Article

Multilingual Open Information Extraction: Challenges and Opportunities

Daniela Barreiro Claro[1,*], Marlo Souza [1], Clarissa Castellã Xavier[2] and Leandro Oliveira[1]

[1] FORMAS Research Group, Computer Science Department, Federal University of Bahia; {dclaro, msouza1}@ufba.br; leo.053993@gmail.com
[2] FORMAS Research Group, Federal Institute of Rio Grande do Sul (IFRS); clarissacastella@gmail.com
* Correspondence: dclaro@ufba.br; Tel.: +55-71-32836336

Version May 6, 2019 submitted to Preprints

Abstract: The number of documents published on the Web in other languages than English grows every year. As a consequence, the necessity of extracting useful information from different languages increases, pointing out the importance of Open Information Extraction (OIE) techniques research. Different OIE methods have been dealing with features from a unique language. On the other hand, few approaches tackle multilingual aspects. In such approaches, multilingualism is restricted to process text in different languages, not on exploring cross-linguistic resources, which results in low precision due to the use of general rules. Multilingual methods have been applied to a vast amount of problems in Natural Language Processing achieving satisfactory results and demonstrating that knowledge acquisition for a language can be transferred to other languages to improve the quality of the facts extracted. We state that a multilingual approach can enhance OIE methods, being ideal to evaluate and compare OIE systems, and as a consequence, to applying it to the collected facts. In this work, we discuss how the transfer knowledge between languages can increase the acquisition from multilingual approaches. We provide a roadmap of the Multilingual Open IE area concerning the state of the art studies. Additionally, we evaluate the transfer of knowledge to improve the quality of the facts extracted in each language. Moreover, we discuss the importance of a parallel corpus to evaluate and compare multilingual systems.

Keywords: Multilingual; Open Information Extraction; Parallel Corpus

1. Introduction

Textual data is the main form of data published in the Web, and the number of published documents increases daily. As much as the Web is a valuable source of information and knowledge, the sheer amount of available pages renders it impossible for a person to explore all of the available information on any subject.

Despite the fact that movements like the Semantic Web [1] and Linked Open Data [2] have urged for the publication of data on the Web in a machine-readable form, it is undeniable that the a great part on the Web is in textual form. As such, it is of great importance to have methods for extracting useful information from texts.

Information Extraction (IE), also called Text Analysis, studies computational methods for identifying structured semantic information from unstructured sources such as documents or web pages. Commonly, IE methods aim to identify semantic information expressed in natural languages, such as discursive entities and their relations, and store it in a standard, computational-friendly, representation for further usages, such as relational tuples.

Notwithstanding IE is a vast area of investigation, covering topics such as Named Entity Recognition, Opinion Mining, traditionally, the term has been commonly employed to refer to one of its central tasks, called Relation Extraction. Relation Extraction methods aim to identify facts expressed in natural language which can represent semantic relations between entities. These entities have

 doi:10.20944/preprints201905.0029.v1

numerous applications in building knowledge representation models that report relations between words, like ontologies, semantic networks, and thesauri, among others.

According to Fader *et al.* [3] "typically, IE systems learn an extractor for each target relation from labeled training examples". These methods are dependent on the domain of application and their adaptation to a new domain requires extensive manual labor. Moreover, this approach is not scalable to corpora with a large number of target relationships or where the target relationships cannot be specified in advance [4].

Recently, the IE problem of domain adaptation and automatic annotation of data across domains has been tackled by several researchers with the use of techniques such as distant supervision for labeling data [5] or learning transferable representations between domains [6].

Distant supervised methods for corpora generation employ techniques to exploit knowledge bases, such as Freebase [7] or Wikidata [8], to annotate texts. While it has become a standard technique annotated data generation in Information Extraction, c.f. [9–12], it is well-recognized that the produced corpora contain a great deal of noise and are subjected to knowledge gaps, in which information that is not available in the source knowledge base cannot be identified in the generated corpora. As such, these techniques require the existence of robust knowledge bases containing annotated information on the target relations.

As Banko et al. [13] discuss in their seminal work, however, in an open context such as that of the Web as a Corpus [14], it is not feasible to enumerate all potential relations of interest for extraction. For specific domains of applications, the existence of such knowledge bases is not guaranteed. In fact, a common application of Information Extraction systems lies on the creation or completion of knowledge bases, aiming to acquire knowledge from existing textual resources in a domain of interest [15–17].

On the other hand, the application of representation learning techniques for IE is also a recent approach to deal with data sparsity and domain adaption in IE [6,18,19]. Representation learning, or feature learning, is a set of machine learning techniques for "learning representations of the data that make it easier to extract useful information when building classifiers or other predictors" [20]. Recently, it has become an important area of research in the area of Machine Learning due to the improvement observed in the application of machine learning techniques, particularly neural-based ones, to diverse tasks in areas such as Artificial Intelligence, Natural Language Processing, Computer Vision, etc.

While representation learning has been applied for domain adaptation in Information Extraction, so far these methods produce systems that are not robust in target domains, when compared to the state of the art for domain-specific IE. Moreover, these methods do not tackle the problem of scalability for IE or when the number of relations is not known before-hand, thus making them unusable in a general context of the Web as a Corpus.

Another approach to tackle these problems have been proposed in the literature, which became known as Open Information Extraction, or Open IE. Introduced in the work of Banko et al. [21], the Open IE approach is a "domain-independent extraction paradigm that uses some generalized patterns to extract all the potential relationships between entities" [22].

It should be noted that, in contrast with traditional IE methods, Open IE approach aims to identify not only a set of previously known semantic relations expressed in a textual fragment but also any semantic relation among concepts, entities, events and also those expressed through attributes. Xavier et al. [23] note that the notion of semantic relation under the Open IE paradigm is broader than that usually employed in the IE literature. In fact, it considers not only the identification of relation instances among entities in a particular domain of discourse, or concrete tuples [13], such as *(Aristotle, was born, Stagira)*, but also for relations "implying properties of general classes" [13], as in *(Philosopher, is author of, book)*. It has been argued, both by Xavier et al. [23] and Wu and Weld [24], that Open IE deals with semantic relations between nominals or concepts, a broader notion than that of relations between entities.

 doi:10.20944/preprints201905.0029.v1

Although nearly half of the Web's content is written in English - W3Techs[1] estimates that around 54% of the 10 million most accessed websites were written in English - the percentage of contents in other languages has been increasing in the last decades. As such, it is of great importance to developing robust Open IE methods for different languages.

Multilingual methods are Natural Language Processing (NLP) methods tailored to work with linguistic resources in multiple languages or to explore linguistic phenomena across different languages. As Faruqui and Kuman [25] point out, multilingual methods may be useful to develop or improve the performance of NLP systems in languages for which computational linguistic resources are unavailable or suffer from low accuracy, by exploring the resources built for other languages. For Information Extraction, multilingual methods are even more critical since content written in different languages are complementary in a sense they present different facts and points of view on the same topic [26].

While multilingual methods have been widely studied in the area of Natural Language Processing for many tasks [27–31], similarly for Information Extraction [32–35], few methods have been proposed to explore multilingual information to the problem of Open IE. This is particularly relevant since Open IE aims to be applicable in a broad context such as that of the Web as a Corpus, which could greatly benefit from extracting information in multiple languages.

In this work, we investigate the area of Multilingual Open Information Extraction, exploring two of the most important systems of the state of the art for the Portuguese and English languages. We conducted a systematic mapping study of the Multilingual Open IE and carried out an initial experiment on parallel corpora and relation extraction systems to improve the effectiveness of Open IE systems. Our results encourage the investigation of transferable methods to achieve cross-language knowledge acquisition.

This paper is organized as follows. Section 2 presents the definitions of Open IE area, giving some examples. Section 3 describes our systematic mapping study and organize our results. Section 4 presents an experiment to handle transferable knowledge acquisition in two languages. Section 5 discusses some challenges and points out some research directions, and finally, section 6 concludes our work and point out some envisioning work.

2. Open Information Extraction

Open Information Extraction allows discovering new facts in a large and heterogeneous set of documents [13]. There is no necessity to previously define the fact to be extracted [36]. An Open IE system performs the task of extracting relationship triples (facts) in raw texts written in natural language in the format:

$$triple = (arg1, rel, arg2) \tag{1}$$

where, $arg1$ and $arg2$ are noun phrases that have a semantic relationship determined by rel such as verb phrases.

Taking the sentence "*The table is in the center of the room*", the fact (*The table, is in the center of, the room*) must be extracted without predefining the relation "*is in the center of*" nor the arguments "*The table*" and "*the room*". Open IE has as a major strength the possibility to extract a high number of facts comparing with traditional IE. However, Open IE diminish the precision of the extracted facts. Open IE strengths are: i) domain independence; ii) unsupervised extraction and iii) more scalability for a large amount of texts [37].

