systems	1	Q1	Elsevier Ltd.	Environmental Science; Social Sciences	United Kingdom
BMC Medical Informatics and Decision Making	1	Q1	BioMed Central Ltd.	Medicine	United Kingdom
International Journal of Digital Earth	1	Q1	Taylor and Francis Ltd.	Computer Science; Earth and Planetary Sciences	United Kingdom
Pattern Recognition Letters	1	Q1	Elsevier	Computer Science	Netherlands
Population, Space and Place	1	Q1	John Wiley and Sons Ltd.	Social Sciences	United Kingdom
Region	1	Q2	European Regional Science Association	Economics, Econometrics and Finance; Social Sciences	Belgium

Table 1. *Cont.*

Journal	No.	Quartile	Publisher	Field(s)	Publisher Country
World Wide Web	1	Q2	Springer New York	Computer Science	United States
Yaogan Xuebao/Journal of Remote Sensing	1	Q2	Science Press	Earth and Planetary Sciences; Physics and Astronomy; Social Sciences	China
Cadernos de Saude Publica	1	Q2	Fundacao Oswaldo Cruz	Medicine	Brazil
Canadian Geographer	1	Q1	Wiley-Blackwell Publishing Ltd.	Earth and Planetary Sciences; Social Sciences	United Kingdom
(Electronic) Journal of Information Technology in Construction	1	Q2	International Council for Research and Innovation in Building and Construction	Computer Science; Engineering	Sweden
International Journal of Applied Engineering Research	1	Q2	Research India Publications	Engineering	India
International Journal of Health Geographics	1	Q1	BioMed Central Ltd.	Business, Management and Accounting; Computer Science; Medicine	United Kingdom
International Journal of Image and Data Fusion	1	Q2	Taylor and Francis Ltd.	Computer Science; Earth and Planetary Sciences	United Kingdom
Journal of Data and Information Quality	1	Q2	Association for Computing Machinery	Computer Science; Decision Sciences	United States

Table 2. Publication sources of the selected conference articles.

Conference Proceedings/Book Series	No.	Publisher Country	Field(s)
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	6	Germany	Computer Science; Mathematics
Proceedings of the International Conference on Document Analysis and Recognition, ICDAR	3	United States	Computer Science
International Conference on Information and Knowledge Management, Proceedings	1	United States	Business, Management and Accounting; Decision Sciences
NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference	1	United States	Computer Science
Proceedings – 2010 IEEE 7th International Conference on Services Computing, SCC 2010	1	United States	Computer Science; Mathematics
Proceedings – International Conference on Pattern Recognition	1	United States	Computer Science
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	1	United States	Computer Science

Regarding the number of publications, the top journals include the ISPRS International Journal of Geo-Information (with 4 articles), the Applied Sciences (Switzerland) journal (with 2 articles), the International Journal of Geographical Information Science (with 2 articles), the Transactions in GIS journal (with 2 articles), and the Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University (with 2 articles), of which the first four are ranked as Q1. Within the top journal publishers are MDPI AG (Switzerland), MDPI Multidisciplinary Digital Publishing Institute (Switzerland), Taylor and Francis Ltd. (United Kingdom), Wiley-Blackwell Publishing Ltd. (United Kingdom), and Wuhan University (China). The most influential journals, in terms of the number of citations, are also ranked as Q1 (Table 3), as was expected. Most of the conference papers originate from the book series Lecture Notes in Computer Science (with 6 papers), followed by the International Conference on Document Analysis and Recognition (with 3 papers), with the latter being more influential as far as the number of citations is concerned.

Table 3. Main sources by number of citations.

Journal Articles	**No. of Publications**	**No. of Citations**
International Journal of Health Geographics	1	228
BMC Medical Informatics and Decision Making	1	70
International Journal of Geographical Information Science	2	38
ISPRS International Journal of Geo-Information	4	28
International Journal of Digital Earth	1	18
Canadian Geographer	1	16
Conference Articles	**No. of publications**	**No. of citations**
Proceedings of the International Conference on Document Analysis and Recognition, ICDAR	3	28
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	6	17
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	1	12
Proceedings - 2010 IEEE 7th International Conference on Services Computing, SCC 2010	1	12
Proceedings - International Conference on Pattern Recognition	1	10
International Conference on Information and Knowledge Management, Proceedings	1	6

Overall, the main research fields identified in the analysis were computer science (34%), Earth and planetary sciences (16%), social sciences (15%), mathematics (9%), engineering (5%), medicine (4%), and physics and astronomy (4%), with the remaining areas corresponding to decision sciences, materials science, business, management and accounting, chemical engineering, economics, econometrics and finance, and environmental science. The top publishing countries comprise the United Kingdom (27%), the United States (24%), Germany (15%), and Switzerland (15%). Considering the authors' affiliations, the first position is held by China (with 38%), followed by the United States (12%), the United Kingdom (10%), and India (8%).

In the considered time period, spanning from 2002 to 2021, the number of published articles has been steadily increasing since 2017 (Figure 3). This trend can be explained by the big volume of unstructured address data that has been created by the rapid development of mobile internet and location-based services and the increasing need for effective address matching methods, in order to facilitate geocoding and promote geospatial management [2,9]. In countries like China, rapid urban expansion has also led to an

increased concern with the improvement of address quality and the retrieval of standard address data [19].

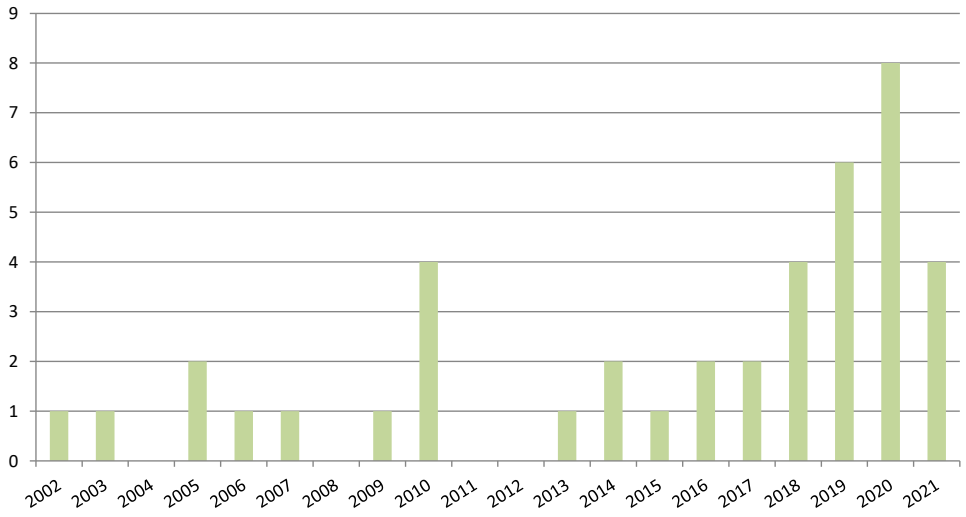
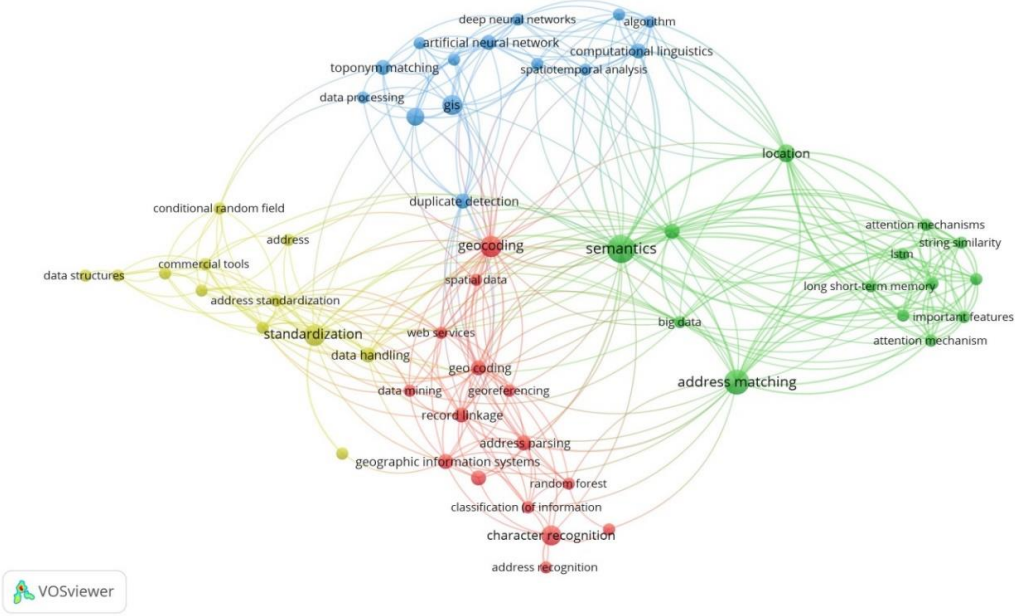


Figure 3. Number of articles published over time.

