



INSTITUTO POLITÉCNICO NACIONAL



Centro de Investigación en Computación
Laboratorio de Procesamiento de Lenguaje Natural

**Open information extraction using constraints
over part-of-speech sequences**

T E S I S

**QUE PARA OBTENER EL GRADO DE
DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

PRESENTA

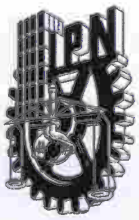
M. en C. ALISA ZHILA

DIRECTOR:

Dr. Alexander Gelbukh

México, D.F.

Diciembre de 2014



INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D. F. siendo las 16:00 horas del día 12 del mes de noviembre de 2014 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Open information extraction using constraints over part-of-speech sequences"

Presentada por la alumna:

ZHILA

Apellido paterno

ALISA

Nombre(s)

Apellido materno

Con registro:

A	1	1	0	8	9	8
---	---	---	---	---	---	---


aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron **APROBAR LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Director de Tesis

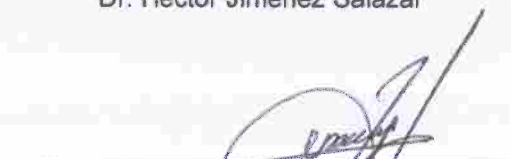

Dr. Alexander Gelbukh


Dr. Sergio Suárez Guerra


Dr. Grigori Sidorov


Dr. Héctor Jiménez Salazar



Dr. Luis Villaseñor Pineda


Dr. Francisco Miram Calvo Castro



PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN
EN COMPUTACIÓN
DIRECCIÓN


Dr. Luis Alfonso Villa Vargas

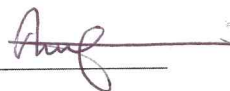


INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México, D.F. el día 28 del mes de noviembre del año 2014, el (la) que suscribe Alisa Zhila alumno (a) del Programa de Doctorado en Ciencias de la Computación con número de registro A110898, adscrito al Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Alexander Gelbukh y cede los derechos del trabajo intitulado Open Information Extraction using Constraints over Part-of-Speech Sequences, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección alisa.zhila@gmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Alisa Zhila 

Nombre y firma

Resumen

En la década de 2010 unos cuantos exabytes de datos se producen a diario. Aproximadamente entre $\frac{1}{5}$ y $\frac{1}{3}$ de estos datos es texto. Para hacer uso de esas enormes cantidades de datos, se requiere detectar, extraer, estructurar, y procesar la información importante de una manera rápida y escalable. Extracción de información abierta (Open IE) es una solución para la detección, extracción, y la estructuración inicial de la información.

Open IE es un paradigma de la extracción de información independiente del dominio del texto realizada de manera no supervisada. Por eso el rendimiento de alta velocidad y escalabilidad son sus mayores ventajas. Eso convierte Open IE en un campo muy atractivo para la investigación y para aplicación a otras tareas de procesamiento del texto.

En este trabajo se realizó una amplia investigación sobre diversos métodos para Open IE y sobre su aplicación a otras tareas, y hemos contribuido de varias maneras.

Hemos introducido un método para Open IE que requiere un pre-procesamiento mínimo del texto de entrada que asegura su velocidad y robustez. Además, hemos propuesto este método para el idioma español. El método resultó ser superior a otros métodos al menos en uno de los dos aspectos: ya sea en términos de precisión o en términos de robustez. Como una contribución adicional, introducimos un método para comparar el rendimiento de los sistemas de Open IE implementados para diferentes idiomas mediante la comparación de las salidas sobre los conjuntos de datos paralelos.

Hemos presentado un método para Open IE con pre-procesamiento semántico adicional, que permite la interpretación semántica de las extracciones, lo cual no era posible con otros métodos. Se demostró que este método tiene una precisión muy alta, aunque a un alto costo del rendimiento. Lo más importante es que logramos demostrar la estructuración semántica de extracciones con un nuevo procedimiento para presentación de extracciones en el formato RDF/XML, que es un formato estándar mantenido por el W3C.

Además, hemos demostrado que Open IE puede servir para medir la informatividad de documentos de dominio arbitrario extraídos de la Web “como son”, sin pre-procesamiento adicional. Eso manifiesta que Open IE puede servir para una tarea compleja que tiene un impacto directo en un usuario final.

Y por último, tenemos a disposición del público el software y los recursos de evaluación desarrollados como parte de este trabajo.

Abstract

In 2010's several exabytes of data are produced daily. Approximately between $\frac{1}{5}$ and $\frac{1}{3}$ of these data is text. To make use of such huge amounts of textual data, we need to be able to detect, to extract, to structure, and to process important information conveyed through this data flow in a fast and scalable manner. Open information extraction (Open IE) is a solution for detection, extraction, and initial structuring of information.

Open IE is open-domain and relation-independent paradigm for information extraction performed in an unsupervised manner. This makes high-speed performance and scalability its main advantages, converting Open IE into a very perspective field for research and for applications to other text data processing tasks.

In this work we have conducted an extensive research on various methods for Open IE and on its application to other tasks, and contributed in several ways.

First, we have introduced an Open IE method requiring minimal pre-processing of input that assures its speed and robustness. Additionally, we proposed this method for Spanish language and showed it to be superior to other methods at least in one of the two aspects: either in terms of precision or in terms of robustness. As an additional contribution, we also introduced a method for performance comparison of Open IE systems implemented for different languages by comparing outputs for parallel datasets.

Next, we introduced a method for Open IE with additional semantic pre-processing that allows semantic interpretation of the extracted relations, which was not possible with other Open IE methods. We showed that this method has a very high precision, although at a cost of yield. Most importantly, we demonstrated the application of this method to semantic structuring of extractions by introducing a novel procedure of extraction presentation in RDF/XML format that is a standard format maintained by W3C.

Further, we showed that Open IE can serve for measuring of Web document informativeness that is one of the aspects of document quality. Not only did we show that it can be applicable to the arbitrary domain documents extracted from the Web “as is”, without additional pre-processing, we also showed that Open IE can serve for a complex text processing task that has a direct impact on an end-user.

And the last but not least, we made publicly available the software and evaluation resources developed as parts of this work.

Acknowledgments

I would like to thank Honorato Aguilar-Galicia for the initial version of labeled dataset FactSpCIC, my father Vladimir Zhila for proofread and my mother Svetlana Zhila for continuous moral support, and my advisor Dr. Alexander Gelbukh for all the opportunities that he has given to me during my time in the Center for Computing Research and his titanic endeavors to make this work complete. I would also like to thank the members of my Tutorial Committee: Dr. Grigori Sidorov, Dr. Sergio Suárez Guerra, Dr. Hiram Calvo Castro, Dr. Luís Villaseñor-Pineda, and Dr. Héctor Jiménez-Salazar, for their advice and helpful observations. The financial support from Consejo Nacional de Ciencia y Tecnología (CONACyT, National Council of Science and Technology) and Instituto Politécnico Nacional in the form of BEIFI scholarship made possible the fulfillment of this work.

Contents

Resumen	1
Abstract	2
Acknowledgments	3
List of Figures	7
List of Tables	8
Glossary	9
1 Introduction	11
1.1 Open Information Extraction	12
1.2 Motivation	13
1.3 Hypothesis	14
1.4 Objectives	14
1.4.1 General objective	14
1.4.2 Particular objectives	15
1.5 Contributions	15
1.5.1 Theoretical contributions	15
1.5.2 Practical contributions	15
1.6 Structure of the document	16
2 State of the Art	18
2.1 Main Strategies for Open IE methods	19
2.2 Selected Rule-Based Algorithms for Open IE	21
2.2.1 REVERB's algorithm	21
2.2.2 DEPOE's algorithm	21
2.2.3 Limitations	22

2.3	Open IE for Languages Other than English	23
2.4	Evaluation Technique for Open IE Performance	24
2.5	Applications of Open IE	25
3	Framework	26
3.1	Levels of Automatic Language Analysis	27
3.2	Generic Rule-based Open IE Algorithm	30
3.3	Relevant Language-Specific Properties of Spanish	31
4	Open Information Extraction based on Rules over Part-of-Speech Tags	33
4.1	POS-tag Patterns for Information Detection	34
4.1.1	Verb relation pattern	34
4.1.2	Noun argument pattern	35
4.1.3	Patterns for complex syntactic structures	35
4.2	Our Algorithm	36
4.3	ExtrHech System	37
4.4	Experiments and Results	38
4.4.1	Experiments on Spanish language dataset	39
4.4.2	Robustness evaluation	40
4.4.3	Discussion of experiments on Spanish language dataset	41
4.4.4	Experiment on parallel Spanish and English datasets	42
4.4.5	Experiment on Raw Web dataset	43
4.4.6	Comparative table for various Open IE methods	45
4.5	Limitations	46
4.6	Errors in Open Information Extractions	47
4.6.1	Main types of errors	47
4.6.2	Main issues that cause errors and possible solutions	50
5	Named-Entity-Driven Open Information Extraction	58
5.1	Motivation	59
5.2	Named-Entity-Driven Open Information Extraction with Post-Processing Rules	62
5.2.1	Open information extraction constrained by named entities	62
5.2.2	Reported speech extraction	63
5.2.3	Detection of target relations in post-processing	64
5.2.4	Illustration of the method	66

5.3	Experiments of Performance Evaluation	67
5.4	Conversion into RDF/XML format	69
5.4.1	Format Validation	71
5.5	Discussion	71
6	Application to Measuring Informativeness of Web Documents	73
6.1	Motivation	74
6.2	Previous Work in Text Quality Evaluation	75
6.3	Building the Ground Truth Dataset	76
6.3.1	The dataset	76
6.3.2	Ground truth ranking by human annotators	77
6.4	Automatic Measurement of Document Quality	79
6.5	Experiment and Results	80
6.6	Discussion	82
7	Conclusions and Future Work	84
7.1	Conclusions	85
7.2	Contributions	85
7.2.1	Theoretical contributions	86
7.2.2	Practical contributions	86
7.3	Limitations	87
7.4	Future Work	87
7.5	Publications	88
7.6	Awards and Invited Talks	89
	Bibliography	90
	Appendix	98
A	Running ExtrHech	98
B	Regular Expressions for Open Information Extraction	99

List of Figures

1.1	Statistics on the use of languages in the Internet. Images extracted from Wikipedia [63]	13
3.1	The basic NLP pipeline	27
3.2	A syntactic parsing dependency tree for sentence “ <i>Yesterday, New York based Foo Inc. reported that they had acquired Bar Corp.</i> ” . . .	29
4.1	Processing pipeline of EXTRHECH system	38
4.2	Structure of a verb-based N-ary relation	53
5.1	Structure of the extraction.	67
5.2	An RDF graph corresponding to the extraction.	70
6.1	Process of corpus generation	76
6.2	Screenshot of the MaxDiff questionnaire tool	79
6.3	Diagram of the factual density estimation	80
6.4	Rank correlation	82

List of Tables

2.1	Extraction rules of dependency-parsing based Open IE algorithm . . .	22
4.1	Example of a POS-tagged sentence by Freeling-2.2.	38
4.2	Comparison of performance of rule-based Open IE systems for Spanish	40
4.3	Comparison of pre-processing robustness for rule-based Open IE systems for Spanish	41
4.4	Comparison of extraction robustness for rule-based Open IE systems for Spanish	41
4.5	Performance comparison of REVERB and EXTRHECH systems over a parallel dataset.	43
4.6	Performance of EXTRHECH on the grammatically correct dataset and the dataset of noisy sentences extracted from the Web	44
4.7	Comparative data for various Open IE systems.	45
4.8	Distribution of error types by the number of returned extractions for different datasets.	49
5.1	Performance of the named-entity-first Open IE method and ReVerb Open IE system at the same recall level.	68
6.1	Classification of the documents in the dataset by the types of text content	77
6.2	Human annotator ranking and factual density ranking	81
6.3	Result of correlation tests between factual density algorithm ranking and human annotator ranking for 50 document dataset	82

Glossary

Constituent A fragment of a sentence that functions as a single unit within a **syntactic** structure of a sentence.

Gold Standard Set of sample instances with correct labeling. Normally such instances are labeled by human experts.

Ground Truth see **Gold Standard**.

Lemma A dictionary **wordform** of the word. For example, for the lemma of a verb is its infinitive.

Named Entity, NE A conventional text unit, normally, a word or a phrase, that signifies an instance of one of the following classes: **Person**, **Location**, **Money unit**, **Organization**, etc. A different list of named entity classes may be defined for a particular task or application.

Named Entity Recognition, NER A task of automatic identification of named entities and assignment of a correct NE class.

Natural Language Processing, NLP A field that lies in the intersection of computer science, artificial intelligence, and linguistics that researches on methods and algorithm for automatic processing and understanding texts written in a natural (human) language. For example, the work at hand is a work in NLP field.

Parsing see **Syntactic parsing**.

Part-of-Speech, POS A category to which a word is assigned in accordance with its syntactic functions and morphological characteristics.

POS-tagging A process of assigning a part-of-speech label or tag to a word.

POS tag A tag or a label assigned to a word that indicates its part-of-speech. Normally it also includes other grammatical information. Examples are NN for noun, NNP for proper noun, VBD for past tense verb, JJ for adverb.

Semantics In linguistics, a branch of the field that studies meaning of text at different levels. Most often at the level of words, phrases, and sentences.

Sentence splitting An natural language processing task consisting in automatic detection of limits of a sentence.

Syntax A branch of linguistics that studies and analyzes which arrangements and orders of words and phrases create well-formed sentences in a language. Also these arrangements and orders.

Syntactic parsing An NLP task consisting in automatic detection of a correct syntactic structure of a sentence. Normally, its output is a syntactic tree that is a graph of syntactic dependencies in the sentence and each node is assigned its syntactic role, e.g., VP for verb phrase, PP for prepositional phrase, NP for noun phrase.

Syntactic Chunk Entire subtree of a syntactic tree that corresponds to the constituents at the next-to-the-root level.

Token A separable contiguous sequence of characters between spaces or punctuation marks. Punctuation marks are tokens as well.

Word sense disambiguation (WSD) Task consisting of choosing the meaning of a given word.

Wordform A particular form of a word. Words in text are present in the one of their wordforms. For example, *containing* is word form of the verb *contain*.

Yield In NLP, the size of a method's output, e.g., the number of returned extractions.

Chapter 1

Introduction

Open information extraction serves for automatic analysis of vast amounts of texts of arbitrary domain and extraction of information in the form of tuples consisting of a relation and its arguments. Various approaches to Open information extraction are designed to perform in a fast and unsupervised manner. In this work we will explore methods based on minimal pre-processing and see whether they can work for Spanish language.

In this chapter we will discuss the following topics:

- What is Open IE;
- Why methods for Open IE are needed for Spanish language;
- Hypothesis;
- The objectives of the current work;
- Contributions of the work;
- Structure of this document.

1.1 Open Information Extraction

In the world of massive availability of textual information, the humanity needs methods for its real-time processing. The first step towards its fulfillment is to be able to detect and extract potentially important, informative, and worthy text fragments. Open Information Extraction (**Open IE**) is a task of extraction of such fragments from large amounts of natural language text and their representation in shallow semantic form [50], in particular, a form of triples consisting of a relation and its arguments. The key characteristics of Open IE are (1) domain independence, (2) unsupervised extraction, and (3) scalability to large amounts of text. For example, analysis of the sentence

“Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806.”
(“Benito Jarez was born in San Pablo Guelatao, Oaxaca, in 1806.”)

by methods of Open IE should return extractions

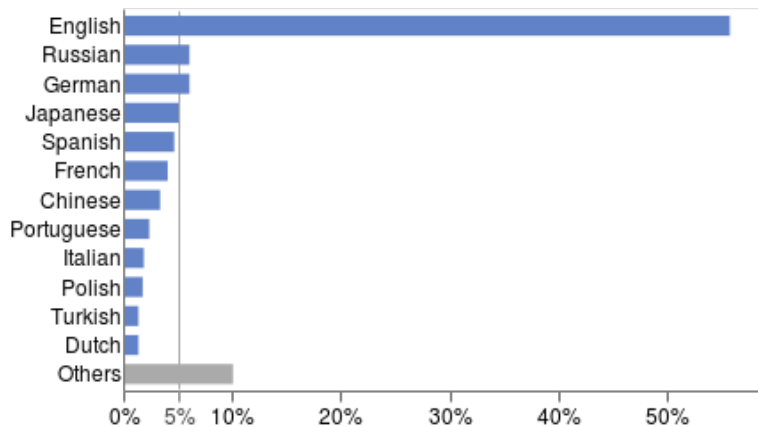
$\langle \text{Arg1} = \textit{Benito Juárez} \rangle \langle \text{Rel} = \textit{nació en} \rangle \langle \text{Arg2} = \textit{San Pablo Guelatao, Oaxaca} \rangle$

and

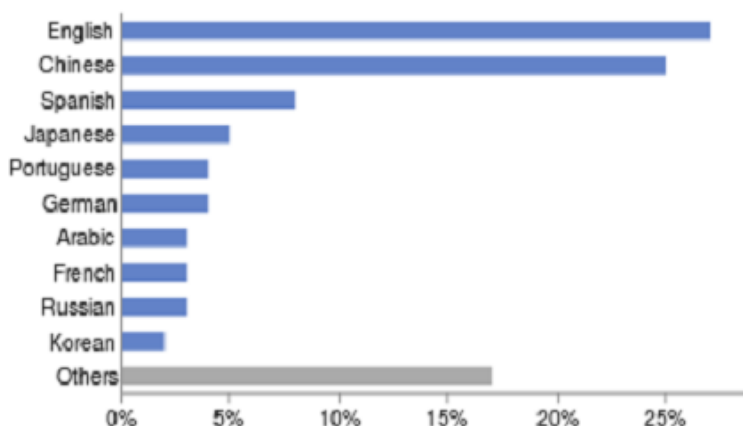
$\langle \text{Arg1} = \textit{Benito Juárez} \rangle \langle \text{Rel} = \textit{nació en} \rangle \langle \text{Arg2} = \textit{1806} \rangle.$

Open IE was introduced by Banko et al. [6] 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 domain and relation independence and 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 pre-specified [7].

Output of Open IE can be used either directly for text summarization [26] or be a part of more complex natural language processing tasks as ontology population, question answering [55], machine reading, document classification [52, 18], knowledge graph construction, automatic quality evaluation of texts, web page semantic parsing, opinion mining [51], etc.



(a) Content languages for websites as of November 2014



(b) Internet users by language as of December 2013

Figure 1.1: Statistics on the use of languages in the Internet. Images extracted from Wikipedia [63]

1.2 Motivation

Generally, to the date not much work has been done in Open IE for languages other than English. It turns out to be really remarkable, given that the super-goal of Open IE is the ability to process texts directly from the Web at real time speed. This goal looks very appealing for current state of information processing. However, the Internet is written in many languages other than English. The recent statistics provided by Q-Success, Software Quality Management Consulting, shows that Spanish is the fifth most frequent language for the content on the Web [15]. Spanish is also the third by the number of Internet users who speak Spanish as estimated by Miniwatts Marketing Group [14]. Figures 1.1a and 1.1b visually demonstrate the corresponding statistics.

Giving that the non-English speaking audience of the Internet is the majority, and Web content for most of the languages other than English is constantly growing [61], the researchers must look for Open IE methods for other languages. In particular, the portion of the Web content in Spanish language has grown from 4.4% to 4.8% from November 2013 to November 2014, which turns out to be the fastest growing language among the top 5 Web content languages.

However, work on Open IE for languages other than English is extremely limited. Even in a quite recent and relatively comprehensive survey on multilingual information extraction and summarization [49] there is actually no information on multi-language Open IE. We will get back to this issue in Section 2.3.

To sum up, the great potential of Open IE and the significance and fast growth of Spanish as a language of Web use, motivated this work on Open IE for Spanish language.

1.3 Hypothesis

Open IE based on rules with minimal level of input pre-processing is possible. It can achieve at least the same or better performance as the Open IE methods that require deeper pre-processing of the input and similar performance as similar Open IE methods for English language. Additionally, one can compare the outputs of methods designed for different languages using a parallel dataset.

