ISSN: 0975-766X
CODEN: IJPTFI
Available Online through Review Article
www.ijptonline.com

A LITERATURE SURVEY ON ENTITY EXTRACTION TECHNIQUES IN BIO-MEDICAL DATASETS

[1]R.Sivashankari, [2]Dr. B.Valarmathi
School of Information Technology and Engineering, VIT University, Vellore, India.
Email: sivashankari.r@vit.ac.in

Received on 10-11-2016 *Accepted on: 28-11-2016*

Abstract:

The term "Entity Extraction" is a renowned technique now and broadly employed as a part of opinion mining and sentiment analysis. This technique is one of the information retrieval approaches, where the precise data have been obtained from the huge volume of information resources. Entity extraction is a traditional approach of named entity extraction, which is basically, retrieves the person name, product names, organization names, services, political issues titles, news article titles, locations and local and global time and etc. This entity extraction is very interesting filed and plays a major role in bio-medical field to recognize the drug entities, understand biological data in bioinformatics and categorize bio-medical documents, document summarization and explore the bio-medical terms and etc. Since the volume of online data has been increasing day by day, the entity extraction process needs more efficient techniques to find the entity replication, document replication, semantics between entities and documents and etc. This paper presents a study of various approaches employed in entity extraction on unstructured data.

Keywords: Named entity extraction, Machine learning, Data mining, Natural Language Processing, latent Dirichlet Allocation and Latent Semantic Information.

1. Introduction

Entity extraction is the process of locating a term (consists of single or sequence of words) in a sentence or in a phrase or document or collection of documents. Extraction of online data is a non-trivial task, because it is the process of retrieving structured data or useful data from unstructured data and this activity cannot be done as a manual work. Because the volume of online data is huge and time consuming to carry out this process is more. Thus the extraction process has been automated with the help of machine learning, data mining, statistical approaches and probabilistic approaches. The following section elaborates the detailed analysis of various approaches deployed in entity extraction and comparative analysis is made for researchers to relate the approaches which are related with entity extraction.Table 1 and Table 2 represents the various methods used for the entity extraction in biomedical field.

2. Machine learning models for entity extraction

Ayat et al. [1] Proposed two probabilistic approaches to find the similarity index between the entities in the databases. An entity e_i in database D is called x-tuple, if e_i contains several possible tuples and all the tuples are mutually exclusive. Mutual exclusion allows the values of two tuples are disjoint. For each tuple, the likelihood of truth is calculated in terms of probability. The maximum probability of a tuple shows more accurateness of values in the tuples. The similarity function is calculated to measure entity resolution. Entity resolution is the process of identifying multiple tuples that represent the same entity and convert those multiple tables into a single table. To calculate the similarity function, Context Free and Context Similarity Free probabilistic approaches were used. For the experimental purpose, UCI repository, Cora dataset, and Restaurant datasets are used. This proposed method has compared with other similar functions such as softTFIIDF and Flexi-softTFIDF, the experimental results shows that the proposed method outperforms compare than other functions.

Bellare et al [2], suggested a new approach on active algorithm called “IWAL active algorithm” for entity matching which determines whether the extracted entities are addressing the same object or different objects. Active algorithm is faster than supervised algorithms in labeling the data. Supervised learning classifier has to be trained with more samples compare to active algorithm. In active algorithm, the learner or the designer is free to choose the samples to classify interactively and the number samples taken by them is very less compare to supervised algorithms. One of the popular active algorithm “*Improved Weightage active algorithm*” introduced in [2], which improves recall compared with existing active algorithms. By using IWAL, the recall rate has been increased in entity matching. For this entity matching evaluation four different data sets such as Business listings used in the production system at Yahoo, Record linkage dataset from the UCI machine learning repository, DBLP-ACM and Scholar-DBLP are used.

Bhattacharya et al [3] recommended a combined entity resolution approaches to find the redundancy of research authors' names in research articles. Here, entity refers to author name of the research paper and entities are clustered if they refer the same object. For any two author names, similarity score has to be calculated using a greedy agglomerative clustering algorithm and based on the similarity score, the clusters are generated. Another approach used in [3] is a Common neighbor score, which is employed to every two clusters of entities to find the number of common entities between them. If the number of common entities is high, then the probability of two clusters refers to the same entity, thus the two clusters are grouped with each other.

