COMPARISON OF OPEN INFORMATION EXTRACTION FOR ENGLISH AND SPANISH

Zhila A. (alisa_zh@mail.ru), **Gelbukh A.** (www.gelbukh.com)

Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Open Information Extraction (IE) is the task of extracting relational tuples representing facts from text, with no prior specification of relation, no prespecified vocabulary, or a manually tagged training corpus. Part-of-speech based systems are shown to be competitive with parsing-based systems on this task and work faster for large-scale corpora. Nevertheless, implementation of such a system requires language-specific information. So far, all work has been done for English. We present a relation extraction algorithm for Open IE in Spanish, based on POS-tagged input and semantic constraints. We provide a description of its implementation in an Open IE system for Spanish ExtrHech. We compare its performance with Open IE systems for English, including a comparison on a parallel English-Spanish dataset, and show that the performance is comparable with the state-of-the-art systems, while the system is more robust to noisy input. We give a comparative analysis of errors in extractions for both languages.

Keywords: open information extraction, fact extraction, cross-lingual, Spanish, English

1. Introduction

Traditional Information Extraction (IE) is focused on detection of precise, prespecified type of information that would satisfy requests narrowed to a certain domain or area of activity. For example, an IE system could be trained for extraction of information of some certain classes, e.g. $ACQUIRE(argument_1, argument_2, \ldots, argument_n)$ with a fixed number of arguments n. Normally, an IE system would learn an extractor from a large tagged corpus for a specific relation marked up by human annotators or in a semi-supervised manner [8, 10, 14]. Although this approach might be efficient for a certain target relation or classes (e.g., *people* or *cities* as in [10]), it requires very expensive resources for training and, more importantly, this approach does not scale to large corpora such as the Web, where the number of possible relations is very large or where the target relations cannot be specified beforehand.

Open information extraction (Open IE) was introduced by Banko et al. [2] in 2007 as a new extraction paradigm that facilitates domain independent discovery of relations in text and can be readily scaled to a large and versatile corpus such as the Web. An Open IE system extracts *all* possible relations and assertions without requiring any prior specification of relations, manually tagged training corpora, example seeds tailored for the target relations, or any other relation-specific input. This guarantees scalability, and the system can satisfy unanticipated user needs. Open IE is necessary when the number of relations is large and the relations are not prespecified [3]. Consequently, it can serve purposes distinct from the traditional IE: fact extraction at sentence level (e.g. <*Mozart*, *was born in*, *Salzburg*>), new perspective on search as question answering (e.g. "*Where was Mozart born?*") in an unrestricted form [6], or assessment of the quality of text documents at the Web scale [7].

Independency from relation pre-specification is achieved through the implementation of a compact set of relation-independent lexico-syntactic patterns that allow identification of *arbitrary* relations [2]. However, the patterns are language dependent. All previous work in this field has been done for English [4, 6, 12, 15]; no language-related issues not specific for English have been addressed.

We present an Open IE for Spanish *ExtrHech*, compare its performance with that of a similar Open IE system for English ReVerb [6] on a parallel dataset, and perform analysis of errors for both languages.

The paper is organized as follows. Related work is reviewed in Section 2. Section 3 presents the Open IE approach for Spanish and describes the *ExtrHech* system. Section 4 describes the experimental results for two datasets in Spanish and a parallel English-Spanish dataset. In section 5, the analysis of errors is presented. Section 6 draws the conclusions and outlines future work.

2. Related Work

Open IE is the task of extracting arbitrary relations with their corresponding arguments from text without pre-specification of relations or manually tagged training corpora. The first step of any Open IE system is extraction of relations from a sentence. For example, in a sentence "*The policeman saw a boy who was crossing the street*", two assertions can be identified: <*the policeman, saw, a boy*> and <*a boy, was crossing, the street*>. A large corpus of text such as the Web is highly redundant, and many assertions are expressed repeatedly in different forms. After being encountered many times in various sources, an assertion has a significantly higher probability to be true.

The basic idea is that most sentences contain highly reliable syntactic clues to their structure [2]. There are three major approaches to relation identification in Open IE.