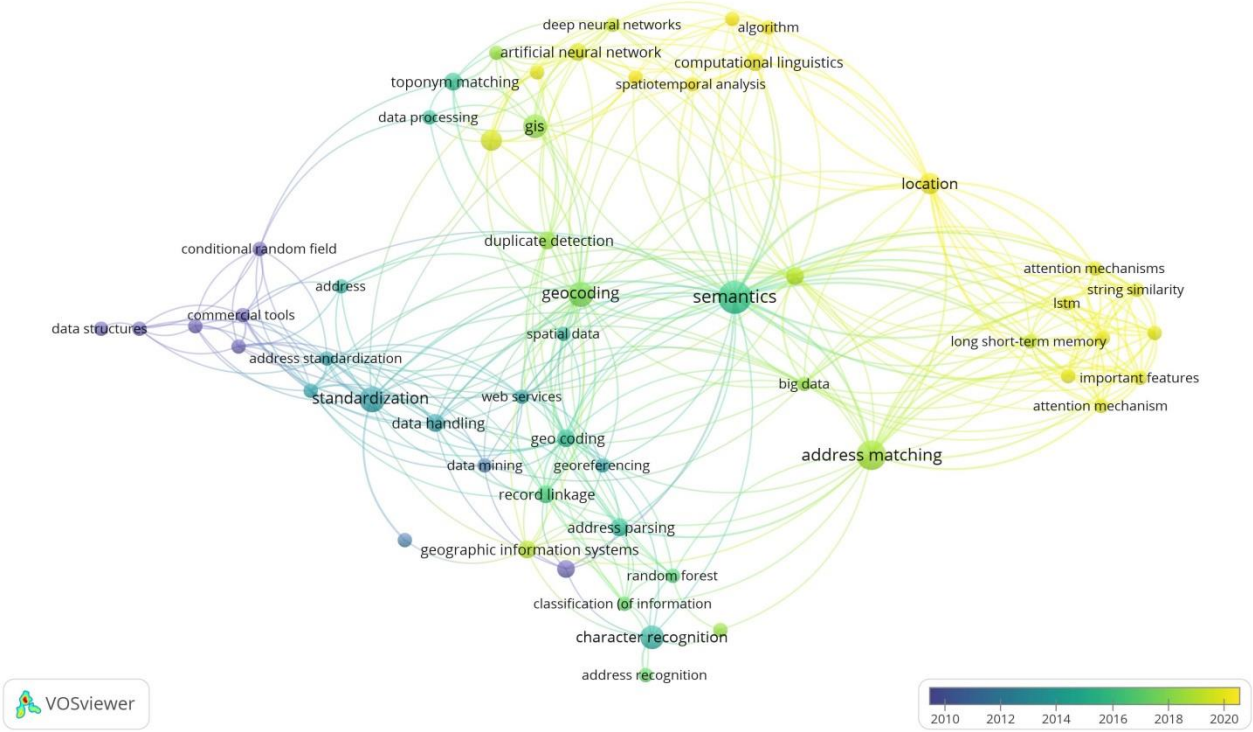
3.1.2. Keyword Occurrence Analysis

The keyword co-occurrence analysis was performed using VOSviewer, as depicted in Figure 4a. A keyword is represented by a circle and its importance by the circle's size, with circles of the same color belonging to the same cluster. The number of times two connected nodes are referenced together is represented by the thickness of the link connecting the circles. In particular, a full counting method was used, involving 55 screened keywords, with a minimum threshold of 2 occurrences. In Figure 4b, an overlay visualization is also included in order to reveal changing trends in keywords. Earlier occurrences are depicted in blue, and the more recent ones in yellow. Computation linguistics, attention mechanisms, LSTMs, and location-based services emerge as some of the most recent research topics.



(**a**)

Figure 4. *Cont.*



(**b**)

Figure 4. (**a**) Network visualization of keyword occurrence. (**b**) Network overlay of keyword occurrence. Note: Circles represent keywords and their size represents the keywords' importance. Clusters are identified by circles sharing the same color and the thickness of the link connecting the circles corresponds to the number of times two connected nodes are mentioned together. An earlier time is represented by a color closer to blue and a more recent year by a color closer to yellow.

As shown in Table 4, four clusters were identified, based on VOSviewer default clustering technique [17]: address matching and NLP, in green (e.g.,: Xu et al. [20]), GIS/geocoding and machine learning, in blue (e.g.,: Peng et al. [21]), address standardization, in yellow (e.g.,: Churches et al. [6]) and address recognition and parsing, in red (e.g.,: Wei et al. [22]).

Table 4. Co-words obtained in each cluster by VOSviewer.

Clusters	Co-Words
Address matching and NLP	address matching(s), big data, important features, location, location-based services, long short-term memory (lstm), natural language processing systems, search engines, semantic representation, semantics, string similarity
GIS/geocoding and machine learning	accuracy assessment, algorithm, artificial neural network, China, computational linguistics, data processing, deep neural networks, duplicate detection, geographic information retrieval, gis, machine learning, recurrent neural networks, spatiotemporal analysis, toponym matching
Address standardization	address, address standardization, article, artificial intelligence, commercial tools, conditional random field, data cleansing, data handling, data source, data structures, information analysis, standardization
Address recognition and parsing	address parsing, address recognition, character recognition, classification (of information), data mining, geocoding, geographic information systems, georeferencing, hidden markov models, neural networks, random forest, record linkage, spatial data, web services

3.1.3. Co-Authorship Analysis

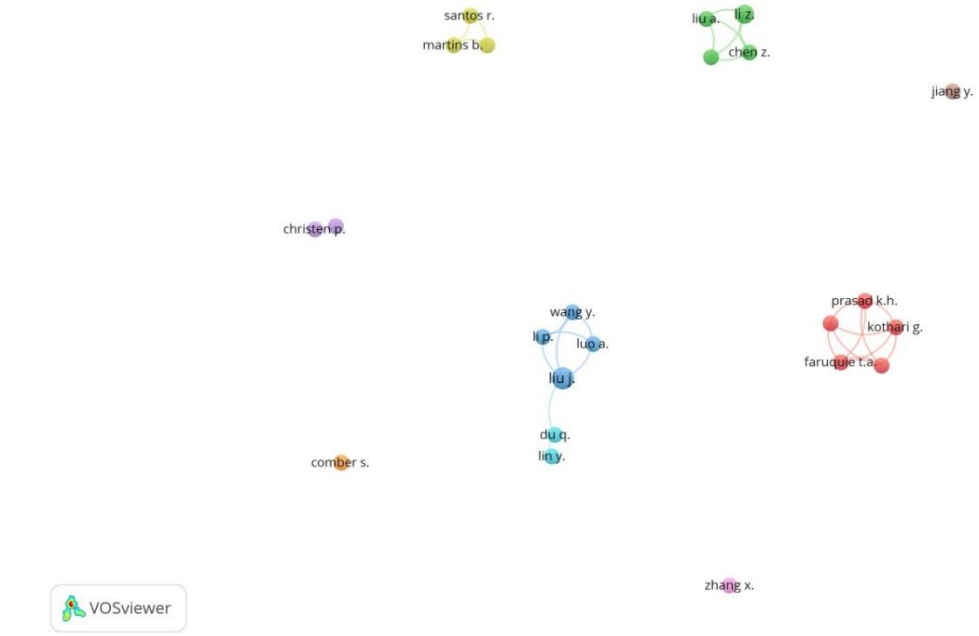
VOSviewer was also used to perform the analysis on co-authorship. A full counting method and a minimum of 2 documents and 2 citations were chosen, resulting in a total of 23 authors. As shown in Table 5 and Figure 5a,b, 9 clusters were found, which appear to be organized around collaborating authors' countries of origin (and, in most cases, to the corresponding address model structures and the language in which addresses are written), degree of collaboration between researchers (link strength), and average year of publication: cluster 1 corresponds to authors from India (2010); clusters 2, 3, and 6 to Chinese researchers who published articles in 2020, 2021, and 2018–2019, respectively, with the latter exhibiting a weaker link strength than the former; cluster 4 includes Portuguese authors (2018); cluster 5 involves Australian researchers (2004); cluster 7 refers just an English author (2019); and, finally, clusters 8 and 9 engage Chinese researchers publishing in different years (2006 and 2010), with these earlier works being focused on more traditional approaches to address matching and parsing. Overall, Chinese authors lead the way, with 19 papers that represent 46% of the published articles, half of them published between 2019 and 2021, after a peak in 2016 (11%). Nevertheless, the most cited authors are the Australians Peter Christen and Tim Churches, which may be influenced by the year and field of research of the corresponding papers, since different publication and citation cultures may put in disadvantage papers from more recent time periods and specific subfields [23].

Table 5. Author co-authorship ranked by link strength.

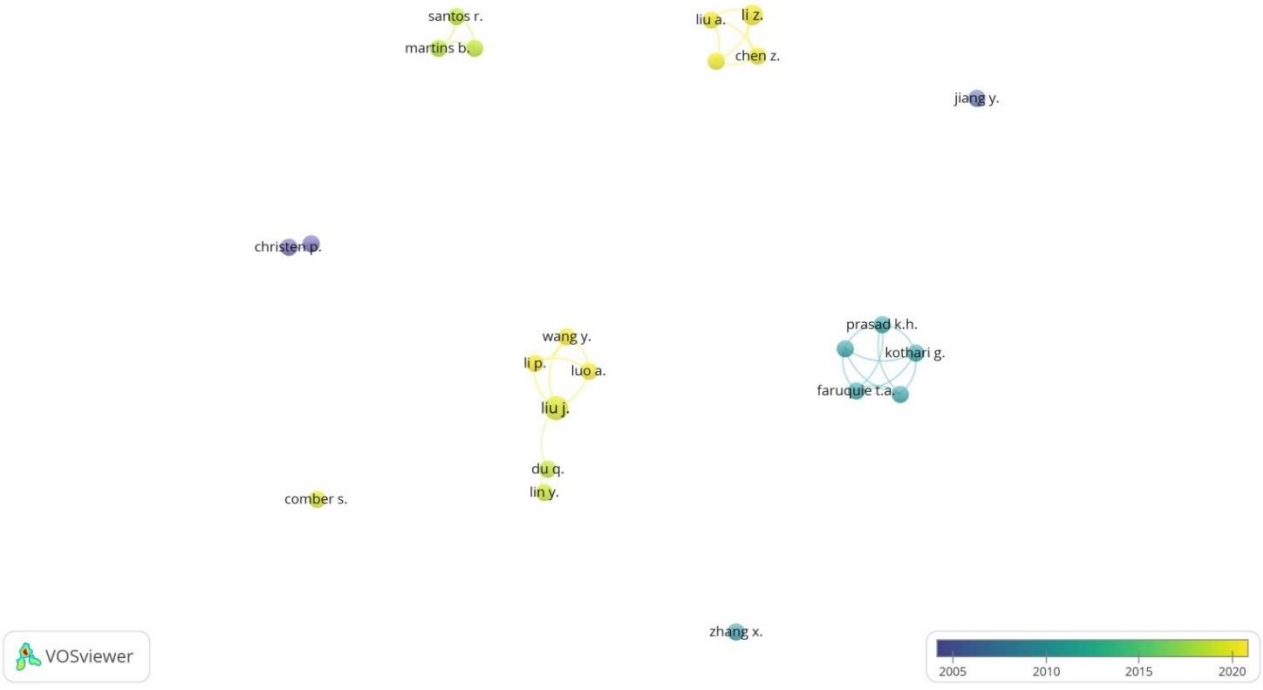
Author	Cluster	Link Strength	Documents	Citations	Average Publication Year
Tanveer A. Faruquie	1	8	2	22	2010
Govind Kothari	1	8	2	22	2010
Mukesh K. Mohania	1	8	2	22	2010
K. Hima Prasad	1	8	2	22	2010
L. Venkata Subramaniam	1	8	2	22	2010
Jing Liu	3	7	4	26	2019
Zhigang Chen	2	6	2	3	2020
Zhixu Li	2	6	3	3	2020
An Liu	2	6	2	3	2020
Shuangli Shan	2	6	2	3	2020
Pengpeng Li	3	6	2	4	2021
An Luo	3	6	2	4	2021
Yong Wang	3	6	2	4	2021
Bruno Martins	4	4	2	42	2018
P. Murrieta-Flores	4	4	2	42	2018
Rui Santos	4	4	2	42	2018
Peter Christen	5	2	2	80	2004
Tim Churches	5	2	2	80	2004
Qingyun Du	6	2	2	36	2018
Yue Lin	6	1	2	14	2019
Sam Comber	7	0	2	10	2019
Yan Jiang	8	0	2	21	2006
Xiaoxun Zhang	9	0	2	25	2010

