

Article

Multilingual Open Information Extraction: Challenges and Opportunities

Daniela Barreiro Claro^{1,*} , Marlo Souza¹ , Clarissa Castellã Xavier² and Leandro Oliveira¹

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

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

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

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

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

53 As Banko *et al.* [13] discuss in their seminal work, however, in an open context such as that of the
54 Web as a Corpus [14], it is not feasible to enumerate all potential relations of interest for extraction.
55 For specific domains of applications, the existence of such knowledge bases is not guaranteed. In
56 fact, a common application of Information Extraction systems lies on the creation or completion of
57 knowledge bases, aiming to acquire knowledge from existing textual resources in a domain of interest
58 [15–17].

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

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

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

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

85 Although nearly half of the Web's content is written in English - W3Techs¹ estimates that around
86 54% of the 10 million most accessed websites were written in English - the percentage of contents in
87 other languages has been increasing in the last decades. As such, it is of great importance to developing
88 robust Open IE methods for different languages.

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

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

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

107 This paper is organized as follows. Section 2 presents the definitions of Open IE area, giving
108 some examples. Section 3 describes our systematic mapping study and organize our results. Section
109 4 presents an experiment to handle transferable knowledge acquisition in two languages. Section 5
110 discusses some challenges and points out some research directions, and finally, section 6 concludes our
111 work and point out some envisioning work.

112 2. Open Information Extraction

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

$$117 \text{triple} = (\text{arg1}, \text{rel}, \text{arg2}) \quad (1)$$

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

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

126 The accuracy of Open IE is still low compared to the Traditional IE, since the number of invalid
127 extractions is high. An extraction is said to be invalid or incorrect when one or more elements of
the extracted triplet $t = (\text{arg1}, \text{rel}, \text{arg2})$ does not correspond to the information contained in the

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

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

135 Another case occurs when the extracted triple is valid but uninformative. The extraction is said
 136 to be uninformative when the semantic relation expressed by the triple does not correspond to the
 137 information presented in the sentence. Considering Table 1, for instance, it is simple to notice that the
 138 relation extraction from lines 1 and 2, although still maintain a binary format, they are uninformative.
 139 In this example, a semantic meaning presented in the extraction does not represent what is written in
 140 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)

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

145 2.1. Open IE with different languages

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

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

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

162 From the standpoint of the Portuguese language, the first proposed methods were: DepOE
 163 [37] and ArgOE [45]. Both methods are multilingual and perform extractions for texts in English,
 164 Spanish, and Galician as well as Portuguese. In addition, LSOE [46] uses morphosyntactic patterns
 165 and also stands out by extracting facts in an unsupervised way. Similar to this approach is the method
 166 (nicknamed, SGC_2017) [47]. SGC_2017 proposes an adaptation of the ReVerb [36] to the Portuguese
 167 language and a syntactic restriction to identify nominal phrases. SGC_2017 presents an inferential
 168 approach to extract new facts using a binary SVM classifier between the transitive and symmetric
 169 classes. Next, we have the InferPORToie [48] that enhances their inferential approach and allows

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

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

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

189 3. Multilingual Open IE: a systematic mapping study

190 A Systematic Mapping Study (SMS) provides an overview of the scope of the area and allows
191 the discovery of research gaps, forums, relevant authors, research groups, and trends [58,59]. SMS is
192 organized into three groups of activities: planning, conducting and reporting [59]. The first group
193 aims to identify the reasons for this study, followed by the research questions and then definitions of
194 the protocol. The second group of activities organizes the selection of primary studies, the extraction,
195 and the data summarization. Finally, the third group defines the threats during the study activities.
196 Multilingual Open IE is a new research topic with the first published work, that we are aware of, in
197 2012 [37]. Although some secondary studies have been published, SMS discovers quantitative and
198 qualitative data on primary studies not yet presented in those secondary studies. While a traditional
199 review can present the bias of a group or researcher, SMS aims to determine the gaps and to observe
200 relevant aspects of the area diminishing (or eliminating) the bias vision. Our study begins with a
201 general question about the state of the art in Multilingual Open IE:

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

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

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

213 The next step in planning is to determine the search engine to retrieve the primary studies. The
214 search method to find primary studies is carried by automatically search in electronic databases through
215 a set of keywords. As discussed in [60], the term "information extraction" was avoided because of
216 a large number of recovery results. Considering semantic terms from multilingual approaches, we

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

- 220 • "multi lingual" OR "crosslingual" OR "multilingual" OR "multi-lingual"
- 221 • "open information extraction"
- 222 • "relation extraction"

223 We performed our search into five databases: Science Direct², IEEE Xplore³, ACM Digital Library⁴,
224 Scopus⁵, and Google Scholar⁶ with the search strings described.

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

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

235 We choose to remove the studies written in languages other than English due to the fact that
236 English written papers have a broad audience in academic community. Texts such as MSc or Ph.D.
237 reports, technical reports, or any content which were not evaluated by a program committee were also
238 removed. Review studies (secondary studies) were removed as well, since we are concerned only with
239 primary studies. An important observation is that some studies use the terms Open IE and relation
240 extraction, but do not deal with those type of systems. In this SMS we consider either the paper which
241 performs traditional IE, with some insights about Open IE systems. This was considered because of the
242 restriction of the number of papers which cover multilingual approaches. However, almost all papers
243 which do not deal with Traditional IE nor Open IE, but refers to Named Entity, Information Retrieval
244 were removed by F5 filter. After executing the string query, the filtering step starts. The removal of
245 primary studies was conducted in two stages. In the first stage, we read the abstracts of each paper to
246 identify occurrences outside the scope of our SMS. In the second stage, with the remaining papers,
247 we filtered based on a paper full reading. Figure 1 presents the values of each filter applied to each
248 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>

