

COMPARISON OF OPEN INFORMATION EXTRACTION FOR SPANISH AND ENGLISH

ALISA ZHILA

ALEXANDER GELBUKH

CENTER FOR COMPUTING RESEARCH, INSTITUTO POLITÉCNICO NACIONAL, MEXICO

Dialogue-2013

OUTLINE

Introduction

- Open Information Extraction (Open IE)
- Applications of Open IE
- Approaches to Open IE
- Problem

Open IE for Spanish

Experiments & Results

Error Analysis

Conclusions and Future Work

TRADITIONAL IE

- **Find all, say, acquisitions: *who bought what***
- **Target relations are predefined:**
 - Relations: *acquisition(arg1, arg2, ..., argN)*
 - args: *people, companies, currency...*
- **Hand-labeled **lexicalized** training examples**
- **Lots of training data**
- **Tuned linguistic technologies (NER, parsing, ...)**
- **Extensive human involvement**

Used in: Domain-specific information extraction from relatively small homogeneous corpora

WHAT IS OPEN IE? 1/2

Introduced by Banko *et al.* in 2007

Arbitrary relations, not predefined:

Born in, comes from, makes a deal with, ...

Extracted tuples are called “assertions”:

<Argument1, Relation, Argument2>

McCain fought hard against Obama, but finally lost the election

- *<McCain, fought against, Obama>*
- *<McCain, lost, the election>*

WHAT IS OPEN IE? 2/2

Unlexicalized, domain-independent:

looks only at POS/syntactic structure

No need in extensive hand-labeled training dataset:

uses heuristics or distant supervision

Fast and scalable to the Web:

appropriate for a large heterogeneous corpus

Can serve even undefined user needs:

users can interactively refine their need

APPLICATIONS OF OPEN IE

Different from traditional IE!

- **Common-sense knowledge collection**
- **New perspectives in QA systems**
- **New approach to IR [Etzioni, 2011]**
- **Machine Reading: automatic, unsupervised understanding of text [Etzioni et al., 2006]**
- **Web text quality automatic assessment [Horn & Zhila et al., 2013 @ NoDaLiDa]**

APPROACHES TO OPEN IE

1. ML-based

TextRunner (Banko, 2007), WOE^{pos} & WOE^{parse} (Wu & Weld, 2010)

Shortcomings: Extracts incoherent relations

“The Mark 14 was central to the torpedo scandal of the fleet.”

~~←was central torpedo→~~

2. Syntactic and context analysis

OLLIE (Mausam, 2012), FES (Aguilar, 2012)

Shortcomings: slow, computational resource demanding

3. POS analysis and syntactic constraints

ReVerb (Fader et al., 2011)

Shortcomings: only verb-based relations

Advantages: fast, easy to implement, accurate, efficient

PROBLEM

- **Requires language-specific information**
e.g. Typical POS sequence in a relation
- **Was implemented for English only**
“simple canonical ways in which verbs express relationships **in English**” [Etzioni et al., 2011]

3. POS analysis and syntactic constraints
What are peculiarities of application of this method to another language?
ReVerb (Fader et al., 2011)

WHY IS IT IMPORTANT?

- **Different morphology** (different POS-tagging)
- **Different grammar** (i.e. word order)
- **In general:**
 - **Languages are different**
 - **No work on languages other than English**
 - **We cannot expect the same behavior**

OUTLINE

Introduction

Open IE for Spanish

- Architecture of ExtrHech system

Experiments & Results

Error Analysis

Conclusions

ARCHITECTURE OF EXTRHECH OPEN IE SYSTEM FOR SPANISH 1/2



EAGLES POS-tag set for Spanish from Freeling-2.2

Syntactic constraints as regular expressions

1. "Relation phrase"-first approach: looks for **verb phrase**
$$\text{VREL} \rightarrow (\text{V W}^*\text{P}) | (\text{V})$$
2. Looks for **noun phrases** to the left and right
$$\text{NP} \rightarrow \text{N} (\text{PREP N})?$$
3. Rules for
 - Coordinating conjunctions
 - Relative clauses
 - Participles

ARCHITECTURE OF EXTRHECH OPEN IE SYSTEM FOR SPANISH 2/2

- **Does not resolve zero subject**

“Cerró la entrada.”

