EasyChair Preprint

№ 4424

Analysis on Semantic level Information Retrieval and Query Processing

S K Liji and Muhamed P Ilyas

October 19, 2020

Noname manuscript No.
(will be inserted by the editor)

Analysis on Semantic level Information Retrieval and Query Processing

Liji S K · Muhamed Ilyas P

Received: date / Accepted: date

Abstract Query processing and Information Retrieval plays important application of Natural Language Processing (NLP) and Data Mining. It aims to retrieve relevant documents for natural language queries. Nowadays large amounts of unstructured data are scattered across the web. So Information Retrieval from these large volumes of unstructured data using natural languages become more crucial and challenging task. The relevant Information Retrieval from such a large amount of unstructured data needs knowledge about the semantic information or contextual information. The semantic information retrieval from unstructured data uses the methods from Data Analytics, Natural Language Processing and Machine Learning etc. Here we propose a survey on different models for Information Retrieval, Information Retrieval using Natural Languages and emphasis on semantic level Information Retrieval. And also perform the comparison and analysis of various models.

Keywords Natural Language Processing · Information Retrieval · Query Processing · Machine Learning · Deep Learning · Neural Networks · Ontology · Word Embedding · Document Embedding

1 Introduction

Human-computer interaction is a mission of mankind since the development of modern computers. The interaction between computers and humans using natural languages are possible with NLP and data mining techniques. NLP researchers aim to gather knowledge about how human beings understand

Liji S K
Sullamussalam Science College,
Tel.: +919645252257
E-mail: liji.s.k@gmail.com

Muhamed Ilyas P
Sullamusslam Science College

and use native languages. The applications of Natural Language Processing includes natural language text processing and summarization, machine translation, Information Retrieval (IR), Query Processing, and Automatic Speech Recognition (ASR) etc. Information Retrieval and Query processing are the two mechanisms available now for information access. Nowadays a huge amount of unstructured data are scattered across the web and it is growing at an exponential rate also very large numbers of people engaged in information retrieval simultaneously, as they use web search engines. Information Retrieval identified as a powerful form of information access than traditional database management systems.

Different approaches to Information Retrieval are NLP based approach, statistical approach and pattern matching approach etc. NLP based Information Retrieval is the most reliable method for human-computer interaction. Using NLP techniques the native languages like Malayalam, Kannada, Tamil will be analysed and processed. The earlier IR systems such as LUNAR and BASEBALL[1] were based on NLP techniques. These systems processed the questions by using NLP techniques, then convert them into standard database query and retrieve the results. Most of the Information Retrieval systems in early stages are keyword similarities based or using statistical techniques. Sometimes irrelevant information is retrieved and not use any query reformation techniques. Information retrieval systems can also be classified by the domain at which they operate. There are a few research works are done in the field of Malayalam Information Retrieval[2–4]as now, most of the Malayalam Information Retrieval systems, exist today are keyword based. No effective semantic level work exist in Malayalam Information Retrieval and Query Processing, because Malayalam is an agglutinative and morphologically rich language. Due to the complexity, development of an Information Retrieval System for Malayalam is a tedious and time-consuming task.

This paper organized as follows. An Introduction about Information Retrieval and Query Processing is in section 1, section 2 discusses various methodologies and architecture of Information Retrieval models, section 3 contains the comparative study and analysis about different models and the last section describes the conclusion and direction for future research work.

2 Architecture and Methodologies of Related Works

The literature reviews related to this survey mainly focuses on the following perspective, Semantic level Information Retrieval and Query Processing, Information Retrieval and Query processing in Natural languages and different approaches to Semantic level Information Retrieval. First, we taken all the research papers exists in this field since 2015, it contained about more than 70 papers. After the content filtering reviewed about 22 papers till 2020. The detailed architecture and methodologies of different works are described as

follows.
In a work Nadia Soudani, Ibrahim Bounhas, Yahya Slimani[5] described an Arabic semantic IR, using a text mining approach. They proposed a generic semantic search approach on Semantic Spaces. They make a comparative experimental study of NLP tools for Arabic and use of linguistic resources thereby the effect of them on the semantic search performance and the importance of the linguistic choices in alienating semantic search engines results. A module of QR is integrated with the System based on a knowledge-based approach for Arabic Semantic Disambiguation by use of a dictionary. The process of Word Sence Disambiguation (WSD) is done based on a Sense Recognition algorithm. Different Semantic Information Retrieval approaches are experimented relying on Semantic Spaces. Tests were made with the use of different Morphological Analyzers and different linguistic resources. The Mean Average Precision for the system varies from 0.97 to 7.52.