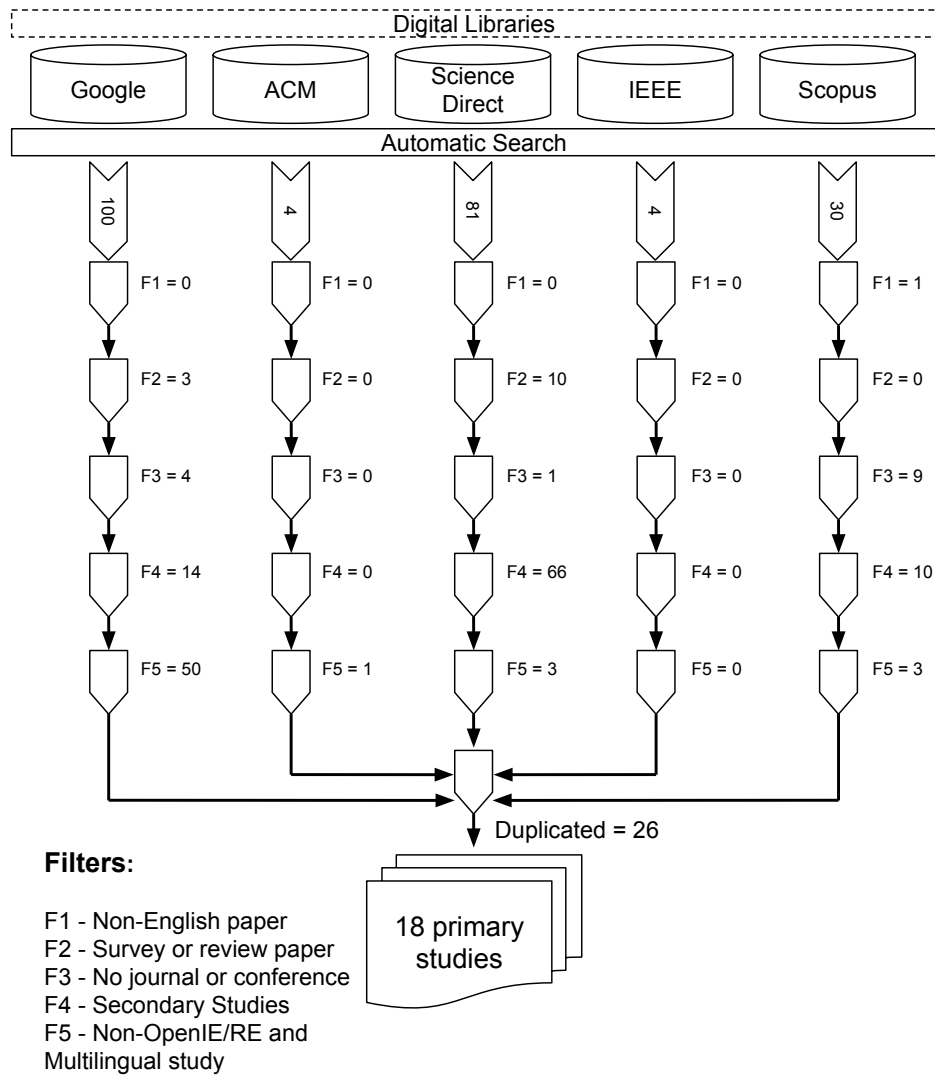


Figure 1. Filters applied in SMS process

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

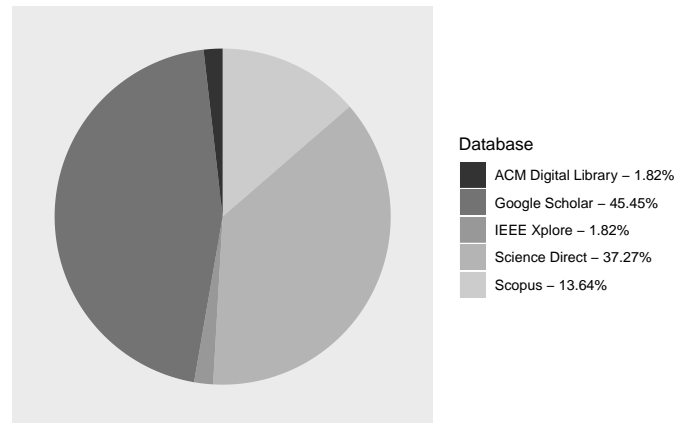


Figure 2. Percentage of each database in primary studies selection.

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

255 We initiate the analysis phase reading each selected paper and filling a form with such structure
 256 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

257 After obtaining the data from each selected primary study and filling the form (Table 2), the next
 258 step was to summarize the data to recover useful information. The objective of this step was to answer
 259 the RQ set following the same order.

260 3.1. Answer to RQ1: What are the sources of publications in Multilingual Open IE area?

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

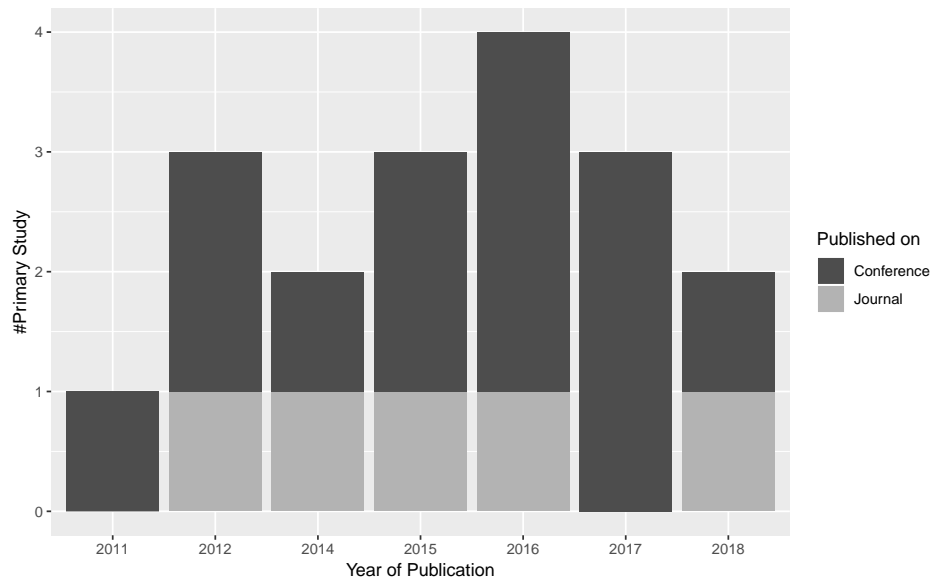


Figure 3. Distribution over the years of selected primary studies per paper type (journal and conference)