(“[He] closed the entrance.”)

- **Does not resolve inverse word order**

Object/Indirect Object – Verb – Subject

“De la médula espinal nacen los nervios periféricos”

(“Out of the spinal cord come peripheral nerves”)

OUTLINE

Introduction

Open IE for Spanish

Experiments & Results

- For different Spanish datasets
- For parallel English-Spanish dataset
- Performance comparison

Error Analysis

Conclusions

EXPERIMENT OVER TWO SPANISH DATASETS 1/2

FACT-SPA-CIC

- **68 sentences in Spanish**
- **Manually selected from school textbooks**
- **Grammatically and orthographically correct**

RAW WEB TEXT

- **159 sentences**
- **randomly extracted from Web (with language detection filter)**
- **36 sentences (22%) either grammatically incorrect or incoherent**

“cronista cumple del diego video diego el 10”

(“journalist accomplishes of the [D]iego video [D]iego 10 [points]”)

PERFORMANCE FOR SPANISH DATASETS 2/2

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

$$\mathbf{Precision} = \frac{\textit{correct assertions}}{\textit{all extracted assertions}}$$

$$\mathbf{Recall} = \frac{\textit{correct assertions}}{\textit{all possible assertions}}$$

- **correct assertions** as evaluated by two human annotators
- **all possible (correct) assertions** = all expected extractions + assertions returned by the system that both annotators considered correct