1. *Self-supervised learning* involves three steps: automatic labeling of relations using heuristics and distant supervision; learning of a relation phrase extractor; and extraction, for which a candidate pair of arguments is detected and then a relation extractor is applied to detect a relation between these arguments. Examples of such systems are TextRunner [2], WOE^{pos}, and WOE^{parse} [15]. One of its shortcomings is that potential arguments are detected before the relation is defined and cannot be backtracked. Therefore, a noun that actually belongs to a relation phrase can be marked as an argument. For example, in the relation "*to make a deal with*", *deal* can be incorrectly recognized as an argument. Consequently, the output of such systems contains many incoherent or uninformative extractions.
2. *Context analysis*, implemented in OLLIE system [12]. This approach overcomes various limitations of the other approaches. First, it extracts not only relations

expressed via verb phrases, but also relations mediated by adjectives, nouns, etc. Second, it is not limited to binary relations and can detect more than two arguments of a relation. Yet deeper context analysis requires syntactic parsing, which is time- and resource-consuming and makes real-time processing at Web scale impractical. Syntactic parsing analysis using heuristic rules is also implemented in the Open IE for Spanish FES [1].

3. *Syntactic and lexical constraints* implemented in the form of rules, as in ReVerb [6]. In contrast to the first approach, it initially detects a verb phrase, and then searches for its possible arguments, which reduces incoherent and uninformative extractions. It also has more light-weight implementation and faster execution time than context-analysis systems because it is based on part-of-speech analysis.

These approaches have been evaluated only for English. However, their relation extraction algorithms are language dependent: they use either part-of-speech or syntactic dependency information. It is not known how language affects implementation and output of Open IE.

We present a relation extraction algorithm for Open IE in Spanish, following the third approach. Since the datasets used for evaluation of the Open IE systems in [2, 6, 12] (300 to 500 sentences) are not available, we have created two comparable datasets for evaluation of our system.

3. ExtrHech, an Open IE system for Spanish

ExtrHech, an Open IE system for Spanish, is a POS-tag based system using syntactic and lexical constraints; see Figure 1.

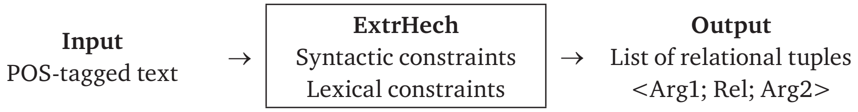

Fig. 1. Processing pipeline of ExtrHech

The system takes a POS-tagged text as an input. For POS-tagging we used Freeling-2.2 [13] that uses EAGLES POS-tagset for Spanish.

ExtrHech performs sentence-by-sentence processing. First, it looks for a verb phrase that is limited to be either a single verb or a verb immediately followed by dependent words till a preposition (*nació en*) or a preposition followed immediately by an infinitive (*sirven para acentuar*). The expression for a verb phrase is:

$$\text{VREL} \rightarrow (\text{V W*P})|(\text{V}),$$

where V stands for a single verb possibly preceded by a reflexive pronoun (*se caracterizaron*), or a participle (*relacionadas*); V W*P stands for a verb with dependent words, where W can be a noun, an adjective, an adverb, a pronoun, or an article, and P stands for a preposition possibly immediately followed by an infinitive. Here * stands for zero or more occurrences, | stands for choice of a variant; ? (see below) stands for zero or one occurrence.

Next, the system looks to the left of the verb phrase for a noun phrase that could be a first argument in a relation. Then it searches to the right from the verb phrase for a second argument. The following expression describes noun phrases in ExtrHech:

$$\text{NP} \rightarrow \text{N (PREP N)?},$$

where N stands for a noun optionally preceded by an article (*los gobernantes*), an adjective or ordinal number (*los primeros homínidos*), a number (*3.5 millones*), or their combination, optionally followed by a single adjective (*el epíteto heroico*), a single participle (*las fuentes consultadas*), or both (*los documentos escritos antiguos*). PREP stands for a single preposition. In our system a noun phrase can be either a single noun with optional modifiers or a noun with optional modifiers followed by a dependent prepositional phrase, consisting of a preposition and another noun with its optional modifiers (*la historia de la civilización romana*).

