Journal of Education and Science (ISSN 1812-125X), Vol: 29, No: 1, 2020 (242-252)

A Survey Study on Relation Extraction for Web Pages

Ghada A.K. Alsaigh[1*]; Ghayda A.A. Al-Talib[2]; Alaa Y. Taqa[3]

[1] Central Library, University of Mosul, Mosul, Iraq
[2] Department of Computer Science, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq
[3] Department of Computer Science, College of Education For Pure Science, University of Mosul, Mosul, Iraq

Email: [1*] ghalsaigh@gmail.com, [2] ghaydatalib@yahoo.com, [3]alaa.taqa@gmail.com

(Received January 14, 2019; Accepted April 10, 2019; Available online March 01, 2020)

DOI: 10.33899/edusj.2020.164377,

Abstract:

Natural language means a language that is used for communication by human. Natural Language Processing (NLP) helps machines to understand the natural language. The natural language for the web pages consists of many semantic relations between entities. Discovering significant types of relations from the web is challenging because of its open nature.

In this paper we survey several important types of semantic relations. This paper also covers the relation extraction (RE) approaches which are divided into: supervised approach, which contains Feature base and Kernel base, and the unsupervised approach. Three relation extraction algorithms are discussed: Support Vector Machine (SVM), Genetic algorithm and Naive Bayes classifier

This survey would be useful for three kinds of readers First the Newcomers in the field who want to quickly learn about relation extraction. Second the researchers who want to know how the various relation extraction techniques developed over time. Third the trainers who just need to know which RE technique works best in different settings

Keywords: relation extraction, web pages, NLP

دراسة مسحية لاستخراج العلاقة من صفحات الويب

غادة عبد الكريم الصائغ[1*] و غيداء عبد العزيز الطالب[2] و الاء ياسين طاقة[3]

[1*] المكتبة المركزية, جامعة الموصل, الموصل ، العراق
[2] قسم علوم الحاسوب, كلية علوم الحاسوب والرياضيات, جامعة الموصل، الموصل ، العراق
[3] قسم علوم الحاسوب, كلية التربية للعلوم الصرفة, جامعة الموصل ، الموصل، العراق

الملخص

اللغة الطبيعية تعني اللغة التي يستخدمها الإنسان للتواصل. تساعد معالجة اللغات الطبيعية (NLP) الآلات على فهم اللغة الطبيعية. تتكون اللغة الطبيعية لصفحات الويب من العديد من العلاقات الدلالية بين الكيانات. يعد اكتشاف أنواع مهمة من العلاقات من الويب تحديا صعبًا بسبب طبيعة الويب المفتوحة.

في هذا البحث ، تم مسح عدة أنواع مهمة من العلاقات الدلالية كما يتناول البحث أيضًا اساليب استخراج العلاقة (RE) التي تنقسم إلى: اسلوب خاضع للإشراف ، والذي يحتوي على قاعدة الميزات وقاعدة البذرة ، والاسلوب غير الخاضع للإشراف والذي تم فيه مناقشة ثلاث خوارزميات لاستخراج العلاقة: دعم ناقل الماكينة (SVM) ، الخوارزمية الجينية ومصنف Naive Bayes

يعد هذا البحث نافعًا لثلاثة أنواع من القراء أولاً الوافدين الجدد في هذا المجال الذين يريدون أن يتعلموا بسرعة موضوع استخراج العلاقة. ثانياً ، الباحثون الذين يريدون أن يعرفوا كيف تطورت أساليب استخراج العلاقة المختلفة مع مرور الوقت. ثالثًا ، المدربين الذين يحتاجون فقط إلى معرفة تقنية استخراج العلاقة التي تعمل بشكل أفضل في بيئات مختلفة

الكلمات المفتاحية: استخراح العلاقة ، صفحات الويب ، معالجة اللغة الطبيعية NLP

1-Introduction:

Through the World Wide Web increasing information and texts, knowledge are available and found in the digital archives, it has seen that web content has been kept in HTML "Hyper Text Markup Language"[1]. In this case the web is for human use because of the displaying content as syntax based HTML. Query ambiguity reduces HTML retrieval quality. For example “bank” may be border of a water body or monetary establishment. Web pages have more information, as HTML tags, hyperlinks and anchor text with the regular text content visible in a browser. These characteristics that are placed on pages are useful for classification [2]. There has been an increasing demand in "Information Extraction" (IE), which recognizes relevant information (usually of predefined types) from text documents in a specific subject and it gathers it in a structured format [3]**.** One of the purposes of relation extraction is to specify the named entities, and to extract the relationship between entities and the events [4]**.**

Relation extraction is defined as the process of discovering and describing the "semantic relations" between entities of text [5]**.** Most algorithms of relation extraction begin with some linguistic analysis, parsing the text to find relations directly from the sentences. [6].

