A Review of Open Information Extraction Techniques

Sally Mohamed Ali, Hamdy M. Mousa, Mahmoud Hussein

Dept. of Computer Science, Faculty of Computers and Information
Menoufia University, Egypt
smbm222@yahoo.com, hamdimmm@hotmail.com, fci_3mh@yahoo.com

***Abstract*—Nowadays, massive amount of data flows all the time. Approximately between 20 or 30 percent of these data is text. This data is always organized in semi-structured text, which cannot be used directly. To make use of such huge amounts of textual data, there is a need to detect, extract, and structure the information conveyed through this data in a fast and scalable manner. This can be performed using Information Extraction Techniques. However, the task of information extraction is one of the main challenges in Natural Language Processing and there are limitations for its implementation on a large scale of data. Open Information Extraction (OIE) is an open-domain and relation-independent paradigm to perform information extraction in an unsupervised manner. This technique can lead to high-speed and scalable performance. The review of previous research proposals reveals that there are OIE experiments among different languages, such as English, Portuguese, Spanish, Vietnamese, Chinese, and Germany. This paper reviews the OIE techniques, compare their performance in some languages, and then integrates these results with the languages complexity levels to reveal the relationship between the suitable model and the language complexity level.**

Keywords—Open Information Extraction; Natural Language Processing

I. Introduction

A. *Definition and Evolution of Open Information Extraction*

Information extraction is benefit many fields such as collect product information from different websites, automatic answering of questions, contact information search, find and link a specific information in journal articles, and removal of the noisy data [1]. In order to make wider use of information extraction, researchers have introduced Open Information Extraction (OIE), which is a relation-independent paradigm that extracts a large set of relational tuples in a much more general domain of articles. Open Information Extraction (OIE) is also an open-domain paradigm for information extraction performed in an unsupervised manner.

The OIE task is an unsupervised one that has no idea about the types of entities to be mined up front. Furthermore, weakly supervised methods either expand a small set of initial relations or they use other knowledge bases from external sources in order to learn the relations in a corpus [2]. OIE has been shown to be a useful paradigm for a wide range of semantic tasks, including question answering, summarization, and text comprehension and has consequently drawn consistent attention over the last years [3]. The main properties of OIE systems are as follows [4]: These systems are domain independent, rely on unsupervised extraction methods, and scalable to large amounts of text.

B. *Current challenges and motivations*

Even after more than one decade of research in the area of OIE, there is only a very little work on evaluating and comparing results among different OIE systems in a large-scale, objective, and reproducible fashion. Also, most of the previous work focuses on the English language and some exceptions in other languages [5]. In this paper, a review of the OIE modules is accomplished with give rise to three different languages: English, Spanish, and Chinese. The paper specifically focuses on the different techniques that used in these languages, compares between the results, which reached in these languages to address most effective OIE module in each language, and integrates these results with the languages complexity levels to reveal the relationship between the suitable method and the language complexity level. This paper aims at clarifying the using of all OIE modules and promoting OIE in other languages by paving the way to choose the most suitable method for each language.

This article is organized as follows. Section 2 describes different methodologies used in OIE models. Section 3 presents the using of OIE in some languages. Then, section 4 discusses the effect of languages complexity on OIE implementation. Finally, conclusions and future work are presented in section 5.

II. Different Methodologies of OIE

An Open IE system performs the task of extracting relationships (or facts) in raw texts written in natural language with the triple format for any binary relation found in the text.:(arg1, rel, arg2) where, arg1and arg2are noun phrases that have a semantic relationship determined by rel which is the relation that can be a verb or verb +pronoun for example [6]. The first generation of OIE was known as data-based OIE that includes a shallow syntax and dependency methods. Recently, the second generation of OIE has emerged and it is known as rule-based OIE. Also, the second generation includes shallow syntax and dependency methods. Depending on the reviewing of previous research, this research adopts four categories of OIE as shown in Figure 1. [7].

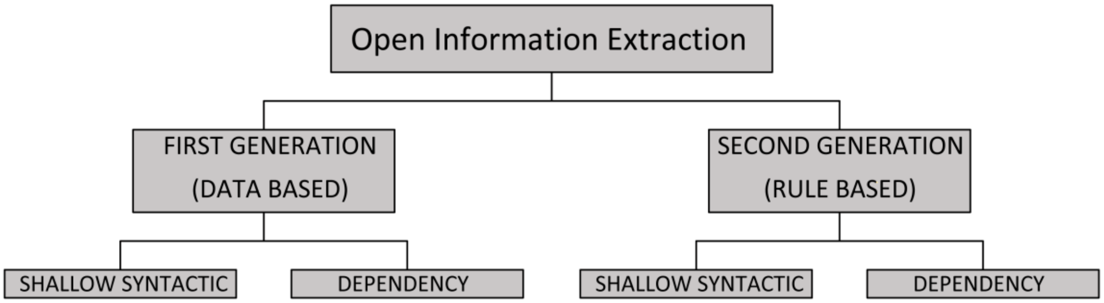

Open Information Extraction
FIRST GENERATION (DATA BASED)
SECOND GENERATION (RULE BASED)
SHALLOW SYNTACTIC
DEPENDENCY
SHALLOW SYNTACTIC
DEPENDENCY