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

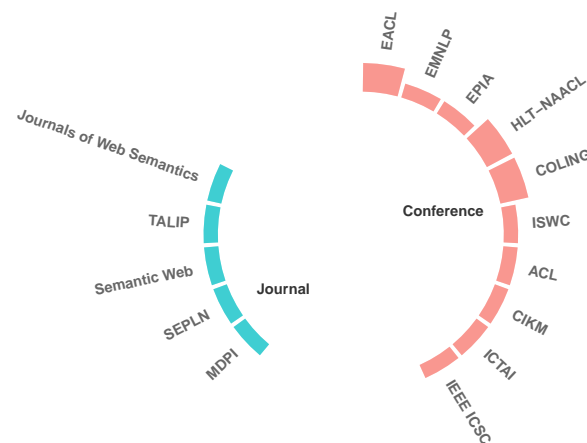


Figure 4. Distribution of the selected papers on Journals and Conferences

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

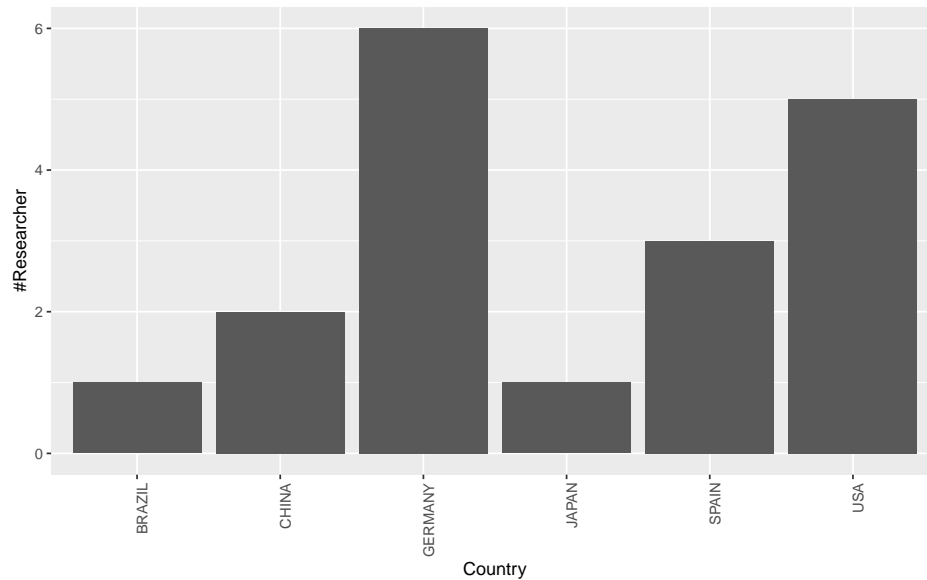


Figure 5. Distribution of the countries of the researchers in primary studies.

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

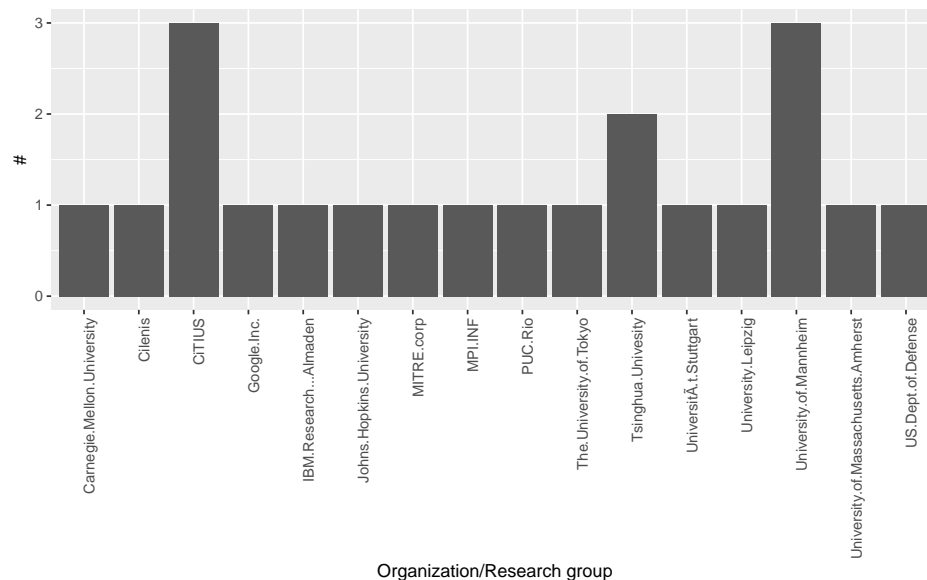


Figure 6. Distribution of the affiliation authors in primary studies.

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

278 3.2. Answer to RQ2: What are the types of contributions made by Multilingual Open IE studies?

279 Figure 7 depicts the contribution types of Multilingual Open IE methods. It is noteworthy that the
 280 *METHOD* group corresponds to studies with novel methods or approaches to the task of “Multilingual
 281 Open IE.” This kind of research tries to evolve the state of the art with new ideas comparing its work
 282 with other systems, in spite of the fact that Multilingual Open IE is a young research area that lacks
 283 benchmark materials. Some studies also produce contributions in *RESOURCE* and *TOOL*. *RESOURCE*
 284 represents studies that have created datasets for evaluation or other resources for training or testing of
 285 Open IE systems. There is also a relevant portion of studies that present a new approach to Multilingual
 286 Open IE through a *TOOL*. In these studies, authors present, compare and provide an implementation
 287 of their work.

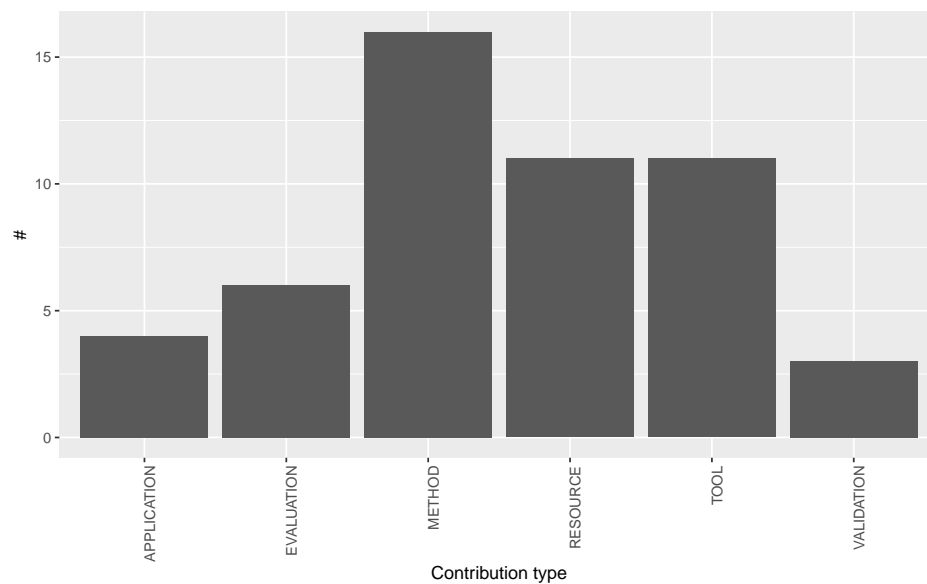


Figure 7. Number of the type of contributions.