The relation extraction system in (Figure 1), which is inspirited by [7], enters as input the text in a document, and produces a list of (entity, relation, entity) as its output.

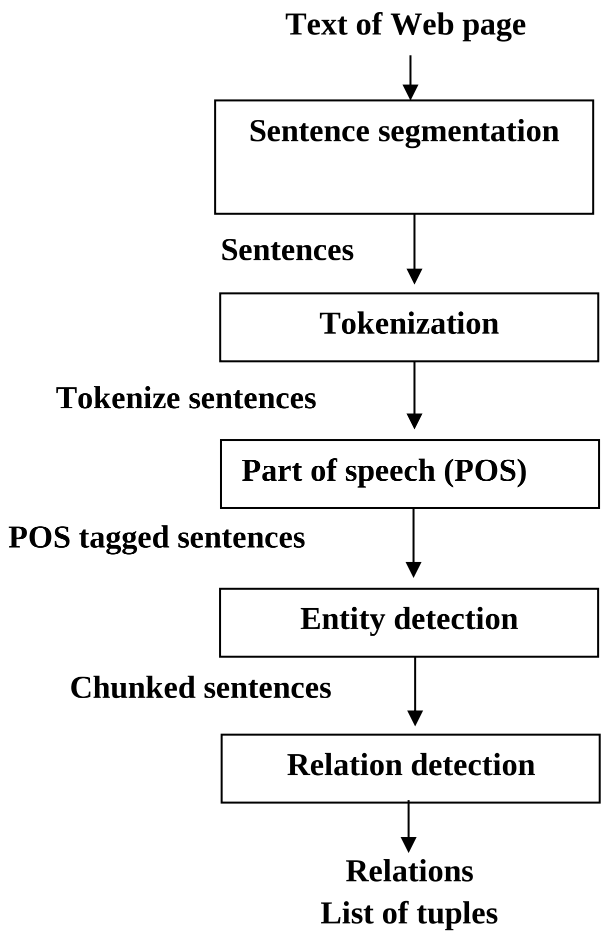

Figure 1: Simple Pipeline Architecture for an Information Extraction System.

2-Data Source:

This research do a review about the web documents which derive its information from several sources such as: Wikipedia, ACE RDC 2003 and 2004, Social Networks (Twitter & Facebook), Clueweb09 dataset, MEDLINE, PharmGKB database and PubMed. Web document can be:

2.1 XML document "eXtensible Markup Language" is a typical format, it is used to share and transfer information in different fields, because it can transfer the content of logical structures into documents, and it is autonomous from platform [8].

2.2 HTML document Hypertext Markup Language (HTML) is the standard markup language it aims at producing web pages and web applications [9]. A document may contain many links, a technical text or a short answer to a special question [10].

3-Text relation

It is the relations between the words in the sentence. This relation can be a relation of syntax, lexical and semantic relation. Syntax relation describes how words are grouped and connected to each other in a sentence [11]. While A lexical relation is a pattern of association that exists between lexical units in a language [12].

3.1-Semantic Relations

The primary aim of recent researches is to extract relevant documents. Web development to the next generation called the "Semantic Web" [13], the attention will move from looking for documents to getting facts, useful information [12]. The increasing capability of finding the information in the form of entities, contained within documents, leads to the important results in extracting relations between these entities. [14] Relationships are fundamental to semantics because they join the meanings to the words, terms and entities [15]. The description of word semantic relationships is shown in the following:

● **Synonyms**

Synonyms relation means a word with the same or nearly the same meaning as another in the same language [16], as shown in (Figure 2)[17]:

