information

MDPI

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, Salvador - BA 40170-110, Brazil
[2] FORMAS Research Group, Federal Institute of Rio Grande do Sul, Porto Alegre - RS 90030-040, Brazil
* Correspondence: dclaro@ufba.br; Tel.: +55-71-32836336

Received: 1 May 2019; Accepted: 29 June 2019; Published: 2 July 2019

Abstract: The number of documents published on the Web in languages other than English grows every year. As a consequence, the need to extract useful information from different languages increases, highlighting the importance of research into Open Information Extraction (OIE) techniques. Different OIE methods have dealt with features from a unique language; however, few approaches tackle multilingual aspects. In those approaches, multilingualism is restricted to processing text in different languages, rather than exploring cross-linguistic resources, which results in low precision due to the use of general rules. Multilingual methods have been applied to numerous 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 argue that a multilingual approach can enhance OIE methods as it is ideal to evaluate and compare OIE systems, and therefore can be applied to the collected facts. In this work, we discuss how the transfer knowledge between languages can increase acquisition from multilingual approaches. We provide a roadmap of the Multilingual Open IE area concerning 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 are 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 such as 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 most material on the Web is in textual form. 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. IE methods usually 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 being 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, Relation Extraction. Relation Extraction methods aim to identify facts expressed in natural language which can represent semantic relations between entities. These entities have numerous applications in building knowledge representation models that report relations between words, such as 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 work. 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 to annotate texts, such as Freebase [7] or Wikidata [8]. Despite the fact that it has become a standard technique to annotate 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 lays 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, this 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, for 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 has been proposed in the literature, which has become known as Open Information Extraction, or Open IE. Introduced in the work of Banko et al. [13], the Open IE approach is a "domain-independent extraction paradigm that uses some generalized patterns to extract all the potential relationships between entities" [21].

It should be noted that, in contrast to traditional IE methods, the 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. [22] note that the notion of semantic relation in 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. [22] and Wu and Weld [23] that Open IE

deals with semantic relations between nominals or concepts, a broader notion than that of relations between entities.

While nearly half of the content of the Web is written in English—W3Techs (https://w3techs.com/technologies/overview/content_language/all) 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 recent decades. Given thus, it is of great importance to develop 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 [24] 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 is complementary in the sense that they present different facts and points of view on the same topic [25].

It is worth noting that, with the rise in robust statistical machine translation (SMT) systems, multilingual methods and resources have gained a great deal of attention in the area. The reason for this, we believe, is the emergence of robust bilingual and multilingual parallel corpora in recent decades, such as EuroParl [26], MultiUN [27] and JRC-Acquis [28]. These corpora are constructed in such a way that texts in different languages are aligned so that the pieces (such as sentences or words) can be easily connected to all available languages. While these have been created mainly to develop SMT systems, they have also been applied recently to many other cross-lingual applications such as parallel terminology extraction [29,30], cross-lingual information retrieval [31], and cross-lingual question answering [32].

Discussing the great usefulness of multilingual parallel texts, Mihalcea and Simard [33] argue that any translation can be seen as a semantic representation of a text, in the lack of a better alternative representation, and, as such, may be manipulated for various purposes. On the other hand, what many scholars have realized is that translations across languages can also be used as *bridges* to transfer linguistic annotation from one language to another [34–36]. Multilingual corpora can be seen as a tool to develop more robust NLP systems and resources for different languages.

While multilingual methods have been widely studied in the area of Natural Language Processing for many tasks [37–41] and either for similar tasks on Information Extraction [42–45], few methods explore multilingual information to Open IE. It 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 corpus and relation extraction systems to improve the effectiveness of Open IE systems. Our results reinforce the investigation of transferable methods to achieve cross-language knowledge acquisition.

This paper is organized as follows: Section 2 presents the definitions of the Open IE area, giving some examples. Section 3 describes our systematic mapping study and presents its results. Section 4 describes an experiment to handle transferable knowledge acquisition in two languages and shows the results. Both our Systematic Mapping Study (SMS) and Experiment lead to the discussions in Section 5, which point out some challenges and research directions. Finally, Section 6 concludes and proposes further work.

2. Open Information Extraction

Open Information Extraction enables the discovery of new facts in a large and heterogeneous set of documents [13]. There is no need to previously define the fact to be extracted [3]. Open IE systems extract semantic triples (facts) from texts written in natural language in the format:

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

where $arg1$ and $arg2$ are the noun phrases that have a semantic relationship delimited by rel as a verb phrase.

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*". A major strength of Open IE is the possibility to extract a high number of facts compared to traditional IE. However, Open IE diminishes 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 [46].

The accuracy of Open IE is still low compared to 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 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*'—and is 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 easy to notice that the relation extraction from lines 1 and 2, while maintaining a binary format, is uninformative. In this example, the semantic meaning presented in the extract does not represent what is written in the sentences.

Table 1. Uninformative relations examples [47].

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. Some of this research concerns the particularities of each language.

Open IE with Different Languages

In the context of the English language, TextRunner [48] stands out because it was the first Open IE system. After TextRunner, a new system WOE (Wikipedia-based Open Extractor) [23] 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 [3] is an approach which 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 [46], 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*) [49].

Currently, Claus-IE [50] and CSD (Contextual Sentence Decomposition) [51,52] 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, there is the MinIE [53] system based on the main characteristics of Claus-IE, but designed to overcome one of the main gaps of the method: extraction of so-called super-specific facts.

For Portuguese language, the first proposed methods were: DepOE [46] and ArgOE [54]. Both are multilingual and perform extractions for texts in English, Spanish, and Galician as well as Portuguese. In addition, LSOE (Lexical-Syntactic pattern-based Open Extractor) [55] uses morphosyntactic patterns and also stands out for extracting facts in an unsupervised way. Similar to this approach is the method (nicknamed, SGC_2017) [56]. SGC_2017 proposes an adaptation of ReVerb [3] for 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. Furthermore, there is InferPORToie [57] that enhances the inferential approach and provides better results than the others. Considering this new generation of approaches that use dependency parsers, we have DependentIE [58]. DependentIE is a method that uses a dependency analyzer for the Portuguese language. Improving the DependentIE approach, there is DPToie [59] whose results for the Portuguese language currently stands above the other works.

Finally, regarding 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 [60] and ZORE [61] use a *shallow parser* with a set of syntactic constraints to perform extractions. It is worth noting that CORE, according to the authors, was the first Open IE system for the Chinese language. On the other hand, ClausORE [62] and GCORE [63] are characterized by the use of a dependency parser and a heuristic, respectively. ClausORE adopts an Open IE approach for the extraction of n-ary facts, while GCORE uses an Open IE approach to extract binary facts. Another method for the Chinese language is C-COERE [64], which, unlike the other methods, uses a semi-supervised learning approach combined with syntactic trees. For the German language, the GerIE method [65] uses dependency analysis to extract facts in textual documents. For the Vietnamese language, the method vnOIE [66], also based on dependency analysis and, according to the authors, is the first Open IE system for this language.

Although some Open IE research has been dealing with different languages, few initiatives have emerged for Multilingual Open IE approaches. Thus, we carried out a Systematic Mapping Study on this 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 [67,68]. SMS is organized into three groups of activities: planning, conducting and reporting [68]. The first group aims to identify the reasons for this study, followed by the research questions and then the 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, as far as we are aware, in 2012 [46]. Although some secondary studies have been published, SMS discovers quantitative and qualitative data on primary studies not yet presented in the 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) a biased 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 covers a broad domain concerning multilingual Open IE. We outlined a set of secondary questions to support the identification of relevant aspects of this domain. The set of RQs helps to carry out our mapping process, and each RQ is described as follows:

- RQ1: What are the sources of publications in the area of Multilingual Open IE ?
- RQ2: What are the types of contributions made by Multilingual Open IE studies?
- RQ3: What are the types of applications made for 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?