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

291 3.3. Answer to RQ3: What are the types of applications made by Multilingual Open IE studies?

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

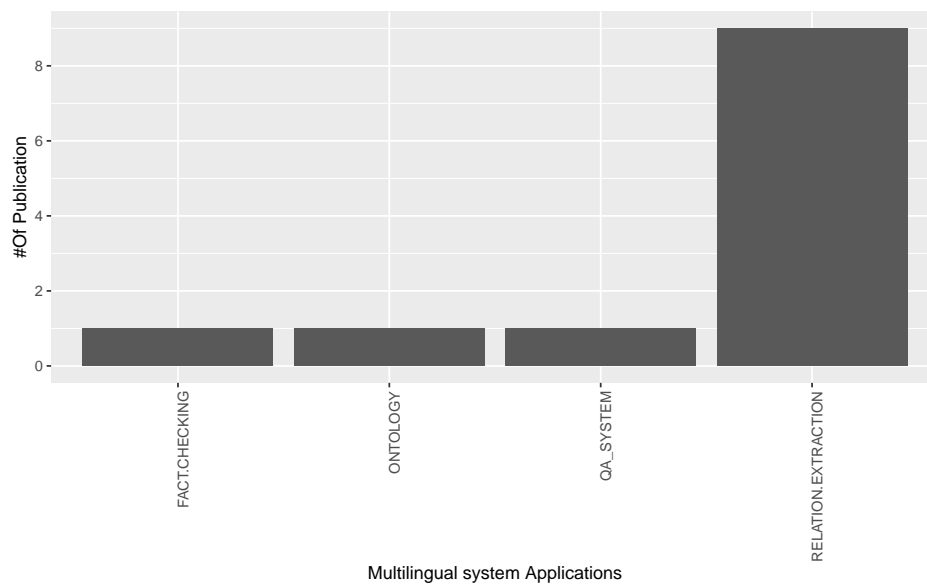


Figure 8. Number of primary studies per type of applications.

295 **3.4. Answer to RQ4: What are the available Multilingual Open IE datasets?**

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

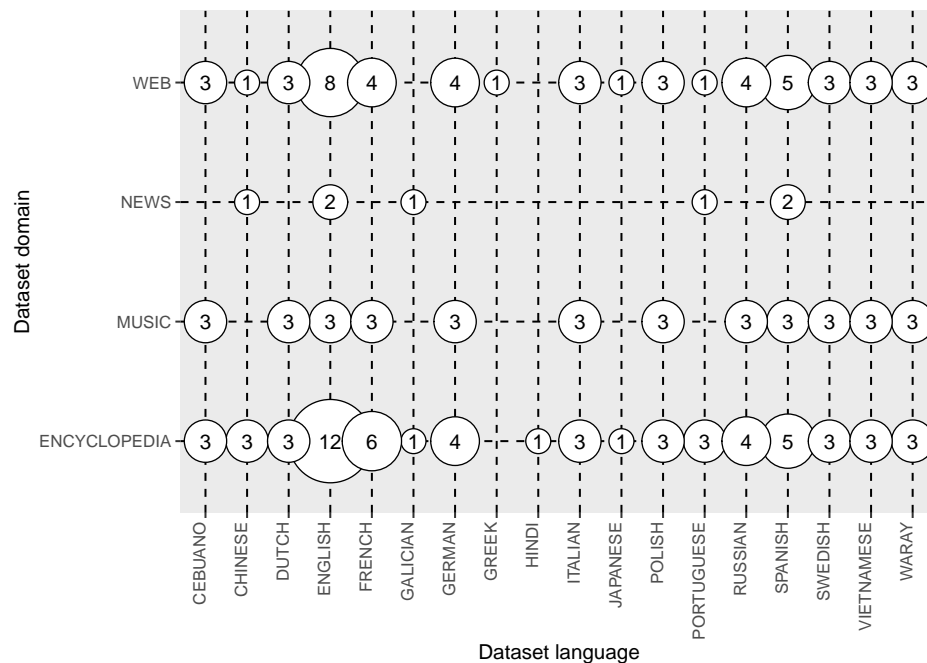


Figure 9. Dataset Language per Domain.

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

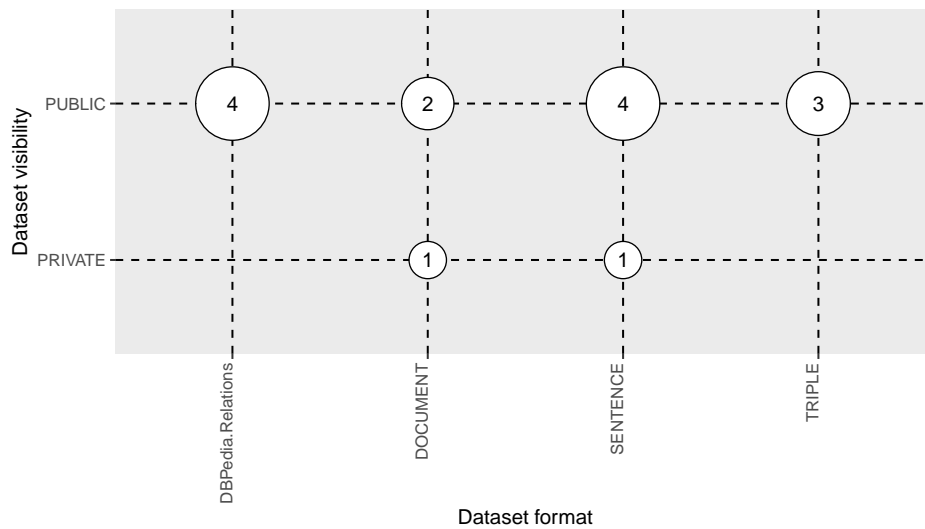


Figure 10. Map of the dataset format and visibility from primary studies. For X axis the format file of the dataset. For Y axis the visibility of datasets.

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

307 3.5. Answer to RQ5: What are the tools used in Multilingual Open IE systems?