If a noun is followed by a participle clause terminating with another noun, then the participle phrase is resolved into a separate relational tuple. For an example,

(1) *Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.*
"The Egyptians were characterized by their beliefs related with death."

gives two relational tuples:

(2) <Arg1 = *Los egipcios;* Rel = *se caracterizaron por;* Arg2 = *sus creencias*>
(3) <Arg1 = *sus creencias;* Rel = *relacionadas con;* Arg2 = *la muerte*>,

with (2) corresponding to the main verb of sentence (1) and (3) corresponding to the participle clause.

ExtrHech also resolves coordinating conjunctions for verbal and noun phrases into independent relations or arguments correspondingly. Relative clauses are also resolved into independent assertions. Lexical constraints currently limit the length of relational phrases to prevent over-specifying of a relation. We use EAGLES POS-tag set and properly treat reflexive pronouns for verbal phrases. Currently we do not tackle anaphora, zero subject construction, and free word order. Still ExtrHech's precision is comparable with that of other Open IE systems (see Section 4.1).

4. Experimental Results

4.1. Experiments on Different Spanish Datasets

We analyzed ExtrHech's performance on two datasets.[1] The first one, FactSpCIC [1], contains 68 grammatically and orthographically correct and consistent

[1] Both datasets are available on www.gelbukh.com/resources/spanish-open-fact-extraction.

sentences manually selected from school textbooks. The second one contains 159 sentences randomly extracted from CommonCrawl 2012 corpus [9], which is a corpus of web crawl texts from over 5 billion web pages. It contains the sentences in their original form as they were crawled from the Internet. As evaluated by a human judge, 36 sentences (22% of the corpus) were either grammatically incorrect or incoherent.

Two human judges independently evaluated each extraction as correct or incorrect. For FactSpCIC dataset, they agreed on 89% of extractions (Cohen's kappa k = 0.52), which is considered to be moderate agreement [11]. For the raw Web text dataset of 159 sentences, they agreed on 70% of extractions (Cohen's kappa k = 0.40), which is considered the lower bound of moderate agreement. The number of correct extraction was calculated as an average for the two judges.

Precision of the system is the fraction of returned extractions that are correct. *Recall* is the fraction of correct extractions in the number of all possible correct extractions. To estimate the latter, we made a list of all extractions that the system is expected to return. Then, this set was extended by the extractions returned by the system that both annotators considered correct. This gives a lower bound estimation of all possible extractions that could be detected in the datasets, which gives the *upper bound* for recall; see Table 1.

Table 1. Performance of ExtrHech system on a grammatically correct and on a noisy datasets

Dataset	Precision	Recall
FactSpCIC (grammatically correct)	87%	70%
Raw Web text (noisy)	55%	49%

4.2. Experiments on Parallel Spanish and English Dataset

To analyze differences in performance between systems for Spanish and English, we formed a parallel Spanish-English dataset. The original Spanish FactSpCIC dataset was manually translated into English by a professional human translator. Then, the fact extractor for Spanish ExtrHech was run on the 68 sentences in Spanish, and ReVerb was run on the English translation.

The evaluation of extraction for Spanish is presented in Section 4.1. The output in English was also evaluated by two human judges. The judges agreed on 85% of extractions (Cohen's kappa $\kappa = 0.60$), similar to 89% agreement for Spanish, which is the upper bound for the moderate agreement range and is slightly higher than $\kappa = 0.52$ for Spanish.

Precision and recall for the extraction in English were calculated as described in Section 4.1.1; see Table 2, where the number of correct extractions is averaged by the two human judges.

Table 2. Comparison of Open IE systems for Spanish and English on a parallel dataset

System	Precision	Recall	correct extractions	found extractions	expected extractions
ExtrHech (Spanish)	87%	70%	99.5	115	137
ReVerb (Englist)	76%	50%	71	93	139

For the parallel dataset, precision and recall for the English ReVerb system are lower than those for the Spanish ExtrHech. Yet this might be due to overadjustment of ExtrHech to the dataset, which was also used during development of the system (note that no learning was involved). However, the total number of assertions extracted by ReVerb is lower than the amount of extractions by ExtrHech. Thus the Spanish extractor is more robust than the English one. Higher number of expected extractions for the English dataset (139) is due to the absence of the zero-subject phenomenon in English.

To show that ExtrHech performs at the level comparable with the state-of-the-art Open IE systems, we provide the comparative data on performance of the Open IE systems described in Section 2 based on the data provided in [1, 6, 12] and ExtrHech for different datasets; see Table 3.