In another work Shengxian Wan, Yanyan Lan et al[6] proposed a new deep neural network architecture for semantic matching with multiple positional sentence representations named MV-LSTM. They use a bidirectional long short term memory Bi-LSTM. Then model the interaction between the representations, using three operations-Cosine, bi-linear and tensor layer. Then use k-max poling strategy for selecting top k strongest interactions and produce the result by MLP. Learn the model by Backpropagation and Stochastic Gradient Descent. They demonstrate the experiment on semantic matching for QA and sentence completion. The analysis shows that the MV-LSTM achieve 11.4 % result than the baseline method.

In a work, Saravanakumar Kandasamy and Aswani Kumar Cherukuri [7] proposed a method to improve open domain question answering. There are two components query processing and document processing. Query processing uses POS tagging, Named Entity Recognition, Parsing, Keyword extraction, Finding synsets, and Similarity measurement to create alternate queries. Document processing use URL weight calculation and Latent Semantic Analysis to correct answer retrieval. The precession of the system is 0.77 and Mean Reciprocal Rank(MRR) is 0.79.

Piyush Arora, Jennifer Foster, and Gareth J.F. Jones[8] use a query expansion (QE) methods in information retrieval on WebAP dataset. The different approaches they used are Pseudo Relevance Feedback (PRF), using Robertson term selection and Word Embeddings (WE) of query words to address the query-document term mismatch issue. The embedding of each word is performed by using a feed-forward neural network by predicting a word by its context. The Normalized Discounted Cumulative Gain (NDCG) of the system is 0.16 and Mean Reciprocal Rank (MRR) is 0.36.

Nouha Othman, Rim Faiz, Kamel Smaili [9] discussed a Community Question Answering (CQA) system. They used a word embedding based method to

bridge the lexical gap between the questions. Model the semantic information of words in language vector space by using Word2Vec model. The questions are then ranked by using cosine similarity. The previous question with high similarity score with the new queried question will be returned and the find the corresponding answer. The Mean Average Precision (MAP) ranges from 0.39 to 0.45 on different models.

In a work, Shenghui Wang and Rob Koopman[10] compared word embeddings Word2Vec and GloVe with their own Ariadne approach. They used a neural network-based document embedding method, Doc2Vec with Ariadne approach in the context of Information Retrieval on Medline dataset. The average recall of the Doc2Vec and Ariadne methods is 93.3% and 86.3 % respectively. However, they have shown that Ariadne performs equally well as Doc2Vec in a specific Information Retrieval. If the application is to provide contextual information of a word, Ariadne might be a better choice.

Prathyusha Kanakam, S. Mahaboob Hussain and D. Suryanarayana[11] proposed an algorithm to querying the semantic web. It uses SPARQLquerying language as well as Linked Open Data Quality Assessment(LODQA) for semantic search that converts natural language user's queries to machine-understandable format. The Web Ontology Language(WOL) is used to describe relationships among classes and classifications. Then by using SPARQL to retrieve most accurate results from these ontologies. In this work, the entire approach follows High-Performance Linguistics (HPL) algorithmic process for the proposed system.

In a work, Reshma PK and Lajish VL[12] proposed a semantic Information Retrieval model for University domain using ontology by the help of Protege. Ontology is used to compare conceptual information across two knowledge bases on the web, it formally describes a list of terms which represent important concepts, such as classes of objects and the relationships between them to represent an area of knowledge. Ontology Web Language (OWL) is used to build ontologies. The different steps for building Ontologies are ontology capture, ontology coding and integration with existing Ontologies. The different tasks are define classes and class hierarchy,define object properties and then define instance of ontologies, finally querying with DLQuery.The precision and recall parameters of the system are evaluated as 87% and 56 % respectively.