In order to better understand the connections between the different authors and their research, an analysis based on the references of each paper was also undertaken. Of the three different types of citation-based approaches available in VOSviewer, bibliographic coupling [24] was chosen, which measures the similarity between papers based on the number of references they share [25]. This approach is less affected by changes over time since references remain stable [25] and it outperformed alternative methods in a comparative study by X. Liu [26]. In Figure 6, a bibliographic coupling analysis of authors is depicted, based on a full counting method and a minimum of 2 documents and 2 citations, pointing to the existence of bibliographic coupling relations between almost all of the

researchers at hand, in spite of their relative isolation in terms of the co-authorship analysis, outside each cluster.



(**a**)



(**b**)

Figure 5. (**a**) Network visualization of author co-authorship analysis. (**b**) Author co-authorship analysis by year overlay visualization. Note: An author's name is represented by a circle and its importance by the circle's size, with circles of the same color belonging to the same cluster. The thickness of the link connecting the circles represents the number of times two connected nodes are mentioned together. An earlier time is represented by a color closer to blue and a more recent year by a color closer to yellow.

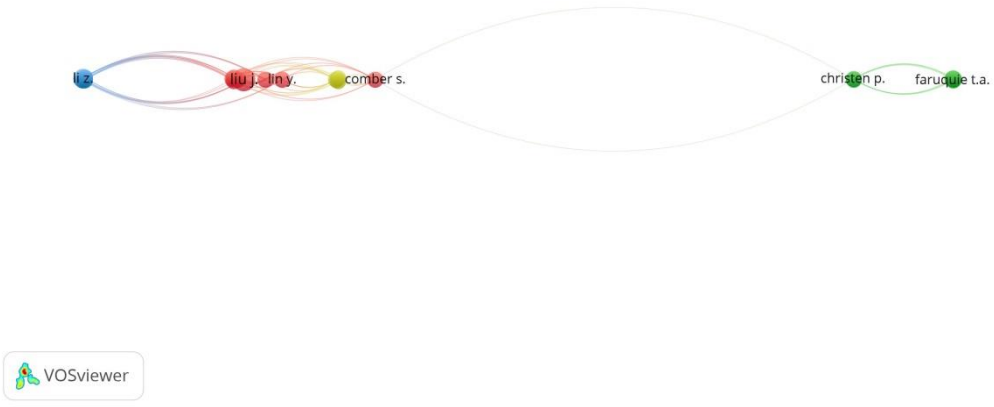


Figure 6. Author bibliographic coupling visualization network. Note: Each circle represents an author, with larger circles representing researchers that have more publications. The relative strength of the relations between authors is represented by the color of the clusters and by how close they are to each other on the visualization.

3.1.4. Application and Methods Analysis

An evaluation of the main application areas, methods and algorithms used in the papers under study was also undertaken using VOSviewer. Two separate keyword analyses using a full counting method and a minimum threshold of 1 occurrence were performed. In each of the analyses, only keywords related to applications or methods/algorithms were taken in consideration.

Figure 7 shows that the top 5 application domains consist of geographical information systems (GIS)/Census [27], POIs/Spatial Analysis [2], GIS/Urban Planning [9], GIS/Health Care [28], and Location Based Services [29]. Taking into account the average publication year (Figure 8), it is possible to observe that the most recent application domains consist of disease control (covid-19), location-based services, and GIS/census/urban planning, in which geocoding, with an increasing importance in people's daily lives, stands as a common feature.

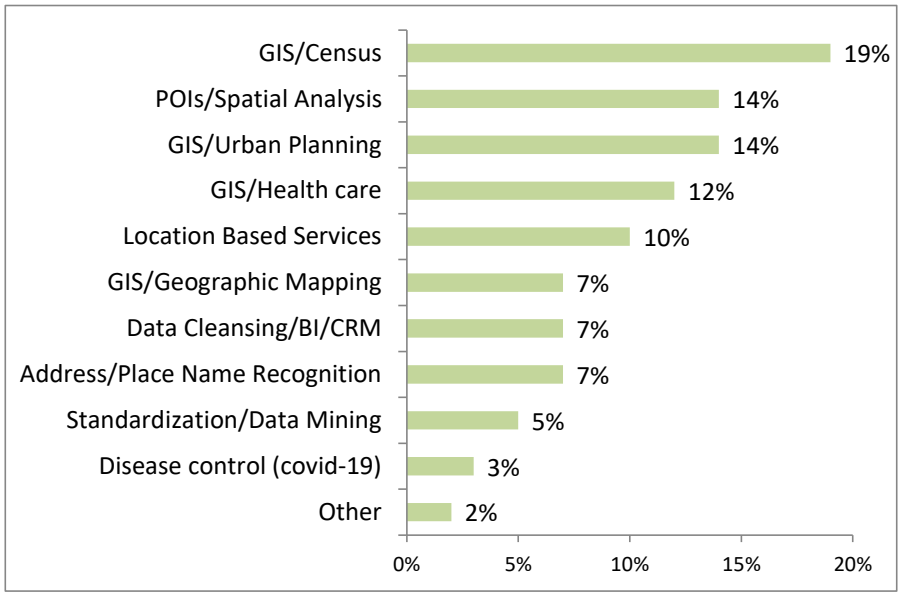


Figure 7. Keywords' occurrences by application area (%).

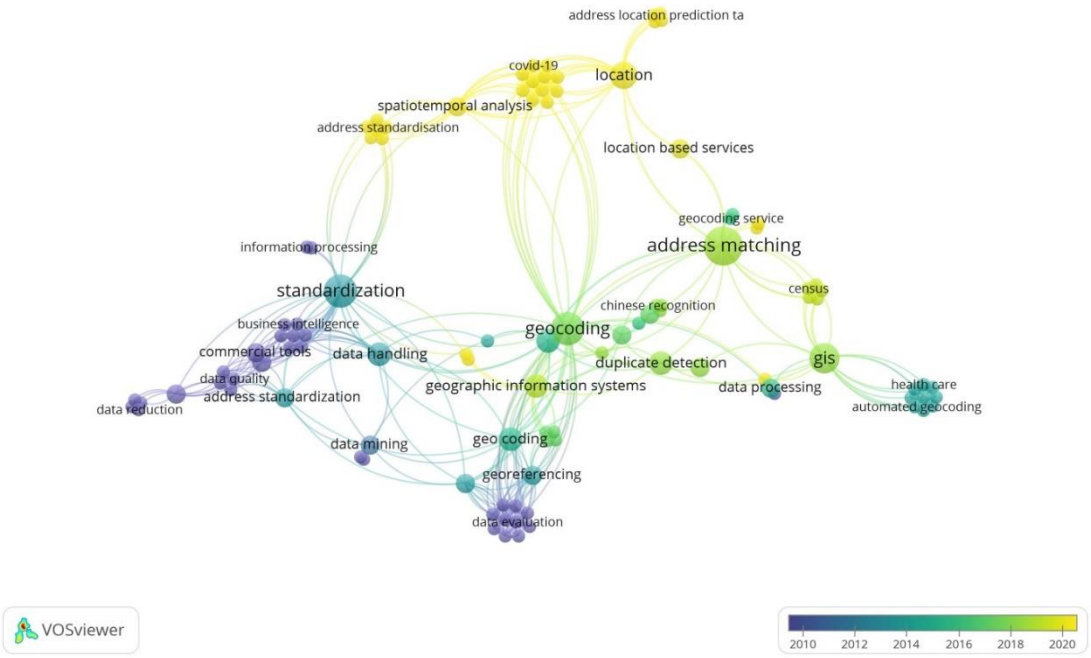


Figure 8. Application domains by average publication year.

As far as methods/algorithms are concerned, Figure 9 shows the growing importance of deep learning algorithms in the field of address matching since 2018, such as recurrent neural networks [30], long short-term memory networks [31], gated recurrent units [30], bidirectional encoder representations from transformers [32,33], and Graph Convolutional Networks [34]. Recurrent neural networks (RNNs) were originally conceived to spot patterns in data sequences like character strings, through "a recurrent hidden state whose activation at each time step is dependent on the previous time step" [35] (p. 331). Long short-term memory networks (LSTM) and gated recurrent units (GRU) consist of two well-known extensions of RNNs, which are able to handle RNN's difficulties in modelling long-term dependencies (i.e., long sequences). Graph Convolutional Networks (GCN) are a special case of Graph Neural Networks (GNN), which were also originally introduced as extensions of RNNs [36]. Bidirectional encoder representations from Transformers (BERT) [32,33] consist of a simpler network architecture, based solely on attention mechanisms (which assign higher weights to the most important features), not requiring the sequential processing of data.