Banjula et al [4] recommended a combined machine learning approaches to recognize the entities. To perform entity extraction, POS Tagger, feature set and decision tree techniques are used to extract the named entities. The feature set encodes the term to experiment for the entity. In this regard, the default notations of entities in the data are considered as the extractors. That is, a single character; term starts with a capital letter, alpha numeric terms, and only number terms are extracted as candidate entities. Unix Dictionary used to look up the term for entity recognition. If a term is not present in present in the dictionary, the term is considered as a candidate entity. Since dictionary does not contain the terms like person name, locations and etc. Even though the most of the extracted candidate terms are entities, there are few extracted terms are not entities. To remove those candidate terms, a popular machine language approach decision tree has employed. To make a decision tree as a classifier, the set of all rules was given to it to learn from the training set. With the help of rules, the classifier predicts the extracted candidate terms could be an entity or not.

Bashir et al [5] proposed a genetic programming (GP) to extract the opinion entities from the customer review corpus. In GP, three approaches are followed to extract the target entities. First, similarities of opinion entities are identified the entities whose opinions are same with respect to particular feature(s) in the review sentences. Since the data set is very large, every word or term in the review sentence cannot be compared with other words. To reduce the time complexity, the opinion words which are closer synonym with keywords in the user queries are considered to determine the similarity of opinionated entities. Second, the frequency of selected keywords in the review corpus is computed. The third approach is, the selected keywords frequency distribution (PL2) is calculated and the average score of all the keywords frequency distribution (avgPL2) is calculated. Based on the score of each opinion entity the polarity of that entity is determined. If the difference between positive score and negative is greater than 0, then the entity has a positive opinionated entity. Otherwise the entity has a negative opinionated. Based on the opinion score obtained by each entity, the rank has been assigned. If a positive score of any entity is higher than among other entities, then that entity has been ranked as number one. The experimental result shows that GP performance is better than RankBoost and RankSVM approaches.

Brill et al [6] recommended Semi-Markov model, a modified approach of Hidden Markov Model (HMM) to determine the similarity of extracting entities in the contiguous words in the sentences. A sentence may have a collection of entities such as a person's name, time, location and others. Each entity can be a single or sequence of words (terms). To segment a sentence into a set of terms based on their identified entity type, SMM probabilistic model has deployed. In this model a word (term) entity type depends on the previous word (term) entity type. A person name length is fixed maximum of two. The following example shows the SMM based similarity entity extraction in a sentence.

For example, “Rohit Agarwall house is located nearby Kamala house”

(Rohit Agarwal) $_{\text{Person}}$, (house) $_{\text{Location}}$ (is located) $_{\text{others}}$ and (Kamala house) $_{\text{Location}}$. {(1 ,2, Person), (3, 3, Location), (4, 5, Others), (6, 7, Location)}

In the above sentence, SMM found that Agarwal is the similar entity type of previous word Rohit and *house* cannot be adjoined with Rohit Agarwal because the person name length limited to 2. The proposed approach SMM is experimented with HMM-VP and HMM-VP $_{(4).}$ The experimental result shows that the proposed approach is better than compared with other approaches.

Ding et al [7] used Conditional Random Field to extract the product name (comment target) from review corpus. In this approach the set of fixed features of the entities identified manually. An entity is spotted by features around it. The set of fixed features occurrences in a sentence determines the position of an entity. The probability of fixed features occurrence is computed by CRF. The features of the entities are POS tagger, word and ontology.

Eka et al [8] recommended combined approaches of regular expression and logical regression to extract entities from SMS text messages. The extracted entities from the text message include name, date, location, telephone number and time. A set of regular expressions manually generated separately for each entity. Since the text messages are in Sweden language Granska POS Tagger is used to understand the entities. To evaluate the extracted entities using regular expression and POS Tagger, logical regression machine learning approach is used as a classifier. Set of features are annotated for logical regression to build a classifier from the training set. This set consists of a list of words which are frequently occurring in more than five documents, top twenty words which frequently precede entity classes, list of diagrams which precede entities, list of suffixes of entities and list of suffixes of named classes. LIBLINEAR package is used for logical regression.

Guo et al [9] proposed a rule based entity extraction engine called Jabari semantic to extract entities in sensor data. The proposed method used GATE /ANNIE information retrieval approach to extract candidate entities from the selected dataset. Jabari used multiple dictionaries to extract maximum entities in an efficient way. Guo also used regular expressions to extract various representations of date used in reviews. Since the user entered review is unstructured and not written in grammatically Hence, extracting entities becomes a non-trivial task.