Pratibha Bajpai, Parul Verma and Syed Q. Abbas[13] discussed the development of English to Hindi Cross-Language Information Retrieval (CLIR) system. They experimented the system with Google and Bing search results documents. They used a two-level word sense disambiguation model to perform disambiguation of Hindi words to the English language. To optimize the translation and disambiguation model by adding a new valuable component analyzer in the basic CLIR architecture. The MAP of Google and Bing queries are 0.45 and 0.35 respectively.

D Thenmozhi and Chandrabose Aravindan[14] developed a Tamil- English Cross-Language Information Retrieval (CLIR) system in the agriculture domain, using Ontology and Word Sense Disambiguation. The MAP of the system is 95.36 percnt. Sumit Kumar Mishra, V.K. Singh [10] also build a semantic Information Retrieval system for legal cases using Ontology merger and extended GAIA methodology, which contains information about legal cases. This model provides reasoning capability too.

Piyush Mital, Saurabh Agrawal al[15] proposed a graph-based question answering system on Wikipedia documents. They create an information extraction and retrieval system from unstructured natural language text documents to structured graphs along with natural language querying. They used the NLP techniques such as, semantic role extraction, phrase chunking, concept extraction etc to better understand input query and generate elements that constitute the graph. The Precision, Recall and Average accuracy of the system was 85.45%, 86.28% and 80.1% respectively.

DwaipayanRoy, Debasis Ganguly al[16] proposed a word embedding base query expansion technique for Information Retrieval on Wikipedia documents. They used two models, i)Word2Vec ii)fastest used subword information for learning. The similarity between the word is calculated with Jaccard similarity. The query terms are matched with embedded word vectors using Indexing Unit Composition(IUC) method. The MAP for Word2Vec and fastText of the system is evaluated as 0.23 and 0.24 respectively. Also, they conclude that Word2Vec works well on stemmed collection and fastText on unstemmed collection.

Shomi Khan, Khadiza Tul Kubra, Md Mahadi Hasan Nahid[17] attempted for improving answer extraction for Bangali Question Answering system. In their work, demonstrated a wed document hierarchy and wordnet for answer retrieval using semantic matching with Anaphora-Catephora-resolution. Wordnet is referred to as a lexical database. The average accuracy of the system is observed as 74%.

In a paper, Bo Xu, Hongfei Lin, Yuan Lin [18] proposed a novel query expansion framework based on learning-to-rank methods for biomedical information retrieval. In the framework, they incorporated the MeSH thesaurus into the co-occurrence-based term selection method to select the candidate expansion terms. To refine the expansion terms, define and extract both the corpus-based term features and the resource-based term features to represent the terms as feature vectors, which are taken as the inputs for learning-to-rank methods to learn the term-ranking models. Different approaches to learning-to-rank are investigated for training the term-ranking models. The Average MAP of the system evaluated as 0.35. In another work [28], they proposed to optimize the pseudo-relevance feedback method, a classic query expansion method, using

learning-to-rank methods to refine the set of expansion terms.

Manasamithra P, H.C Vijayalakshmi[19] proposed a method for convert natural language query to system understandable query using a hybrid approach. Which include keyword-based and semantic-based methods by using an efficient data structure- B-tree to store keywords which act as a knowledge base. The semantic analysis is carried out by using dependency parser. The system has experimented with an employee database. The analysis has shown that the execution time reduced by almost 86% while using B-tree.

In a work, Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, Hong Cheng[20] proposed a semantic question answering system over knowledge graphs. They use a novel systematic method to understand natural language questions using a large number of templates by exploiting the knowledge graph and a large text corpus. The templates are executed by using semantic graphs. To select the target templates, use Semantic Dependency Graph(SDG). The proposed effective strategies to perform entity level and structural level disambiguation during the conversion of natural language queries to structured queries. Finally, a SPARQL query can be constructed, then the corresponding answer will be returned. They conduct the study with Wikipedia text corpus- Dbpedia and freebase. The average precision of Dbpedia and freebase are 84.67% and 82.19% respectively.

Fan fang, Bo-wen Zhang, and Xu-cheng yin [21] developed a Semantic Sequential Dependence Model (SSDM) for Biomedical article search, which is a combination of semantic information and the conventional Sequential Dependence Model (SDM). The synonyms are obtained automatically through the word embeddings, here used word2vec and skip-gram models. They used the neural network-based, SSDM language model. The create a thesaurus by using KNN classification algorithm. Afterwards, the query keywords are extracted and replaced with the synonyms from the thesaurus. Then the synonyms are used to generate possible sequences with the same semantics as the original query and these sequences are input into SDM to obtain the retrieved results.