Probabilistic based approaches for segmenting and labelling sequence data, such as Hidden-Markov Models (HMMs) and Conditional Random Fields (CRFs) [3], on their turn, have been mostly used before 2015 (it should be noticed, however, that CRFs have continued being used in combination with other, more advanced approaches [37]). A Hidden Markov Model [38–40] consists of a finite set of unobserved (hidden) states, a matrix of transition probabilities between those states, a collection of observable facts, and an observation (or emission) matrix comprising the probabilities with which each hidden state emits an observation. Conditional Random Fields (CRFs) are inherently conditional and assume that the output labels are not independent [41,42].

Semantics, which aims at understanding natural language contents such as addresses [9], consists of the most important node of the network depicted in Figure 9, reflecting its central role and growing relevancy in this research field. In order to better understand the relative importance of each node and the interactions between them, two centrality measures were additionally considered: the eigenvector centrality, which characterizes the global centrality of a node in a network, and the betweenness centrality, which can be described as the number of times that a certain node needs another one to reach a third node through the shortest path [43]. For that effect, a Pajek (*.net) file containing the network of keyword occurrences related to methods/algorithms was extracted from VOSviewer

and used as an input to Gephi. The obtained results are included in Table 6, confirming the global centrality of semantics.

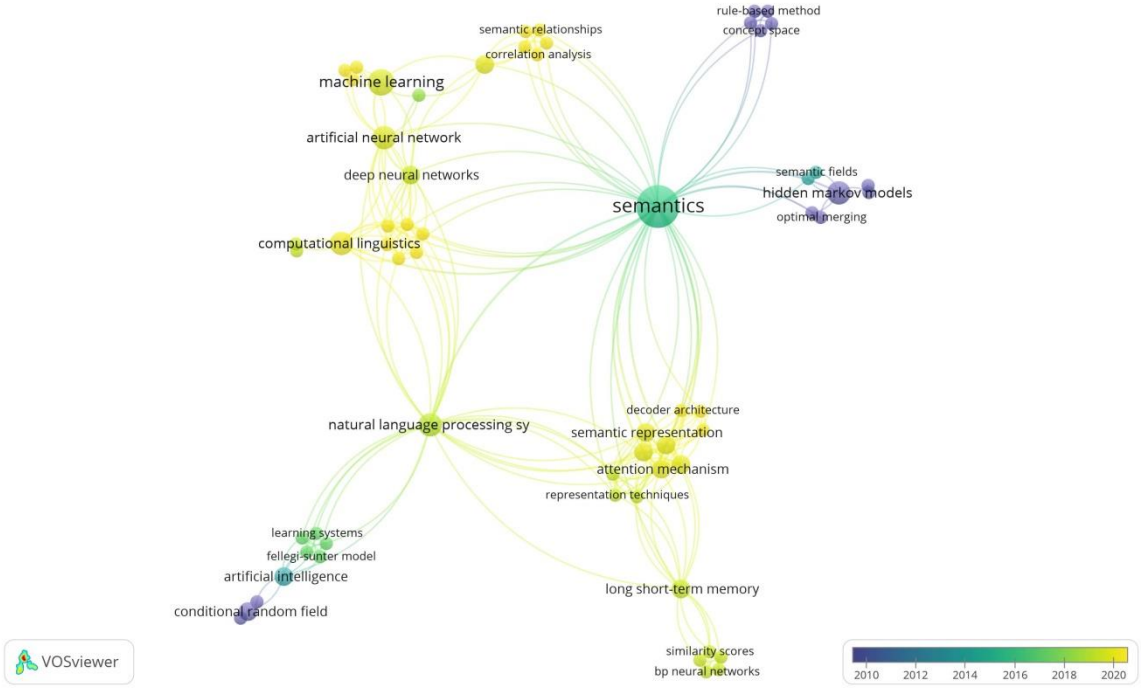


Figure 9. Network overlay of keyword occurrences related to methods/algorithms.

Table 6. Main algorithms/methods-related keywords (by no. of occurrences, average publication year, eigenvector, and betweenness centrality measures).

Keyword	No.	Avg. Pub. Year	Eigenvector (max.)	Betweenness (max.)
semantics	20	2016	1.00	0.24
natural language processing (nlp)	5	2017	0.81	0.13
attention mechanisms	4	2020	0.54	0.00
long short-term memory (lstm)	4	2019	0.54	0.07
artificial neural networks	5	2019	0.47	0.05
representation techniques	1	2019	0.46	0.00
vector spaces	1	2019	0.46	0.00
deep neural networks	3	2019	0.46	0.01
computational representations	4	2020	0.43	0.03
bidirectional encoder representations from transformer (bert)	2	2020	0.42	0.00
cluster analysis	2	2020	0.42	0.00
numerical model	1	2020	0.42	0.00
decoder architecture	1	2020	0.34	0.00
gcn	1	2020	0.34	0.00
word-embeddings	1	2020	0.34	0.00
recurrent neural networks	2	2020	0.25	0.01
correlation analysis	1	2021	0.16	0.00
gru	1	2021	0.16	0.00
spatial correlation	2	2021	0.16	0.00
machine learning	5	2019	0.12	0.00
concept space	1	2009	0.12	0.00
latent semantics	1	2009	0.12	0.00

Table 6. *Cont.*

Keyword	No.	Avg. Pub. Year	Eigenvector (max.)	Betweenness (max.)
rule-based approach	2	2009	0.12	0.00
supervised learning methods	2	2014	0.12	0.00
artificial intelligence	2	2014	0.12	0.04
hidden markov models (hmm)	6	2012	0.12	0.03
fellegi-sunter model	1	2017	0.12	0.00
learning algorithms	1	2017	0.12	0.00
learning systems	1	2017	0.12	0.00
probabilistic record linkage	1	2017	0.12	0.00
optimal merging	1	2005	0.10	0.00
bp neural networks	1	2019	0.07	0.00
conditional random fields (crf)	3	2013	0.07	0.01
lm-lstm-crf model	1	2019	0.07	0.00
word2vec	1	2020	0.06	0.00
latent variable	1	2019	0.04	0.00
linear chain	1	2019	0.04	0.00
random forests	3	2017	0.03	0.01
conditional functional dependencies	1	2018	0.02	0.00
decision trees	1	2018	0.02	0.00
back propagation algorithm	1	2015	0.02	0.00
ripple down rules	1	2010	0.01	0.00
learning address parsers	1	2006	0.01	0.00
probabilistic gis	1	2006	0.01	0.00
dirichlet process	2	2010	0.01	0.00
variational techniques	1	2010	0.01	0.00
contrast learning	1	2021	0.00	0.00
ensemble learning	1	2018	0.00	0.00
trees (mathematics)	1	2016	0.00	0.00
word-level-tree	2	2016	0.00	0.00
bi-gru neural network	1	2020	0.00	0.00
likelihood functions	1	2017	0.00	0.00
probabilistic modeling	1	2017	0.00	0.00
automated geocoding	1	2014	0.00	0.00
deep learning	1	2018	0.00	0.00
suffix trees	1	2007	0.00	0.00
symbolic object	1	2005	0.00	0.00

3.2. Discussion and Future Research

3.2.1. Detailed Literature Review

The present section aims to perform a more detailed discussion of the different address matching algorithms based on the full text of the selected articles (also summarized in Appendix A), with a view to extend the previously presented keyword-based analysis. This more detailed literature review will be organized around the three main methods which have been found to be the most relevant: string similarity-based methods, address element-based methods, and deep learning methods [9].

String similarity-based methods consist of a standard approach for address matching and generally involve the computation of a similarity metric between the addresses under comparison. Three main methods can be identified: character-based, vector-space based, and hybrid approaches [44]. Character-based methods comprehend edition operations, like sub-sequence comparisons, deletions, insertions, and substitutions. One of the best-known character-based methods is the Levenshtein edit distance metric [45], consisting of the minimum number of insertions, substitutions, or deletions which are required to convert a string into another (for instance, the edit distance between the toponyms Lisboa and Lisbonne is three, since it requires two insertions and one substitution) [44]. Another example of a character-based method is the Jaro metric [46], specifically conceived for matching short strings, like person names, with a more advanced version being latter

proposed (Jaro–Winkler similarity) [47], in order to give higher scores to strings matching from the beginning up to a given prefix length. Regarding vector-space approaches, the calculation of the cosine similarity between representations based on character n-grams (i.e., sequences of n consecutive characters) consists of a common approach, alongside the Jaccard similarity coefficient [44]. Lastly, hybrid metrics, while combining the advantages of the two previous approaches, also allow for small differences in word tokens and are more flexible in what concerns to word order and position [44]. Nevertheless, in terms of performance, there is not a best technique. The available metrics are task-dependent and, according to the study developed by Santos et al. [44], involving the comparison of thirteen different string similarity metrics, the differences in terms of performance are not significant, even when combined with supervised methods, to avoid the manual tuning of the decision threshold (one of the most important factors to obtain good results).

Address element-based methods, on their turn, rely on address parsing, a sequence tagging task which has been traditionally approached using probabilistic methods mainly based on Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [3], alongside other less common approaches not always involving machine learning methods.

