Expert Systems With Applications 112 (2018) 372–387

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A systematic mapping study on open information extraction

Rafael Glauber, Daniela Barreiro Claro*

Formalisms and Semantic Applications Research Group (FORMAS), LASiD/DCC/IME, Federal University of Bahia, Salvador, Bahia, Brazil

ARTICLE INFO

Article history:
Received 14 November 2017
Revised 4 June 2018
Accepted 21 June 2018
Available online 27 June 2018

Keywords:
Systematic mapping study
Open information extraction
Open relation extraction
Open knowledge acquisition
Open relation mapping

ABSTRACT

Open information extraction (Open IE) is a task for extracting relationship triples in plain texts without previously determining these relationships. The Open IE systems are generally applied to solutions on the web-scale such improving question answering systems, ontology constructions, document filtering and clustering. Since 2007, within the first Open IE system *TEXTRUNNER*, other related works have been proposed in this area. Despite other secondary studies on Open IE, useful information available to initiate new research in the area is limited. Thus, we propose a review of the literature in Open IE by a systematic mapping study. We have retrieved 2484 articles about Open IE in Science Direct, IEEE Xplore, ACM Digital Library, Scopus and Google Scholar databases. Among them, 2411 were filtered by exclusion criteria proposed in our systematic mapping protocol. The remaining 73 papers represent the state-of-the-art from the past seven years. Different researchers have proposed important contributions and have pointed out some open problems for Open IE. As a result, we summarized these contributions and identified significant gaps that could be envisioned as future works.

1. Introduction

Nowadays the Web publishes many data in texts written in a natural language format. The number of those texts are growing every day. Nevertheless, it is hard for humans to extract useful information from such a large data repository. In general, knowledge available in texts is expressed by relationships between entities. The named entity recognition area has high rates of precision (Nadeau & Sekine, 2007). Identifying relationships between entities is not a trivial task. Information extraction (IE) is a research area that automates the extraction of information from textual documents (de Abreu, Bonamigo, & Vieira, 2013).

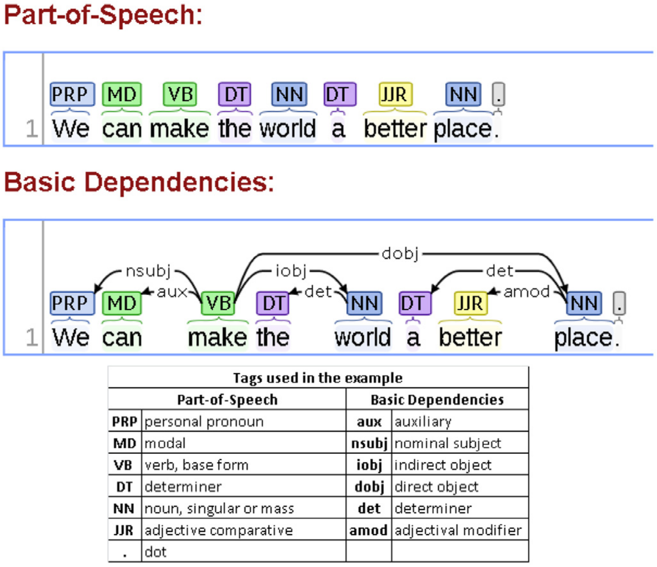
During the first years of research, IE was applied to a small number of texts and specific domains. Common applications were *Biotext* (Liu, Shi, & Sarkar, 2007) to identify links among proteins, genes and diseases. Another application is the improvement of question and answering systems (QA) (Jijkoun, De Rijke, & Mur, 2004). According to Fader, Soderland, and Etzioni (2011) and Xavier, de Lima, and Souza (2015), traditional IE is based on training an extractor with some target relationships previously defined. The main drawback of traditional IE is its low coverage and its well fitting when applied to a particular domain. When texts are from different domains, human intervention may be required. To overcome this problem, open information extraction (Open IE) has emerged to extract facts without determining a set of relationships previously. This new topic of research in IE had the first study published in 2007 (Banko, Cafarella, Soderland, Broadhead, & Etzioni, 2007). After the *TEXTRUNNER* system, many other studies have been issued. The increase of research on Open IE has raised to some revision studies on this area (Gamallo, 2014; Garcia, 2016; Konstantinova, 2014; Xavier et al., 2015). Although traditional revisions provide broad content on a research topic, typically data and information are unsystematically collected. A known effect of these revisions is to present the vision of a single researcher or a group on a topic. In these cases, the lack of rules for collecting data can lead to biased the discussions.

According to Petersen, Feldt, Mujtaba, and Mattsson (2008), a systematic mapping study (SMS) is a methodology frequently used in medical investigation. In recent years, it has frequently been applied in computer science, mainly, in software engineering. An SMS aims to classify, to conduct a thematic analysis, to identify forums and potential research gaps (Petersen et al., 2008). The difference from traditional revision is a rigid definition of rules for conducting the study. To the best of our knowledge, there is no mapping study about "open information extraction". To organize the literature on Open IE, we follow a systematic mapping process proposed in Petersen, Vakkalanka, and Kuzniarz (2015).

The mapping study of this paper is organized as follows: Section 2 presents the principles of the area; Section 3 discusses some related works on mapping study; Section 4 describes the

* Corresponding author.
E-mail addresses: rglauber@dcc.ufba.br (R. Glauber), dclaro@ufba.br (D. Barreiro Claro).

https://doi.org/10.1016/j.eswa.2018.06.046
0957-4174/© 2018 Elsevier Ltd. All rights reserved.


Part-of-Speech:
PRP MD VB DT NN DT JJR NN .
1 We can make the world a better place.
Basic Dependencies:
nsubj
aux
iobj
dobj
det
det
amod
PRP MD VB DT NN DT JJR NN .
1 We can make the world a better place.

Tags used in the example

Part-of-Speech		Basic Dependencies	
PRP	personal pronoun	aux	auxiliary
MD	modal	nsubj	nominal subject
VB	verb, base form	iobj	indirect object
DT	determiner	dobj	direct object
NN	noun, singular or mass	det	determiner
JJR	adjective comparative	amod	adjectival modifier
.	dot		

Fig. 1. Part-of-speech and basic dependencies generated by Stanford CoreNLP. The verb "make" is the root of the dependency tree.

main configuration of our SMS; Section 5 presents our main results and some discussions about them; In Section 6 we present some threats to validity and how to avoid them. Finally, Section 7 concludes our study with some discussions.

2. Principles

An Open IE system performed the task of extracting relationship triples (facts) in raw texts written in natural language in the format:

$$triple = (arg1, rel, arg2) \quad (1)$$

where, *arg*1 and *arg*2 are noun phrases that have a semantic relationship determined by *rel* such as verb phrases. Authors in (Gamallo, 2014) propose two main steps to organize Open IE methods: (a) Systems that use shallow analysis or dependency analysis for sentences annotation and (b) systems that use machine learning or handcrafted rules for extract relationship triples. Shallow analysis are conducted by algorithms that identify constituent parts of a sentence. For example, part-of-speech (POS) tagger that identifies word-by-word morphological classes contained in a sentence. Another example is the algorithm of noun or verb phrase chunking that identifies the constituting parts of a sentence. These algorithms consider the sentence as a tree, but only allow access to parts of it separately. On the other hand, dependency-analysis algorithms are able to provide the link among all the words of a sentence. All members of a sentence are accessed from the root (usually a verb). Fig. 1 presents two examples using Stanford CoreNLP (Manning et al., 2014) for POS tagger and basic dependencies[1].

Dependency-analysis algorithms have a higher computational cost against shallow analysis. However, the methods that have used dependency analysis has been presenting best results (*e.g. ClausIE* Del Corro & Gemulla, 2013 and *OLLIE* Schmitz, Bart, Soderland, & Etzioni, 2012 according to Rodríguez, Merlino, Pesado, & García-Martínez, 2016). The first Open IE system called *TEXTRUNNER* (Banko et al., 2007) uses a machine learning method to the extraction task. However, from *REVERB* system (Fader, Soderland, & Etzioni, 2011) there was a new trend on Open IE systems with a higher concentration of methods using handcrafted rules (usually implemented by regular expressions). The combination of these two main steps creates four new categories for Open IE: (i) Shallow analysis and machine learning method, (ii) shallow analysis and handcrafted method, (iii) dependency analysis and machine learning method and (iv) dependency analysis and handcrafted method. Gamallo (2014) presents examples for each category. Despite acknowledging such kind of organization, methods such as *REVERB* are difficult to classify. This method has a syntactic constraint to accomplish the extractions (handcrafted rules) followed by a lexical constraint to increase its precision (machine learning method).

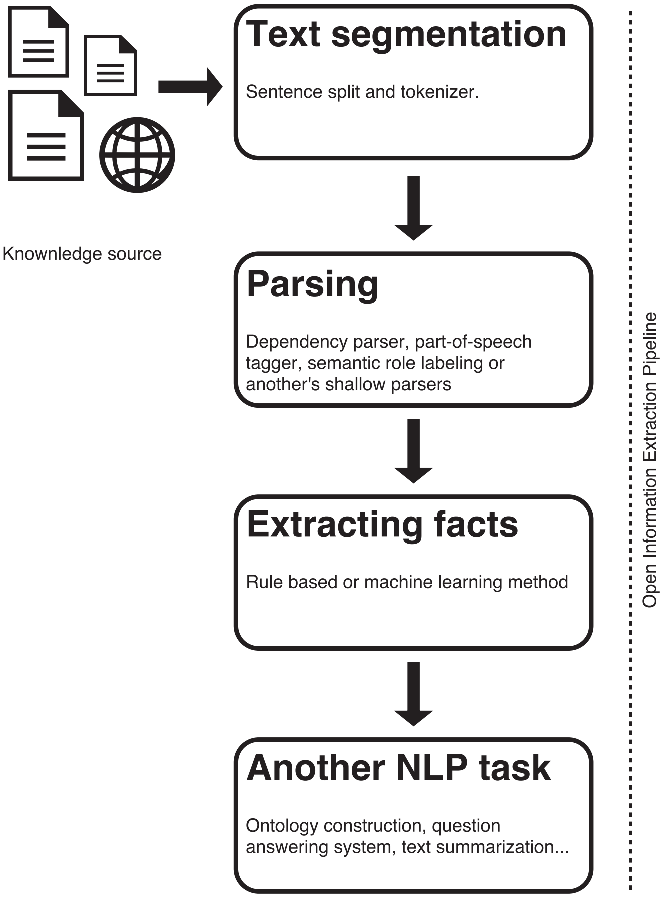

Knowledge source
Text segmentation
Sentence split and tokenizer.
Parsing
Dependency parser, part-of-speech tagger, semantic role labeling or another's shallow parsers
Extracting facts
Rule based or machine learning method
Another NLP task
Ontology construction, question answering system, text summarization...
Open Information Extraction Pipeline

Fig. 2. Open information extraction pipeline.