Our planning step delimits the search method to recover the primary studies. Our search method finds primary studies within an automatically search in electronic databases through our set of keywords. As discussed in [69], the term "information extraction" was avoided because of the large number of results. Regarding semantic terms from multilingual works, we explore four types of multilingual entries to retrieve studies on this topic: "multi lingual", "crosslingual", "multilingual" and "multi-lingual". All of them recover a set of multilingual aspects. Exploring these entries, we can consider multilingual aspects to be more global than crosslingual. However, with our search engine, we had 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 conducted our search in five databases: Science Direct (http://www.sciencedirect.com), IEEE Xplore (http://ieeexplore.ieee.org/Xplore/home.jsp), ACM Digital Library (http://dl.acm.org), Scopus (http://www.scopus.com), and Google Scholar (http://scholar.google.com). These databases are the main vehicles in the Computer Science domain. Other databases such as Web of Science and DBLP were not described in our SMS due to Google Scholar indexed almost all single scientific repository, i.e., papers returned from DBLP are duplicate entries covered by our SMS. This set has been used in other systematic mapping studies [67,69].

Our inclusion criteria retrieved works with a recent impact on this research area. Queries were performed through the databases in February 2019, and we retrieved published papers from 2007 to 2019. Our exclusion criteria (F–filters) for primary studies were:

- F1: Remove non-English written paper.
- F2: Remove survey or review paper.
- F3: Remove paper not published in journals or conferences.
- F4: Remove paper that has some "openie" or "relation extraction" terms, but do not deal with this topic.
- F5: Remove the non-multilingual paper.
- Duplicated: Remove one of the duplicate occurrences.

Works written in languages other than English were removed firstly because of the difficulty in understanding the language that they were written in and secondly due to the fact that texts in English have a broader public in the academic community. Manuscripts such as academic reports, technical reports, or any text which had not been evaluated by a program committee were removed. Review studies (secondary studies) were also removed as we were interested in only primary studies. It is

important to note that some works use the term Open IE or Relation Extraction; however, they do not deal with these approaches. We include papers which present Relation Extraction or Open IE approaches. This was to restrict the number of papers that cover multilingual approaches. Papers with no multilingual aspects were removed. Nevertheless, papers which do not deal with Relation Extraction or Open IE, but refer to Named Entity or Information Retrieval were removed by F4 filter. The set of Filters starts after performing the string query. The removal of primary studies was carried in two stages. In the first stage, we read the abstract of each paper to identify occurrences outside the scope of our SMS. In the second stage, with the remaining studies, we read each paper fully to filter. Figure 1 displays each filter applied into each database.

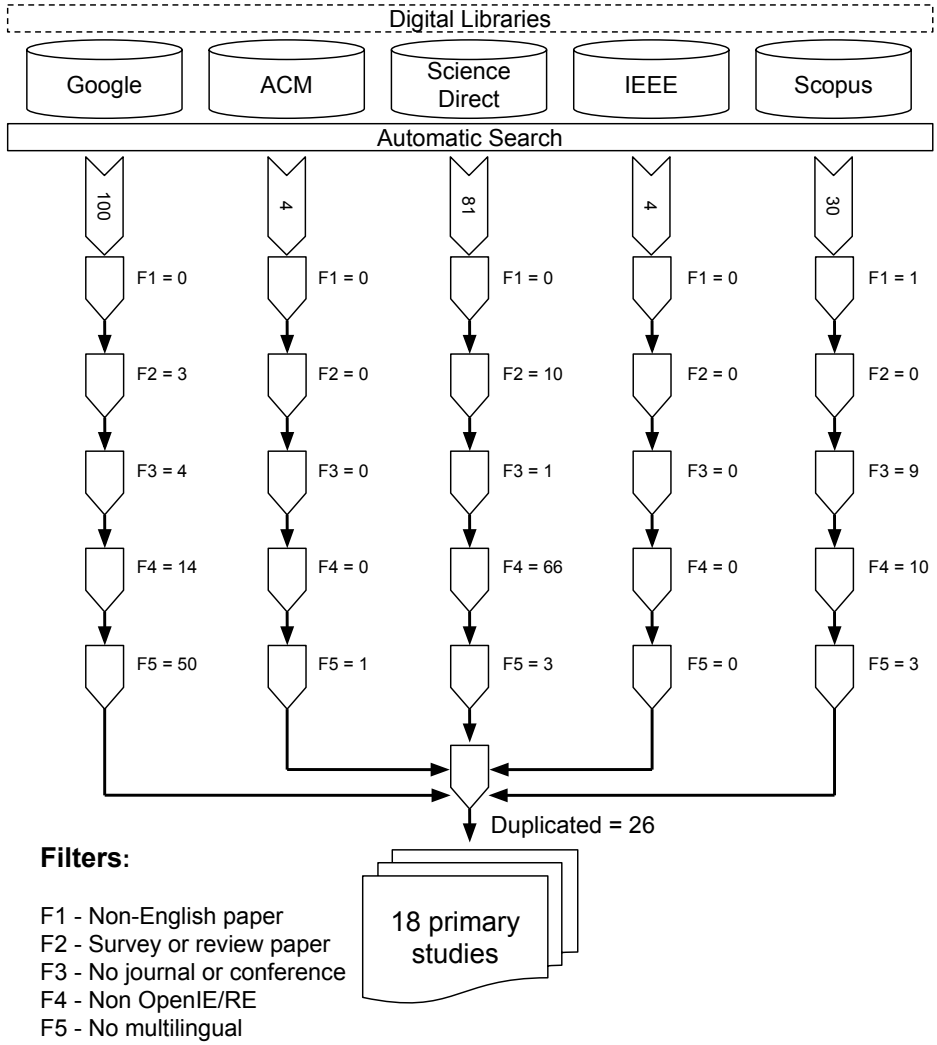

Figure 1. Filters applied in the SMS process. (Authors' own).

After executing the filter step, 18 primary studies were selected. We identified each contribution from each work. The studies recovered are summarized for each database in Figure 2. Google Scholar and Science Direct are the most representative databases in our work.

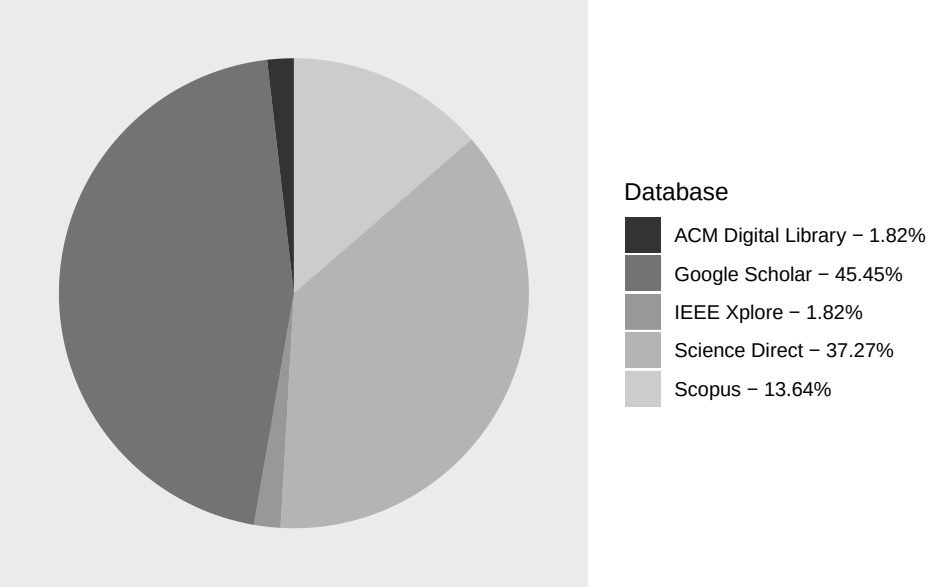

Figure 2. Percentage of the selected primary study per database. (Authors' own).

Our analysis phase started by reading each selected work and filling a form with the structure described in Table 2.

Table 2. Data extraction form manipulated by our SMS, adapted from [69].

Data Item	Value	RQ
General study ID	Integer	
Article Title	Title of the Article	
Author list	List of the Author's name	
Year of publication	Calendar year	RQ1
Research center	Author's affiliation	RQ1
Country	Country of the Research Centre or Organization	RQ1
Affiliation	Affiliation of the authors	RQ1
Publication	Source of publication: conference or journal	RQ1
Dataset visibility	Public or Private	RQ4
Dataset language	English, Chinese, Portuguese...	RQ4
Dataset source	Corpus name employed to create the dataset	RQ5
Dataset format	Sentence, document, triple, ...	RQ4
Dataset domain	Domain of the Corpus	RQ4
Evaluation	Evaluation measures used in the study	RQ6
Contribution type	Tool, Resource, Method, Application, Validation or Evaluation	RQ2
NLP task	NLP tasks employed in the study	RQ5
NLP tool	NLP tools employed in the study	RQ5
Other tool	Other tools employed in the study	RQ5
Extract method	Training data or handcrafted rules based	RQ5
Application	Construction of ontology, text summarization...	RQ3

After collecting the data from each study and filling in the form (Table 2), the next step is to extract useful information from the collected data. Each piece of useful information extracted aims to answer the set of RQs in the order.