Table 3. Comparative data for various Open IE systems

System	Approach	Dataset (# of sentences)	Precision	Recall	Running Time
ExtrHech (Spanish)	syntactic and lexical constr. over POS-tagged text	FactSpaCIC (68)	0.87	0.73	< 5 min
		raw Web text (159)	0.55	0.49	
ReVerb (English)	syntactic and lexical constr. over POS-tagged text	FactSpaCIC (68), translated	0.76	0.50	< 5 min
		Yahoo (500)	0.87 0.60	at 0.20 at 0.50	
Text-Runner (English)	self-learning on POS-tagged text	Yahoo (500)	< 0.64	at >0	< 5 min
WOE^{parse} (English)	self-learning on parsed text	Yahoo (500)	0.87	at 0.15	hours
OLLIE (English)	context analysis of parsed text	news, Wikipedia, biology textbooks (300)	0.66–0.85	N/A (various yield levels from [11])	N/A, probably hours
FES (Spanish)	heuristic rules on parsed text	FactSpaCIC (68)	0.87	0.91	hours

Table 3 shows that for the dataset of raw Web sentences, ExtrHech's 0.55 precision and 0.49 recall are slightly lower than those of ReVerb system for the Yahoo dataset (0.60 and 0.50 correspondingly). However, the Yahoo dataset is not available, so we do not know whether it is raw Web text including incorrect and incoherent sentences common for texts on the Internet that hinder fact extraction.

ExtrHech speed is at the same level as that of other POS-tag based systems. It is much faster than syntactic parsing based systems, which perform significantly slower although with better precision.

5. Error Analysis

To analyze errors in assertion extraction for ReVerb and ExtrHech for the parallel English-Spanish dataset FactSpaCIC, first, we compared the distributions of the types of errors found in extractions. The type classification was modified from [6] to clearly distinguish between error types and their reasons. Table 4 shows the fractions of each type of errors in the total number of extractions for both systems.

Table 4. Distribution of the types of errors in all extractions

System and total number of extractions	Incorrect relation phrase	Incorrect arguments	Correct relation phrase, incorrect arguments	Incorrect argument order
ExtrHech (Spanish), 115	0.09	0.22	0.16	0.04
ReVerb (English), 93	0.12	0.26	0.13	–

Very similar distributions can be seen for the first 3 types of errors for both languages. Incorrect argument order was not observed for English because of highly dominant direct word order.

Causes of errors in assertion extractions from the parallel FactSpaCIC dataset are shown in Table 5.

Table 5. Causes of errors in assertion extraction from parallel FactSpaCIC, percent of all errors

	ExtrHech	ReVerb
N-ary relation	24%	41%
Underspecified noun phrase	10%	9%
Incorrect POS-tagging	10%	5%
Incorrect coordinative conjunction	43%	14%
Incorrect relative clause	19%	9%
Non-contiguous relation	5%	–
Over-specified relation phrase	5%	–
Inverse word order	14%	–
Infinitive	–	9%
Underspecified relation phrase	–	5%
Over-specified noun phrase	–	5%
No extraction	–	23%

One of the main issues for both languages is N-ary relations, i.e. relations requiring more than two arguments (e.g. "*The boy gave a book to the girl*"). Other issues frequent for both languages are incorrect relative clause resolution and incorrect coordinating conjunction resolution; however, both are more typical for ExtrHech system. Relative clauses can be more common and complicated for Spanish language because relative pronouns readily take prepositions (e.g. *en el cual* vs. less common *in which*), although this needs linguistic proof. Resolution of coordinating conjunction is implemented differently in each system. The English-language system would sometimes either leave out all but the first of coordinated elements or consider all coordinated elements as one argument. Consequently, they were either not counted as extracted assertions at all or considered as one correct extraction.

Interestingly, for neither of the systems incorrect POS-tagging is among top causes of errors, due to the high precision of the modern POS-taggers.

Several issues are encountered only for one language. Non-contiguous relation phrases, although present in English too, are more common in Spanish since they can be caused by free word order. Over-specification vs. under-specification of relational phrases is causes by differences in system implementation. Another observation is that ReVerb does not attempt detecting facts in 5 sentences from the dataset. In this experiment, ExtrHech showed more robust behavior.

