



# Open Information Extraction for Spanish Language based on Syntactic Constraints

Alisa Zhila, Alexander Gelbukh

Natural Language Processing Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Mexico



## Open Information Extraction

Huge variety of textual information on the Web:



### Problem: Need to process arbitrary information

- Arbitrary relations are numerous
- It is not possible to make an exhaustive list of all relations and their arguments
- Traditional Information Extraction (IE) methods require large training corpora and training for each relation and its arguments

### Solution: Open Information Extraction

- Introduced by Michele Banko et al. in 2007
- Extracts information based on specific syntactic patterns without requiring a pre-specified vocabulary or large manually tagged training corpora
- Relations are extracted in the form of tuples:  $\langle \text{Argument 1}; \text{Relational phrase}; \text{Argument 2} \rangle$

### Example:

"Man who drove van full of kids is charged with attempted murder"

Extractions:  $\langle \text{Man}; \text{drove}; \text{van full of kids} \rangle$

$\langle \text{Man}; \text{is charged with}; \text{attempted murder} \rangle$

- Open IE is performed using various approaches
- All approaches are language-dependent

In this work: **approach based on syntactic constraints**

### Features:

- Rule based
- Fast, scalable to the Web
- Easy implementation

### Limitations:

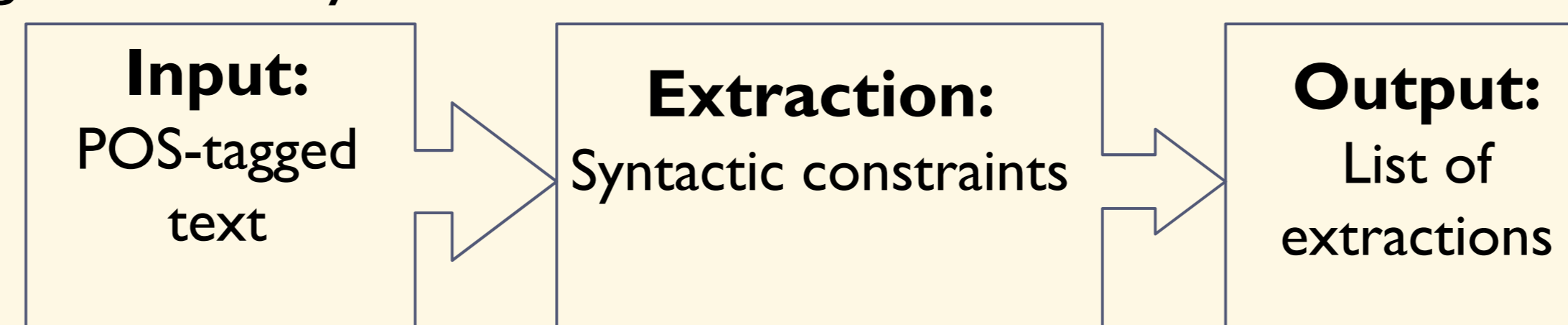
- Only verb-based relations
- Only 2 arguments

## Open IE Based on Syntactic Constraints

### Basic Algorithm (Fader et al. 2011)

Search for a verb-containing relation phrase and the nearest noun phrases to the left and to the right

Implemented in **ReVerb** and shown to work for English on grammatically correct texts.



## Open IE for Spanish

### I. Spanish vs. English

#### Similarities:

- Predominantly Subject-Verb-Object word order
- Analytic languages:
  - no grammatical cases for nouns;
  - verb-noun relations are conveyed by prepositions

#### Sample Differences in Spanish:

- Reflexive pronouns: *se realizaron* ("were carried out")
- infinitives are not preceded by "to"
- adjectives usually follow nouns
- oblique case pronouns precede verbs: *lo veo* / "I see it"

### II. Syntactic Rules for Spanish

**Verb Phrase:**  $VREL \rightarrow V [W^* P]$

V: non-infinitive verb optionally preceded by a reflexive pronoun or a participle

W: noun | adjective | adverb | preposition | article

P: preposition | infinitive | gerund

**Noun Phrase:**  $NP \rightarrow Np [PREP Np]$

Np: noun with or w/o article | adjective | number

PREP: preposition

Implemented in **ExtrHech** Open IE system for Spanish

## Experiments

### I. Spanish vs. English

**Dataset:** 300 parallel sentences from the English-Spanish part of News Commentary Corpus

**Settings:** **ReVerb** was run on the English part of the dataset **ExtrHech** was run on the Spanish part.

**Evaluation:** by human annotators

Inter-annotator agreement measured by Cohen's kappa

\* indicates substantial agreement between the annotators

System	Precision	Recall	Correct extractions	Total extractions	Cohen's kappa
ExtrHech	0.59	0.48	218	368	0.60*
ReVerb	0.56	0.44	201	358	0.68*

**Results:** stable performance for English and Spanish

### II. "Raw" Web Texts vs. News Articles

**Datasets:** (1) 159 unprocessed sentences randomly extracted from the "Raw" Web CommonCrawl 2012 corpus

(2) 300 sentences from News Commentary Corpus

**Settings:** **ExtrHech** run on both datasets to compare the performance for unedited and totally arbitrary texts from the Web vs. edited news articles

**Evaluation:** by human annotators

+ indicates the lower bound of moderate agreement

Dataset	Precision	Recall	Cohen's kappa
"Raw" Web	0.55	0.49	0.40+
News	0.59	0.48	0.60*

**Results:** almost as good for "raw" Web texts

## Discussion

- Promising fast and stable performance for other SVO word order languages with good POS-taggers
- Promising performance at Web scale: robust on raw Web texts