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 them into conference and journal papers Table 3 to have a better overview of this topic.

Table 3. Conference and journal names and acronyms.

Acronym	Conference and Journal Names
EACL	European Chapter of the Association for Computational Linguistics
EMNLP	Conference on Empirical Methods in Natural Language Processing
EPIA	Portuguese Conference on Artificial Intelligence
HLT-NAACL	Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics
COLING	International Conference on Computational Linguistics
ISWC	International Semantic Web Conference
ACL	Association for Computational Linguistics
CIKM	Conference on Information and Knowledge Management
ICTAI	International Conference on Tools with Artificial Intelligence
IEEE ICSC	IEEE International Conference on Semantic Computing
TALIP	ACM Transactions on Asian Language Information Processing
SEPLN	Spanish Society for Natural Language Processing
MDPI Information	Information Journal
Semantic Web	The Semantic Web Journal
–	Journal of Web Semantics

As observed in Figure 3, most papers were published as conference paper, depicting how recently the Multilingual Open IE topic is. Three conferences published two papers, and all the others published just one paper. As can be observed in Figure 4, it is important to highlight the three top conferences: European Chapter of the Association for Computational Linguistics (EACL), the North American Chapter of the Association for Computational Linguistics (NAACL), the International Conference on Computational Linguistics (COLING) in the area of Computational Linguistics received more than one paper on a topic, and demonstrating its relevance to the NLP community. It is important to note that, while work on an Multilingual Open IE figure in the most important conferences in NLP, showing its importance to the area, we have identified a few works published in important conferences focusing on evaluation or empirical/reproducible methods in NLP, such International Conference on Language Resources and Evaluation (LREC) or Empirical Methods in NLP (EMNLP). We believe this outlines a reduced availability of resources to Multilingual methods and a lack of clear methodological framework to guide the development and improvement of methods for its tasks. Thus, it indicates a new challenge of this emerging area.

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

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

Both Spain and Germany have had a significant impact on the Multilingual Open IE area given their number of published papers.

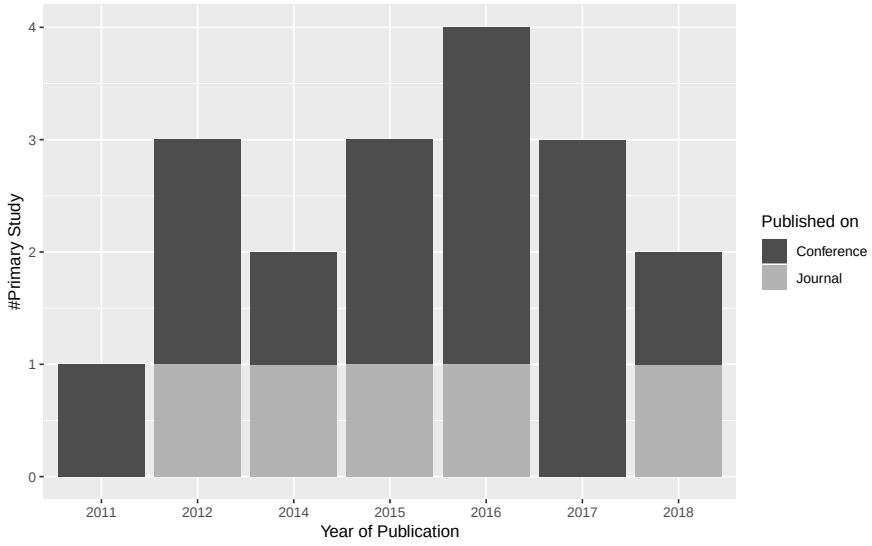

Figure 3. Distribution over the years of selected studies per paper type (journal and conference). (Authors' own).

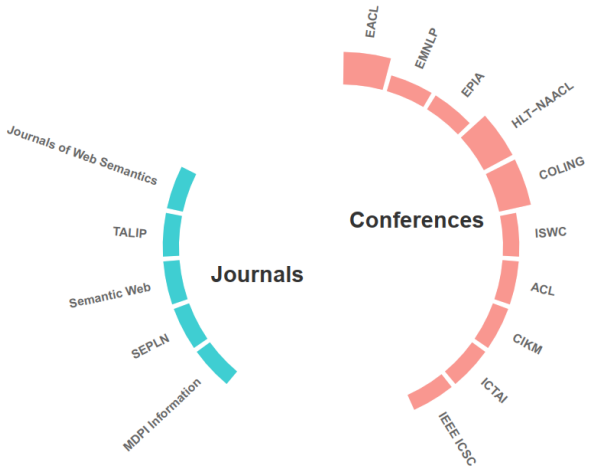

Figure 4. Distribution of the selected papers on Journals and Conferences. (Authors' own).

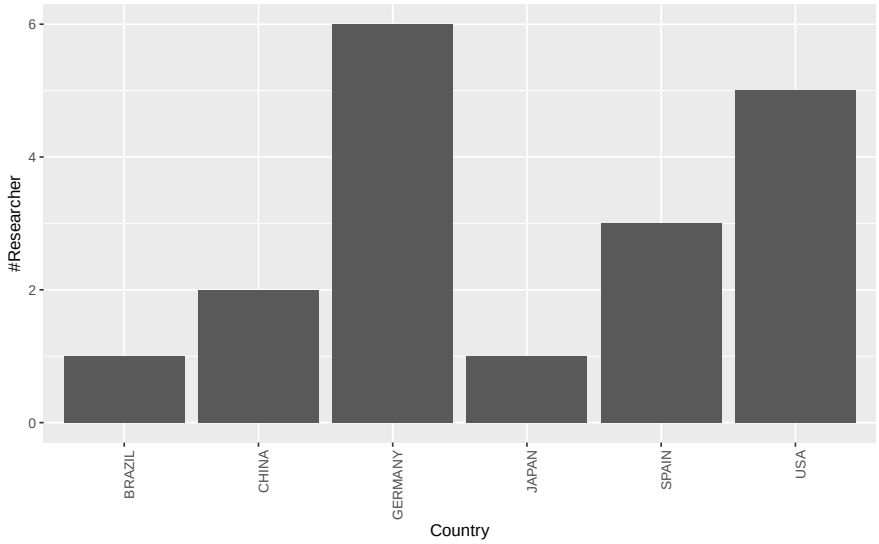

Figure 5. Distribution of the countries of the researchers in primary studies. (Authors' own).

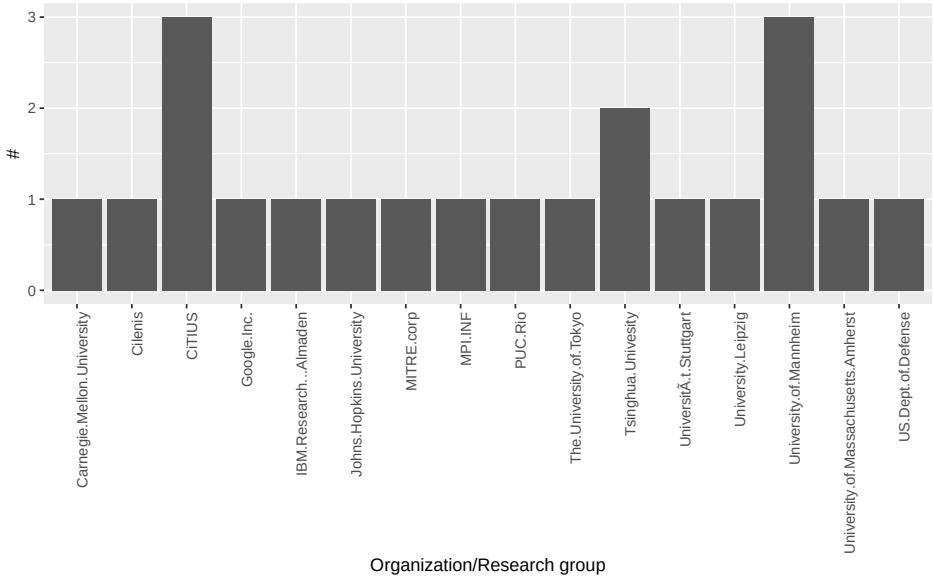

Figure 6. Distribution of the affiliation authors in primary studies. (Authors' own).