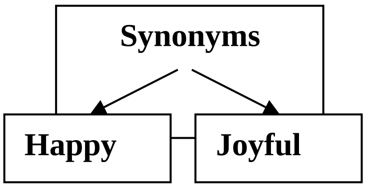

Figure 2: The Synonyms Relation

● **Antonyms**: are words that have contrasting and opposite in meaning to another as shown in (Figure 3)

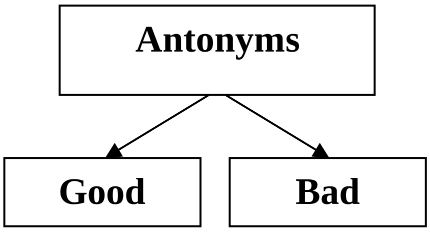

Figure 3: The Antonyms relation of two different words

or they could be opposite by adding the following prefixes to form opposites of words: un-, il-, im-, in-, ir-as shown in table 1 [6].

Table (1): Opposite by adding a prefix

Word	Opposite
Happy	Unhappy
Legal	Illegal
Polite	Impolite
Compatible	Incompatible
Regular	Irregular
Normal	Abnormal

● **Metonyms**: are words used in place of another word which has strong relation. as shown in (Figure 4):

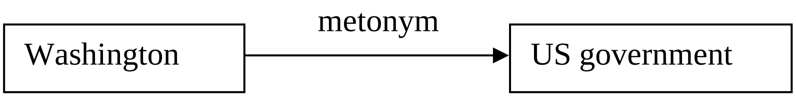

Figure 4: The Metonyms Relation

● **Hyponym and Hypernymy**: The term hyponym means a subcategory of a more general class: Like a relationship between "dog" and "animal". While Hypernymy is the state or quality of being a hypernym or superordinate (a general class under which a set of subcategories is subsumed). as shown in (Figure 5) [17]**.**

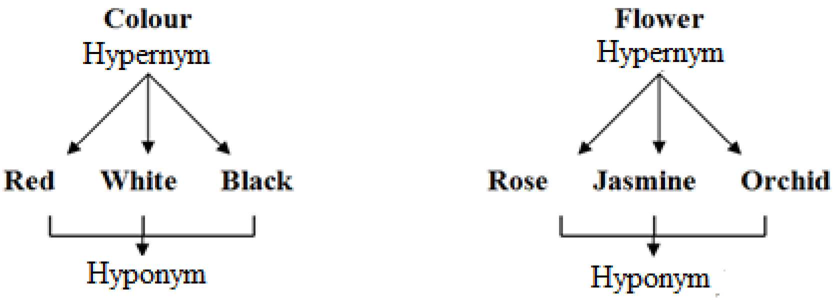

Figure 5: The Hyponym & Hypernymy Relations

● **Polysemy** It means a word, phrase, or concept which has more than one meaning or connotation, as shown in (Figure 6) [18]

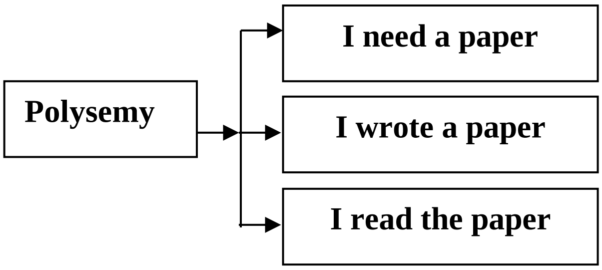

Figure 6: The Polysemy Relation

In this example "paper" in the first sentence refers to a piece of paper, in the second sentence it means a research paper and in the third one it denotes to a newspaper

● **Homonyms** Words that are similar in forms or sounds, but they are different in meanings and origins as shown in (Figure 7) [16].