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

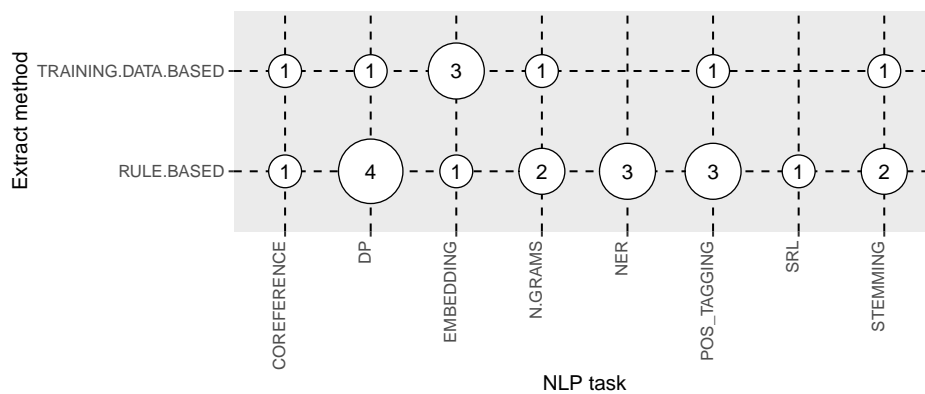


Figure 11. Map of natural language processing tasks and extraction method approach applied in Open IE studies. For X axis the NLP taggers. For the Y axis the extraction method approach.

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

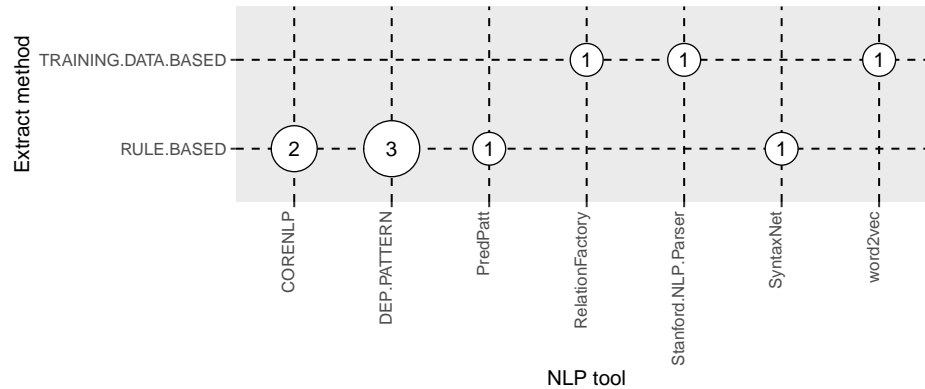


Figure 12. NLP Tool per extract method.

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

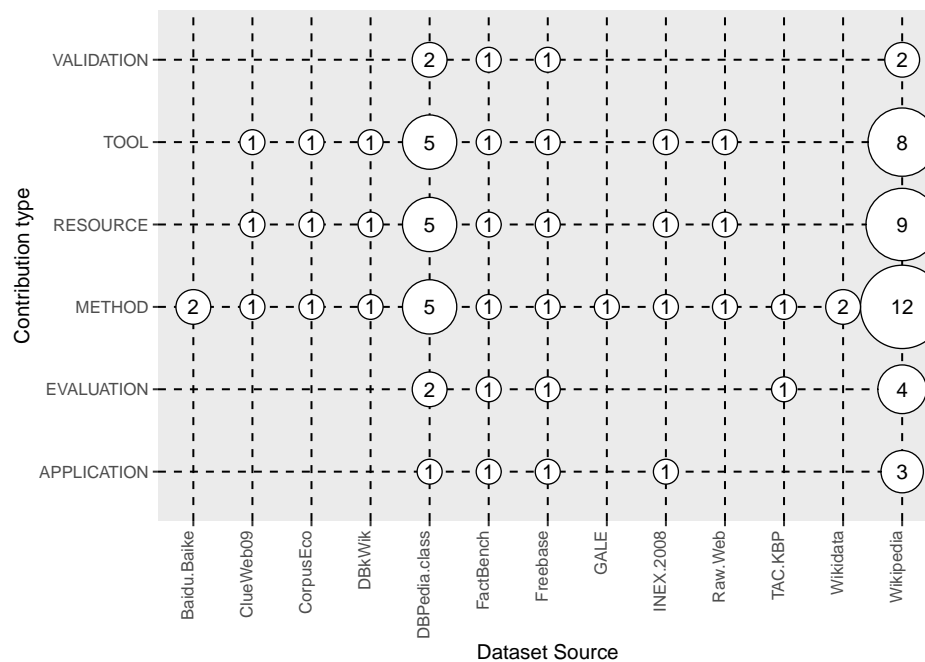


Figure 13. Map of Dataset Sources and types of contribution identified. For X axis the Dataset Sources bases. For Y axis the type of contributions.

324 3.6. Answer to RQ6: How are Multilingual Open IE systems evaluated?

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

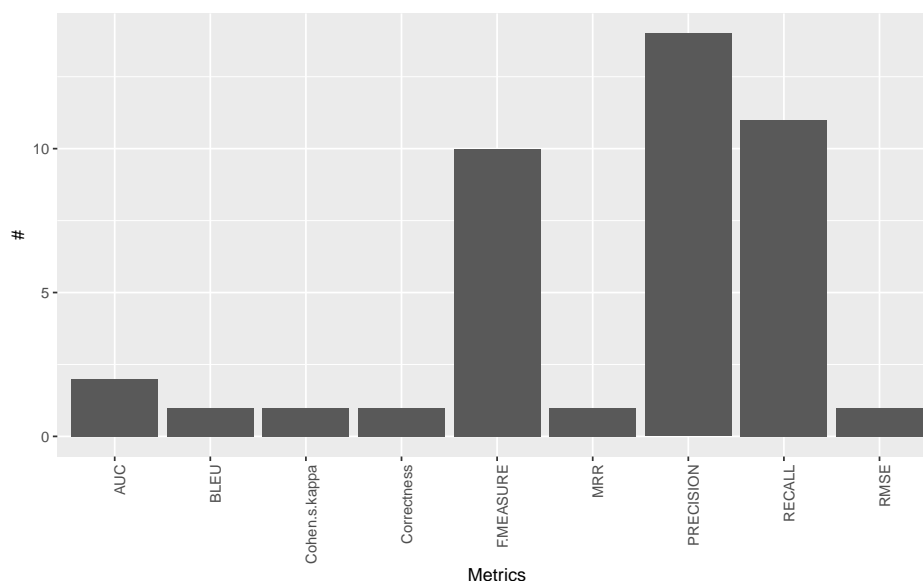


Figure 14. Evaluation Metrics.