Fig. 1. Open Information Extraction Models Categories [7]

A. *Data-based OIE (First Generation)*

This method is considered as the first generation of OIE generates patterns based on training data represented by means of dependency tree or Part of Speech (PoS) tagged text. A PoS-tagging is a process scans all words in a sentence and assigns a tag to clarify its type to each word [7]. The dependency parsing is a set of directed syntactic relations between the words in the sentence [8]. The root of the dependency parsing is either a non-copular verb or the subject complement of a copular verb. The examples for this type are Text Runner and OLLIE [7].

1) Training data and shallow syntax

The example for this type is Text Runner model. This model has two phases to extracting generic relationships as shown in Figure 2. In the first phase, a syntactic parser is applied to several thousand sentences, generating the corresponding syntactic dependencies. For each parsed sentence, then applies a set of heuristic constraints to label the sentence as a positive example of a relationship. Second phase, the labelled sentences are mapped into a feature vector, with domain- independent features that can be evaluated at extraction time without the use of a parser. Examples of included features are: the sequence of PoS tags between two entities, the PoS tag to the left of the first entity, the PoS tag to the right of second entity. The features are used to train a Naïve Bayes classifier[7].

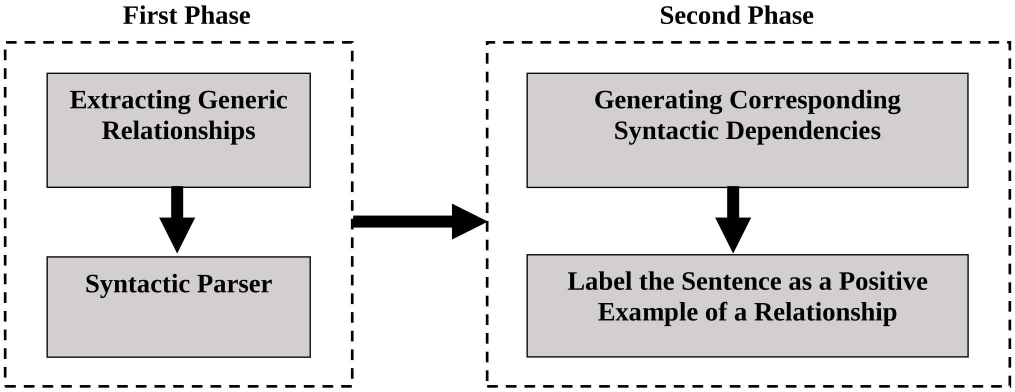

First Phase
Second Phase
Extracting Generic Relationships
Syntactic Parser
Generating Corresponding Syntactic Dependencies
Label the Sentence as a Positive Example of a Relationship

Fig. 2. The text runner model's stages [7]

2) Training data and dependency parsing

Training data and dependency parsing methods take a sentence as input and perform PoS tagging, syntactic chunking, and dependency parsing, and then return a set of relation triples [9]. OLLIE is an example of this category. As shown in Figure 3, this model collects sentences from a corpus containing words including variations of the verb. For each sentence, OLLIE (Open Language Learning for Information Extraction) [10] computes the syntactic dependencies connecting the two relationship arguments and the relational word. Next, it annotates the relation node in the syntactic dependency path with the exact relation word and the PoS-tag. Then by checking some constraints over the syntactic dependency tree, the model generates extraction patterns which mean the types of relation used in information extraction process. for patterns fails to match the constraints the model generates semantic and lexical patterns by removing the relational then aggregates the patterns based on the syntactic structure. After that, the relational word is replaced into a list of words with which the pattern was seen. The extraction templates are generated by replacing, the corpus associated with each sentence the relational word with rel, and by normalizing auxiliary verbs [7].