EXPERIMENT OVER PARALLEL ENGLISH-SPANISH DATASET

**Gramatically correct dataset FactSpaCIC of 68 sentences
was translated into English**

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

- ReVerb turned out to be less robust:
More unattempted sentences

COMPARISON OF PERFORMANCE FOR VARIOUS OPEN IE SYSTEMS

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

OUTLINE

Introduction

Open IE for Spanish

Experiments & Results

Error Analysis

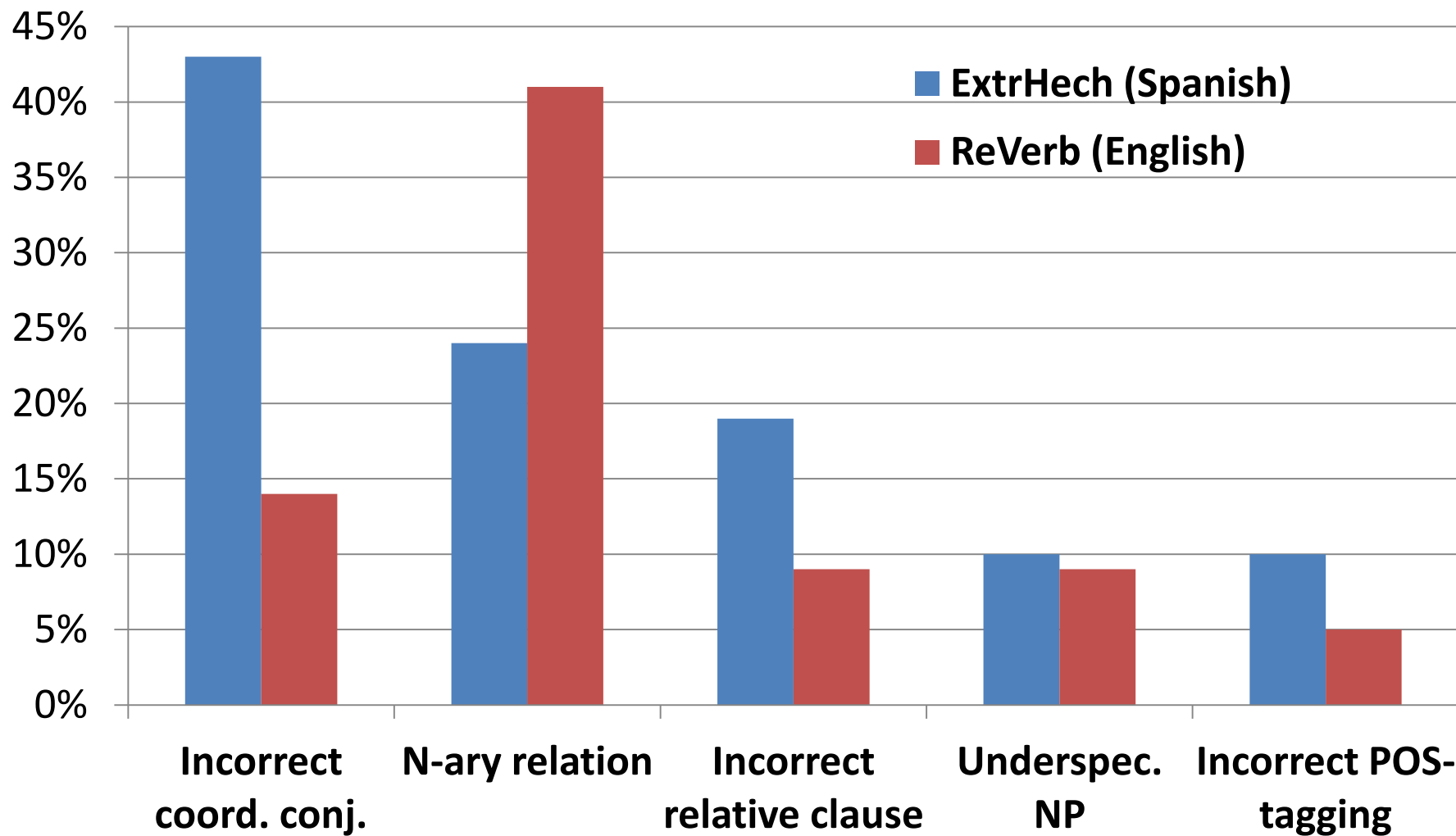
Conclusions

ERROR ANALYSIS

Performed:

- **For Spanish language system ExtrHech:**
over FactSpaCIC (68 sent., grammatically correct) and Raw Web (159 sent.) datasets
- **For English language system ReVerb:**
over the English translation of FactSpaCIC (68 sent., gram. correct)

