Open Information Extraction from Real Internet Texts in Spanish Using Constraints over Part-Of-Speech Sequences: Problems of the Method, Their Causes, and Ways for Improvement

Alisa Zhila, Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

Abstract: Usually we do not know the domain of an arbitrary text from the Internet, or the semantics of the relations it conveys. While humans identify such information easily, for a computer this task is far from straightforward. The task of detecting relations of arbitrary semantic type in texts is known as Open Information Extraction (Open IE). The approach to this task based on heuristic constraints over part-of-speech sequences has been shown to achieve high performance with lower computational and implementation cost. Recently, this approach has gained spread and popularity. However, Open IE is prone to certain errors that have not yet been analyzed in the literature. Detailed analysis of the errors and their causes will allow for faster and more focused improvement of the methods for Open IE based on this approach. In this paper, we analyze and classify the main types of errors in relation extraction that are specific to Open IE based on heuristic constraints over part-of-speech sequences. We identify the causes of the errors of each type and suggest ways for preventing such errors, with corresponding analysis of their cost and scale of impact. The analysis is performed for extractions from two Spanish-language text datasets: the FactSpaCIC dataset of grammatically correct and verified sentences and the RawWeb dataset of unedited text fragments from the Internet. Extraction is performed by the ExtrHech system.

Key Words: Open Information Extraction, relation extraction, rule-based methods, error analysis, classification of errors, Spanish, Internet texts, Web texts.

Revista Signos. Estudios de Lingüística, Vol. 49, No. 90, March, 2016.

Extracción abierta de información a partir de textos de internet en español utilizando reglas sobre categorías de palabras en secuencias: Problemas del método, sus causas y posibles mejoras

Resumen: Usualmente, el dominio de un texto arbitrario en Internet se desconoce, así como la semántica de las relaciones que transmite. Mientras que los humanos identifican fácilmente esta información, para una máquina esta tarea está lejos de ser sencilla. La tarea de detectar las relaciones semánticamente arbitrarias en el texto, se conoce como extracción abierta de información (*Open Information Extraction* en inglés). El método para esta tarea basado en reglas heurísticas sobre secuencias de etiquetas de categorías gramaticales de palabras ha demostrado un alto rendimiento con un bajo costo computacional. A pesar de la amplia popularidad de tal enfoque, es propenso a ciertos errores son específicos de este enfoque. Tales errores no han sido analizados en la literatura. En este trabajo, analizamos y clasificamos los principales tipos de errores en la extracción de información. Estos son específicos para el enfoque basado en reglas heurísticas sobre secuencias de etiquetas de categorías gramaticales de palabras. También identificamos las causas para cada tipo de error y sugerimos posibles soluciones, con un correspondiente análisis de su costo y la magnitud del impacto. Hemos realizado el análisis de extracciones a partir de dos conjuntos de textos en español: FactSpaCIC, un conjunto de oraciones gramaticalmente correctas y verificadas, y RawWeb, un conjunto de fragmentos de texto procedentes de Internet sin corrección alguna. La extracción se llevó a cabo con el sistema ExtrHech.

Palabras Clave: extracción abierta de información, extracción de relaciones, métodos basados en reglas, análisis de errores, español, textos en el Internet, textos en la Web.

INTRODUCTION

The amount of textual information on the World Wide Web is constantly increasing. In order to make use of this colossal and growing resource, there is a need to process this information automatically and to convert it into a structured form that allows for its convenient use. The task of automatic extraction of information from texts and its conversion into a structured form is called Information Extraction (IE).

Traditional approaches to IE assume that texts are searched for information of a predefined type. For example, an IE system could be targeted at extraction of all instances that satisfy the relation <COMPANY 1; AQUIRED; COMPANY 2>, as in <*CT Corp; bought; a 40 percent stake in Carrefour's Indonesian unit*>. This is achieved by training the IE system on a large corpus manually annotated for a specific relation and its possible arguments by human annotators (Kim & Moldovan, 1993; Riloff, 1996; Soderland, 1999). Although this approach might be efficient for certain target relations, it requires very expensive resources for training and, more importantly, does not scale to a much larger corpus such as the Web, where the number of possible relations is very large or where target relations cannot be specified beforehand.

Open Information Extraction (Open IE) overcomes these restrictions and thus can be used to extract arbitrary information from texts without requiring specification of a target relation or its possible arguments. In Open IE, a unit of extracted information is presented as a "relation tuple," most commonly a triple <Argument 1; Relation; Argument 2>. For example, a relation tuple corresponding to the phrase "*Benito Juárez nació en San Pablo Guelatao, Oaxaca*" ("Benito Juárez was born in San Pablo Guelatao, Oaxaca") is <*Benito Juárez; nació en; San Pablo Guelatao, Oaxaca*>. An entire tuple can be also referred to as a relation, although it is more common to speak about a relation and its arguments.

In Open IE, the range of extracted information is not restricted by a predefined vocabulary or list of relations, as is the case with the traditional IE. This is achieved by automatic identification of so-called relation phrases—phrases that convey relations between entities in a sentence (Banko, Cafarella, Soderland, Broadhead, Etzioni, 2007). This makes Open IE suitable for automatic processing of Web texts.

In addition, web-scale text processing requires high-speed performance to cope with the huge amount of texts in the Web. Within Open IE, this is achieved by employing various heuristics in the form of constraints based on rules over parts-of-speech (POS) tags (Fader, Soderland, Etzioni, 2011, Zhila & Gelbukh, 2013).

This approach to information extraction attracts significant attention due to its high potential in multilingual applications, since the only language preprocessing required for it

is POS-tagging, which is often more reliable and is available for a greater number of languages than, for example, syntactic parsers (Zhila & Gelbukh, 2013).

The requirement of high processing speed implies that the extraction rules should be sufficiently simple and general to allow for fast application and processing. This, however, inevitably leads to errors and imprecisions specific to this approach. Though Fader et al. (2011) mentioned this issue, no detailed study of the errors has been reported so far. Such analysis, without doubt, would accelerate and focus the improvement of the methods based on this approach to Open IE.

In this paper, we give extensive analysis and classification of errors in relation extraction that are specific to Open IE based on POS tagging and syntactic constraints. The analysis is performed for extractions from texts in the Spanish language. The experiments were conducted for two datasets: FactSpaCIC, a dataset of grammatically correct sentences collected from school textbooks (Aguilar Galicia, 2012), and RawWeb, a dataset of sentences randomly collected directly from the Web without any preprocessing except language detection (Horn, Zhila, Gelbukh, Lex, 2013). In the experiments, extraction was performed using ExtrHech, a state-of-the-art Open IE system for Spanish that implements an extraction method following the approach based on constraints over POS-tag sequences.