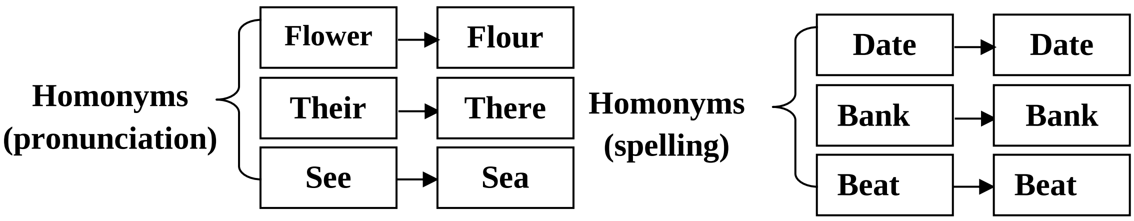

Figure 7: The Homonyms Relation

4-Relation Extraction (RE)

The aim of relation extraction is to discover semantic relations between entities [19]. This means confront in open-domain of the web. This relation must be able to deal with a very, huge and rapid growth in scale, multiple styles of documents and more types of relations that are exist [20]. To find these relations, a system should not expect a specific set of relation types, nor rely on a rigid set of relation argument types. It also must efficiently capable to deal with a huge size of data [21]. A huge size of hand labeled data is needed when the supervised learning algorithms are used but annotating training data is undesirable and time overwhelming job [22]. On the Web, manually labeling data of each subject area are stubbornly, the number of subjects of interest is simply very large. Relation extraction with automated labeling is called "unsupervised relation extraction". [23].

4.1- Supervised Relation Extraction Approach

Supervised approaches concentrate on relation extraction at particular area. These approaches need labeled data where each pair of entity that are mentioned, labeled with one of the pre-defined relation types. [24].

4.1.1 Feature Based Approach

The feature-based methods are used to find useful lexical feature, syntactic structured feature and so on. As shown in Table 2

Table 2: Feature based method

Title	Author(s)	year	Application	Features
"A distributed meta-learning system for Chinese entity relation extraction"	"Lishuang Li, Jing Zhang, Liuke Jin, Rui Guo, Degen Huang"	2015	Chinese languages	distributed meta learning system (lexical)
"Extracting logical structures from HTML tables	Yeon-Seok Kim, Kyong-Ho Lee	2008	HTML Tables	semantic coherency
Exploiting aspectual features and connecting words for summarization-inspired temporal-relation extraction	Bonnie J. Dorr, Terry Gaasterland	2007	NLP	Tense of the sentences (syntax)

The cost in Lishuang Li e.al. [25] predication phase when combine the feature and kernel based calculation is lower than other but the computational cost in the training phase is bigger compared to the other.

The feature based approach is an excellent method for extracting the logical structures of HTML tables and moving them into XML documents Yeon-Seok & Yeon-Seok [8] using area segmentation and structure analysis algorithm, as well as semantic coherency feature. While Bonnie.& Gaasterland [26] use feature based approach to identify tense of the sentences at Penn Treebank tags for parse tree. The work extracts, reanalysis, and reinterpretation of both temporal and non temporal relations between two events.

4.1.2 Kernel based approach

Kernels based approach compares the structure of two patterns using the syntax tree from the node at the top "root" to the lowest node "child". This approach still has restrictions in measuring patterns of multiple types, which decrease the act of new relation extraction. The main advantage of kernel based methods is that such explicit feature engineering is avoided [27] as shown in Table 3

Table-3: kernel based method

Title	Author(s)	Year	Application	Features
"Construction of semantic bootstrapping models for relation extraction"	"Zhang Chunyun, Weiran Xu, Zhanyu Ma, Sheng Gao, Qun Li, Jun Guo"	2015	Text Analysis Conference	POS
"Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel"	"Maengsik Choi, Harksoo Kim"	2013	Social Network	Name entity
"Tree kernel-based semantic relation extraction with rich syntactic & semantic information"	"Zhou Guodong, Qian Longhua, Fan Jianxi"	2010	Newspapers, newswires, and broadcasts.	Semantic relation

The framework of Zhang et.al.[28] exploit "trigger words" as the semantic restrict to lead the "bootstrapping iterations". It widen a work on usual model of bootstrapping in extraction of the relation by construct a noble way for explaining trigger words, pattern representation, similarity method and evaluation method. Furthermore, a noble "bottom up kernel" algorithm was defined to calculate if the result's pattern from a new sentence is relation form or not. Maengsik & Harksoo [29] use SVM algorithm on social network application to identify name entity by using kernel based approach on social network. Zhou et.al. [3] combine different types of syntactic and semantic information into one tree structure; and they also extract such varieties via nobel context-sensitive convolution tree kernel.