The accuracy of Open IE is still low compared to the Traditional IE, since the number of invalid extractions is high. An extraction is said to be invalid or incorrect when one or more elements of the extracted triplet $t = (arg_1, rel, arg_2)$ does not correspond to the information contained in the

[1] https://w3techs.com/technologies/overview/content_language/all

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

original sentence. For example, taking the sentence "*A deal has been negotiated with another company*", the information (, *'has been negotiated with'*, *'another company'*) is an incorrect extraction because it lacks the first argument (*arg1*) describing what has been negotiated. Similarly, in the sentence "*Added tickets, hotels (touristic superior/first), with coffee, tour guide, transfers and travel insurance.*", the extraction (*'tour guide'*, *'transfers'*, *'travel insurance'*) is also an incorrect extraction, since the word *transfer* in the original sentence does not act as a descriptor of a relation between '*tour guide*' and '*travel insurance*' - being misclassified as a verb in the sentence.

Another case occurs when the extracted triple is valid but uninformative. The extraction is said to be uninformative when the semantic relation expressed by the triple does not correspond to the information presented in the sentence. Considering Table 1, for instance, it is simple to notice that the relation extraction from lines 1 and 2, although still maintain a binary format, they are uninformative. In this example, a semantic meaning presented in the extraction does not represent what is written in the sentences.

Table 1. Uninformative relations examples[38]

Sentence	Uninformative extraction
"After the defense of Bahia rebound, Maurinho kicked and scored."	(*defense of Bahia, rebound, Maurinho*)
"The star symbol of PT will frame the scenario of the candidate's programs Luiz Inácio Lula da Silva."	(*PT, will frame, Luiz Inácio Lula da Silva*)

Over the past few years, many researchers have worked on Open IE approaches, concentrating their efforts on languages other than English. Moreover, some of these researches concern the particularities of each language. Thus, various Open IE solutions have been developing, and new challenges have either emerged.

2.1. Open IE with different languages

In the context of the English language, TextRunner [39] stands out because it was the first Open IE system. After TextRunner, a new system WOE [24] has emerged. WOE operates in two modes: WOE*pos* which uses *Part-of-Speech tagger* (POS tagger) and WOE*parse* which uses a dependency parser. Then a new generation of Open IE systems emerged, focusing on learning patterns that express relationships. ReVerb [36] is its primary approach that uses lexical and syntactic patterns to extract arguments and relations expressed by verbs in English sentences.

A new generation of methods began using dependency and constituency structure analysis and a set of rules for detecting useful parts (clauses) in a sentence. One of the first examples of this approach is DepOE [37] which uses a rule-based parser to extract multi-domain text facts. Another system that is representative of this generation is OLLIE (*Open Language Learning for Information Extraction*) [40].

Currently, Claus-IE [41] and CSD [42,43] methods may be considered the state of the art on Open IE for the English language. Both use a dependency/constituency parser to extract facts, or basic propositions, from textual documents based on syntactic patterns. Claus-IE stands out with the best results in terms of *precision* and *recall*. More recently, we have the MinIE [44] systembased on the main characteristics of Claus-IE, but intending to overcome one of the main gaps of the method: extraction of so-called super-specific facts.

From the standpoint of the Portuguese language, the first proposed methods were: DepOE [37] and ArgOE [45]. Both methods are multilingual and perform extractions for texts in English, Spanish, and Galician as well as Portuguese. In addition, LSOE [46] uses morphosyntactic patterns and also stands out by extracting facts in an unsupervised way. Similar to this approach is the method (nicknamed, SGC_2017) [47]. SGC_2017 proposes an adaptation of the ReVerb [36] to the Portuguese language and a syntactic restriction to identify nominal phrases. SGC_2017 presents an inferential approach to extract new facts using a binary SVM classifier between the transitive and symmetric classes. Next, we have the InferPORToie [48] that enhances their inferential approach and allows

 doi:10.20944/preprints201905.0029.v1

better results than the others. Considering the new generation that use dependency parsers, we have DependentIE [49]. DependentIE is a method that uses a dependency analyzer for the Portuguese language. Improving the DependentIE approach, there is DPToie [50] whose results for the Portuguese language currently outstands the other approaches.

Finally, from the standpoint of other languages, such as Chinese, German and Vietnamese, for example, some methods have been recently proposed in the literature. In the Chinese language the methods CORE [51] and ZORE [52] use a *shallow parser* with a set of syntactic constraints to perform extractions. It is worthy of notice that CORE, according to the authors, was the first Open IE system for the Chinese language. On the other hand, ClausORE [53] and GCORE [54] are characterized by the use of a dependency parser and an heuristic, respectively. ClausORE adopts an Open IE approach for extraction of n-ary facts, while GCORE uses an Open IE approach for extracting binary facts. Another method for the Chinese language is C-COERE [55] which unlike the other methods, uses a semi-supervised learning approach combined with syntactic trees. For the German language, the GerIE method [56] uses dependency analysis to extract facts in textual documents. For the Vietnamese language, the method vnOIE [57], also based on dependency analysis and, according to the authors, is the first Open IE system for this language.

Although some Open IE researches have been facing different languages, few initiatives have emerged to Multilingual Open IE approaches. Thus, we carried a Systematic Mapping Study to conduct this new Open IE direction.

3. Multilingual Open IE: a systematic mapping study

A Systematic Mapping Study (SMS) provides an overview of the scope of the area and allows the discovery of research gaps, forums, relevant authors, research groups, and trends [58,59]. SMS is organized into three groups of activities: planning, conducting and reporting [59]. The first group aims to identify the reasons for this study, followed by the research questions and then definitions 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. Multilingual Open IE is a new research topic with the first published work, that we are aware of, in 2012 [37]. 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 Multilingual Open IE:

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

This MRQ is not entirely explored due to its extensive coverage. We defined a set of secondary questions to help the identification of relevant aspects of Multilingual Open IE. The set of RQ is the baseline of the mapping process, and each RQ is defined as follows:

- RQ1: What are the sources of publications in Multilingual Open IE area?
- RQ2: What are the types of contributions made by Multilingual Open IE studies?
- RQ3: What are the types of applications made by Multilingual Open IE studies?
- RQ4: What are the available Multilingual Open IE datasets?
- RQ5: What are the tools used in Multilingual Open IE systems?
- RQ6: How are Multilingual Open IE systems evaluated?

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 through a set of keywords. As discussed in [60], the term “information extraction” was avoided because of a large number of recovery results. Considering semantic terms from multilingual approaches, we

 doi:10.20944/preprints201905.0029.v1

search for four types of multilingual approaches which we considered to retrieve papers on this topic. Thus, we remain with three relevant keywords which were combined to retrieve primary studies on multilingual Open IE:

- "multi lingual" OR "crosslingual" OR "multilingual" OR "multi-lingual"
- "open information extraction"
- "relation extraction"

We performed our search into five databases: Science Direct[2],IEEE Xplore[3], ACM Digital Library[4], Scopus[5], and Google Scholar[6] with the search strings described.

Our inclusion criteria recover studies that have a recent impact on this research area. Queries were executed on the databases in February 2019 and we recover published papers from 2007 to 2019. The exclusion criteria (F–filters) for primary studies are:

- F1: Remove non-English written paper.
- F2: Remove survey or review paper.
- F3: Remove paper not published in journals or conferences.
- F4: Remove secondary study.
- F5: Remove paper that has some “OpenIE” or "relation extraction" and "multilingual" terms, but is not study on this topic.
- Duplicated: Remove one of the duplicate occurrences.

We choose to remove the studies written in languages other than English due to the fact that English written papers have a broad audience in academic community. Texts such as MSc or Ph.D. reports, technical reports, or any content which were not evaluated by a program committee were also removed. Review studies (secondary studies) were removed as well, since we are concerned only with primary studies. An important observation is that some studies use the terms Open IE and relation extraction, but do not deal with those type of systems. In this SMS we consider either the paper which performs traditional IE, with some insights about Open IE systems. This was considered because of the restriction of the number of papers which cover multilingual approaches. However, almost all papers which do not deal with Traditional IE nor Open IE, but refers to Named Entity, Information Retrieval were removed by F5 filter. After executing the string query, 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 based on a paper full reading. Figure 1 presents the values of each filter applied to each database.