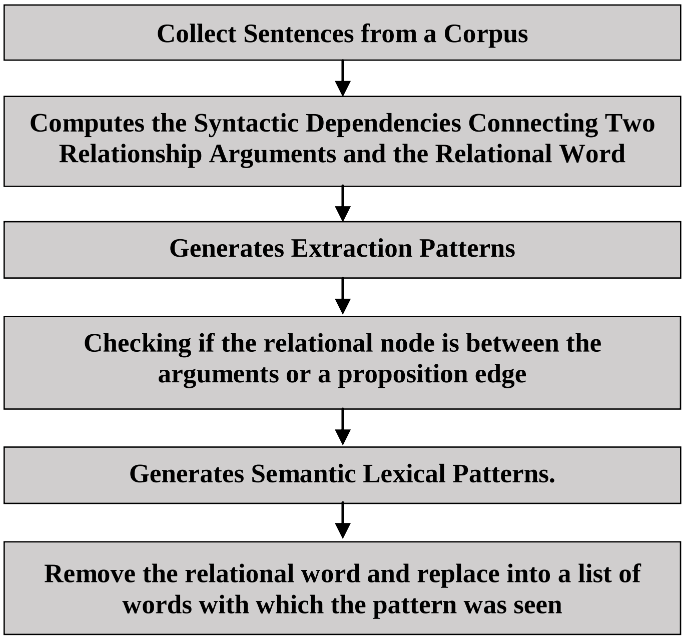

Collect Sentences from a Corpus
Computes the Syntactic Dependencies Connecting Two Relationship Arguments and the Relational Word
Generates Extraction Patterns
Checking if the relational node is between the arguments or a proposition edge
Generates Semantic Lexical Patterns.
Remove the relational word and replace into a list of words with which the pattern was seen

Fig. 3. OLLIE model's stages [7]

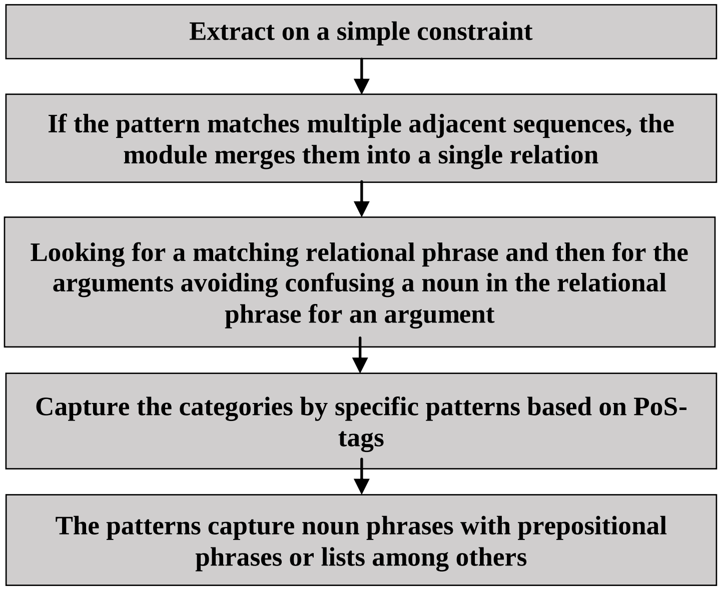

Extract on a simple constraint
If the pattern matches multiple adjacent sequences, the module merges them into a single relation
Looking for a matching relational phrase and then for the arguments avoiding confusing a noun in the relational phrase for an argument
Capture the categories by specific patterns based on PoS-tags
The patterns capture noun phrases with prepositional phrases or lists among others

Fig. 4. ReVerb model's stages [7]

A. Rule-based OIE (Second Generation)

This method relies on hand-crafted heuristics based on textual features, such as PoS-tagged or dependency parse trees.The example of this type are clauseIE and ExtrHech[7].

1) Rule-based and shallow syntax:

Rule-based and shallow syntax rely on lexica-syntactic patterns and hand-crafted from PoS tagged text [11]. The model extracts relationships based on a simple constrain which is every relational is a verb or a verb followed by a preposition or a verb followed by nouns, adjectives, or adverbs. If there are multiple possible matches for a single verb, the longest possible match is chosen. If the pattern matches multiple adjacent sequences, the module merges them into a single relation phrase, and the system looks first for a matching relational phrase and second for the arguments (e1, e2) such that avoiding the confusion with a noun in the relational phrase. These categories are then captured by specific patterns based on PoS-tags. The patterns capture noun phrases with prepositional phrases or lists among others [7].

2) Rule-based and dependency parsing

Rule-based and dependency parsing make the use of hand-crafted heuristics operating on dependency parses [11]. ClausIE is an example for this category. As shown in Figure 5, this model reasons over the information given by a dependency parser to extract relationships. Then the ClausIE identifies the clause type and the verb type using two insights. Once the clause type is identified, an extraction rule can be applied. The second insight is that each occurrence of a verb in the language sentence can be classified into the number of types. Also, the verb type along with the presence of a direct object, indirect object or a compliment, is uniquely determined by the type of the constituents and the type of the clause. ClausIE uses these observations to detect the clause type. It then applies rules specific to each clause to extract relationships [7].

