Open Information Extraction for Knowledge Graph Construction

Iqra Muhammad, Anna Kearney, Carrol Gamble, Frans Coenen, and Paula Williamson.

Department of Computer Science, The University of Liverpool, Liverpool, L693BX, UK
iqra@liverpool.ac.uk, frans.coenen@liverpool.ac.uk

Abstract. An open information extraction approach for knowledge graph construction is presented. The motivation for the work is that large quantities of scholarly documents are available within many domains of discourse, and the subsequent challenge is to identify the most relevant articles concerning a particular topic. The proposed approach takes a document corpus and identifies triples within this corpus which are then processed to generate a literature knowledge graph. The extraction of triples is conducted using an open information extraction approach. The proposed OIE4KGC approach was evaluated using a bespoke clinical research methodology dataset and a benchmark dataset. A f-score of 51% was achieved on a clinical research methodology dataset and a f-score of 37% was achieved on the benchmark dataset.

Keywords: Open Information Extraction · Literature Knowledge Graph Construction

1 Introduction

The number of available scientific papers has been increasing at an exponential rate. In 2009, it was estimated that the 50 million mark in the number of scientific papers was passed [2]. One solution is online article repositories, which typically feature some limited form of search facility. One example is the abstracts stored in the MEDLINE[1] repository, which can be accessed (searched) using the PubMed[2] interface. However the search functionality supported by these systems is typically inadequate for efficiently searching large repositories. An alternative solution, which is advocated in this paper, is to store the document corpus in a literature knowledge graph [14, 15] where the vertices represent concepts and documents, and the edges represent relationships between concepts, or concepts and documents. This, it is suggested, will provide a better organisation of the data and consequently provide for more effective information retrieval and knowledge understanding.

To create a literature knowledge graph, information extraction techniques are applied to the unstructured text in the document corpus. The extracted information can then be used to build the desired literature knowledge graph. However, there are many challenges to building effective information retrieval systems regardless of the domain

[1] https://www.nlm.nih.gov/bsd/medline.html
[2] https://www.ncbi.nlm.nih.gov/pubmed/

of discourse. A particular issue is that the vocabulary in many domains tends to be extensive, compounded by the fact that there are often semantic variations for the same concept and that the relationships between concepts are often complex. This is especially the case in the clinical domain.

Traditional information extraction techniques for building knowledge graphs tend to use a pre-defined schema, an agreed set of specific concept types and relation types for vertices and relations; and typically operate using domain-specific supervised learning approaches that require training data [12, 23–26]. However, the training data and a schema specific to a domain of discourse is typically not available in many cases. An alternative is to use domain independent Open Information Extraction (OIE) models that are already pre-trained on general datasets. The knowledge graphs constructed using OIE do not require a pre-defined schema. Open information extraction techniques make use of a set of patterns to extract triples. Each triple consists of two arguments, a subject and an object, and a predicate (relation) linking the arguments, which can then be used to construct a knowledge graph [8].

This paper presents the Open Information Extraction for Knowledge Graph Construction (OIE4KGC) approach; a novel process for generating a literature knowledge graph from a given corpus using the concept of OIE. The focus for the work is clinical trial's methodology literature; an essential resource in facilitating clinical trials research. The dataset used for evaluation purposes consisted of 400 abstracts on recruitment strategies for clinical trials, selected from the ORRCA dataset[3] [4]. The abstracts in this dataset represent recruitment strategies, adopted by clinical trials, when recruiting patients for trials.

The rest of this paper is structured as follows. Section 2 considers previous work directed at the concept of creating knowledge graphs from unstructured text. In Section 3 the proposed OIE4KGC approach is described. Section 4 then presents the evaluation of the proposed OIE technique for knowledge graph generation. Finally, Section 5 concludes the paper with a summary of the main findings and directions for future work.

2 Literature Review

This literature review section presents an overview of existing work on open information extraction and knowledge graph construction relevant to the work presented here. It has been divided into two sections. The first, Section 2.1 considers OIE techniques; and the second, Section 2.2, describes some of the existing work on knowledge graphs.