An Open IE system is part of a pipeline of different natural language process (NLP) applications such as QA systems and inference (Angeli, Premkumar, & Manning, 2015). Fig. 2 presents a general pipeline for Open IE task.

Considering the sentence *Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building.* obtained from *ClausIE* (Del Corro & Gemulla, 2013) an Open IE system that extract triples as follows:[2]

- (Bell, is, a telecommunication company)
- (Bell, is based, in Los Angeles)
- (Bell, makes, electronic products)
- ...
- (Bell, distributes, building products)

In Open IE systems, the most common context is limited to a single sentence. Their arguments are fragments found in the pro-

[1] See CoreNLP demo site in http://nlp.stanford.edu:8080/corenlp/.

[2] See ClausIE demo site in https://gate.d5.mpi-inf.mpg.de/ClausIEGate/ClausIEGate?inputtext=.

cessed text, and verbs establish their relations. Works such as *TEXTRUNNER* and *REVERB* extract only facts from sentences with verb. Systems such as *OLLIE* and *ClausIE* make non-verb-mediated extractions. Considering the sentence *Obama, the president of the US* retrieved from (Schmitz, Bart, Soderland, & Etzioni, 2012), recent systems can extract triple as follows:

- (Obama, is, the president of the US)

The synthetic relationship improves the results in noun sentences, but have limitations in a small set of verbs (usually, *to be* or *to have*).

There are two aspects of the facts extracted by an Open IE system. The first one is the coherence between the extracted fact and the original sentence. Some experts in the text domain can be queried to establish this agreement. In Del Corro and Gemulla (2013) two experts arbitrate the comparison between an extracted fact and the original sentence. The second aspect is discussed by Bast and Haussmann (2014) and it concerns the informativeness of an extracted fact. For instance, consider our previous sentence *Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building. REVERB* system extracts the fact:

- (a telecommunication company, is based, in Los Angeles)

Although this fact does not present useful information when compared with the original sentence, it is coherent. Open IE systems do retrieve many facts without useful information. On the other hand, traditional IE systems, where relationships are retrieved based on a manual set, achieve more accurate results. Traditional IE systems are hard to use in a large corpus due to the limited and predefined set of relationships (Fader, Soderland, & Etzioni, 2011). Moreover, it is hard to increase or change such limited set of relationships (Fader, Soderland, & Etzioni, 2011). This tradeoff is a problem for Open IE systems which have three key goals (Del Corro & Gemulla, 2013): (i) domain independence, (ii) unsupervised extraction, and (iii) scalability to plenty of texts. In this case, we can add a fourth goal: (iv) informativeness and coherent facts.

3. Related work

Despite a systematic mapping study (SMS) and a systematic literature review (SLR) share some common procedures, they are different regarding goals and data analysis (Kitchenham & Charters, 2007; Kitchenham, Budgen, & Brereton, 2010; Petersen et al., 2015). As a consequence, the whole process suffers some distinctions to achieve a specific goal. Different characteristics of an SRL and an SMS process can be observed in research questions, search processes, search strategies, quality evaluations, and results (Kitchenham, Budgen, & Brereton, 2010). However, an important feature for classifying studies as SMS or SLR is the specificity of research questions. In an SMS process, research questions are more general and results are different observed and returned. A mapping study provides an overview of the scope of the area and allows to discover research gaps, forums, relevant authors, research groups, and trends (Petersen et al., 2015). SMS is important to understand how the area is structured, common practices in existing works and open problems.

Many researchers on a number of areas uses different guidelines and methods to sample their systematic mapping studying, such as (Dusse et al., 2016; Enríquez, Domínguez-Mayo, Escalona, Ross, & Staples, 2017). Dusse et al. (2016) proposes an SMS and follow the guidelines of Kitchenham and Charters (2007) and Petersen et al. (2015) to conduct the mapping study of their area. They propose an SMS to analyze the available information visualization tools and their applications in emergency management activities. Different from our approach, we conducted the SMS to the Open IE area. On the other hand, Enríquez, Domínguez-Mayo, Escalona, Ross, and Staples (2017) conducted an SMS following the guidelines from Kitchenham and Charters (2007). They propose to analyze works whose area are entity reconciliation in the context of Big Data. Different from our approach, our SMS follows the recommendations described in Petersen et al. (2015).

4. Systematic mapping process

SMS is organized into three groups of activities: planning, conducting and reporting (Petersen et al., 2015). The first group aims to identify the reasons for this study, followed by the research questions and then definitions and test of the protocol. The second group of activities organizes the selection of primary studies, the extraction, and the data summarization. Finally, the third group defines the threats during the study activities. Our timeline is presented in Fig. 3.

Open IE is a new research topic with the first published work in 2007. Although some secondary studies have been published, SMS discovers quantitative and qualitative data on primary studies not yet presented in those secondary studies. While a traditional review can present the bias of a group or researcher, SMS aims to determine the gaps and to observe relevant aspects of the area diminishing (or eliminating) the bias vision. Our study begins with a general question about the state of the art in Open IE:

- **Main Research Question (MRQ):** What is the state-of-the-art of Open Information Extraction?

We consider this MRQ as in-depth, and we do not have the ambition to provide all information to answer it. However, we have established a set of secondary questions to help the identification of relevant aspects of Open IE. The set of RQ is the baseline of the mapping process, and each RQ is defined as follows:

- **RQ1:** What are the synonym terms that define the Open IE area?
- **RQ2:** What are the sources of publications in Open IE area?
- **RQ3:** What are the types of contributions made by Open IE studies?
- **RQ4:** What are the Open IE systems?
- **RQ5:** How are Open IE systems used?
- **RQ6:** How are Open IE systems evaluated?
- **RQ7:** What are the tools used in Open IE systems?
- **RQ8:** What are the open problems in Open IE area?

The next step in planning is to determine the search engine to retrieve the primary studies. The search method to find primary studies is carried by automatically search in electronic databases. Keywords are also a critical task in our SMS. There is a tradeoff between using specific query strings that eliminate relevant primary studies and using non-specific query strings that are too massive effort to filter out studies from other areas. Thus we elaborate preliminary versions of the protocol with non-specific keywords and then start polishing our protocol. Beyond the selection studies, we identified the related terms. The term “information extraction” was avoided because of a large number of recovery results. This term is employed in studies such as named entity recognition, coreference resolution, and sentiment analysis. Other terms such as “data extraction” or “term extraction” were not suitable to find Open IE studies. Moreover, terms were either searched in *Semantic Scholar*[3] to identify semantic similar terms; however no different term was found. Thus, we attain four relevant keywords which were used to retrieve primary studies on Open IE:

[3] https://www.semanticscholar.org/.

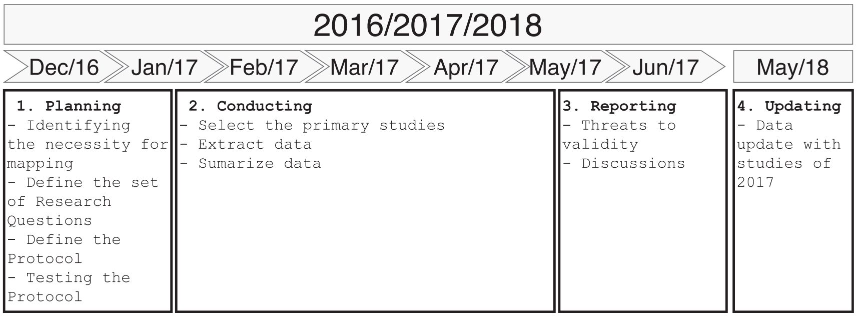


Fig. 3. The timeline of our systematic mapping process.

Table 1
String query used per database.

Database	#Article	Query
Science Direct	47	pub-date > 2010 and TITLE - ABSTR - KEY("open information extraction") or TITLE - ABSTR - KEY("open relation extraction") or TITLE - ABSTR - KEY("relation extraction") or TITLE - ABSTR - KEY("relation discovery")
IEEE Xplore	233	(((("open information extraction") OR "open relation extraction") OR "relation extraction") OR "relation discovery") and refined by Year: 2011 - 2017
ACM Digital Library	164	{"query": { ("open information extraction" "open relation extraction" "relation extraction" "relation discovery") } "filter": {"publicationYear":{ "gte":2010 }}, {owners.owner = HOSTED}}
Scopus	226	TITLE - ABS - KEY ("open information extraction" OR "open relation extraction" OR "relation extraction" OR "relation discovery") AND DOCTYPE (ar) AND PUBYEAR > 2010 AND (LIMIT - TO (LANGUAGE , "English"))
Google Scholar	1813	"open information extraction" OR "open relation extraction" custom range 2011 - present

- *"open information extraction"* OR
- *"open relation extraction"* OR
- *"relation extraction"* OR
- *"relation discovery"*

We performed our search into five databases: *Science Direct*,[4] *IEEE Xplore*,[5] *ACM Digital Library*,[6] *Scopus*,[7] and *Google Scholar*[8] with the search strings detailed in Table 1. In *Google Scholar* we do not employ the terms "relation extraction" nor "relation discovery" because of the large number of results (more than 9k entries).

Our *inclusion criteria* aims to recover high-quality studies that have a recent impact on this research area. Queries were executed on the databases at February 2017, May 2018 with studies published till the end of 2017 (data update). Each database allows different search configurations. However, we are all considering that the keywords must appear in at least one of the article title, abstract or keywords fields. The *exclusion criteria* (F–filters) for primary studies are:

- **F1:** Remove non-English written paper.
- **F2:** Remove paper not published in journals or conferences.
- **F3:** Remove short paper.
- **F4:** Remove survey or review paper.
- **F5:** Remove paper that have some "Open IE" terms, but are not studies on the topic.
- **Duplicated:** Remove one of the duplicate occurrences.

We choose to remove the studies written in other languages than English due to English written papers have a large audience in an academic community. Texts such as MSc or Ph.D. reports, technical reports, or any content which were not evaluated by a program committee were either removed. Partial results published in *short paper* format was either removed. The review studies (secondary studies) were also removed. An important observation is about some studies that use the terms Open IE, but do not deal with open systems. Almost all of them perform traditional IE, with some insights about Open IE systems. The F5 filter removed this kind of study. After executing the string query (Table. 1), the filtering step starts.

The removal of primary studies was conducted in two stages. In the first stage, we read the abstracts of each paper to identify occurrences outside the scope of our SMS. In the second stage, with the remaining papers, we filtered the occurrences within a paper full reading. Fig. 4 presents the values of each filter applied to each database. After performing the filter step, 73 primary studies were selected. The complete list of selected primary studies is in https://www.bibsonomy.org/user/sysmapopenie. For each primary study, we try to identify the contributions.

Despite we have identified duplicate studies, before removing them, they were summarized per database. Fig. 5 depicts the percentage of studies per database of all retrieved studies including the duplicate ones. It is noteworthy that *Scopus* and *Google Scholar* are the most representative databases in our work. After this summarization, all duplicate occurrences were removed by the last filter step, gathering a total of 73 papers.