CAUSES OF ERRORS FOR BOTH SYSTEMS 1/3



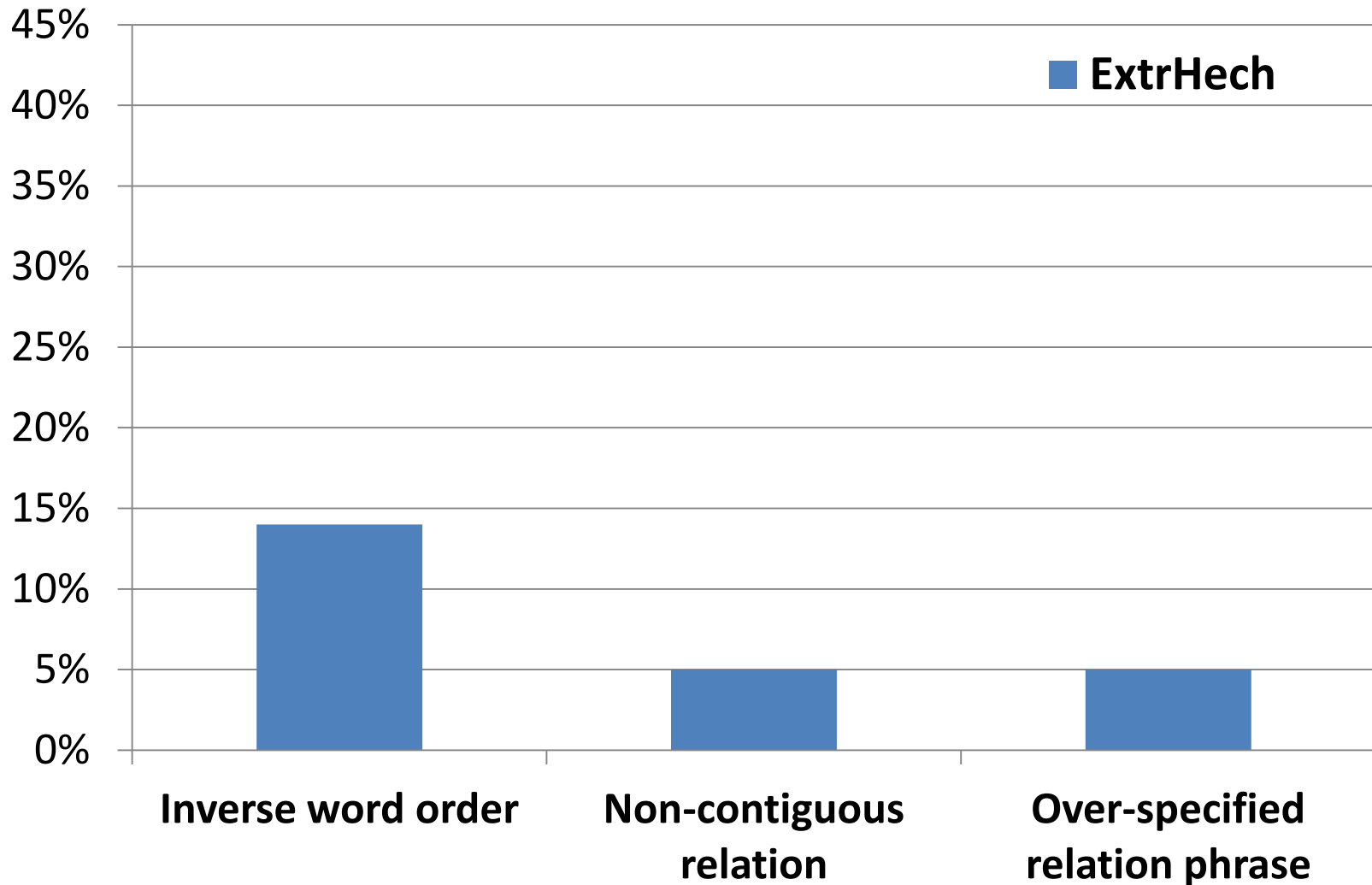
CAUSES OF ERRORS FOR BOTH SYSTEMS 2/3

Cause	ExtrHech	ReVerb	Example
Incorrect coordinative conjunction resolution	43%	14%	<p><u>The hypothalamus</u> is responsible for certain body functions such as temperature control and <u>receives the signal of</u> sleep, hunger and thirst</p> <p><certains body functions; receives the signal of; sleep , hunger and thirst></p>
N-ary relation	24%	41%	<p>...<u>crevices and folds</u> that <u>give it the appearance of a peeled walnut</u></p> <p><crevices and folds; give; it></p>

CAUSES OF ERRORS FOR BOTH SYSTEMS 3/3

Cause	ExtrHech	ReVerb	Example
Incorrect relative clause resolution	19%	9%	El lugar en el que florecieron las culturas <El lugar; florecieron; las culturas>
Under-specified noun phrase	10%	9%	<u>The data from the consulted sources</u> must be registered in index cards. <Arg1=the consulted sources>
Incorrect POS-tagging	10%	5%	Archaeology uses new techniques to ... study <u>the material remains and tracks and signs</u> that man made in the past <the material; signs ^V ; that ^{PN} man>