Liang Pang, Yanyan Lan et al[22] proposed a new deep learning architecture named DeepRank for relevance ranking in Information Retrieval. In their approach, they simulate a human judgement process in relevance ranking. The relevance label is generated by three steps 1) relevant locations are detected by using a query-centric context 2) local relevance ie relevance between query and each query-centric context is determined by using Convolutional Neural Network(CNN) and two-dimensional gated recurrent units(2D-GRU) 3) finally local relevances are aggregated by Recurrent Neural Network(RNN) to output a global relevance score. The DeepRank model is trained by using the Stochastic Gradient Decent(SGD) method. The experiment is evaluated with LETOR4.0 and large scale Chinese click trough data and the MAP for the same is evaluated as 0.49 and 0.41. respectively.

In a work, Ming Zhu, Aman Ahuja et al[23] discussed the development of a neural network model for ranking documents for question answering in health care domain. The proposed model perform deep attention at word, sentence and document level. They also construct a large health care question-answering data set. They use a neural network model, HAR-a Hierarchical Attention Retrieval model for retrieving answers for health-related queries. The different components of the HAR model are 1) Word embedding-create a k-dimension word vector. 2) Encoder-use a bi-directional RNN(Bi-RNN) to encode the inter-document temporal dependencies within query and document words. 3) Compute the relevance of each query word w.r.t each word in the document by using a bi-directional attention mechanism. 4) Query inner attention mechanism used to encode variable-length queries into fixed-size embedding by the self-attention mechanism. 5) Finally use a document hierarchical inner attention mechanism to get a fixed dimensional representation document by using sentence level embedding. Then they use a negative sampling mechanism for optimization of the results. They use health care data set and named it as HeaithQA. The MRR of the system is evaluated as 87.87% and recall as 96.84%.

Zhuyin Dai, Jamie Callan[24] proposed a contextual neural language model-BERT, to provide deeper text understanding for Information Retrieval. BERT (Bi-directional Encoder Representation from Transformers) used for ad-hoc document retrieval. The input for BERT is the concatenation of query and documents tokens, with a special token['SEP'] separating the two segments. Tokens are embedded then separate the query from document embeddings and added to token embedding. The position embedding is also added for word orders. The tokens are gone through several layers of transformations. At each layer, a new contextualized embedding is generated for each token by finding the weighted-sum of all other token embeddings. The weights are calculated by several-attention matrices. Words with strong attention are considered as more close to the target word. Then the output embedding of the first token is used for all query-document pairs. It then inputs into a Multi-Layer Perceptron(MLP) to predict the relevance posibility. This can be augmented with search knowledge. They used two standard datasets- Robust-04-news corpus and Clueweb09-B. About the accuracy, the nDCG of Robust-04 and Clueweb09-B are 0.52 and 0.29 respectively. It is shown that BERT performs well on Robust-04 than Clueweb09-B dataset.

Yuan Zhang, Dong Wang, Yan Zhang[25] developed a Graph Embedding-based ranking model for Product Search(REPS) for e-commerce search. The system integrated the click-graph features into a unified neural ranking framework. In their model, they first introduce a simple neural network architecture as the base model, then plugged a graph embedding technique for better retrieval performance. First, they represent terms of queries and product description as vectors. Then input these vectors to CNN layers for semantic feature ex-

traction, max-pooling layers are used for dimension reduction. Finally use Multi-Layer Perceptron(MLP) to transform semantic feature vectors into the same vector space as query and output relevance score. They used graph embedding during training phase using CNN or RNN. Evaluate the model using the CIKM Cup-2016 Track-2 data set. The MRR, MAP, NDCG of the model is evaluated as 0.49,0.46,0.53 respectively.