We organize the primary studies into categories per contributions. We hypothesized that a categorization of these studies might reveal some useful information about Open IE. Our list of contribution categories are as follows:

- **Tool:** Studies that present a tool that performs an Open IE task. In this category, we consider studies that present a tool by its name or its alias. We search for tools that publicly make available source or binary files.
- **Resource:** Studies that have created datasets for evaluation or other resources for training or testing of Open IE systems.

[4] http://www.sciencedirect.com.
[5] http://ieeexplore.ieee.org/Xplore/home.jsp.
[6] http://dl.acm.org.
[7] http://www.scopus.com.
[8] http://scholar.google.com.

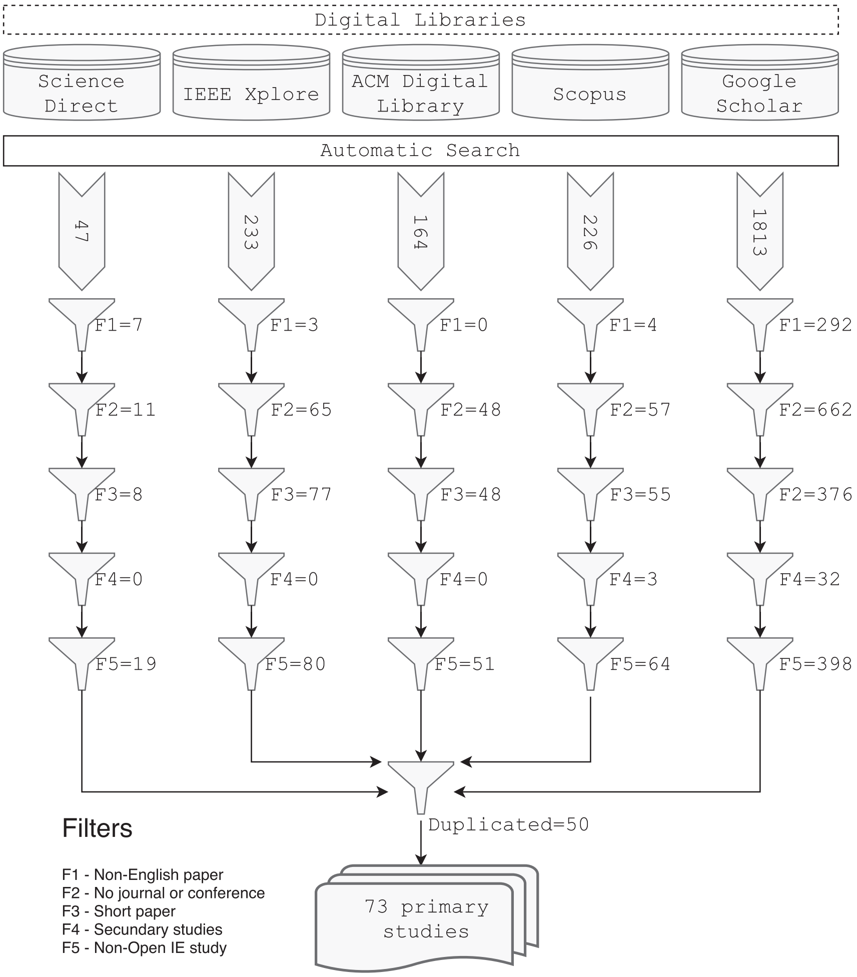


Fig. 4. Distribution of the primary studies per digital database and the set of exclusion criteria values (F1.F5 and duplicated).

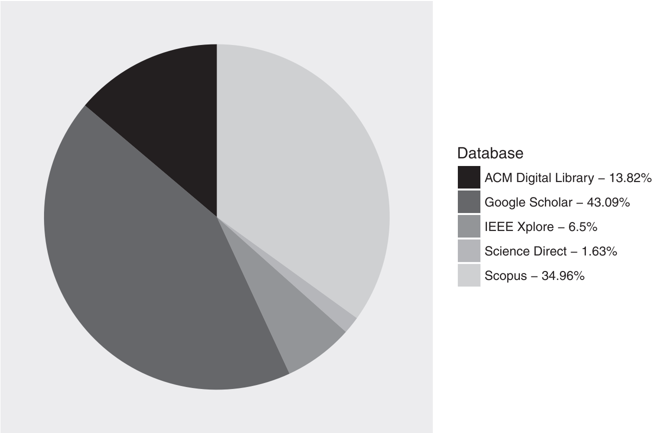


Fig. 5. Percentage of each database in primary studies selection.

- **Method:** Studies that propose new methods or approaches to construct Open IE systems.
- **Application:** Studies that use Open IE systems for some natural language processing task.
- **Validation:** Studies that evaluate the results in different Open IE systems.
- **Evaluation:** Studies that contribute with new methods or evaluation measures for Open IE systems.

The selected primary studies provide a large set of information about Open IE. In the conducting step of our SMS, we create a form to register a large set of information about those primary studies. We consider different types of information about the selected studies. The first one concerns the researchers, research groups, and countries from affiliation. Secondly, the information about datasets such as language, domain, size, and whether they are publicly accessible. We also collect information about the evaluation measures and other Open IE systems used to compare each approach. Moreover, we gather the NLP tasks employed and any additional information from the primary studies. Table 2 presents the form filled in our SMS within each RQ. There is no field for RQ8 because the open problems were obtained by the author's annotations or the list of future works. After reading all studies, it was possible to conclude which problems have not yet been solved. Afterwards, we initiated the analysis phase with all those data.

5. Results

After obtaining the data from each selected primary study and filling the form (Table 2), the next step was to summarize the data to recover useful information. The objective of this step was to answer the RQ set. We organize the presentation of the summarized data according to RQ order. Before starting with our questioning, we present a word cloud using the full text of our 73 selected studies (Fig. 6). The font size indicates the frequency in which the term is used in all primary studies. We choose to use all the text of each article, and we observe terms with different origins. Terms

Table 2
Data extraction form used by our SMS.

Data item	Value	RQ
General study ID	Integer	
Article title	Name of the article	
Author list	Author's name list	
Year of publication	Calendar year	
Research center	Authors affiliation	RQ2
Country	Country of the headquarters of the research centre	RQ2
Keyword	Keyword study indexing	RQ1
Keyword class	Our keyword category	RQ1
Source	Source of publication: conference or journal	RQ2
System name	Open IE system name when the authors define	RQ4
Dataset visibility	Public or private	RQ6
Dataset language	English, Chinese, Portuguese...	RQ6
Dataset source	Corpus name used to create the dataset	RQ6
Dataset format	Sentence, document, triple or question and answering match	RQ6
Dataset domain	Domain of the Corpus	RQ6
Measure	Evaluation measures used in the study	RQ6
Against system	The system's name used in comparison with other proposal	RQ6
Contribution type	Tool, Resource, Method, Application, Validation or Evaluation	RQ3
NLP task	NLP tasks used in the proposed study	RQ7
NLP tool	NLP tools used in the proposed study	RQ7
Other tool	Other tools used in the proposed study	RQ7
Extract method	Training data or handcrafted rules based	RQ7
Ontology	Ontology used in the proposal (if applied)	RQ7
Application	Construction of ontology, text summarization...	RQ5

Fig. 6. Word cloud from the full text of the 73 selected primary studies in our SMS.

such as *Precision* and *Recall* are related to the evaluation of selected studies. *TEXTRUNNER, OLLIE, REVERB* and *ClausIE* are Open IE systems' names.

5.1. Reply to RQ1: What are the synonym terms that define the Open IE area?

The term "open relation extraction" (Open RE) is also used in papers which present some contribution for Open IE task. Since the earliest study published in the area (Banko, Etzioni, & Center, 2008) "Open RE" is also used as a synonym for Open IE. In Petroni, Del Corro, and Gemulla (2015) the authors confirm both usages as synonyms; however, the researchers present an important difference among both tasks. Open IE extracts facts found in texts in a purely syntactic way using as an argument parts of the text. Open RE is the task of extracting new facts from knowledge bases or even raw texts. Knowledge bases cited by the authors can be constructed by Open IE systems. In synthesis, an Open RE extracts new facts learned from an open set of relations.

Another task called "open relation mapping" (Open RM) is defined by authors in Liu et al. (2013). Authors define this task as a mining of relations obtained by Open IE systems. Using clustering algorithm, semantic relations groups are created using facts extracted from an Open IE system. For example, when extracting facts like "Obama, was born in, Honolulu" or "Obama, flies back to, Honolulu" the mapping process manages to establish a link between these two facts. Semantic relation groups allow, for example, QA systems to match with any doubts in different formats (*e.g.* "Where is the hometown of Obama?", "Where was Obama born?").

Among the 73 studies, we found another term: "open knowledge acquisition" (Open KA). According to Kim and Myaeng (2016), "knowledge acquisition is an activity or a process of building or extracting knowledge from other sources for a knowledge base, and can be done either manually or automatically.". We believe that the concept of Open KA is similar to Open RE presented by Petroni et al. (2015). We do not find more terms related to Open IE in the selected primary studies. Among the selected studies, only authors in Petroni et al. (2015) discuss the application of the terms "Open IE" and "Open RE" as synonyms.

After identifying the differences (or similarities) among the four terms, we compared each term used to retrieve the study with the term inside the paper. This comparison is presented in Fig. 7. First, there is a higher concentration in the use of the "Open IE" term which can indicate it as a general term. On the other hand, even studies retrieved by the term "Open RE", most of the authors use the term "Open IE" inside the text. Although we believe that the term "Open IE" can cover this research area, this kind of branches are not completely clear yet.

Fig. 8 presents a hierarchy of the four terms identified in this SMS. Our classification is defined considering that the three tasks (Open RE, Open KA, and Open RM) extract information as an Open IE. These tasks are applications at a later stage in NLP pipeline.