3.2. Answer to RQ2: What Are the Types of Contributions by Multilingual Open IE Studies?

The contribution types of Multilingual Open IE methods are described in Figure 7. The *METHOD* group corresponds to works including novel methods or techniques for "Multilingual Open IE" task. New approaches boost the comparison with other works, though Multilingual Open IE is a growing research area that lacks benchmark materials. Other studies provide some contributions in *RESOURCE* and *TOOL*. *RESOURCE* represents works which have created datasets for evaluation or other resources for training or testing Open IE systems. There is also a number of studies that present a new approach through a *TOOL*. In this type of work, the authors present an implementation of their approach.

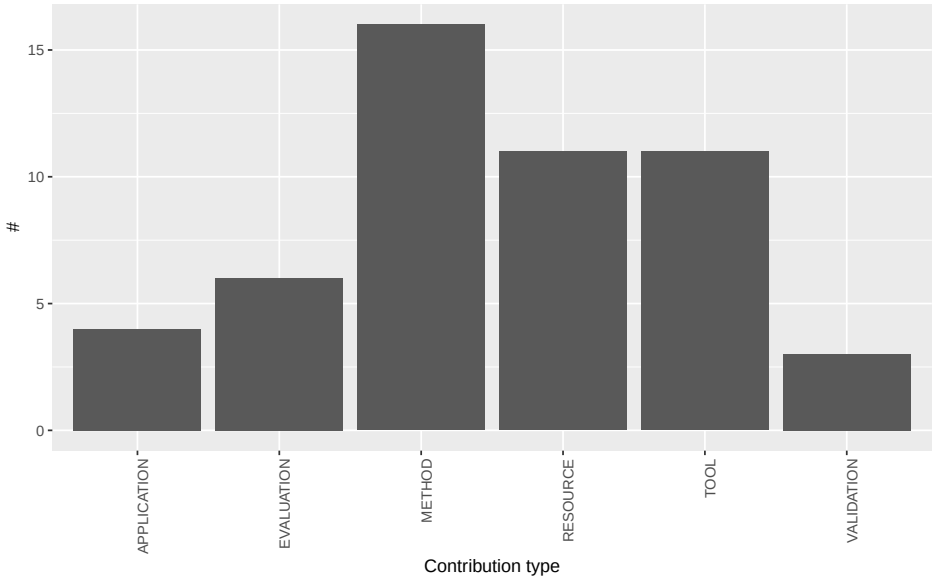

Figure 7. Type of contributions. (Authors' own).

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

We can observe that most approaches on Multilingual Open IE are concerned with proposing methods and tools, for which they evaluate on a specific resource created for such. Again, the focus on

proposing methods and tools, not on validation or evaluation, and the diversity of resources used in the area indicates a lack of established methodology for Multilingual Open IE.

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.

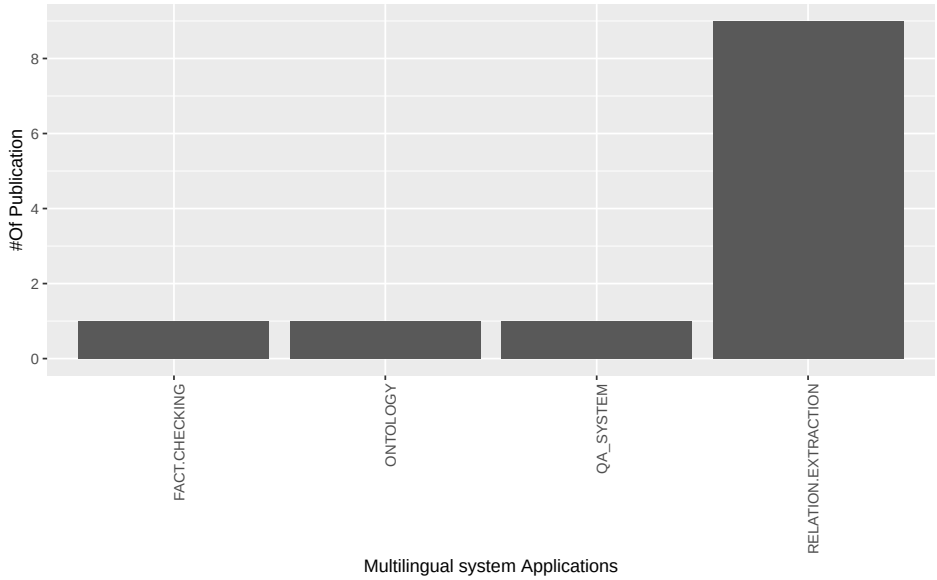

Figure 8. Number of primary studies per type of applications. (Authors' own).

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). Even though English carries a large number of datasets, the distribution among other languages is equivalent, and it covers 18 different dataset languages with at least three occurrences. However, no multilingual corpora were found, outlining a potential opportunity in developing resources for the area. Notice that the focus of the area on the English language (particularly, based on prose using in journalistic or encyclopedic texts) and its particular characteristics, although natural from an engineering point of view due to the availability of resources, may induce important bias in the area [70]. The use of multilingual resources, as proposed in this work, may come as a solution to this problem, leading to more robust and linguistically supported methods and applications.

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.

From these datasets, it is noticeable that they were built from documents and sentences and also from triples and relations. This highlights that these 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 and this may influence the relation extraction task.

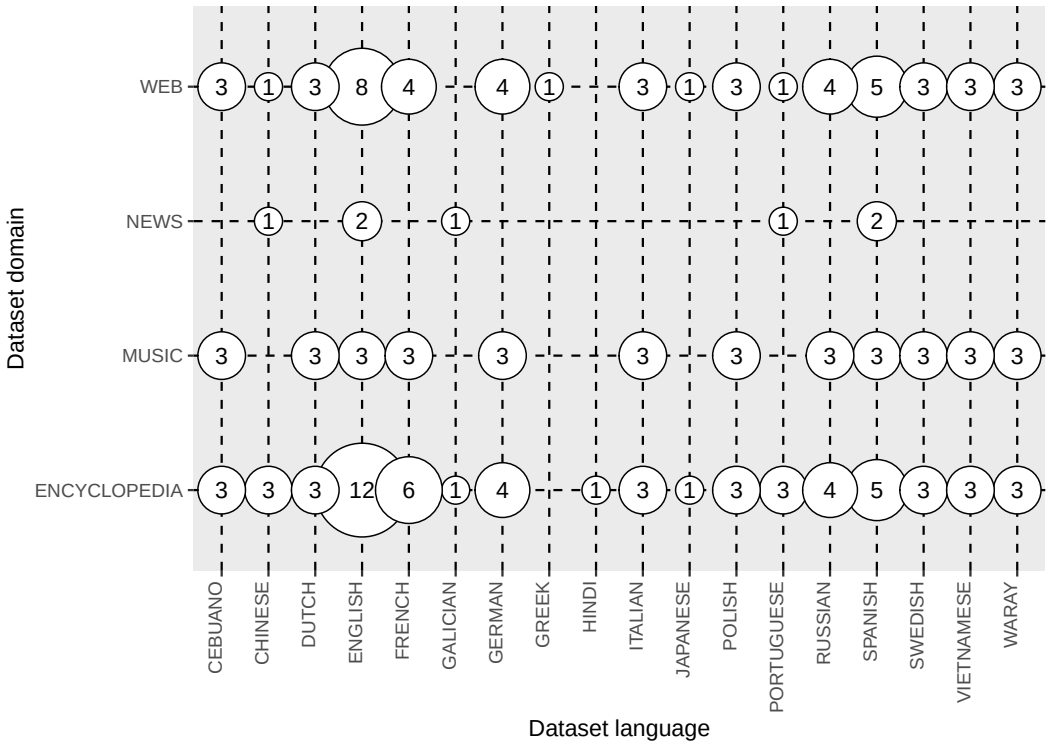

Figure 9. Dataset language per domain. (Authors' own).