2.1 Open Information Extraction

Recently, many attempts have been made at using OIE techniques for converting unstructured text to structured text. Existing techniques based on the idea of OIE use a set of patterns to convert a sentence into relational triples. OIE techniques can be divided

[3] The ORRCA (Online Resource for Recruitment Research in Clinical Trials) dataset is part of a PhD with the University of Liverpool's Biostatistic's department. This dataset will be released publicly on the author's website

into three categories: (i) learning-based, (ii) rule-based and (iii) inter-proposition-based:

Learning-based systems Learning-based open information extraction systems use training data, from which a model is learned for producing relational triples. One of the first systems directed at learning-based OIE was TextRunner [5]. Using TextRunner, a small sample of sentences are first parsed using Penn Treebank after which a dependency parser is applied to identify and label a set of "extractions" as positive and negative training examples. In [28] an open information extraction system is used, and relies on a bootstrapping approach based on a wikipedia dataset. In [8] an OIE system, called RnnOIE[4] founded on a deep-learning based approach was presented. RnnOIE is a model pre-trained on the OIE2016 dataset. The reported experiments demonstrated that RnnOIE, was able to outperform many state of the art benchmarks. The RnnOIE tool was therefore adopted with respect to the OIE4KGC approach described in this paper.

Rule-based systems A number of approaches to OIE make use of hand-crafted extraction rules. One example is REVERB [7], this is a "shallow extractor" that makes use of hand-craft extraction rules. REVERB addresses the problems of uninformative and incoherent extractions. Another rule-based approach is PredPratt [16], which used a set of non-lexicalised rules, defined over universal dependency parses, for extracting predicate-argument structures. The disadvantage is the need to hand-craft the rule set. The approach was therefore deemed inappropriate for the knowledge graph generation application.

Inter-proposition relationship based systems OIE systems extract a list of relational triples also called propositions, where each proposition consists of a single predicate and a number of arguments from an input sentence. The majority of the above mentioned OIE systems are not capable of capturing the complete expression in a sentence as they ignore the context under which a proposition is complete. An example of such a scenario is the relational triple ⟨ *Barack Obama, was a , good president* ⟩ from the sentence "Democrats believe that Barack Obama was a good president"; this triple is inappropriate since the input sentence is not asserting this proposition. Such shortcomings can be handled by OLLIE [27], which adds an additional attribute context to the extracted relation triple or proposition, showing that a proposition is reported by some entity (*AttributedTo believe; Democrats*). This idea of adding additional attributes to an extracted relation triple or proposition is referred to as inter-propositional relation. Another similar state-of-the-art approach was proposed in [13] where a nested representation for OIE was presented. This approach was able to capture high-level dependencies, allowing for an improved representation of the meaning of a sentence. However, for the purpose of building literature knowledge graphs this limitation of learning based systems was accepted; additional attributes could always be added at a later date.

2.2 Knowledge Graphs

As noted in the introduction to this paper, knowledge graphs are labelled graphs where vertices represent concepts and edges represent relations between them. Previous work

[4] https://github.com/gabrielStanovsky/supervised-oie

on the automatic construction of knowledge graphs can be divided into two categories: (i) Domain-Specific Knowledge graphs and (ii) Literature Knowledge graphs:

Domain Specific Knowledge Graphs Domain Knowledge Graphs, as the name implies, are domain-specific, meaning that the text used in the construction of the knowledge graph is limited to a specialised field like biology, physics, computer science or any other domain of discourse. One of the first few attempts at creating a knowledge graph in the biomedical science domain involved the use of rdf-extraction from excel sheets in [21]. A recent, frequently cited work [22] focused on the construction of a knowledge graph for the domain of biomedical sciences.

Literature Knowledge Graphs Literature knowledge graphs act as a storage mechanism for representing concepts and relations in the literature associated with some domain of interest. A well-known literature knowledge graph, is that used within Semantic Scholar is presented in [15]. Another well-known literature knowledge graph was created by Microsoft and comprised author vertices, concept vertices, paper vertices and edges connecting them [29].

3 Open Information Extraction for Knowledge Graph Construction (OIE4KGC)

In this section the proposed OIE4KGC approach is presented. Sub-section 3.1 first gives a problem definition for literature knowledge graph construction. There are two kinds of vertices in the envisioned knowledge graph: (i) concept vertices and (ii) document vertices. The vertices are linked by edges representing relationships. The proposed approach is illustrated in Figure 1 with an example taken from the ORRCA dataset used for evaluation purposes. From the figure it can be seen that OIE4KGC comprises four main components: (i) Triple Extraction. (ii) Triple Filtering, (iii) Concept Linking and (iv) Merging of vertices and Knowledge graph population. Each is discussed in further detail in Sub-sections 3.2, 3.3 and 3.4. The pseudocode for the OIE4KGC is given in Algorithm 1; this will be referred to in the following sub-sections.