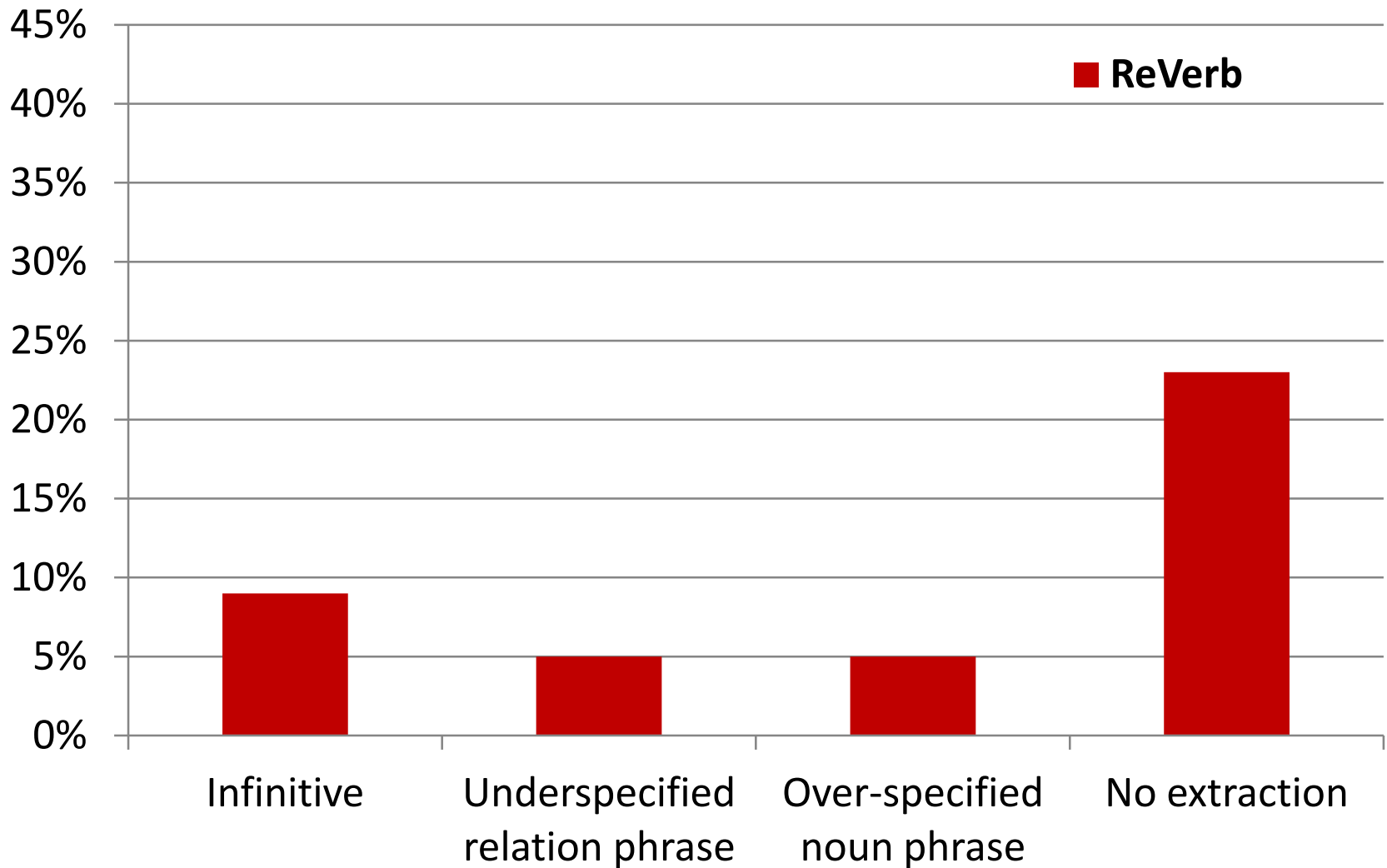
CAUSES OF ERRORS FOR SPANISH SYSTEM 1/2



CAUSES OF ERRORS FOR SPANISH SYSTEM 2/2

Cause	Extr Hech	ReVerb	Example	Intuition
Inverse word order	14%	—	De la médula espinal nacen los nervios periféricos. <la médula espinal; nacen; los nervios periféricos>	Sp
Non-contiguous relation	5%	—	<u>bajo</u> cuyo <u>nombre</u> <u>pueden entrar</u> los sextantes <nombre; pueden entrar; los sextantes>	Sp
Over-specified relation phrase	5%	—	La Botánica ha logrado analizar las características de la vegetación <Rel = ha logrado analizar las características de>	sys