In a work, Ping Wang, Tian Shi, Chandan K. Reddy[26] proposed a deep learning-based TRanslate-Edit Model for Question-to-SQL (TREQS) generation for Question Answering on Electronic Medical Records, which adapts the widely used sequence-to-sequence model to generate SQL query for a given query, and performs the required edits using an attentive-copying mechanism and task-specific look-up tables. They created a large-scale Question-SQL pair dataset-MIMICSQL from the publicly available Electronic Medical Records (EMR), it contains two sets, the first set contain template questions and the second consists of natural language questions. Finally Conducted an extensive set of experiments on MIMICSQL dataset for both template questions and natural language questions to demonstrate the effectiveness of the proposed model. They adopt an RNN sequence-to-sequence (Seq2Seq) framework for the Question-to-SQL generation, the encoder reads a sequence of word embeddings of input tokens and turns them into a sequence of encoder hidden states and the testing is performed with implement a beam search algorithm for the SQL generation. Their model gains a significant performance improvement on both development and testing dataset and 30 per cent, on average more accurate than other models. The average accuracy of the system was evaluated as 0.97.

3 Comparative Study and Analysis of Different Models

The comparative analysis of various models discussed in the previous section is tabulated in Table1. Which contain Author, Paper names, Domain and language in which the experiment is conducted, methodologies used and accuracy of the results.

From the literature, it is clear that a few works are done in native languages such as Arabic, Tamil, Bengali etc. Most of the semantic Information Retrieval works are done in the field of English. NLP techniques, Machine Learning and Deep Learning techniques are used for semantic processing. The different methods used are Ontology, Word Sence Disambiguation, CNN, RNN, Word embedding and Document embedding. The recent works are based on deep learning and context level word embedding.

Table 1 Comparative study and Analysis of Different Models.

Sl No	Authors	Domain	Language	Methods	Accuracy
1	Nadia Soudani et al[5]	Semantic space	Arabic	Word Sence Disambiguation	MAP-0.97 to 7.5
2	Shengxian Wan et al [6]	Yahoo Answers	English	LSTM,Bi-LSTM,MLP.	MV-LSTM- 11.4% more
3	Saravanakumar et al [7]	Open Domain	English	Latent Semantic Analysis	Precission- 0.77 MRR - 0.79
4	Piyush Arora et al [8]	WebAP	English	Pseudo Relevance Feedback Word Embedding	NDCG-0.16 MRR-0.36
5	Nouha Othman et al [9]	Yahoo-Webscope	English	Word Embedding Cosine Similarity	MAP- 0.39 - 0.45
6	Shenghui Wang et al [10]	Medline	English	Document Embedding Ariadine	Average recall Doc2Vec-93.3%,Ariadine-86.3%
7	Prathyusha et al[11]	Semantic Web	English	SPARQL,Ontology	-
8	Reshma PK et al[12]	University Data,	English	Ontology,DLQuery	Precision-87% , Recall-56%
9	Pratibha Bajpai et al[13]	Google, Bing	Hindi	Word Sence Disambiguation Component Analyser	MAP- Google-0.45,Bing-0.35.
10	D Thenmozhi et al[14]	Agriculture	Tamil	Ontology Word Sence Disambiguation	MAP-95.36%
11	Piyush Mital et al [15]	Wikipedia	English	Wikipedia Semantic Role extraction	Precision-85.45% Recall-86.28%
12	Dwaipayan Roy et al[16]	Wikipedia	English	Word Embedding- Word2Vec,Fast Text	MAP-Word2Vec-0.23, Fast Text-0.24
13	Shomi Khan et al[17]	Bangali database	Bangali	AnaphoraCatephoraresolution	Avg.Accuracy-75%
14	Bo Xu et al [18]	TREC genomics	English	Pseudo-relevance feedback Learning-to-rank	Avg.MAP-0.35
15	Manasamithra P et al[19]	Employee data	English	Dependency parser B-tree	Time reduced Time reduced-86%
16	Weiguo Zheng et al[20]	Wikipedia DBpedia,Freebase	English	Semantic Dependency Graph,SPARQL	Average precision- Dbpedia-84.67,Freebase-82.19
17	Fan fang et al[21]	MEDLINE	English	Word embedding- Word2Vec,skip-gram	MAP-0.34
18	Liang Pang et al[22]	LETOR4.0, Chinese Click	English	CNN,RNN, 2D-GRU	MAP-LETOR4.0- 0.49 Chinese Click-0.41.
19	Ming Zhu et al[23]	HealthQA	Englush	Word embedding, MLP	MRR-87.87%, Racall-96.84%
20	Zhuyun Dai et al[24]	Robust-04 Clueweb09-B	English	Word embedding, MLP.	NDCG-Robust-04- 0.52 Clueweb09-B-0.29
21	Yuan Zhang et al[25]	E-commerce data- CIKM Cup-2016.	English	Graph Embedding CNN,RNN,MLP.	MRR- 0.49, MAP- 0.46 NDCG- 0.53
22	Ping Wang et al[26]	EMR	English	TRanslate-Edit Model LSTM,RNN.	Avg. accuraccy-0.9