We classify the errors into four groups, as well as analyze and describe the causes of these errors. These causes are grouped into eleven classes basing on their nature. For each class, we suggest a solution. Further, we analyze to what extent each class affects information extraction and suggest the optimal strategies for improvement of methods for Open IE based on rules over POS-tag sequences.

The paper is organized as follows. In Section 1, we overview different approaches to Open IE and formulate the problem addressed in this paper. In Section 2, we describe the formal grammar expressions used for relation extraction from texts in Spanish language. In Section 3, we present the tools and datasets used in our work. We describe ExtrHech, an Open IE system for Spanish, the datasets used, and our evaluation methodology. In Section 4, we analyze the errors in information extraction and introduce the classification of the types of errors in Open IE. In Section 5, we describe issues and problems that cause those errors, and suggest solutions to these problems. In Section 6, we give further analysis

and discussion. We show that improvement of different issues affects the output to a larger or smaller extent, and the cost of the improvement also varies. We propose solutions that would produce higher performance improvement at a lower cost. Finally, in Section 7 we draw conclusions and outline future work.

1. Related Work and Problem Statement

Automatic information extraction is an important stage of text processing that embraces detection of potentially useful and informative fragments of text and their extraction in a structured or semi-structured form. Open IE is the task of extracting arbitrary relations with their corresponding arguments from text.

The task was defined by Etzioni, Banko, Soderland, and Weld (2008), who proposed an approach to it based on semi-supervised learning of relation patterns. This approach was explored and implemented in such systems as TextRunner (Banko et al., 2007), WOE^{pos} and WOE^{parse} (Wu & Weld, 2010), and OLLIE (Mausam, Schmitz, Bart, Soderlund, Etzioni, 2012). Their drawback is frequent incoherent or uninformative extractions.

Another category comprises Open IE methods based on relation extraction heuristics or rules. According to the classification of Open IE methods given by Gamallo (2014), it can be divided into two sub-categories. Methods of the first subcategory apply rules over deep syntactic parsing: ClausIE (Del Corro & Gemulla, 2013), CSD-IE (Bast & Haussmann, 2013), KrakeN (Akbik & Loser, 2012), DepOE (Gamallo, Garcia, Fernández-Lanza, 2012), and FES (Aguilar, 2012). These methods are prone to slow performance (Aguilar, 2012), are less robust to grammar errors in texts due to the nature of automatic syntactic parsing, and their implementation is not easily available for many languages—again, due to the limited availability of syntactic parsers, especially for commercial use.

Methods of the other subcategory rely on rules over POS tags, such as those used in ReVerb (Fader et al., 2011), ExtrHech (Zhila & Gelbukh, 2014), and LSOE (Castella Xavier, Souza, Strube de Lima, 2013). The POS tag-based approach has been shown to be very promising in terms of speed, ease of implementation, and portability to other languages (Zhila & Gelbukh, 2013). This makes it appropriate for information extraction on the Web scale.

Although fast and robust, the systems based on rules over POS-tag sequences are not error-free. Fader et al. (2011) showed that, although the ReVerb Open IE system for English achieves good precision and usually avoids incoherent extractions that are typical for the systems of the former subcategory, it still generates at least 20% of erroneous extractions.[1]
Fader et al. (2011) also attempted to describe the errors present in extractions generated by their system. However, as we show in Sections 4 and 5, their analysis was superficial: they do not distinguish between errors and their causes and do not provide any exhaustive classification of errors.
In this paper, we present a detailed analysis of types of errors along with their causes and possible ways of correction for Open IE for Spanish. We classify the errors into those caused by the rule-based nature of the approach, the POS tagging at the preprocessing stage, the diversity of style and level of grammaticality of texts found in the Internet, and to the implementation of the system. A clear vision of the types of errors and their causes will allow faster and more focused improvement of the method. It will also clarify performance boundaries for POS tag-based methods for Open IE, and show what improvements are achievable without using the costly deep syntactic parsing.

2. Grammar Formalisms for Information Extraction

The heuristic or rule based approach to Open IE detects a relation in a text in two steps: first, a relation phrase, which is assumed to convey a binary relation, is detected; then a pair of relation arguments is found. The first argument is considered an agent of the relation, and the second, a general object of the relation.
In our formalism, a relation phrase is limited to be either a single verb (e.g., *estudia* "studies") or a verb immediately followed by dependent words until a preposition (e.g., *atrae la atención de* "attracts attention of" or *nació en* "was born in") optionally followed by infinitive (e.g., *sirven bien para acentuar* "serve well to emphasize"). The corresponding formal grammar expression for a verb phrase is:

(1) $VREL \rightarrow V\,(W^*\,P)?$

where the question mark indicates that the expression in parentheses is optional and the asterisk indicates zero or more occurrences. The variable *V* stands for either a verb optionally preceded by a reflexive pronoun (e.g., *se caracterizaron* "were characterized") or a participle (e.g., *relacionadas* "related"); *W* stands for a noun, an adjective, an adverb, a pronoun, or an article; *P* stands for a preposition optionally followed by an infinitive.

Note that with this definition, a relation phrase is not equivalent to a verb, a phrasal verb, a phraseological unit, or the grammatical predicate of a sentence. The heuristics works in a majority of cases, yet extracted phrases might not correspond to a syntactic unit of the text.

Another rule of our formal grammar describes noun phrases:

(2) $NP \rightarrow N\ (PREP\ N)?$

where *N* stands for a noun optionally preceded by a determiner (e.g., *los gobernantes* "the governors"), an adjective (e.g., *los primeros homínidos* "the first hominids"), a number (e.g., *3.5 millones* "3.5 million"), or their combination. The noun can optionally be followed by a single adjective (e.g., *el epíteto heroico* "heroic epithet"), a single participle (e.g., *las fuentes consultadas* "the consulted sources"), or both (e.g., *los documentos escritos antiguos* "the ancient written documents"). *PREP* stands for a single preposition (e.g., *de* "of"). In our grammar, a noun phrase can be either a single noun with optional modifiers (e.g., *el pueblo griego* "the Greek people") or a noun with optional modifiers followed by a dependent prepositional phrase that consists of a preposition and another noun with its corresponding optional modifiers (e.g., *la historia de la civilización romana* "the history of Roman civilization").

To extend the coverage of our grammar, we use a rule for coordinating conjunctions:

(3) $COORD \rightarrow Y \mid COMMA\ Y?$

where a bar denotes a choice of a variant; *Y* stands for a coordinator (e.g., *y* "and", *o* "or", *pero* "but"), and COMMA stands for a comma. A coordinating conjunction can be either a coordinator or a single comma optionally followed by a coordinator (e.g., ",_*y*", ",_and").