1.4 Objectives

Taking into account the motivation and hypothesis stated above, the objectives of our work are:

1.4.1 General objective

To design a method of Open IE from texts that would have the following properties:

- have the main characteristics of Open IE approach, i.e., (1) domain independence, (2) unsupervised extraction, and (3) scalability to Web text;
- achieve better performance than existing methods for Spanish language and at least the same or better performance than similar methods for English language;

- guarantee high robustness;
- be useful for more complex NLP tasks.

1.4.2 Particular objectives

- design rules for Open IE targeted at Spanish language that fulfill the general objective requirements;
- implement the designed method so that the system would show the level of performance defined in the general objective;
- evaluate performance of the system;
- compare to performance of other systems;
- design and implement a method for integration of our Open IE system in a more complex NLP task.

1.5 Contributions

This work has several contributions to the field of natural language processing.

1.5.1 Theoretical contributions

At the theoretical level we have contributed:

- a robust and high-performance rule-based method for Open IE for Spanish language text that require minimal input pre-processing and that is useful for various complex NLP tasks;
- a method of its application to measurement of quality of Internet texts;
- a method based on deeper input pre-processing that leads to deeper semantic interpretation of extracted relations.

1.5.2 Practical contributions

At the practical level we have contributed:

- a software system for Open IE for Spanish language, EXTRHECH¹ based on our method;
- a labeled parallel English-Spanish version of FactSpCIC dataset²;
- a labeled parallel English-Spanish dataset of 300 sentences from news articles³;
- a labeled dataset of 159 sentences in Spanish extracted randomly from the Web ⁴.

All datasets have been used for Open IE performance evaluation.

1.6 Structure of the document

In this section we provide a roadmap of the document to facilitate the reader’s navigation through the document.

Chapter 2 overviews the existing approaches to solution of Open IE task in English and in other languages. We also provide a detailed pseudocode for selected algorithms, against which we compare our method in particular. We also overview existing approaches to Open IE application to other NLP tasks.

Chapter 3 contains a description of a basic NLP pipeline and explains the concept of “levels” of NLP. We also provide a pseudocode of a generic algorithm of rule-based approach to Open IE. Additionally, we provide an overview of some grammatical differences between Spanish and English languages to make clear the non-triviality of the transfer of methods for Open IE.

Chapter 4 introduces our method for Open IE for Spanish language. We also describe various experiments for performance evaluation of our method against other methods for Spanish language as well as a similar method for English language. We also describe the experiment for random Internet texts and show the performance robustness of our method for this type of texts.

Chapter 5 introduces a rule-based method based on named-entity tagged input. We show that this method allows deeper semantic interpretation of the relations between arguments and more semantically granular partition of complex

¹Available for download from https://bitbucket.org/alisa_ipn/extrhech.git

²<http://www.gelbukh.com/resources/spanish-open-fact-extraction#FactSpCIC>

³<http://www.gelbukh.com/resources/spanish-open-fact-extraction#news>

⁴<http://www.gelbukh.com/resources/spanish-open-fact-extraction#RawWeb>

extraction components. Although returning lower yield, this is a highly precise method.

Chapter 6 shows that our method of Open IE for Spanish is applicable to measurement of informativeness of textual contents of arbitrary Web documents. We discuss that it is a good indication of its our method appropriateness for the Web processing.

Chapter 7 summarizes and discusses the results achieved at all stages of our work and derive the contributions of our work. We also outline future perspectives for these direction of research.

Appendix provides information on the execution of EXTRHECH system.

Chapter 2

State of the Art

Open information extraction is a relatively new paradigm for information extraction. This name, Open IE, was introduced in [6] in the early days of this approach. A lot of productive and novel research has been done since then in this field. To begin with, a number of different approaches to its solution varying from rule-based methods to methods incorporating machine learning approaches has been suggested. Despite their variety, all methods for Open IE are language dependent. There is a direction of research on the mulch-lingual Open IE. Further, thanks to its properties such as domain and relation independence, Open IE has been considered for numerous applications varying from ontology population to text quality measure.

In this chapter we will provide an overview of the following topics:

- Main strategies for Open IE and their advantages and disadvantages;
- Evaluation of performance of Open IE methods;
- Work in Open IE for languages other than English;
- Application of Open IE to other tasks.

2.1 Main Strategies for Open IE methods

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: $\langle the\ policeman \rangle \langle saw \rangle \langle a\ boy \rangle$ and $\langle a\ boy \rangle \langle was\ crossing \rangle \langle the\ street \rangle$. 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.

Several approaches to Open IE can be distinguished. The basic idea is that most sentences contain highly reliable syntactic clues to their structure [6].

1. Chronologically the first one was introduced in the works of [6] and [22] that were the pilot works in Open IE. Their approach is based on semi-supervised machine learning principles and includes three main steps: 1) manual labeling of a training corpus for seed relation phrases and features; 2) further semi-supervised learning of relations; 3) automatic extractions of relations and their arguments. This approach is implemented in `TEXTRUNNER` [7], `WOEpos`, and `WOEparse` [64]. In these systems, the detection of a relation starts from the potential arguments expressed as noun phrases, i.e., before the relation phrase is detected. Therefore, a noun that actually belongs to a relation phrase can be marked as an argument. Consequently, the relation phrase cannot be backtracked. Let’s consider relation “*to make a deal with*”. Here *deal* can be erroneously extracted as an argument, although it is a part of the relation. This makes the approach prone to incoherent and uninformative extractions.
2. Rule-based approach includes systems based on rules over outputs of various levels of automatic linguistics analysis. `FES-2012` system for Spanish language [1] applies rules to the fully parsed sentences. However, in the same work the authors show that this approach is too slow to be scaled to a Web-sized corpus and is not robust. Another system implementing rule-based approach is `DEPOE` [25]. In this systems, the rules are applied to the output of shallow dependency parsing. In `REVERB` system [24] syntactic constraints are applied over POS tags and syntactic chunks. Another recent clause-based method for Open IE [19] is based on dependency parsing and a small set of

domain-independent lexica. It is implemented in system CLAUSIE. The methods of this approach show high results in terms of precision-recall, speed of performance, and, consequently, scalability to a Web-sized corpus.

3. Ultimately, the approach based on the deep automatic linguistic analysis is implemented in OLLIE [41]. This system combines various approaches: it uses output of a rule-based Open IE system to bootstrap learning of the relation patterns and then additionally applies lexical and semantic patterns to extract relations that are not expressed through verb phrases. Such a complex approach leads to high precision results with a high yield. However, there is a tradeoff between the accurate output and the cost of implementation and computation, and complexity of the training stage. 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 impractical at Web scale.

Considering the prospective of each of the approaches, the research on the first strategy has been practically abandoned when the second and the third types of strategies were suggested. It has been shown in [24] that the method for Open IE based on rules over POS-tags and syntactic chunks yields much higher performance than the methods from the first cohort. In [41], the authors show that their approach is even superior to both $\text{WOE}^{\text{parse}}$ system implementing the first strategy and REVERB system implementing the second strategy. However, their approach has been shown slower in performance and quite costly in implementation that makes doubtful its advantages over the rule-based approaches.

All these approaches, excluding FES-2012 system, have been evaluated only for English. However, their relation extraction algorithms are language dependent. They use output of various linguistic analysis such as part-of-speech or syntactic dependency information as well as immediate lexical information to define patterns or constraints for relations. Therefore, there has been done little to no research on whether and how these approaches to Open IE will function for other languages.

2.2 Selected Rule-Based Algorithms for Open IE

2.2.1 ReVerb’s algorithm

Here we will outline the Open IE algorithm introduced by Fader *et al.* as it is presented [24]. The algorithm takes as **input** a POS-tagged and syntactically chunked sentence. This requires the following pre-processing of the input sentence:

1. POS-tagging;
2. Syntactic chunking.

Essentially, the algorithm consists of rules over POS-tags and syntactic chunks:

Algorithm 1 REVERB’s Open IE algorithm

Search for a verb, a verb followed immediately by a preposition or an infinitive marker “to”, or a verb followed by nouns, adjectives, or adverbs ending in a preposition or infinitive marker “to”

if detected **then**

 Mark it as a relation phrase

 Search for a noun phrase to the left of the relation phrase

if detected **then**

 Search for another noun phrase to the right of the relation phrase

if detected **then return** the extraction triple

else return false

else return false

else return false

The first step defines a pattern for a relation phrase which limits it to be either a verb (e.g., *reported*), a verb followed immediately by a preposition (e.g., *born in*), or a verb followed by at least one or several nouns, adjectives, or adverbs ending in a preposition (e.g., *opened a new office in*).

Although relation phrases are detected via matching against POS-tag sequences, the arguments are searched as whole noun phrases that has been detected at the pre-processing stage of syntactic chunking.

The actual implementation of this algorithm in REVERB includes lexical and syntactic constraints over detected relation phrases.

2.2.2 DepOE’s algorithm

Another rule-based Open IE algorithm is introduced in [25]. Since it has been implemented for Spanish as well as for English, we provide the outline of this algorithm.

As its **input** the algorithm requires dependency-parsed text which implies corresponding pre-processing. Then, the algorithm works as follows:

Algorithm 2 DEPOE’s dependency-parsing based Open IE algorithm

```

Identify verb clauses in a sentence
if detected then
  for each verb clause do
    Identify the verb participants, including their functions: subject, direct
    object, attribute, and prepositional complements
    if detected then
      Apply a rule return extraction triple
    else return false
else return false

```

The rules are shown in Table 2.1:

Table 2.1: Extraction rules of dependency-parsing based Open IE algorithm

Rule Pattern	Extraction
subj-vp-dobj	$\langle \text{Arg1} = \text{subj} \rangle \langle \text{Rel} = \text{vp} \rangle \langle \text{Arg2} = \text{dobj} \rangle$
subj-vp-vprep	$\langle \text{Arg1} = \text{subj} \rangle \langle \text{Rel} = \text{vp+prep}(\text{prep from vprep}) \rangle$ $\langle \text{Arg2} = \text{np}(\text{from vprep}) \rangle$
subj-vp-dobj-vprep	$\langle \text{Arg1} = \text{subj} \rangle \langle \text{Rel} = \text{vp+dobj+prep} \rangle \langle \text{Arg2} = \text{np}(\text{from vprep}) \rangle$
subj-vp-attr	$\langle \text{Arg1} = \text{subj} \rangle \langle \text{Rel} = \text{vp} \rangle \langle \text{Arg2} = \text{attr} \rangle$
subj-vp-attr-vprep	$\langle \text{Arg1} = \text{subj} \rangle \langle \text{Rel} = \text{vp+attr+prep}(\text{from vprep}) \rangle$ $\langle \text{Arg2} = \text{np}(\text{from vprep}) \rangle$

No language specific adjustments are described in [25] where this algorithm was introduced.

2.2.3 Limitations

We would like to note that both of these algorithms require relation phrases to contain a verb. This generally limits this approach to extraction of relations expressed via a verb and cannot infer from a phrase “*Research Professor Dr. Alexander Gelbukh*” that a person named *Alexander Gelbukh* works as *Research Professor* and holds a *doctorate degree* because these relations are not explicitly expressed via verbs.

Nevertheless, the research community considers it an acceptable trade-off between the simplicity and limitations of the algorithm.

In Section 3.2 we will show that the generalization of REVERB’s algorithm turns out to be a basic sequence for rule-based methods

2.3 Open IE for Languages Other than English

As we mentioned, all approaches described in Section 2.1 require language dependent information for their implementation. The third approach, which is the approach based on the deep automatic linguistic analysis, directly uses lexical information for the context analysis. The other two approaches employ morphological and syntactic information that varies among the languages. Additionally, descriptions of all the methods introduce either lexical patterns as in [6] or lexical constraints as in [24] to detect or filter out relations correspondingly. Naturally, any lexical information is language dependent.

Among the described strategies to solve the task of Open IE, the work for languages other than English has mainly been done using the rule-based approach. Indeed, we have already briefly discussed in Section 2.1 that the rule-based approach generally outperformed the semi-supervised learning approach and has a lot of advantages in ease of implementation and performance speed against the third approach.

To the best of our knowledge, there are only two works to the date that have suggested their Open IE methods for languages other than English. Curiously, both works deal with Spanish language.

The first method was introduced by H. Aguilar-Galicia in [1] and implemented in FES-2012 system. As we mentioned in Section 2.1, his method lies within rule-based method. In particular, it applies extraction detection rules over syntactically parsed text. However, its employment of the full syntactic parsing does not scale to a Web-sized corpus. Aguilar-Galicia shows that the particular implementation of his method ran slowly on the machine on which FES-2012 was developed. We hope this disadvantage shall be overcome with the growth of available computational capacities and optimization of syntactic parsing methods for Spanish. More important shortcoming was the non-robustness of the method. Of the 68 testing sentences approximately a third were not parsed correctly leading to incorrect or no extractions. The test dataset consisted of sentences from secondary school student books.

The paper [25] is another work that claims to have its variants of Open IE solutions for Spanish, Portuguese, and Galician languages. Its method also lies within the rule-based strategy and is based on rules over shallow dependency parsing. The method has been implemented in DEPOE system. However, no experimental results with languages other than English or any language specific details are reported.

2.4 Evaluation Technique for Open IE Performance

Practically all works in Open IE share the same method of evaluation. Normally, the authors would report **precision** and **recall** of their systems.

Precision Precision of an Open IE system is the fraction of the number of returned correct extractions among the total number of returned extractions:

$$\text{Precision} = \frac{\text{correct extractions}}{\text{all returned extractions}} \quad (2.1)$$

Recall Recall of an Open IE system is the fraction of returned correct extractions among all *possible* or *expected* correct extractions:

$$\text{Recall} = \frac{\text{correct extractions}}{\text{all possible correct extractions}} \quad (2.2)$$

We would like to note that there is no commonly shared or publicly available standard dataset for evaluation of Open IE task. Therefore, the divisor in the equation (2.2) for recall calculation becomes difficult or infeasible to estimate and to some extent deliberate. It really depends on the judges and on the given instructions what fragments of text should be considered as *possible* or *expected* correct extractions. Therefore, not all authors report recall for their systems [41, 19]. Instead, they report the total number of extractions.

Additionally to reporting precision and recall for one given state of the system, a few works [64, 24, 41, 19] provided series of numbers or even curves corresponding to different confidence levels for extraction. However, few authors actually explained how they estimated the confidence level of each extraction. For example, Del Corro and Gemulla in [19] simply take their Dependency Parser confidence level. Fader *et al.* and Mausam *et al.* claim that they estimated extraction confidence level based on some confidence classifier preliminary trained on labeled examples. However, a detailed description of the exact implementation and training features used in those works is unavailable.

2.5 Applications of Open IE

In general, the domain- and relation-openness of Open IE methods makes them extremely useful in the settings when a relation cannot be defined in advance, possible semantic classes of arguments are not known, or user needs cannot be known. In particular, Open IE seeks applications in machine reading [21], text summarization [48, 49], new perspective on search as question answering [20], automatic text quality evaluation [28, 36] and many others.

An approach to adaptation of Open IE to domain-specific relations is suggested in [56]. The eventual goal of their attempt to relation detection is to map the extractions against a given domain-specific ontology, which is seen to be a part of a Question Answering task. In their work Soderland *et al.* modify the original Open IE system TEXTRUNNER for a higher recall, i.e. to return larger chunks of texts than conventional Open IE extraction tuples. Then, they apply domain adaptation rules to the output of the system in two stages. First, they introduce rules to detect domain specific classes, i.e. named-entities and semantic classes. This is done by introducing lists of class-specific key words that are manually learned from a training/development set and extended by synonyms. The output of this step is the extractions enriched with semantic and NE tags on certain terms. At the second stage, they apply domain relation mapping rules that are a set of constraints on tuple arguments and on the context to the left and right of the tuple. This extended context is returned due to the modification of their Open IE system for a higher recall. The Open IE adaptation described in the paper is demonstrated for NFL football domain and corresponding 13 relations.

Chapter 3

Framework

Before we can proceed to the main goal of this work, in this chapter we would like to set the reader in the environment of rule-based Open IE. First, we will describe the framework of the rule-based approach to Open IE and provide sufficient information on the types of automatic language analysis needed for this task. Then, we will overview the particular properties of Spanish language that make the task of Open IE non-trivial.

The chapter covers the following topics:

- Different levels of automatic linguistic analysis;
- Generic Open IE algorithm based on rules over part-of-speech tag sequences;
- Some language-specific properties of Spanish.

3.1 Levels of Automatic Language Analysis

Basic Natural Language Processing (NLP) can be considered as a pipeline of different manipulations with a given text. Some of these manipulations are due purely to the algorithmic nature of the computer processing, while others correspond to actual linguistic analysis. The basic NLP pipeline is shown in Figure 3.1. The stages of this pipeline are considered to be basic building blocks for any complex and high-level NLP task such as Information Extraction in particular [3]. However, not all of them are necessary as we will see further in the description of our Open IE method in Chapter 4.

Let us now describe each processing stage and exemplify it by analyzing a text fragment “*Yesterday, New York based Foo Inc. reported that they had acquired Bar Corp. We’ve learned it from the news.*”

Tokenization First, we need to determine the limits of the tokens, i.e., words, punctuation signs, and numbers, whichever are present. For our example sentence the expected output is: *Yesterday , New York based Foo Inc. reported that they had acquired Bar Corp . We ’ve learned it from the news .*

Sentence splitting Now that we have determined separate tokens, we can detect limits of different sentences: *Yesterday , New York based Foo Inc. reported that they had acquired Bar Corp {.} We ’ve learned it from the news {.}*

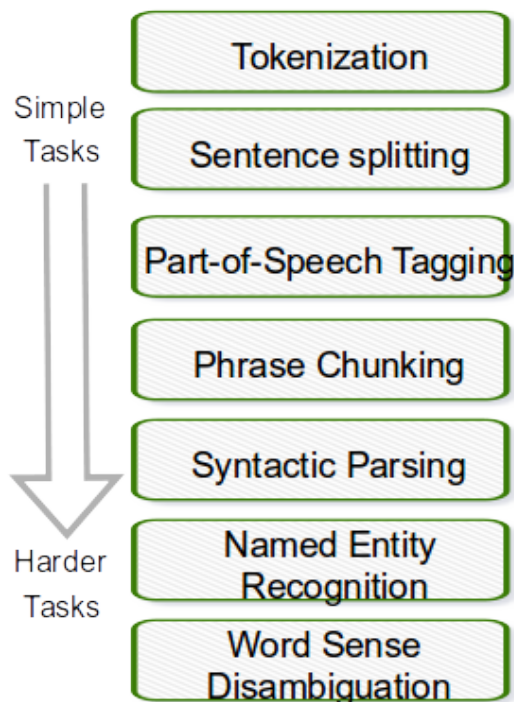


Figure 3.1: The basic NLP pipeline

While naturally performed by humans without any effort, these tasks can give hard times to an automatic processing. As we illustrate in this example, it is actually not obvious which full stop punctuation mark serve for abbreviation and which indicate the end of a sentence. Note, that the full stop at *Corp.* actually serves for both.

Fortunately, these task have been mostly reliably resolved for the Indo-European languages with a large number of speakers including Spanish [9]. In contrast, for Chinese, where word and sentence limits are not marked, these are still open issues.

Further follow the tasks that are actually considered to be stages of linguistic analysis of a language as it is thought of in non-computational linguistics.

Part-of-speech tagging In linguistics known as **morphological analysis**, this is a task of determination of a part of speech for each word as presented in the text. For our example the output would look like: *Yesterday*^JJ , ^P *New*^NNP *York*^NNP *based*^VBN *Foo*^NNP *Inc.*^NNP *reported*^VBD ... This is a non-trivial task because one wordform can correspond to different parts-of-speech as *reports*^(Verb, 3rd person, singular) and *reports*^(Noun, plural). Currently, highly accurate POS-taggers are available for most European languages, e.g., 97 – 98% POS-tagging accuracy for the languages supported by Freeling-3.0 package [45]. However, POS-tagging in other languages still performs with lower accuracy, e.g. ~88% for Bengali [17] or 94.33% for Chinese [39].