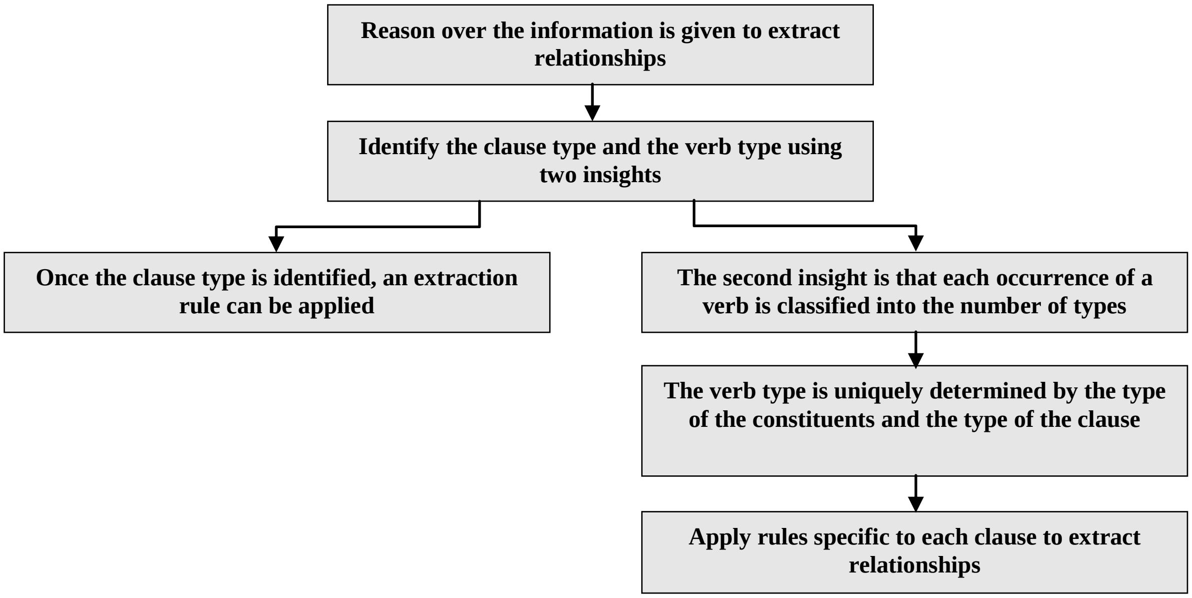

Reason over the information is given to extract relationships
Identify the clause type and the verb type using two insights
Once the clause type is identified, an extraction rule can be applied
The second insight is that each occurrence of a verb is classified into the number of types
The verb type is uniquely determined by the type of the constituents and the type of the clause
Apply rules specific to each clause to extract relationships

Fig. 5. ClausIE model's stages

III. Using OIE in different languages

Many languages are used on the internet. According to the number of used people, English is the most used language and followed by Chinese and Spanish. Table 1 shows the ranking of the languages by the number of users. However, in this study, the first three languages have been chosen to investigate the OIE application. This paper aims to compare the use of OIE in different languages and the effect of languages complexity on applying it.

The langue complexity include different dimension such as phonological, morphological, syntactic, and semantic complexity. language is more complex if it has more marked members in its phonemic inventory, or if it makes more extensive use of inflectional morphology [12]. However, by reviewing number of research [13], [14] and [15] try to rank the langue's by their difficulty it could be concluded for the chosen languages that the Chinese has the most complexity followed by the Spanish which have medium complexity then the English has low complexity. The following section presents the different OIE models introduced in the three selected languages (English, Spanish, and Chinese).

TABLE I. NUMBER OF INTERNET USERS FOR DIFFFRENT LANGUAGES [16]

Rank	Language	No. of Internet users	Percentage
1	English	1,052,764,386	25.3%
2	Chinese	804,634,814	19.4%
3	Spanish	337,892,295	8.1%
4	Arabic	219,041,264	5.3%
5	Portuguese	169,157,589	4.1%
6	Indonesian / Malaysian	168,755,091	4.1%
7	French	118,626,672	2.9%
8	Japanese	109,552,842	2.8%
9	Russian	108,014,564	2.7%
10	German	84,700,419	2.2%
11–36	Others	950,318,284	22.9%

1) Using OIE in English

Many researches have applied OIE on the English language. OIE was first introduced by Text Runner, developed at the University of Washington Turing Center headed by Oren Etzioni [17]. Other methods introduced later such as Reverb, OLLIE, Clause IE, helped to shape the OIE task by characterizing some of its aspects. At a high level, all of these approaches make use of a set of patterns to generate the extractions. Depending on the particular approach, these patterns are either hand-crafted or learned [5].

2) Using OIE in Spanish (español)

Spanish is one of the top three spoken languages and in top five for the content languages on the Internet. Therefore, there is no doubt that it should have corresponding methods for its automatic processing Open IE for the Spanish language that outperforms the systems implementing the similar rule-based strategy. It also shows good results compared to the more complex method based on the deep automatic linguistic analysis and definitely has gained in time [18].