In what concerns the application of HMMs in the context of residential addresses, the hidden states correspond to each segment of the address and the observations consist of the tokens assigned to each word of the input address string (after the application of some cleaning procedures), which may be based on look-up tables and hard-coded rules [6]. For instance, the address "17 Epping St Smithfield New South Wales 2987", after cleaning and tokenization, would turn into the following:

'17' (NU), 'epping' (LN), 'street' (WT), 'smithfield' (LN), 'nsw' (TR), '2987' (PC)

where 'NU' would stand for other numbers, 'LN' for locality (town, suburb) names, 'WT' for wayfare type (street, road, avenue, etc.), 'TR' for territory (state, region), and 'PC' for postal (zip) code [6] (p. 6). In order to determine, by statistical induction, the most likely arrangement of hypothetical "emitters" behind the observed sequence, a set of training examples is used to learn both the transition matrix and the observation matrix, through the maximum likelihood approach. Since it is computationally infeasible to evaluate the probability of every possible path (for N states and T observations, there would be N^T different paths), the Viterbi algorithm is used to find the most probable path through the model [48]. As such, the most probable sequence of states, based on previously trained transition and emission matrices, will present the highest probability of occurring, as illustrated below, in which the observation symbols are in brackets and the emission probabilities are underlined [6] (p. 7):

Start -> Wayfare Number (NU) -> Wayfare Name (LN) -> Wayfare Type (WT) -> Locality (LN) -> Territory (TR) -> Postal Code (PC) -> End

$$0.9 \times 0.9 \times 0.95 \times 0.1 \times 0.95 \times 0.92 \times 0.95 \times 0.8 \times 0.4 \times 0.94 \times 0.8 \times 0.85 \times 0.9 = 1.18 \times 10^{-2}$$

One of the main drawbacks of traditional HMMs is the fact that they do not support multiple simultaneous observations for one token. Even in more advanced versions of HMMs such as entropy Markov Models [49], in which the current state depends both on the previous state and on existing observations, there is a weakness called the label bias problem [50]: "transitions leaving a given state to compete only against each other, rather than against all transitions in the model" [41] (p. 2).Within the present literature review, four of the considered articles propose HMM-based methods: the already mentioned one by Churches et al. [6], aiming at the preparation of name and address data for record linkage purposes, through a combined approach using lexicon-based tokenization and HMMs, with the obtained experimental results confirming it as an alternative to rule-based systems that is both feasible and cost-effective; a second paper by the same authors [51], in which a geocoding system based on HMMs and a rule-based matching engine (*Febrl*)

for spatial data analysis is proposed and tested on small datasets of randomly selected addresses from different sources, with experimental results pointing to exact matches rates between 89% and 94%, depending on the source and considering the total exact matches obtained at various levels (address level, street level and locality level); the paper by X. Li et al. [40], in which an HMM-based large scale address parser is proposed, obtaining an accuracy of 95.6% (F-measure), after being tested on data from various sources with varying degrees of quality and containing billions of registers, of which 20% were synthetically rearranged in order to reproduce normal address variations; and, finally, the paper by Fu et al. [52], in which an HMM-based segmentation and recognition algorithm is proposed for the development of automatic mail-sorting systems involving handwritten Chinese characters (a problem which will be further addressed in the present literature review), with experimental results confirming its effectiveness.

Conditional Random Fields (CRFs) consist of a recent innovation in the field of text segmentation. CRFs are conditional by nature and assume no independence between output labels, illustrating real world addresses, in which zip codes, for instance, are related to city names, localities, and even streets [3]. Having all the advantages of Maximum entropy Markov models (MEMMs), CRFs also solve the label bias problem by letting the probability of a transition between labels also depend on past and future elements and not only on the current address element [3]. "The essential difference between CRFs and MEMMs is that the underlying graphical model structure of CRFs is undirected, while that of MEMMs is directed" [41] (p. 2). Considering, as an example, the address "3B Records, 5 Slater Street, Liverpool L1 4BW", an HMM parser would erroneously predict the first and second labels as standing for number ('3B') and street ('Records'), respectively, whereas the CRF parser, when reaching the actual property number (5), would give a higher score to the current label in order to revise it to a property number and the previous label (3B Records) to a business name [3]. Another recent approach to address parsing is based on so-called "word-embeddings", the name given to the vector representation of words [3]. An implementation of such method is word2vec [53], an unsupervised neural network language which aims to make predictions about the next words by modeling the relationships between a given word and the words in its context, based on two possible architectures: the continuous skip-gram model (Skip-Gram) and the continuous bag-of-words model (CBOW) [53]. The latter is usually chosen over the former, since it is trained by inferring the meaning of a particular word from its context [9].

A practical comparison between HMMs, CRFs, and a CRF augmentation with word2vec is undertaken in Comber and Arribas-Bel [3]. The VGI based *Libpostal* library (https://github.com/openvenues/libpostal (accessed on 9 November 2021)), which trains a CRF model on 1 billion street addresses from OSM data, was used for the segmentation task. Although the obtained results are broadly consistent in terms of precision, the classifiers using the HMM technique present lower recall values than the ones obtained by the CRF, meaning that both methods are capable of distinguishing true positives from false positives, but the CRF is able to classify a greater proportion of matches [3]. The augmented version of the CRF model does not outperform the results obtained by the original one but presents the advantage of not committing the user to a particular string distance and its biases [3]. In another recent work by the same author [54], a predictive model for address matching is proposed, based on recent innovations in machine learning and on a CRF parser for the segmentation of address strings. The biggest contribution of the paper at hand, however, is the thorough documentation of all the steps required to execute the proposed model's workflow. In other papers included in the present literature review, CRFs are used as a benchmark model, e.g.,: Dani et al. [55] or in combination with other methods, which will be further addressed [56–58].

Other less recent approaches have been proposed for address parsing/segmentation, namely within address standardization studies aiming at minimizing the size of labelled training data. One such example is the work by Kothari et al. [59], in which a nonparametric Bayesian approach to clustering grouped data, known as hierarchical Dirichlet process

(HDP) [60], is used with a view to discover latent concepts representing common semantic elements across different sources and allow the automatic transfer of supervision from a labeled source to an unlabeled one. The obtained latent clusters are used to segment and label address registers in an adapted CRF classifier, with experimental results pointing to a considerable improvement in classification accuracy [59]. A similar approach is proposed by Guo et al. [61], in which paper a supervised address standardization method with latent semantic association (LaSA) is presented, with a view to capture latent semantic association among words in the same domain. The obtained experimental results show that the performance of standardization is significantly improved by the proposed method. Expert systems have also been proposed, namely by Dani et al. [55], in which paper a Ripple Down Rules (RDR) framework is proposed with a view to enable a cost-effective migration of data cleansing algorithms between different datasets. RDR allows the incremental modification of rules and to add exceptions without unwanted side effects, based on a failure driven approach in which a rule is only added when the existing system fails to classify an instance [55]. After comparison with traditional machine learning algorithms and a commercial system, experimental results show that the RDR approach requires significantly less rules and training examples to reach the same accuracy as the former methods [55].

Tree-based models have been proposed to handle automatic handwritten address recognition, which consists of a particular address parsing/segmentation task mostly studied by Chinese researchers, due to the greater complexity of the Chinese language (larger character set, different writing styles, great similarity between many of the characters) [62]. In the paper by Jiang et al. [62], a suffix tree is proposed to store and access addresses from any character. In relation to previous approaches also based on a tree data structure, the proposed suffix tree is able to deal with noise and address format variations. Basically, a hierarchical substring list is firstly built, after which the obtained input radicals are compared with candidate addresses (filtered by the postcode) with a view to optimize a cost function, combining both recognition and matching accuracy [62]. A correct classification rate of 85.3% is obtained in the experimental results. However, according to Wei et al. [22], the recognition accuracy of character-level-tree (CLT) models is dependent on the completeness of the address list on which they are based. In order to overcome this limitation, the authors propose a structure tree built at the word level (WLT), in which each node consists of an address word and the path from the roof to the leaf corresponds to a standardized address format. After initial recognition by a character classifier, segment candidate patterns are mapped to candidate address words based on the WLT database. In the final phase (path matching), candidate address words' scores are summed in order to obtain the address recognition result [22]. The obtained experimental results show that the proposed method outperforms four benchmarking methods, including the previously mentioned suffix tree. Address tree models for address parsing and standardization are also proposed in the papers by Tian et al. [63], Liu et al. [64] and Li et al. [65]. In the first two, the address tree model is mainly used for rule-based validation and error recognition, by providing information about the hierarchy of Chinese addresses and, in the case of the paper by Li et al. [65], latent tree structures are designed with a view to capture rich dependencies between the final segments of an address, which do not always follow the same order.

Within the address element-based methods, it is also worth highlighting geocoding as a means to enhance address standardization, through the correction of misspellings and the filling of missing attributes, some of the most common errors found in postal addresses [66]. After successful matching with a record from a standardized reference database (like Google Maps or OSM), reverse geocoding can be performed to obtain a valid and complete representation of the queried address. In the case of geocoded databases like GNAF, for instance [51], geographic coordinates can be used to calculate the spatial proximity between different records for conducting distance-based spatial analyses and for record linking purposes between different databases (up to the house number). Another important application of address geocoding relates to the matching of historical address

records (such as census records) with contemporary data, by attaching grid references to the former in order to perform longitudinal spatial demographic analyses [27]. However, the successful automated geocoding of residential addresses depends on a number of factors, namely population densities (with positional error increasing as population density decreases) [27,67], the completeness of an address (existence or not of a number and street name), and changes in street names, among others [27]. These limitations can be tackled by the previous standardization and enrichment of addresses [68] and the choice of the most adequate geocoding method, including the use of property data [67] or the use of hybrid geocoding approaches [28].

With the advancement of deep learning methods, various authors have recently proposed the adoption of the previously mentioned extensions to RNNs (namely, LSTMs, GRUs, and GCNs), in order to better cope with nonstandard address records and highly dissimilar toponyms. LSTMs and GRUs are both composed of gates, which consist of neural networks that regulate the flow of information from one time step to the next, thereby helping to solve the short memory problem. In particular, GRUs have two gates—update and reset gates—and LSTMs, three gates—input, forget, and output gates [30]. The amount of fresh information added through the input gate in LSTMs is unrelated to the amount of information retained through the forget gate. In GRUs, the retention of past memory and the input of new information to memory are not mutually exclusive. The GRU stores both long-term dependence and short-term memories in a single hidden state, whereas the LSTM stores the former in the cell state and the latter in the hidden state. Because there are fewer weights and parameters to update during training, GRUs are faster to train than LSTMs [30].