3.1 Problem Definition

The aim is to construct a literature knowledge graph $G = \{V, E\}$ where the set of vertices V represent documents (abstracts) or concepts, and the set of edges E represent relationships. Given, a corpus of n documents (abstracts), $D = \{D_1, \ldots, D_n\}$, where each document is comprised of m sentences $S = \{S_1, \ldots, S_m\}$, the task is to find an appropriate set of triples T from each sentence in each document and use these triples to construct G. A triple T takes the form $\langle a_s, r, a_o \rangle$, where a_s is the subject argument, a_o is the object argument and r is a predicate (relation) between them; each is represented by a string. The arguments represent concepts which may potentially be included in the eventual knowledge graph.

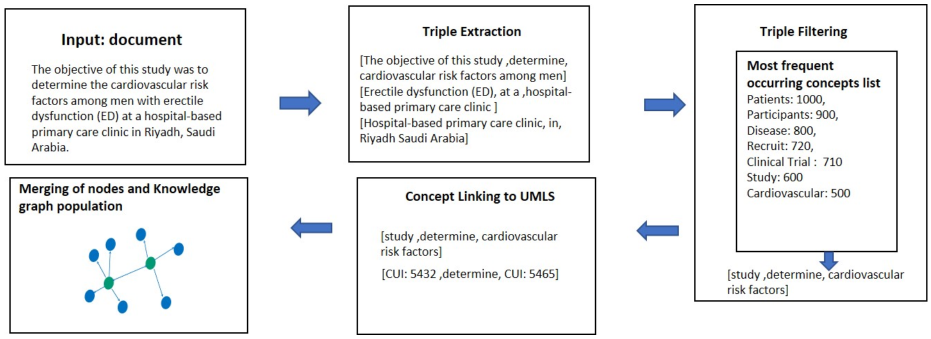

Fig. 1. Stages involved in the construction of a literature knowledge graph using OIE4KGC

3.2 Triple Extraction

Triple Extraction is the first stage in the proposed approach. A variety of tools are available that can be used to extract triples from unstructured text, both supervised and unsupervised. The assumption was, as in the case of the clinical research methodology scenario used as the focus for this paper, that training data was not available. Hence a semi-supervised approach was required; a pre-trained OIE tool of the form discussed in Section 2. For the proposed approach RnnOIE [8] was used because as reported in [8], RnnOIE had been shown to outperform other state of the art tools for OIE using benchmark datasets.

The first step in the application of any OIE tool is the pre-processing of the data so as to identify sentences, $S = \{S_1, \ldots, S_m\}$ (line 6 in the pseudo code given in Algorithm 1) for every document D_i in the corpus D (line 4 in Algorithm 1). With respect to the proposed approach the Spacy's sentence segmentation tool was used[5]. The second step (line 9) was to apply RnnOIE to each sentence [8]. In this manner the predicates and arguments in each sentence could be identified without requiring any domain specific knowledge. In Figure 1 three triples are identified. The first of these $\langle$*the objective of this study, determine, cardiovascular risk factors among men*$\rangle$, where *determine* is the relation (predicate), and *the objective of this study* and *cardiovascular risk factors among men* are its arguments. The identified argument and relation strings, as illustrated in the example, were expected to include unnecessary words which should be removed. This was done (line 11) using Spacy's Noun Chunker so that only the informative noun phrases for arguments were retained. Thus, the above example will be reduced to $\langle$*study, determine, cardiovascular risk factors*$\rangle$.