3) Using OIE in Chinese (中国)

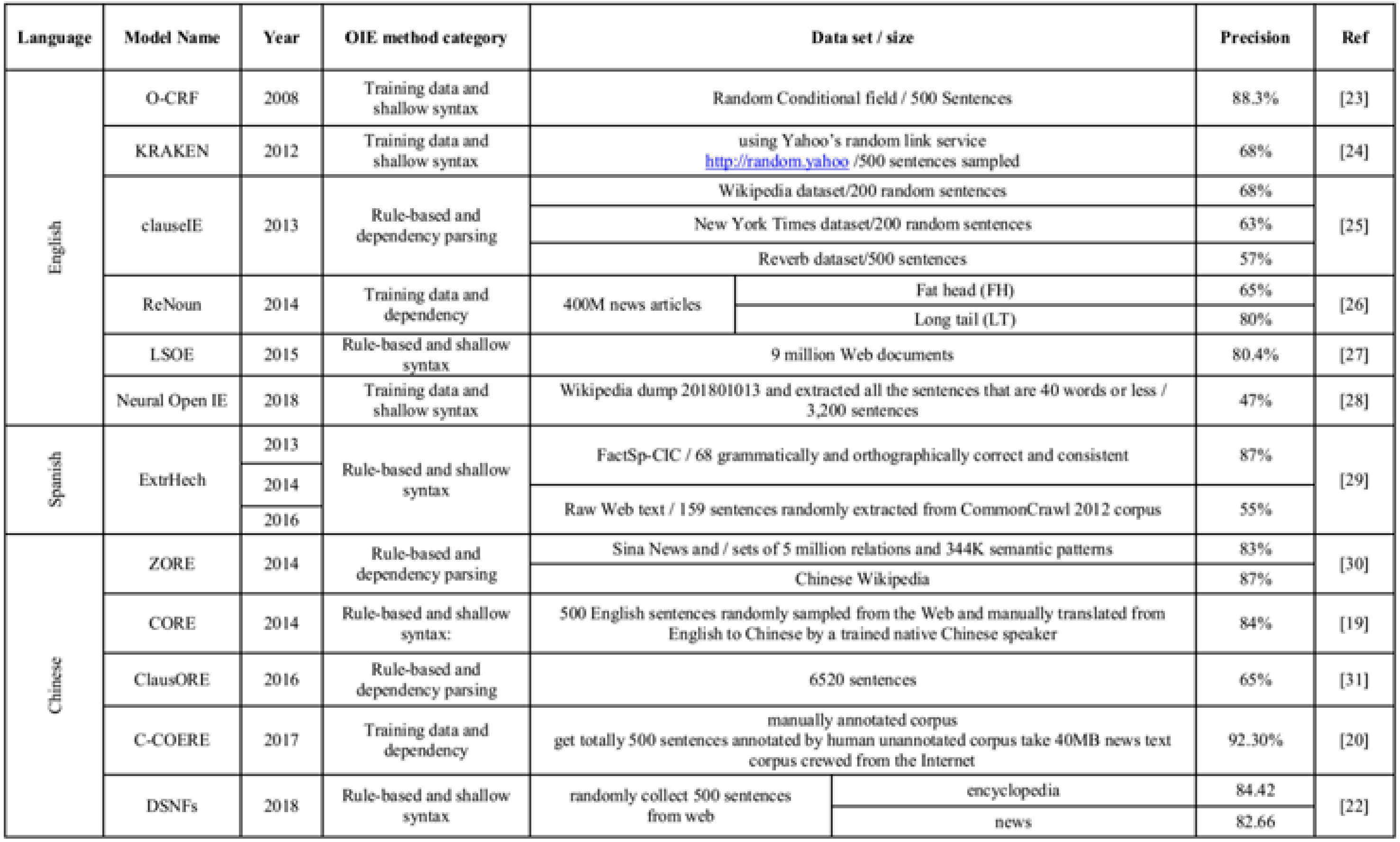
In Chinese language, a number of papers have implemented open information extraction. One of the researches explores Chinese open relation extraction which utilizes a series of NLP techniques to extract relations embedded in Chinese sentences [19]. Another one constructs the entity relation graph with the extracted tuples and makes a visual display [20].

IV. ACCURACY OF OIE IN DIFFERENT LANGUAGES

Open IE approaches are essential when the number of relations of interest is massive or unknown. On the other hand, while these new techniques to deal with the problem are getting more sophisticated, and the variety of data considered increases, many of the evaluations in this line of work are isolated and seldom based on a rather small sample. Open IE systems were predominantly evaluated by hand on small-scale corpora that consist of only a few hundred sentences, thereby ignoring one of the fundamental goals of Open IE: scalability to large amounts of text. Moreover, none of the datasets that were used for assessing the performance of different systems is widely agreed upon. The performance of the OIE module can be evaluated by the performance of precision that can be defined as:

$$\text{Precision} = \frac{Number\ correct}{Number\ correct + number\ incorrect} \quad [21]$$

Because of the simplicity of English morphology, the Open IE systems in English have extracted billions of assertions as the basis for both common-sense knowledge and novel question-answering systems. Also, the performance Open IE system in Spanish is similar in English. On the other hand, Chinese open relation extraction is not well established, because of the complexity of Chinese linguistics makes it harder to operate, and the methods for English are not compatible with that for Chinese. The diversities between Chinese and English linguistics are mainly reflected in morphology and syntax [22]. Table 2 collects the previous OIE models in the three investigate languages and the precision evaluation of them.