We also use a relative pronoun rule:

(4) $QUE \rightarrow PR$

where *PR* stands for a relative pronoun (e.g., *que* "that", *cual* "which", etc.); this rule is used for resolution of relative clauses; see Section 3.1.

These simple rules are implemented in ExtrHech system in the form of regular expressions over sequences of POS tags for Spanish language.

3. Tools and Datasets

3.1 ExtrHech, an Open IE System for Spanish

In this section we give implementation details of ExtrHech[2], an Open IE system for Spanish based on constraints over POS-tag sequences. Its processing pipeline is shown in Figure 1.

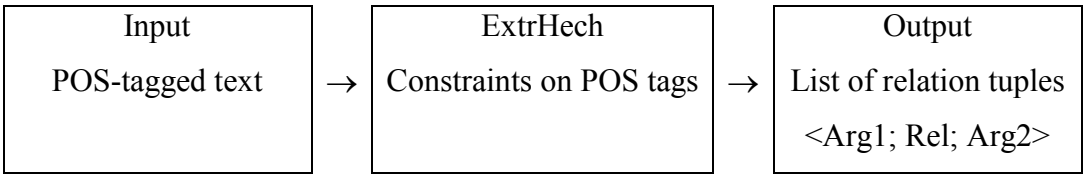


Figure 1. Processing pipeline of ExtrHech.

The system takes a POS-tagged text as input. For POS-tagging, we use Freeling-2.2 (Padró, Collado, Reese, Lloberes, Castellón, 2010), which employs the EAGLES POS tagset for Spanish. An example of a POS-tagged sentence is given in Table 1, where the first row shows the words as they appear in the sentence, the second row shows lemmatized forms of the words, and the third row shows the corresponding POS tags from EAGLES POS tagset.

Table 1. Example of a sentence POS-tagged by Freeling-2.2:
("The Arabic numbering system comes from India").

Word	*La*	*numeración*	*arábiga*	*procede*	*de*	*India*	.
Lemma	El	numeración	arábigo	proceder	de	india	.
POS tag	DA0FS0	NCFS000	AQ0FS0	VMIP3S0	SPS00	NP00000	Fp

Freeling-2.2 requires the Spanish language input to be encoded in the ISO encoding. Nevertheless, texts from Internet are encoded in different encodings. In the RawWeb

dataset described in Section 3.2, the majority of texts extracted from the Internet are encoded in UTF-8. We use an extra preprocessing step to convert UTF-8 texts into ISO.

ExtrHech performs sentence-by-sentence processing. Given a sentence, it first searches for a verb phrase that matches a regular expression (1) described in Section 2. In the current version, ExtrHech chooses the longest match. Next, it looks to the left of the verb phrase for a noun phrase that could be the first argument of the relation. In case of success, it searches to the right of the verb phrase for a second argument. The patterns for noun phrases are implemented in the form of the regular expression (2) from Section 2.

Additionally, if a noun is followed by a participle clause that ends in another noun, the participle phrase is also converted into a relation tuple. Consider the following sentence:

Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.

("The Egyptians were characterized by their beliefs related with (the) death."[3])

ExtrHech generates two relation tuples:

<Arg1 = *Los egipcios*; Rel = *se caracterizaron por*; Arg2 = *sus creencias*>

<Arg1 = *sus creencias*; Rel = *relacionadas con*; Arg2 = *la muerte*>,

the first one corresponding to the main verb of the sentence and the other one to the participle clause.

ExtrHech processes coordinating conjunctions for verb relations and noun phrase arguments with a rule implemented according to the expression (3). For example:

La civilización China nos heredó el papel, la pólvora y la brújula.

("The Chinese civilization gave us (the) paper, (the) powder, and (the) compass."[3])

Resolution of the coordinating conjunctions results in extraction of three tuples:

<Arg1= *La civilización China*; Rel = *nos heredó*; Arg2 = *el papel*>

<Arg1= *La civilización China*; Rel = *nos heredó*; Arg2 = *la pólvora*>

<Arg1 = *La civilización China*; Rel = *nos heredó*; Arg2 = *la brújula*>

Relations are also extracted from relative clauses that are detected by a rule corresponding to the expression (4). A relative pronoun (e.g., *que*, *cual*, etc.) is discarded, and the left argument of the relation tuple is looked for to the left of the relative pronoun. For instance:

Los primeros griegos se organizaron en grupos que tenían lazos familiares.

("The first Greeks were organized in groups that had family relations.")

two relations are detected:

<Arg1 = *Los primeros griegos*; Rel = *se organizaron en*; Arg2 = *grupos*>,

<Arg1 = *grupos*; Rel = *tenían*; Arg2 = *lazos familiares*>,

the first one being detected in the main clause and the second one, in the relative clause.

For Spanish language, the current version is adjusted to the EAGLES POS tagset to treat properly reflexive pronouns for verb phrases.

The current version of ExtrHech has various limitations. It does not resolve anaphora, zero subject construction, and free or inverse word order, such as (Indirect)Object–Verb–Subject instead of Subject–Verb–(Indirect)Object, which can occur in Spanish. The inverse order-related limitation and a possible solution are discussed in Sections 5 and 6.

Zhila & Gelbukh (2013) compared performance of ExtrHech with that of systems based on a similar method for English language. Zhila (2014) showed that ExtrHech outperforms other Open IE systems for Spanish, such as DepOE and FES.

3.2 Datasets

In this paper, we ran ExtrHech on two datasets.[4] The first one, FactSpaCIC (Aguilar Galicia, 2012), comprises 68 sentences from textbooks for the secondary level school, manually checked and verified to be grammatically and orthographically correct and consistent. The second dataset, RawWeb, consists of 159 sentences[5] randomly extracted from CommonCrawl 2012 (Kirkpatrick, 2011), a corpus of web crawl data composed of over 5 billion web pages. This corpus is the initial source of Web texts in our work. RawWeb contains the sentences in their original form as they were collected from the Internet. According to the judgement of a professional linguist and translator, 36 sentences (22% of the dataset) are grammatically incorrect or incoherent (e.g., *cronista cumple del diego video diego el 10* "journalist satisfies of diego video diego the 10"), have orthographical errors that hinder understanding, or are formed of just one word. We kept these sentences in the dataset, because they exemplify incoherent textual data that can be encountered as a part of a poorly edited article or social network comment.[6]

The total size of our datasets, 227 sentences, is comparable with the 300–500 sentences used in other similar works (Mausam et al., 2012, Fader et al., 2011, Wu & Weld, 2010).