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

332 4. Some experiments on transferable knowledge in Multilingual Open IE

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

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

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

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

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

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

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

364 4.1. Dataset

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

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

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

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

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

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

390 4.2. Experiment : analyzing cross-lingual extraction complementarity

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

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

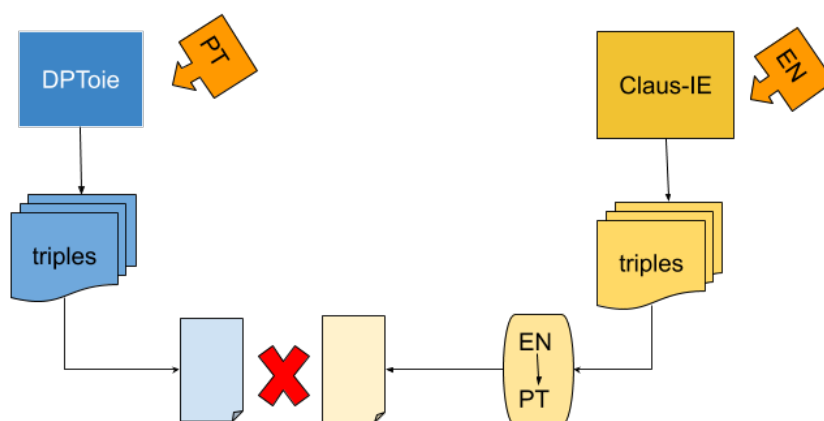


Figure 15. Experiment on relation extractions intersection for two languages: PT and EN

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

401 *“It is naturally important that food can also flow freely”*

402 which have the pronoun ‘it’ as subject, can be translated to:

403 *“É naturalmente importante que os produtos alimentares possam também circular livremente.”*

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

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

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

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

418 *“The report proposes twelve representatives for the board of the new Food Authority, two of whom would be
 419 representatives of the food industry.”*

420 is aligned with

421 *“O senhor deputado propõe para o Conselho de Administração da Autoridade Alimentar Europeia doze
 422 representantes, dois dos quais em representação da indústria alimentar.”⁷*

423 As a result, the extraction (‘o relatório’, ‘propõe’, ‘12 representantes...’) obtained from the English
 424 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.”

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

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

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

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

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

436 For example, in the sentence

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

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

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

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

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

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

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

458 5. Challenges and Opportunities

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

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

465 Another critical issue is the evaluation process on Open IE systems. Although the most common
466 metrics used to evaluate IE systems are *Precision*, *Recall*, and *F-Measure*, it is usually not feasible to
467 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*').

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

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

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

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

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

490 6. Conclusions and Future Directions

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

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

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

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

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

508 .

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

512 Abbreviations

513 The following abbreviations are used in this manuscript:

514

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
515 RQ	Research Question
QA	Querying Answering
RDF	Resource Definition Framework
DP	Dependency Parser
SRL	Semantic Role Labeling
NER	Named Entity Recognition

516 References

- 517 1. Berners-Lee, T.; Hendler, J.; Lassila, O.; others. The semantic web. *Scientific american* **2001**, *284*, 28–37.
- 518 2. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic services, interoperability and*
519 *web applications: emerging concepts*; IGI Global, 2011; pp. 205–227.
- 520 3. Fader, A.; Soderland, S.; Etzioni, O. Identifying relations for open information extraction. Proceedings of
521 the Conference on Empirical Methods in Natural Language Processing. Association for Computational
522 Linguistics, 2011, pp. 1535–1545.
- 523 4. Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; Mausam, M. Open information extraction: The second
524 generation. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence.
525 AAAI Press, 2011, Vol. 1, pp. 3–10.
- 526 5. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant supervision for relation extraction without labeled data.
527 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International
528 Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for
529 Computational Linguistics, 2009, pp. 1003–1011.
- 530 6. Nguyen, T.H.; Grishman, R. Employing word representations and regularization for domain adaptation
531 of relation extraction. Proceedings of the 52nd Annual Meeting of the Association for Computational
532 Linguistics (Volume 2: Short Papers), 2014, Vol. 2, pp. 68–74.
- 533 7. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database
534 for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD international conference on
535 Management of data. AcM, 2008, pp. 1247–1250.
- 536 8. Vrandečić, D.; Krötzsch, M. Wikidata: a free collaborative knowledge base **2014**.
- 537 9. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. Joint
538 European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2010, pp.
539 148–163.
- 540 10. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance multi-label learning for relation
541 extraction. Proceedings of the 2012 joint conference on empirical methods in natural language processing
542 and computational natural language learning. Association for Computational Linguistics, 2012, pp.
543 455–465.
- 544 11. Krause, S.; Li, H.; Uszkoreit, H.; Xu, F. Large-scale learning of relation-extraction rules with distant
545 supervision from the web. International Semantic Web Conference. Springer, 2012, pp. 263–278.
- 546 12. Nguyen, T.V.T.; Moschitti, A. End-to-end relation extraction using distant supervision from external
547 semantic repositories. Proceedings of the 49th Annual Meeting of the Association for Computational
548 Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational
549 Linguistics, 2011, pp. 277–282.
- 550 13. Banko, M. Open Information Extraction for the Web. PhD thesis, Seattle, WA, USA, 2009. AAI3370456.
- 551 14. Kilgarrieff, A.; Grefenstette, G. Web as corpus. Proceedings of Corpus Linguistics 2001. Corpus Linguistics.
552 Readings in a Widening Discipline, 2001, pp. 342–344.
- 553 15. Yangarber, R.; Grishman, R.; Tapanainen, P.; Huttunen, S. Automatic acquisition of domain knowledge
554 for information extraction. Proceedings of the 18th conference on Computational linguistics-Volume 2.
555 Association for Computational Linguistics, 2000, pp. 940–946.