Syntactic chunking Also known as **shallow parsing**, syntactic chunking actually is partial parsing. A chunker assigns a partial syntactic structure to a sentence by dealing only with syntactic “chunks”, simplified constituents. For example, [*Yesterday*]^NounPhrase [,]^O [*New York*]^NounPhrase [*based*]^VerbPhrase [*Foo Inc.*]^NounPhrase ... A lot of work has been done on syntactic chunking, and although its performance might vary depending on text genres, it is actually quite stable across different classifiers used for chunking. CONLL 2000 task showed results as high as 92-92% for F-measure¹ [59] while [66] shows that for a general genres its quality reaches $F = \sim 88\%$.

Syntactic parsing Full syntactic parsing is known to be one of the most complex stages of the basic NLP pipeline. The task is to fully determine all syntactic dependencies of a sentence, also known as building its syntactic dependency tree. We show the dependency tree corresponding to our example on Figure 3.2. While syntactic parsing is useful for many NLP tasks [53, 54], syntactic parsing implementations involve computationally costly algorithms with the complexity up to $O(n^3)$ for rule-based chart parsing. Although there is a trade-off between speed and accuracy for syntactic parsing, generally, the

¹**F-measure** is a measure of accuracy and is calculated as a harmonic mean of precision and recall. It is not commonly used for Open IE evaluation due to the difficulties of recall calculation mentioned in Section 2.4.

accuracy is not higher than ~89% while the processing time for a file of 1364 tokens takes dozens of minutes [12].

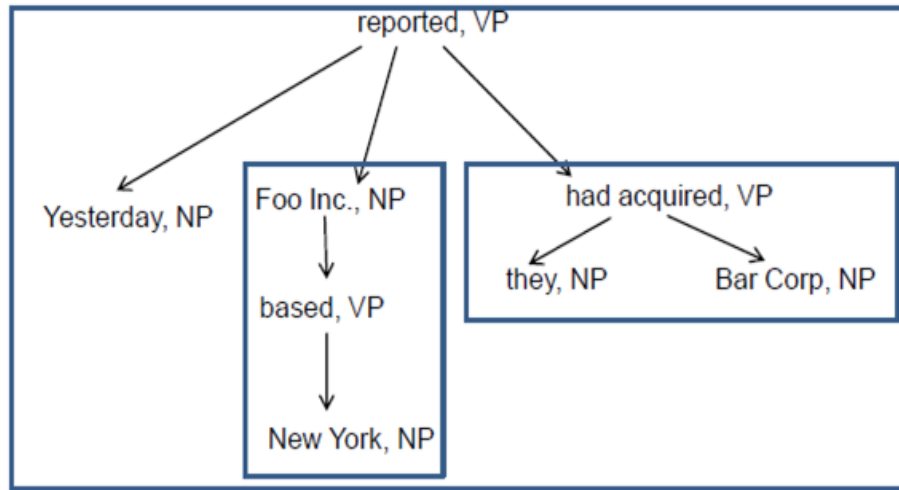


Figure 3.2: A syntactic parsing dependency tree for sentence “*Yesterday, New York based Foo Inc. reported that they had acquired Bar Corp.*”

The NLP pipeline stages described above have direct correspondence to the stages of linguistic analysis. Additionally, their implementation actually depends on the previous task making these chain of the stages a pipeline indeed. In the framework of linguistics what comes next is semantic analysis. In theory, semantics is the study of meaning of a text unit: a word, phrase, entire sentence, and even larger fragments. It is a very general task and in the field of NLP it has been split into numerous tasks each working with different lexical units and different aspects of semantics. The following two stages of the basic NLP pipeline lie in the area of semantic analysis, which also includes other tasks such as recognizing textual entailment [46].

Named entity recognition Often abbreviated as NER, this task seeks to locate and classify elements in text into pre-defined categories such as **Person**, **Organization**, **Location**, **Time**, **Money unit**, etc. In our example a couple named entities can be detected: *Yesterday*, [*New York*]^**Location** based [*Foo Inc.*]^**Organization** reported that they had acquired [*Bar Corp.*]^**Organization**. The accuracy of this task varies depending on the named entity class that we want to determine.

Word sense disambiguation Abbreviated as WSD, it is a task of automatically choosing an adequate sense for a given word in a given context out of a set

of senses. For instance, for a verb *report* a well-known sense inventory WordNet [57] returns six different meanings, e.g., (1) “*announce as the result of an investigation or experience or finding*”, (2) “*announce one’s presence*”, (3) “*complain about; make a charge against*”, etc. For our example in particular, one would need to determine which one of the returned 6 senses corresponds to the use of the verb in the context “*Yesterday, New York based Foo Inc. reported that they had acquired Bar Corp.*” WSD is believed to be an open task and there is still a lot of ongoing research [60].

Each next stage of the pipeline normally relies on the output of either all or at least some previous stages. Moreover, the difficulty of the tasks increases while moving down the pipeline. Yet syntactic parsing shows the highest performance time. Therefore, when performing a higher-level NLP task such as information extraction in our case, it is always an advantage to use fewer basic NLP pipeline stages.

3.2 Generic Rule-based Open IE Algorithm

Here we will outline the core steps of a rule-based algorithm for Open IE:

- First, search for a verb-containing relation phrase in a sentence;
- If detected, search for a noun-containing fragment immediately to the left of the relation phrase;
- If detected, search for another noun-containing fragment immediately to the right of the relation phrase;
- If found, return the components in the form of a triple
 $\langle \text{Argument1} \rangle \langle \text{Relation} \rangle \langle \text{Argument2} \rangle$.

In this general case we do not specify whether a verb or a noun phrase is detected based on POS-tags or dependency-parsing syntactic tags or any others. We also avoid using terms “verb phrase” and “noun phrase” because they are generally used in the domain of syntactic parsing.

From this description it becomes obvious that the main assumption for information detection is that important information is conveyed in a language through Subject-Verb-Object word order. Object-Verb-Subject word order also might be considered. However, in the latter case the interpretation of the semantic roles of

the arguments, i.e., which is the agent and which is the object of the relation, is not trivial.

3.3 Relevant Language-Specific Properties of Spanish

As we have already mentioned in Section 1, 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.

As it is known, most of the methods for natural language processing are developed and tested primarily for English. Next language that receives substantial research work on different language analysis levels is Chinese: [47] at word segmentation level, [58, 68] in grammatical extraction and dependency parsing, [67] in machine translation—just to name a few presented at a single international conference on computational linguistics ACL in 2014. Partially, it can be explained by the fact that language characteristics of Chinese are very different from those of English which is true.

However, the fact that both English and Spanish belong to the same language family, namely, Indo-European language family does not mean that the methods developed for English can be trivially transferred onto Spanish language. Just to start from the fact that they belong to different language groups within the same language family: Romance for Spanish and Germanic for English.

Further we will illustrate some differences in grammar that makes it obvious why the adaptation of NLP methods for Spanish language is not a trivial task. By no means do we intend to give a full comparative analysis of these languages which is a task of Comparative Linguistics.

Sample differences between Spanish and English are:

Infinitives In English infinitives are marked by a special particle *to* that makes 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.

Reflexive pronouns Reflexive pronouns are pronouns that refer back to the subject of the sentence or clause. In Spanish they stick to the infinitive (e.g., *peinarme*, “*to brush myself*”) while get detached and moved in front of the

verb in other forms (e.g., *me peiné*, “(I) brushed myself”, 1st person singular, past tense).

Clitic A clitic is a morpheme that has syntactic characteristics of a word, but depends phonologically on another word or phrase. It is a general case for reflexive pronouns that also includes any oblique case pronouns in Spanish when used with infinitives and imperative voice: *¡Dámelo!* vs. “Give it to me!”, or *dárselo* vs. “to give it to him/her”.

Oblique case pronouns When not accompanied by an infinitive or imperative verb, oblique case pronouns are written separately in Spanish as well as in English. However, the word order is different: in Spanish they precede verbs (*lo veo*) whereas in English pronouns follow verbs (“I see it”).

Adjective order Unlike in English where adjectives strictly precede nouns, in Spanish adjective can both precede and (more often) follow them as in *diversas tumbas egipcias*, “various Egyptian tombs”. This phenomenon complicates detection of noun phrases in Spanish.

Despite a lot of differences between them, the languages have a few features in common:

Word order Here we refer to the order of main sentence components: subject, verb, object. Both languages share predominantly subject-verb-object word order, although Spanish is more likely than English to have indirect object - verb - subject word order.

Analytic languages Both languages are of the same language type, namely, analytic. An analytic language is a language that conveys grammatical relationships without using inflectional morphemes. In particular, it means that both languages use word order or prepositions to convey syntactic relation between verbs and nouns rather than grammatical cases for nouns.

These two similarities are very important for our hypothesis that the Open IE algorithm described in Section 3.2 can be adapted for Spanish without major changes apart from specific to the lexicon.

Chapter 4

Open Information Extraction based on Rules over Part-of-Speech Tags

In this chapter we introduce the novel method for Open IE for Spanish language that outperforms the systems implementing 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 a gain in time.

This chapter covers the following topics:

- POS-tag patterns for detecting extraction components;
- The algorithm for Open IE for Spanish;
- Implementation of the algorithm in EXTRHECH system;
- Evaluation of performance and comparison with other systems;
- Discussion of the results;
- Limitations of the method;
- Illustration of the errors and discussion of their reasons.

4.1 POS-tag Patterns for Information Detection

In our method for Open IE for Spanish we follow the same action sequence for extraction detection as presented in Section 3.2. We will repeat it here:

- First, search for a verb-containing relation phrase in a sentence;
- If detected, search for a noun-containing fragment immediately to the left of the relation phrase;
- If detected, search for another noun-containing fragment immediately to the right of the relation phrase;
- If found, return the components in the form of a triple
 $\langle \text{Argument1} \rangle \langle \text{Relation} \rangle \langle \text{Argument2} \rangle$.

Further we will describe the patterns over POS-tag that we introduce for detection of potentially important information written in Spanish language and its extraction.

4.1.1 Verb relation pattern

In our method, a verb 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*”) or until an infinitive (e.g., *sirven bien para acentuar*, “*serve well to emphasize*”). The corresponding formal expressions for the verb phrase is:

$$\text{VREL} \rightarrow (\text{VW} * \text{P}) | (\text{V}) \quad (4.1)$$

where the expression in the second brackets stands for a single verb, and the expression in the first brackets stands for a verb with dependent words. Formal word V denominates a verb possibly preceded by a reflexive pronoun (e.g., *se caracterizaron*, “*were characterized*”), or a participle (e.g., *relacionadas*, “*related*”). $\text{VW} * \text{P}$ matches a verb with dependent words, where W stands for either a noun, an adjective, an adverb, a pronoun, or an article, and P stands for a preposition optionally immediately followed by an infinitive or a gerund (*sigue siendo*, “*continues to be*”). Special symbol $*$ signifies zero or more matches, $|$ stands for a choice of a variant. The whole match is referred to as **relation phrase**.

4.1.2 Noun argument pattern

Another formal pattern describes noun phrases:

$$\text{NP} \rightarrow \text{N}(\text{PREPN})? \quad (4.2)$$

where **N** matches a noun optionally preceded by either an article (*la dinámica*, “the dynamics”), an adjective, an ordinal number (*los primeros ganadores*, “the first winners”), a number (*3 casas*, “3 houses”), or their combination, optionally followed by either a single adjective (*un esfuerzo criminal*, “a criminal effort”), a single participle, or both (*los documentos escritos antiguos*, “the ancient written documents”). The whole expression matched by **N** can be preceded by an indefinite determinant construction, *uno de* (i.e., “one of”). **PREP** matches a single preposition. Hence, an entire noun phrase is either a single noun with optional modifiers or a noun with optional modifiers followed by a prepositional phrase that is a preposition and another noun with its corresponding optional modifiers (*una larga lista de problemas actuales*, “a long list of current problems”). Special symbol ? signifies 0 or 1 matches.

4.1.3 Patterns for complex syntactic structures

To amplify the coverage of our method, we also introduce a rule for coordinating conjunctions:

$$\text{COORD} \rightarrow \text{Y}|\text{COMMA}? \quad (4.3)$$

where the formal word **Y** stands for a coordinator: *y* (“and”), *o* (“or”), *pero* (“but”)—and formal word **COMMA** stands for a comma. A coordinating conjunction can be either a coordinator or a comma optionally followed by a coordinator as in ‘, *y*’.

We also introduce a pattern for relative pronouns:

$$\text{QUE} \rightarrow \text{PR} \quad (4.4)$$

Formal word **PR** stands for relative pronouns, e.g., *que* (“that”), *cual* (“which”), and serves for resolution of relative clauses.

4.2 Our Algorithm

In the previous section we introduced the few rules that define the gist of our method. Here we will outline our algorithm for Open IE for Spanish language [69]. It takes as **input** POS-tagged text, therefore, only POS-tagging is necessary for pre-processing. Similar to the algorithms described in Section 2.2, it executes sentence by sentence processing.

Algorithm 3 Our algorithm for Open IE based on rules over POS-tags

```

Identify a potential verb relation by matching against pattern (4.1)
if detected then
    Search to the left of the verb phrase for a potential first argument by matching
    against the noun phrase pattern (4.2)
    if detected then
        Search to the right of the verbal phrase for a second argument matching
        against pattern (4.2)
        if detected then return extraction triple
        else return false
    else return false
else return false

```

This is the core of the Open IE algorithm. In the actual implementation we also added additional rules for syntactically more complex cases.

Participle clauses If a noun within a detected noun phrase is followed by a participle clause terminating with another noun, the participle phrase is resolved into an independent relational tuple. Consider the following sentence:

Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.
 (“*The Egyptians were characterized by their believes related with (the) death*”)

As an output, the algorithm returns two relational tuples:

$$\langle \text{Arg1} = \textit{Los egipcios} \rangle \langle \text{Rel} = \textit{se caracterizaron por} \rangle \langle \text{Arg2} = \textit{sus creencias} \rangle$$

$$\langle \text{Arg1} = \textit{sus creencias} \rangle \langle \text{Rel} = \textit{relacionadas con} \rangle \langle \text{Arg2} = \textit{la muerte} \rangle$$

The first of these extractions corresponds to the main verb of the sentence, while the other one corresponds to the participle clause.

Coordinating conjunctions We also added processing of coordinating conjunctions within noun phrase arguments with a rule implemented according to pattern (4.3). 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.*”)

The correct resolution of the coordinating conjunctions results in three extractions:

$\langle \text{Arg1} = \text{La civilización China} \rangle \langle \text{Rel} = \text{nos heredó} \rangle \langle \text{Arg2} = \text{el papel} \rangle$
 $\langle \text{Arg1} = \text{La civilización China} \rangle \langle \text{Rel} = \text{nos heredó} \rangle \langle \text{Arg2} = \text{la pólvora} \rangle$
 $\langle \text{Arg1} = \text{La civilización China} \rangle \langle \text{Rel} = \text{nos heredó} \rangle \langle \text{Arg2} = \text{la brújula} \rangle$

Relative clauses Information is also extracted from relative clauses that are detected by a rule corresponding to pattern (4.4). A relative pronoun is filtered out, and the left argument of a relational tuple is searched to the left of the relative pronoun. Thus, in the sentence:

Los primeros griegos se organizaron en grupos que tenían lazos familiares.
 (“*The first Greeks were organized in groups that had family relations.*”)

two facts are detected:

$\langle \text{Arg1} = \text{Los primeros griegos} \rangle \langle \text{Rel} = \text{se organizaron en} \rangle \langle \text{Arg2} = \text{grupos} \rangle$
 $\langle \text{Arg1} = \text{grupos} \rangle \langle \text{Rel} = \text{tenían} \rangle \langle \text{Arg2} = \text{lazos familiares} \rangle$

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

These additional rules help our method to return more extraction, by this increasing its recall. In the same time, they make the extracted components more semantically and syntactically simple. This is an important property, because as we mentioned in Section 2.5, one of the applications of Open IE is knowledge base population that requires normalized concept as its input.

4.3 ExtrHech System

We have implemented our algorithm for Open IE for Spanish language in EXTRHECH system. The processing pipeline of the system is shown in Figure 4.1.

The system takes as input a POS-tagged text. For POS-tagging we use Freeling-2.2 [44], which uses the EAGLES POS tag set for Spanish [33]. An example of

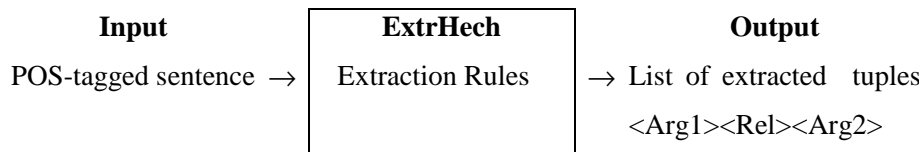


Figure 4.1: Processing pipeline of EXTRHECH system

POS-tagged analysis of sentence “*La numeración arábica procede de India.*” (“*The Arabic numbering comes from India*”) follows in Table 4.1, where the first row shows the words as they are in the sentence, the second row presents the lemmatized forms of the corresponding words, and the third row contains the POS tags according to EAGLES POS tag set.

Table 4.1: Example of a POS-tagged sentence by Freeling-2.2.

Word	La	numeración	arábica	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 ISO encoding. Nevertheless, an arbitrary text can be encoded in a different encoding. In the case of the texts extracted from the Internet, the encoding is mostly UTF-8 (as is the case with the texts in the Raw Web dataset described in Section 4.4). Therefore, we have developed an additional pre-processing module that converts a UTF-8-encoded text into an ISO encoded text which then can be passed to Freeling-2.2.

EXTRHECH performs sentence-by-sentence processing. The rules are implemented in the form of regular expressions over sequences of POS tags, which are provided in appendix B. Thereby, rule matching is executed as regular expression matching.

The source code of the system is available from a BitBucket.org repository at https://bitbucket.org/alisa_ipn/extrhech.git.

4.4 Experiments and Results

In this section we describe the experiments conducted to evaluate the performance of our method implemented in EXTRHECH system. We conducted a series of experiments on different datasets¹.

¹All materials are available on the following webpage: <http://www.gelbukh.com/resources/spanish-open-fact-extraction/>

4.4.1 Experiments on Spanish language dataset

We compared performance of our method against other systems for Spanish language: FES-2012 based on complex heuristics over full syntactic parsing described in [1] and DEPOE based on fewer rules over dependency parsing [25]. Unfortunately, FES-2012 is unavailable for downloading. Hence, we could not actually replay it on an arbitrary dataset. Instead, we will refer to the results presented in [1] obtained from his experiments on FactSpCIC dataset [1]. FactSpCIC contains 68 grammatically and orthographically correct and consistent sentences manually selected from school textbooks.

DEPOE system is available for download from here². Therefore, we ran EX-TRHECH and DEPOE on the sentences from FactSpCIC datasets.

Further, two human judges independently evaluated each extraction as correct or incorrect. In the instructions for annotators we followed [23] and we instructed our annotators to:

1. Identify whether the information conveyed by an extraction actually is contained in the sentence. Given the sentence

“Roma no imponía ideas políticas o credos en sus territorios.”
(“Rome did not imposed political ideas or creeds in its territories.”),

the extraction $\langle Roma \rangle \langle imponía \rangle \langle ideas políticas \rangle$, which states exactly the opposite, was considered incorrect.

2. Detect whether the order of the arguments adequately corresponds to the assumption that **Arg1** corresponds to **agent/experiencer** and **Arg2** corresponds to **object/indirect object** of the relation. Let’s consider the sentence

“El sistema nervioso periférico lo conforman los nervios”
(“Nerves comprise the peripheral nervous system.”).

Then, the extraction $\langle sistema nervioso periférico \rangle \langle conforman \rangle \langle los nervios \rangle$ was instructed to be labeled as incorrect.

3. Consider incoherent and uninformative extractions as incorrect. For example, the sentence is

²<http://gramatica.usc.es/~gamallo/prototypes/DepOE-beta.tar.gz>

Table 4.2: Comparison of performance of rule-based Open IE systems for Spanish

System	Precision	Recall	Correct Extr's	Returned Extr's	Returned Extr's	Rules based on
EXTRHECH	0.87	0.73	99	115	137	POS-tags
DEPOE	0.80	0.29	39	49	137 ³	Syntactic tags
FES-2012	0.66	0.72	120	182	166	Syntactic tags

“El cerebro es capaz de llevar a cabo varias acciones al mismo tiempo”
(“The brain is able to perform various actions at the same time.”).