Gruetze et al [10] proposed combined approaches for entity linking in the documents. The proposed combined approach named as Coherent and Efficient Named Entity Linking (CohEEL). To perform effective entity linking, two approaches are employed. One is Candidate classification and Judicious Neighborhood Exploration (JNE) respectively. In candidate classification, the communication between various entities across the documents are identified. To incorporate this higher–order scoring functions has employed. An iterative exploration model has been introduced to discover the relationship between the entities and group the entities which are related to each other.

Konkol et al [11] proposed combined approaches of the Hyperspace Analogue to Language (HAL) and Latent Semantic models to enhance the excellence of entity extraction from different language sentences. In this HAL approach, the words are converted into a vector space. The words in the dataset are categorized into two type's namely Local and Global contexts to build a vector space. The local context of a particular word is determined by four words of its left and right side. In case of global context the entire document is considered for a word or entity. The vector space contains the similarities of two words in the corpus. Since the vector space is large, LDA approach is used to perform the reduction. The reduction is carried out with transitive rules. Random Indexing (RI) is employed to find the similar value in the matrix for similar words. By using RI, co-occurrence words in the dataset are found. Another approach Purandare and Pedersen (P&P) has employed to create the cluster of words. All the words in a cluster have the same meaning. The proposed approach results analyzed with LDA, CRF, S-HAL and clustering methods and proposed methods scores high F1-score than other methods.

Korkontzelos et al [12] suggested a combined approach of genetic programming, DrugBank dictionary and regular expressions to extract drug entities from biomedical dataset. Regular expressions are used to check any overlapping term is exists in a sentence. If so, the overlapping terms are removed and multinomial logistic regression is used to classify the entities. DrugBank is then used to check whether the extracted term is a drug entity or not and extracted entities are experimented. Here, to evaluate the generated regular expression programming is used to evaluate the work of regular expression.

Michael [13] suggested a collection of neural network approaches to extract and rank the entities from the sentences. Voted perceptron, boosting algorithm and Maximum Entropy model is used to predict. And assign the tag to the word, boosting algorithm methods are used to perform the extraction as well as a ranking of the entities. Voted perceptron's has used to classify the candidate and boosting algorithm predicts entities in the sentences.

Florian et al [14] suggested combined approaches to recognize the entities from the question answering dataset. This Robust Risk Minimization classifier (RRM), Transformation based Classifier (TBL), Hidden Markov Model (HMM) and Maximum Entropy Classifier (ME) to recognize the named entities. In ME classifier, weights are assigned to the words in the training sentences. The classifier is trained with weights and Viterbi algorithm has used to assign weights to the words. In HMM, the named entity boundaries are computed using statistical models. The selected candidate entities region occurs within the named entity region, then the candidate, entity be selected as an entity. Among four probabilistic models used for named entity extraction, RRM and ME recognize the more entities than the other two methods.

Mccallum et al [15] used CRF graphical model to extract the entities from the news articles. CRF is a finite state machine (FSM) model which is trained with manually annotated features set. A feature in the feature set is a sequence of ordered word blocks or *WebListing* or a lexicon. These ordered word blocks predicts whether the candidate sequence is an entity or not. That is particular (i-1)th word in the candidate sequence is an entity type then and ith word categorized as an entity type.

Nothman et al [16] presented a system to extract the entities from Wikipedia. The implemented system performs various operations, which are related to entities in the Wikipedia. Wikipedia uses "MediaWiki markup" markup language to represent its data in online. The proposed system performs five tasks. First, Wikipedia entities are categorized as conceptional, relational, thematic, administrative, product, fictional characters and facilities. Second, identifies interlink between two articles written in two different language languages. Third, extracts all outgoing links on a particular article, Fourth, naming the links to the outgoing links and Fifth, mapping entities into the target scheme. The implemented system is evaluated the Wikipedia articles which have written in nine different languages.

Song et al [17] proposed PKDE4J approach to extract entities and extract entity relation from bio-medical data sets. For entity extraction, entity dictionary has prepared with a Trie data structure which performs fast entity matching. In data pre-processing, all the abbreviations of entities are modified with original entity names. Penn Treebank is used to tokenize the bio-medical data's. Split the data's into sentences, the Stanford NLP splitting algorithm is used and Stanford POS Tagging is used to assign the tags to the words in the sentences. For normalization of words, the Stanford Lemmatization technique has used. Entities in the sentences are extracted using the soft - TFIDF technique. This soft-TFIDF is a weighted edit distance approach to find the similarities in the matching words. After entity extraction process, the relationships among the entities are identified using syntactical parsing.