---

2 http://www.sciencedirect.com
3 http://ieeexplore.ieee.org/Xplore/home.jsp
4 http://dl.acm.org
5 http://www.scopus.com
6 http://scholar.google.com

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

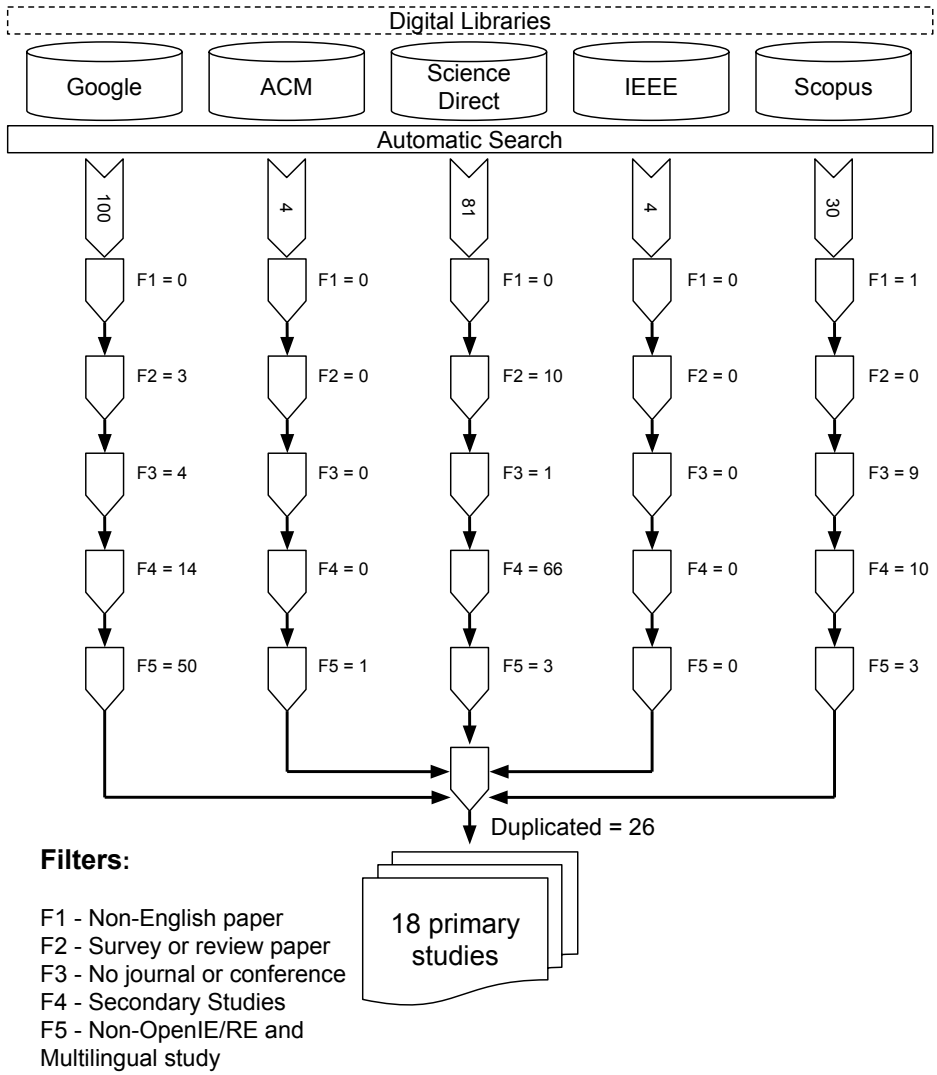

Figure 1. Filters applied in SMS process

After performing the filter step, 18 primary studies were selected. For each primary study, we try to identify each contribution. We summarize the retrieved studies by source database, without considering duplicates, in Figure 2. It is noteworthy that Google Scholar and Science Direct are the most representative databases in our work.

 doi:10.20944/preprints201905.0029.v1

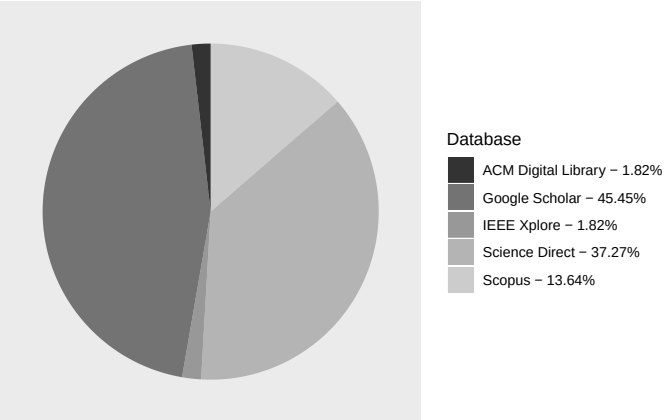

Figure 2. Percentage of each database in primary studies selection.

After this summarization, all duplicate occurrences were removed by the last filter step, gathering a total of 18 papers.

We initiate the analysis phase reading each selected paper and filling a form with such structure described in Table 2.

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	RQ1
Research center	Authors affiliation	RQ1
Country	Country of the Research Centre	RQ1
Affiliation	Affiliation of the authors	RQ1
Source	Source of publication: conference or journal	RQ1
Dataset visibility	Public or Private	RQ4
Dataset language	English, Chinese, Portuguese...	RQ4
Dataset source	Corpus name used to create the dataset	RQ5
Dataset format	Sentence, document, triple, ...	RQ4
Dataset domain	Domain of the Corpus	RQ4
Measure	Evaluation measures used in the study	RQ6
Contribution type	Tool, Resource, Method, Application, Validation or Evaluation	RQ2
NLP task	NLP tasks used in the proposed study	RQ5
NLP tool	NLP tools used in the proposed study	RQ5
Other tool	Other tools used in the proposed study	RQ5
Extract method	Training data or handcrafted rules based	RQ5
Application	Construction of ontology, text summarization...	RQ3

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 following the same order.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

3.1. Answer to RQ1: What are the sources of publications in Multilingual Open IE area?

Our first research question concerns a quantitative measure of the set of selected papers. In our opinion, it was essential to observe the distribution of papers over the years to follow the number of publications. We divided into conference and journal, to have a better overview of this topic.

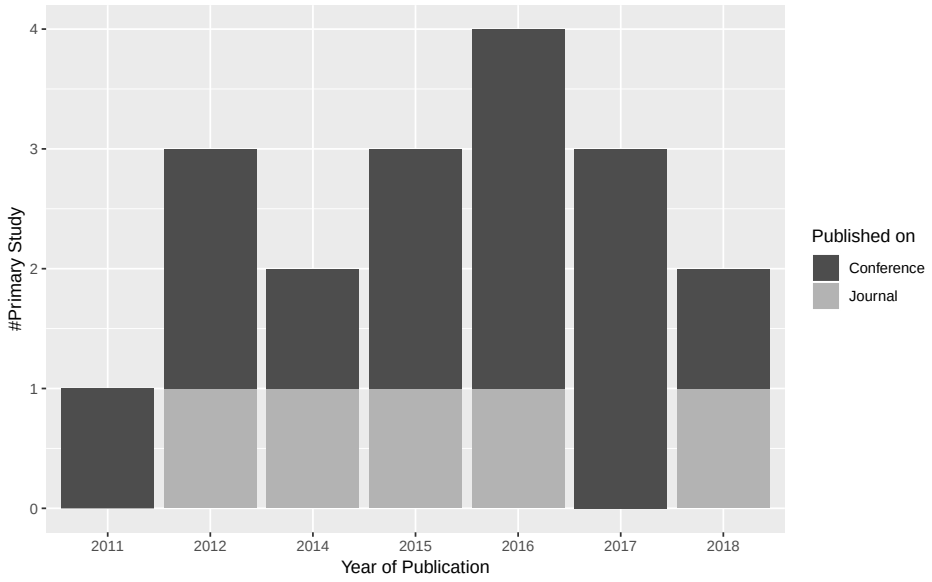

Figure 3. Distribution over the years of selected primary studies per paper type (journal and conference)

As observed in Figure 3, almost paper's publication is in a conference, depicting the youngest area the Multilingual Open IE topic is. Three conferences have published two papers, and all others have published a sole paper, even in journal vehicle. As observed in Figure 4, it is important to highlight the three top conferences (EACL, NAACL, COLING) in the area of Computational Linguistics received more than one paper on the topic, demonstrating its relevance to the NLP community.

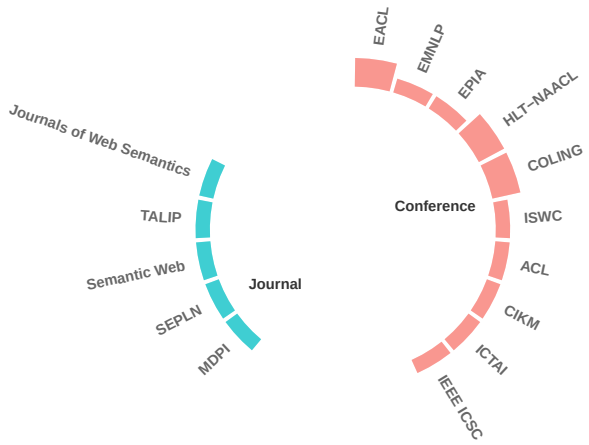

Figure 4. Distribution of the selected papers on Journals and Conferences

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

Figure 5 depicts two countries with high number of participating authors among the 18 selected studies. Both Germany and USA are publishing within this research area (Multilingual Open IE). It is noteworthy that Spain has a high number of researchers working in this topic.