Within the present literature review, several of the considered papers propose these types of methods, namely the ones by Santos et al. [35], Lin et al. [9], J. Liu et al. [58], Shan et al. [7,29], P. Li et al. [69], and Chen et al. [70]. To take into account contextual information both from previous and future tokens, by processing the sequence in two directions, bidirectional LSTM (BiLSTM) or GRU layers are also being employed in the great majority of these studies. The best performing models further connect the encoder and decoder through an attention mechanism in order to assign higher weights to the most important features [7,9,29]. With a view to reduce overfitting and enhance the classification models' generalization abilities, a dropout regularization layer is also normally added [9,35,58]. The ESIM model [71] consists of an illustrative example of a deep learning architecture based on the principles previously described. After address tokenization (with the help of gazetteers and dictionaries, in the case of more complex languages, with no natural separators) and the obtaining of vector representations of the different (labelled) address pairs (based on word2vec), the ESIM model is employed through the following four layers [9]:

- An input encoding layer, that encodes the input address vectors and extracts higher-level representations using the bidirectional long short-term memory (BiLSTM) model;
- A local inference modelling layer, that makes local inference of an address pair using a modified decomposable attention model [72];
- An inference composition layer, responsible for making a global inference between two compared address records based on their local inference, in which average and max pooling are used to summarize the local inference and output a final vector with a fixed length;
- Finally, a prediction layer, based on a multilayer perceptron (MLP) composed of three fully connected layers with rectified linear unit (ReLU), tanh and softmax activation functions, is used to output the predictive results of address pairs (that is, whether there is a match or not).

In terms of performance, all of the previously presented deep learning methods achieve a greater matching accuracy than the traditional text-matching models. In the case of the BiLSTM model proposed by Lin et al. [9], the precision, recall, and F1 score on the test set all reached 0.97, against the 0.92 scores achieved by the second-best performing model (Jaccard similarity coefficient + RF method). The deep neural network based on GRUs, to categorize

toponym pairs as matches or non-matches, proposed by Santos et al. [35], also outperforms traditional text-matching methods, achieving an increase of almost 10 points in most of the evaluation metrics (namely, accuracy, precision, and F1). The LM-LSTM-CRF+BP neural networks model proposed by J. Liu et al. [58] achieves an accuracy and F1 score of 87%, compared with average scores of 70% by the benchmark methods (word2vec and edit distance). The address GCN method proposed by Shan et al. [7] also presents better results, on both precision (up to 8%) and recall (up to 12%), than the existing methods, which include the DeepAM model previously proposed by the same author, based on an encoder-decoder architecture with two LSTM networks [29]. The Bi-GRU neural network proposed by P. Li et al. [69] presents a similar performance to that shown by a Bi-LSTM neural network (F1 score of 99%) and a higher performance than unidirectional GRU and LSTM neural networks (F1 score of 93%), as it would be expected. Finally, the attention-Bi-LSTM-CNN network (ABLC) proposed by Chen et al. [70] achieves an improvement of 4–10% more accuracy than the baseline models, which include the previously mentioned ESIM model, presenting the second-best overall performance.

In two of the most recent studies included in the present literature review also based on deep learning methods, bidirectional encoder representations from Transformers (BERT) are proposed instead. The first one is the study by Xu et al. [20], which proposes a method for fusing addresses and geospatial data based on BERT (in what concerns the learning of addresses' semantics) and a K-Means high-dimensional clustering algorithm, enhanced by innovative fine-tuning techniques, to create a geospatial-semantic address model (GSAM). The computational representation extracted from GSAM is then employed for predicting address location, based on a neural network architecture to minimize the Euclidean distance between the predicted and real coordinates. In the second study [37], a new address element recognition method is proposed, for dealing with address elements with shortchange periods (streets, lanes, landmarks, points of interest names, etc.) which still have not been included in a segmentation dictionary. A model based on BERT is first applied to obtain the vector representations of the dataset and learn the contextual information and model address features, followed by the use of a CRF to predict the tags, with new address elements being recognized according to the tag [37].

In terms of performance, the GSAM model [20] achieves a classification accuracy above 0.97, against a minimum expected accuracy of 0.91 by other methods; the BERT-CRF model [37] achieves the highest F1 score on generalization ability (0.78), when compared to benchmark models combining word2vec, BiLSTM, and CRF methods (with an average F1 score of 0.41), as well as an equally high F1 score on the testing dataset (0.95).

Although related to POIs' locations and descriptions, two final articles (both published in 2021) are worth mentioning, due to the combined use of the previously presented approaches and spatial correlation/reasoning methods. The first of this studies [2] presents a method for identifying POIs in large POI datasets in a quick and accurate manner, based on: an enhanced address-matching algorithm, combining string, semantic, and spatial similarity, within an ontology model describing POIs' locations and relationships, in order to support the transition from semantic to spatial; a grid-based algorithm capable of achieving compact representations of vast qualitative direction interactions between POIs and performing quick spatial reasoning, through the fast retrieval of direction relations and quantitative calculations. The second of the studies [8] proposes an unsupervised method to segment and standardize POIs' addresses, based on a GRU neural network combined with the spatial correlation between address elements for the automatic segmentation task, and a tree-based fuzzy matching of address elements for the standardization task, with experimental results pointing to a relatively high accuracy.

3.2.2. Research Gaps

Within the more recently published papers considered in the present literature review, the most relevant opportunities for further work can be summarized as follows: the use of representative and large enough datasets [20]; the inclusion of duplicate place names,

in order to enable the application of the proposed methodology to a national address database [9]; to improve accuracy, different weights might be assigned to the address-element vectors depending on their hierarchy [9]; the need to fine tune the weight ratio of fused features, such as coordinates and the semantic representation of addresses, alongside the improvement of the underlying concatenation method and measurement metrics [20]; the adoption of systematic approaches for tuning hyper-parameters and experimenting with different architectures [35]; the need to involve more complex spatial objects and relations [2,8]. Some of the limitations highlighted in less recent studies, however, should also be taken in consideration in the application of the most recent methods, like the need to tackle privacy and confidentiality issues [51] when using personal quasi-identifiers such as addresses (especially, residential ones). Another concern that should be addressed and which was tackled in some of the earlier studies [55,59,61] is related to the minimization of human labelling when generating both training and test data. Lastly, no references have been found about the use of genetic programming (GP) [73] in the field of semantic address matching. GP has several advantages over other machine learning methods, including the ability to provide results that can be easily interpreted, based on programs, rules, or functions, as well as the ability to easily incorporate specific knowledge about a problem, despite its efficiency issues, which are primarily due to a time-consuming fitness function computation [74]. In Figure 10, the main research gaps are illustrated.

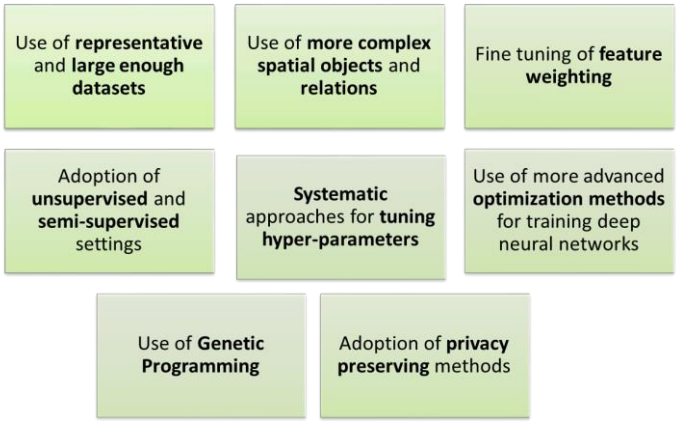


Figure 10. Main research gaps identified in the present SLR.

4. Conclusions

In this study, a systematic literature review based on Scopus and Web of Science, covering a time span of 20 years was undertaken in order to better understand how past and current limitations to address matching have been and can be overcome through the adoption of automated approaches to address matching. For the screening of the articles initially found, the PRISMA guidelines were followed, resulting in a final set of 41 relevant papers from high ranked Journals and Conference Proceedings. VOSviewer, a bibliometric analysis tool, was used to perform cluster analysis on the relationships between authors and popular research topics. The number of published articles has been increasing since 2017, a trend that may be closely related to the application of deep learning methods in this field. Chinese authors lead the way, with 19 papers that represent 46% of the published articles, half of them published between 2019 and 2021, after a peak in 2016 (11%). Disease control (covid-19), location-based services, and GIS/census/urban planning stand as some of the most recent application domains in the field under study. The research seems to confirm that probabilistic methods (such as HMMs and CRFs) have been outplaced by NLP methods based on semantics, encoder-decoder architectures, and attention mechanisms. There also seems to exist some evidence pointing to the very recent adoption of hybrid approaches with an increased use of spatial constraints and entities. It should be noted, however, that this review has some limitations, such as the subjectivity of the search query and screening procedures. As such, a more effective search query should be considered

in future research with a view to avoid the exclusion of potentially relevant papers. In spite of its limitations, the present review presented a concise and detailed overview of the research being produced in the field of automated address matching, within a considerably long time span, of 20 years. Future studies can develop upon its main findings, mainly in what concerns the improved use of the identified deep learning algorithms, in terms of the adoption of unsupervised or semi-supervised settings, optimization strategies for deep neural network training and/or systematic approaches to hyper-parameter tuning, as well as of privacy preserving methods, namely when dealing with residential addresses acting as quasi-identifiers in record linkage processes. The use of more complex spatial objects and relations, as means to enhance address matching and standardization, consists of another important gap to address, namely in domains not limited to POIs' retrieval. Lastly, no references have been found to evolutionary-based approaches in the field of semantic address matching, which may also be a potential research gap to address in future studies.

Author Contributions: Conceptualization, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; methodology, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; software, Paula Cruz; validation, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; formal analysis, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; investigation, Paula Cruz; resources, Paula Cruz; data curation, Paula Cruz; writing—original draft preparation, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; writing—review and editing, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; visualization, Paula Cruz; supervision, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; project administration, Paula Cruz, Leonardo Vanneschi, Marco Painho and Paulo Rita; funding acquisition, Leonardo Vanneschi. All authors have read and agreed to the published version of the manuscript.