3.3 Triple Filtering

The aim of the triple filtering stage was to filter the triples extracted from each of the abstracts in a given corpus during the Triple Extraction Stage (Stage 1) and removed redundant and uninformative words within arguments. As a result some arguments would

[5] https://spacy.io/

Algorithm 1 OIE for Knowledge graph Pseudocode

1: Input D, Output G
2: D = A set of Documents, G = Empty Knowledge graph database
3: L = Lexicon of most frequently occurring words in D
4: **for** $D = \{1, 2, \ldots, i\}$ **do**
5: $\quad G = G$ plus vertex representing D_i
6: $\quad S =$Set of Sentences in D_i
7: $\quad T = \oslash$ (Set to hold triples)
8: $\quad$ **for** $S = \{1, 2, \ldots, i\}$ **do**
9: $\quad\quad T_i$ = Set of triples in S_i
10: $\quad\quad$ **for** $t = \{1, 2, \ldots, i\}$ **do** where $t_i = \langle a_s, r, a_o \rangle$ in T_i
11: $\quad\quad\quad t_i = t_i$ with noun chunking applied
12: $\quad\quad\quad t_i = t_i$ with only nouns that are retained after checking L
13: $\quad\quad\quad t_i = t_i$ annotated with CUI
14: $\quad\quad$ **end for**
15: $\quad\quad T = T \cup T_i$
16: $\quad$ **end for**
17: $\quad G = G$ incorporating $T = \{T_1, \ldots, T_i\}$
18: **end for**
19: Exit with G

be "empty". Informative words were considered to be words that appear frequently in the given corpus. To this end a "Most frequent occurring concepts list" was constructed; a lexicon L of the 100 most frequently occurring nouns in the corpus. A fragment of such a lexicon is given in Figure 1; alongside each entry is its associated occurrence count. For each argument in each triple the words appearing in L were retained (line 12). In Figure 1 the triple $\langle$*study, determine, cardiovascular risk factors*$\rangle$ remains unchanged because all these word are present in L. The objective of the triple filtering stage was also to ensure, that if the triples are converted to knowledge graph embeddings, the embeddings would not be sparse.

3.4 Linking of Clinical Concepts to UMLS

The arguments contained in the triples retained after Stage 2 express concepts to be included in the knowledge graph. Several arguments might express the same concept, whilst at the same time a single argument can express different concepts depending on context. To construct a useful knowledge graph these ambiguities need to be resolved. The central idea for achieving this was to annotate each argument with additional information, for example with its synonyms and/or hyponyms, that would allow disambiguation. In the case of the clinical research methodologies application domain this was achieved by annotating each argument with the relevant Concept Unique Identifier (CUI) held in the Unified Medical Language System (UMLS) Metathesaurus [3]. Using the words and phrases held in the metathesaurus, the arguments in the identified triples were annotated with a CUI indicating the sense of the argument (line 13). For example the word "study" has the CUI 5432, while the phrase "cardiovascular risk factors" has the CUI 5465 (as indicated in Figure 1). These CUI annotations were then used for disambiguating purposes in Stage 4.

3.5 Merging of Vertices and Knowledge Graph Population

The final stage in the proposed approach is the construction of the desired knowledge graph (line 17 in the pseudo code). A knowledge graph can be represented using a variety of graph database models, with respect to the work presented in this paper Neo4j was used[6]. Custom data-structures were created for concept vertices, document vertices and for relations. The arguments within each triple represent concepts to be included in the knowledge graph. Figure 2 shows a toy example of a literature knowledge graph generated using the proposed OIE4KGC approach. In the figure there are two types of vertices: (i) concept vertices (blue) and (ii) document vertices (green). Each concept vertex has two properties, (i) the argument string (the concept name) and (ii) the associated CUI based ID that links the argument string to the UMLS Metathesaurus sense (included to add additional information). Each document vertex references a document (abstract). A document vertex in the knowledge graph also has two properties: (i) the title string of the abstract and (ii) a unique identifier (it cannot be assumed that each document will have a unique title). There are two kinds of edge in the knowledge graph: (i) edges linking concepts and (ii) edges linking concepts and documents. Edges linking documents and concepts have the label "mentions" indicating that the document mentions the indicated concept. Edges linking a pair of concepts indicate relations extracted using OIE.