4.2- Unsupervised Relation Extraction Approach

It refers to the task of automatically finding interesting relations between entities in large text corpora Yulan [30], as shown in Table 4

Ya-nan et.al. [4] used a proposed "statistical score S" to calculate the familiar association between strong related events and clip relations with low S value. Ying. et.al. [31] investigated Social Network using unsupervised feature based to extract name entity feature by disambiguation system. The main advantage is the collection of the unsupervised features extracted from broad resources that can effectively improve the robustness of a disambiguation system.

Bonan et.al. [21] used an algorithm handles polysemy of relation instances on Clueweb09 dataset and achieves a significant improvement in recall while maintaining the same level of precision.

Yulan et.al. [30] worked on Wikipedia, their work can abstract away from different surface realizations of text. These relations expressed in different "dependency structures" with redundant information from the growing size of Web pages.

Table 4: Unsupervised Approach

Title	Author(s)	Year	application	Feature
"Mining Large-scale Event Knowledge from Web Text"	"Ya-nan Cao, Peng Zhang, Jing Guo, Li Guo"	2014	NLP	lexico-syntactic and lexico semantic
A robust web personal name information extraction system	Ying Chen, Sophia Yat Mei Lee, Chu-Ren Huang	2012	Social Network	Name entity
Towards Large-Scale Unsupervised Relation Extraction from the Web	Bonan Min, Shuming Shi, Ralph Grishman, Chin-Yew Lin	2010	Clue web09	POS
"Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web"	"Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang and Mitsuru Ishizuka"	2009	Wikipedia	Surface pattern

5- Relation Extraction Algorithms

Throughout this section three algorithms (Support Vector Machines, Genetic algorithm and Naive Bayes classifier) have been discussed in relation extraction.

5-1 Support Vector Machines (SVM)

Support Vector machine is "Vector space based machine -learning method" used to extract a decision limits between two classes. These classes are a long way from any point in the training data. separately from executing linear classification, SVMs are able to run a non-linear classification in efficient manner using what is called the "kernel trick", implied mapping their inputs into high-dimensional feature spaces. [32]. Table 5 illustrates the different use of SVM in relation extraction.

Table 5: Support Vector Machine SVM in relation extraction

Title	Author(s)	Year	Application	Algorithm
"A distributed meta-learning system for Chinese entity relation extraction"	"Lishuang Li, Jing Zhang, Liuke Jin, Rui Guo, Degen Huang"	2015	Chinese languages	SVM
"Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel"	"Maengsik Choi, Harksoo Kim"	2013	Social Network	SVM
"Compensating for Annotation Errors in Training a Relation Extractor"	"Bonan Min, Ralph Grishman"	2012	different web article from ACE2005	SVM, Baseline algorithm & purify
"Tree kernel-based semantic relation extraction with rich syntactic and semantic information"	"Zhou Guodong, Qian Longhua, Fan Jianxi"	2010	Newspapers, newswires, and broadcasts.	SVM

Bonan & Ralph [19] found that "one-pass annotation" is a powerful in cost than annotation with effective assurance. While Zhou et.al [33] found that correctly unifying multi type of syntactic and semantic information into a one tree structure; and clipping such differences via a good context-sensitive convolution tree kernel.

5-2 Genetic Algorithm (GA)

Christy & Thambidurai[34] show that Genetic Algorithm well performed in mining rules and features optimization of a text.

Ines et.al.[35] deploy genetic algorithm and get a high precision but low recall and they combine the benefits of ML algorithms with "rule-based" techniques to find the related arabic named entities. The effect of each algorithm used linguistic module to create important results against previous one but the method unable to capture some of the relations that exist between words that are far from the named entity locations, especially in sentences which are long and complex. Table 6 illustrates the GA in relation algorithm

Table 6: Using Genetic Algorithm

Title	Author(s)	Year	Application	Algorithm
"A hybrid method for extracting relations between Arabic named entities"	"Ines Boujelben, Salma Jamoussi, Abdelmajid Ben Hamadou"	2006	Arabic Named entity	Genetic Algorithm
"Efficient Information Extraction Using Machine Learning and Classification Using Genetic and C4.8 Algorithms"	"Christy , A. & Thambidurai, P."	2006	Text	Genetic Algorithm

5-3 Naive Bayes classifier

Naive Bayes classifier is a method which learns both annotated and not annotated documents in a "semi-supervised algorithm". Suresh & Kumar, [36] applied the Naive Bayes classifier on Q/A systems using "lexico-syntactic and lexico semantic feature". They reach the high precision and recall (the ideal case).