5.2. Reply to RQ2: What are the sources of publications in Open IE area?

Our second RQ retrieves information from the primary studies about research groups, countries and the evolution of the area during last years. First, we present the number of published studies over the past years. Fig. 9 depicts growth in the number of publications except between the years 2012, 2015 and 2017. Comparing the extreme values (2011 and 2017), we observe that the number of publications grew over 100%. It is worth noting that the most of

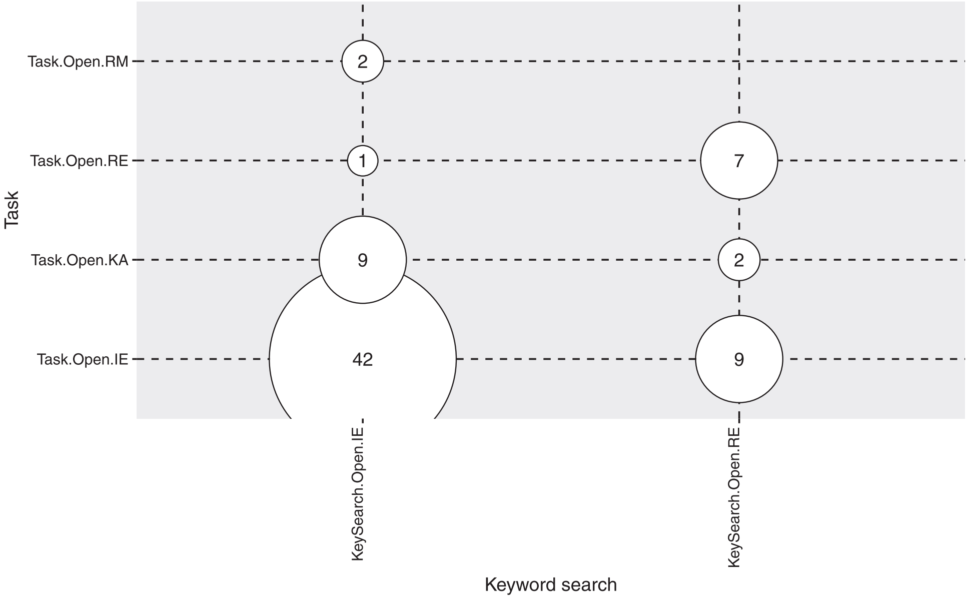


Fig. 7. Comparison among the terms used to retrieve studies and the terms employed inside the paper.

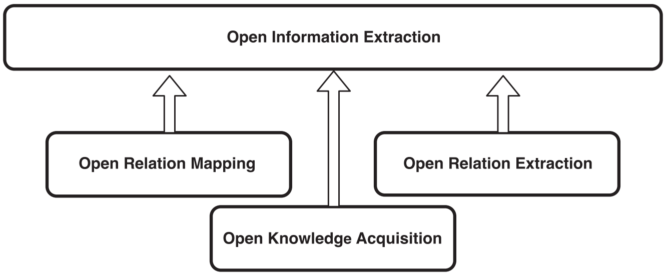


Fig. 8. Four terms identified among the 73 primary studies and the hierarchy among them.

those primary studies were published at conferences. This fact can confirm that research on Open IE is quite young.

Studies on Open IE have a higher number of published papers at the conference *Empirical Methods on Natural Language Processing* (EMNLP) with eleven occurrences in Fig. 10. Open IE area is multidisciplinary, and the distribution of publications in conferences on data mining, information retrieval, NLP, and ontology confirms this information.

From the affiliation of each paper, the authors' name, research group and country headquarters of the institution are collected. Fig. 11 depicts the country with the highest number of participating authors among the 73 selected studies. The country with the highest number of publications is the USA where this research area (Open IE) has emerged (Banko et al., 2007).

Fig. 12 presents the organizations and the number of studies in this area. The research group with the highest number of publications is the *KnowItAll* at the University of Washington. This research group was responsible for introducing this research area (Banko et al., 2007). The MPI-INF research institute deserves prominence due to the number of published studies in Open IE. Although headquartered in Germany has a significant impact on Open IE for English texts, it is important to mention the *ClausIE* system (Del Corro & Gemulla, 2013) due to its contribution to this area.

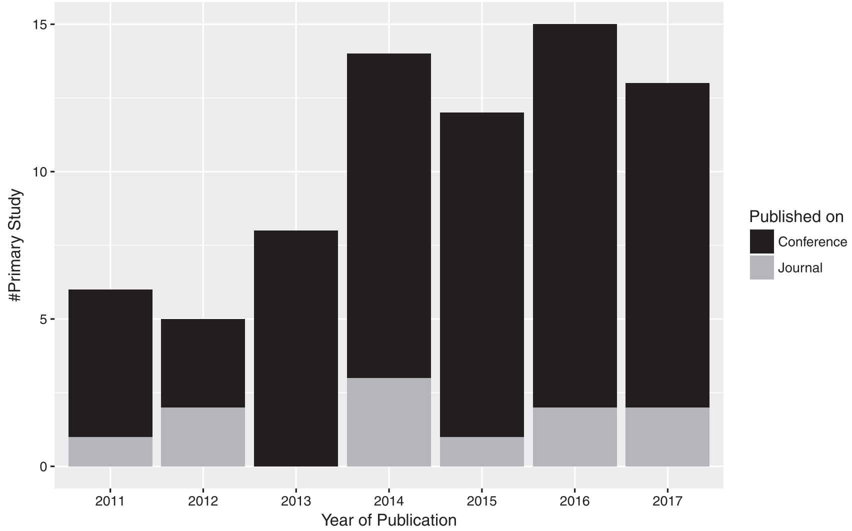


Fig. 9. Distribution over the years of selected primary studies per paper type (journal and conference).

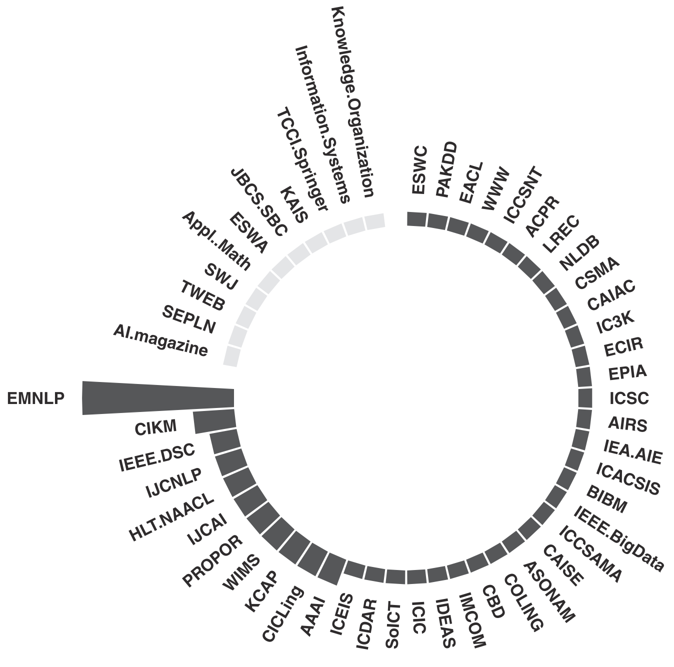


Fig. 10. Distribution of primary studies between selected journal (smallest group) and conferences (largest group) in our systematic mapping. EMNLP conference has eleven primary study entries and journals have only one entry each.

5.3. Reply to RQ3: What are the type of contributions made by Open IE studies?

Through Fig. 13 it is observed two major contributions. The *METHOD* group corresponds to studies with novel methods or approaches to the task of “Open IE”. This kind of research tries to evolve the state of the art with new ideas comparing its work with other systems, in spite of Open IE is a young research area that lacks benchmark materials. Some studies also produce contributions in *RESOURCE* and *EVALUATION*. There is also a relevant portion of studies that present a new approach to Open IE through a *TOOL*. In these studies, authors present, compare and provide an implementation of their work. These studies are relevant because new studies can evaluate their results by reducing the risk of introducing errors in the experimental phase when dealing with new implementations. The *VALIDATION* subgroup is a special research case in which a methodology to compare the results of different studies is proposed. This kind of research is still incipient, faced with the difficulty to build the resources for an evaluation method that covers different domains of text, data volume, or even multiple languages.

The second group of contributions is *APPLICATION*. This study group applies an Open IE system on some NLP task. The intersection between *APPLICATION* and *METHOD* groups represent the studies that propose new approaches and concludes the research with its application. Another part of this group are direct applications of other methods such as *ClausIE, OLLIE*, and *REVERB*. We observe that the exclusive application subgroup has only nine studies. Namely, only 15% of them deals exclusively with the application of Open IE systems which is a small number. Few applications is another argument that reinforces the little maturity of this young research area.

5.4. Reply to RQ4: What are the Open IE systems?

In Fig. 13, we present the groups of contributions identified from the primary studies. The *TOOL* group recognizes the methods that have been appointed and makes available to the community. Named Open IE tools are presented in Table 3. We organize this presentation by year in descending order.

An important aspect of any NLP tool is the input language. In this study, we took care of retrieving the input language for the Open IE system. It is noticeble that English texts are the largest amount of systems followed by some Chinese systems. There are two multilingual systems *DepOE +* and *ArgOE* proposed by a research group at the University of Santiago de Compostela.

5.5. Reply to RQ5: How are Open IE systems used?

The two most common applications are the construction of ontologies and improve QA (Querying Answering) systems. In Fig. 14, other applications are observed such as: Document filtering, summarize, and clustering in text mining applications. These applications can improve their results by the use of extracted facts from Open IE systems. Another Open IE usage is linked data proposed by Perera and Nand (2015). A particular case of application is to use extracted facts from Open IE system as input to another IE system. For example, traditional IE system described

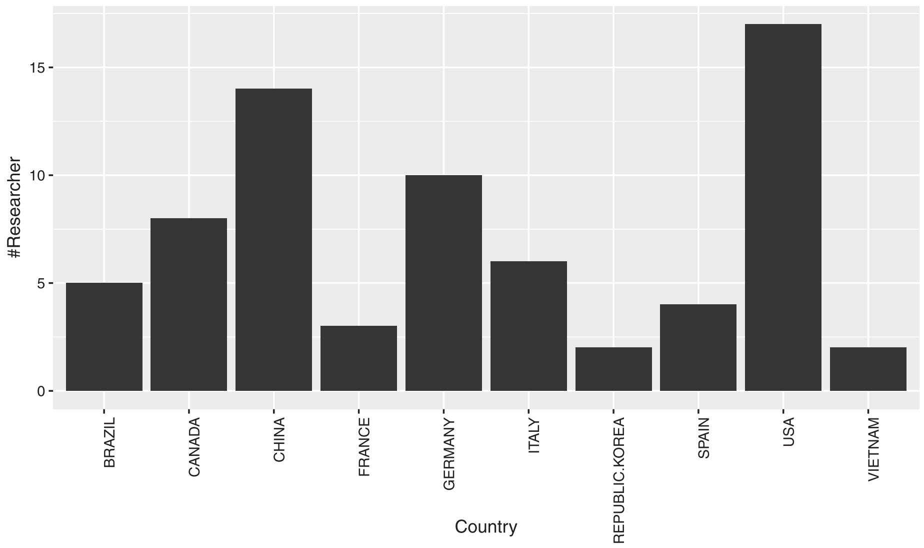


Fig. 11. Distribution of the countries of the researchers in primary studies.