3.3 Evaluation methodology

Evaluation of the quality of Open IE output is an area where human judgment is necessary (Fader et al., 2011, Banko et al., 2007). Following the widely accepted methodology for evaluation by human annotators, we ran ExtrHech on our two datasets, and then two human judges independently evaluated each extraction as correct or incorrect. The judges were postgraduate students in computer science and natural language processing who qualified for this task. For FactSpaCIC, a dataset of grammatically correct sentences, they agreed on 89% of the extractions, with Cohen's kappa agreement coefficient $\kappa = 0.52$, which is considered to be moderate agreement (Landis & Koch, 1977) and is typical for this kind of research (Fader et al., 2011). For the RawWeb dataset of 159 sentences, the judges agreed on 70% of the extractions, with $\kappa = 0.40$, which is considered at lower bound of moderate agreement. We defined the number of correct extractions by averaging by the two judges.

4. Main Types of Errors

The main novelty of our work is the analysis of the errors in Open IE specific to the approach based on heuristic constraints over sequences of POS tags. Apart from a very brief analysis of incorrect extractions in (Fader et al., 2011), where errors are not distinguished from their sources, no detailed study of errors in Open IE and their sources has been reported so far. In this paper, we distinguish between errors and their sources and provide corresponding classifications for both, based on the analysis of errors found in the extractions from the FactSpaCIC and RawWeb datasets.

To build our classification of error types, we start from the only type included in the error analysis in (Fader et al., 2011), namely, "Correct relation phrase, incorrect arguments." Then, we add a classification based on the component location of the error in an extracted tuple. We also introduce classes for special cases. This component-based approach to classification guarantees that the classification is complete, that is, each possible error falls into at least one of the types.

Incorrect relation phrase. This type comprises errors resulting in incorrectly detected relation phrase. For example, consider the sentence:

Louis Botha llevó a cabo numerosas manifestaciones públicas.

("Louis Botha organized (*lit.* brought to accomplishment) numerous public manifestations."[3])

The relation detected in this sentence by ExtrHech is:

<Arg1 = *Louis_Botha*; Rel = *llevó a*; Arg2 = *cabo numerosas*> (incorrect).

The relation phrase in this extraction is incorrect: it should be "*llevó a cabo*", i.e., include the noun *cabo* of the idiom *llevar a cabo*, lit. "to bring to accomplishing", which is not supported by the current algorithm. In addition, since a verb relation phrase is the first element of the relation tuple looked for by ExtrHech, incorrect detection of the relation phrase in most of the cases leads to incorrect detection of the arguments. In this example, the second argument should be *numerosas manifestaciones públicas*.

Incorrect argument(s). In this case, at least one of the arguments of the relation tuple is detected incorrectly. For the newspaper article title:

Opositor a la guerra de Irak liberado de arresto militar.

("Opponent of the war of Iraq liberated from military arrest."),

the relation generated by our system is:

<Arg1 = *Irak*; Rel = *liberado de*; Arg2 = *arresto militar*> (incorrect).

The first argument of this extraction *Irak* is incorrect, it should be *Opositor a la guerra de Irak*. In this example, the argument is underspecified (incomplete), i.e., shorter than it should be.

Correct relation phrase, incorrect argument(s). This type of errors is introduced to better understand when errors in arguments are not provoked by the reasons causing incorrect relation phrase detection. Consider the sentence:

María Eva estaba embarazada de dos meses. ("Maria Eva was pregnant at two months.")

and the corresponding extraction:

<Arg1= *María Eva*; Rel = *estaba embarazada de*; Arg2 = *meses* > (incorrect).

The right argument of the relation omits the numeral *dos*. In this particular case, the rules erroneously ignored the numeral POS-tag.

Incorrect order of arguments. In the sentence:

De la médula espinal nacen los nervios periféricos.

("From the spinal cord come the peripheral nerves."[3])

the system detected the relation tuple:

<Arg1 = *la médula espinal*, Rel = *nacen*, Arg2 = *los nervios periféricos* > (incorrect).

As we have mentioned earlier, the first (the left) argument is expected to be the agent or experiencer of the relation, while the second (the right) argument is expected to be the object of the relation. Therefore, the correct order of the arguments is:

<Arg1 = *los nervios periféricos*, Rel = *nacen de*, Arg2 = *la médula espinal* > (correct).

Although complete, this classification is overlapping because the errors included into the "Correct relation phrase, incorrect arguments" category must also be classified as "Incorrect argument(s)" type. However, classifying errors into these classes is useful to better detect and distinguish between their sources.

The distribution of error types by the number of extractions per dataset is shown in Table 5.

Table 5. Error types by the number of extractions for different datasets.

Error type / Dataset	Incorrect relation phrase	Incorrect argument(s)	Correct relation, incorrect argument(s)	Incorrect argument order
FactSpaCIC	9%	22%	16%	3%
RawWeb	21%	45%	26%	6%

As one can see, the distributions of the error types for the two datasets are similar. We have mentioned above that some errors in arguments can result from errors in relation phrase detection. One can observe that the number of extractions with erroneous relation phrases and extractions with correct relations but incorrect arguments together is nearly the same as the number of extractions with incorrect arguments. Further, the number of extractions with incorrect relation phrases is nearly half of the number of extractions with incorrect arguments for both datasets. Hence, solving the problems in the relation phrase detection is very likely to correct a substantial number of errors in argument detection. Therefore solving the problems in relation phrase extraction should be given high priority.

On the other hand, errors in argument order are the least common for both datasets.

5. Main Sources of Errors, and Possible Solutions

In this section, we analyze the issues that cause the errors in information extraction. Some of the issues, namely, N-ary relations, non-contiguous relation phrase, overspecified relation phrase, and incorrect POS-tagging, have been introduced by Fader et al., (2011). However, after a thorough analysis of each error in the extractions returned by ExtrHech on the two datasets, we identified several other major issues. We also describe the previously mentioned issues in more detail. Our list is not exhaustive, and some other issues might be detected in a larger dataset or a dataset that includes texts with unusual language structures, such as poems. However, we assume that those issues are not common and could be included into the group "Other" of our classification. Below, we give a list of the identified issues with examples and suggest solutions.

1) **Underspecified noun phrase**. For the sentence:

La agrupación de seres humanos en un mismo espacio favoreció el intercambio de conocimientos.

("The grouping of human beings in the same place favored the interchange of knowledge.") one would expect the following extracted relation:

<Arg1 = *la agrupación de seres humanos en un mismo espacio*;
Rel = *favoreció el intercambio de*; Arg2 = *conocimientos*>.

However, the system generated the extraction:

<Arg1 = *un mismo espacio*; Rel = *favoreció el intercambio de*; Arg2 = *conocimientos*>,

where the first argument *un mismo espacio* is underspecified: it is a dependent part of the complete argument *la agrupación de seres humanos en un mismo espacio*.