4 Conclusions and Direction for Future Work.

The semantic level Information Retrieval systems are used for retrieval of relevant answers for natural language queries from large-sized unstructured data. The most recent semantic models used the methods of deep learning and context level embedding etc. There is no such semantic level Information Retrieval system exists in the field of Malayalam till now. But there are semantic level models available for other native languages. Here we going to propose a semantic level Malayalam Query-Processing system for heath related Question Answering. The system becomes very helpful for people seeking answers to

their health-related queries.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Green BF, Wolf AK, Chomsky C, and Laughery K. "Baseball: An automatic question answerer". In Proceedings of Western Computing Conference, Vol. 19, 1961, pp. 219224. Proceedings of AFIPS Conference, Vol.42, 1973, pp. 441450
2. Arjun Babu, Sindhu L. "An Information Retrieval System for Malayalam Using Query Expansion Technique",IEEE,978-1-4799-8792-4/15, 2015.
3. Archana S.M., Naima Vahab, Rekha Thankappa, C. Raseek, "A Rule-Based Question Answering System in Malayalam corpus using Vibhakthi and POS Tag Analysis", International Conference on Emerging Trends in Engineering, Science and Technology,ICE TEST 2015.
4. Liji S K and Lajish V L,"An Efficient Malayalam Query Processing System for University Enquiry ", Proceedings of the Eighth National Conference on Indian Language Computing (NCILC 2018), CUSAT, Kerala, March 2018
5. Nadia Soudani, Ibrahim Bounhas, Yahya Slimani."Semantic Information Retrieval: A comparative experimental study of NLP Tools and Language Resources for Arabic", IEEE,DOI 10.1109/ICTAI.2016.133,2375-0197/16, 2016.
6. Shengxian Wan, Yanyan Lan all,"A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations",Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence@2016.
7. Saravanakumar Kandasamy and Aswani Kumar Cherukuri,"nformation Retrieval for Question Answering System Using Knowledge-Based Query Reconstruction by Adapted Lesk And Latent Semantic Analysis", Research Gate, International Journal of Computer Science and ApplicationsJanuary 2017.
8. Piyush Arora, Jennifer Foster, and Gareth J.F. Jones."Query Expansion for Sentence Retrieval Using Pseudo Relevance Feedback and Word Embedding". Springer International Publishing. DOI: 10.1007/978-3-319-65813-1 8, 2017.
9. Nouha Othman, Rim Faiz and Kamel Smaili."A Word Embedding based Method for Question Retrieval in Community Question Answering".ICNLSSP - International Conference on Natural Language, Signal and Speech Processing, ISGA, Dec 2017.
10. Shenghui Wang and Rob Koopman. "Semantic embedding for information retrieval"Workshop on Bibliometric-enhanced Information Retrieval.,BIR 2017
11. Prathyusha Kanakam, S. Mahaboob Hussain and D. Suryanarayana."HPL Algorithm For Semantic Information Retrieval with RDF And SPARQL".Research Article.Jr. of Industrial Pollution Control 33,2017.
12. Reshma PK, Lajish VL. "Ontology-Based Semantic Information Retrieval Model for University The domain" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 2018 pp. 12142-12145, Number 15 2018.
13. Pratibha Bajpai, Parul Verma and Syed Q. Abbas."English-Hindi Cross-Language Information Retrieval System: Query Perspective".Pratibha Bajpai et al. / Journal of Computer Science. DOI:10.3844/jcssp.2018.705.713, 2018.
14. D Thenmozhi and Chandrabose Aravindan. "Ontology-based TamilEnglish cross-lingual information retrieval system". Indian Academy of Sciences.Sdhan, https://doi.org/10.1007/s12046-018-0942-7,2018
15. Piyush Mital, Saurabh Agrawal al."Graph-based Question Answering System", IEEE,978-1-5386-5314-2/18/ 2018.
16. DwaipayanRoy, Debasis Ganguly al."Using Word Embeddings for Information Retrieval: How Collection and Term Normalization Choices Affect Performance". 3269206.3269277 CIKM '18, ACM.,Torino, Italy, October 2018