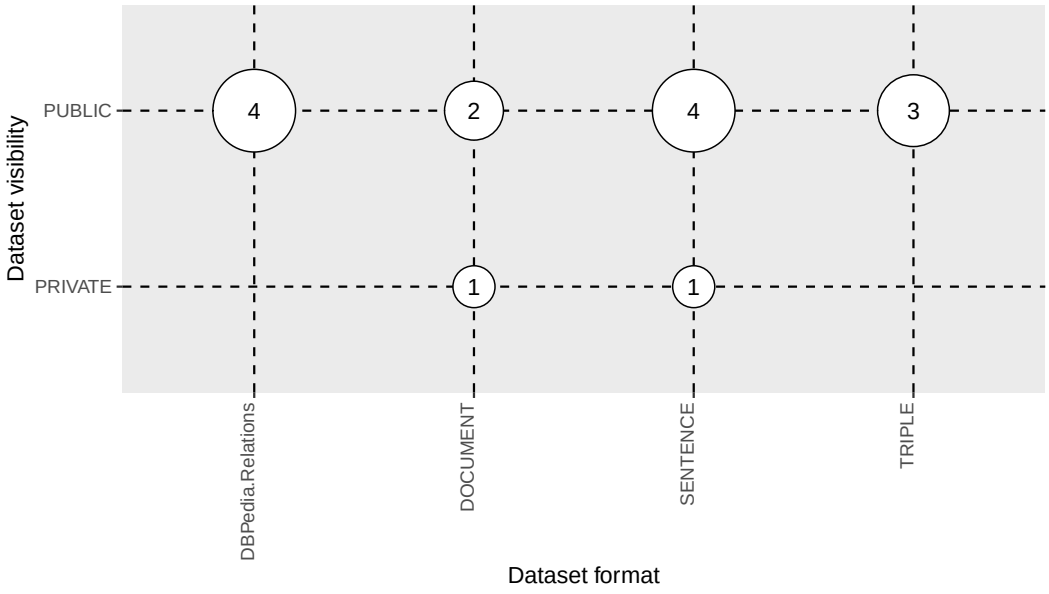

Figure 10. Mapping the dataset format and visibility from primary studies—for the *x*-axis, the format file of each dataset; for the *y*-axis, the visibility of the dataset. (Authors' own).

3.5. Answer to RQ5: What Are the Tools Used in Multilingual Open IE Systems?

RQ5 shows the main tools used in multilingual systems. From our mapping study, we identified eight taggers. Works in [46,54,71–73] use a Dependency Parser (DP). It is worth mentioning that DPs usually use a POS tagger as an auxiliary tool. In [74], a POS tagger is used without DP. Apart from 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 [71,72] use co-reference identification techniques to improve their Open IE system performances. We also

find systems that use Word Embeddings [42–45] and Semantic Role Labeling (SRL) [73]. In Figure 11, we observed that a POS Tagger and a DP were the most frequent combinations being widely used in rule-based methods while systems based on machine learning were less frequent, outlining a new challenge to multilingual methods.

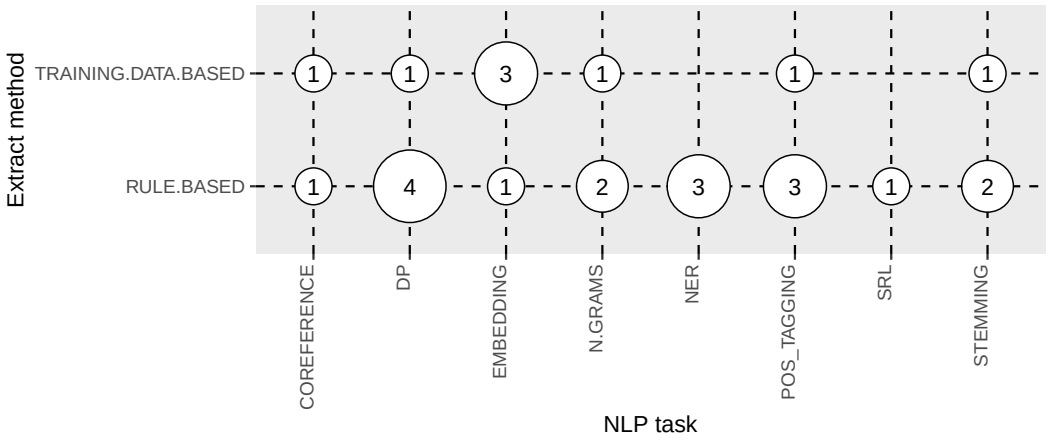

Figure 11. Mapping the NLP task and each extraction approach applied in Open IE studies—for the *x*-axis, the NLP tagger; for the *y*-axis, the extraction approach. (Authors' own).

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 use the DepPattern were from the same research group, CITIUS. We found two papers which use the CoreNLP system. Other tools such as Pred Patt, Relation Factory, SyntaxNet, and Word2Vec were less frequent.

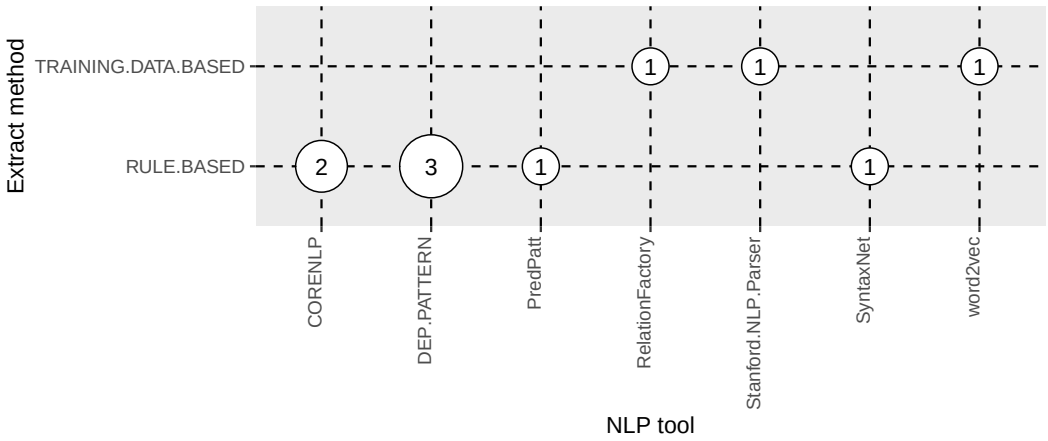

Figure 12. NLP Tool per extract method. (Authors' own).

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 analyzed here. Potential opportunities emerge to multilingual methods through exploring different dataset domains, linguistic styles and an overall greater linguistic variation.

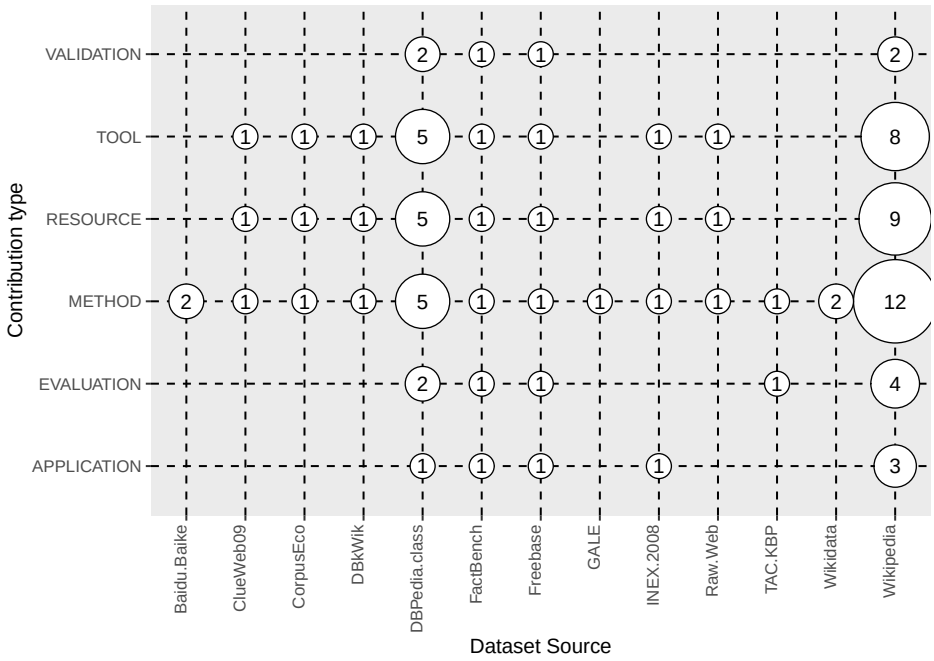

Figure 13. Map of Dataset Sources and types of contribution identified—for the *x*-axis, the Dataset Sources bases; for the *y*-axis, the type of contributions. (Authors' own).