Weninger et al [18], proposed Web structure mining approach *Hybrid List Extraction* to extract entity pages from World Wide Web (WWW). This approach performs an extraction of web link lists, extraction of mapping entity pages and attributes of entities. The web pages are connected with each other through web links and two web pages are said to be connected if and only if both pages are under the same web list. A same page can be referred by two links. These two links are called as parallel links or parallel paths. In the web structure mining algorithm, shortest path technique is used to find the parallel links.

Kang et al [19] presented a machine learning algorithm to rank the interrelated entities from the user queries. A user query contains a pair of entity and its facet (features of the entities). The entities are retrieved from the queries and plotted in a semantic graph to group them and assign the ranking to those entities. , Pairwise Comparison Model (PCM) is used to find the preferences between two entities. Boosting algorithm has used to rank the entities based on the result obtained from the PCM.

Tianlei et al [20] recommended a statistical based framework model to link all the scattered information about a particular entity in the biography documents in the Wikipedia. To link all the related entities data's of the internet, two knowledge models, temporal knowledge and relational knowledge are deployed. First, the candidate entities are retrieved by using manually annotated heuristics string matching rules. This rule extracts entities from title page, the alias name for the extracted entity, title page contains only capital letters, redirect page title, title of hyperlink page and etc. One of the popular ontology knowledge base is *freebase*, which is adopted for the extracting entity linking.

Table 1: various techniques and data sets used in the research articles.

SNO	Paper	Proposed method	Purpose	Language	Dataset		Quality Metrics			
1	[1]	Context Free , Context Similarity Free	ERS	E	Census , Cora and Restaurant Datasets		F1-Socre >80% of all the datasets (UCI, Restaurants, Cora, Synthetic)			
2	[2]	Importance Weighted Active Learning algorithm	EE	E	Dataset	Entities	F1-socre-0.93 Precision-85%			
					Business	3958				
					Person	574913				
					DBLP-ACM	494437				
					Scholar-DBLP	589326				
3	[3]	Jaccard Coefficient, Relational Clustering Algorithm	ERS	E	Dataset	Author names	Dataset	F1		
					CiteSeer	1165	CiteSeer	0.995		
					arXiv	9200	arXiv	0.985		
					BioBase	831991	BioBase	0.819		
4	[4]	Decision Tree, POS Tagger	EE	E	100 Reuters news articles		Approach	P	R	F1
							Dictionary	0.862	0.808	83.7

						POS Tagger	0.918	0.476	62.6
						Feature Set	0.980	0.627	76.4
5	[5]	Genetic Algorithm	ER	E	TripAdvisor.com - Hotels reviews, Edmunds.com- Car reviews.	Accuracy-96.6%			
6	[6]	Hidden Markov Model	EE	E	Student names- 16623 from Roster Finder project, http://www.softwarejobs.com – 159 job titles	Dataset	P	R	F1
						Student names	74.8	84.6	79.4
						Job titles	36.0	51.9	42.5
7	[7]	Conditional Random Field	EE	A	COAE 2008 review corpus. Documents- 474 8177 camera reviews, car reviews, notebook reviews and cell phone reviews	Dataset	P	R	
						Cell Phone Data	86.56%	77.23 %	
						Notebook Data	83.16%	73.97 %	
8	[8]	Regular expression, Logistic regression	EE	E & G	4,500 text messages of 11 participants	Approach	F1		
						Regular Expression	76.76		
						Logistic regression	77.73		
9	[9]	Jabari semantic engine	EE	E	30 text documents consists of intelligence reports and tactical messages	P	R		
						85.5%	54.8%.		
10	[10]	Coherent and Efficient Named Entity Linking	EL	E	100 randomly picked Reuters articles from the CoNLL-YAGO dataset 335 Wikipedia articles 50 short text snippets and was introduced in the AIDA	Datasets	P	R	F1
						News	91.1 %	74.0%	81.7%
						Encyclopedic	89.2 %	72.63 %	83.3%
						Micro	90.5 %	40.4%	85.7%
11	[11]	Hyperspace Analogue to Language & Latent semantics.	EE	E,S,D &C	250,000 tokens from coNLL Corpora.	Language	F1		
						English	89.44		
						Spanish	83.08		
						Dutch	83.01		
						Czech	74.08		
12	[12]	Genetic Programming	EE	E	UKPMC database -2 million documents.	P	R		
						>97%	>93%		
13	[13]	Boosting algorithm, Voted Perceptron	EE	E	the full 53,609 sentences of training data, and decoded the 14,717 Sentences of test data.	Approaches	P	R	F
						Maximum Entropy Model	0.84	0.86	0.85
						Boosting Model	0.87	0.87	0.87
						Voted Perceptron	0.87	0.88	0.87