6- Evaluation Metrics

A common motivated way of evaluating results of Machine Learning experiments is using Recall, Precision and F1-measure [37]. Precision measures as shown in equation (1) is the percentage of the correct retrieved items on the number of the whole retrieved items [38]. The good system produces a high precision in retrieving correct items [39].

$$Precision = P = \frac{No.of\ relevant\ retrieved\ items}{No.of\ retrieved\ items} \quad (1)$$

Recall, on the other hand, is a percentage of the total number of the correct items as computed in equation (2). The higher the Recall rate, indicates less missing correct items [40]

$$Recall = R = \frac{No.of\ relevant\ retrieved\ items}{No.of\ relevent\ items} \quad (2)$$

Finally F1 measure: is the average of the precision and recall. The F-measure measure is prompt because in many studies this measure is the best measurement of the result of the classifier [40]. Equation (3) depends on Precision and Recall

$$F1\ measure = \frac{2PR}{P+R} \quad (3)$$

Table 7 illustrates the evaluation metrics for different algorithms that have been used in relation extraction to extract a specified feature for a given application

Table 7: Evaluation Results

Title	Author(s)	Approach	Identification feature	Algorithm	precision	Recall	F1
Concept relation extraction using Naïve...	Suresh & Zayaraz (2015)	rule base approach	lexico-syntactic and lexico semantic	Naive Bayes classifier	96	99%	97.4
A hybrid method for extracting relations ..	Ines et.al. (2014)	rule base approach	Name Entity	Genetic Algorithm	84.8	67.6	75.22
Mining Large-scale Event ..	Ya-nan et.al (2014).	pattern based	lexico-syntactic and lexico semantic	Statistical Score S	89	83%	85.9
Tree kernel-based semantic…	Zhou et.al. (2010)	tree kernel based	semantic relation	SVM	83.1	73.5	77.8

Conclusion

This survey paper discussed importance of relation extraction techniques in natural language processing field. Also it discussed different approaches which are widely used for relation extraction task then it discussed the evaluation criteria metrics. It is obvious that the naïve bayes classifer, using "lexico-syntactic and lexico semantic features", gives the best evaluation measures near the ideal case. On the other hand, it is very important to reduce the time to extract web relations accurately without loosing efficiency.

The use of pattern based with local dependency tree increases the accuracy and recall of event-arguments extraction process.

Supervised approaches for the more can do well when the domain is more restricted. While the unsupervised approaches appear to be more appropriate for unrestricted domain relation extraction systems, due to they are capable of simply grew with the database size and can scale to new relations easily.

Rule sets have a benefit of sentence structure and grammar to capture more specific information. Moreover, these rule sets can be sets in an ontology that allows modification of relationships and inference over them.[41]

This work suggests that future work in this area could apply fuzzy logic which is a principal component of soft computing.

References:

1- Eichler K., Hemsen H. & Neumann G., Proceedings of the 6th edition of the language resources and evaluation conference:1674-1679 (2008).
2- Leela Devi B. & Sankar A., International Journal of Computer Applications, 69(2):41-46 (2013).
3- Zhou GD., Zhang M., Ji DH. & Zhu QM., Information Processing and Management, 44:1008–1021 (2008).
4- Ya-nan C., Peng Z., Jing G. & Li G. Procedia Computer Science, 29:478-487 (2014).
5- Jing J., Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2-7 August:1012-1020 (2009)
6- Veda C. S., VLDB Journal, 2:455-488 (1993).
7- Steven B., Ewan K. & Edward L., Natural Language Processing with Python. O'Reilly Media, Inc. Sebastopol, California, USA. (2014).
8- Yeon-Seok K. & Kyong-Ho L. , Computer Standards & Interfaces, 30.:296–308 (2008).
9- Brooks, DR., An Introduction to HTML and JavaScript for Scientists and Engineers, Springer-Verlag London Limited . (2007)
10- Eissen, SM & Stein, B., Annual Conference on Artificial Intelligence, S. Biundo, T. Fruhwirth, and G. Palm (Eds.): KI, LNAI 3238:256–269, Springer-Verlag Berlin Heidelberg (2004)
11- Vladimir L., arXiv 0802.4181v1 (2008).
12- Mridha M.F. , Aloke K. S. & Jugal K. D., International Journal of Advanced Computer Science and Applications , 4(1): 17-21 (2014)
13- Zheng X., Xiangfeng L., Shunxiang Z., Xiao W., Lin M. & Chuanping Hu. , Future Generation Computer Systems, 37:468–477 (2014).
14- Sujatha, T., Ramesh N G, Suresh B., International Journal of Soft Computing and Engineering, 2(3): 213-218 (2012).
15- Sheth A., Arpinar I.B., Kashyap V. In: Nikravesh M., Azvine B., Yager R., Zadeh L.A. (eds) Enhancing the Power of the Internet. Studies in Fuzziness and Soft Computing, vol 139. Springer, Berlin, Heidelberg (2004)
16- The Oxford English Dictionary. Oxford University Press (2017).
17- Zapata, AA., Inges, vol(4), 7p. (2008)
18- Mojela, V.M., Lexikos (17):433-439 (2007)
19- Bonan M., Shuming S., Ralph G. & Chin-Yew L., Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural language learning, 1027-1037 (2012)
20- Bryan Rink, Sanda Harabagiu, Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc.,18:594-600 (2011)
21- Bonan Min , Shuming Shi, Ralph Grishman & Chin-Yew Lin (2010). Towards Large-Scale Unsupervised Relation Extraction from the Web. Int. J. on Semantic Web & Information Systems, 8(3):1-23 (2012)
22- Haibo Li, Yutaka Matsuo, and Mitsuru Ishizuka. Semantic Relation Extraction Based on Semi-supervised Learning . Cheng P.-J.et al. (Eds.): AIRS, LNCS 6458, Springer-Verlag Berlin Heidelberg, 270–279.(2010)

23- Doug Downey, OrenEtzioni, Stephen Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction. Artificial Intelligence, 174:726-748 (2010)
24- Sachin P., Girish K. P. & Pushpak B., arXiv:1712.05191 (2017)
25- Lishuang L., Jing Z., Liuke J., Rui G. & Degen H., Neurocomputing, 149:1135-1142 (2015).
26- Bonnie J. Dorr & Terry G. , Information Processing and Management, 43:1681-1704 (2007)
27- Dmitry Z., Chinatsu A. & Anthony R., Journal of Machine Learning Research, 3:1083-1106 (2003)
28- Zhang, C., Weiran X., Zhanyu M., Sheng G., Qun L. & Jun G., Knowledge-Based Systems 83.:128–137 (2015).
29- Maengsik C. & Harksoo K., Information Processing and Management, 49:303-311 (2013).
30- Yulan Y., Naoaki O., Yutaka M., Zhenglu Y. & Mitsuru I., Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, (2009)
31- Ying C., Lee S. & Chu-Ren H. ,Expert Systems with Applications, 39:2690-2699 (2012).
32- Govindarajan, M & Romina, M., The International Journal Of Engineering And Science (IJES), Volume 2(12):11-15 (2013)
33- Zhou G., Qian L. & Fan J., Information Sciences 180 :1313–1325 (2010).
34- Christy , A. & Thambidurai, P., Information Technology Journal, 5 (6): 1023-1027, (2006)
35- Boujelben I., Jamoussi S. & Hamadou A., Journal of King Saud University –Computer and Information Sciences, 26:425-440 (2014).
36- Suresh G. & Zayaraz, G., Journal of King Saud University – Computer and Information Sciences, 27:13-24 (2015).
37- Powers, D. & Ailab. , J. Mach. Learn. Technol. 2: 2229-3981 (2011).
38- Christopher D. M., Prabhakar R. & Hinrich S., Introduction to Information Retrieval, Cambridge University Press..(2008)
39- Soderland, S. Machine Learning :34: 233 (1999).
40- Morgan K. Proceeding of DARPA broadcast News Workshop (1999).
41- Adrien C., Nigam H. Sh., Yael G., Mark M. & Russ B. A. Journal of Biomedical Informatics, 43:1009-1019 (2010)