Then, the extraction $\langle El \text{ cerebro} \rangle \langle es \rangle \langle capaz \rangle$ is uninformative and, thereby, incorrect.

For FactSpCIC dataset, the annotators agreed on 89% of extractions with Cohen’s kappa $\kappa = 0.52$, which is considered to be moderate agreement [32]. The number of correct extraction was calculated as an average for the two judges.

Precision and recall of the systems were calculated following the formulas provided in Section 2.4. As the reader remembers, for its calculation recall needs the number of all possible correct extractions in the divisor. To estimate the latter, we made a list of all extractions that the method is expected to return. Then, this set was extended by the extractions returned by the systems 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. Aguilar-Galicia’s thesis [1] contained all the necessary data, i.e., the dataset with expected extractions and the returned extractions, so that we could evaluate precision and recall using the same instructions for the annotators as for the two other systems. The comparison is presented in Table 4.2.

4.4.2 Robustness evaluation

As we state throughout the work, our goal was not only to introduce a high performance method for Open IE, but also a method that would be highly robust. However, this aspect of Open IE performance has not been evaluated previously. Consequently, no measure for its evaluation has been suggested.

Here, we would like to introduce two measures of robustness.

We call the first one **pre-processing robustness** and calculate it as fraction of the number of attempted sentences, i.e. sentences that were pre-processed correctly enough to be passed to the extraction stage, in the total number of input sentences

Table 4.3: Comparison of pre-processing robustness for rule-based Open IE systems for Spanish

System	Robust _{pre-proc}	Attempted sentences
EXTRHECH	0.93	63
FES-2012	0.87	59
DEPOE	0.72	49

Table 4.4: Comparison of extraction robustness for rule-based Open IE systems for Spanish

System	Robustness	Sentences w/ correct extraction
EXTRHECH	0.84	57
FES-2012	0.79	54
DEPOE	0.57	39

as in equation 4.5:

$$\text{Robustness}_{pre-proc} = \frac{\text{attempted sentences}}{\text{all input sentences}} \quad (4.5)$$

The results of evaluation of pre-processing robustness on FactSpCIC dataset are shown in Table 4.3. We remind that the total number of sentences in this dataset is 68.

The other one is properly **extraction robustness**. It is calculated as the number of sentences with at least one correct extraction divided by the total number of input sentences as in equation 4.6:

$$\text{Robustness} = \frac{\text{sentences with correct extraction}}{\text{all input sentences}} \quad (4.6)$$

The results of evaluation of this measure on the same FactSpCIC dataset are shown in Table 4.4

4.4.3 Discussion of experiments on Spanish language dataset

As we can observe from Table 4.2, our Open IE method for Spanish implemented in EXTRHECH system shows higher performance in terms of precision and recall compared to the other systems.

The main shortcoming of DEPOE system is that it is not adapted for Spanish

language. As Algorithm 2 shows, it simply takes the same sequence of actions and applies it to several languages, one of which is Spanish, replacing the syntactic parser for a corresponding language. We calculated recall of the system using the same number of expected extraction as it was estimated for EXTRHECH system. Even if we could estimate a better expectation for DEPOE system, the total number of returned extractions is very low: as few as 49 extractions. It happens because their algorithm does not take into account the requirement of semantic granularity of extraction components and most of their extraction simply repeat the whole sentence:

“Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.” →
⟨Los egipcios⟩⟨se caracterizaron por⟩⟨sus creencias relacionadas con la muerte⟩.

Such wordy extractions will require very substantial post-processing to be used in any further application.

The rules implemented in FES-2012, actually, were able to process complex syntactic structures and to return extractions with adequately concise components. Possibly, that is why the number of expected extractions is high, 166. However, the system have shortcomings in resolution of verbs with reflexive pronouns:

“Los habitantes de la antigua Roma se ocupaban en diversos trabajos.” →
*(“The inhabitants of ancient Rome had various occupations.”) ⟨habitantes de
 antigua Roma⟩⟨ocupaban⟩⟨en diversos trabajos⟩ (~ ⟨The inhabitants of ancient
 Rome⟩⟨used⟩⟨in various occupations⟩).*

The extraction does not convey the information contained in the sentence.

Tables 4.3 and 4.4 also show that our method demonstrates higher robustness than the other two methods. As one can expect, it is due to the fact that EXTRHECH’s method only requires POS-tagging at the pre-processing stage which is based on a more robust algorithm than syntactic parsing which is used by the two other methods.

4.4.4 Experiment on parallel Spanish and English datasets

We have also compared EXTRHECH’s performance to the performance of REVERB, which implements similar rule-based methods for Open IE for English language. Since these systems are designed for different languages, we ran our experiment on a parallel dataset.

Table 4.5: Performance comparison of REVERB and EXTRHECH systems over a parallel dataset.

System	Precision	Recall	Correct Extractions	Returned Extractions	Rules based on
EXTRHECH	0.59	0.48	218	368	POS-tags
REVERB	0.56	0.44	201	358	POS-tags & syntactic chunks

We took 300 parallel sentences from the English-Spanish part of News Commentary Corpus [10]. Then, we ran the extractors over the corresponding languages. After that, two human annotators labeled each extraction as correct or incorrect. For the Spanish part of the dataset, the annotators agreed on 80% of extractions (Cohen’s kappa $\kappa = 0.60$), whereas for the English part they agreed on 85% of extractions with $\kappa = 0.68$. For both datasets their respective κ coefficients indicate substantial agreement between the annotators.

By manual revision of the sentences in the datasets, we made a list of all expected correct extractions. Their number was used to estimate the recall.

We also would like to note that on the contrast to REVERB, our system does not have a confidence score mechanism. To make the comparison between the systems appropriate, we ran REVERB extractor with the confidence score level set to 0 that means that the system returns all extractions that match the rules. Hence, the systems were in equivalent conditions. The results of the experiment are shown in Table 4.5.

As we see, on a parallel dataset of texts from News Commentary Corpus, both systems show a very similar performance. Based on this observation, we can conclude that the algorithm suggested in [24] can be easily adopted for other languages with dominating SVO word order and an available POS-tagger.

4.4.5 Experiment on Raw Web dataset

One of the most important goals of Open IE systems is to be able to process large amounts of texts directly from the Web. Texts on the Web often lack grammatical and orthographical correctness or coherence. In this work we also evaluated the performance of our system on a dataset of sentences extracted immediately from the Internet “as is”. For this dataset, we took 200 random data chunks detected by a sentence splitter from CommonCrawl 2012 corpus [30], which is a collection of web texts crawled from over 5 billion web pages. However, 41 from those 200 chunks

Table 4.6: Performance of EXTRHECH on the grammatically correct dataset and the dataset of noisy sentences extracted from the Web

Dataset	Precision	Recall
News Commentary	0.59	0.48
Raw Web	0.55	0.49

were not samples of textual information in human language but rather pieces of programming codes or numbers. We took out these chunks because they are trivial for our research. In a real life scenario they could be easily detected and eliminated from the Web data stream. After this, our dataset consisted of 159 sentences written in human language. We will refer to this dataset as Raw Web text dataset.¹ Of 159 sentences of the dataset, 36 sentences (22% of the dataset) were grammatically incorrect or incoherent, as evaluated by a professional linguist.

We ran EXTRHECH system over this dataset and asked two human judges to label extractions as correct or incorrect. The annotators agreed on 70% of extractions with Cohen’s $\kappa = 0.40$, which indicates the lower bound of moderate agreement between judges.

Precision and recall were calculated in the same manner as in the experiments described above. We compare these numbers to the results obtained for the dataset of grammatically correct sentences from News Commentary Corpus in Table 4.6.

We can observe that our system’s performance has not lowered significantly when processing “noisy” texts compared to edited newspaper texts. An interesting observation is that texts from the Internet are poorer in facts than the news texts. The number of expected extractions was manually evaluated by a human expert for both datasets. The ratio extractions:to sentences for the news dataset was 1.5:1, while for the Raw Web dataset it was only 1.03:1.

Now we will briefly discuss the issue arising due to various encoding standards used for such Spanish language characters as *á*, *é*, *ñ*, etc. While applying Freeling morphological analyzer to the dataset, we encountered an issue that the sentences came in various encodings. As we mentioned in Section 4.3, Freeling-2.2 analyzer works properly only with ISO encoded input. Therefore, we had to convert each sentence from the dataset into ISO encoding. While most of the sentences were in UTF-8 encoding and were converted in a single pass, the encoding of about 3% of the sentences was initially corrupted, therefore, they were not processed correctly by the POS-tagger. Although the issue is manageable at the scale of a small dataset, it might affect the speed and quality of fact extraction when working at web scale.

Table 4.7: Comparative data for various Open IE systems.

System	Approach	Dataset (# of sent.)	Precision	Recall	Running Time
EXTRHECH (Spanish)	rules over	FactSpCIC (68)	0.87	0.73	seconds
	POS-tags	Raw Web texts (159)	0.55	0.49	seconds
REVERB (English)	rules over POS-tags synt. chunks	FactSpCIC (68), translated	0.76	0.50	seconds
TEXTRUNNER (English)	self-learning on POS-tags	Yahoo (500)	0.30	0.42	seconds
WOE ^{PARSE} (English)	self-learning on full synt. parsing	Yahoo (500)	0.7	0.6	seconds
OLLIE (English)	context analysis analysis on full synt. parsing	news, Wikipedia, biology textbooks (300)	0.66	N/A	hours
FES-2012 (Spanish)	rules over full synt. parsing	FactSpCIC (68)	0.66	0.72	hours
DEPOE (Spanish)	rules over full synt. parsing	FactSpCIC (68)	0.80	0.29	seconds
CLAUSEI (English)	rules over synt. clauses	Yahoo (500), Wikipedia(200), NYT (200)	0.70	N/A	N/A

4.4.6 Comparative table for various Open IE methods

Table 4.7 provides a comparative analysis of performance of various Open IE systems implementing various approaches to Open IE described in Section 2.1. Not all of the systems were available for download that is why we provide some numbers according to the corresponding papers where the results were published indicating the name and size of a used dataset. The Precision/Recall data are taken at the confidence score level of 0, that means the highest recall or the maximum yield. Therefore, we took the precision numbers at the highest recall/yield level, whichever was provided.

The comparison of performance for the systems designed for different languages on different datasets is indirect, because there are a variety of reasons for the differences in the results. However, EXTRHECH’s speed is at the same level as that of other POS tag-based systems. It is also much faster than syntactic parsing-based systems, which perform significantly slower, although with better precision. Thus,

performance of EXTRHECH is of the same order or higher as that of similar state-of-the-art systems.

4.5 Limitations

As we have mentioned in Section 2.2, the compared methods for Open IE as well as ours all share the same limitation that they only detect relations expressed via verbs.

In our approach to Open IE in Spanish, we do not allow pronouns to be potential arguments of a relation. It was mainly done because of a wide use of a neutral pronoun *lo* (“*this*”, “*which*” or no direct translation) as a head of relative clauses in Spanish language, e.g., *lo que dio valor al poder judicial* (“*___ that gave value to the judiciary*”). Including pronouns for potential argument matches would return a lot of uninformative relations as $\langle lo \rangle \langle dio\ valor\ a \rangle \langle el\ poder\ judicial \rangle$. This issue can be solved only by introducing anaphora resolution techniques which involves processing on a super-sentence level. Although seemingly feasible, this modification will necessarily slow down the extraction speed which is critical while working with large scale corpora. As mentioned in Section 2.1, high velocity performance is one of the main advantages of the approach to Open IE based on syntactic constraints compared to the others. Hence, any modifications that would affect its speed should be considered with caution.

Another language dependent limitation is related to the order of the processing. As earlier described in Section 4.4, an extracted triple is expected to correspond semantically to $\langle agent/experience \rangle \langle relation \rangle \langle general\ object/circumstance \rangle$. This is expected to be correct for a direct word order, i.e., Subject – Verb – (Indirect) Object, which is a dominant word order for Spanish. Yet the inverted word order, i.e. (Indirect) Object – Verb – Subject (e.g., *De la médula espinal nacen los nervios periféricos*, i.e., literally **“From the spinal cord arise peripheral nerves”*), also occasionally takes place in grammatically correct and stylistically neutral Spanish texts. However, the occurrence of this construction is less than 10% according to [13].

Additionally, our method does not resolve anaphora and zero subject construction, which occur in Spanish. Their resolution require substantial additional processing and deeper linguistic analysis (see Section 3.1), which contradicts our hypothesis of sufficiency of only POS-tagging only.

4.6 Errors in Open Information Extractions

In this section we provide a detailed analysis of errors found in returned extractions. First, we will suggest classification of types of the errors and then analyze possible causes for the errors. Finally, we will discuss possible directions for improvement and their cost of implementation and how they will affect the speed and performance of the current Open IE method.

Apart from a very brief analysis of incorrect extractions in [24], where errors are not distinguished from their causes, no substantial study of such errors and their reasons has been reported. In this work, we distinguish between errors and their causes and provide corresponding classifications for both based on the analysis of errors found in FactSpCIC and Raw Web datasets.

4.6.1 Main types of errors

To build our classification of error types, we started from the following item included in error analysis in [24]:

- correct relation phrase, incorrect arguments.

This is the only item that describes an actual type of error as opposed to error causing issues such as “N-ary relation” or “non-contiguous verb phrase”, which do not describe errors by themselves but rather language phenomena that could have caused errors.

In continuation, we introduce the classification based on the components of extracted tuples where an error occurs. We have added other classes to make the classification complete. This means that each possible error falls into at least one of the suggested error types, which are the following:

Incorrect relation phrase In the errors of this type, a detected relation phrase is incorrect. For an example, consider the fragment:

*...la brujita, la del circo o la holandesa que comentaba Montse por los
foros...*
(“...the witch, the one from the circus or the dutchess that Montse
commented on the forums...”)

The fact that is detected in this sentence by EXTRHECH system looks as follows:

$\langle \text{Arg1} = \textit{la holandesa} \rangle \langle \text{Rel} = \textit{comentaba Montse por} \rangle \langle \text{Arg2} = \textit{por los foros} \rangle$
(incorrect).

Obviously, the relation phrase in this extraction is incorrect because it includes the syntactic subject “*Montse*”. Since the relation phrase is the first component of a relational tuple that is searched, incorrect detection of the relation phrase in most of the cases leads to incorrect detection of the arguments as in the example above.

Incorrect argument(s) In this case, at least one of the arguments of the extracted tuple is detected incorrectly. For example, for a sentence:

Opositor a la guerra de Irak liberado de arresto militar
 (“*Opponent of the war of Iraq liberated from military arrest*”)

the fact detected by our system reads:

$\langle \text{Arg1} = \textit{Irak} \rangle \langle \text{Rel} = \textit{liberado de} \rangle \langle \text{Arg2} = \textit{arresto militar} \rangle$ (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, i.e., shorter than it should be.

Correct relation phrase but incorrect arguments We consider this type of errors for better understanding of the cases when errors in arguments are not provoked by issues causing incorrect relation detection at the same time. Consider an example sentence:

La soldada tapa resguarda un rico cóctel cardiosaludable
 (“*The soldered tap preserves a rich heart-healthy cocktail*”)

and the corresponding extraction:

$\langle \text{Arg1} = \textit{tapa} \rangle \langle \text{Rel} = \textit{resguarda} \rangle \langle \text{Arg2} = \textit{un rico cóctel cardiosaludable} \rangle$
(incorrect).

The left argument of the relation is detected incorrectly. The correct argument should be *la soldada tapa*. Yet, in this particular example the rule was unable to detect the complete argument because of the incorrect POS tag of the word *soldada* that was tagged as a noun (“*female soldier*”) instead of being tagged as an adjective (“*soldered*”).

Table 4.8: Distribution of error types by the number of returned extractions for different datasets.

Dataset	Error type			
	Incorrect argument(s)	Correct relation, incorrect argument(s)	Incorrect relation phrase	Incorrect argument order
FactSpCIC	22%	16%	9%	3%
Raw Web	45%	26%	21%	6%

Incorrect order of arguments For example, from the sentence:

Vaya susto que se llevó tu hija!
 (“What [a] fright that got herself your daughter!”)

the system detected a relation tuple:

$\langle \text{Arg1} = \text{susto} \rangle \langle \text{Rel} = \text{se llevó} \rangle \langle \text{Arg2} = \text{tu hija} \rangle$ (incorrect).

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

$\langle \text{Arg1} = \text{tu hija} \rangle \langle \text{Rel} = \text{se llevó} \rangle \langle \text{Arg2} = \text{susto} \rangle$ (correct).

Although complete, this classification is overlapping because the errors included into **Correct relation phrase, but incorrect arguments** category must also be classified as **Incorrect argument(s)**. However, classifying errors into these classes helps better error detection and more precise distinction between issues causing the errors.

The distribution of error types by the number of extracted facts for each dataset is presented in Table 4.8.

As one can see, for both the dataset of grammatically correct sentences FactSpCIC and the Raw Web text datasets, the distribution of the error types is similar. The majority of the errors is due to the errors in argument detection. More than a half of them do not co-occur with errors in relation phrases. However, for the grammatically correct dataset they occur more often than errors in relation detection, whereas the situation is vice versa for Raw Web dataset. It might be explained by the fact that grammatical errors in the original sentences impede their correct processing and the extraction process fails at the stage of relation detection, which

is the first one in our algorithm. Errors in argument order are the least common for both datasets because, [13] indicates, the inverse word order is not a common construction in Spanish.

4.6.2 Main issues that cause errors and possible solutions

In this section we analyze the issues that provoke the errors in extraction. A few of the issues, namely,

- N-ary relations
- Non-contiguous relation phrase
- Overspecified relation phrase
- Incorrect POS-tagging

were mentioned in [24]. However, after a thorough analysis of each error in extractions returned by EXTRHECH from our two datasets, we could identify several other major issues, as well as describe the previously mentioned issues with more detail. We do not claim this list to be exhaustive, although it covers most of the issues that cause errors in news articles, textbook, and some Web forum comments. Possibly, other issues might be detected in a larger Web-based dataset. Yet one can assume that those issues are not very common and could be included into the group Others of the current classification. Below we provide a list of the identified issues with examples and suggest possible solutions.

1 Underspecified noun phrase For example, 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.*”)

we would expect an extraction of the tuple:

$\langle \text{Arg1} = \textit{la agrupación de seres humanos en un mismo espacio} \rangle \langle \text{Rel} = \textit{favoreció el intercambio de} \rangle \langle \text{Arg2} = \textit{conocimientos} \rangle$.

Yet the extraction returned by the system is:

$$\langle \text{Arg1} = \text{un mismo espacio} \rangle \langle \text{Rel} = \text{favoreció el intercambio de} \rangle \langle \text{Arg2} = \text{conocimientos} \rangle,$$

where the first argument *un mismo espacio* is underspecified, i.e., is just a fragment of the complete component *la agrupación de seres humanos en un mismo espacio*.

To overcome the underspecification of noun phrases, simple POS-based rules are not always sufficient. Syntactic chunking could provide better detection of complete constituents. Yet introduction of an additional pre-processing procedure would inevitably increase the running time of the extraction.

2 Overspecified verb phrase

For example, consider the sentence:

La Botánica ha logrado analizar las características de la vegetación.
 (“The Botany has achieved analyzing the characteristics of the vegetation.”)

The returned extraction is:

$$\langle \text{Arg1} = \text{La Botánica} \rangle \langle \text{Rel} = \text{ha logrado analizar las características de} \rangle \langle \text{Arg2} = \text{la vegetación} \rangle.$$

Although the relation phrase *ha logrado analizar las características de* is extracted in accordance with the longest match for a verbal phrase rule, in fact, the relation phrase should be shorter: *ha logrado analizar*, and the correct extracted tuple should be:

$$\langle \text{Arg1} = \text{La Botánica} \rangle \langle \text{Rel} = \text{ha logrado analizar} \rangle \langle \text{Arg2} = \text{las características de la vegetación} \rangle.$$

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

The solution suggested in [24] is, first, to perform a massive relation extraction on a large corpus, and then, to consider only the relations with frequencies above a certain threshold as valid relations. This is done by creating a dictionary of valid relations and comparing a newly extracted relation against the dictionary (“lexical constrain”). In the mentioned work, the size of the dictionary was about 2 million relation phrases. However, about 23% of missed

extractions were filtered out by these constraints. Generally, this solution affects the ability of a system to extract arbitrary relations which is important for massive Web-scale text processing.