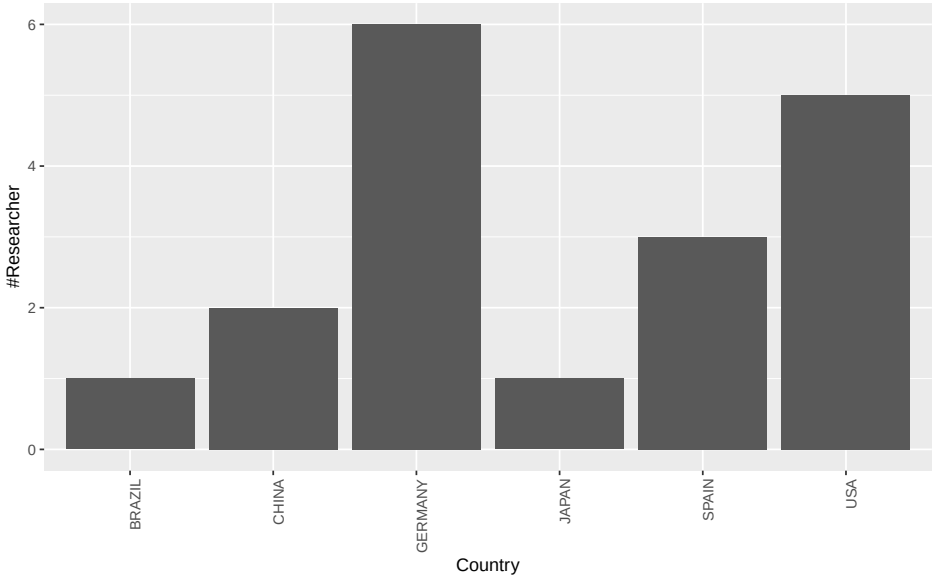

Figure 5. Distribution of the countries of the researchers in primary studies.

Observing organizations and research groups whose studies are in Multilingual Open IE area, Figure 6 deserves prominence to two groups: CITIUS from the University of Santiago de Compostela in Spain and the Mannheim Center for Empirical Multilingualism Research at the University of Mannheim in Germany.

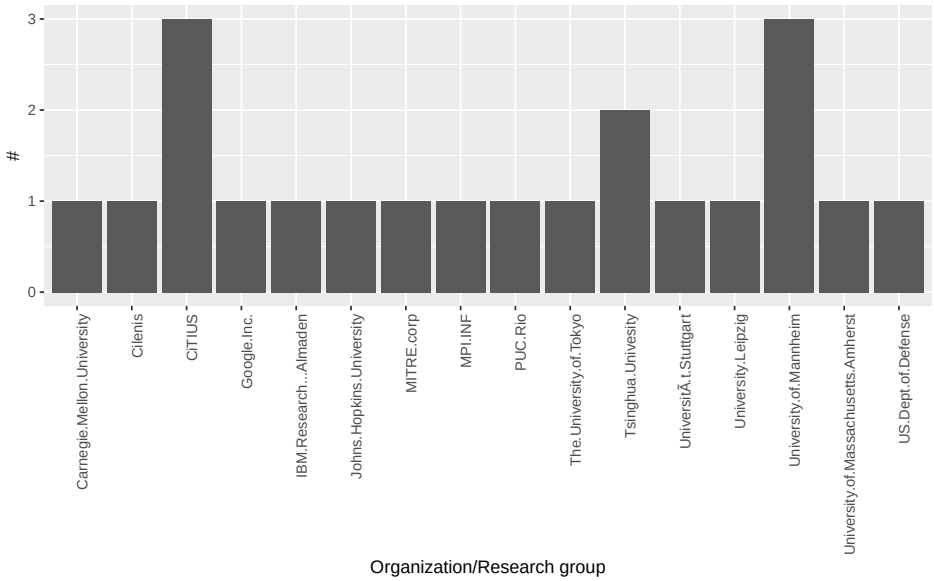

Figure 6. Distribution of the affiliation authors in primary studies.

Both Spain and Germany have significant impact on Multilingual Open IE area due to their number of published papers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

3.2. Answer to RQ2: What are the types of contributions made by Multilingual Open IE studies?

Figure 7 depicts the contribution types of Multilingual Open IE methods. It is noteworthy that the *METHOD* group corresponds to studies with novel methods or approaches to the task of "Multilingual 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 the fact that Multilingual Open IE is a young research area that lacks benchmark materials. Some studies also produce contributions in *RESOURCE* and *TOOL*. *RESOURCE* represents studies that have created datasets for evaluation or other resources for training or testing of Open IE systems. There is also a relevant portion of studies that present a new approach to Multilingual Open IE through a *TOOL*. In these studies, authors present, compare and provide an implementation of their work.

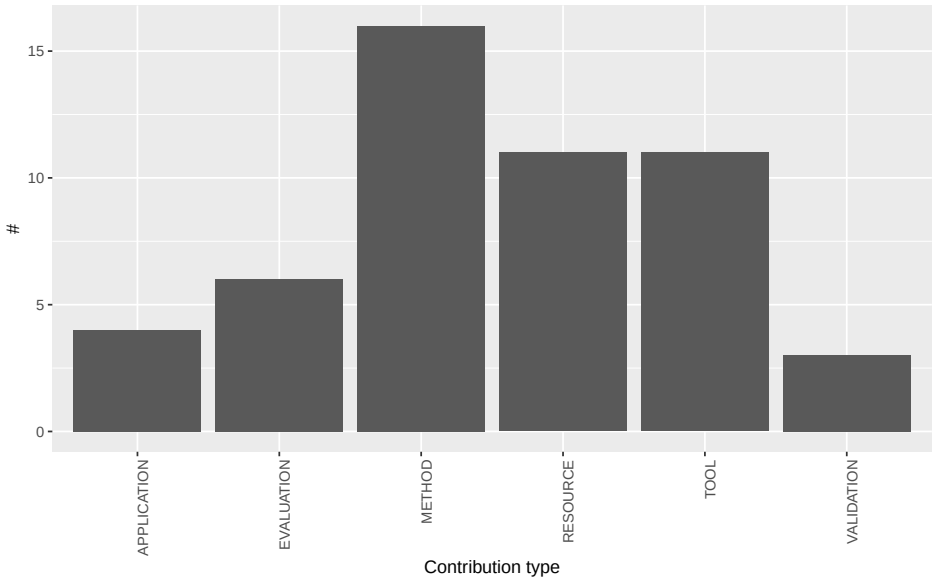

Figure 7. Number of the type of contributions.

The *VALIDATION* corresponds to studies which evaluate the results and *EVALUATION* corresponds to studies that evaluates new measures for Multilingual Open IE systems. *APPLICATION* represents studies which uses a Multilingual Open IE task for some NLP Task.

3.3. Answer to RQ3: What are the types of applications made by Multilingual Open IE studies?

As observed in Figure 8, most multilingual systems are applied to relation extraction tasks. Other possible applications retrieved by our mapping study are: ontology construction, improving QA (Querying Answering) systems and fact checking.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

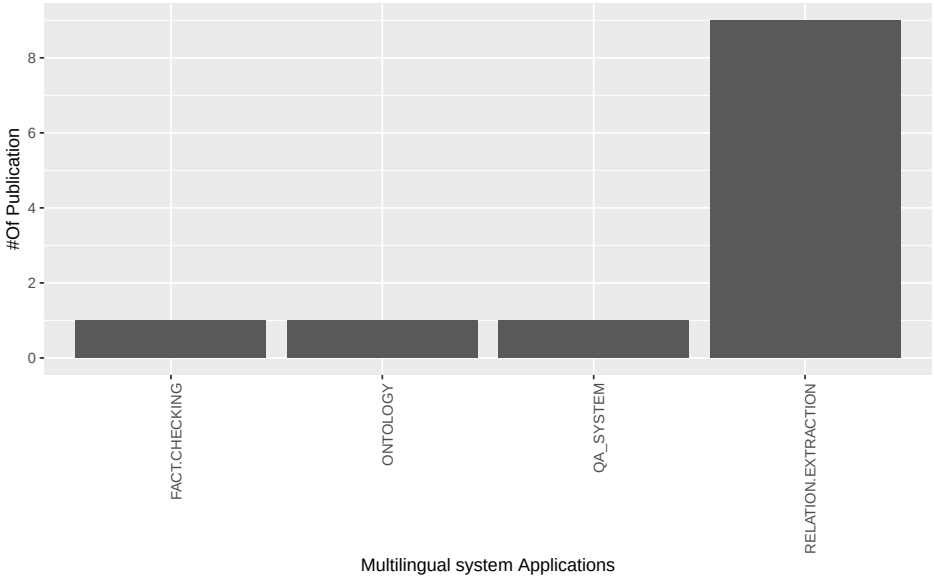

Figure 8. Number of primary studies per type of applications.

3.4. Answer to RQ4: What are the available Multilingual Open IE datasets?

Among the datasets analyzed, we found that there was a high concentration of studies focused on English texts (Figure 9). Many of these datasets are created from sentences retrieved from Encyclopedias (usually Wikipedia) and the Web.