Funding: The work by Leonardo Vanneschi, Marco Painho and Paulo Rita was supported by Fundação para a Ciência e a Tecnologia (FCT) within the Project: UIDB/04152/2020—Centro de Investigação em Gestão de Informação (MagIC). The work by Prof. Leonardo Vanneschi was also partially supported by FCT, Portugal, through funding of project AICE (DSAIPA/DS/0113/2019).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Application and Methods' Analysis

Id. [Ref.]	Authors, Pub. Year	Application	Methods
1 [9]	Lin et al., 2019	Geocoding	Word2vec; bi-directional LSTM (ESIM model)
2 [52]	Fu et al., 2005	Handwritten address character string segmentation and recognition	Hidden Markov Model (HMM)
3 [55]	Dani et al., 2010	Address standardization; data quality improvement	Ripple Down Rules (RDR); conditional random field (CRF)
4 [51]	Christen et al., 2006	Geocoding	Learning address parser based on hidden Markov models and a rule-based matching engine
5 [62]	Jiang et al., 2007	Address recognition system	Suffix tree based system
6 [75]	Song, 2013	Location-based services	Natural language understanding
7 [61]	Guo et al., 2009	Address standardization	Free-text address standardization method with latent semantic association (LaSA).
8 [69]	P. Li et al., 2020	Geocoding	Bidirectional gated recurrent unit (GRU) neural network
9 [27]	Walford, 2019	Geocoding historical census records	Four-stage semi-automated method to geocode historical census addresses

Id. [Ref.]	Authors, Pub. Year	Application	Methods
10 [76]	Verma and Kaur, 2015	Character recognition from handwritten document	Neural Networks
11 [58]	J. Liu et al., 2019	Financial anti-fraud	LM-LSTM-CRF
12 [77]	Choi et al., 2017	Probabilistic record linkage; Entity resolution	Similarity functions (ex. Jaro-Winkler); Fellegi-Sunter model
13 [29]	Shan et al., 2019	Location-based services	Encode-decoder architecture with two LSTM networks and an attention mechanism
14 [54]	Comber, 2019	Spatial socio-economic applications	CRF; string similarity functions; Random Forest
15 [28]	Shah et al., 2014	Geocoding for public health research	Geocoding methods
16 [57]	Weinman, 2017	Historical Map Alignment and Toponym Recognition	Semi-Markov CRF text recognizer; Caffe-based CNN
17 [7]	Shan et al., 2020	Location-based services	Encode-decoder architecture with two LSTM networks and an attention mechanism; GCN
18 [20]	Xu et al., 2020	Management and application of non-standard addresses	Bidirectional encoder representations from Transformers (BERT); high-dimensional clustering algorithm to fuse semantic and geospatial information
19 [64]	Q. Liu et al., 2018	Handwritten address character string recognition	Deep neural network for character recognition (CNNs); domain specific knowledge for address recognition
20 [40]	X. Li et al., 2014	Record linkage	HMM
21 [1]	Javidaneh et al., 2020	Evaluate the influence of formal addressing systems on spatial knowledge acquisition	Agent-based simulation of spatial knowledge acquisition
22 [5]	Lee et al., 2020	Geocoding	Regex for address parsing; support vector machine (SVM), random forest (RF), extreme gradient boosting (XGB) for address matching
23 [44]	Santos et al., 2017	Geographical information retrieval	13 different string similarity metrics; supervised machine learning methods for combining the scores (Support Vector Machines, Random Forests, Extremely Randomized Trees, Gradient Boosted Trees)
24 [3]	Comber and Arribas-Bel, 2019	Record linkage	word2vec; CRFs
25 [65]	H. Li et al., 2019	Parsing of non-standard addresses	Neural structured prediction models with latent variables (latent tree structures and regular chain structures)
26 [6]	Churches et al., 2002	Record linkage	HMM
27 [37]	Zhang et al., 2020	Location-based services	BERT; CRF
28 [22]	Wei et al., 2016	Recognition of handwritten non-standard address	Word-level-tree (WLT) based method
29 [56]	Tang et al., 2010	Toponym resolution	Geo-parsing approach based on CRF; geo-coding approach based on partial fuzzy matching
30 [78]	Nagabhushan et al., 2005	Postal automation	Symbolic knowledge base supported address validation system

Id. [Ref.]	Authors, Pub. Year	Application	Methods
31 [35]	Santos et al., 2018	Toponym recognition	Bidirectional GRUs
32 [59]	Kothari et al., 2010	Address cleansing (with transfer of supervision)	Hierarchical Dirichlet process
33 [63]	Tian et al., 2016	Geocoding	Address tree model; Lucene fuzzy matching
34 [21]	Peng et al., 2020	COVID-19 Epidemic Prevention and Control	Word segmentation weighted address matching algorithm considering a variety of semantics
35 [8]	Luo et al., 2021	Address standardization of POIs	GRU; spatial correlation
36 [70]	Chen et al., 2021	Address semantic matching	Attention-Bi-LSTM-CNN
37 [66]	Koumarelas et al., 2018	Enhancing address matching	CRF; geocoding; similarity measures
38 [68]	Cortes et al., 2021	Improving geocoding matching rates of structured addresses	Regular expressions and dictionary-based methods for address standardization and enrichment; geocoding
39 [67]	Cayo and Talbot, 2003	Evaluation of positional error in automated geocoding of residential addresses	GIS
40 [2]	Cheng et al., 2021	Locating POIs in large datasets	Combination of multiple similarities (string, semantic and spatial); grid-based spatial reasoning algorithm
41 [79]	Florczyk et al., 2010	Urban management	Compound geocoding architecture, based on gazetteers, cadastral services and address geocoding services

References

1. Javidaneh, A.; Karimipour, F.; Alinaghi, N. How Much Do We Learn from Addresses? On the Syntax, Semantics and Pragmatics of Addressing Systems. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 317. [CrossRef]
2. Cheng, R.; Liao, J.; Chen, J. Quickly Locating POIs in Large Datasets from Descriptions Based on Improved Address Matching and Compact Qualitative Representations. *Trans. GIS* **2021**, 1–26. [CrossRef]
3. Comber, S.; Arribas-Bel, D. Machine Learning Innovations in Address Matching: A Practical Comparison of Word2vec and CRFs. *Trans. GIS* **2019**, *23*, 334–348. [CrossRef]
4. Sun, Y.; Ji, M.; Jin, F.; Wang, H. Public Responses to Air Pollution in Shandong Province Using the Online Complaint Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 126. [CrossRef]
5. Lee, K.; Claridades, A.R.C.; Lee, J. Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 5628. [CrossRef]
6. Churches, T.; Christen, P.; Lim, K.; Zhu, J.X. Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models. *BMC Med. Inform. Decis. Mak.* **2002**, *2*, 9. [CrossRef] [PubMed]
7. Shan, S.; Li, Z.; Yang, Q.; Liu, A.; Zhao, L.; Liu, G.; Chen, Z. Geographical Address Representation Learning for Address Matching. *World Wide Web.* **2020**, *23*, 2005–2022. [CrossRef]
8. Luo, A.; Liu, J.; Li, P.; Wang, Y.; Xu, S. Chinese Address Standardisation of POIs Based on GRU and Spatial Correlation and Applied in Multi-Source Emergency Events Fusion. *Int. J. Image Data Fusion* **2021**, *12*, 319–334. [CrossRef]
9. Lin, Y.; Kang, M.; Wu, Y.; Du, Q.; Liu, T. A Deep Learning Architecture for Semantic Address Matching. *Int. J. Geogr. Inf. Sci.* **2019**, *34*, 559–576. [CrossRef]
10. Wang, J.; Deng, H.; Liu, B.; Hu, A.; Liang, J.; Fan, L.; Zheng, X.; Wang, T.; Lei, J. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine over the Past 20 Years: Bibliometric Study on Pubmed. *J. Med. Internet Res.* **2020**, *22*, e16816. [CrossRef] [PubMed]
11. Melo, F.; Martins, B. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Trans. GIS* **2017**, *21*, 3–38. [CrossRef]
12. Kayed, M.; Dakrory, S.; Ali, A.A. *Postal Address Extraction from the Web: A Comprehensive Survey*; Springer: Dordrecht, The Netherlands, 2021. [CrossRef]
13. Barrington-Leigh, C.; Millard-Ball, A. The World's User-Generated Road Map Is More than 80% Complete. *PLoS ONE* **2017**, *12*, e0180698. [CrossRef] [PubMed]