To prevent underspecification of noun phrases, either more complex POS-based rules should be introduced or syntactic analysis can be added to the preprocessing stage. Yet introduction of syntactic parsing as an additional preprocessing procedure will inevitably increase the running time and computational cost of information extraction.

2) **Overspecified verb phrase**. Consider the sentence:

La Botánica ha logrado analizar las características de la vegetación.

("The Botany has achieved analyzing the characteristics of the vegetation.")

The corresponding extracted tuple reads:

<Arg1 = *la Botánica*; Rel = *ha logrado analizar las características de*;

Arg2 = *la vegetación*>.

The relation phrase *ha logrado analizar las características de* is extracted in accordance with the condition of the longest match for a verb phrase. In fact, the relation phrase should be shorter: *ha logrado analizar*. The correct corresponding extracted tuple is:

<Arg1 = *la Botánica*; Rel = *ha logrado analizar*;

Arg2 = *las características de la vegetación*>.

This example shows how an error in relation detection leads to an error in argument detection: overspecification of the relation led to underspecification of the argument.

The solution suggested by Fader et al., (2011) is, first, to perform a massive information extraction on a large corpus, and then to consider only the relations with frequencies above a certain threshold as valid relations. They compiled a dictionary of about 2 million "valid" relations and looked up each proposed relation in the dictionary ("lexical constraint"). However, 23% of missed extractions in their work were filtered out by this constraint. This solution affects the ability of the system to extract arbitrary relations from the Web.

3) **Non-contiguous verb phrase**. As our analysis shows, this issue is closely related to the free word order in Spanish. For example, in the phrase:

bajo cuyo nombre pueden entrar los sextants

("under whose name can appear the sextants")

the relation phrase should be *pueden entrar bajo el nombre de*. It is non-contiguous in the source text, some parts of it preceding other parts. This is a complex language phenomenon difficult to treat even by modern syntactic parsers.

4) **N-ary relation or preposition**. This issue comprises two similar phenomena. First, there are prepositions requiring more than one object, such as *entre* "between":

En América la agricultura inició entre el 8000 y el 5000 a.C.

("In America the agriculture began between (the) 8000 and (the) 5000 B.C.")

In this case the expected extraction is

<Arg1 = *la agricultura*; Rel = *inició entre*; Arg2 = *el 8000 y el 5000 a.C.*>,

where the coordinate structure *el 8000 y el 5000 a.C.* is governed by the preposition *entre* and should not be broken into two arguments. Yet according to our current treatment of conjunctions described in Section 3.1, the system splits the right argument into two:

<Arg1 = *la agricultura*; Rel = *inició entre*; Arg2 = *el 8000*>

<Arg1 = *la agricultura*; Rel = *inició entre*; Arg2 = *el 5000*>,

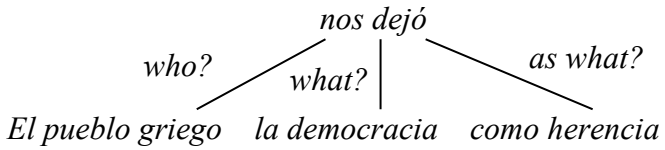
which leads to two incorrect extractions. This issue can be resolved by introducing an auxiliary dictionary of prepositions that suppresses argument splitting by a conjunction.

Non-binary relations are another source of extraction errors, that is, the relations that connect more than two entities. For example, in the sentence:

El pueblo griego nos dejó como herencia la democracia.

"The Greek people left us as heritage (the) democracy."[3]

the relations between the components are:



Therefore, the extraction:

<Arg1 = *el pueblo griego*; Rel = *nos dejó*; Arg2 = *herencia*>

is incorrect. The current approach to information extraction considers only binary relations. Hence, this issue cannot be covered within this approach.

5) **Conditional subordinate clause** or an adverb that affect the semantics of the original sentence. Consider the sentence:

Los primeros homínidos eran recolectores y sólo comían carne cuando encontraban los restos abandonados por otros animales. ("The first hominids were gatherers and only ate meat when (they) found the leftovers abandoned by other animals.")

The sentence tells us that the first hominids ate meat only under some specific conditions, namely, when they encountered leftovers. Therefore, the extraction:

<Arg1 = *los primeros homínidos*; Rel = *sólo comían*; Arg2 = *restos*>

does not reflect the meaning conveyed by the sentence, although it seemingly makes sense. In fact, the relation presented in the sentence is non-binary and should be skipped or ignored within the current approach to Open IE. This can be solved by introducing an auxiliary dictionary of conditional phrases, which is not resource-consuming.

6) **Incorrectly resolved relative clause**. This issue concerns resolution of prepositional relative clauses, such as in:

El lugar en el que florecieron las culturas más desarrolladas del México antiguo

("The place in (the) which flourished the most developed cultures of ancient Mexico"[3]),

<Arg1 = *el lugar*; Rel = *florecieron*; Arg2 = *las culturas*>.

The correct extraction would be:

<Arg1 = *las culturas más desarrolladas del México antiguo*; Rel = *florecieron en*;
Arg2 = *el lugar*>.

The source of this type of errors is the complex syntactic structure of the sentence itself. Therefore, their solution requires syntactic analysis.

7) **Incorrectly resolved coordinate structures**. As described in Section 3.1, ExtrHech breaks coordinate structures into separate extractions. It usually does it correctly when the conjunction occurs either between relation phrases or between the arguments of a relation. However, when a conjunction occurs within an argument, it may be resolved incorrectly:

los cambios climáticos que crearon un ambiente propicio para la reproducción y la selección de plantas. ("the climatic changes that created an environment appropriate for the reproduction and the selection of plants.")

The correct pair of relations conveyed by this phrase is:

<Arg1 = *los cambios climáticos*; Rel = *crearon un ambiente propicio para*;
Arg2 = *la reproducción de plantas*>, <Arg1 = *los cambios climáticos*;
Rel = *crearon un ambiente propicio para*; Arg2 = *la selección de plantas*>.

Yet the system erroneously extracts the following tuples:

<Arg1 = *los cambios climáticos*; Rel = *crearon un ambiente propicio para*; Arg2 = *la reproducción*>, <Arg1 = *los cambios climáticos*; Rel = *crearon un ambiente propicio para*; Arg2 = *la selección de plantas*>,

where the first extraction lacks the dependent part of the second argument, thus leaving it underspecified. This phenomenon can be treated by selective application of relation component splitting at coordinating conjunctions. This can be done by syntactic parsing the sentences where a conjunction occurs between noun phrases. Selective syntactic parsing triggered by a certain POS-tag sequence is a better solution than parsing of all sentences. Further analysis is needed to determine how many sentences include a coordinating conjunction between noun groups and whether this approach will solve a significant amount of errors at a lower cost than the syntactic parsing of all sentences.