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>14</td><td>[14]</td><td>Robust Linear Classifier, Hidden Markov Model, Maximum Entropy , Transformation Based Learning</td><td>EE</td><td>E, G</td><td>IBM question answering system</td><td><table><tr><td rowspan="2">Approaches</td><td colspan="2">E</td><td colspan="2">G</td></tr><tr><td>P</td><td>R</td><td>P</td><td>R</td></tr><tr><td>HMM</td><td>0.82</td><td>0.74</td><td>-</td><td>-</td></tr><tr><td>TBL</td><td>0.88</td><td>0.81</td><td>0.69</td><td>0.68</td></tr><tr><td>ME</td><td>0.90</td><td>0.85</td><td>0.68</td><td>0.67</td></tr><tr><td>RRM</td><td>0.92</td><td>0.85</td><td>0.70</td><td>0.71</td></tr></table></td></tr>
<tr><td>15</td><td>[15]</td><td>Conditional Random Fields</td><td>EE</td><td>E, G</td><td>news articles
with tagged entities</td><td><table><tr><td>Language</td><td>P</td><td>R</td><td>F1</td></tr><tr><td>E</td><td>89.84</td><td>88.10</td><td>88.96</td></tr><tr><td>G</td><td>75.97</td><td>61.72</td><td>68.11</td></tr></table></td></tr>
<tr><td>16</td><td>[16]</td><td>semi-supervised machine learning</td><td>EE</td><td>E,G,S, F,I,D, P, PO,R</td><td>2300 Wikipedia articles</td><td>F1=85%</td></tr>
<tr><td>17</td><td>[17]</td><td>PKDE4J- Entity dictionary based approach</td><td>EE &EL</td><td></td><td>CRAFT - 67 full-text articles of 21,000 sentences,
GENETAG-2000 Sentences ,
AnEM -500 Biomedical documents.</td><td>F1-Score-83.7%</td></tr>
<tr><td>18</td><td>[18]</td><td>HyLiEn link algorithm</td><td>EE</td><td>E</td><td>106 universities
1,206,445-Web pages</td><td><table><tr><td>P</td><td>R</td><td>F</td></tr><tr><td>99.7%</td><td>99.9%</td><td>99.4%</td></tr></table></td></tr>
<tr><td>19</td><td>[19]</td><td>Web-scale entity ranking algorithm</td><td>ER</td><td>E</td><td>6000 query entities ,

33000 entities–facet pairs</td><td><table><tr><td>P</td><td>R</td></tr><tr><td>0.89</td><td>0.99</td></tr></table></td></tr>
<tr><td>20</td><td>[20]</td><td>KeEL- Statistical framework model</td><td>EL</td><td>E</td><td>1000 biography pages in
Wikipedia</td><td><table><tr><td>P</td><td>R</td><td>F1</td></tr><tr><td>0.916</td><td>0.891</td><td>0.903</td></tr></table></td></tr>
</table>

E- English, A-Arabian, G- German, S- Spanish, D-Dutch C-Czech, I-Italian, D-Dutch, P-Polish, PO-Portuguese, R-Russian

ERS-Entity Resolution, ER-Entity Ranking, EL-Entity Linking, EE-Entity Extraction

P- Precision, R- Recall, F1-F-Score

Table 2: Techniques used in research papers

SNO	Techniques Used	The papers referred
1	Probabilistic Models	[1] ,[5],[6]

2	IWAL active algorithm	[2]
3	Clustering Algorithm	[3]
4	Decision Tree	[4]
5	Genetic Programming	[5], [12]
6	Hidden Markov Model	[6],[14]
7	Conditional Random Field	[7],[15]
8	POS Tagger	[4],[7],[8],[17]
9	Logic regression	[8],[12]
10	Regular expression	[8],[9],[12]
11	Latent semantics.	[11]
12	Bootstrapping algorithm	[13],[16]
13	Maximum Entropy Model	[13],[14]

3. Conclusion

The Entity Recognition field has been flourishing for many years. It is one of the information retrieval techniques and popular in sentiment analysis. This paper has given an outline of the strategies utilized to extract entities, link entities, map entities, resolve entity names conflict, entity attribute extraction and etc. Most of the research papers are used machine learning techniques to recognize the entities and few used statistical models. In recent years, this entity extraction approaches are not using any annotated corpus and lexicon to detect the entities. Thus, the approaches are designed as fully automated for entity extractions.