Language	Model Name	Year	OIE method category	Data set / size		Precision	Ref
English	O-CRF	2008	Training data and shallow syntax	Random Conditional field / 500 Sentences		88.3%	[23]
	KRAKEN	2012	Training data and shallow syntax	using Yahoo's random link service http://random.yahoo /500 sentences sampled		68%	[24]
	clauseIE	2013	Rule-based and dependency parsing	Wikipedia dataset/200 random sentences		68%	[25]
				New York Times dataset/200 random sentences		63%	
				Reverb dataset/500 sentences		57%	
	ReNoun	2014	Training data and dependency	400M news articles	Fat head (FH)	65%	[26]
					Long tail (LT)	80%	
	LSOE	2015	Rule-based and shallow syntax	9 million Web documents		80.4%	[27]
	Neural Open IE	2018	Training data and shallow syntax	Wikipedia dump 201801013 and extracted all the sentences that are 40 words or less / 3,200 sentences		47%	[28]
Spanish	ExtrHech	2013 2014 2016	Rule-based and shallow syntax	FactSp-CIC / 68 grammatically and orthographically correct and consistent		87%	[29]
				Raw Web text / 159 sentences randomly extracted from CommonCrawl 2012 corpus		55%	
Chinese	ZORE	2014	Rule-based and dependency parsing	Sina News and / sets of 5 million relations and 344K semantic patterns		83%	[30]
				Chinese Wikipedia		87%	
	CORE	2014	Rule-based and shallow syntax:	500 English sentences randomly sampled from the Web and manually translated from English to Chinese by a trained native Chinese speaker		84%	[19]
	ClausORE	2016	Rule-based and dependency parsing	6520 sentences		65%	[31]
	C-COERE	2017	Training data and dependency	manually annotated corpus get totally 500 sentences annotated by human unannotated corpus take 40MB news text corpus crewed from the Internet		92.30%	[20]
	DSNFs	2018	Rule-based and shallow syntax	randomly collect 500 sentences from web	encyclopedia	84.42	[22]
					news	82.66	

TABLE II. OIE MODELS IN DIFFERENT LANGUAGES

V. Discussion

In the best of our knowledge, this research is the first attempt to compare the OIE modules and their use in specific languages. Furthermore, this comparison study integrates the performance of the OIE models and the morphological complexity of the languages. Three languages have been included in this study, while the scarcity of OIE implementation in many languages. In order to discuss the languages morphological complexity levels and their effect on OIE precision in different models, the selected languages have been sorted by their complexity as follows: Chinese, Spanish and English. For further explanation, English morphology is simpler comparing with other languages, because many words give a comprehensive meaning without a suffix or prefix. For example, “a cat” gives a meaning for a type of an animal, but in Chinese, there is no one word can give the same meaning wherein every word in the chinse needs to add another word in the left or the right to give a comprehensive meaning [32]. Also, in English, infinitives are marked by a special particle to make identifying them slightly easier. In contrast, in Spanish, infinitives are indicated by any particles, hence, their morphological form is an only indicator of it part-of-speech [18]. Figure 6 shows a comparison between the evaluation of OIE models in these languages and their complexity levels. This comparison reveals that the shallow syntactic approach resulted in the highest precision in the English language, which has the low morphological complexity. Also, the rule-based and shallow syntactic category result in a highest precision with Spanish language while the training data and dependency parsing category resulted in the highest precision in the Chinese language which has most morphological complexity. Obviously, the using a variety of categories in the English language reflects a large number of OIE implementation in English and the simplicity of its morphology while in the other languages the complexity of their morphologies limits the implementation of different categories.