14. Yassine, M.; Beauchemin, D.; Laviolette, F.; Lamontagne, L. Leveraging Subword Embeddings for Multinational Address Parsing. In Proceedings of the 2020 6th IEEE Congress on Information Science and Technology (CiSt), Agadir-Essaouira, Morocco, 5–12 June 2021.
15. Goldberg, D.W.; Wilson, J.P.; Knoblock, C.A. From Text to Geographic Coordinates: The Current State of Geocoding. *URISA J.* **2007**, *19*, 33–46.
16. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *PLoS Med.* **2021**, *18*, 372. [CrossRef]
17. Van Eck, N.J.; Waltman, L. Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]
18. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Icwsm* **2009**, 361–362.
19. Lin, Y.; Kang, M.; He, B. Spatial Pattern Analysis of Address Quality: A Study on the Impact of Rapid Urban Expansion in China. *Environ. Plan. B Urban Anal. City Sci.* **2019**, *48*, 728–740. [CrossRef]
20. Xu, L.; Du, Z.; Mao, R.; Zhang, F.; Liu, R. GSAM: A Deep Neural Network Model for Extracting Computational Representations of Chinese Addresses Fused with Geospatial Feature. *Comput. Environ. Urban Syst.* **2020**, *81*, 101473. [CrossRef]
21. Peng, M.; Li, Z.; Liu, H.; Meng, C.; Li, Y. Weighted Geocoding Method Based on Chinese Word Segmentation and Its Application to Spatial Positioning of COVID-19 Epidemic Prevention and Control. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomat. Inf. Sci. Wuhan Univ.* **2020**, *46*, 808–815.
22. Wei, X.; Lu, S.; Wen, Y.; Lu, Y. Recognition of Handwritten Chinese Address with Writing Variations. *Pattern Recognit. Lett.* **2016**, *73*, 68–75. [CrossRef]
23. Bornmann, L.; Wohlrabe, K. *Normalisation of Citation Impact in Economics*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 120. [CrossRef]
24. Babalola, A.; Musa, S.; Akinlolu, M.T.; Haupt, T.C. A Bibliometric Review of Advances in Building Information Modeling (BIM) Research. *J. Eng. Des. Technol.* **2021**. [CrossRef]
25. Baraibar-Diez, E.; Luna, M.; Odriozola, M.D.; Llorente, I. Mapping Social Impact: A Bibliometric Analysis. *Sustainability* **2020**, *12*, 9389. [CrossRef]
26. Liu, X. Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 1852–1863. [CrossRef]
27. Walford, N.S. Bringing Historical British Population Census Records into the 21st Century: A Method for Geocoding Households and Individuals at Their Early-20th-Century Addresses. *Popul. Space Place* **2019**, *25*, e2227. [CrossRef]
28. Shah, T.I.; Bell, S.; Wilson, K. Geocoding for Public Health Research: Empirical Comparison of Two Geocoding Services Applied to Canadian Cities. *Can. Geogr.* **2014**, *58*, 400–417. [CrossRef]
29. Shan, S.; Li, Z.; Qiang, Y.; Liu, A.; Xu, J. *DeepAM: Deep Semantic Address Representation for Address Matching*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 3. [CrossRef]
30. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
31. Hochreiter, S.; Urgen Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 17351780. [CrossRef]
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
34. Thekumparampil, K.K.; Wang, C.; Oh, S.; Li, L.J. Attention-Based Graph Neural Network for Semi-Supervised Learning. *arXiv* **2018**, arXiv:1803.03735.
35. Santos, R.; Murrieta-Flores, P.; Calado, P.; Martins, B. Toponym Matching through Deep Neural Networks. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 324–348. [CrossRef]
36. Gori, M.; Monfardini, G.; Scarselli, F. A New Model for Learning in Graph Domains. *Proc. Int. Jt. Conf. Neural Netw.* **2005**, *2*, 729–734. [CrossRef]
37. Zhang, H.; Ren, F.; Li, H.; Yang, R.; Zhang, S.; Du, Q. Recognition Method of New Address Elements in Chinese Address Matching Based on Deep Learning. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 745. [CrossRef]
38. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
39. Ghahramani, Z. An Introduction to Hidden Markov Models and Bayesian Networks. *Int. J. Pattern Recognit. Artif. Intell.* **2001**, *15*, 9–42. [CrossRef]
40. Li, X.; Kardes, H.; Wang, X.; Sun, A. HMM-Based Address Parsing with Massive Synthetic Training Data Generation. *Int. Conf. Inf. Knowl. Manag. Proc.* **2014**, 33–36. [CrossRef]

41. Lafferty, J.; Mccallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data Abstract. In Proceedings of the 18th International Conference on Machine Learning 2001, San Francisco, CA, USA, 28 June–1 July 2001; pp. 282–289.
42. Blei, D.M.; Ng, A.Y.; Jordan, M.I.; Wallach, H.M.; Hinton, G.E.; Osindero, S.; Teh, Y.-W. Conditional Random Fields: An Introduction. *Neural Comput.* **2004**, *18*, 1–9. [CrossRef]
43. Borgatti, S.P. Centrality and Network Flow. *Soc. Netw.* **2005**, *27*, 55–71. [CrossRef]
44. Santos, R.; Murrieta-Flores, P.; Martins, B. Learning to Combine Multiple String Similarity Metrics for Effective Toponym Matching. *Int. J. Digit. Earth* **2017**, *11*, 913–938. [CrossRef]
45. Levenshtein, V.I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710. [CrossRef]
46. Jaro, M.A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Am. Stat. Assoc.* **1989**, *84*, 414–420. [CrossRef]
47. Winkler, W.E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proc. Sect. Surv. Res. Am. Stat. Assoc.* **1990**, 354–359.
48. Forney, G.D. The Viterbi Algorithm. *Proc. IEEE* **1973**, *61*, 268–278. [CrossRef]
49. McCallum, A.; Freitag, D.; Pereira, F. Maximum Entropy Markov Models for Information Extraction and Segmentation. In Proceedings of the 17th International Conference on Machine Learning, 2000, San Francisco, CA, USA, 29 June–2 July 2000.
50. Wang, M.; Haberland, V.; Yeo, A.; Martin, A.; Howroyd, J.; Bishop, J.M. A Probabilistic Address Parser Using Conditional Random Fields and Stochastic Regular Grammar. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016. [CrossRef]
51. Christen, P.; Willmore, A.; Churches, T. A Probabilistic Geocoding System Utilising a Parcel Based Address File. In *Data Mining*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3755, pp. 130–145. [CrossRef]
52. Fu, Q.; Ding, X.Q.; Liu, C.S.; Jiang, Y. A Hidden Markov Model Based Segmentation and Recognition Algorithm for Chinese Handwritten Address Character Strings. *Proc. Int. Conf. Doc. Anal. Recognit. ICDAR* **2005**, *2005*, 590–594. [CrossRef]
53. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
54. Comber, S. Demonstrating the Utility of Machine Learning Innovations in Address Matching to Spatial Socio-Economic Applications. *Region* **2019**, *6*, 17–37. [CrossRef]
55. Dani, M.N.; Faruquie, T.A.; Garg, R.; Kothari, G.; Mohania, M.K.; Prasad, K.H.; Subramaniam, L.V.; Swamy, V.N. A Knowledge Acquisition Method for Improving Data Quality in Services Engagements. In Proceedings of the 2010 IEEE International Conference on Services Computing, Miami, FL, USA, 5–10 July 2010; pp. 346–353. [CrossRef]
56. Tang, X.; Chen, X.; Zhang, X. Research on Toponym Resolution in Chinese Text. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomat. Inf. Sci. Wuhan Univ.* **2010**, *35*, 930–935.
57. Weinman, J. Geographic and Style Models for Historical Map Alignment and Toponym Recognition. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 957–964. [CrossRef]
58. Liu, J.; Wang, J.; Zhang, C.; Yang, X.; Deng, J.; Zhu, R.; Nan, X.; Chen, Q. *Chinese Address Similarity Calculation Based on Auto Geological Level Tagging Jing*; Springer International Publishing: Cham, Switzerland, 2019; Volume 2. [CrossRef]
59. Kothari, G.; Faruquie, T.A.; Subramaniam, L.V.; Prasad, K.H.; Mohania, M.K. Transfer of Supervision for Improved Address Standardization. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2178–2181. [CrossRef]
60. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc.* **2006**, *101*, 1566–1581. [CrossRef]
61. Guo, H.; Zhu, H.; Guo, Z.; Zhang, X.X.; Su, Z. Address Standardization with Latent Semantic Association. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 1155–1163. [CrossRef]
62. Jiang, Y.; Ding, X.; Ren, Z. A Suffix Tree Based Handwritten Chinese Address Recognition System. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 23–26 September 2007; Volume 1, pp. 292–296. [CrossRef]
63. Tian, Q.; Ren, F.; Hu, T.; Liu, J.; Li, R.; Du, Q. Using an Optimized Chinese Address Matching Method to Develop a Geocoding Service: A Case Study of Shenzhen, China. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 65. [CrossRef]
64. Liu, Q.; Wang, D.; Lu, H.; Li, C. *Handwritten Chinese Character Recognition Based on Domain-Specific Knowledge*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 2, pp. 221–231. [CrossRef]
65. Li, H.; Lu, W.; Xie, P.; Li, L. Neural Chinese Address Parsing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 3421–3431.
66. Koumarelas, I.; Kroschk, A.; Mosley, C.; Naumann, F. Experience: Enhancing Address Matching with Geocoding and Similarity Measure Selection. *J. Data Inf. Qual.* **2018**, *10*, 1–16. [CrossRef]
67. Cayo, M.R.; Talbot, T.O. Positional Error in Automated Geocoding of Residential Addresses. *Int. J. Health Geogr.* **2003**, *2*, 1–12. [CrossRef] [PubMed]

68. Cortes, T.R.; da Silveira, I.H.; Junger, W.L. Improving Geocoding Matching Rates of Structured Addresses in Rio de Janeiro, Brazil. *Cad. Saude Publica* **2021**, *37*, e00039321. [CrossRef]
69. Li, P.; Luo, A.; Liu, J.; Wang, Y.; Zhu, J.; Deng, Y.; Zhang, J. Bidirectional Gated Recurrent Unit Neural Network for Chinese Address Element Segmentation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 635. [CrossRef]
70. Chen, J.; Chen, J.; She, X.; Mao, J.; Chen, G. Deep Contrast Learning Approach for Address Semantic Matching. *Appl. Sci.* **2021**, *11*, 7608. [CrossRef]
71. Chen, Q.; Ling, Z.; Jiang, H.; Zhu, X.; Wei, S.; Inkpen, D. Enhanced LSTM for Natural Language Inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1657–1668. [CrossRef]
72. Parikh, A.P.; Täckström, O.; Das, D.; Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. *arXiv* **2016**, arXiv:1606.01933. [CrossRef]
73. Koza, J.R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*; MIT Press: Cambridge, MA, USA, 1992.
74. Araujo, L. Genetic Programming for Natural Language Processing. *Genet. Program. Evolvable Mach.* **2020**, *21*, 11–32. [CrossRef]
75. Song, Z. Address Matching Algorithm Based on Chinese Natural Language Understanding. *J. Remote Sens.* **2013**, *17*, 788–801.
76. Verma, A.; Kaur, G. Character Recognition from Handwritten Document Using Neural Networks. *Int. J. Appl. Eng. Res.* **2015**, *10*, 37574–37579.
77. Choi, S.C.T.; Lin, Y.; Mulrow, E. Comparison of Public-Domain Software and Services for Probabilistic Record Linkage and Address Standardization. *Lect. Notes Comput. Sci.* **2017**, *10344*, 51–66. [CrossRef]
78. Nagabhushan, P.; Angadi, S.A.; Anami, B.S. Symbolic Data Structure for Postal Address Representation and Address Validation through Symbolic Knowledge Base. *Lect. Notes Comput. Sci.* **2005**, *3776*, 388–394. [CrossRef]
79. Florczyk, A.J.; López-Pellicer, F.J.; Muro-Medrano, P.; Nogueras-Iso, J.; Zarazaga-Soria, F.J. Semantic Selection of Georeferencing Services for Urban Management. *Electron. J. Inf. Technol. Constr.* **2010**, *15*, 111–121.