CAUSES OF ERRORS FOR ENGLISH SYSTEM 1/2



CAUSES OF ERRORS FOR ENGLISH SYSTEM 2/2

Cause	Extr Hech	ReVerb	Example	Intuition
Infinitive	–	9%	such as <u>to interpret</u> what the eyes see, <u>think</u> , and <u>control</u> many of the body's movements <the eyes; control many of; the body 's movements>	Eng
Under-specified relation phrase	–	5%	<u>a peaceful nation of navigators</u> who was in contact with <u>Egypt</u> <a peaceful nation of navigators; was in; contact>	sys
Over-specified noun phrase	–	5%	The mammoths migrated from <u>Africa</u> 3.5 million years ago <Arg2 = Africa 3.5 million years>	sys/Eng
No extraction	–	23%	–	sys

OUTLINE

Introduction

Open IE for Spanish

Experiments & Results

Error Analysis

Conclusions & Future Work

CONCLUSIONS

- **Open IE based on POS-tagged input & syntactic constraints adapted to Spanish**
- **First cross-lingual comparative study of Open IE**
- **Performance for Spanish is comparable to English**
 - for system based on the same approach
- **Detailed analysis of errors:**
 - POS-tagging accuracy of 95+% is sufficient for this task
 - Inverse word order is not the biggest problem
- **Good news for Russian (and other European languages): the approach should work as well**

FUTURE WORK

- **Run the system over a large corpus**
- **Most frequent assertions will be considered “facts”**
- **Cluster relation phrases and arguments**
- **Map relations to some ontology**

THANK YOU! QUESTIONS?

APPENDIX

DIFFERENCES IN IMPLEMENTATION

- **Different POS-tag set :**

EAGLES vs Penn Tree

- **Different verb phrase treatment:**

- Reflexive verbs in Spanish: *Juan se lava la cara.*

- **Based on regular expressions**

- **Differences in implementation of coordinative conjunction resolution,**

Purely engineering details

REGEX EXAMPLES

Verb phrase:

VREL → **(V W*P)|(V)**

W can be 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)?))'`

PROBLEM

3. POS analysis and syntactic constraints

ReVerb (Fader et al., 2011)

- **Requires language-specific information**
e.g. Typical POS sequence in a relation
- **Was implemented for English only**
“simple canonical ways in which verbs express relationships **in English**” [Etzioni et al., 2011]

What are peculiarities of application of this method to another language?

APPROACHES TO OPEN IE 1/3

Learning based systems:

TextRunner (Banko, 2007), WOE^{pos} & WOE^{parse} (Wu & Weld, 2010)

- Automatically labeled sentences (using heuristics or distant-supervision)
- Learn relation phrase extractor
- Argument-first:
Detect arguments (Arg1, Arg2) and then identifies a relation

Shortcomings:

- Noisy training corpus
- Doesn't work well for long sentences
- Detects incoherent relations:
(Faust; made; a deal) instead of (Fauts; made a deal with; the devil)

APPROACHES TO OPEN IE 2/3

Syntactic-analysis based systems:

OLLIE(Mausam, 2012), FES(Aguilar, 2012)

- Deeper syntactic and context analysis
- Detects relations that are not expressed via a verb

Shortcomings:

- High computational capacity
- Slow

APPROACHES TO OPEN IE 3/3

POS analysis and syntactic constraints based systems:

ReVerb (Fader et al., 2011)

- Does not need labeled corpus
- POS-tagging and rules
- “Relation phrase”- first
- Fast in implementation and execution

Shortcomings:

- Detects only verb-based relations
- Works on a sentence-level