Since, the dataset volumes are too large and trend is big data, preparing lexicons for such a huge volume of data’s will consume more time.

References

1. Ayat, N., Akbarinia, R., Afsarmanesh, H., & Valduriez, P. (2014). Entity resolution for probabilistic data. Information Sciences, 277, 492-511. doi:10.1016/j.ins.2014.02.135
2. Bellare, K., Iyengar, S., Parameswaran, A., & Rastogi, V. (2013). Active Sampling for Entity Matching with Guarantees. ACM Trans. Knowl. Discov. Data ACM Transactions on Knowledge Discovery from Data TKDD, 7(3), 1-24. doi:10.1145/2513092.2500490.

3. Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. ACM Trans. Knowl. Discov. Data ACM Transactions on Knowledge Discovery from Data TKDD, 1(1). doi:10.1145/1217299.1217304

4. Baluja, S., Mittal, V. O., & Sukthankar, R. (2000). Applying Machine Learning for High-Performance Named-Entity Extraction. Computational Intell Computational Intelligence, 16(4), 586-595.

5. Bashir, S., Afzal, W., & Baig, A. R. (2016). Opinion-Based Entity Ranking using learning to rank. Applied Soft Computing, 38, 151-163. doi:10.1016/j.asoc.2015.10.001

6. Brill, E., & Pop, M. (1999). Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging. Text, Speech and Language Technology Natural Language Processing Using Very Large Corpora, 27-42. doi:10.1007/978-94-017-2390-9_3

7. Ding, S., & Jiang, T. (2010). Comment Target Extraction Based on Conditional Random Field & Domain Ontology. 2010 International Conference on Asian Language Processing. doi:10.1109/ialp.2010.81

8. Eka Tobias, Kirkegaarda Camilla, Jonssonb Håkan, Nuguesa Pierre(2011) , Named entity extraction for short text messages, Procedia - Social and Behavioral Sciences, Computational Linguistics and Related Fields,Volume 27, 2011, Pages 178–187

9. Guo, J. K., Brackle, D. V., Lofaso, N., & Hofmann, M. O. (2015). Extracting Meaningful Entities from Human-generated Tactical Reports. Procedia Computer Science, 61, 72-79. doi:10.1016/j.procs.2015.09.153

10. Gruetze, T., Kasneci, G., Zuo, Z., & Naumann, F. (2016). CohEEL: Coherent and efficient named entity linking through random walks. Web Semantics: Science, Services and Agents on the World Wide Web. doi:10.1016/j.websem.2016.03.001

11. Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent semantics in named entity recognition. Expert Systems with Applications, 42(7), 3470-3479.

12. Korkontzelos, I., Piliouras, D., Dowsey, A. W., & Ananiadou, S. (2015). Boosting drug named entity recognition using an aggregate classifier. Artificial intelligence in medicine, 65(2), 145-153.

13. Michael Collins. (2002). Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron. In Proceedings of the 40th Meeting of the ACL, pages 489–496, Philadelphia, PA

14. Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named entity recognition through classifier combination. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 168-171). Association for Computational Linguistics.

15. Mccallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 -. doi:10.3115/1119176.1119206
16. Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence, 194, 151-175
17. Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of Biomedical Informatics, 57, 320-332. doi:10.1016/j.jbi.2015.08.008
18. Weninger, T., Johnston, T. J., & Han, J. (2013). The parallel path framework for entity discovery on the web. ACM Transactions on the Web (TWEB), 7(3), 16.
19. Kang, C., Yin, D., Zhang, R., Torzec, N., He, J., & Chang, Y. (2015). Learning to rank related entities in Web search. Neurocomputing, 166, 309-318.
20. Tianlei, Z., Xinyu, Z., & Mu, G. (2015). KeEL: knowledge enhanced entity linking in automatic biography construction. The Journal of China Universities of Posts and Telecommunications, 22(1), 57-71.
21. Hsu, Y. Y., & Kao, H. Y. (2015). Curatable named-entity recognition using semantic relations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(4), 785-792.