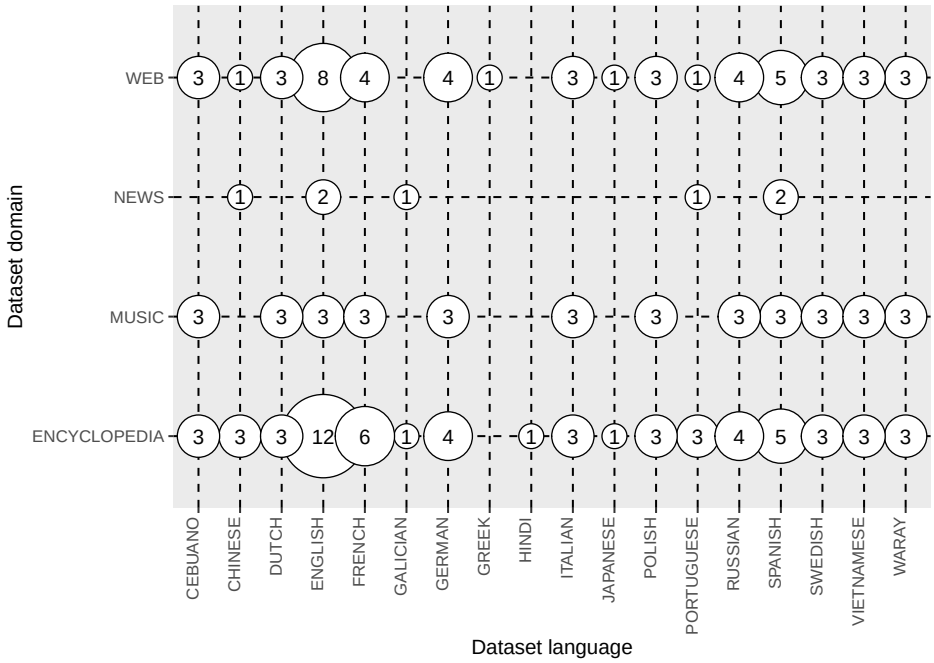

Figure 9. Dataset Language per Domain.

Most of the studies deliver their datasets in a public manner (Figure 10). The advantage of having public datasets is that other researchers can access and use them to compare their approaches, which may encourage the advance of the state of the art.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

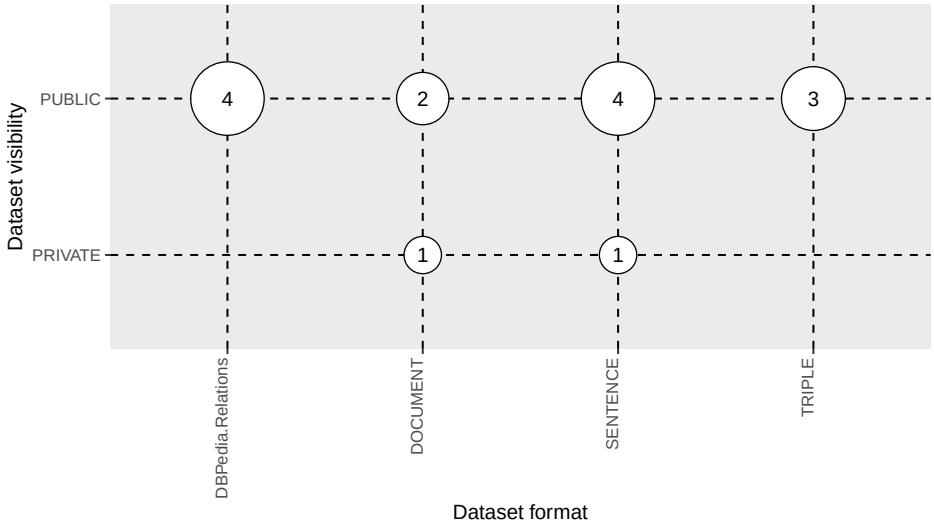

Figure 10. Map of the dataset format and visibility from primary studies. For X axis the format file of the dataset. For Y axis the visibility of datasets.

From these datasets, it is noticeable that they were built from documents and sentences and also from triples and relations. This fact can highlight that those extraction systems may vary their input type, depending on the dataset format. Moreover, extracting information from sentences and documents may be harder than from semi-structured data. This fact may influence the relation extraction task.

3.5. Answer to RQ5: What are the tools used in Multilingual Open IE systems?

RQ5 depicts the main tools used in multilingual systems. From our mapping study, we identified eight taggers. Works in [37,45,61–63] use a Dependency Parser (DP). It is worth mentioning that DPs usually use a POS tagger as an auxiliary tool. In [64], a POS tagger is used without DP. Besides them, works that employ other NLP tools, such as N-Grams extractors, Named Entity Recognizers (NER) and Stemmers, were also identified in our mapping study. On the other hand, some authors [61,62] use co-reference identification techniques to improve their Open IE system performances. We also find systems that use Word Embeddings [32–35] and Semantic Role Labeling (SRL) [63]. In Figure 11, we observed that a POS Tagger and a DP were the most frequent combination, thus being widely used in rule-based methods. On the other hand, systems based on machine learning were less frequent.

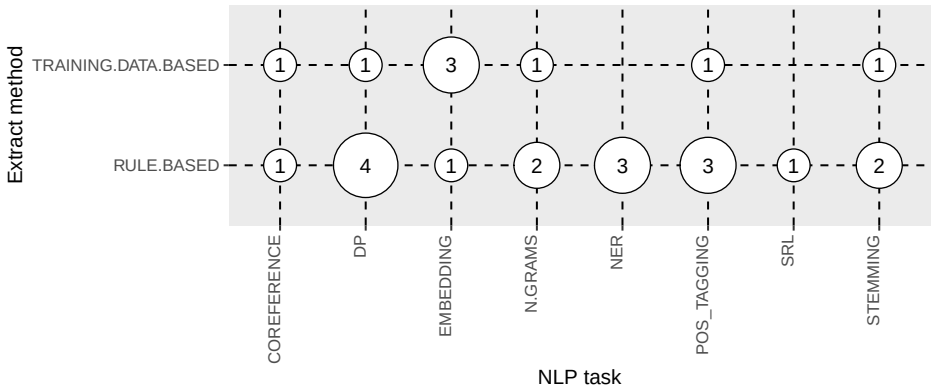

Figure 11. Map of natural language processing tasks and extraction method approach applied in Open IE studies. For X axis the NLP taggers. For the Y axis the extraction method approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

The most commonly used tool/system identified from our mapping study was DepPattern (Figure 12). It is important to note that the three papers which carry the DepPattern were from the same research group, CITIUS. Besides, we found two papers which use the CoreNLP system. Other tools such as Pred Patt, Relation Factory, SyntaxNet, and Word2Vec were less frequent.

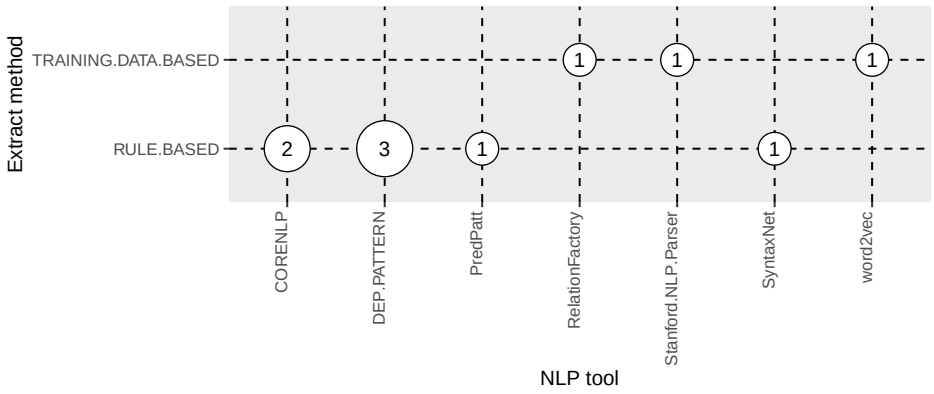

Figure 12. NLP Tool per extract method.

Figure 13 presents the identified data sources and the relationship with the type of contribution of primary studies. DBPEDIA Class and Wikipedia are the most popular sources of data in the works analised here.

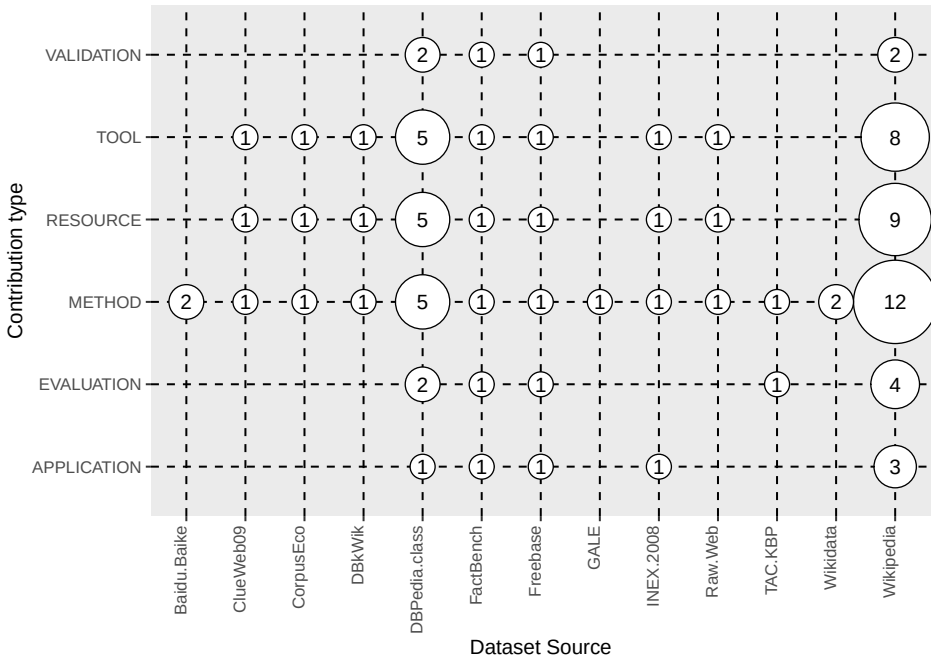

Figure 13. Map of Dataset Sources and types of contribution identified. For X axis the Dataset Sources bases. For Y axis the type of contributions.