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

bajo cuyo nombre pueden entrar los sextantes
 (“under whose name can appear the sextant”) (literally)

the relation phrase should be *pueden entrar bajo el nombre de*, which is non-contiguous in the given fragment.

This problem is quiet difficult to solve because even syntactic parsing does not show high accuracy for non-contiguous constituents. To the best of our knowledge, no Open IE method has resolved this issue to the moment.

4 N-ary preposition Some prepositions require more than one object, such as *entre* or “between”:

La agricultura inició entre el 8000 y el 5000 a.C.
 (“The agriculture began between (the) 8000 and (the) 5000 B.C.”)

In this case the expected extraction should be

$\langle \text{Arg1} = \textit{la agricultura} \rangle \langle \text{Rel} = \textit{inició entre} \rangle \langle \text{Arg2} = \textit{el 8000 y el 5000 a.C.} \rangle$,

where the coordinating conjunction in the fragment *el 8000 y el 5000 a.C.* is governed by the preposition *entre*, which converts the relation at hand into an N-ary relation between the subject *la agricultura* and two time points: *el 8000 a.C.* and *el 5000 a.C.*

This type of preposition is known to be difficult to handle in any type of linguistic analysis. According to syntactic analysis, *between*-constituent is a compound one and should be considered as a tree with a root *between X* and a leaf *and Y*. The issue becomes even more complex if we think of how this relation should be presented in an ontology. To the best of our knowledge, it is not present in the known public ontologies. Therefore, dealing with N-ary prepositions is an open issue.

5 N-ary relation Although similar to the previous item, this issue has different language nature. Non-binary or N-ary relations 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”)

the relations between the components are as shown in Figure 4.2:

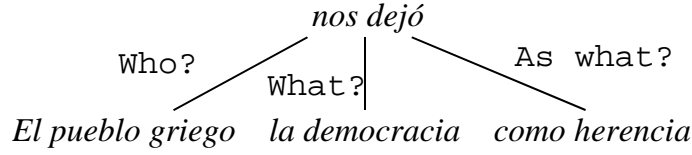


Figure 4.2: Structure of a verb-based N-ary relation

Therefore, the extraction:

$\langle \text{Arg1} = \textit{el pueblo griego} \rangle \langle \text{Rel} = \textit{nos dejó} \rangle \langle \text{Arg2} = \textit{herencia} \rangle$ (incorrect)

cannot be considered correct.

Current approaches to Open IE consider only binary relations. Therefore, handling of this issue requires a different approach to Open IE.

6 Conditional subordinate clause or an adverb that affect semantics of the input sentence Let’s 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 when found the leftovers abandoned by other animals”).

This sentence tells us that the first hominid ate meat only under some specific conditions, such as when they encountered leftovers. Therefore, the extraction:

$\langle \text{Arg1} = \textit{los primeros homínidos} \rangle \langle \text{Rel} = \textit{sólo comían} \rangle \langle \text{Arg2} = \textit{carne} \rangle$,

which means “The first hominids only ate meat”, drastically changes the meaning conveyed by the sentence, although it seemingly makes sense.

This problem can be resolved by introducing an auxiliary dictionary of conditional phrases and adverbs. However, additional research is needed to know how frequent such phrases are and how reliable and resource-consuming this solution can be.

7 Incorrectly resolved relative clause This concerns the resolution of prepositional relative clauses as in the fragment:

El lugar en el que florecieron las culturas más desarrolladas del México antiguo
 (“The place in (the) which flourished the cultures most developed of ancient Mexico”)

and corresponding extractions

$\langle \text{Arg1} = \text{el lugar} \rangle \langle \text{Rel} = \text{florecieron} \rangle \langle \text{Arg2} = \text{las culturas} \rangle$.

The correct fact should look like:

$\langle \text{Arg1} = \text{las culturas más desarrolladas del México antiguo} \rangle \langle \text{Rel} = \text{florecieron en} \rangle \langle \text{Arg2} = \text{el lugar} \rangle$.

These errors are quite difficult to solve because even methods based on syntactic parsing do not resolve them successfully.

8 Incorrectly resolved coordinating conjunction As described in Section 4.2, our method includes rules for coordinating conjunctions analysis. Fragments with coordinating conjunctions are mostly analyzed correctly when the conjunction occurs between either relation phrases or arguments of a relation. However, when a conjunction occurs inside of an argument, it may be resolved incorrectly. Consider the following fragment:

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 facts conveyed by this phrase is:

$\langle \text{Arg1} = \text{los cambios climáticos} \rangle \langle \text{Rel} = \text{crearon un ambiente propicio para} \rangle \langle \text{Arg2} = \text{la reproducción de plantas} \rangle$
 and
 $\langle \text{Arg1} = \text{los cambios climáticos} \rangle \langle \text{Rel} = \text{crearon un ambiente propicio para} \rangle \langle \text{Arg2} = \text{la selección de plantas} \rangle$.

Yet the system erroneously extracts the following information:

$\langle \text{Arg1} = \text{los cambios climáticos} \rangle \langle \text{Rel} = \text{crearon un ambiente propicio para} \rangle \langle \text{Arg2} = \text{Arg2} = \text{la reproducción} \rangle$
 and
 $\langle \text{Arg1} = \text{los cambios climáticos} \rangle \langle \text{Rel} = \text{crearon un ambiente propicio para} \rangle \langle \text{Arg2} = \text{la selección de plantas} \rangle$,

where the first extraction lacks the dependent part of an argument, and thus is underspecified.

9 Inverse word order This phenomenon occurs when a dominating direct word order, which is Subject-Verb-(Indirect)Object in most of European languages, is inverted, i.e., (Indirect)Object-Verb-Subject order occurs:

De la médula espinal nacen los nervios periféricos.
 (“From the spinal cord originate the peripheral nerves”).

Currently our system is designed to resolve only the direct word order, which leads to incorrect extractions as in:

$\langle \text{Arg1} = \text{la médula espinal} \rangle \langle \text{Rel} = \text{nacen} \rangle \langle \text{Arg2} = \text{los nervios periféricos} \rangle$

corresponding to *incorrect argument order* error type.

10 Incorrect POS-tagging This issue arises when a word is incorrectly POS-tagged at the input pre-processing stage. For example, in a sentence:

La soldada tapa resguarda un rico cóctel cardiosaludable
 (“The soldered tap preserves a rich heart-healthy cocktail”)

the word *soldada*, which is an adjective meaning “soldered” in this sentence, was tagged as a noun meaning “female soldier”. Hence, the left argument could not match the noun argument pattern 4.2 described in Section 4.1 and, consequently, the extraction suffered the *underspecified argument* error:

$\langle \text{Arg1} = \textit{tapa} \rangle \langle \text{Rel} = \textit{resguarda} \rangle \langle \text{Arg2} = \textit{un rico cóctel cardiosaludable} \rangle$.

Extraction errors caused by this issue cannot be tackled at the level of information extraction algorithm because they are caused by issues at the pre-processing stage when language analysis tools are used. However, as the robustness analysis showed in Subsection 4.4.2, POS-tagging pre-processing is much less prone to errors than other types of deeper language pre-processing such as syntactic parsing. Therefore, we consider it a minor issue.

The issues listed above provoked errors detected in both grammatically correct FactSpCIC and Raw Web datasets. Below we describe some issues that did not occur in the grammatically correct dataset FactSpCIC, yet they were detected in the larger Raw Web dataset.

11 Grammatical errors in input text Grammatical errors mainly in syntax and punctuation lead to incorrect extraction by our method. For example a sentence:

*En aquellos dias como en casa hay jardin con muchos arbolitos nos encanta
andar trepado prrrrr es una delicia*
(“*In those days as at home there is a garden with many trees we love
climbing trees prrrrr it’s a delight*”)

lacks various punctuation marks, which hinders its understanding even by human readers. In this case the system was not able to detect correct information. Since grammatical errors essentially pertain to human written texts, so far there is no obvious solution for this issue.

12 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 their own classes because of their low counts. For example, in a sentence:

Louis Botha llevó a cabo numerosas manifestaciones públicas.
(“*Louis Botha organized (lit. brought to accomplishing) numerous public
manifestations*”)

the returned extraction looks as follows:

$\langle \text{Arg1} = \textit{Louis_Botha} \rangle \langle \text{Rel} = \textit{llevó a} \rangle \langle \text{Arg2} = \textit{cabo numerosas} \rangle$ (incorrect).

. Here the idiom *llevar a cabo* has not been detected correctly. However, in the final implementation of our system we adjusted parameters of the input pre-processing tool to recognize idioms (we remind the reader that we used the POS module from Freeling-2.1 NLP package). Eventually, this example in particular was successfully resolved. However, in this case we rely on the processing of idioms realized at the pre-processing stage. Generally, idioms and fixed collocations can be resolved by introduction of corresponding lists or dictionaries.

To finalize, low counts of errors caused by these two last issues do not justify endeavors targeted specifically at their solution.

Chapter 5

Named-Entity-Driven Open Information Extraction

In the previous chapter we have discussed a method for Open IE that requires minimal linguistic pre-processing of input. However, semantic interpretation of extractions is an open question. In this chapter we will introduce modifications to the Open IE method as well as some post-processing rules that will lead us to shallow semantic interpretation of the extracted relations.

Further we will cover the following topics:

- Why semantic interpretation of extracted relations should be the next step in Open IE;
- RDF/XML format as a means for semantic interpretation;
- The modified Open IE method and its performance evaluation;
- Post-processing of extractions.

5.1 Motivation

Methods for open information extraction (Open IE) have proved to be efficient for extraction of information from large amounts of unstructured text such as the Web or other large text corpora. Based on language specific syntactic patterns that operate on sentence level, these methods detect potentially important information and extract it in the form of tuples that represent a relation and its arguments. For example, from a sentence “*Woman who drove van full of kids is charged with attempted murder*” the following information could be identified: $\langle \textit{Woman} \rangle \langle \textit{drove} \rangle \langle \textit{van full of kids} \rangle$, and $\langle \textit{Woman} \rangle \langle \textit{is charged with} \rangle \langle \textit{attempted murder} \rangle$.

Methods for “traditional” information extraction (IE) are targeted at extraction of information about certain predefined relations, normally, with predefined semantic classes for arguments, determined by a certain domain. For example, the target relations can be `Time`, `Location`, or more complex `Acquisition`, `ToBeBornIn`. Restrictions on semantic classes of arguments can be illustrated by the description `ToBeBornIn(HUMAN; LOCATION)`. Domains might be financial, medical, legal domains, etc. Examples of returned extractions are `LocatedIn(Big Ben, London)`, `HasJobTitle(Marissa Mayer, CEO)`.

IE system JASPER [4] extracted information particularly about earnings from corporate reports. SCISOR system [29] included an IE module that extracted information about corporate merges and acquisition from online news. These and other early IE systems were based on so called Knowledge Engineering approach [5]. Knowledge Engineering assumes a rule based approach for detection of a specific type of information. The rules are normally written in an iterative manner by manual inspection of a sample corpus. The shortcomings of these systems is that they are restricted only to the type of information they are designed for and therefore cannot take any arbitrary information into account. A good review of such systems can be found in [49]. On the other hand, the clear predefined semantics of returned instances of relations made it suitable for population of domain-specific ontologies with a very narrow range of relations and concept classes. [40] introduce a method of IE for the bacterium ontology population by learning 10 shallow semantic relations as `p_of`, `s_of`, `t_by`, etc.

In contrast to “traditional” IE, methods for Open IE are able to extract arbitrary information about arbitrary entities. This is gained by detection of syntactic patterns that are universal for meaningful parts of text in a given language. In many approaches, the detection is guided by rules based on matching of the sequences of

Part-Of-Speech (POS) or syntactic tags to domain- and relation-independent patterns. The state of the art method for Open IE was introduced by Fader et al. in [24]. They suggested the following algorithm for detection and extraction of relevant information in text.

- In a preliminary POS-tagged text, first, a verb with its immediate dependent words is detected and considered to be a potential relation phrase.
- Next, lexical constraints may be applied over the detected verb phrase.
- Then, a noun phrase is searched to the left of the verb phrase.
- If found, another noun phrase is searched to the right from the end of the verb phrase.
- If all three components are found, the tuple is extracted.

Due to its relatively simple algorithm that does not involve deep linguistic analysis this approach shows good results in terms of speed of performance, and, consequently, scalability to a Web-sized corpus.

In general, the unrestrictedness of Open IE methods makes them extremely useful in the settings when a relation cannot be defined in advance, possible semantic classes of arguments are not known, or user needs cannot be known. In particular, Open IE seeks applications in machine reading [21], text summarization [48, 49], new perspective on search as question answering [20], automatic text quality evaluation [28, 36] and many others.

However, the open-domain and open-relation approach of Open IE hinders its application to the tasks where knowing the type or semantics of a relation is important. Mainly, such appealing for Open IE applications as ontology population and semantic indexing of documents require mapping of extractions to some predefined relations. Since Open IE methods provide no restrictions on the arguments of a relation, a lot of extracted entities would be difficult to map onto an ontology if needed, or construct an elementary triple appropriate for the representation in RDF format useful for semantic indexing of a document. Some examples of such extractions from Reuters News Corpus [34] are:

⟨Nothing⟩ ⟨could be further from⟩ ⟨the truth⟩
⟨the proverbial straw⟩ ⟨breaks⟩ ⟨the camel's back⟩
⟨a young age⟩ ⟨languished in⟩ ⟨jails⟩

This happens because both extracted arguments has no lexical or other restrictions apart from being noun phrases or pronouns. On the other hand, lexical constraints over verb relation phrases exclude any relations that have not been encountered frequently enough in a corpus where the statistics was gathered. To be able to take into account relation phrases typical to various domains, either analysis of an extremely large and comprehensive corpus is needed or statistics should be gathered on a domain specific corpus.

An approach to adaptation of Open IE to domain-specific relations is suggested in [56]. The eventual goal of their attempt to relation detection is to map the extractions against a given domain-specific ontology, which is seen to be a part of a Question Answering task. In their work Soderland *et al.* modify the original Open IE system TEXTRUNNER for a higher recall, i.e. to return larger chunks of texts than conventional Open IE extraction tuples. Then, they apply domain adaptation rules to the output of the system in two stages. First, they introduce rules to detect domain specific classes, i.e. named-entities and semantic classes. This is done by introducing lists of class-specific key words that are manually learned from a training/development set and extended by synonyms. The output of this step is the extractions enriched with semantic and NE tags on certain terms. At the second stage, they apply domain relation mapping rules that are a set of constraints on tuple arguments and on the context to the left and right of the tuple. This extended context is returned due to the modification of their Open IE system for a higher recall. The Open IE adaptation described in the paper is demonstrated for NFL football domain and corresponding 13 relations.

Detection and interpretation of relations between concepts in a text is known as text graph construction. One of the commonly accepted frameworks for data structure representation is RDF, Resource Description Framework. RDF/XML format is one of the standard formats for representation of semantic graphs. In an RDF graph all nodes are entities which are connected by relations. The core structures of an RDF graph are triples otherwise known as RDF statements that consist of a subject, a predicate and an object. Each component consists of a lexical form and a datatype IRI, International Resource Identifier.

In this chapter we introduce the named entity driven approach to Open IE and a method of extraction post-processing for shallow semantic interpretation of the relations. Our Open IE algorithm shows high precision although trading off for lower recall. Importantly, it facilitates extraction post-processing that allows us to skip additional multi-phase processing as in [56] or trained learning as in [40]. The

precision for relation detection is as high as 93%. We also describe an algorithm for presentation of post-processed extractions in RDF/XML format which essentially is a data graph of extractions.

5.2 Named-Entity-Driven Open Information Extraction with Post-Processing Rules

In this work we propose a method for Open IE that we call named-entity-first open information extraction and a number of post-processing rules for refinement of extraction components and interpretation of relations between arguments. Our method suggests two stages. First, we apply our Open IE technique to a text. Then, the extracted tuples that usually contain complex information pieces are analyzed for more granular relations and converted into extended tuples with more granular arguments and predicates. Importantly, the post-processing rules use the output of linguistic analysis performed at the previous IE stage. By this, no additional linguistics processing, i.e. POS-tagging, named entity recognition, syntactic parsing or others that are known to be resource-consuming. For example, a broader context is normally needed for reliable linguistic analysis while extractions essentially are short fragments.

In this work, we also describe the implementation of a module for presentation of processed extractions in RDF/XML. This is done keeping in mind that the eventual goal of information extraction is to serve in a real-life application that naturally would require extraction conversion into some standardized format.

In the description of the method, we would use examples from Reuters News Corpus [34] as well as extractions from current news articles, different parts of which were used for the development and testing of the method. We restricted ourselves to the domain of newspaper articles because timely and accurate processing of the information conveyed in the large amount of news is important for many applications for business and leisure. And fast processing of large amount of texts is the main advantage of Open IE.

5.2.1 Open information extraction constrained by named entities

We have observed that the most important information in general purpose texts such as news articles (as opposed to narrow-domain texts such as scientific articles, man-

uals and reports) normally is conveyed through description of events performed by a named entity, e.g. “*Evert pledged to [...] slash fuel costs*”, “*The Mexican president shared his thoughts about the U.S. and immigration*”, “*Yahoo has acquired various companies*”, or “*U.S., Europe Impose New Sanctions on Russia*”. In all these examples, a verb phrase relation connects a named entity of a certain type –**Person**, **JobTitle**, **Organization**, **Location**– with another argument. Therefore, in this work we introduce an algorithm for Open IE constrained by named entities. We emphasize that it does not change Open IE’s principle of open-relation approach.

Input. As its input, the algorithm takes POS-tagged and shallow-parsed text with detected named entities. In this work we use ANNIE module of GATE NLP platform [16] for this pre-processing. Also, we only considered **Person**, **JobTitle**, **Organization** and **Location** classes of named entities as more likely agents/subjects of events.

The algorithm works on a sentence level and is as follows:

- First, detect a named entity optionally preceded by dependent words that form a noun phrase chunk, for example, “*The BBC’s correspondent Steve Rosenberg*” looking from right to left.
- Then, search for a verb with dependent words that follows immediately the named entity structure.
- Last, we look for a noun phrase chunk that follows the verb phrase detected on the previous step.

Output. The output is a list of three-component tuples that represent relations such as the following:

$\langle \textit{Belarus} \rangle \langle \textit{would open an economic office in} \rangle \langle \textit{Taipei} \rangle$

or

$\langle \textit{Federal Government} \rangle \langle \textit{plans to increase} \rangle \langle \textit{fees} \rangle$

A first component is an agent/subject, the second is a predicate, and the last one is an object of the relation.

5.2.2 Reported speech extraction

Since in this work our target domain were news articles, we focused on the patterns that are common for this type of texts. In particular, one of the most common grammatical structures that conveys important information is reported speech.

These structures normally include a named entity as a first argument: “*Ukraine **said** Russian tanks had flattened a small border town*”, “*Steve Rosenberg **says** the van belonged to Denis Pushilin*”.

The input is the same as in Section 5.2.1. Then, if a speech verb is detected after the first component, the algorithm follows a different detection sequence:

NP with NE - speech verb - NP - VP - NP,

where NE stands for a named entity, **speech verb** matches against a list of common speech verbs –*say, warn, admit, note, stress, reveal*– optionally followed by preposition *that*, NP - VP - NP is a triple similar to the one described in Section 5.2.1, with that difference that the first component need not contain a named entity. Each component is detected in the left to right order.

Output of this sequence is a five component tuple, e.g.,

⟨*Clinton’s spokesman Mike McCurry*⟩⟨*said that*⟩⟨*the president*⟩⟨*has been on*⟩⟨*the phone*⟩.

The first component is an agent of the reported speech, followed by a speech verb, then a subject of the subordinate relation, a predicate of this relation, and an object.

5.2.3 Detection of target relations in post-processing

As it can be seen from the sample extractions above, the granularity of components of the extracted tuples is not fine enough to be considered semantic units. For example, *Clinton’s spokesman Mike McCurry* exposes an easily detectable **HasJobTitle** relation, *plans to increase* contains actually two actions: *plans* and *to increase*; the relational component *would open an economic office in* contains a syntactic object *an economic office in* (we call it “inner object” for convenience).