- 556 16. Mooney, R.J.; Bunescu, R. Mining knowledge from text using information extraction. *ACM SIGKDD*
557 *explorations newsletter* **2005**, *7*, 3–10.
- 558 17. Socher, R.; Chen, D.; Manning, C.D.; Ng, A. Reasoning with neural tensor networks for knowledge base
559 completion. *Advances in neural information processing systems*, 2013, pp. 926–934.
- 560 18. Plank, B.; Moschitti, A. Embedding semantic similarity in tree kernels for domain adaptation of relation
561 extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*
562 *(Volume 1: Long Papers)*, 2013, Vol. 1, pp. 1498–1507.
- 563 19. Nguyen, T.H.; Grishman, R. Relation extraction: Perspective from convolutional neural networks.
564 *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015,
565 pp. 39–48.
- 566 20. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE*
567 *transactions on pattern analysis and machine intelligence* **2013**, *35*, 1798–1828.
- 568 21. Banko, M.; Etzioni, O.; Center, T. The Tradeoffs Between Open and Traditional Relation Extraction. *ACL*;
569 *Association for Computational Linguistics*: Stroudsburg, PA, USA, 2008; Vol. 8, pp. 28–36.
- 570 22. Li, H.; Bollegala, D.; Matsuo, Y.; Ishizuka, M. Using graph based method to improve bootstrapping relation
571 extraction. In *Computational Linguistics and Intelligent Text Processing*; Springer: Berlin, Germany, 2011;
572 Vol. 2, pp. 127–138.
- 573 23. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open information extraction based on lexical semantics. *Journal of*
574 *the Brazilian Computer Society* **2015**, *21*, 1–14.
- 575 24. Wu, F.; Weld, D.S. Open Information Extraction Using Wikipedia. *Proceedings of the 48th Annual Meeting*
576 *of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg,
577 PA, USA, 2010; ACL '10, pp. 118–127.
- 578 25. Faruqui, M.; Kumar, S. Multilingual Open Relation Extraction Using Cross-lingual Projection. *arXiv*
579 *preprint arXiv:1503.06450* **2015**, *abs/1503.06450*, 1351–1356.
- 580 26. Steinberger, R. A survey of methods to ease the development of highly multilingual text mining
581 applications. *Language Resources and Evaluation* **2012**, *46*, 155–176.
- 582 27. Bel, N.; Koster, C.H.; Villegas, M. Cross-lingual text categorization. *International Conference on Theory*
583 *and Practice of Digital Libraries*. Springer, 2003, pp. 126–139.
- 584 28. Bering, C.; Drozdowski, W.; Erbach, G.; Guasch, C.; Homola, P.; Lehmann, S.; Li, H.; Krieger, H.U.;
585 Piskorski, J.; Schäfer, U.; others. Corpora and evaluation tools for multilingual named entity grammar
586 development. *Proceedings of Multilingual Corpora Workshop at Corpus Linguistics*, 2003, pp. 42–52.
- 587 29. Boyd-Graber, J.; Blei, D.M. Multilingual topic models for unaligned text. *Proceedings of the Twenty-Fifth*
588 *Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 75–82.
- 589 30. Hassan, S.; Mihalcea, R. Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings*
590 *of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 1192–1201.
- 591 31. Al-Rfou, R.; Perozzi, B.; Skiena, S. Polyglot: Distributed word representations for multilingual nlp. *arXiv*
592 *preprint arXiv:1307.1662* **2013**.
- 593 32. Lin, Y.; Liu, Z.; Sun, M. Neural relation extraction with multi-lingual attention. *Proceedings of the 55th*
594 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp.
595 34–43.
- 596 33. Verga, P.; Belanger, D.; Strubell, E.; Roth, B.; McCallum, A. Multilingual relation extraction using
597 compositional universal schema. *arXiv preprint arXiv:1511.06396* **2015**.
- 598 34. Zhang, S.; Duh, K.; Van Durme, B. Mt/ie: Cross-lingual open information extraction with neural
599 sequence-to-sequence models. *Proceedings of the 15th Conference of the European Chapter of the*
600 *Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 64–70.
- 601 35. Wang, X.; Han, X.; Lin, Y.; Liu, Z.; Sun, M. Adversarial multi-lingual neural relation extraction. *Proceedings*
602 *of the 27th International Conference on Computational Linguistics*, 2018, pp. 1156–1166.
- 603 36. Fader, A.; Soderland, S.; Etzioni, O. Identifying Relations for Open Information Extraction. *Proceedings of*
604 *the Conference on Empirical Methods in Natural Language Processing*; Association for Computational
605 *Linguistics*: Stroudsburg, PA, USA, 2011; EMNLP '11, pp. 1535–1545.
- 606 37. Gamallo, P.; Garcia, M.; Fernandez-Lanza, S. Dependency-based Open Information Extraction. *Proceedings*
607 *of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*; Association for
608 *Computational Linguistics*: Stroudsburg, PA, USA, 2012; ROBUST-SUP '12, pp. 10–18.