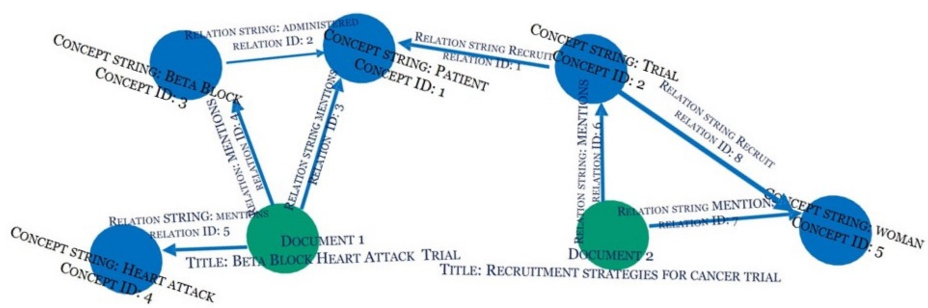

Fig. 2. A toy example of a literature knowledge graph generated using OIE4KGC

To generate the literature knowledge graph each triple was processed in turn. For each triple two new vertices were created, v_s and v_o, connected by the given relation r, and each connected to the document vertex created for D_i. These were then compared to the knowledge graph G so far. There are four options:

1. If v_s and v_o match two vertices v_1 and v_2 in G: merge v_s and v_o with v_1 and v_2 adding the relation r if not already in existence.

[6] https://neo4j.com/

2. If v_s matches a vertex v_1 in G, but v_o does not match any vertex in G: merge v_s with v_1.
3. If v_o matches a vertex v_2 in G, but v_s does not match any vertex in G: merge v_o with v_2.
4. Otherwise (v_s and v_o do not match any vertices in G): do nothing.

To facilitate the merging Neo4j has a merge utility.

4 Evaluation

This section describes the evaluation conducted to assess the performance of the proposed approach. The evaluation was centred on the RnnOIE OIE tool [8] central to the proposed OIE4KGC approach. For the evaluation two data sets were used: (i) the ORRCA data set [4] and (ii) the Reverb ClauseIE dataset[7] [18]. For the ORRCA data set 100 sentences were randomly chosen and a "gold standard2" set of triples identified by manual inspection of the 100 sentences. The ClauseIE data set is a benchmark dataset of 500 sentences manually labelled for OIE; 100 sentences were randomly selected from the ClauseIE data set. The evaluation metric used was F-score, the harmonic mean of precision and recall. Note that precision was defined as the number of correct triples divided by the total number of triples extracted by RnnIE tool, whilst recall was defined as the number of correct triples divided by the number of triples in the gold standard for selected 100 sentences. It should be noted that the objective of this evaluation was to assess the RnnOIE tool at the sentence level.

The results obtained are given in Table 1. From the table it can be seen RnnOIE was able to achieve an F-score of 51% using the ORRCA data set and 37% that using the ClauseIE dataset. It can also be be seen from Table 1 that the precision was better using the ORRCA dataset compared to the ClauseIE dataset. This difference in precision can be accounted for by the structural differences in sentences in both datasets. Triples extracted from the ClauseIE dataset have numerical values in the arguments; which, using the proposed approach, results in a triple being discarded. It is also note-worthy that the sentences in the ORRCA dataset are longer than in the case of the ClauseIE dataset; the average number of words in each sentence for the ORRCA dataset was 30 compared to 10 for the ClauseIE dataset. From the results it can be concluded that RnnOIE is appropriate for clinical document collections as exemplified by the ORRCA dataset, and appropriate for inclusion in the proposed OIE4KGC approach advocated in this paper.

Dataset	Precision	Recall	F-score
ClauseIE dataset	0.473	0.311	0.375
ORRCA dataset	0.783	0.391	0.512

Table 1. Table showing the performance of the RnnIE tool on the ORRCA and ClauseIE datasets

[7] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/clausie/

5 Conclusion and Future Work

This paper has presented the Open Information Extraction for Knowledge Graph Construction (OIE4KG) approach for constructing literature knowledge graphs. The focus of the work was a clinical trials methodological articles collection. Open information extraction was used for the extraction of triples from the document collection. The RnnOIE too was evaluated using two datasets, ORRCA and ClauseIE. The F-score of 51% percent using the ORRCA dataset suggests that OIE tools such as RnnOIE can be successfully used to construct literature knowledge graphs in the clinical domain. In terms of future research, the intention is to focus on canonicalizing the knowledge graph and using the knowledge graph embeddings for tasks like document retrieval and document ranking.