8) **Inverse word order**. This phenomenon occurs when the dominating word order Subject–Verb–(Indirect)Object is inversed to (Indirect)Object–Verb–Subject:

De la médula espinal nacen los nervios periféricos.

("From the spinal cord originate the peripheral nerves")

Currently our system is designed to process only the direct word order. This leads to incorrect extractions:

<Arg1 = *la médula espinal*; Rel = *nacen*; Arg2 = *los nervios periféricos*>.

This illustrates the "Incorrect argument order" type of errors.

9) **Incorrect POS-tagging**. This issue occurs at the preprocessing stage and results in errors in extractions at a later stage. For example, in the sentence:

La soldada tapa resguarda un rico cóctel cardiosaludable

("The soldered tap preserves a rich heart-healthy cocktail")

the word *soldada*, which in this sentence is an adjective ("soldered"), was tagged as a noun ("soldier"). Consequently, the left argument could not be matched by the expression (2) and the extracted tuple was affected by underspecification of the argument:

<Arg1 = *tapa*; Rel = *resguarda*; Arg2 = *un rico cóctel cardiosaludable*>.

Currently, information extraction errors caused by this issue are inevitable, because even the state-of-the-art morphological analyzers commit errors in POS-tagging. The good news is that, as we show in the next section, it is one of the least common sources of errors.
The issues listed above caused errors on both datasets: FactSpaCIC and RawWeb. Below we describe some issues that did not occur in the grammatically correct dataset FactSpaCIC (possibly due to its limited size), yet they occurred on the (larger) RawWeb dataset.

10) **Grammatical errors.** Grammatical errors in the original text, mainly in syntax and punctuation, also lead to incorrect extractions generated by our system. For example, the sentence:

En aquellos dias como en casa hay jardin con muchos arbolitos nos encanta andar trepado prrrr es una delicia ("In those days as at home there is a garden with many trees we love climbing trees prrrr it's a delight"[3])

lacks various punctuation marks, which hinders its understanding even by human readers. In this case, the system was unable to generate correct extractions. Since grammatical errors in real-world texts are inevitable, so far there is no obvious solution for this issue.

11) **Others**: idioms, relations involving adjectives, etc. About 7.5% of errors were caused by issues that were classified neither into one of the above classes nor into specially introduced classes, because of their low frequency. For example, in the sentence:

Maradona prepara la lista para enfrentar a España

("Maradona prepares the list for withstanding (to) Spain")

the last preposition *a* is not matched by the rule (1). Low frequency of errors caused by such issues does not justify endeavors targeted specifically at their solution.

6. Discussion

In this section, we analyze to what extent errors of each type are caused by each issue, and solutions to which issues would produce higher improvement at a lower cost.
Figure 2 shows what types of errors are caused by each issue, where numbers 1 to 11 correspond to the classification presented Section 5. The columns labeled with R

correspond to the errors encountered in the RawWeb dataset. The columns labeled with F correspond to the FactSpaCIC dataset (issues 10 and 11 were not encountered in FactSpaCIC; the former, due to *a priori* grammatical correctness of the dataset, and the latter, likely because of its simplified school textbook language and small size).

As one can observe, issues 2, 3, 4, and 6 have quite similar distributions of the types of errors between the datasets. This is a good indication of what types of errors could be treated by targeting at the issues of this group. The distributions for issues 1, 7, 8, and 9 are also quite similar between the datasets, despite an additional type of errors encountered for each issue in the RawWeb dataset. This can be explained, first, by a larger size of this dataset and, second, by the wider variation of language constructions used in Web forums or comments. Hence, these two groups of issues support the hypothesis that certain issues are prone to raise errors of certain types in any dataset.

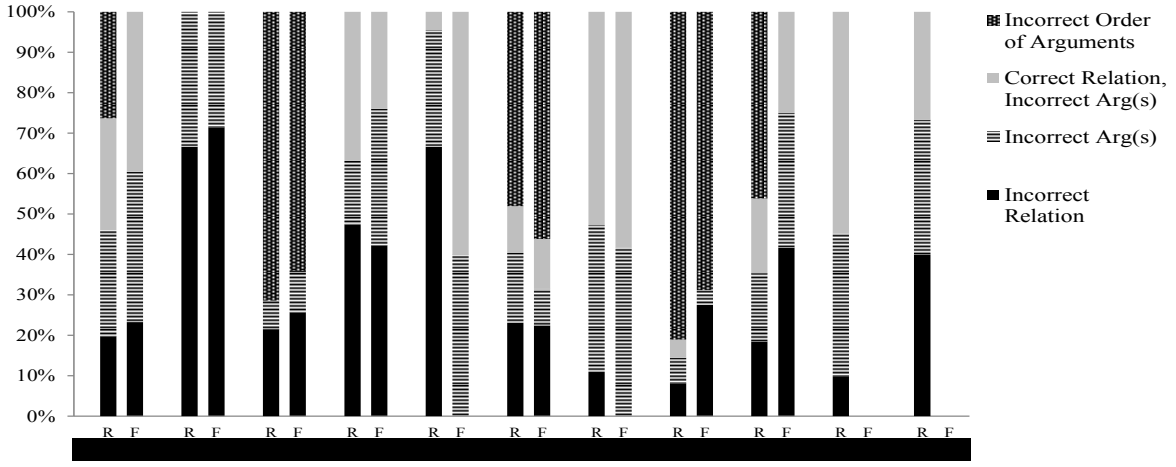


Figure 2. Distribution of types of errors by issues for each dataset. R stands for the RawWeb dataset, F for FactSpaCIC. The issues are indicated with numbers: 1: underspecified noun phrase, 2: overspecified verb phrase, 3: non-contiguous verb phrase, 4: N-ary relation, 5: conditional clause, 6: relative clause, 7: coordinate structure, 8: inverse word order, 9: incorrect POS-tagging, 10: grammatical errors, 11: others.

From Figure 2, it might seem that the issue 5 has different distributions of error types for the two datasets. However, Figure 3a shows that the total count of errors caused by the

issue 5 (conditional clause) is as low as 2, i.e., the FactSpaCIC dataset simply does not provide sufficient data on this issue. Since conditional clauses are ubiquitous in texts, we can expect that the errors caused by this issue will generally have a distribution similar to the one in the RawWeb dataset, where 10 errors of this type were observed (Figure 3b).