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 gold standards in all used languages.

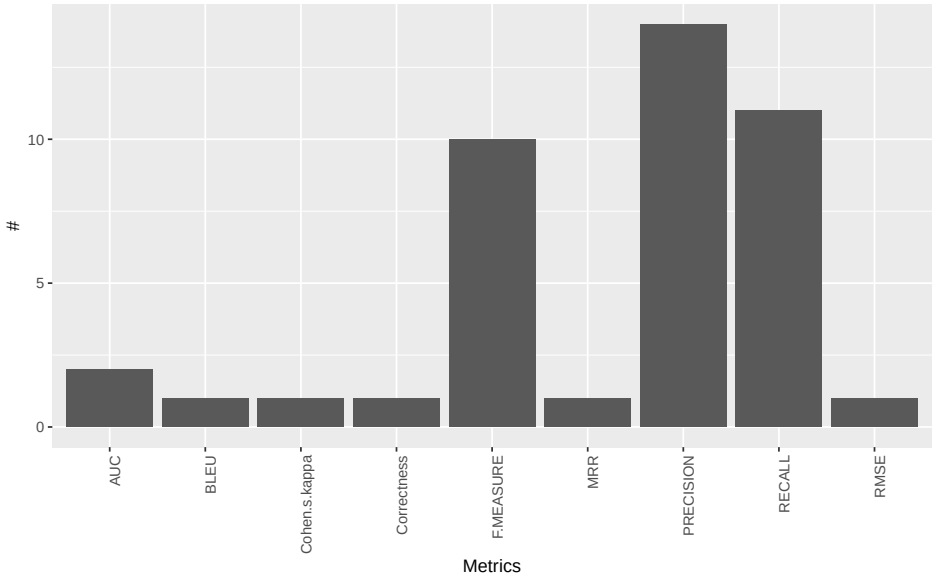

Figure 14. Evaluation metrics. (Authors' own).

In some works, the authors employ Machine Translation tools to perform 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, in addition 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 draw 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. [69]; only a few of these works perform extraction on different languages.

More importantly, while we identified some multilingual works on Open IE, only a few of these works use multilingual information to perform Open IE in a given language. Others, such as [54], 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 in different languages present complementary facts and points of view on the same topics [25]. This intuition is corroborated by previous works which observed improvement in performance in tasks such as Information Retrieval [75], word analogy identification [76], dependency parsing [77,78], and sequence labelling problems such as NER and chunking [79] 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 see 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.

Note that, while multilingual extraction has been proposed by Faruqui and Kumar [24], based on a cross-lingual projection of extractions, their work differs in the sense that we advocate that performing Open IE in, and across, multiple languages are advantageous to the task of 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 vital clues to structure the information in a text. Therefore, we believe that Open IE systems developed for different languages identify different relations, and the results are complementary.

4.1. Dataset

In our experiment, we used the Portuguese-English section of the Europarl parallel corpus as input data [26]. Europarl is multilingual corpus composed of European Parliament 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 composed of sentences aligned per language. For the Portuguese-English variant, it contains 1,960,407 sentences.

In this work, we randomly selected 1000 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. Note that, in this work, we limit our analysis to the Portuguese and English languages since we did not have access to enough human judges fluent in other languages to perform the necessary evaluations. More experiments with other languages are the subject of future studies.

On these 1000 sentences, we applied two Open IE systems for English and Portuguese, namely Claus-IE [50], for the English, and DPToie [59], for the Portuguese. Both systems are based on the extracting clauses based on the dependency structure of the sentence to identify basic propositions in natural language sentences [46]. 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 508 relations for the Portuguese language.

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

On the extracts performed by the Claus-IE, the judges agreed on 87% of the annotations, achieving substantial agreement ($kappa = 0.73$). On the extractions performed by DPToie, the judges agreed on 91% of the annotations, achieving near perfect agreement ($kappa = 0.82$). We therefore believe that manual evaluation gives trustworthy information on the quality of the extracts 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 produced by two Open IE systems for two languages. With this, we want to evaluate how much novel information extracted in one language can boost the extractions for the other.

In order to compare the extractions, we 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).

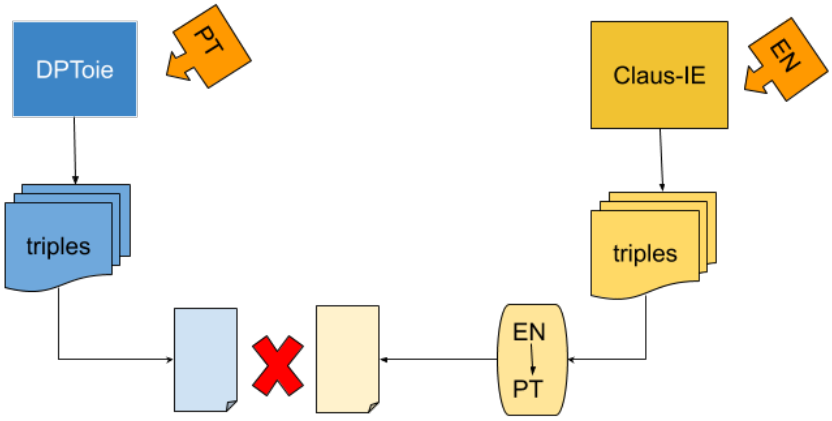

Figure 15. Experiment on relation extractions intersection for two languages: Portuguese (PT) and English (EN). (Authors' own).

It is important to note that, due to differences in the syntax of the languages, not all translations resulted in a valid extraction. For example, Portuguese has sentences without subjects, known as non-existent subject, when the main verb is impersonal. 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 a 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, unlike DPToie, we only maintained in the translated dataset those extractions containing both arguments. The translated set of extractions therefore contains 180 relations.

No intersection was observed between the two sets, i.e., no exact match was found between the relations of the two sets. Analyzing partial matches in 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, one coinciding in the second argument, and nine coinciding in a 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."* ("The deputy proposes for the Administrative Council of the European Food Authority twelve representatives, two of which representing the food industry.")

As a result, the extraction (*'The report', 'proposes', 'twelve representatives ...'*) obtained from the English sentence presents a mismatch with the one extracted from the Portuguese sentence (*'o senhor deputado', 'propõe', 'para o Conselho ...'*) ((*'The deputy', 'proposes', 'for the Administrative Council ...'*)), 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 can be seen 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'*) ((*'Phillip Whitehead', 'has managed', 'to expertly combine in a balanced report the many opinions which are around in our Parliament'*).).

Moreover, there are several extractions performed by Claus-IE for which no similar extraction has been performed by DPToie, i.e., no extraction of the DPToie system matched (exactly or partially) the same information. An example is 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. We believe this is evidence for our idea that the difference in linguistic structure in the sentences in different languages may help a multilingual system to extract more valid information from a sentence, since, in each language, it privileges a certain structuring of the information in the sentence.

Adjusting our analysis with 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. We conclude that exploring multilingual resources—in our case, a parallel corpus—has the potential to improve the performance of Open Information Extraction methods. In our experiments, we observed that using a bi-lingual parallel corpus and two mono-lingual Open IE systems, we were able to extract a broader set of valid relations from the same texts than the ones extracted by each system separately.

It is important to notice that the different systems can extract different information from the text not necessarily because of differences in the methods, but also inherent differences in the structure and use of the languages involved. 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). We believe this difference in length can indicate different strategies for the systems in choosing the predicate structure underlying the sentence—different structuring of the information in the sentence for each language.

5. Challenges and Opportunities

From our initial experiments, we believe that exploring multilingual resources has great potential to increase the performance of Open IE methods. 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 note a significant limitation of our approach to Multilingual Open IE concerning the transfer of knowledge from one language to another. The method investigated here relies heavily on Machine Translation Systems and, as much as Machine Translation has advanced in the last few years, these systems are not yet reliable, especially in an open context such as that of the Web as a Corpus.