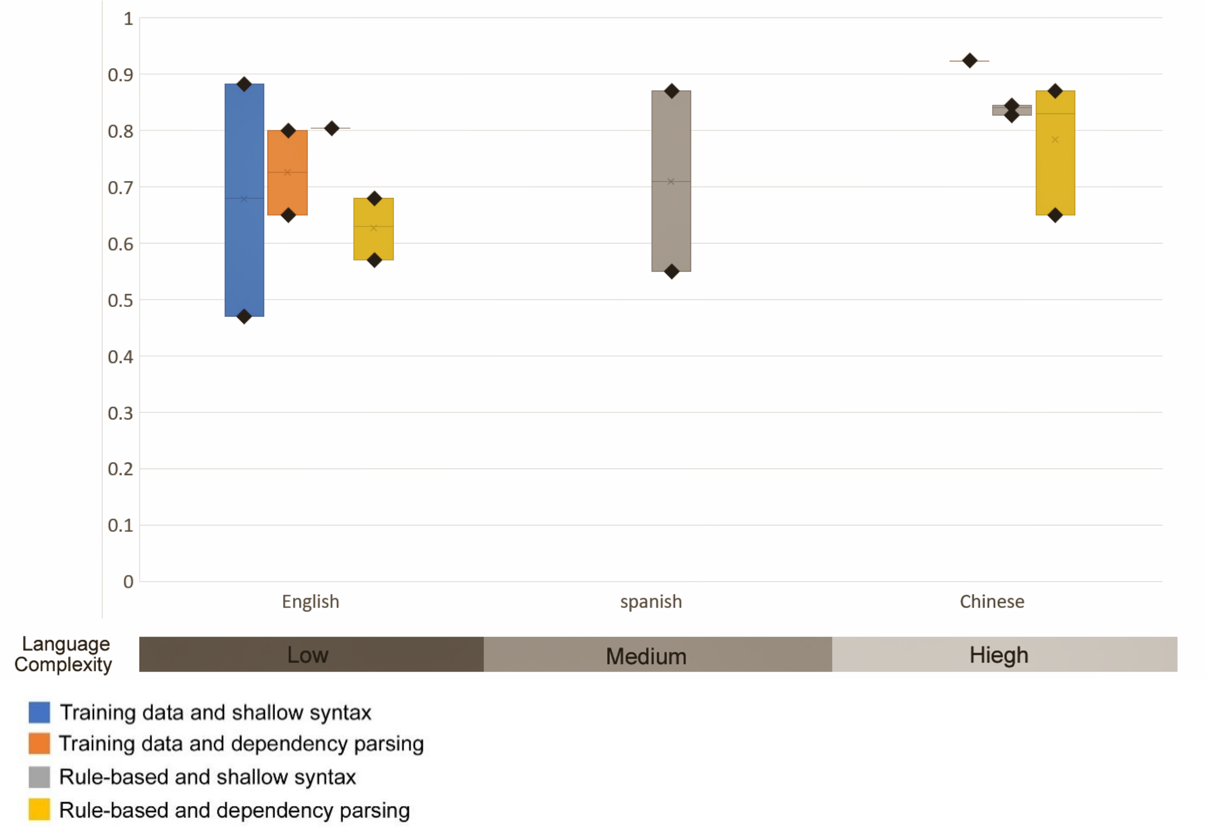


Fig. 6. Comparison between OIE models using their precisions in different languages

VI. Conclusion

This paper reviewed the existing approaches of OIE, which are divided into four main categories depending on the methodology to extract possible relations and compared the performance of these approaches among specific languages. Three languages (English, Spanish, and Chinese) have been selected depending on the amount of use on the internet.;. In order to compare the different OIE categories, the evolution of previous models in the selected languages has been collected. Also, in this comparison, the morphological complexity has taken into account to reveal its effect on the OIE models performance. The evaluation of OIE models in these languages and the complexity level them. This comparison reveals that the shallow syntactic approach resulted in the highest precision in the English language, which has the low morphological complexity and the rule-based and shallow syntactic category resulted in highest precision with Spanish language, while the training data and dependency-parsing category resulted in the highest precision in the Chinese language, which has most morphological complexity. This paper aims to paving the way to the new implementation of the OIE in the languages, which has a limited OIE implementation until now. The research methodology can help in choosing the most suitable category to use in the new implementation.

In the future work, this study should apply among all languages that have OIE implementation taken into consideration the different constrains and factors may affect the performance.

REFERENCES

[1] J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li, “Information Extraction: Methodologies and applications,” *Emerg. Technol. Text Min. Tech. Appl.*, pp. 1–33, 2008.

[2] C. C. Aggarwal, *Machine Learning for Text*. 2017.

[3] T. Falke, G. Stanovsky, I. Gurevych, and I. Dagan, “Porting an Open Information Extraction System from English to German,” *Proc. 2016 Conf. Empir. Methods Nat. Lang. Process.*, pp. 892–898, 2016.

[4] F. Pereira, P. Machado, E. Costa, and A. Cardoso, “Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015 Coimbra, Portugal, September 8–11, 2015 Proceedings,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9273, no. April 2016, 2015.

[5] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, “A Survey on Open Information Extraction,” in *COLING*, 2018.

[6] R. Glauber and D. B. Claro, “PT US CR,” *Expert Syst. Appl.*, 2018.

[7] D. S. Batista and C. Gaspar, “Large-Scale Semantic Relationship Extraction for Information Discovery,” 2016.