By analyzing extractions from news articles, we have concluded that several post-processing rules will bring the extractions to more granular units that are more appropriate for further applications as we previously discussed. These rules give us a fast and easy analysis on semantic level: not only do they detect the limits of finer grained semantic units in a compound extracted component, but also they provide interpretation for connecting relations, either full semantic or shallower lexico-semantic one.

The semantic interpretation can be determined by the following rules:

1. **HasJobTitle** This rule relies on GATE’s named entity recognizer. If an argument of a tuple has a job title entity in front of another named entity, we

split this component and establish a generic relation **HasJobTitle** between the parts. For example, from component $\langle \textit{Assistant Secretary of State John Pelletreau} \rangle$ we get:

$$\langle \textit{John Pelletreau} \rangle \langle \textbf{HasJobTitle} \rangle \langle \textit{Assistant Secretary of State} \rangle.$$

2. **BelongsTo** A possessive apostrophe before named entities was resolved into **BelongsTo** relation. Combined with the previous rule, component $\langle \textit{Ukraine's president Petro Poroshenko} \rangle$ can be analyzed as follows:

$$\langle \textit{Petro Poroshenko} \rangle \langle \textbf{HasJobTitle} \rangle \langle \textit{president} \rangle$$

$$\langle \textit{president} \rangle \langle \textbf{BelongsTo} \rangle \langle \textit{Ukraine} \rangle$$

3. **Date** Since there is a vast amount of works on time construction detection [27], we implemented only a very basic version of date relation detection, based on time words:

$$\langle \textit{package} \rangle \langle \textit{will arrive} \rangle \langle \textbf{Date-on: Sunday} \rangle$$

The semantic analysis of other structures cannot be easily performed based only data that we receive from the pre-processing analysis, i.e., lexical properties, named entity tags, and syntactic chunking tags, and using only context provided by an extraction. Therefore, we restricted ourselves to shallower lexic-semantic analysis:

1. **Inner Object** As we described above, complex relation phrases that contain a syntactic object are also split into parts. From $\langle \textit{remained the biggest market for} \rangle$ we get:

$$\langle \textit{remained} \rangle \langle \textbf{InnerObject: the biggest market for} \rangle.$$

2. **Compound relation with TO** If a relation phrase contains a *to + infinitive* structure, we split it into two parts as in $\langle \textit{plans to increase} \rangle \Rightarrow \langle \textit{plans} \rangle \langle \textbf{to: increase} \rangle$.
3. **Prepositional Relations** This rule clarifies the type of connection between a relational phrase and the second argument based on a connecting preposition. After processing by this rule, extraction $\langle \textit{China} \rangle \langle \textit{will be in} \rangle \langle \textit{the market} \rangle$ will be transformed to:

$\langle \textit{China} \rangle \langle \textit{will be} \rangle \langle \textit{in the market} \rangle$

The level of interpretation by the last two types of rules is similar to the relations considered in [40].

We would like to note, that deep semantic analysis of extractions was not the goal of the current work. Instead, we wanted to “normalize” the components returned by Open IE procedure. By normalization we mean to bring compound components to simpler semantic units and to describe or characterize the relations between the components on a high-level to make possible their presentation in a standard RDF/XML format. We will discuss this and other issues in Section 5.4.

5.2.4 Illustration of the method

Finally, we provide an example of the sequential application of the Open IE method described in Sections 5.2.1 and 5.2.2, and post-processing rules from Section 5.2.3 over a sentence from a news article. We show that it results in a rather substantial analysis of the extraction.

Example. “*The BBC’s correspondent Steve Rosenberg says the van belonged to Denis Pushilin*”.

1. Part-Of-Speech tagging. In our implementation this step is done by GATE NLP platform. For English language, Penn Tree Bank POS-tag set is used.
The^DT *BBC*^NNP *'s*^POS *correspondent*^NN *Steve*^NNP *Rosenberg*^NNP *says*^VBZ *the*^DT *van*^NN *belonged*^VBD *to*^TO *Denis*^NNP *Pushilin*^NNP.
2. Named entity recognition. This is also done by means of GATE NLP platform.
[The BBC]^Organization *'s* *correspondent* *[Steve Rosenberg]*^Person *says the van belonged to* *[Denis Pushilin]*^Person.
3. Syntactic chunking is also performed through GATE NLP platform. *[The BBC’s correspondent Steve Rosenberg]*^NounPhrase *says*^VerbPhrase *[the van]*^NounPhrase *belonged*^VerbPhrase *[to Denis Pushilin]*^PrepositionalPhrase.
4. Information extraction. Due to the presence of a speech verb *says* in the text, the extraction will follow the Reported Speech rule, returning the tuple: $\langle \textit{The BBC’s correspondent Steve Rosenberg} \rangle \langle \textit{says} \rangle \langle \textit{the van} \rangle \langle \textit{belonged to} \rangle \langle \textit{Denis Pushilin} \rangle$.

5. Post-processing rules. Afterward, the post-processing rules are applied to the extraction. $\langle \textit{BelongsTo: The BBC's} \rangle \langle \textit{HasJobTitle: correspondent} \rangle \langle \textit{Steve Rosenberg} \rangle \langle \textit{says} \rangle \langle \textit{the van} \rangle \langle \textit{belonged} \rangle \langle \textit{To: Denis Pushilin} \rangle$.

In the end we get a graph of the extraction relations as shown in Figure 5.1.

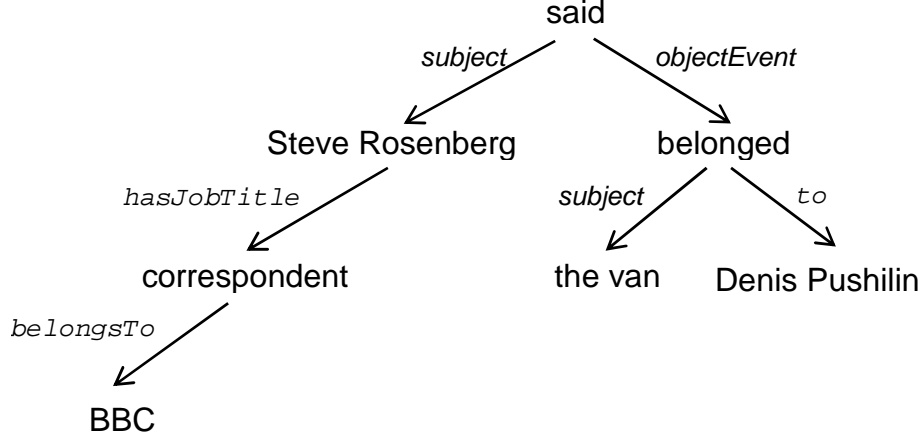


Figure 5.1: Structure of the extraction.

5.3 Experiments of Performance Evaluation

We evaluated our method of Open IE and post-processing rules on a test set of 100 news articles randomly chosen from a subset from Reuter News corpus. We evaluated the general output of extraction system and then evaluated the correctness of the non-trivial rules, namely `HasJobTitle` and reported speech rule. The total number of returned extractions was 306.

The evaluation procedure is as follows. Two human annotators evaluated the extractions as correct or incorrect. The annotators were instructed to judge the following: (1) whether an extraction was meaningful and corresponded to the information conveyed in the original article; (2) whether the partition of a tuple did not violate logical or syntactical structure of the fragment; (3) whether the extraction was logically complete. Some examples of incorrect extractions follow: “ $\langle \textit{Bill Clinton} \rangle \langle \textit{put} \rangle \langle \textit{U} \rangle$ ” does not make sense; in “ $\langle \textit{Kiko Narvaez} \rangle \langle \textit{made} \rangle \langle \textit{it two with} \rangle \langle \textit{a spectacular effort} \rangle$ ” the extraction partition does not correspond to logical relations, and “ $\langle \textit{Iraq} \rangle \langle \textit{said} \rangle \langle \textit{on Sunday} \rangle$ ” is not logically and syntactically complete.

The annotators agreed on 90% of extractions with Cohen’s kappa coefficient $\kappa = 0.80$, which indicates substantial agreement between the annotators.

First, we calculated precision and recall for the output of the extraction system. Precision was calculated as a fraction of correct extractions among all returned extractions. Calculating recall in our settings was a less straightforward procedure. Recall is defined as a fraction of returned correct extractions among all possible (i.e., expected) correct extractions. However, it is not clear how to estimate the number of all possible extractions for an Open IE system. As an approximation to this number, we took the number of all extractions returned by ReVerb system on the same dataset and multiplied it by the approximation of precision value for the case when ReVerb’s Recall $\rightarrow 1$ based on the graphs provided in [24], which is approximately 0.2. We took this approach for estimation of the total number of possible extractions, because the Open IE approach that underlies ReVerb has a similar extraction algorithm and does not have restrictions on non-named entities in the arguments. Consequently, we estimated the number of all possible extractions to be equal to 467. The precision and recall for our system is given in Table 5.1. We also provide the data for the same recall level for ReVerb system.

Table 5.1: Performance of the named-entity-first Open IE method and ReVerb Open IE system at the same recall level.

System	Precision	Recall
NE 1st	0.79	0.51
ReVerb	~ 0.60	~ 0.50

We observe a very high precision-recall ratio compared to the ones achieved in [24]. However, due to the imposed restriction on named entities in extraction arguments, Recall = 0.51 is the highest achievable level of recall. In other words, our system shows higher precision by filtering out potential extractions that do not contain named entities.

Further, we evaluated performance of the post-processing rules that detect semantics: **HasJobTitle** and reported speech rules. In this case we calculate accuracy as a relation of the number of correctly detected and interpreted relations between correct arguments to the number of all returned relations of the corresponding type.

The accuracy for **HasJobTitle** relation is as high as:

$$\text{accuracy_HasJobTitle} = 0.93 \quad (5.1)$$

The achievement of such a high accuracy is due to the efficiency of the underlying named entity recognizer that is used at the linguistic analysis stage and the easily detectable syntactic form when a job title precedes a **Person** named entity.

However, the accuracy of the reported speech relation is relatively low:

$$\text{accuracy_ReportedSpeech} = 0.22 \quad (5.2)$$

This is due to the fact that reported speech is more often an N-ary relation: “*Somebody told something to somebody else*”; or introduce a dependent clause: “*Somebody said on Sunday that X had happened*”. We discuss in Section 5.5 the current difficulties with processing of non-binary relations.

5.4 Conversion into RDF/XML format

Although extractions in text format can be convenient for human readers or annotators, any further machine processing application will require presentation of extractions in some standard data presentation format. In this work we focus on RDF standard introduced and maintained by W3C (The World Wide Web Consortium) and used for Semantic Web applications, and on its RDF/XML format in particular. As we mentioned throughout the work, Open IE has a lot of applications to larger NLP tasks: semantic indexing of text documents, mapping of extracted information onto an ontology, structural modeling of text information presented in a web page, just to name a few.

Here we suggest a procedure of conversion of extractions processed by our post-processing rules into RDF statements, by this building an RDF graph of the extraction. The output of this procedure is corresponding RDF statements in RDF/XML format.

As we showed in Section 5.2, the method of Open IE with the post-processing rules suggested in this work essentially builds a relation graph for an extraction. After the post-processing, each final component of the extraction can be considered as a node in a graph. However, to be considered an IRI (International Resource Identifier) as required by RDF standard, it should be normalized. Here we suggest the following normalization procedure:

- Eliminate all determiners.
- Eliminate lexical preposition. We have already used them for shallow relation interpretation at the post-processing stage.
- Delete spaces between proper nouns of a named entity keeping all first letters capitalized, e.g., *Steve Rosenberg* \Rightarrow *SteveRosenberg*.

- Delete spaces inside other multi-word components converting the first word into lowercase and capitalizing all the following words, e.g., *Deputy Finance Minister* \Rightarrow *deputyFinanceMinister* or *is planning* \Rightarrow *isPlanning*.

At the current moment, the problem of mapping of an arbitrary concept to a universal comprehensive ontology is not solved and such an ontology does not exist, although there is a few promising projects such as WordNet [57], ProBase [65], NELL [11]. An ontology roughly corresponds to an RDF vocabulary. In the absence of a universal comprehensive RDF vocabulary, we suggest using a very high-level syntactic based vocabulary that includes the following items: **verb**, **subject**, **object**, **objectEvent**, and items corresponding to prepositional relations.

Using this vocabulary, a graph that corresponds to the extractions from “*The BBC’s correspondent Steve Rosenberg says the van belonged to Denis Pushilin.*” is depicted in Figure 5.2.

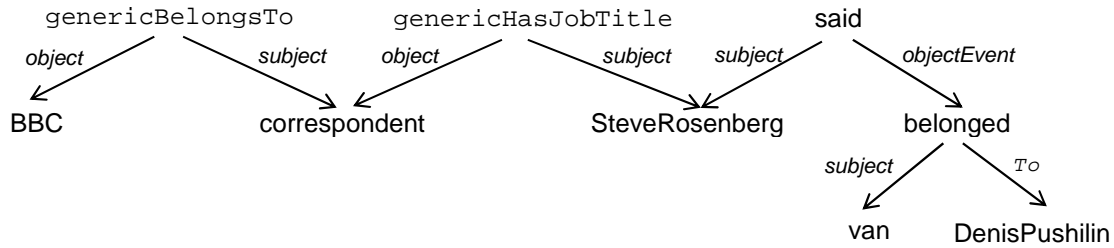


Figure 5.2: An RDF graph corresponding to the extraction.

As we previously mentioned, a standard RDF statement is a triplet. However, the original sentence contains more complex relations. Therefore, they are split into triples. This is done by introducing “dummy” nodes that do not correspond to any lexical expression. In the end, we get an RDF/XML representation as shown in Listing 5.1.

Listing 5.1: RDF/XML representation of information extracted from sentence “*The BBC’s correspondent Steve Rosenberg says the van belonged to Denis Pushilin.*”

```

<rdf:Description rdf:about="doc:eventSpeech1">
<dpp:verb>docTerm:says</dpp:verb>
<dpp:subject>docTerm:SteveRosenberg</dpp:subject>
<dpp:objectEvent>event1</dpp:objectEvent>
</rdf:Description>

<rdf:Description rdf:about="doc:event1">
<dpp:verb>docTerm:belonged</dpp:verb>

```

```
<dpp:subject>docTerm:van</dpp:subject>
<dpp:to>DenisPushilin</dpp:to>
</rdf:Description>
```

```
<rdf:Description rdf:about="doc:jobTitle1">
<dpp:subject>docTerm:SteveRosenberg</dpp:subject>
<dpp:object>docTerm:correspondent</dpp:object>
</rdf:Description>
```

```
<rdf:Description rdf:about="doc:belongTo1">
<dpp:subject>docTerm:correspondent</dpp:subject>
<dpp:object>docTerm:BBC</dpp:object>
</rdf:Description>
```

The numbering is given to the statements to make them unique.

5.4.1 Format Validation

The output of this conversion to RDF/XML format passes validation test by the official W3C RDF/XML validator at <http://www.w3.org/RDF/Validator/> and can be constructed for any arbitrary text processed by our Open IE method with the corresponding post-processing.

5.5 Discussion

Existing methods of Open IE proved to be very promising and efficient for extracting arbitrary information from texts of an arbitrary domain. They have doubtless advantages against methods of traditional information extraction in that they are not restricted to target relations or predefined classes. Also, they are fast and scalable to the Web. Additionally, the recent approaches to Open IE including the one suggested in the current work do not require time and resource consuming preliminary training of the algorithm.

The novelty of the method that we propose in this chapter is that we introduce: (1) a named-entity-first approach to detection of potentially informative extractions; (2) an additional detection pattern for correct reported speech extraction; (3) post-processing rules based only on the linguistic data already used for the extraction stage. Applied to an extraction, these rules allow building of a shallow semantic

graph of the extraction. The main advantage of the post-processing rules is that their application does not need any additional linguistic analysis of a text fragment, because all the analysis has already been done at the pre-processing for the extraction stage.

Our work generally lies in the direction towards adaptation of methods of Open IE to practical application. As some of the principle applications of news text analysis we see ontology population and document semantic indexing. These tasks in particular require higher precision and allow sacrificing recall because fewer high precision extraction can give better overview of what a document is about although leaving out some detail.

As we discussed in Section 5.4, any further processing of extractions inevitably requires their presentation in a particular format. Although some tools work with custom formats, the common practice is to use one of the existing standardized data presentation formats. Considering Open IE as a necessary step for automatic Knowledge Base population and Knowledge Graph construction, which is used for Question Answering, Semantic Indexing, Search, Recommendation Systems and so on, we elaborated a method for presentation of extractions in RDF/XML format which is a standard format for data presentation recognized and maintained by W3C (The World Wide Web Consortium). The method is a straightforward conversion of the results of analysis performed by the post-processing rules into RDF statements following the requirements of RDF standard.

We remind the reader, that one of the main requirements of the RDF framework is that an RDF statement must be a triplet, i.e., a binary relation connecting two arguments. Thanks to the form of the extraction patterns, the extracted information comes in the form of tuples that are either triples or can be converted into triples as we show in Section 5.4.

Nevertheless, many of relations that are encountered in the real world, and likely the majority of the relations, are N-ary. An N-ary relation connects more than two arguments, e.g., “*John buys a “Lenny the Lion” book from books.example.com*”. Here a relation *buys* connects an individual *John*, an object “*Lenny the Lion*” book and a seller *books.example.com*. Our extractor can analyze this phrase and return an extraction that can be further processed by the post-processing rules returning in the end: $\langle John \rangle \langle buys \rangle \langle \text{InnerObject: a “Lenny the Lion” book} \rangle \langle \text{from: books.example.com} \rangle$. Yet conversion of this relation into a set of triples is an open question. Although there are general guidelines suggested in [43], currently there is no general solution for this problem.

Chapter 6

Application to Measuring Informativeness of Web Documents

In Section 2.5 we talked about different applications of Open IE to more complex task. One of the tasks where Open IE has found efficient application is measuring quality of texts, of texts on the Web in particular. Indeed, Open IE is quite appropriate for this task: it is domain- and relation-open and scalable to the Web. In this chapter we will show that Open IE is applicable to measure the quality of textual contents of arbitrary Web documents.

In this chapter we will discuss:

- The importance of the problem of automatic measurement of text quality;
- Overview previous attempts to solve this task;
- Describe the method for text quality assessment;
- Describe our experiments and results.

6.1 Motivation

Assessment of information quality becomes increasingly important because nowadays decision making is based on information from various sources that are sometimes unknown or of questionable reliability. Besides, a large part of the information found in the Internet has low quality: the Internet is flooded with meaningless blog comments, computer-generated spam, and documents created by copy-and-paste that convey no useful information.

As one might assume, talking about the quality of the Internet content on the whole is too general and practically impossible, because the content on the Web is of an extremely versatile form and serves to very different needs. It is unreasonable to compare the quality of an online encyclopedia to the quality of a photo storing resource because the assessment of the quality of any object depends on the purpose to which the object serves. Hence, we restrict ourselves to the scope of text documents and assume that the general purpose of a text document is to inform a reader about something. Therefore, the quality of a text document can be related to its *informativeness*, i.e. the amount of useful information contained in a document. [36] suggest that informativeness of a document can be measured through factual density of a document, i.e. the number of facts contained in a document, normalized by its length.

Due to the lack of a standard corpus, previous works on the estimation of Web quality considered only Wikipedia articles, in fact assessing their informativeness. No special studies were performed about human judgment on text informativeness. Therefore, [36, 8, 35, 37] considered Wikipedia editors' choice of *featured* and *good* articles as a reasonable extrapolation of high judgment on their informativeness. The work [36] showed the feasibility of factual density application as measurement of informativeness on the base of automatic prediction of the featured/good Wikipedia articles.

In this work we have conducted experiments to estimate the adequacy of application of factual density to informativeness evaluation in the “real” Internet, i.e. not limited to particular web-sites with a particular form of content but rather covering a wide variety of web-sources, which a user could browse through while looking for information. For this purpose we created a dataset of 50 randomly selected documents in Spanish language from CommonCrawl corpus [30], which is a large extraction of texts from the Internet. We assessed factual density automatically using our Open IE system for Spanish language, EXTRHECH, which is adequate for Web-scale ap-

plications. Further, 13 human annotators ranked 50 documents according to their informativeness using the MaxDiff [38] technique. The automatic ranking produced by EXTRHECH system correlates with the ground truth ranking by human annotators with Spearman’s ρ coefficient of 0.41 (coinciding rankings would have 1, and the random baseline is 0.018).