3.6. Answer to RQ6: How are Multilingual Open IE systems evaluated?

We have found nine metrics that are employed to evaluate multilingual systems (Figure 14). Among them, the most frequent are Precision, Recall, and F-Measure. In these cases, the most cited difficulty was the lack of large similar golden standards in all used languages.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

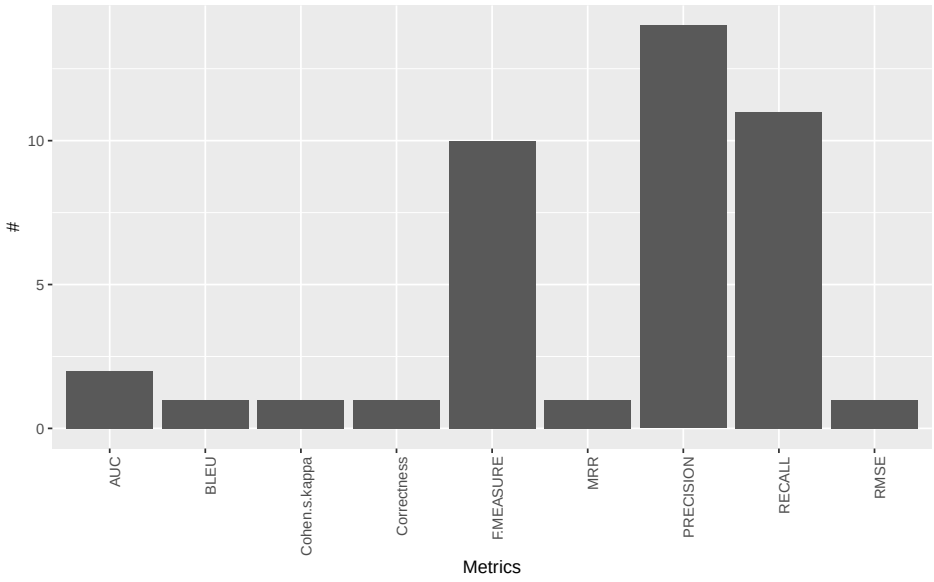

Figure 14. Evaluation Metrics.

In some works, the authors employ Machine Translation tools for performing analysis of the results obtained by Open IE systems. We believe that this is not a reliable approach since no machine translation tool is 100% accurate. In such cases, additionally to the results, it is necessary to deduct the translator error rate from the evaluation.

4. Some experiments on transferable knowledge in Multilingual Open IE

From the systematic mapping study described in the previous section, we can obtain some conclusions.

Firstly, there is still a research gap regarding the topic of multilingual methods for Open IE, while multiple Open IE methods and systems have been proposed for several different languages, c.f. [60], only few of these works perform extraction on different languages.

Most importantly, while we identified some multilingual works on Open IE, only a small fraction of these works use multilingual information to perform Open IE in a given language. Others, like [45], provide several different monolingual models for each different language.

As we have discussed, however, it is our intuition that multilingual information can be used to improve Open IE methods for a given language, since it is well known that corpora written on different language present complementary facts and points of view on the same topics [26]. This intuition is corroborated by previous works which observed improvement in performance in tasks such as Information Retrieval [65], word analogy identification [66], dependency parsing [67,68], and sequence labelling problems such as NER and chunking [69] when exploring multilingual information, such as cross-lingual representations, cross-lingual word clusters, or training the methods on multilingual data.

As such, we perform an exploratory experiment to evaluate how Open IE can benefit from multilingual information. Our hypothesis in this experiment is that we can explore the variation in the linguistic structure between languages to identify information in a target language. To validate this hypothesis, we perform an experiment measuring the degree of complementary of the information extracted by Open IE systems in different languages for a parallel corpus and how we can use the extractions provided by a system in one language to improve the extractions made by a system in the target language.

Notice that, as much as multilingual extraction has been proposed by Faruqui and Kumar [25], based on cross-lingual projection of extractions, their work differs from our in the sense that we advocate that performing Open IE in, and across, multiple languages is advantageous to the task of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

Open IE. Their work, on the other hand, is based on the use of methods and systems developed for a single language, English in their experiments, to obtain extraction for another language. We believe, however, that the different structure and cultural aspects latent in the languages are important clues to structuring the information in a text. As such, we believe that Open IE systems developed for different languages identify different relations and their results are complementary.

4.1. Dataset

In our experiment, we used as input data the Portuguese-English section of the Europarl parallel corpus [70]. Europarl is multilingual corpus composed of European Parliament's debates, dating back to 1996. The corpus is composed of texts in 21 languages with up to 60 million words in each language. The parallel corpus is aligned by sentences linking each composing language. For the Portuguese-English variant, it contains 1,960,407 sentences.

In this work, we randomly selected 1,000 Portuguese-English aligned sentences from the corpus in order to evaluate whether we could explore multilingual resources in order to improve Open IE in a target language: in our case the Portuguese language.

On these 1,000 sentences, we applied two Open IE systems for English and Portuguese, namely Claus-IE [41], for the English language, and DPToie [50], for the Portuguese language. Both systems are based on the extracting clauses based on the dependency structure of the sentence to identify basic propositions in natural language sentences [37]. We choose these two systems since they are based on similar methods and, in fact, DPToie was influenced by Claus-IE. After applying the systems to the selected sentences, we obtained 434 relations for the English sentences and 530 relations for the Portuguese language.

Two human judges performed a manual evaluation of the corrected and informativeness of the extractions. The agreement between the judges was evaluated by Cohen's Kappa.

On the extractions performed by the Claus-IE, the judges agreed on 87% of the annotations, achieving a substantial agreement ($\kappa = 0.74$). On the extractions performed by the DPToie, the judges agreed on 91% of the annotations, achieving near perfect agreement ($\kappa = 0.83$). As such, we believe that the manual evaluation gives trustworthy information on the quality of the extractions that will be used in our experiment.

From the judges' evaluations, we selected all extractions which both judges agreed to be correct, obtaining a final set of 210 extractions for the English language and 218 extractions for the Portuguese language.

4.2. Experiment : analyzing cross-lingual extraction complementarity

In this experiment, we aim to evaluate how much intersection there is in the extractions performed by two Open IE systems for two languages. With this, we want to evaluate how many novels information extracted in one language can boost the extractions for the other.

In order to compare the extractions, we have automatically translated the extracted relations from the English language to Portuguese using Google Translate API and compared the resulting translated extractions with the one obtained from the application of the DPToie in the Portuguese sentences (Figure 15).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

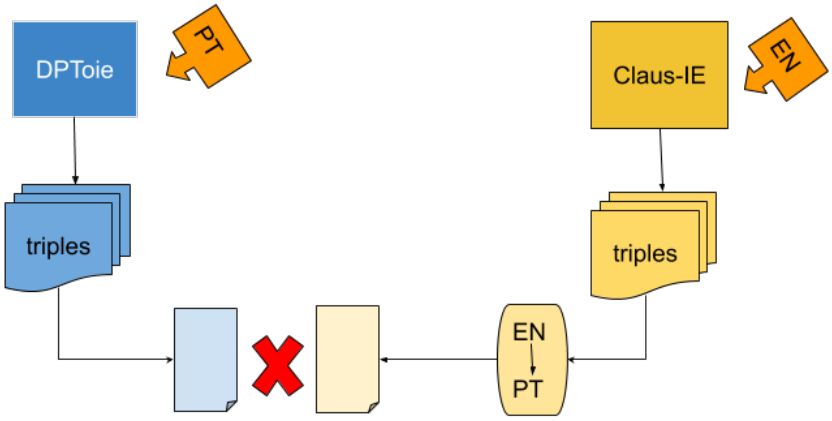

Figure 15. Experiment on relation extractions intersection for two languages: PT and EN

It is important to notice that due to differences in the languages syntax, not all translations resulted in a valid extraction. For example, Portuguese language admits sentences without subjects, known as non-existent subject, when the main verb is impersonal. As such, sentences such as:

"It is naturally important that food can also flow freely"

which have the pronoun '*it*' as subject, can be translated to:

"É naturalmente importante que os produtos alimentares possam também circular livremente."

containing as main verb '*é*' (is) - in this context an impersonal verb - and, thus, does not contain a subject. Naturally, the related extraction from the English (*'it', 'is' , 'important naturally that food can also flow freely'*) cannot be properly translated into Portuguese.

Since Claus-IE extracts tuples with only one argument, differently than DPToie, we only maintained in the translated dataset those extractions containing both arguments. The translated set of extractions contains, thus, 180 relations.

No intersection has been observed between the two sets, i.e. no exact match has been found between the relations of the two sets. Analyzing partial matches the extractions, we observed 29 extractions in the translated set of relations that have the same relation descriptor to some relation in the native Portuguese extractions, 34 coinciding in the first argument with some Portuguese language relation, 1 coinciding in the second argument, and 9 coinciding in relation descriptor and one argument.