[8] L. Del Corro, “Methods for Open Information Extraction and Sense Disambiguation on Natural Language Text,” 2016.

[9] D. Vo and E. Bagheri, “Open Information Extraction,” vol. 1, no. 1, 2016.

[10] M. Schmitz, R. Bart, S. Soderland, and O. Etzioni, “Open Language Learning for Information Extraction,” *EMNLP-CoNLL*, no. July, pp. 523–534, 2012.

[11] P. Gamallo, “An overview of open information extraction,” *OpenAccess Ser. Informatics*, vol. 38, pp. 13–16, 2014.

[12] P. Juola, “Assessing linguistic complexity,” no. May, pp. 89–108, 2008.

[13] C. Bentz, T. Ruzsics, A. Koplenig, and T. Samardži, “A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora,” pp. 142–153, 2016.

[14] Effective Language Learning, “Language Difficulty Ranking,” 2014. [Online]. Available: https://www.effectivelanguagelearning.com/wp-content/w3tc/pgcache//language-guide/language-difficulty/_index.html_gzip. [Accessed: 18-Feb-2019].

[15] Glossika, “The Glossika Blog,” 2018. [Online]. Available: https://blog.glossika.com/rank-of-language-difficulty/. [Accessed: 18-Feb-2019].

[16] Internet World Stats, “Top Ten Internet Languages - World Internet Statistics,” *Internet World Stats*, 2016. [Online]. Available: https://www.internetworldstats.com/stats7.htm. [Accessed: 18-Feb-2019].

[17] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam, “Open Information Extraction: The Second Generation.,” *{IJCAI} 2011, Proc. 22nd Int. Jt. Conf. Artif. Intell. Barcelona, Catalonia, Spain, July 16-22, 2011*, vol. 11, pp. 3–10, 2011.

[18] Q. U. E. Para and O. El, “T e s i s,” 2014.

[19] Y.-H. Tseng *et al.*, “Chinese Open Relation Extraction for Knowledge Acquisition,” *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguist. Vol. 2 Short Pap.*, pp. 12–16, 2014.

[20] X. Wu and B. Wu, “The CRFs-Based Chinese Open Entity Relation Extraction,” *Proc. - 2017 IEEE 2nd Int. Conf. Data Sci. Cyberspace, DSC 2017*, pp. 405–411, 2017.

[21] S. C. de Abreu, T. L. Bonamigo, and R. Vieira, “A review on Relation Extraction with an eye on Portuguese,” *J. Brazilian Comput. Soc.*, vol. 19, no. 4, pp. 553–571, 2013.

[22] S. Jia, S. E, M. Li, and Y. Xiang, “Chinese Open Relation Extraction and Knowledge Base Establishment,” *ACM Trans. Asian Low-Resource Lang. Inf. Process. TALLIP Homepage Arch.*, vol. 17, no. 3, p. 15, 2018.

[23] M. Banko and O. Etzioni, “The Tradeoffs Between Open and Traditional Relation Extraction,” *Proc. 46th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. Conf.*, no. June, pp. 28–36, 2008.

[24] A. Akbik and A. Löser, “KrakeN : N-ary Facts in Open Information Extraction,” *Proc. Jt. Work. Autom. Knowl. Base Constr. Web-scale*

Knowl. Extr., pp. 52–56, 2012.

[25] L. Del Corro and R. Gemulla, “ClausIE : Clause-Based Open Information Extraction,” *Proc. 22nd Int. Conf. World Wide Web*, no. i, pp. 355–365, 2013.

[26] M. Yahya, S. E. Whang, R. Gupta, and A. Halevy, “ReNoun : Fact Extraction for Nominal Attributes,” *Proc. EMNLP 2014, Doha, Qatar*, pp. 325–335, 2014.

[27] C. C. Xavier, V. L. Strube de Lima, and M. Souza, “Open information extraction based on lexical semantics,” *J. Brazilian Comput. Soc.*, vol. 21, no. 1, 2015.

[28] L. Cui, F. Wei, and M. Zhou, “Neural Open Information Extraction,” 2018.

[29] A. Zhila and A. Gelbukh, “Open Information Extraction for Spanish Language based on Syntactic Constraints,” *Proc. ACL 2014 Student Res. Work.*, pp. 78–85, 2014.

[30] L. Qiu and Y. Zhang, “ZORE: A Syntax-based System for Chinese Open Relation Extraction,” *Emnlp*, pp. 1870–1880, 2014.

[31] J. Xu, L. Gan, L. Deng, J. Wang, and Z. Yan, “Dependency parsing based Chinese open relation extraction,” *Proc. 2015 4th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2015*, no. Iccsnt, pp. 552–556, 2016.

[32] J. L. Packard, “The {Morphology} of {Chinese}: {A} {Linguistic} and {Cognitive} {Approach},” 2000.