- 609 38. Souza, E.N.P.; Claro, D.B. Extração de Relações utilizando Features Diferenciadas para Português.
610 *Linguamática* **2014**, *6*, 57–65.
- 611 39. Etzioni, O.; Banko, M.; Soderland, S.; Weld, D.S. Open information extraction from the web. *Communications*
612 *of the ACM* **2008**, *51*, 68–74.
- 613 40. Mausam.; Schmitz, M.; Bart, R.; Soderland, S.; Etzioni, O. Open Language Learning for Information
614 Extraction. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing
615 and Computational Natural Language Learning; Association for Computational Linguistics: Stroudsburg,
616 PA, USA, 2012; EMNLP-CoNLL '12, pp. 523–534.
- 617 41. Del Corro, L.; Gemulla, R. ClausIE: Clause-based Open Information Extraction. Proceedings of the 22Nd
618 International Conference on World Wide Web; ACM: New York, NY, USA, 2013; WWW '13, pp. 355–366.
619 doi:10.1145/2488388.2488420.
- 620 42. Bast, H.; Haussmann, E. Open information extraction via contextual sentence decomposition. ICSC; IEEE,
621 Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on:rvine, CA, USA, 2013; pp.
622 154–159.
- 623 43. Bast, H.; Haussmann, E. More Informative Open Information Extraction via Simple Inference.
624 Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval
625 - Volume 8416; Springer-Verlag New York, Inc.: New York, NY, USA, 2014; ECIR 2014, pp. 585–590.
626 doi:10.1007/978-3-319-06028-6_61.
- 627 44. Gashteovski, K.; Gemulla, R.; Del Corro, L. MinIE: Minimizing Facts in Open Information Extraction.
628 Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association
629 for Computational Linguistics, 2017, pp. 2630–2640.
- 630 45. Gamallo, P.; Garcia, M., Multilingual Open Information Extraction. In *Progress in Artificial Intelligence: 17th*
631 *Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*;
632 Pereira, F.; Machado, P.; Costa, E.; Cardoso, A., Eds.; Springer International Publishing: Cham, 2015; pp.
633 711–722. doi:10.1007/978-3-319-23485-4_72.
- 634 46. Xavier, C.C.; de Lima, V.L.S.; Souza, M. Open Information Extraction based on lexical-syntactic patterns.
635 Intelligent Systems (BRACIS), 2013 Brazilian Conference on. IEEE, 2013, pp. 189–194.
- 636 47. Sena, C.F.L.; Glauber, R.; Claro, D.B. Inference Approach to Enhance a Portuguese Open Information
637 Extraction. Proceedings of the 19th International Conference on Enterprise Information Systems - Volume
638 1: ICEIS; INSTICC, ScitePress: Porto, Portugal, 2017; pp. 442–451. doi:10.5220/0006338204420451.
- 639 48. Sena, C.F.L.; Claro, D.B. InferPortOIE: A Portuguese Open Information Extraction system with inferences.
640 *Natural Language Engineering* **2019**, *25*, 287–306. doi:10.1017/S135132491800044X.
- 641 49. de Oliveira, L.S.; Glauber, R.; Claro, D.B. DependenteIE: An Open Information Extraction system on
642 Portuguese by a Dependence Analysis. *Encontro Nacional de Inteligência Artificial e Computacional* **2017**.
- 643 50. de Oliveira, L.S.; Claro, D.B. DptOIE: A Portuguese Open Information Extraction system based on
644 Dependency Analysis. *Computer Speech and Language - under review* **2019**.
- 645 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
646 relation extraction for knowledge acquisition. Proceedings of the 14th Conference of the European Chapter
647 of the Association for Computational Linguistics, volume 2: Short Papers, 2014, pp. 12–16.
- 648 52. Qiu, L.; Zhang, Y. ZORE: A Syntax-based System for Chinese Open Relation Extraction. Proceedings of
649 the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for
650 Computational Linguistics, 2014, pp. 1870–1880. doi:10.3115/v1/D14-1201.
- 651 53. Xu, J.; Gan, L.; Deng, L.; Wang, J.; Yan, Z. Dependency parsing based Chinese open relation extraction.
652 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015, Vol. 01,
653 pp. 552–556. doi:10.1109/ICCSNT.2015.7490808.
- 654 54. Wang, Y.; Zhou, G.; Tian, F.; Nan, Y.; Ma, J. GCORE: A Gravitation-Based Approach for Chinese Open
655 Relation. 2015 International Conference on Computer Science and Mechanical Automation (CSMA), 2015,
656 pp. 86–91. doi:10.1109/CSMA.2015.24.
- 657 55. Wu, X.; Wu, B. The CRFs-Based Chinese Open Entity Relation Extraction. 2017 IEEE Second International
658 Conference on Data Science in Cyberspace (DSC), 2017, pp. 405–411. doi:10.1109/DSC.2017.40.
- 659 56. Bassa, A.; Kroll, M.; Kern, R. GerIE-An Open Information Extraction System for the German Language.
660 *Journal of Universal Computer Science* **2018**, *24*, 2–24.

- 661 57. Truong, D.; Vo, D.T.; Nguyen, U.T. Vietnamese Open Information Extraction. Proceedings of the Eighth
662 International Symposium on Information and Communication Technology; ACM: New York, NY, USA,
663 2017; SoICT 2017, pp. 135–142. doi:10.1145/3155133.3155171.
- 664 58. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic Mapping Studies in Software Engineering.
665 Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering;
666 BCS Learning & Development Ltd.: Swindon, UK, 2008; EASE'08, pp. 68–77.
- 667 59. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies
668 in software engineering: An update. *Information and Software Technology* **2015**, *64*, 1 – 18.
669 doi:https://doi.org/10.1016/j.infsof.2015.03.007.
- 670 60. Glauber, R.; Claro, D.B. A Systematic Mapping Study on Open Information Extraction. *Expert Systems with
671 Applications* **2018**. doi:https://doi.org/10.1016/j.eswa.2018.06.046.
- 672 61. Garcia, M.; Gamallo, P. Entity-centric coreference resolution of person entities for open information
673 extraction. *Procesamiento del Lenguaje Natural* **2014**, *53*, 25–32.
- 674 62. Nunes, T.; Schwabe, D. Building Distant Supervised Relation Extractors. 2014 IEEE International
675 Conference on Semantic Computing. IEEE, 2014, pp. 44–51.
- 676 63. Akbik, A.; Danilevsky, M.; Kibrom, Y.; Li, Y.; Zhu, H.; others. Multilingual information extraction with
677 PolyglotIE. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics:
678 System Demonstrations, 2016, pp. 268–272.
- 679 64. Duc, N.T.; Bollegala, D.; Ishizuka, M. Cross-language latent relational search between japanese and english
680 languages using a web corpus. *ACM Transactions on Asian Language Information Processing (TALIP)* **2012**,
681 *11*, 11.
- 682 65. Vulić, I.; Moens, M.F. Monolingual and cross-lingual information retrieval models based on (bilingual) word
683 embeddings. Proceedings of the 38th international ACM SIGIR conference on research and development
684 in information retrieval. ACM, 2015, pp. 363–372.
- 685 66. Upadhyay, S.; Faruqui, M.; Dyer, C.; Roth, D. Cross-lingual models of word embeddings: An empirical
686 comparison. *arXiv preprint arXiv:1604.00425* **2016**.
- 687 67. Xiao, M.; Guo, Y. Distributed word representation learning for cross-lingual dependency parsing.
688 Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014, pp.
689 119–129.
- 690 68. Täckström, O.; McDonald, R.; Uszkoreit, J. Cross-lingual word clusters for direct transfer of linguistic
691 structure. Proceedings of the 2012 conference of the North American chapter of the association for
692 computational linguistics: Human language technologies. Association for Computational Linguistics, 2012,
693 pp. 477–487.
- 694 69. Yang, Z.; Salakhutdinov, R.; Cohen, W. Multi-task cross-lingual sequence tagging from scratch. *arXiv
695 preprint arXiv:1603.06270* **2016**.
- 696 70. Koehn, P. Europarl: A parallel corpus for statistical machine translation. MT summit, 2005, Vol. 5, pp.
697 79–86.