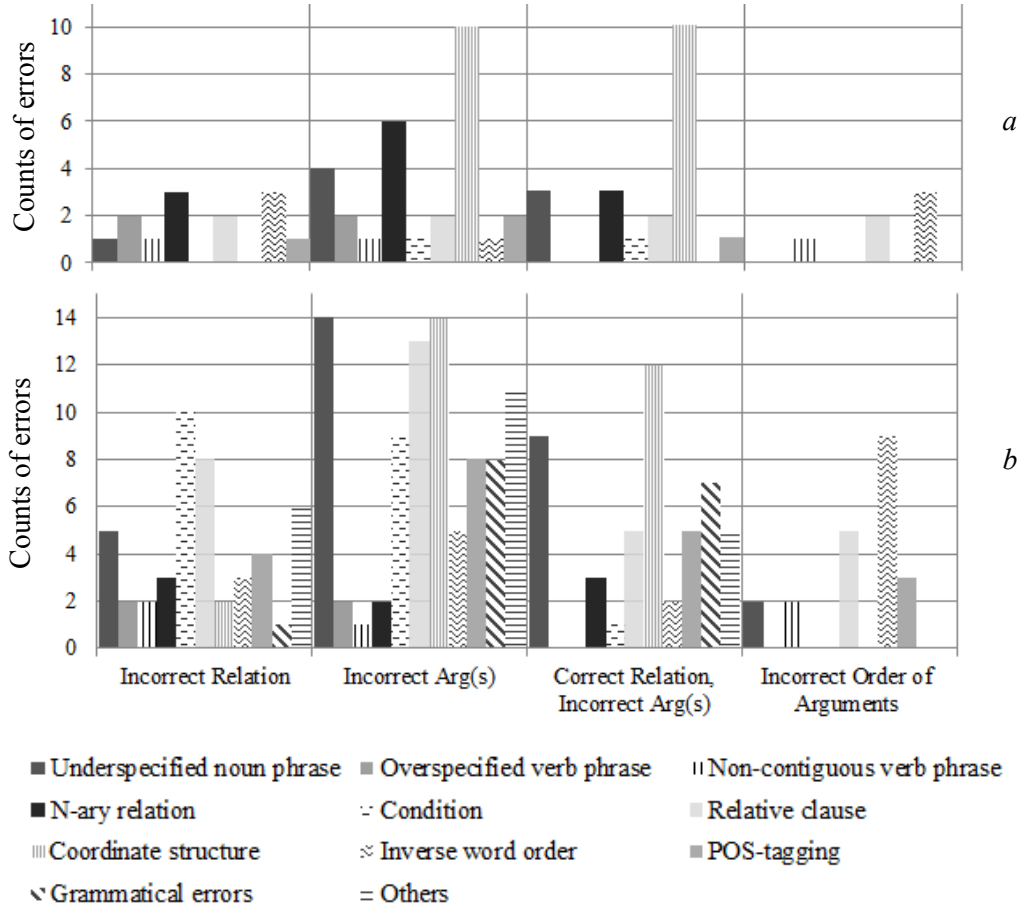


Figure 3. Counts of errors raised by each issue in FactSpaCIC (a) and RawWeb (b).

As expected, incorrect extractions caused by grammatical errors in original sentences (issue 10) are only encountered in RawWeb. Other issues (11) embrace diverse phenomena that were not encountered in FactSpaCIC, but were encountered in RawWeb with a frequency insufficient to allow their detailed description and classification.

Now that we have discussed the similarities and differences in distributions of errors between the datasets, let us take a closer look at the sources of errors and possible solutions.

There are several groups of issues that mostly cause errors of a particular type. As shown in Figure 2, overspecified verb phrases (2) and N-ary relations and prepositions (4) mostly affect the relation phrase detection and, consequently, the argument detection (as explained in Section 4). Therefore, elimination of the corresponding sources will significantly reduce the number of errors of these types. However, it is not clear whether these issues can be solved by introducing syntactic analysis alone or whether they need an introduction of lexical knowledge, e.g., a list of N-ary prepositions (which seems feasible) or even a list of verbs with N-ary government, such as *dar algo a alguien* "give something to somebody" (which is an open problem).

Non-contiguous verb phrases (3) and inverse word order (8) mostly cause errors in the order of arguments. This can be approached by adding rules for inverse order detection, e.g., checking for a preposition before the left noun phrase. We have tried a straightforward implementation of this rule, but it raised incorrect extractions in other sentences, as in:

Uno de los inspectores se dejó llevar por su experiencia

(similar in structure to "One of the inspectors has acted by the rules"[3])

that looks similar to a passive construction, and the corresponding erroneous extraction:

<Arg1 = *su experiencia*; Rel = *se dejó llevar por*; Arg2 = *de los inspectores*>,

which is, again, an incorrect argument order error, now found in a sentence with the direct word order. An efficient solution must be more complex. However, as Figure 3 shows, the incorrect argument order is the least common error type, primarily caused by the inverse word order phenomenon. According to Brown & Rivas (2011), Verb–Subject word order is much less common for Spanish than Subject–Verb word order (for example, 4% vs. 47% for Puerto Rican Spanish). This leads to the conclusion that this issue can be left alone and only be approached as a side effect of a solution to a major issue.

Incorrect resolution of coordinate structures (7) mostly affects argument detection. This is an interesting observation since the mechanism of coordinate structure resolution is the same for conjunctions between verb phrases and noun phrases.

A similar situation is observed for grammatical errors (10): although they are expected to affect all components of the extractions, detection of arguments is affected more frequently than that of relation phrases. To verify this, further experiments with a larger dataset are

needed. Grammatical errors are inherent to informal communication. A preprocessing stage of intelligent automatic grammar correction could solve this problem. However, this lies far beyond the area of information extraction.

Noun phrase under-specification (1), relative clause resolution (6), and incorrect POS-tagging (9) have more or less even distributions of errors of all types. As discussed in Section 5, incorrect POS-tagging happens at the pre-processing stage of morphological analysis and is out of the scope of information extraction. Improvement of relative clause resolution and noun phrase boundary detection is much more complicated since these are complex language phenomena. Deeper syntactic analysis can help treat them correctly.

Other issues (11) that include a variety of problems might require a specific approach to each of them ranging from dictionaries of idioms to introduction of specific rules for syntactic parsing. Yet in our experiments their counts are low, which suggests that they are not common and their handling will not give substantial improvement.

To summarize, many errors could be eliminated by introducing syntactic analysis. However, syntactic parsing is computationally expensive. It would require conversion of a POS-tagged based system into a syntax-based system, losing the advantages in speed. A possible compromise between preserving the high-speed performance of a POS tag-based approach and higher precision of a syntax-based approach is "triggered" syntactic parsing. In this case, only sentences with a special structure, e.g., with a coordinating conjunction or a relative pronoun, will be fully parsed, and corresponding more complex rules will be applied. This is a promising direction for future work.