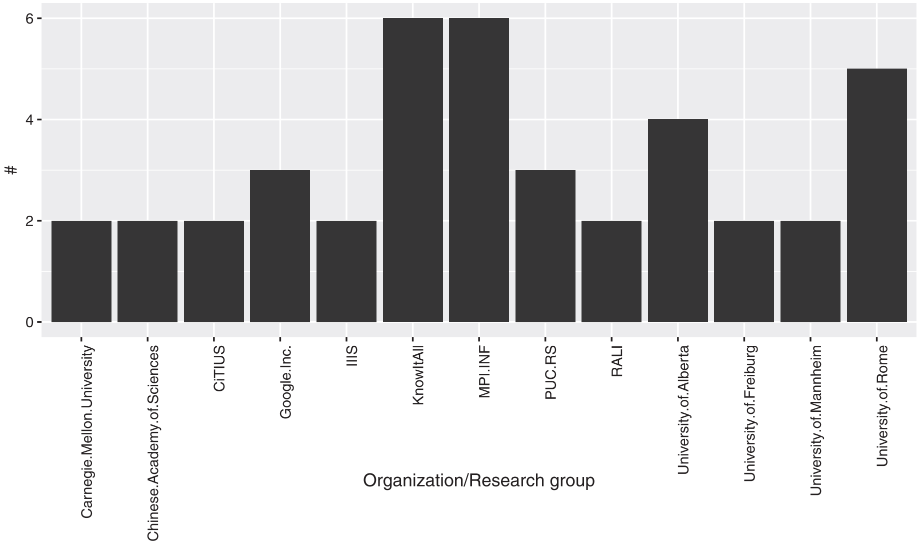


Fig. 12. Distribution of the country affiliation authors in primary studies.

Table 3
List of named Open IE tools retrieved from selected primary studies in the SMS.

System	Paper	Year	Language
MinIE	Gashteovski, Gemulla, and Del Corro (2017)	2017	English
RelP	de Abreu and Vieira (2017)	2017	Portuguese
C-COERE	Wu and Wu (2017)	2017	Chinese
vnOIE	Truong, Vo, and Nguyen (2017)	2017	Vietnamese
R-OpenIE	Lin et al. (2016)	2016	English
SemIE	Tan et al. (2016)	2016	English
REALTEXT	Perera and Nand (2015)	2015	English
ClausORE	Xu, Gan, Deng, Wang, and Yan (2015)	2015	Chinese
CORE	Petroni et al. (2015)	2015	English
GCORE	Wang, Zhou, Tian, Nan, and Ma (2015)	2015	Chinese
ArgOE	Gamallo and Garcia (2015)	2015	Multilingual
LSOE	Xavier et al. (2015)	2015	English
Stanford OpenIE	Angeli et al. (2015)	2015	English
WEBCHILD	Tandon, De Melo, and Weikum (2014)	2014	English
CORE	Tseng et al. (2014)	2014	Chinese
DepOE +	Garcia and Gamallo (2014)	2014	Multilingual
LEGALO	Presutti, Nuzzolese, Consoli, Gangemi, and Reforgiato Recupero (2014)	2014	English
ReNoun	Yahya et al. (2014)	2014	English
AWAKE	Boschee et al. (2014)	2014	English
ZORE	Qiu and Zhang (2014)	2014	Chinese
ClausIE	Del Corro and Gemulla (2013)	2013	English
EXEMPLAR	Mesquita, Schmidek, and Barbosa (2013)	2013	English
CSD-IE	Bast and Haussmann (2013)	2013	English
SONEX	Merhav, Mesquita, Barbosa, Yee, and Frieder (2012)	2012	English
OLLIE	Schmitz, Bart, Soderland, and Etzioni (2012)	2012	English
OntExt	Mohamed, Hruschka Jr, and Mitchell (2011)	2011	English
REVERB	Fader, Soderland, and Etzioni (2011)	2011	English
RDROIE	Kim, Compton, and Kim (2011)	2011	English

in Liu, Ling, An, and Hu (2014) using *SEMREP* data to accomplish its task. *OLLIE* system uses high confidence facts extracted by *REVERB*.

We have recovered the list of Open IE systems employees in application studies and, we present in Fig. 15. We consider *NELL* project (Betteridge et al., 2009) which uses a semi-supervised method as an Open IE system. *REVERB* method is the most popular method among the primary studies selected in our SMS. It is easy to understand, and it was published in 2011. Although it performs for English language, other languages such as Portuguese (Sena, Glauber, & Claro, 2017) and French (Gotti & Langlais, 2016) still holds either attention.

5.6. Reply to RQ6: How are Open IE systems evaluated?

Since the first studies, Open IE systems evaluate their efficiency by the use of information retrieval measures. Precision, recall, and f-measure are the most popular measures applied to this area as demonstrated in Fig. 16. The evaluation of a new proposal or idea might present their results comparing to the state of the art. Fig. 16 also presents some primary study that does not compare their proposal with any study (see NONE values in *X*-axis). There is a greater number of studies that use *Precision* and *NONE*. This choice corresponds to almost 50% of the studies. However such results could be superficial even worse when the work is not com-

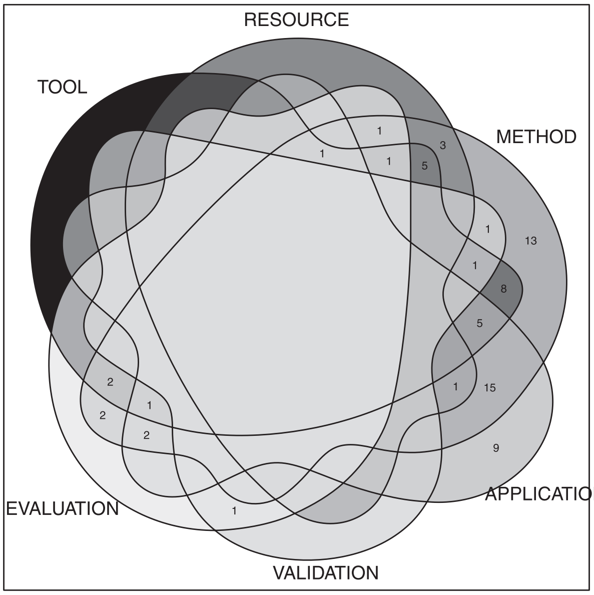


Fig. 13. Number of primary studies by types of contributions and the intersections among contribution categories.

pared with any state of the art. On the other hand, the difficulty of identifying studies in languages other than English persists. We observe that it is usual for authors to evaluate their proposals with more than one measure as well as information retrieval systems.

There is an important aspect of the evaluation procedure: Facts are extracted without informativeness (Mesquita, Schmidek, & Barbosa, 2013). In this paper, we select an efficiency evaluation proposed by Léchelle and Langlais (2016) that uses a question and answering system as an evaluation mechanism (indirect mode - “QA measure”). Researchers made available the material used and presented efficient results comparing to some state-of-the-art systems.

In addition to a higher concentration of studies that present Open IE tools for English texts (Fig. 17), we also identify a larger quantity of datasets in English texts. Most of the datasets were created with sentences from: Wikipedia ,[9] New York Times [10] and ClueWeb09 .[11]

We observe that most primary studies prefer not to publish their datasets as shown in Fig. 18. For a public visibility of a dataset, it is necessary to make it available for open access (preferably on a website). The lack of this type of resource forces researchers to construct their own datasets. Without public datasets, it is hard to fairly compare related works and moreover to advance the state of the art. Despite the fact that different domains and dataset languages strengths the generalization of evaluation, the lack of public datasets states the immaturity of this domain.

5.7. Reply to RQ7: What are the tools used in Open IE systems?

The RQ7 specifies which tools are used in Open IE systems. Based on the selected primary studies, ten taggers for sentences were identified. In Fader, Soderland, and Etzioni (2011) part-of-speech (POS tagger) and noun phrase chunking (Chunking) are used. Authors in Schmidek and Barbosa (2014) adds dependency parsing (DP). Different methods of Open IE make the combination of taggers with shallow or dependency analyzer. We also find other shallow analyzers with less frequency, such as: semantic role labeling (SRL) (Christensen, Soderland, Etzioni et al., 2011), sentence constituent identification (SCI) (Bast & Haussmann, 2013) and structured string-tree correspondence (SSTC) (Tan, Lim, Soon, & Tang, 2016). Primary studies such as *ClausIE* and *Stanford Open IE* (Angeli et al., 2015) uses taggers like POS tagger, but their proposals are based on DP for clauses identification (useful parts of a sentence). Therefore, we consider this type as a DP-based method according to the classification proposed by Gamallo (2014). In Fig. 19 POS tagger and DP analyzers are the most frequent in combination with the rule-based method. The rule-based method overpowers Open IE approaches since *REVERB* system which assigns methods that use machine learning limitations to the quantity of extracted facts.

There is a large concentration of studies using the Stanford CoreNLP and Apache Open NLP[12] tools as presented in Fig. 20. We

[9] https://dumps.wikimedia.org/.
[10] https://catalog.ldc.upenn.edu/ldc2008t19.
[11] https://lemurproject.org/clueweb09/.
[12] https://opennlp.apache.org/.

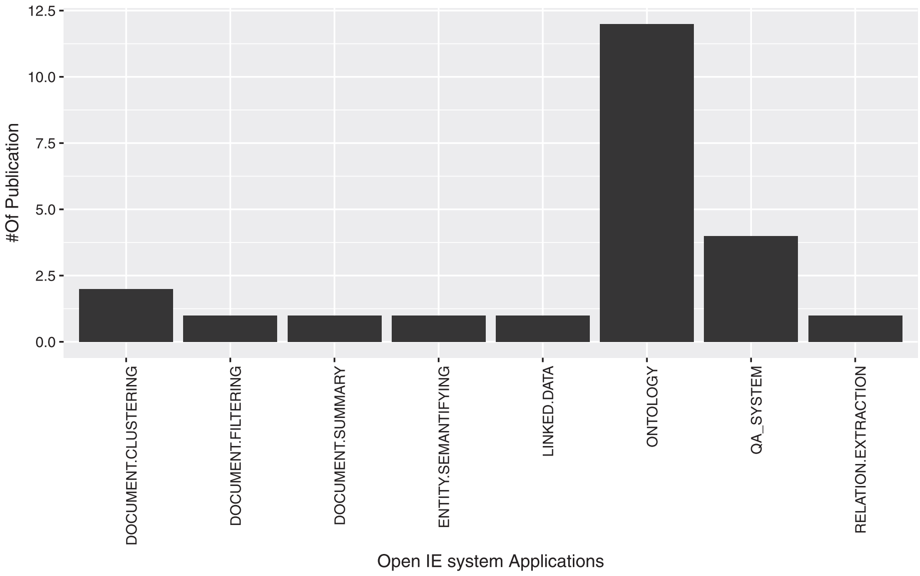


Fig. 14. Number of primary studies per type of applications.