Despite the fact that any domain-specific text can be considered input to a machine translation system, high accuracy translations are difficult to achieve due to potential domain-specific terminology. The problem of domain transfer between languages is not different 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 for the development of this area.

Notice that, in our experiments, as we chose to translate the extracted triples using a Machine Translation System, the quality of translation may suffer, as systems such as Google Translate—used in this work—use the sentence as contextual clues to compute the translation of each word and its place in the translated sentence. In this regard, a word alignment model, such as that used by [24], could be used to perform the translation between the languages. In this work, we chose to use Google Translate due to the fact that it is readily available for a great number of target languages—thus our approach can be replicated for different contexts—and we did not have the computational resources available to train the state-of-the-art MT methods to compute these word alignment models.

It is important to notice, however, that, in our experiments, we have observed that the Google Translator system performed adequately on the translation of the triples and the main disparities between the two sets of extractions were mainly due to differences in the original sentences for each language in the corpus, not in the translation process. It is not clear, however, that the system will have similar performance to other language pairs, especially considering other languages with fewer available resources than Portuguese.

Another critical issue regarding our results and conclusions lies in the evaluation process of 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. 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, a methodology that we follow in this work.

In fact, as Xavier et al. [22] point out, the notion of "relation" in the Open IE literature is vague and may have a different meaning for different authors. The evaluation methodology employed here may be biased by subjective interpretations of the human judges on the task at hand.

Furthermore, 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, such as Portuguese, or even skip the evaluation in these languages. We have identified 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 to improve Open IE evaluations and prove more reliable results.

A great limitation of the approach presented in this work concerns the existence of high-performance Open IE system for various languages.While Open IE systems have been proposed for several different languages in the last decade, e.g., for Chinese [62,63], Vietnamese [66], Portuguese [56,58,80], German [65,81] and Spanish [54], there is still a need for the development (or porting) of such systems for many languages in order for our approach to be able to exploit most of the available information in the Web as a Corpus context.

6. Conclusions and Future Directions

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

In our systematic mapping study, we identified some important research gaps in the area of Multilingual Open Information Extraction, such as the glaring lack of *benchmarks*, evaluation methodologies and a lack of applications to well-studied multilingual tasks in NLP, which could be exploited for extrinsic evaluations. It is also clear that, while multilingual Open Information Extraction tools and methods have been proposed, fewer explore multi- or cross-lingual information to improve their extractions.

From our empirical investigations, on the other hand, we conclude that exploring multilingual resources such as parallel or multilingual corpus can increase the performance of Open IE systems by identifying complementary information to be extracted, as pointed out by [25]. As we have discussed, however, the multilingual methods identified in the literature are still 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 Open Information Extraction aims to identify structured semantic information from unstructured sources, we believe multilingual Open IE systems may be used to further improve machine translation systems by providing semantic representations for the information as presented in different languages [33].

As for future directions, we remark a long way to transfer relation extractions from one language to another. Preliminary tools such as POS taggers, chunkers, and DP analyzers need to be improved for single languages to be incorporated into multilingual perspectives. In addition, the experiments presented in this work were limited to deal with texts in two languages. Exploring further the information expressed in other languages in the corpus may provide us with more insight on how different linguistic features influence the representation of information in a language, and how an Open IE system can benefit from these differences.

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

Author Contributions: Conceptualization, D.B.C., M.S. and C.C.X.; Data curation, M.S.; Methodology, D.B.C. and M.S.; Software, L.O.; Validation, D.B.C. and M.S.; Visualization, L.O.; Writing—original draft, D.B.C., M.S. and L.O.; Writing—review & editing, D.B.C., M.S. and C.C.X.

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:

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. The semantic web. *Sci. Am.* **2001**, *284*, 28–37. [CrossRef]
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: Hershey, PA, USA, 2011; pp. 205–227.
3. Fader, A.; Soderland, S.; Etzioni, O. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 1535–1545.
4. Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; Mausam, M. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*; AAAI Press: Menlo Park, CA, USA, 2011; Volume 1, pp. 3–10.
5. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data. In *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: Menlo Park, CA, USA, 2009; pp. 1003–1011.
6. Nguyen, T.H.; Grishman, R. Employing word representations and regularization for domain adaptation of relation extraction. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 68–74.
7. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 9–12 June 2008; pp. 1247–1250.
8. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledge Base. Available online: https://ai.google/research/pubs/pub42240.pdf (accessed on 1 June 2019).
9. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin, Germany, 2010; pp. 148–163.

10. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance multi-label learning for relation extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 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. In *International Semantic Web Conference*; Springer: Berlin, Germany, 2012; pp. 263–278.
12. Nguyen, T.V.T.; Moschitti, A. End-to-end relation extraction using distant supervision from external semantic repositories. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2, Portland, OR, USA, 19–24 June 2011; pp. 277–282.
13. Banko, M.; Cafarella, M.J.; Soderland, S.; Broadhead, M.; Etzioni, O. Open Information Extraction from the Web. In Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07), Hyderabad, India, 6–12 January 2007; pp. 2670–2676.
14. Kilgarriff, A.; Grefenstette, G. Web as corpus. In Proceedings of the Corpus Linguistics 2001, Lancaster, UK, 29 March–2 April 2001; pp. 342–344.
15. Yangarber, R.; Grishman, R.; Tapanainen, P.; Huttunen, S. Automatic acquisition of domain knowledge for information extraction. In Proceedings of the 18th conference on Computational linguistics-Volume 2, Saarbrücken, Germany, 31 July–4 August 2000; pp. 940–946.
16. Mooney, R.J.; Bunescu, R. Mining knowledge from text using information extraction. *ACM SIGKDD Explor. Newsl.* **2005**, *7*, 3–10. [CrossRef]
17. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with neural tensor networks for knowledge base completion. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 926–934.
18. Plank, B.; Moschitti, A. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 1498–1507.
19. Nguyen, T.H.; Grishman, R. Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, CO, USA, 5 June 2015; pp. 39–48.
20. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
21. 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; Volume 2, pp. 127–138.
22. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open information extraction based on lexical semantics. *J. Braz. Comput. Soc.* **2015**, *21*, 1–14. [CrossRef]
23. Wu, F.; Weld, D.S. Open Information Extraction Using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 118–127.
24. Faruqui, M.; Kumar, S. Multilingual Open Relation Extraction Using Cross-lingual Projection. *arXiv* **2015**, arXiv:1503.06450.
25. Steinberger, R. A survey of methods to ease the development of highly multilingual text mining applications. *Lang. Resour. Eval.* **2012**, *46*, 155–176. [CrossRef]
26. Koehn, P. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the Tenth Machine Translation Summit, Phuket, Thailand, 12–16 September 2005; pp. 79–86.
27. Eisele, A.; Chen, Y. *MultiUN: A Multilingual Corpus from United Nation Documents*; LREC: Stockholm, Sweeden, 2010.
28. Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufis, D.; Varga, D. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *arXiv* **2006**, arXiv:cs/0609058.
29. Déjean, H.; Gaussier, É.; Sadat, F. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In Proceedings of the 19th international conference on Computational linguistics-Volume 1, Taipei, Taiwan, 24 August–1 September 2002; pp. 1–7.
30. Yang, C.C.; Luk, J. Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 671–682. [CrossRef]