6. Conclusions and Future Work

We have presented the Open IE system for Spanish language, ExtrHech, based on syntactic and lexical constraints. It takes a POS-tagged text as an input and outputs a list of extracted binary relations per sentence.

It ExtrHech performs at the precision and recall levels comparable with the state-of-the-art systems for English based on similar approach, i.e. syntactic and lexical constraints and POS-tagging. 87% precision and 70% recall were obtained for the dataset with grammatically correct sentences, and 55% precision and 49% recall were observed on the raw Web text dataset, which included incorrect or incoherent sentences. Although the recall of ExtrHech is lower than that of the syntactic parsing based systems, the precision is at the same level, and the speed is much higher.

We also performed the analysis of errors in extractions made by ReVerb and ExtrHech system from the parallel English-Spanish dataset of 68 grammatically correct sentences. It shows that the major error causes are common for both languages. Interestingly, incorrect POS-tagging is not among the major issues for extraction errors. There are sets of issues that are typical either for one language. Some of them are related to the language properties, others are caused by systems' implementation differences. However, ExtrHech was more robust on the dataset used in the experiment.

Future work includes detailed analysis on how POS-tagger accuracy affects POS-tag based Open IE. We also plan to conduct a comparative experiment for an English-Spanish parallel or comparable dataset containing incoherent or incorrect sentences to better understand the robustness in different languages. Additionally, we will continue improving ExtrHech's handling of the inverse word order, relative clauses, and coordinating conjunctions.

Acknowledgements. The work was partially supported by European Union via 7th FWP (*Web Information Quality Evaluation Initiative*, WIQ-EI, project 269180), Government of Mexico via CONACYT (50206-H), IPN (SIP 20131702), and Mexico City Government (ICYT PICCO10-120).

References

1. *Aguilar Galicia H.* (2012), Extracción automática de información semántica basada en estructuras sintácticas, MSc thesis, IPN, Mexico.
2. *Banko M., Cafarella M. J., Soderland S., Broadhead M., and Etzioni O.* (2007), Open information extraction from the Web, *Proc. IJCAI 2007*, pp. 2670–2676.
3. *Banko, M., & Etzioni, O.* (2008). The Tradeoffs between Open and Traditional Relation Extraction. *Proc. ACL 2008*.
4. *Etzioni O., Banko M., Soderland S, and Weld D. S.* (2008), Open information extraction from the web, *Commun. ACM* 51(12):68–74.
5. *Etzioni, O.* (2011). Search needs a shake-up. *Nature*, 476(7358), pp. 25–26.
6. *Fader A., Soderland S., and Etzioni O.* (2011), Identifying relations for open information extraction, *Proc. EMNLP 2011*, pp. 1535–1545.
7. *Horn C., Zhila A., Gelbukh A., Kern, R., and Lex E.* (2013), Using Factual Density to Measure Informativeness of Web Documents, *Proc. NoDaLiDA 2013,* in print.
8. *Kim J., Moldovan, D.* (1993), Acquisition of semantic patterns for information extraction from corpora, *Proc. of 9th IEEE AIA*, pp. 171–176.
9. *Kirkpatrick M.* (2011), New 5 Billion Page Web Index with Page Rank Now Available for Free from Common Crawl Foundation, readwrite.com/2011/11/07/common_crawl_foundation_ announces_5_billion_page_w.
10. *Kozareva, Z., Hovy, E., & Rey, M.* (2010). Not All Seeds Are Equal: Measuring the Quality of Text Mining Seeds. *Proc. HLT 2010,* pp. 618–626.
11. *Landis J. R., Koch G. G.* (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics* 33(1):159–174.
12. *Mausam, Schmitz M., Bart R., Soderlund S., and Etzioni O.* (2012), Open Language Learning for Information Extraction, *Proc. EMNLP 2012*.
13. *Padró L., Collado M., Reese S., Lloberes M., and Castellón I.* (2010), FreeLing 2.1: Five Years of Open-Source Language Processing Tools, *Proc. LREC 2010*.
14. *Soderland S.* (1999), Learning Information Extraction Rules for Semi-Structured and Free Text, *Machine Learning* 34(1–3):233–272.
15. *Wu F., Weld D. S.* (2010), Open information extraction using Wikipedia, *Proc. ACL 2010*, pp. 118–127.