6.2 Previous Work in Text Quality Evaluation

Evaluation of the quality of Web text content has been mainly performed with metrics capturing content quality aspects like objectivity [35], content maturity, and readability [62]. These methods are based on selection of appropriate features for document presentation. For example, in [35] stylometric features were used to assess the content quality. Character trigram distributions were exploited in [37] to identify high quality *featured/good* articles in Wikipedia. [8] considered simple word count as an indicator for the quality of Wikipedia articles. [36] proposed factual density as a measure of document informativeness and showed that it gives better results for Wikipedia articles than other methods. Wikipedia articles were taken into consideration mainly due to the lack of a standard corpus in this field of work. For evaluation purposes, those Wikipedia articles that have the featured article or good article template in the wikitext were considered to be of a high quality or more informative. No specially designed human annotation or evaluation was involved, and no scale or ranking of informativeness was introduced.

To asses factual density of a text document, [36] apply Open IE methods. She treats each extraction as a fact stated in the document. The, she proposes a factual density measure

$$\text{Score}_{\text{factdens}}(d) = \text{fc}(d)/\text{size}(d), \quad (6.1)$$

where $\text{fc}(d)$ is the fact count for a document d , and $\text{size}(d)$ is its length in characters including white spaces. Lex *et al.* showed taht this text quality measure was adequate for Wikipedia documents.

Our task was to prove whether this measure is adequate for arbitrary Web texts as well.

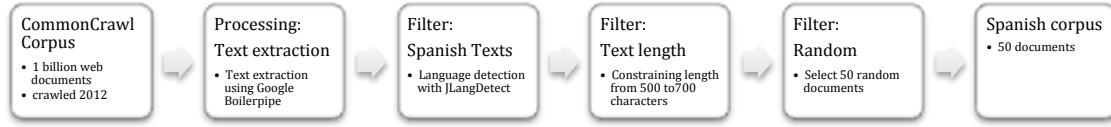


Figure 6.1: Process of corpus generation

6.3 Building the Ground Truth Dataset

Since no special corpus for informativeness evaluation previously existed, we aimed at creation of such a corpus of texts extracted from the Internet broader than Wikipedia.

6.3.1 The dataset

For the purpose of evaluation of the factual density as a measure of informativeness, we needed to create a dataset that would be a reasonable projection of texts on the Web and small enough to be able to conduct the experiment with available resources. To create our dataset we performed the following steps:

- We used a 1 billion page subset from the CommonCrawl corpus [30] from 2012, which is a corpus of the web crawl data composed of over 5 billion web pages, as an initial source of Web texts. From that corpus, we extracted textual content of websites using Google’s Boilerpipe framework [31].
- For each article, the language was detected using JLangDetect [42].
- From this dataset, we randomly selected 50 documents in Spanish. In order to avoid the length-based bias on the human annotation stage described in the next subsection (e.g. users might tend to rate longer texts as more informative than shorter ones), we constrained the text length to range from 500 to 700 characters.

In the end, we formed a corpus of 50 text documents in Spanish of similar length that represent a random sample of the textual content from the Web. Figure 6.1 shows the process.

We would like to emphasize that not all textual content presented on the Web is coherent text. The texts encountered on the Internet can consist of pure sequences of keywords, or be elements of web-page menus, for example, “*For more options click here Leave your comment CAPTCHA*”. Lists and instructions are another common

form of texts, characterized by incomplete sentences normally starting with a verb in the infinitive or imperative form, e.g. “*To open a file: – click the Microsoft Office button; – select Open*”. Texts can be sets of short comments or tweets that also tend to be incomplete sentences often lacking grammatical correctness. Commercials and announcements also typically consist of incomplete sentences, e.g. “*Information about the courses, dates, and prices*”, numbers, e.g. “*\$65 \$75 \$95 All prices in US dollars*”, and telephone numbers and addresses. We manually performed a rough classification of the texts from our dataset shown in Table 6.1. Since we used only short documents for the experiment, each document mainly corresponded to only one text type.

Table 6.1: Classification of the documents in the dataset by the types of text content

Type of text	# of docs	Characteristics
keywords	2	sequence of nouns with no verbs
web page menu	1	short phrases, verbs
commercials, announcements	18	addresses, phone numbers, prices, imperatives
coherent narrative: descriptions, news	13	full sentences with subjects, verbs, and objects
comments, tweets	6	short sentences that lack grammatical correctness
instructions, lists	9	phrases starting with infinitives or no verbs
incorrectly detected language	1	impossible to POS-tag for the system and to read for human annotators

In the current work we did not do any additional pre-processing for text type detection. This was not done for several reasons. First, we want to keep the system as simple and fast as possible for the purpose of scalability to large amounts of text. Next, we believe that the factual density approach presented in the chapter will be appropriate for automatic detection of incoherent and uninformative texts. Consequently, there will be no need for additional filtering.

6.3.2 Ground truth ranking by human annotators

To overcome the lack of a standard corpus in the field of web text informativeness assessment, we formed a ground truth ranking of the documents based on inquiry of 13 human annotators. All human annotators are natively Spanish speaking people

with graduate level of education. For the questionnaire, we opted for the MaxDiff (Maximum Difference Scaling) technique [38]. According to MaxDiff technique, instead of ranking all items at once, a participant is asked to choose the best and the worst item from a subset of items at a time. This procedure is repeated until all items are covered.

MaxDiff is considered to be a strong alternative to standard rating scales [2]. The advantage of the MaxDiff technique is that it is easier for a human to select the extremes of a scale rather than to produce a whole range of scaling. Consequently, MaxDiff avoids the problems with scale biases and is more efficient for data gathering than the simple pairwise comparison.

For this purpose, we created a set S of MaxDiff questions. Each question, which will subsequently be called Q , contained 4 different documents d . Four items is few enough for a participant to be able to concentrate and to make a decision, and large enough to be able to cover all 50 documents in a reasonable number of questions.

The set S was created as follows:

- First, we calculated all possible combinations of how 4 documents can be chosen from 50 documents. This resulted in 230,300 possible combinations.
- Then, in a loop, we picked up one random question Q from the set of combinations and added it to the resulting set S , until every document d was included three times in the set S . This ensures that every document is compared at least three times with other documents. Once finished, the resulting set S contained 103 questions Q , which we used for the MaxDiff inquiry.

Further, we created a MaxDiff questionnaire system that displayed one question Q' ($= 4$ documents d') at a time to a participant. A participant was asked to select the most informative document and the least informative document from the set of 4 documents. The interface of the questionnaire system is shown in Figure 6.2. In the experiment, the instructions and the documents were given to the annotators in Spanish language. In Figure 6.2 they are translated into English for convenience of the reader of this article. The selection of the most and the least informative documents was based on the intuitive understanding of the notion of “informativeness” by the participants. Each of the participants rated 25 questions on average. The system ensured that each question Q is (i) rated three times in total and (ii) rated by different users. This resulted in 309 ratings, with each question being answered 3 times.

Rated documents: 7

Which is the most and the least informative document?

Continue

Hint: You can use the 's' key to continue

Digg Digg
A great loss for the financial sector happened today with the partial takeover of the Anglo Irish Bank by the government the existing Board of Directors have all resigned. Anne Heraty who is running the largest Irish recruitment agency CPL also resigned today from her Directors role in the Anglo Irish Bank. I see it as a loss for the whole financial industry, to lose people like Anne Heraty. She is a clear leader of the whole recruitment industry in Ireland, and actually quite unique in her capacity and the leading role in the industry in general.
?

Fast virtual private servers with full root access and dedicated resources. Windows & Linux available. Ideal for Resellers and Developers. From... £8.49 per month Daily.co.uk is one of the fastest growing Web Hosting companies in the UK. We provide low cost domain name registration combined with a world class Web Hosting platform that gives our customers unparalleled control of their web presence. Our award winning VPS Hosting is used by Web Designers and Developers. Many opt for our Linux VPS Hosting service, and others prefer our Windows VPS Hosting service.

more products
Fire Door Sign - This Staircase Must Be Kept Clear of All Obstructions
This fire door sign assists you in conforming to the Building Regulations 1991 act, which advises that all doors designed to be fire resistant should display an appropriate mandatory fire safety sign. All signs comply to BS5499 Part 1 and 5, and are all available in 1mm rigid plastic, self adhesive vinyl, or rigid plastic with self adhesive backing. We can also provide you with custom fire doors signs with your own text, click here . All prices include VAT

Black Anodised Belly Bar
Here we have a superb quality Bellybar. This is made from full surgical grade stainless steel which is then anodised to give it its black finish. In this one we have bright, sparkly synthetic Deep Red stones - one at each end. the smaller ball unscrews for fitting. The clip in the middle is not part of it and is there just to display it. Note - The price for one is £8:00 but if you want to order any 3 from the £8 range please phone Debs in the office as she will arrange a discount for you. 3 for £20 Price: ?8.00

Figure 6.2: Screenshot of the MaxDiff questionnaire tool

We applied the MaxDiff technique to the answers obtained from the user-study. The rank for each document d was calculated proportionally to its MaxDiff scoring $\text{Score}_{\text{maxdiff}}(d)$, which was calculated using the following formula:

$$\text{Score}_{\text{maxdiff}}(d) = R_{\text{pos}}(d) - R_{\text{neg}}(d), \quad (6.2)$$

where $R_{\text{pos}}(d)$ is the number of positive answers and $R_{\text{neg}}(d)$ is the number of negative answers for the same document d .

After calculating the score for each document d , we formed a ground truth ranking of the 50 documents in our dataset.

6.4 Automatic Measurement of Document Quality

The aim of this work is to study the feasibility of automatic factual density estimation for informativeness measurement for text documents on the Web. This section describes the procedure for the automatic ranking.

In the factual density approach to web informativeness assessment, each text

document is characterized by the factual density feature. To calculate the value of factual density, first, the simple fact count is determined for each document d using the Open IE method. That means that only direct information on the number of facts, i.e. fact count $fc(d)$, obtained from a text resource d is taken into account. It is obvious that the fact count in a document is correlated with its length, i.e. longer documents would tend to have more facts than shorter ones. To overcome this dependency, factual density $fd(d)$ is calculated as a fact count in a document $fc(d)$ divided by the document size $size(d)$: $fd(d) = \frac{fc(d)}{size(d)}$.

In this work we used our Open IE system for Spanish EXTRHECH to determine the fact count in a document. The length of a document was calculated as a number of characters including white spaces.

6.5 Experiment and Results

In this work we conducted an experiment to study the appropriateness of factual density measure for assessment of web text informativeness. In order to prove the hypothesis, we compared the ranking based on automatic factual density scoring to the ground truth ranking based on the MaxDiff inquiry of human annotators.

To form the factual density based ranking, 50 documents were fed into a pipeline: Freeling-2.2 POS-tagger, ExtrHech Open IE system, and a script for factual density calculation shown in Figure 6.3.

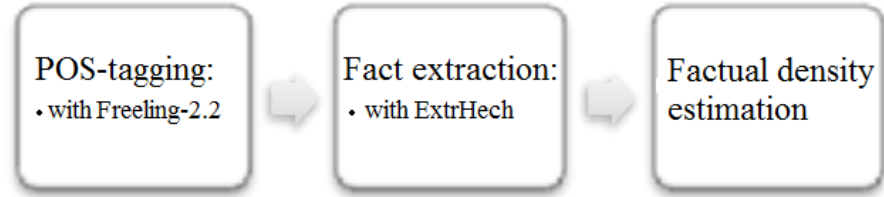


Figure 6.3: Diagram of the factual density estimation

Then, each document d was ranked according to its factual density scoring $\text{Score}_{\text{factdens}}(d)$:

$$\text{Score}_{\text{factdens}}(d) = fc(d)/size(d), \quad (6.3)$$

where $fc(d)$ is the fact count for a document d , and $size(d)$ is its length in characters including white spaces.

Human annotator ranking was formed as described in Section 6.3.2. The rankings are shown in Table 6.2, where HA rank is the human annotator ranking and FD rank is the factual density ranking.

Table 6.2: Human annotator ranking and factual density ranking

Doc ID	HA rank	FD rank	Doc ID	HA rank	FD rank
1	1	3	26	24.5	40
2	2.5	11	27	27	14
3	2.5	29.5	28	29.5	15
4	4.5	2	29	29.5	22
5	4.5	7	30	29.5	32
6	6	6	31	29.5	16
7	7	12	32	33	24
8	8.5	33	33	33	9
9	8.5	1	34	33	43
10	10	5	35	37	47
11	11.5	27	36	37	47
12	11.5	8	37	37	35
13	14.5	42	38	37	34
14	14.5	39	39	37	10
15	14.5	19.5	40	40	17
16	14.5	41	41	41.5	47
17	19	21	42	41.5	47
18	19	4	43	44	18
19	19	29.5	44	44	47
20	19	36	45	44	37
21	19	38	46	46	26
22	22	23	47	47	19.5
23	24.5	13	48	48	25
24	24.5	47	49	49	31
25	24.5	47	50	50	28

Once the rankings were scored, we applied various statistical measures to calculate the correlation between them. Table 6.3 shows the results of the statistical evaluation using Spearman’s ρ coefficient, Pearson product-moment correlation r , and Kendall’s rank correlation τ with the corresponding levels of significance. All these correlation coefficients may vary between 1 for coinciding rankings and -1 for completely opposite rankings. Random baseline for Spearman’s ρ is 0.018. The coefficients should be significantly positive for the rankings to be considered correlated.

In our work we obtained Spearman’s ρ as high as 0.41, Pearson’s r of 0.38, and Kendall’s τ of 0.29. Since all measures give significantly positive correlations with

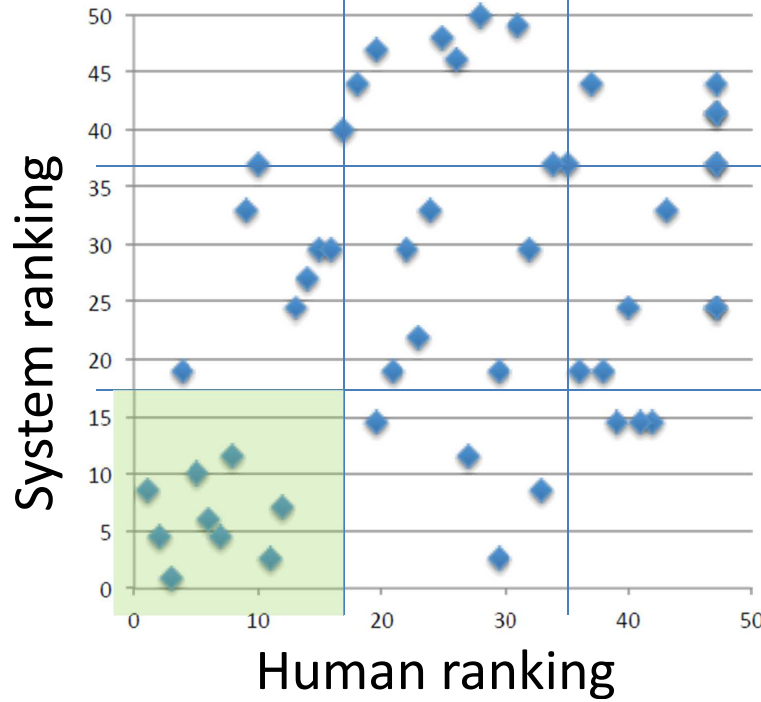


Figure 6.4: Rank correlation

the significance level equal to or higher than 99.49%, we can conclude that medium correlation significantly exists between the two rankings. Consequently, the obtained result show that the factual density measure proves to be feasible as a measure of informativeness for text content on the Web.

Table 6.3: Result of correlation tests between factual density algorithm ranking and human annotator ranking for 50 document dataset

Method	Value	P-Value	Significance Level
Spearman's ρ	0.404	0.00365	99.636%
Pearson's r	0.390	0.00514	99.486%
Kendall's τ	0.293	0.00347	99.653%

6.6 Discussion

Before we conclude this chapter with the description of successfully performed experiment, we would like to discuss the obtained results in detail. Figure 6.4 shows the documents in the space of the two rankings. For any given document (represented as a blue dot), the X-axis shows the ranking in the human questionnaire, and the Y-axis the ranking from the algorithm.

One can observe that although the correlation is not ideal (indeed, the correlation coefficients show moderate correlation in Section 6.5), the correlation in the lower left area (highlighted in green) is higher than the correlation in other areas. This area corresponds to the higher rankings, i.e., “top” quality documents. This observation reveals that although the current method might not achieve the highest accuracy in predicting the entire ranking, it definitely can be used to predict top quality documents. This proves the merits of Open IE as a component of much more complex NLP tasks that can have a direct impact on a user.

In this experiment we studied the adequacy of application of Open IE to factual density estimation and to its usage as a quality measure for arbitrary Web texts. Specifically, this work has been done on the material of Spanish language. As mentioned in Section 6.2, the text quality measuring through factual density estimation with an Open IE system for English was shown to be adequate for English language Wikipedia articles [36]. Our future goal is to show the adequacy of the method for arbitrary English texts as well, especially social media texts.

Chapter 7

Conclusions and Future Work

This chapter concludes this thesis. Here we summarize the final conclusions, reiterate the contributions of the work, state limitations of the proposed methods, and outline the future work.

We will talk about the following:

- Finalizing conclusions of the work;
- Main contributions;
- Limitations of the introduced methods;
- Future work;
- Papers that have been published while accomplishing this work.

7.1 Conclusions

In this work we provided an extensive exploration and analysis of existing methods of Open IE. We introduced a novel method for Open IE that lie in the rule-based area. We showed that a method based on minimal input pre-processing, i.e., just POS-tagging with no syntactic analysis, and a few ingenious heuristics can show at least as good performance as methods based on more complex pre-processing and, hence, more resource consuming. We also showed that this method for Open IE can be successfully applied to the complex task of measurement of informativeness for documents on the Web of an arbitrary domain.

However, semantic interpretation of extractions returned by this approach and their application to ontology population, which is a necessary step in most of knowledge-related NLP tasks, is an open question. Working in this direction, we introduced a different rule-based method for Open IE that takes as its input POS-tagged and NE-labeled text, and we added post-processing rules. We showed that these changes lead to shallow semantic interpretation and to straightforward conversion of extractions into a standard RDF/XML format.

To summarize, we have achieved the objectives that we set for this work:

- We have designed a method for Open IE that is (1) domain-independent, (2) unsupervised, and (3) robust that assures its scalability to the Web.
- Our method shows better performance in terms of precision and robustness than other existing methods for Spanish language and methods based on similar information detection methodology for English language.
- Our method shows higher robustness because it requires minimum input pre-processing.
- We also have proved its usefulness for more complex NLP tasks such as automatic evaluation of text informativeness and suggested a procedure for shallow semantic interpretation of extractions.

Having summarized the main points of the work, we will move on to its contributions.

7.2 Contributions

This work has several contributions to the field of natural language processing.

7.2.1 Theoretical contributions

At the theoretical level we have contributed:

- a robust and high-performance rule-based method for Open IE text that requires only POS-tagged text input. We will remind to the reader, that POS-tagging is one of the most reliable and less resource consuming stages of language processing. We developed this method for Spanish language. We also showed that it can be
- a direct and successful application method of this Open IE approach to such a complex NLP task as quality assessment of arbitrary-domain texts on the Web;
- a novel approach to Open IE methods that includes some semantic processing of the input, namely, NE recognition. This different perspective on the pre-processing of the input allows deeper semantic interpretation of extracted relations.

7.2.2 Practical contributions

Realization of the work required implementation of our method in the form of an Open IE system and elaboration of several datasets for performance evaluation. Therefore, we contributed the following software and resources:

- EXTRHECH, a software system for Open IE for Spanish language¹ based on our method;
- a labeled parallel English-Spanish version of FactSpCIC dataset²;
- a labeled parallel English-Spanish dataset of 300 sentences from news articles³;
- a labeled dataset of 159 sentences in Spanish extracted randomly from the Web ⁴.