31. Xu, J.; Weischedel, R.; Nguyen, C. Evaluating a probabilistic model for cross-lingual information retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, 9–12 September 2001; pp. 105–110.
32. Ferrández, S.; Toral, A.; Ferrández, O.; Ferrández, A.; Munoz, R. Applying Wikipedia's multilingual knowledge to cross–lingual question answering. In *International Conference on Application of Natural Language to Information Systems*; Springer: Berlin, Germany, 2007; pp. 352–363.
33. Mihalcea, R.; Simard, M. Parallel texts. *Nat. Lang. Eng.* **2005**, *11*, 239–246. [CrossRef]
34. Yarowsky, D.; Ngai, G.; Wicentowski, R. Inducing multilingual text analysis tools via robust projection across aligned corpora. In Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, USA, 18–21 March 2001; pp. 1–8.
35. Bentivogli, L.; Pianta, E. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Nat. Lang. Eng.* **2005**, *11*, 247–261. [CrossRef]
36. Hwa, R.; Resnik, P.; Weinberg, A.; Cabezas, C.; Kolak, O. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.* **2005**, *11*, 311–325. [CrossRef]
37. Bel, N.; Koster, C.H.; Villegas, M. Cross-lingual text categorization. In *International Conference on Theory and Practice of Digital Libraries*; Springer: Berlin, Germany, 2003; pp. 126–139.
38. Bering, C.; Drozdzynski, W.; Erbach, G.; Guasch, C.; Homola, P.; Lehmann, S.; Li, H.; Krieger, H.U.; Piskorski, J.; Schäfer, U.; et al. Corpora and evaluation tools for multilingual named entity grammar development. In Proceedings of the Multilingual Corpora Workshop at Corpus Linguistics, Lancaste, UK, 28–31 May 2003; pp. 42–52.
39. Boyd-Graber, J.; Blei, D.M. Multilingual topic models for unaligned text. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009; pp. 75–82.
40. Hassan, S.; Mihalcea, R. Cross-lingual semantic relatedness using encyclopedic knowledge. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 1192–1201.
41. Al-Rfou, R.; Perozzi, B.; Skiena, S. Polyglot: Distributed Word Representations for Multilingual Nlp. *arXiv* **2013**, arXiv:1307.1662.
42. Lin, Y.; Liu, Z.; Sun, M. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 34–43.
43. Verga, P.; Belanger, D.; Strubell, E.; Roth, B.; McCallum, A. Multilingual relation extraction using compositional universal schema. *arXiv* **2015**, arXiv:1511.06396 .
44. Zhang, S.; Duh, K.; Van Durme, B. Mt/ie: Cross-lingual open information extraction with neural sequence-to-sequence models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 64–70.
45. Wang, X.; Han, X.; Lin, Y.; Liu, Z.; Sun, M. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1156–1166.
46. Gamallo, P.; Garcia, M.; Fernandez-Lanza, S. Dependency-based Open Information Extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 10–18.
47. Souza, E.N.P.; Claro, D.B. Extração de Relações utilizando Features Diferenciadas para Português. *Linguamática* **2014**, *6*, 57–65.
48. Etzioni, O.; Banko, M.; Soderland, S.; Weld, D.S. Open information extraction from the web. *Commun. ACM* **2008**, *51*, 68–74. [CrossRef]
49. Schmitz, M.; Bart, R.; Soderland, S.; Etzioni, O. Open Language Learning for Information Extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012; pp. 523–534.
50. Del Corro, L.; Gemulla, R. ClausIE: Clause-based Open Information Extraction. In Proceedings of the 22Nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 355–366. [CrossRef]

51. Bast, H.; Haussmann, E. Open information extraction via contextual sentence decomposition. In Proceedings of the 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, 16–18 September 2013; pp. 154–159.
52. Bast, H.; Haussmann, E. More Informative Open Information Extraction via Simple Inference. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval—Volume 8416*; Springer: New York, NY, USA, 2014; pp. 585–590. [CrossRef]
53. Gashteovski, K.; Gemulla, R.; Del Corro, L. MinIE: Minimizing Facts in Open Information Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 2630–2640.
54. Gamallo, P.; Garcia, M. Multilingual Open Information Extraction. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, 8–11 September 2015*; Pereira, F., Machado, P., Costa, E., Cardoso, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 711–722. [CrossRef]
55. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open Information Extraction based on lexical-syntactic patterns. In Proceedings of the 2013 Brazilian Conference on Intelligent Systems (BRACIS), Fortaleza, Brazil, 19–24 October 2013; pp. 189–194.
56. Sena, C.F.L.; Glauber, R.; Claro, D.B. Inference Approach to Enhance a Portuguese Open Information Extraction. In *Proceedings of the 19th International Conference on Enterprise Information Systems—Volume 1: ICEIS*; INSTICC, ScitePress: Porto, Portugal, 2017; pp. 442–451. [CrossRef]
57. Sena, C.F.L.; Claro, D.B. InferPortOIE: A Portuguese Open Information Extraction system with inferences. *Nat. Lang. Eng.* **2019**, *25*, 287–306. [CrossRef]
58. de Oliveira, L.S.; Glauber, R.; Claro, D.B. DependentIE: An Open Information Extraction system on Portuguese by a Dependence Analysis. In Proceedings of the Encontro Nacional de Inteligência Artificial e Computacional, Uberlândia, Brasil, 2–5 October 2017.
59. de Oliveira, L.S.; Claro, D.B. DptOIE: A Portuguese Open Information Extraction system based on Dependency Analysis. *Comput. Speech Lang.* **2019**, under review.
60. 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. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 12–16.
61. Qiu, L.; Zhang, Y. ZORE: A Syntax-based System for Chinese Open Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1870–1880. [CrossRef]
62. Xu, J.; Gan, L.; Deng, L.; Wang, J.; Yan, Z. Dependency parsing based Chinese open relation extraction. In Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, 19–20 December 2015; Volume 1, pp. 552–556. [CrossRef]
63. Wang, Y.; Zhou, G.; Tian, F.; Nan, Y.; Ma, J. GCORE: A Gravitation-Based Approach for Chinese Open Relation. In Proceedings of the 2015 International Conference on Computer Science and Mechanical Automation (CSMA), Hangzhou, China, 23–25 October 2015; pp. 86–91. [CrossRef]
64. Wu, X.; Wu, B. The CRFs-Based Chinese Open Entity Relation Extraction. In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26–29 June 2017; pp. 405–411. [CrossRef]
65. Bassa, A.; Kroll, M.; Kern, R. GerIE-An Open Information Extraction System for the German Language. *J. Univers. Comput. Sci.* **2018**, *24*, 2–24.
66. Truong, D.; Vo, D.T.; Nguyen, U.T. Vietnamese Open Information Extraction. In Proceedings of the Eighth International Symposium on Information and Communication Technology, Singapore, 15–18 May 2017; pp. 135–142. [CrossRef]
67. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic Mapping Studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*; BCS Learning & Development Ltd.: Swindon, UK, 2008; pp. 68–77.
68. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* **2015**, *64*, 1–18. [CrossRef]
69. Glauber, R.; Claro, D.B. A Systematic Mapping Study on Open Information Extraction. *Expert Syst. Appl.* **2018**. [CrossRef]

70. Bender, E.M.; Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 587–604. [CrossRef]
71. Garcia, M.; Gamallo, P. Entity-centric coreference resolution of person entities for open information extraction. *Proces. Leng. Nat.* **2014**, *53*, 25–32.
72. Nunes, T.; Schwabe, D. Building Distant Supervised Relation Extractors. In Proceedings of the 2014 IEEE International Conference on Semantic Computing, Newport Beach, CA, USA, 16–18 June 2014; pp. 44–51.
73. Akbik, A.; Danilevsky, M.; Kbrom, Y.; Li, Y.; Zhu, H. Multilingual information extraction with PolyglotIE. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, 11–16 December 2016; pp. 268–272.
74. Duc, N.T.; Bollegala, D.; Ishizuka, M. Cross-language latent relational search between japanese and english languages using a web corpus. *ACM Trans. Asian Lang. Inf. Process.* **2012**, *11*, 11. [CrossRef]
75. Vulić, I.; Moens, M.F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 363–372.
76. Upadhyay, S.; Faruqui, M.; Dyer, C.; Roth, D. Cross-lingual models of word embeddings: An empirical comparison. *arXiv* **2016**, arXiv:1604.00425.
77. Xiao, M.; Guo, Y. Distributed word representation learning for cross-lingual dependency parsing. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Baltimore, MD, USA, 26–27 June 2014; pp. 119–129.
78. Täckström, O.; McDonald, R.; Uszkoreit, J. Cross-lingual word clusters for direct transfer of linguistic structure. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, QC, Canada, 3–8 June 2012; pp. 477–487.
79. Yang, Z.; Salakhutdinov, R.; Cohen, W. Multi-Task Cross-Lingual Sequence Tagging from Scratch. *arXiv* **2016**, arXiv:1603.06270.
80. de Abreu, S.C.; Vieira, R. Relp: Portuguese open relation extraction. *KO KNOWLEDGE ORGANIZATION* **2017**, *44*, 163–177. [CrossRef]
81. Falke, T.; Stanovsky, G.; Gurevych, I.; Dagan, I. Porting an open information extraction system from english to german. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 892–898.