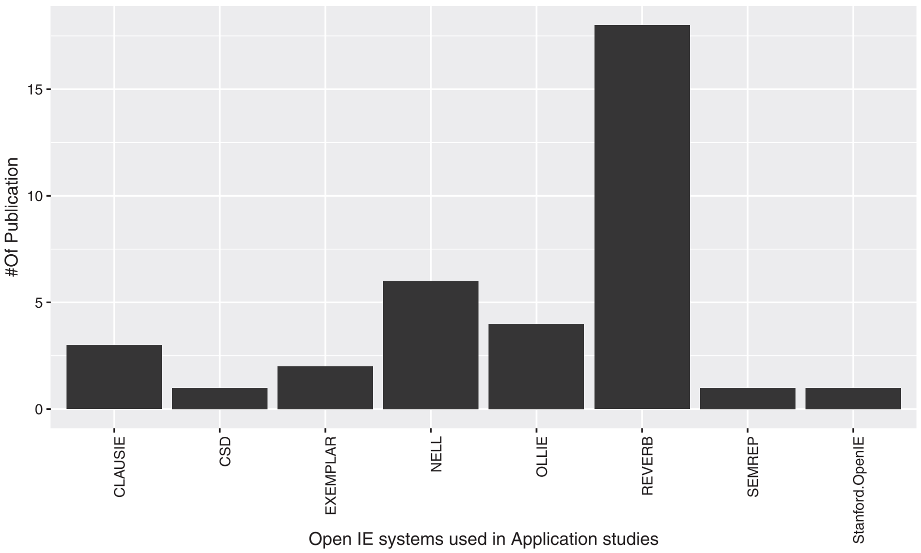


Fig. 15. Distribution of Open IE system application from primary studies.

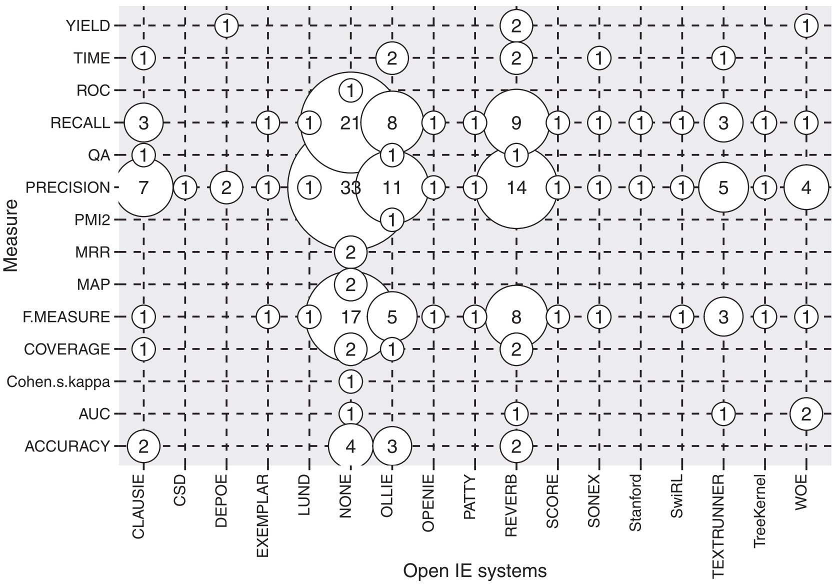


Fig. 16. Map of the evaluation measures from primary studies. For *X* axis the systems. For *Y* axis the measures in evaluation step. *NONE* are occurrences of works that are not compared to any study.

believe that support for multiple languages, multiple tasks of NLP and its respective maintainers guarantee these two projects greater popularity. Other projects which also support multiple languages and multiple tasks such as NLTK[13] is not so popular. In this case, there is a suspicion that version changes in the Python language and an extensive period of candidate versions can justify lesser popularity of this project related to the Open IE area. In this map, specific NLP tools are also localized to a single language, for example, Portuguese with CoGrOO[14], PALAVRAS[15] parser and LemmPORT[16] or HanLP[17] that has support for other languages, but offers models for Chinese.

NLP area has used knowledge bases to enhance tools and approaches to tasks such as word sense disambiguation (WSD) and named-entity recognition (NER) (Della Rocca, Senatore, & Loia, 2017). When the ontology/knowledge models were retrieved from primary studies, we observed different applications from the disambiguation of extracted facts (Dutta & Schuhmacher, 2014) to the construction of new knowledge bases from other knowledge bases (Kim & Myaeng, 2016). Fig. 21 presents the identified knowledge bases and the relationship with the type of contribution of primary studies. DBPEDIA,[18] FREEBASE,[19] YAGO[20] and WordNet[21] are the most popular and concentrated in primary studies for English texts.

[13] https://www.nltk.org/.

[14] http://cogroo.sourceforge.net/download/current.html.

[15] http://visl.sdu.dk/visl/pt/.

[16] https://github.com/rikarudo/LemPORT.

[17] https://github.com/hankcs/HanLP.

[18] https://wiki.dbpedia.org/.

[19] https://developers.google.com/freebase/.

[20] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

[21] https://wordnet.princeton.edu/.

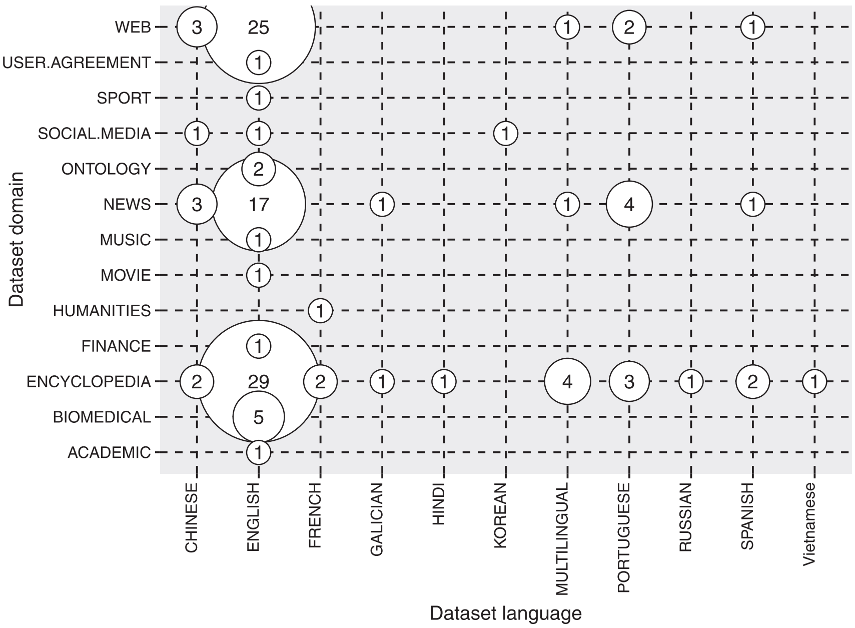


Fig. 17. Map of the dataset properties from primary studies. For *X* axis the language in which the text was written. For *Y* axis the domain of the text.

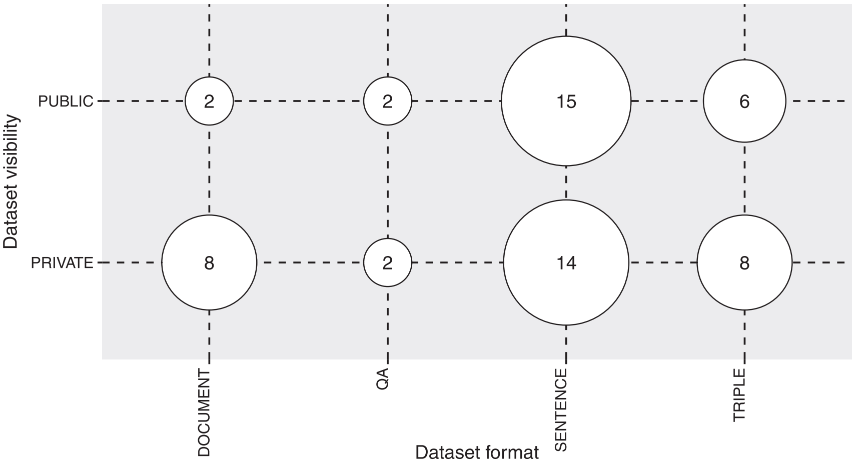


Fig. 18. Map of the dataset format and visibility from primary studies. For *X* axis the format file of the dataset. For *Y* axis the visibility of datasets.

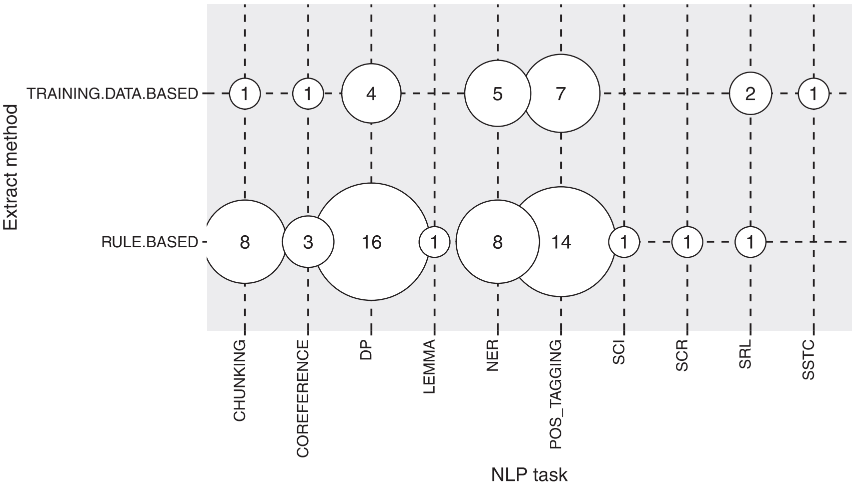


Fig. 19. Map of natural language processing tasks and extraction method approach applied in Open IE studies. For *X* axis the NLP taggers used. For the *Y* axis the extraction method approach.

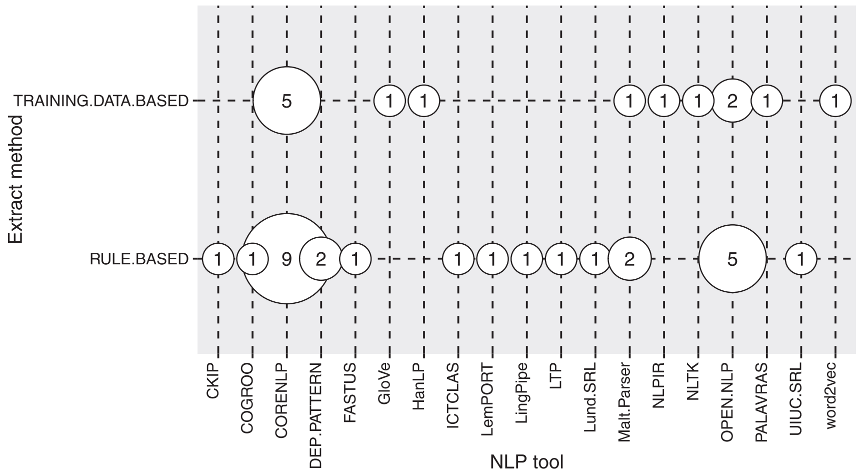


Fig. 20. Map of natural language processing tools and extraction methods applied in Open IE studies. For *X* axis the NLP tools used. For Y axis the extraction method approach.