¹Available for download from https://bitbucket.org/alisa_ipn/extrhech.git

²<http://www.gelbukh.com/resources/spanish-open-fact-extraction#FactSpCIC>

³<http://www.gelbukh.com/resources/spanish-open-fact-extraction#news>

⁴<http://www.gelbukh.com/resources/spanish-open-fact-extraction#RawWeb>

7.3 Limitations

Naturally, the introduced methods have certain limitations. We will discuss them below.

The general limitation of the entire family of methods based only on POS-tagging or shallow syntactically parsed input for Open IE shared by our first method implemented in EXTRHECH as well as by methods suggested by Fader *et al.*, Gamallo *et al.*, Aguilar-Galicia, etc., in [24, 25, 1] is that these methods detect only relations expressed through verbs. Consequently, considering a phrase “*Jugador mexicano el Chicharito Hernández*” (“*Mexican player Chicharito Hernandez*”), these methods does not allow us to conclude that a person named *el Chicharito Hernández* is a player and is from Mexico. Yet our second method for Open IE, which includes NE-recognition as an input pre-processing stage, solves this issue reasonably well.

However, the named-entity driven method to Open IE suffers from much lower recall because it starts relation detection only when found a named entity. Yet it shows a very high precision on the same datasets as other methods.

Concerning our method for Open IE application to automatic text quality assessment, we have showed that it performs best for the top ranked best-quality documents. Yet, it has not achieved such clear correlation with the human-provided ranking for the rest of the documents. Nevertheless, we are quite optimistic about this result, because it is the top ranked documents that normally have significance for any ranking.

7.4 Future Work

This work has solved several issues in the field of open information extraction. However, every novelty brings more questions to be answered and new things to be discovered. As further directions for this work we see the following:

- to perform detailed analysis on how POS-tagger accuracy affects POS-tag based Open IE;
- 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;
- to improve handling of the inverse word order, relative clauses, and coordinating conjunctions in Open IE method for Spanish;

- to modify the named-entity-driven method in order to increase its recall.

7.5 Publications

Publications in JCR-indexed journals:

- Alisa Zhila, Alexander Gelbukh. Automatic Identification of Facts in Real Internet Texts in Spanish using Lightweight Syntactic Constraints: Problems, Their Causes, and Ways for Improvement. *Revista Signos. Estudios de Lingüística*, 87, vol. 48. In print, 2015.
- (*in preparation*) Alisa Zhila, Alexander Gelbukh, Helena Gomez-Adorno. Named-Entity-Driven Open Information Extraction with Post-Processing Rules.

Book chapters:

- Alisa Zhila, Alexander Gelbukh. Análisis de una aplicación multilingüe del agrupamiento de textos (Analysis of a cross-lingual application of context clustering). In: *Avances en Inteligencia Artificial (Advances in Artificial Intelligence)*, Mexican Society for Artificial Intelligence, pp. 45–57, 2012.

Publications in proceedings of international conferences:

- Alisa Zhila, Scott Yih, Chris Meek, Geoffrey Zweig and Tomas Mikolov. Combining Heterogeneous Models for Measuring Relational Similarity. In *Proceedings NAACL-2013*, 2013. Christopher Horn, Alisa Zhila, Alexander Gelbukh, Elisabeth Lex. Using Factual Density to Measure Informativeness of Web Documents. In *Proceedings NoDaLiDa'13*, 2013.
- Alisa Zhila, Alexander Gelbukh. Comparison of Open Information Extraction for Spanish and English. In *Proceedings Dialogue'2013*, 2013.
- Alisa Zhila, Alexander Gelbukh. Exploring context clustering for term translation. In *Computational Linguistics and Intellectual Technologies*, 11, Vol. 1, pp. 716–725, 2012.

Presentations at international conferences:

- Alisa Zhila, Alexander Gelbukh. Open Information Extraction for Spanish Language based on Syntactic Constraints. *ACL SRW*, 2014.

- Alisa Zhila, Alexander Gelbukh. Informativeness and Objectivity of Texts on the Web, Tapia Celebration of Diversity in Computing, 2014.
- Alisa Zhila, Alexander Gelbukh, Christopher Horn. Open Information Extraction for Spanish and Its Application to Measuring Informativeness of Web Documents. Tapia Celebration of Diversity in Computing, 2013.

7.6 Awards and Invited Talks

International awards:

- Microsoft Research Latin America Fellowship, 2012. This award is given to two PhD students in Latin America each year.

Internships:

- Microsoft Research Internship, 2012.
- Oracle MDC Internship, 2014.
- Yahoo Internship, 2014.

Invited talks:

- MICAI 2014, Mexican International Conference on Artificial Intelligence, 2014.
- COMIA 2012, Mexican Conference on Artificial Intelligence, 2012.
- 7° Foro PIFI 2012, Programa Institucional de Formación de Investigadores, 2012.

Bibliography

- [1] Honorato Aguilar-Galicia. Extracción automática de información semántica basada en estructuras sintácticas. Master’s thesis, Center for Computing Research, Instituto Politécnico Nacional, Mexico City, D.F., Mexico, 2012.
- [2] Eric Almquist and Jason Lee. What do customers really want? <http://hbr.org/2009/04/what-do-customers-really-want/ar/1>, apr 2009. [last visited on 09/04/2013].
- [3] Víctor M. Alonso-Rorís, Juan M. Santos Gago, Roberto Pérez Rodríguez, Carlos Rivas Costa, Miguel A. Gómez Carballa, and Luis Anido Rifón. Information extraction in semantic, highly-structured, and semi-structured web sources. *Polibits*, 49:69–75, 2014.
- [4] Peggy M. Andersen, Philip J. Hayes, Alison K. Huettner, Linda M. Schmandt, Irene B. Nirenburg, and Steven P. Weinstein. Automatic extraction of facts from press releases to generate news stories. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC ’92, pages 170–177, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.
- [5] Douglas E. Appelt. Introduction to information extraction. *AI Commun.*, 12(3):161–172, August 1999.
- [6] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007.
- [7] Michele Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36. Association for Computational Linguistics, June 2008.

- [8] Joshua E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1095–1096, New York, NY, USA, 2008. ACM.
- [9] I.A. Bolshakov and A. Gelbukh. *Computational Linguistics: Models, Resources, Applications*. Ciencia de la computación. Instituto Politécnico Nacional, 2004.
- [10] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.
- [11] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.
- [12] Daniel Cer, Marie-Catherine de Marneffe, Daniel Jurafsky, and Christopher D. Manning. Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *7th International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.
- [13] Joseph Clancy Clements. Primary and secondary object marking in spanish. In J. Clancy Clements and Jiyoung Yoon, editors, *Functional approaches to Spanish syntax: Lexical semantics, discourse, and trasitivity*, pages 115–133. London: Palgrave MacMillan, 2006.
- [14] Usage of content languages for websites. <http://www.internetworldstats.com/stats7.htm>, Dec 2013. [last visited on 04/11/2014].
- [15] Usage of content languages for websites. http://w3techs.com/technologies/overview/content_language/all, Nov 2014. [last visited on 04/11/2014].
- [16] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.
- [17] Sandipan Dandapat, Sudeshna Sarkar, and Anupam Basu. Automatic part-of-speech tagging for bengali: an approach for morphologically rich languages in a

- poor resource scenario. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 221–224, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
- [18] Nibaran Das, Swarnendu Ghosh, Teresa Gonçalves, and Paulo Quaresma. Comparison of different graph distance metrics for semantic text based classification. *Polibits*, 49:51–57, 2014.
- [19] Luciano Del Corro and Rainer Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 355–366, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [20] Oren Etzioni. Search Needs a Shake-Up. *Nature*, 476(7358):25–26, August 2011.
- [21] Oren Etzioni, Michele Banko, and Michael J. Cafarella. Machine reading. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1517–1519. AAAI Press, 2006.
- [22] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December 2008.
- [23] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10. AAAI Press, 2011.
- [24] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [25] Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pages 10–18, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [26] Alexander Gelbukh, Grigori Sidorov, and Adolfo Guzmán-Arenas. A method of describing document contents through topic selection. In *Proceedings of*

- SPIRE'99, International Symposium on String Processing and Information Retrieval, Cancun, Mexico, September 22–24*, pages 73–80. IEEE Computer Society Press, 1999.
- [27] Sofía N. Galicia Haro. *Computational Treatment of Temporal Expressions And Named Entities*. Colección de Libros CIDETEC. Centro de Innovación y Desarrollo Tecnológico en Cómputo, I.P.N., 2013.
- [28] Christopher Horn, Alisa Zhila, Alexander Gelbukh, Roman Kern, and Elisabeth Lex. Using factual density to measure informativeness of web documents. In *Proceedings of the 19th Nordic Conference on Computational Linguistics, NoDaLiDa*, 2013.
- [29] P. S. Jacobs and Lisa F. Rau. Scisor: Extracting information from on-line news. *Commun. ACM*, 33(11):88–97, November 1990.
- [30] Marshall Kirkpatrick. New 5 billion page web index with page rank now available for free from common crawl foundation. http://readwrite.com/2011/11/07/common_crawl_foundation_announces_5_billion_page_w, November 2011. [last visited on 25/01/2013].
- [31] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining, WSDM '10*, pages 441–450, New York, NY, USA, 2010. ACM.
- [32] R. J. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159 – 174, 1977.
- [33] Geoffrey Leech and Andrew Wilson. Standards for tagsets. In *Syntactic Word-class Tagging*, pages 55–80. Springer Netherlands, 1999.
- [34] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, December 2004.
- [35] Elisabeth Lex, Andreas Juffinger, and Michael Granitzer. Objectivity classification in online media. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia, HT '10*, pages 293–294, New York, NY, USA, 2010. ACM.

- [36] Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. Measuring the quality of web content using factual information. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '12, pages 7–10, New York, NY, USA, 2012. ACM.
- [37] Nedim Lipka and Benno Stein. Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1147–1148, New York, NY, USA, 2010. ACM.
- [38] Jordan J. Louviere and G.G. Woodworth. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta, 1991.
- [39] Ji Ma, Tong Xiao, Jing Bo Zhu, and Fei Liang Ren. Easy-first chinese pos tagging and dependency parsing. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 1731–1746. Indian Institute of Technology Bombay, 2012.
- [40] Alain-Pierre Manine, Érick Alphonse, and Philippe Bessières. Information extraction as an ontology population task and its application to genic interactions. In *ICTAI (2)*, pages 74–81. IEEE Computer Society, 2008.
- [41] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534. ACL, 2012.
- [42] Shuyo Nakatani. Language detection library for java. <http://code.google.com/p/language-detection/>, 2011. [last visited on 25/01/2013].
- [43] Natasha Noy and Alan Rector. Defining N-ary Relations on the Semantic Web. Technical report, W3C Working Group, 2006.
- [44] Lluís Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. Semantic services in freeling 2.1: Wordnet and ukb. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February 2010. Global Wordnet Conference 2010, Narosa Publishing House.
- [45] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources*

- and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
- [46] Partha Pakray, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh. Semantic textual entailment recognition using UNL. *Polibits*, 43:23–27, 2011.
- [47] Wenzhe Pei, Tao Ge, and Baobao Chang. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [48] Daniele Pighin, Marco Cornolti, Enrique Alfonseca, and Katja Filippova. Modelling events through memory-based, open-ie patterns for abstractive summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 892–901, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [49] Jakub Piskorski and Roman Yangarber. Information Extraction: Past, Present and Future. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization. Volume in the Series: Theory and Applications of Natural Language Processing*. Springer-Verlag, Berlin & New York, 2013.
- [50] Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. Dependency-based semantic parsing for concept-level text analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2014, Part I*, volume 8403 of *Lecture Notes in Computer Science*, pages 113–127. Springer, 2014.
- [51] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Emosenticspace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123, October 2014.
- [52] Steffen Schnitzer, Sebastian Schmidt, Christoph Rensing, and Bettina Harriehausen-Mühlbauer. Combining active and ensemble learning for efficient classification of web documents. *Polibits*, 49:39–45, 2014.
- [53] Grigori Sidorov. Syntactic dependency based n-grams in rule based automatic english as second language grammar correction. *International Journal of Computational Linguistics and Applications*, 4(2):169–188, 2013.

- [54] Grigori Sidorov. Should syntactic n-grams contain names of syntactic relations? *International Journal of Computational Linguistics and Applications*, 5(1):139–158, 2014.
- [55] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014.
- [56] Stephen Soderland, Brendan Roof, Bo Qin, Shi Xu, Mausam, and Oren Etzioni. Adapting open information extraction to domain-specific relations. *AI Magazine*, 31(3):93–102, 2010.
- [57] Michael M. Stark and Richard F. Riesenfeld. Wordnet: An electronic lexical database. In *Proceedings of 11th Eurographics Workshop on Rendering*, 1998.
- [58] Weiwei Sun, Yantao Du, Xin Kou, Shuoyang Ding, and Xiaojun Wan. Grammatical relations in chinese: Gb-ground extraction and data-driven parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 446–456, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [59] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7, ConLL '00*, pages 127–132, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
- [60] Francisco Viveros-Jiménez. *Word sense disambiguation through associative dictionaries*. PhD thesis, Instituto Politécnico Nacional, June 2014.
- [61] Historical trends in the usage of content languages for websites. http://w3techs.com/technologies/history_overview/content_language, Nov 2014. [last visited on 04/11/2014].
- [62] Nicolas Weber, Karin Schoefegger, Jenny Bimrose, Tobias Ley, Stefanie Lindstaedt, Alan Brown, and Sally-Anne Barnes. Knowledge maturing in the semantic mediawiki: A design study in career guidance. In *Proceedings of the 4th European Conference on Technology Enhanced Learning: Learning in the Synergy of Multiple Disciplines*, EC-TEL '09, pages 700–705, Berlin, Heidelberg, 2009. Springer-Verlag.

- [63] Languages used on the internet. Webpage: http://en.wikipedia.org/wiki/Languages_used_on_the_Internet#cite_note-NIUBL-IWS-6, November 2014. [last visited on 04/11/2014].
- [64] Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [65] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492, May 2012.
- [66] Weiqun Xu, Jean Carletta, and Johanna D. Moore. Syntactic chunking across different corpora. In Steve Renals, Samy Bengio, and Jonathan G. Fiscus, editors, *MLMI*, volume 4299 of *Lecture Notes in Computer Science*, pages 166–177. Springer, 2006.
- [67] Frances Yung. Towards a discourse relation-aware approach for chinese-english machine translation. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 18–25, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.
- [68] Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. Character-level chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1326–1336, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [69] Alisa Zhila and Alexander Gelbukh. Automatic identification of facts in real Internet texts in Spanish using lightweight syntactic constraints: Problems, their causes, and ways for improvement. *Revista Signos. Estudios de Lingüística*, 2015.

Appendix A

Running ExtrHech

We have implemented our method for Open IE for Spanish in a system called EXTRHECH. The system comprises a number of interpretable source code files mostly written in Python apart from encoding conversion stage. The files are available for download¹. Here we briefly present the processing pipeline in Listing A.1.

Listing A.1: EXTRHECH processing pipeline

```
cd \path\to\ExtrHech\src

@ chaning encoding
perl .\encoding\utf8-to-ISO-8859-1.pl test_file.txt > test_file_in_ISO.txt

@ POS-tagging with Freeling
\path\to\Freeling-2.2\bin\analyzer.exe -f \path\to\Freeling-2.2\bin\es.cfg --
    outf "tagged" < test_file_in_ISO.txt > ..\intermediate_outputs\POS-tagged\
    test_file.pos0

@ changing format
python tagged_back2line_whole_file.py ..\intermediate_outputs\POS-tagged\
    test_file.pos0 > ..\intermediate_outputs\POS-tagged\test_file.pos

@ extraction
python fact_extr_regexp4.py facts_extr.config \intermediate_outputs\POS-tagged\
    test_file.pos > test_extractions.extr
```

¹https://bitbucket.org/alisa_ipn/extrhech.git

Appendix B

Regular Expressions for Open Information Extraction

In this appendix we provide the regular expressions that implements the rules from Section 4.1 that underlie our algorithm. All expressions use EAGLES POS-tag set. Please, refer to [33] for decoding into particular parts-of-speech.

Expressions, that implement the pattern

$$\text{VREL} \rightarrow (\text{VW} * \text{P}) | (\text{V}) \quad (\text{B.1})$$

are shown in listing B.1.

Listing B.1: Regular expressions for verb relation detection

```
## Verb with dependent adverbs
V = r'(?:\w+\^\w+\^R[GN]\s+)?(?:\w+\^\w+\^P[0|P].[0|C][0|P|S]000\s+)?(?:?:\w+\^\w+\^V[M|S]I....)|(?:?:\w+\^\w+\^V[A|S]I....\s+\w+\^\w+\^VM[P|G]....))(?:\s+\w+\^\w+\^RG)?'

## A noun, an adjective, an adverb, a pronoun, or an article
W = r'(?:(?:\s+\w+\^\w+\^N....)|(?:\s+\w+\^\w+\^A....)|(?:\s+\w+\^\w+\^R.)|(?:\s+\w+\^\w+\^P....)|(?:\s+\w+\^\w+\^D....)|(?:\s+\w+\^\w+\^VMN....(?:\s+\w+\^\w+\^PP...000)?))'

## A preposition optionally immediately followed by an infinitive or a gerund
P = r'(?:(?:\s+\w+\^\w+\^SP...\s+\w+\^\w+\^V.N....(?:?:'+COORD+')'+I+'))|(?:\s+\w+\^\w+\^SP....)|(?:\s+\w+\^\w+\^V.[N|G]....))'

##Verb relation phrase
VREL = '('+V+ W+'*'+P+')|('+V+')'
```

Other formal words, i.e., COORD and I, are explained in listing B.3.

Expressions for noun phrase detection as in

$$\text{NP} \rightarrow \text{N}(\text{PREPN})? \quad (\text{B.2})$$

are provided in listing B.2.

Listing B.2: Regular expressions for noun phrase detection

```
## Numeric
NUM = r'(?:\w+\^\d+\^Z\s+)'

## Centuries
SIG = r'(?:[\w\^]+\^\[. {3,25}\]\^\w+)'

## Proper names
NPROP = r'(?:\s+\w+\^\w+\^NP00000)'

## Currency
USD = r'(?:\w+\^\$_USD\:\d+\^Zm)'

## Indefinite pronouns 'Uno de'
UNOde = r'(?:\w+\^\w+\^PIO..000\s+[D|d][E|e]\^\de\^SPS00\s+)'

## Adverb including 'no'
ADV = r'(?:\s+\w+\^\w+\^R.)'

## Noun with dependent words
N = UNOde+'?(\w+\^\w+\^D[TE]...\s+)?((?:(\w+\^\w+\^A...\s+)|(?:(\w+\^\w+\^
PRO...\s+)))'+NUM+'?(\w+\^\w+\^N.....'+NPROP+'?|'+NUM+'|'+SIG+'|'+
USD+'?)(?:(?:(\w+\^\w+\^A.....)?(\w+\^\w+\^VMP.....)
?(\w+\^\w+\^RG)?\w+\^\w+\^A.....)*'

## Preposition possibly followed by a noun with dependent words
PREP = r'(?:\s+\w+\^\w+\^SP...\s+'N+')'

## Participle clause
PARTICIP = r'(?:\s+\w+\^\w+\^VMP...\s+\w+\^\w+\^SP...\s+'N+PREP+'?)'

## Argument phrase
NP = N+'?'+PREP+'|'+PARTICIP+'?'
```

Additionally, an infinitive pattern used in verb relation detection and expressions for coordinative conjunction and relative pronoun as in

$$\text{COORD} \rightarrow \text{Y}|\text{COMMA}? \quad (\text{B.3})$$

and

$$\text{QUE} \rightarrow \text{PR} \quad (\text{B.4})$$

are implemented as shown in listing B.3.

Listing B.3: Regular expressions for other structures

```
## Infinitive
I = r'(?:\s+\w+\^\w+\^V.N....(?:\s+\w+\^\w+\^PP...000)?)'
```

Relative pronoun

```
QUE = r'(?:\s+\w+\^\w+\^PROC....)'
```

Coordinating conjunctions

```
Y = r'(?:\s+\w+\^\w+\^CC)'
```

```
COMMA = r'(?:\s+,^\^Fc)'
```

```
COORD = Y+'|(?:'+COMMA+Y+'?).'
```