We proceed, then, with a manual evaluation of the extractions to understand how the extractions differ between the extraction sets. The first thing to observe is that the parallel sentences in Portuguese and English in the Europarl corpus are not exactly matched. For example, the sentence

"The report proposes twelve representatives for the board of the new Food Authority, two of whom would be representatives of the food industry."

is aligned with

"O senhor deputado propõe para o Conselho de Administração da Autoridade Alimentar Europeia doze representantes, dois dos quais em representação da indústria alimentar."[7]

As a result, the extraction (*'o relatório', 'propõe', '12 representantes . . .'*) obtained from the English sentence presents a mismatch with the one extracted from the Portuguese sentence (*'o senhor deputado',*

[7] "The deputy proposes twelve representatives for the Administrative Council of the European Food Authority, two of which representing the food industry."

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2019 doi:10.20944/preprints201905.0029.v1

'*propõe*', '*para o Conselho* . . .'), namely on the first argument '*o relatorio*' (the report) and 'o *senhor deputado*' (the deputy), originated from the mismatch between the sentences.

Another reason for the low intersection between the sets of extractions is due to different translations for the terms. For example, in the sentence

"*Commissioner, you have said on many occasions. . .*"

the pronoun 'you' has been translated into '*V. Exa.*' (Your Excellency), while in the Portuguese corpus the same sentence describes '*o senhor comissário*' (commissioner) as subject of the verb '*afirmou*' (said), generating the extractions ('*V. Exa.*', '*afirmou*', . . .) and ('*o senhor*', '*afirmou*', . . .).

Comparing the extractions made by Claus-IE and DPToie, however, it is observable that the difference between the triples in the two systems is systematically different, both based on how the information is expressed in the source language and how it is extracted/structured in the sentence.

For example, in the sentence

"*Mr Whitehead has managed, in a balanced report, to expertly combine the many opinions which are around in our Parliament on the establishment of a food authority.*"

Claus-IE extracted the triple ('*Mr Whitehead*', '*has managed*', '*in a balanced report to combine the many opinions expertly*'), while DPToie extracted the triple ('*Phillip Whitehead*', '*conseguiu*', '*reunir de forma magistral em um relatório equilibrado as muitas opiniões existentes en o nosso Parlamento*')[8].

Moreover, there are several extractions performed by Claus-IE for which no similar valid extraction has been achieved by DPToie, such as the extraction ('*Article 5*', '*should define*', '*the objectives of food legislation*') from the sentence

"*For example, Article 5 should clearly define the objectives of food legislation and . . .*"

Similarly, DPToie was able to extract several triples for which Claus-IE made no corresponding extraction.

Adjusting our analysis by the considerations above, we observed that around 20 out of 180 triples extracted by Claus-IE and translated to the Portuguese language were also extracted by DPToie in the source sentences directly in Portuguese, meaning that around 89% of the extractions are new, or information not extracted by the DPToie system.

It is important to notice that the different systems can extract different information from the text not necessarily because of differences in the method, but also on inherent differences in the structure and use of the involved languages. As such, it is possible to observe that the length of the arguments in the triples extracted by DPToie (mean length of 42 characters) is higher than of those extracted by Claus-IE (mean length of 32 characters), while the lengths of relation descriptors are similar across the systems (11 characters for DPToie and 10 for Claus-IE).

5. Challenges and Opportunities

From our initial experiments, we believe that exploring multilingual resources has a great potential to increase performance for Open IE methods. Notably, we were able to observe that even using a simple method based on translation, it was possible to increase the number of meaningful extractions in a given domain.

While this potential is clear from our experiments, we notice a significant limitation on Multilingual Open IE concerning the transfer of knowledge from one language to another.

Another critical issue is the evaluation process on Open IE systems. Although the most common metrics used to evaluate IE systems are *Precision*, *Recall*, and *F-Measure*, it is usually not feasible to perform such evaluations for the task of Open IE due to the great number of extractions performed.

[8] ('*Phillip Whitehead*', '*has managed*', '*to expertly combine in a balanced report the many opinions which are around in our Parliament*').

As an alternative, some works on Open IE commonly adopt an approach in which human annotators analyze a small subset of sentences and determine the correct relationships to be extracted. These systems are then evaluated to the overlap or the similarity of their extractions to the reference annotations.

For multilingual methods, some works concentrate on the evaluation in one or two languages, even when the proposed methods are designed to deal with more than two. In this case, they usually perform a less detailed evaluation in less representative languages, like Portuguese, or even skip the evaluation on these languages. In that way, we identify a gap in the evaluation and comparison of results of the different Open IE systems in a broad, objective and reproducible way.

For this reason, the construction of a multilingual free and open gold standard is of utmost importance for improving Open IE evaluations and providing more reliable results.

In as much as we believe that the potential of using multilingual methods and data to the problem of Open IE is clear from our experiments, notice that a significant limitation on multilingual Open IE is concerned with the transfer of knowledge from one language to another.

Despite the fact that Machine Translation has advanced in the last few years, machine translation systems are not reliable, especially in an open context such as that of the Web as a Corpus. In this context in which any domain-specific text can be considered as input to the system, high accuracy translations are difficult to achieve, due to potential domain-specific terminology. Thus, the problem of domain transfer between languages is not different in principle from the initial problem that motivates the Open IE approach to Information Extraction. Therefore, we believe that different and more robust methods for multilingual Open IE that do not depend solely on machine translation are necessary to the development of this area.

6. Conclusions and Future Directions

In this work, we investigate approaches to the study of Multilingual Open Information Extraction. To do this, we present a systematic mapping study to analyze the multilingual open information extraction area and perform initial experiments on the use of multilingual resources to improve the performance of Open IE systems.

We conclude from our experiments that exploring multilingual resources such as parallel or multilingual corpora can increase the performance of Open IE systems by identifying complementary information to be extracted, as pointed by [26]. As we discussed, however, the multilingual methods identified in the literature are yet overly dependent on machine translation technology, which we believe is not yet robust enough to be applied in a broad context such as that of the Web as a Corpus. As such, new methods must be developed for the area.

As future directions, we notice that there is a long way to transfer relation extractions from one language to another. Besides, preliminary tools such as POS taggers, chunkers, DP analyzers need to be improved to single languages to be incorporated into multilingual perspectives.

We believe that we show in this work that the area of Multilingual Open IE is still young and there is much work to do, but also that there is much potential for its applications.

Funding: *This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil(CAPES) - Finance Code 001* (https://www.capes.gov.br/) and FAPESB (http://www.fapesb.ba.gov.br/)

.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints201905.0029.v1

OIE	Open Information Extraction
IE	Information Extraction
NLP	Natural Language Processing
POS Tagger	Part-of-Speech Tagger
SMS	Systematic Mapping Study
MRQ	Main Research Question
RQ	Research Question
QA	Querying Answering
RDF	Resource Definition Framework
DP	Dependency Parser
SRL	Semantic Role Labeling
NER	Named Entity Recognition

References

1. Berners-Lee, T.; Hendler, J.; Lassila, O.; others. The semantic web. *Scientific american* **2001**, *284*, 28–37.
2. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*; IGI Global, 2011; pp. 205–227.
3. Fader, A.; Soderland, S.; Etzioni, O. Identifying relations for open information extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 1535–1545.
4. Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; Mausam, M. Open information extraction: The second generation. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence. AAAI Press, 2011, Vol. 1, pp. 3–10.
5. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009, pp. 1003–1011.
6. Nguyen, T.H.; Grishman, R. Employing word representations and regularization for domain adaptation of relation extraction. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, Vol. 2, pp. 68–74.
7. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM, 2008, pp. 1247–1250.
8. Vrandečić, D.; Krötzsch, M. Wikidata: a free collaborative knowledge base **2014**.
9. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2010, pp. 148–163.
10. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance multi-label learning for relation extraction. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012, pp. 455–465.
11. Krause, S.; Li, H.; Uszkoreit, H.; Xu, F. Large-scale learning of relation-extraction rules with distant supervision from the web. International Semantic Web Conference. Springer, 2012, pp. 263–278.
12. Nguyen, T.V.T.; Moschitti, A. End-to-end relation extraction using distant supervision from external semantic repositories. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011, pp. 277–282.
13. Banko, M. Open Information Extraction for the Web. PhD thesis, Seattle, WA, USA, 2009. AAI3370456.
14. Kilgarriff, A.; Grefenstette, G. Web as corpus. Proceedings of Corpus Linguistics 2001. Corpus Linguistics. Readings in a Widening Discipline, 2001, pp. 342–344.
15. Yangarber, R.; Grishman, R.; Tapanainen, P.; Huttunen, S. Automatic acquisition of domain knowledge for information extraction. Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000, pp. 940–946.