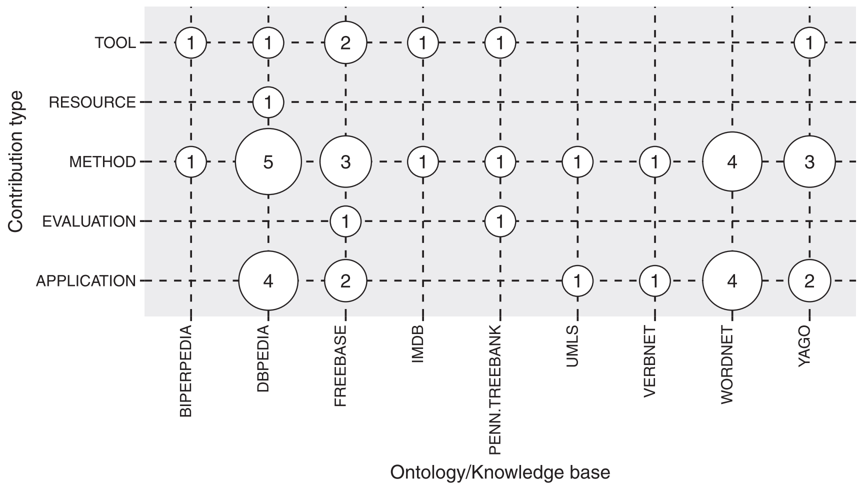


Fig. 21. Map of knowledge bases and types of contribution identified in Open IE studies. For *X* axis the ontology/knowledge bases. For Y axis the type of contributions.

5.8. Reply to RQ8: What are the open problems in Open IE area?

This study did not define fields for our data form to respond to RQ8. The open problems presented in this section deals with issues gathered during the full reading of our 73 selected studies. We do not intend to address particular problems, for instance, NLP tools errors discussed by Etzioni, Fader, Christensen, Soderland, and Mausam (2011) and Del Corro and Gemulla (2013). We are looking for general problems in Open IE area.

A) Fair evaluation among the compared studies – Fig. 13 presents a large concentration of primary studies that contribute to *RESOURCE* and *EVALUATION*. This concentration suggests a lack of standard to evaluate. Fig. 18 depicts an important number of private datasets and Fig. 16 depicts that studies are not been compared with previous ones. The maturity of a research area requires reliable methods of experimentation and comparison among the studies. Authors in Stanovsky and Dagan (2016) have created resources for support at this stage of the research. However, resources is limited to two domains and English methods. The lack of benchmarks for different domains and text languages makes difficult the comparison among studies.

B) Definition of informativeness – In Xavier et al. (2015) discuss that the concept of informativeness is differently among primary studies. Bast and Haussmann (2013) and Del Corro and Gemulla (2013) are more careful than in (Fader, Soderland, & Etzioni, 2011) about informativeness. The difference of concepts make a subjective comparison and impose a bias in results (Xavier et al., 2015). Léchelle and Langlais (2016) propose a mechanism to measure the efficiency of a QA system supplied by different Open IE. The proposal is to measure the efficiency of an Open IE system through the efficiency of a previous task in an NLP pipeline. Improving methods and measures of evaluation beyond the precision of the extracted facts is to create better parameters for improvement of new proposals.

C) One method, different languages (multilingual system) – Table 3 ArgOE (Gamallo & Garcia, 2015) and DepOE + (Garcia & Gamallo, 2014) are Open IE multilingual tools. Both tools were developed by the same research group that shares the same Open IE method category: Dependency parsing and rule-based extraction. Both proposals generalize the extraction rules with a subset of rules to extract facts from different languages. Their results are not impressive for other languages than English. We believe that a general rule is not enough to extract a large number of facts in a multilingual approach. On the other hand, Faruqui and Kumar (2015) propose the use of machine translation (MT). By using a tool that can extract facts in English (OLLIE Schmitz, Bart, Soderland, & Etzioni, 2012), the MT process the input for target language to English and finally the extracted facts to target language. This is a very abstract way of describing the method proposed in Faruqui and Kumar (2015), but we are concerned about the known problems of MT (*e.g.* lexical ambiguity, lexical gaps, and structural grammar differences). These problems can interfere on the results of Open IE systems.

D) Increasing informativeness – We select primary studies that increase informativeness of the extracted facts by coreference (Yahya, Whang, Gupta, & Halevy, 2014) and transitive inference (Bast & Haussmann, 2014). Systems such as *ClausIE* present a high rate of extracted facts with their sophisticated extraction rules but still retrieve a lot of wrong facts (see Tables 2 and 3 in (Rodríguez et al., 2016)). *REVERB* system creates a lexical constraint to eliminate wrong facts, but the use of a large annotated corpus is uninteresting for languages other than English. We can observe that many tools for Open IE and dataset for other languages than English is still scarce. Even if works such as *REVERB* are viable for other languages, it still draws a significant portion of wrong facts.

E) Decreased computational complexity – DP provides Open IE proposals with an increase in computational cost. Some systems have compared their methods with different taggers. WOE*pos* is 30x faster than WOE*parser*; however, WOE*parser* is more accurate (Wu & Weld, 2010). At this point, it is hard to balance accuracy with computational cost. Proposals such as Lin et al. (2016) have DP in an offline step to decrease computational cost and achieve a reduction in execution time. To ensure scalable systems on the web, it is necessary to continue with new research on this problem.

Open IE is a research area that introduces new studies each year and this study applies a systematic mapping method to retrieve useful information for the community. Next section, we present some threats involving the accomplishment of this work.

6. Threats of validity

Our study is about a recent research area involving NLP, IE, and Artificial Intelligence communities. This area has been applied for different domains, and although there is some revision work in the area, to the best of our knowledge, this is the first systematic mapping on Open IE. We consider our study comprehensive and with some threats of validity. In this section, we describe each threat identified to accomplish this work.

Research questions: Our study is centered on a general research question. Defining precisely the highest level of development in an area at a time is a hard task. To make this study viable, we have established a set of secondary questions. The definition of this set was discussed in numerous internal meeting in our research group and revised at different times to validate the protocol. Each research question was created based on aspects considered important to the researcher community. We believe that this set does not cover all subjects for the state of the art. Nevertheless, we believe that this problem does not hinder our study. The subjects dealt within this study are important and collaborate to new studies with useful information.

Publication bias: Our study tried to gather all available primary studies. Studies were retrieved by an automated search and the use of different databases. However, it is important to consider the absence of some important studies. First, even with simple filters, it is possible that some study has been incorrectly removed. Secondly, even using five databases some relevant studies could be not indexed within our choices. We tried to mitigate these problems with a careful review of the protocol.

Form-filling errors: We try to mitigate these problems with careful reviews of the entire execution of the protocol. In some selected studies the data was not located in the article. The experience of each researcher is different. As a consequence, the style of writing is greatly influenced by personal experiences or even by the research area. As a general rule, each research area has the same structure for writing scientific papers. However, it is common for some researchers to omit some aspects judged irrelevant. For instance, materials used in the experiments; Some authors register all the materials used during their experiments. Other authors consider it is less important, and they spend less time describing them.

Duplication of studies: A large quantity of duplicate primary studies was found. Databases as *Scopus* and *Google scholar* do the indexing by gathering other databases which generate this problem. We use the StArt tool (Fabbri et al., 2016) to manage the selection of primary studies and to automate the elimination of duplicate studies avoiding manual errors.

7. Conclusions

The goal of our SMS is to provide useful information about Open IE area. Our research questions present the conferences and journals list, research countries list, method based map, evaluation map, dataset map and others. We have also identified some open problems that have been addressed in primary studies over the recent years. The novelty of this study is that, to the best of our knowledge, this is the first systematic mapping study on Open IE area. Considering the degree of maturity of this area of research, we encourage new primary and secondary studies to raise this area of knowledge.

When we identify other published systematic mappings, we observe that they are more specific studies. In our proposal, we use this method to address an entire area of research. We could only address aspects of Open IE systems evaluation, or even about the NLP tasks employed in this kind of system. However, this area is young and still requires a high organization effort. Even if our pretension is comprehensive, questions such as RQ1 and RQ8 are not entirely exhausted. During our study, we find terms such as Open IE, RE, RM, and KA. This area needs to mature its concepts. The discussion on defining the terms Open IE and Open RE should move forward. If we consider both as synonymous terms, are we considering that no other task will be "Open"?. Soon, we will be able to define a new task as "open entity recognition"?. We like the idea that the "open information extraction" term defines a research area and "open relation extraction", "open knowledge acquisition", "open relation mapping" and others define tasks in this area.

Some aspects have been suppressed in our results by not containing discussions or related results in the selected studies in our SMS. A piece of the selected studies evaluate their proposals by the amount of extracted facts and, we observe that open systems extract a large number of different relationships. This feature hinders the evaluations and applicability of the proposals. Our results point to a small number of applications for Open IE and a concentration in the construction of new knowledge bases. While the traditional relation extraction finds applicability in various intelligent systems of different areas.

Although "informativeness" is a recurring subject in selected studies, we find that much of the studies do not restrict the concrete extracted facts. Few studies use NER annotators as a mechanism to improve informativeness and many valid facts such as {The film, were criticized, by some religious groups}[22] do not express useful information outside of its context, although the raters recognize it as a valid triple. Another critical issue is the large concentration of studies based on verb clauses that extract binary verb-based relationships, only. In Akbik and Löser (2012) proposed a method to recognize n-ary triples, but over the years few studies have dealt with this problem, and the approach of refactoring an n-ary fact in binary facts has been applied more frequently among primary studies.

[22] A triple evaluated as valid in *ClausIE* study.

Ambiguity of natural languages is also a problem embedded in studies of the Open IE area. Although we have deleted our list of gaps in the “Results” section, a problem always discussed in selected studies is the inaccuracies of NLP tools employed in Open IE systems. Tools such as CoreNLP and Open NLP perform machine learning methods with models built with the limited corpus. By employing these tools in texts of different domains and different writing styles, the inaccuracies become more evident. Open IE tasks must overcome the ambiguity problems of natural languages, adding to the problems of NLP tools, we observe a significant amount of invalid facts being extracted.

One concern in this study is to ensure that all relevant primary studies are recovered. Even by applying our query strings in five databases, only the EMNLP conference concentrates the most significant amount of studies; the other forums present only one or two studies. Our concern was that relevant studies were on unindexed forums in the chosen databases. To mitigate this limitation, we apply some of our query strings in Google Scholar. Even limiting the query in this database, the amount of studies recovered is higher than the other databases. Another significant concern is about the filters applied in this study. It is common sense that different researchers and research groups have different write styles. Even the conferences or journals impose to the researchers the description of their results in a specific way. The StArt tool indicates a ranking of the relevant studies based on the similarity between the paper metadata and the query strings which facilitates the identification of relevant studies. This feature of the tool supports the work of the researcher, but still, we worry about not arbitrarily eliminating studies that the title and summary do not clearly define the type of research. To mitigate this concern, we reviewed all removed studies by applied filters.