17. Shomi Khan, Khadiza Tul Kubra, Md Mahadi Hasan Nahid, "Improving Answer Extraction For Bangali Q/A System Using Anaphora-Cataphora Resolution". International Conference on Innovation in Engineering and Technology (ICIET),IEEE, 27-29 December, 2018,978-1-5386-5229-9/18/02018.
18. Bo Xu, Hongfei Lin, Yuan Lin. "Learning to Refine Expansion Terms for Bio-medical Information Retrieval Using Semantic Resources". 10.1109/TCBB..2801303, IEEE/ACM, 2018
19. Manasamithra P, H. C. Vijayalakshmi "NLP for Information Retrieval using B Trees". International Journal of Computer Applications (0975 8887).Volume 182 No.5, July 2018
20. Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, Hong Cheng. "Question Answering Over Knowledge Graphs: Question Understanding Via Template Decomposition". Proceedings of the VLDB Endowment, Vol. 11, No. 11.August.DOI:https://doi.org/10.14778/3236187.3236192. 2018
21. Fan fang, Bo-wen zhang, and Xu-cheng yin. "Semantic Sequential Query Expansion for Bio-medical Article Search", 2169-3536 IEEE, 2018.
22. Liang Pang, Yanyan Lan et al." Deep Rank: A New Deep Architecture for Relevance Ranking in Information Retrieval".arXiv:1710.05649v2 [cs.IR] ,ACM,22 Jul 2019.
23. Ming Zhu, Aman Ahuja et al."A Hierarchical Attention Retrieval Model for Healthcare Question Answering"., IW3C2 (International World Wide Web Conference Committee), ACM ,ISBN 978-1-4503-6674-8/19/05. https://doi.org/10.1145/3308558.3313699,2019
24. Zhuyun Dai, Jamie Callan. "Deeper Text Understanding for IR with Contextual Neural Language Modeling". Association for Computing Machinery. ACM ISBN 978-1-4503-6172-9/19/07, https://doi.org/10.1145/3331184.3331303, 2019.
25. Yuan Zhang, Dong Wang, Yan Zhang. "Neural IR Meets Graph Embedding: A Ranking Model for Product Search".The Web Conference, San Francisco, CA, USA. ACM ISBN 123-4567-24-567,May 2019.
26. Ping Wang, Tian Shi, Chandan K. Reddy. "Text-to-SQL Generation for Question Answering on Electronic Medical Records".IW3C2 (International World Wide Web Conference Committee).,ACM ISBN 978-1-4503-7023-3/20/04,2020.
27. Keet Sugathadasa, Buddhi Ayesha al. "Legal Document Retrieval using Document Vector Embeddings and Deep Learning". Computing Conference 2018 10-12 July 2018.
28. Xu, B., Lin, H., Lin, Y. (2016). "Assessment of learning to rank methods for query expansion". Journal of the Association for Information Science and Technology, 67(6): 1345-1357.2016.
29. Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria. "Recent Trends in Deep Learning Based Natural Language Processing", arXiv, IEEE Computational intelligence magazine,Nov 2018.
30. T. Kawamura, K. Kozaki, T. Kushida, K. Watanabe, and K. Matsumura, "Expanding science and technology thesauri from bibliographic datasets using word embedding," in Proc. IEEE Int. Conf. Tools Artif. Intell., Nov. 2017.
31. Shickel, B., Tighe, P. J., Bihorac, A, Rashidi "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis", JBHI.2767063, IEEE, 2017.
32. Jimmi, Guido Zuccon and Bevan Koopman."Knowledge Graphs And Semantics In Text Analysis And Retrieval".Information Retrieval Journal.https://doi.org/10.1007/s10791-018-9344-z. Springer Nature B.V. 2018.
33. Sumit Kumar Mishra, V.K. Singh. "Building Semantic Information Retrieval System For Legal Cases From Heterogeneous Adapted And Diverse Data Sources Using Extended GAIA Methodology For Multi-Agent System",by IEEE. 978-1-5090-6785-5/18/ 2018 .