16. Mooney, R.J.; Bunescu, R. Mining knowledge from text using information extraction. *ACM SIGKDD explorations newsletter* **2005**, *7*, 3–10.
17. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with neural tensor networks for knowledge base completion. Advances in neural information processing systems, 2013, pp. 926–934.
18. Plank, B.; Moschitti, A. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, Vol. 1, pp. 1498–1507.
19. Nguyen, T.H.; Grishman, R. Relation extraction: Perspective from convolutional neural networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015, pp. 39–48.
20. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 1798–1828.
21. Banko, M.; Etzioni, O.; Center, T. The Tradeoffs Between Open and Traditional Relation Extraction. ACL; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; Vol. 8, pp. 28–36.
22. Li, H.; Bollegala, D.; Matsuo, Y.; Ishizuka, M. Using graph based method to improve bootstrapping relation extraction. In *Computational Linguistics and Intelligent Text Processing*; Springer: Berlin, Germany, 2011; Vol. 2, pp. 127–138.
23. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society* **2015**, *21*, 1–14.
24. Wu, F.; Weld, D.S. Open Information Extraction Using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; ACL '10, pp. 118–127.
25. Faruqui, M.; Kumar, S. Multilingual Open Relation Extraction Using Cross-lingual Projection. *arXiv preprint arXiv:1503.06450* **2015**, *abs/1503.06450*, 1351–1356.
26. Steinberger, R. A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation* **2012**, *46*, 155–176.
27. Bel, N.; Koster, C.H.; Villegas, M. Cross-lingual text categorization. International Conference on Theory and Practice of Digital Libraries. Springer, 2003, pp. 126–139.
28. Bering, C.; Drozdzynski, W.; Erbach, G.; Guasch, C.; Homola, P.; Lehmann, S.; Li, H.; Krieger, H.U.; Piskorski, J.; Schäfer, U.; others. Corpora and evaluation tools for multilingual named entity grammar development. Proceedings of Multilingual Corpora Workshop at Corpus Linguistics, 2003, pp. 42–52.
29. Boyd-Graber, J.; Blei, D.M. Multilingual topic models for unaligned text. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009, pp. 75–82.
30. Hassan, S.; Mihalcea, R. Cross-lingual semantic relatedness using encyclopedic knowledge. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 1192–1201.
31. Al-Rfou, R.; Perozzi, B.; Skiena, S. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662* **2013**.
32. Lin, Y.; Liu, Z.; Sun, M. Neural relation extraction with multi-lingual attention. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 34–43.
33. Verga, P.; Belanger, D.; Strubell, E.; Roth, B.; McCallum, A. Multilingual relation extraction using compositional universal schema. *arXiv preprint arXiv:1511.06396* **2015**.
34. Zhang, S.; Duh, K.; Van Durme, B. Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 64–70.
35. Wang, X.; Han, X.; Lin, Y.; Liu, Z.; Sun, M. Adversarial multi-lingual neural relation extraction. Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1156–1166.
36. Fader, A.; Soderland, S.; Etzioni, O. Identifying Relations for Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; EMNLP '11, pp. 1535–1545.
37. Gamallo, P.; Garcia, M.; Fernandez-Lanza, S. Dependency-based Open Information Extraction. Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; ROBUS-UNSUP '12, pp. 10–18.

38. Souza, E.N.P.; Claro, D.B. Extração de Relações utilizando Features Diferenciadas para Português. *Linguamática* **2014**, *6*, 57–65.
39. Etzioni, O.; Banko, M.; Soderland, S.; Weld, D.S. Open information extraction from the web. *Communications of the ACM* **2008**, *51*, 68–74.
40. Mausam.; Schmitz, M.; Bart, R.; Soderland, S.; Etzioni, O. Open Language Learning for Information Extraction. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; EMNLP-CoNLL '12, pp. 523–534.
41. Del Corro, L.; Gemulla, R. ClausIE: Clause-based Open Information Extraction. Proceedings of the 22Nd International Conference on World Wide Web; ACM: New York, NY, USA, 2013; WWW '13, pp. 355–366. doi:10.1145/2488388.2488420.
42. Bast, H.; Haussmann, E. Open information extraction via contextual sentence decomposition. ICSC; IEEE, Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on: rvine, CA, USA, 2013; pp. 154–159.
43. Bast, H.; Haussmann, E. More Informative Open Information Extraction via Simple Inference. Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416; Springer-Verlag New York, Inc.: New York, NY, USA, 2014; ECIR 2014, pp. 585–590. doi:10.1007/978-3-319-06028-6_61.
44. Gashteovski, K.; Gemulla, R.; Del Corro, L. MinIE: Minimizing Facts in Open Information Extraction. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2017, pp. 2630–2640.
45. Gamallo, P.; Garcia, M., Multilingual Open Information Extraction. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*; Pereira, F.; Machado, P.; Costa, E.; Cardoso, A., Eds.; Springer International Publishing: Cham, 2015; pp. 711–722. doi:10.1007/978-3-319-23485-4_72.
46. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open Information Extraction based on lexical-syntactic patterns. Intelligent Systems (BRACIS), 2013 Brazilian Conference on. IEEE, 2013, pp. 189–194.
47. Sena, C.F.L.; Glauber, R.; Claro, D.B. Inference Approach to Enhance a Portuguese Open Information Extraction. Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS,; INSTICC, ScitePress: Porto, Portugal, 2017; pp. 442–451. doi:10.5220/0006338204420451.
48. Sena, C.F.L.; Claro, D.B. InferPortOIE: A Portuguese Open Information Extraction system with inferences. *Natural Language Engineering* **2019**, *25*, 287–306. doi:10.1017/S135132491800044X.
49. de Oliveira, L.S.; Glauber, R.; Claro, D.B. DependentIE: An Open Information Extraction system on Portuguese by a Dependence Analysis. *Encontro Nacional de Inteligência Artificial e Computacional* **2017**.
50. de Oliveira, L.S.; Claro, D.B. DptOIE: A Portuguese Open Information Extraction system based on Dependency Analysis. *Computer Speech and Language - under review* **2019**.
51. Tseng, Y.H.; Lee, L.H.; Lin, S.Y.; Liao, B.S.; Liu, M.J.; Chen, H.H.; Etzioni, O.; Fader, A. Chinese open relation extraction for knowledge acquisition. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014, pp. 12–16.
52. Qiu, L.; Zhang, Y. ZORE: A Syntax-based System for Chinese Open Relation Extraction. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014, pp. 1870–1880. doi:10.3115/v1/D14-1201.
53. Xu, J.; Gan, L.; Deng, L.; Wang, J.; Yan, Z. Dependency parsing based Chinese open relation extraction. 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015, Vol. 01, pp. 552–556. doi:10.1109/ICCSNT.2015.7490808.
54. Wang, Y.; Zhou, G.; Tian, F.; Nan, Y.; Ma, J. GCORE: A Gravitation-Based Approach for Chinese Open Relation. 2015 International Conference on Computer Science and Mechanical Automation (CSMA), 2015, pp. 86–91. doi:10.1109/CSMA.2015.24.
55. Wu, X.; Wu, B. The CRFs-Based Chinese Open Entity Relation Extraction. 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), 2017, pp. 405–411. doi:10.1109/DSC.2017.40.
56. Bassa, A.; Kroll, M.; Kern, R. GerIE-An Open Information Extraction System for the German Language. *Journal of Universal Computer Science* **2018**, *24*, 2–24.

57. Truong, D.; Vo, D.T.; Nguyen, U.T. Vietnamese Open Information Extraction. Proceedings of the Eighth International Symposium on Information and Communication Technology; ACM: New York, NY, USA, 2017; SoICT 2017, pp. 135–142. doi:10.1145/3155133.3155171.
58. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic Mapping Studies in Software Engineering. Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering; BCS Learning & Development Ltd.: Swindon, UK, 2008; EASE'08, pp. 68–77.
59. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* **2015**, *64*, 1 – 18. doi:https://doi.org/10.1016/j.infsof.2015.03.007.
60. Glauber, R.; Claro, D.B. A Systematic Mapping Study on Open Information Extraction. *Expert Systems with Applications* **2018**. doi:https://doi.org/10.1016/j.eswa.2018.06.046.
61. Garcia, M.; Gamallo, P. Entity-centric coreference resolution of person entities for open information extraction. *Procesamiento del Lenguaje Natural* **2014**, *53*, 25–32.
62. Nunes, T.; Schwabe, D. Building Distant Supervised Relation Extractors. 2014 IEEE International Conference on Semantic Computing. IEEE, 2014, pp. 44–51.
63. Akbik, A.; Danilevsky, M.; Kbrom, Y.; Li, Y.; Zhu, H.; others. Multilingual information extraction with PolyglotIE. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, 2016, pp. 268–272.
64. Duc, N.T.; Bollegala, D.; Ishizuka, M. Cross-language latent relational search between japanese and english languages using a web corpus. *ACM Transactions on Asian Language Information Processing (TALIP)* **2012**, *11*, 11.
65. Vulić, I.; Moens, M.F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, 2015, pp. 363–372.
66. Upadhyay, S.; Faruqui, M.; Dyer, C.; Roth, D. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425* **2016**.
67. Xiao, M.; Guo, Y. Distributed word representation learning for cross-lingual dependency parsing. Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014, pp. 119–129.
68. Täckström, O.; McDonald, R.; Uszkoreit, J. Cross-lingual word clusters for direct transfer of linguistic structure. Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2012, pp. 477–487.
69. Yang, Z.; Salakhutdinov, R.; Cohen, W. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270* **2016**.
70. Koehn, P. Europarl: A parallel corpus for statistical machine translation. MT summit, 2005, Vol. 5, pp. 79–86.