References

1. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction.
2. Jinha, A.E.: Article 50 million: an estimate of the number of scholarly articles in existence. Learned Publishing. 23, 258–263 (2010).
3. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 32, (2004).
4. Kearney, A., Harman, N.L., Rosala-Hallas, A., Beecher, C., Blazeby, J.M., Bower, P., Clarke, M., Cragg, W., Duane, S., Gardner, H., Healy, P., Maguire, L., Mills, N., Rooshenas, L., Rowlands, C., Treweek, S., Vellinga, A., Williamson, P.R., Gamble, C.: Development of an online resource for recruitment research in clinical trials to organise and map current literature. Clinical Trials. 15, 533–542 (2018).
5. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner. Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX - NAACL 07. (2007).
6. Weld, D.S., Hoffmann, R., Wu, F.: Using Wikipedia to bootstrap open information extraction. ACM SIGMOD Record. 37, 62 (2009).
7. Fader, Anthony, Zettlemoyer: Paraphrase-Driven Learning for Open Question Answering. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1608–1618 (2013).
8. Stanovsky, G., Michael, J., Zettlemoyer, L., Dagan, I.: Supervised Open Information Extraction. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). (2018).
9. Cui, L., Wei, F., Zhou, M.: Neural Open Information Extraction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). (2018).
10. Zhan, Junlang, Zhao, Hai: Span Model for Open Information Extraction on Accurate Corpus, https://arxiv.org/abs/1901.10879.
11. Jiang, M., Shang, J., Cassidy, T., Ren, X., Kaplan, L.M., Hanratty, T.P., Han, J.: MetaPAD. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17. (2017).

12. Qin, L., Hao, Z., Yang, L.: Question Answering System based on Food Spot-Check Knowledge Graph. Proceedings of 2020 the 6th International Conference on Computing and Data Engineering. (2020).
13. Bhutani, N., Jagadish, H.V., Radev, D.: Nested Propositions in Open Information Extraction. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016).
14. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., Dsouza, J., Kismihók, G., Stocker, M., Auer, S.: Open Research Knowledge Graph. Proceedings of the 10th International Conference on Knowledge Capture - K-CAP 19. (2019).
15. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.-H., Peters, M., Power, J., Skjonsberg, S., Wang, L., Willhelm, C., Yuan, Z., Zuylen, M., Oren: Construction of the Literature Graph in Semantic Scholar. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). (2018).
16. White, A.S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., Durme, B.V.: Universal Decompositional Semantics on Universal Dependencies. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016).
17. Bhutani, N., Jagadish, H.V., Radev, D.: Nested Propositions in Open Information Extraction. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016).
18. Corro, L.D., Gemulla, R.: ClausIE. Proceedings of the 22nd international conference on World Wide Web - WWW 13. (2013).
19. Zhao, S., Su, C., Sboner, A., Wang, F.: Graphene. Proceedings of the 28th ACM International Conference on Information and Knowledge Management - CIKM 19. (2019).
20. Huang, Z., Yang, J., Harmelen, F.V., Hu, Q.: Constructing Knowledge Graphs of Depression. Health Information Science Lecture Notes in Computer Science. 149–161 (2017).
21. Han, L., Finin, T., Parr, C., Sachs, J., Joshi, A.: RDF123: From Spreadsheets to RDF. Lecture Notes in Computer Science The Semantic Web - ISWC 2008. 451–466 (2008).
22. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics. 41, 706–716 (2008).
23. Haussmann, S., Seneviratne, O., Chen, Y., Ne'Eman, Y., Codella, J., Chen, C.-H., Mcguinness, D.L., Zaki, M.J.: FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation. Lecture Notes in Computer Science The Semantic Web – ISWC 2019. 146–162 (2019).
24. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. (2018).
25. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, Relation, and Event Extraction with Contextualized Span Representations. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). (2019).
26. Silva, V., Freitas, A., Handschuh, S.: Building a Knowledge Graph from Natural Language Definitions for Interpretable Text Entailment Recognition, https://www.aclweb.org/anthology/L18-1542/.
27. Schmitz, Michael: Open Language Learning for Information Extraction, https://www.aclweb.org/anthology/D12-1048.pdf. (2010)
28. Weld, D., Hoffmann, R., Wu, F.: Using Wikipedia to bootstrap open information extraction. ACM SIGMOD Record. 37, 62 (2009).
29. Microsoft Academic Knowledge Graph, http://ma-graph.org/.