CONCLUSIONS

We have analyzed in detail the errors typical to the method of Open IE based on heuristic rules over POS-tags. No detailed description or accurate classification of the errors had been reported before, although some types of errors along with some issues were mentioned by Fader et al. (2011), but not distinguished. We have distinguished between errors and their sources. We have classified all information extraction errors into four types based on the component of an extracted fragment where an error occurred: incorrect relation phrase,

incorrect arguments, incorrect argument order, and incorrect arguments with correct relation phrase. This classification is complete: it covers all possible errors.

We have performed error analysis for two datasets: the FactSpaCIC dataset of grammatically correct verified sentences and the RawWeb dataset of texts directly extracted from the Internet. We have shown that the distributions of types of errors are similar for both datasets.

In addition, we have analyzed what problems and phenomena cause each type of errors. We described ten distinct and frequent issues, as well as minor issues grouped as "others." Among these issues, only the issue "Grammatical errors in the original sentence" is specific to the Web-based dataset. We have also analyzed the distribution of errors by the issues and proposed directions for solutions.

Some directions seem more promising than others. First, a triggered syntactic analysis approach, when full syntactic parsing is performed only for sentences containing certain triggers (e.g., coordinating conjunctions or relative pronouns), could be a good compromise between the speed of a POS-based approach and the precision of a syntax-based one, solving a broad range of problems. Second, introduction of dictionaries of closed word classes, such as N-ary prepositions or subordinating conjunctions for conditional clauses, can solve a few issues at a lower cost in speed and implementation. Yet dictionaries of open word classes, such as verbs with N-ary government, do not seem to be a feasible solution.

In our future work, we plan to experiment with a larger dataset to detect with higher precision which group of issues is more frequent. We also plan to experiment with simplified syntactic representations (Sidorov, 2013, 2014) and more advanced features (Sidorov, Gelbukh, Gómez-Adorno, Pinto, 2014) and rules (Sidorov, Kobozeva, Zimmerling, Chanona-Hernández, Kolesnikova, 2014).

REFERENCES

Aguilar Galicia, H. (2012). *Extracción automática de información semántica basada en estructuras sintácticas*. MSc thesis, Instituto Politécnico Nacional, Mexico.

Akbik, A. and Loser, A. (2012). Kraken: N-ary Extractions in Open Information Extraction. *Proc. AKBC-WEKEX 2012*, 52–56.

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. *Proc. IJCAI 2007*, 2670–2676.

Bast, H. and Haussmann. E. (2013). Open Information Extraction via Contextual Sentence Decomposition. *Proc. ICSC 2013*, 154–159.

Brown, E., Rivas, J. (2011) Verb Word-Order in Spanish Interrogatives: A Quantitative Analysis of Puerto Rican Spanish. *Spanish in Context*, 8(1), 23-49.

Castella Xavier, C., Souza, M., and Strube de Lima, V. (2013). Open Information Extraction based on Lexical-Syntactic Patterns. *Proc. Brazilian Conference on Intelligent Systems*, 189–194.

Del Corro, L. and Gemulla, R. (2013). ClausIE: Clause-based Open Information Extraction. *Proceedings of the World Wide Web Conference (WWW-2013)*, 355–366.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying Relations for Open Information Extraction. *Proc. EMNLP 2011*, 1535–1545.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open Information Extraction from the Web. *Commun. ACM*, 51(12), 68–74.

Gamallo, P. (2014). An Overview of Open Information Extraction, *Proc. SLATE'14*, 13-16.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-based Open Information Extraction. *In ROBUS-UNSUP 2012*, 10–18.

Horn, C., Zhila, A., Gelbukh, A., and Lex, E. (2013). Using Factual Density to Measure Informativeness of Web Documents. *Proc. NoDaLiDA 2013*; *Linköping Electronic Conference Proceedings,* 85, 227–238.

Kim, J., Moldovan, D. (1993). Acquisition of Semantic Patterns for Information Extraction from Corpora, *Proc. of 9th IEEE Conference on AI for Applications*, 171–176.

Kirkpatrick, M. (2011). *New 5 Billion Page Web Index with Page Rank Now Available for Free from Common Crawl Foundation* [online]. Available at: http://readwrite.com/2011/11/07/

Landis, J.R., Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.

Mausam, Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. (2012). Open Language Learning for Information Extraction, *Proc. EMNLP 2012*.

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*.

Riloff, E. (1996). Automatically Constructing Extraction Patterns from Untagged Text. *Proc. AAAI 1996*, 1044–1049.

Sidorov, G. (2013). N-gramas sintácticos no-continuos. *Polibits* 48, 69–78.

Sidorov, G. (2014). Should Syntactic N-grams Contain Names of Syntactic Relations? *International Journal of Computational Linguistics and Applications* 5(1), 139–158.

Sidorov, G., Kobozeva, I., Zimmerling, A., Chanona-Hernández, L., and Kolesnikova, O. (2014). Modelo computacional del diálogo basado en reglas aplicado a un robot guía móvil. *Polibits* 50, 35–42.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas* 18(3), 491–504.

Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1–3), 233–272.

Wu, F., Weld, D.S. (2010). Open Information Extraction Using Wikipedia. *Proc. ACL 2010*, 118–127.

Zhila, A. (2014). *Open Information Extraction using Constraints over Part-of-Speech Sequences*. PhD thesis, Instituto Politécnico Nacional, Mexico.

Zhila, A., Gelbukh, A. (2014). Open Information Extraction for Spanish Language based on Syntactic Constraints. *Proc. ACL SRW 2014*, 78–85.

Zhila, A., Gelbukh, A. (2013). Comparison of Open Information Extraction for Spanish and English. *Computational Linguistics and Intellectual Technologies*, 12(1), 794–802.

NOTES

[1] At a recall level of 20% or higher

[2] Available at https://bitbucket.org/alisa_ipn/extrhech

[3] In English glosses, we try to preserve relevant aspects of the grammatical structure of the examples, even at the expense of awkward English or changes in the meaning.

[4]Available at http://www.gelbukh.com/resources/spanish-open-extraction-extraction.

[5] Initially, we applied a sentence splitter to the corpus and randomly selected 200 resulting chunks. 41 of them were not natural language sentences (e.g., parts of programming code or numerical data) and thus were irrelevant for the information extraction task.

[6] As we explained above, we eliminated from the dataset the chunks that were obviously not in a human language. However, we kept the chunks that doubtlessly represented expressions in a human language written by human users, even if ungrammatical or unintelligible. Technically, it is fast and easy to distinguish between what looks like text and obviously non-language data, whereas deciding whether a chunk represents a correct, meaningful, coherent sentence would require full-blown artificial intelligence, and anyway the threshold would be too vague. Thus, we just kept all chunks that looked like valid text.