Our study recovers some gaps within the Open IE area. We highlight the recurring gaps from the primary studies and, we add some observations. We stated that there is a concentration of methods specialized in texts written in English followed by Chinese. Next, we will evaluate the recurring gaps specifically for Portuguese to propose solutions for this language.

Acknowledgment

The authors would like to thank CAPES for the financial support (Grant no. 1573345).

References

de Abreu, S. C., Bonamigo, T. L., & Vieira, R. (2013). A review on relation extraction with an eye on portuguese.. *Journal of the Brazilian Computer Society, 19*(4), 553–571.

de Abreu, S. C., & Vieira, R. (2017). Relp: Portuguese open relation extraction. *Knowledge Organization, 44*(3), 163–177.

Akbik, A., & Löser, A. (2012). Kraken: N-ary facts in open information extraction. In *Proceedings of AKBC-WEKEX* (pp. 52–56). ACL.

Angeli, G., Premkumar, M. J., & Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of ACL-IJCNLP: 1* (pp. 344–354).

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of IJCAI: 7* (pp. 2670–2676).

Banko, M., Etzioni, O., & Center, T. (2008). The tradeoffs between open and traditional relation extraction.. *ACL, 8*, 28–36.

Bast, H., & Haussmann, E. (2013). Open information extraction via contextual sentence decomposition. In *Proceedings of ICSC* (pp. 154–159). IEEE.

Bast, H., & Haussmann, E. (2014). More informative open information extraction via simple inference. In *Proceedings of ECIR* (pp. 585–590). Springer-Verlag New York, Inc.

Betteridge, J., Carlson, A., Hong, S. A., Hruschka Jr, E. R., Law, E. L., Mitchell, T. M., & Wang, S. H. (2009). Toward never ending language learning.. In *Proceedings of AAAI spring symposium: Learning by reading and learning to read* (pp. 1–2).

Boschee, E., Freedman, M., Khanwalkar, S., Kumar, A., Srivastava, A., & Weischedel, R. (2014). Researching persons & organizations: Awake: From text to an entity-centric knowledge base. In *Proceedings of IEEE big data* (pp. 1030–1039). IEEE.

Christensen, J., Soderland, S., Etzioni, O., et al. (2011). An analysis of open information extraction based on semantic role labeling. In *Proceedings of K-CAP* (pp. 113–120). ACM.

Del Corro, L., & Gemulla, R. (2013). Clausie: Clause-based open information extraction. In *Proceedings of WWW* (pp. 355–366). ACM.

Della Rocca, P., Senatore, S., & Loia, V. (2017). A semantic-grained perspective of latent knowledge modeling. *Information Fusion, 36*, 52–67.

Dusse, F., Júnior, P. S., Alves, A. T., Novais, R., Vieira, V., & Mendonça, M. (2016). Information visualization for emergency management: A systematic mapping study. *Expert Systems with Applications, 45*, 424–437.

Dutta, A., & Schuhmacher, M. (2014). Entity linking for open information extraction. In *Proceedings of NLDB* (pp. 75–80). Springer.

Enríquez, J. G., Domínguez-Mayo, F., Escalona, M., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: a systematic mapping study. *Expert Systems with Applications, 80*, 14–27.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam, M. (2011). Open information extraction: the second generation. In *Proceedings of IJCAI: 11* (pp. 3–10).

Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A., & Belgamo, A. (2016). Improvements in the start tool to better support the systematic review process. In *Proceedings of ease* (p. 21). ACM.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of EMNLP* (pp. 1535–1545). Association for Computational Linguistics.

Faruqui, M., & Kumar, S. (2015). Multilingual open relation extraction using cross-lingual projection. In *Proceedings of HLT-NAACL* (pp. 1351–1356). ACL.

Gamallo, P. (2014). An overview of open information extraction. In *Openaccess series in informatics: 38* (pp. 13–16).

Gamallo, P., & Garcia, M. (2015). Multilingual open information extraction. In *Proceedings of EPIA* (pp. 711–722). Springer.

Garcia, M. (2016). Semantic relation extraction. resources, tools and strategies. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, & A. Branco (Eds.), *Proceedings of PROPOR* (pp. 141–152). Springer.

Garcia, M., & Gamallo, P. (2014). Entity-centric coreference resolution of person entities for open information extraction. *Procesamiento del Lenguaje Natural, 53*, 25–32.

Gashteovski, K., Gemulla, R., & Del Corro, L. (2017). Minie: Minimizing facts in open information extraction. In *Proceedings of EMNLP* (pp. 2630–2640). ACL.

Gotti, F., & Langlais, P. (2016). Harnessing open information extraction for entity classification in a french corpus. In *Proceedings of Canadian conference on artificial intelligence* (pp. 150–161). Springer.

Jijkoun, V., De Rijke, M., & Mur, J. (2004). Information extraction for question answering: improving recall through syntactic patterns. In *Proceedings of COLING* (p. 1284). ACL.

Kim, J., & Myaeng, S.-H. (2016). Discovering relations to augment a web-scale knowledge base constructed from the web. In *Proceedings of WIMS* (pp. 16:1–16:12). ACM.

Kim, M. H., Compton, P., & Kim, Y. S. (2011). RDR-based open ie for the web document. In *Proceedings of K-CAP* (pp. 105–112). ACM.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Technical Report EBSE 2007-001*.

Kitchenham, B. A., Budgen, D., & Brereton, O. P. (2010). The value of mapping studies: A participant observer case study. In *Proceedings of EASE* (pp. 25–33). BCS Learning & Development Ltd..

Konstantinova, N. (2014). Review of relation extraction methods: what is new out there? In *Proceedings of AIST* (pp. 15–28). Springer.

Léchelle, W., & Langlais, P. (2016). An informativeness approach to open ie evaluation. In *Proceedings of CICLING*. Springer.

Lin, H., Wang, Y., Zhang, P., Wang, W., Yue, Y., & Lin, Z. (2016). A rule based open information extraction method using cascaded finite-state transducer. In *Proceedings of PAKDD* (pp. 325–337). Springer.

Liu, F., He, S., Liu, S., Zhou, G., Liu, K., & Zhao, J. (2013). Open relation mapping based on instances and semantics expansion. In *Proceedings of asia information retrieval symposium* (pp. 320–331). Springer.

Liu, M., Ling, Y., An, Y., & Hu, X. (2014). Relation extraction from biomedical literature with minimal supervision and grouping strategy. In *Proceedings of BIBM* (pp. 444–449). IEEE.

Liu, Y., Shi, Z., & Sarkar, A. (2007). Exploiting rich syntactic information for relation extraction from biomedical articles. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; companion volume, short papers* (pp. 97–100).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations* (pp. 55–60).

Merhav, Y., Mesquita, F., Barbosa, D., Yee, W. G., & Frieder, O. (2012). Extracting information networks from the blogosphere. *Transactions on the Web (TWEB), 6*(3), 11.

Mesquita, F., Schmidek, J., & Barbosa, D. (2013). Effectiveness and efficiency of open relation extraction. *New York Times, 500*, 150.

Mohamed, T. P., Hruschka Jr, E. R., & Mitchell, T. M. (2011). Discovering relations between noun categories. In *Proceedings of EMNLP* (pp. 1447–1455). Association for Computational Linguistics.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.

Perera, R., & Nand, P. (2015). A multi-strategy approach for lexicalizing linked open data. In *Proceedings of CICLING* (pp. 348–363). Springer.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering.. In *Ease: 8* (pp. 68–77).

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology, 64*, 1–18.

Petroni, F., Del Corro, L., & Gemulla, R. (2015). Core: Context-aware open relation extraction with factorization machines. In *Proceedings of EMNLP* (pp. 1763–1773).

Presutti, V., Nuzzolese, A. G., Consoli, S., Gangemi, A., & Reforgiato Recupero, D. (2014). From hyperlinks to semantic web properties using open knowledge extraction. *Semantic Web*, 1–28.

Qiu, L., & Zhang, Y. (2014). Zore: A syntax-based system for Chinese open relation extraction.. In *Proceedings of EMNLP* (pp. 1870–1880).

Rodríguez, J. M., Merlino, H. D., Pesado, P., & García-Martínez, R. (2016). Performance evaluation of knowledge extraction methods. In *Proceedings of IEA/AIE* (pp. 16–22). Springer.

Schmidek, J., & Barbosa, D. (2014). Improving open relation extraction via sentence re-structuring.. In *Proceedings of LREC* (pp. 3720–3723).

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open language learning for information extraction. In *Proceedings of EMNLP–CONLL* (pp. 523–534). ACL.

Sena, C. F. L., Glauber, R., & Claro, D. B. (2017). Inference approach to enhance a portuguese open information extraction. In *Proceedings of ICEIS, INSTICC* (pp. 442–451). ScitePress.

Stanovsky, G., & Dagan, I. (2016). Creating a large benchmark for open information extraction. In *Proceedings of EMNLP* (pp. 2300–2305). ACL.

Tan, S. S., Lim, T. Y., Soon, L.-K., & Tang, E. K. (2016). Learning to extract domain-specific relations from complex sentences. *Expert Systems with Applications, 60*, 107–117.

Tandon, N., De Melo, G., & Weikum, G. (2014). Acquiring comparative commonsense knowledge from the web. In *Proceedings of AAAI* (pp. 166–172).

Truong, D., Vo, D.-T., & Nguyen, U. T. (2017). Vietnamese open information extraction. In *Proceedings of SOICT* (pp. 135–142). ACM.

Tseng, Y.-H., Lee, L.-H., Lin, S.-Y., Liao, B.-S., Liu, M.-J., Chen, H.-H., et al. (2014). Chinese open relation extraction for knowledge acquisition.. In *Proceedings of EACL* (pp. 12–16). ACL.

Wang, Y., Zhou, G., Tian, F., Nan, Y., & Ma, J. (2015). Gcore: A gravitation-based approach for chinese open relation. In *Proceedings of CSMA* (pp. 86–91). IEEE.

Wu, F., & Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of annual meeting of the association for computational linguistics* (pp. 118–127). ACL.

Wu, X., & Wu, B. (2017). The crfs-based chinese open entity relation extraction. In *Proceedings of DSC* (pp. 405–411). IEEE.

Xavier, C. C., de Lima, V. L. S., & Souza, M. (2015). Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society, 21*(1), 1.

Xu, J., Gan, L., Deng, L., Wang, J., & Yan, Z. (2015). Dependency parsing based chinese open relation extraction. In *Proceedings of ICCSNT: 1* (pp. 552–556). IEEE.

Yahya, M., Whang, S., Gupta, R., & Halevy, A. Y. (2014). Renoun: Fact extraction for nominal attributes.. In *Proceedings of EMNLP* (pp. 325–335).