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
TRADITIONAL IE

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

\textbf{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 et al., 2013 @ NoDaLiDa]
APPROACHES TO OPEN IE

1. ML-based
TextRunner (Banko, 2007), WOE\textsubscript{pos} & WOE\textsubscript{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” [Etzoni et al., 2011]

3. POS analysis and syntactic constraints
What are peculiarities of application of this method to another language?
OUTLINE

Introduction

Open IE for Spanish
  • Architecture of ExtrHech system
  • Differences in implementation

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**
   \[
   \text{VREL} \rightarrow (V \ W^*P) \mid (V)
   \]
2. Looks for **noun phrases** to the left and right
   \[
   \text{NP} \rightarrow N \ (\text{PREP} \ N)\
   \]
3. Rules for
   - coordinating conjunctions
   - Relative clauses
   - Participles

**Input**
- POS-tagged text

**ExtrHech**
- Syntactic constraints

**Output**
- List of assertions
  \(<\text{Arg1}; \text{Rel}; \text{Arg2}>\)
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”)
DIFFERENCES IN IMPLEMENTATION

- Different POS-tag set:
  EAGLES vs Penn Tree

- Different verb phrase treatment:
  - Reflexive verbs in Spanish: *Juan se lava la cara*.
  - Assertion extracted from participle clauses:
    “... sus creencias relacionadas con la muerte.”
    (“...their beliefs related to death”)

  \(<\text{Arg1 = sus creencias; Rel = relacionadas con; Arg2 = la muerte}>\)

*Purely engineering details*
OUTLINE

Introduction
Open IE for Spanish

Experiments & Results
  • For different Spanish datasets
  • For parallel English-Spanish Datasets
  • 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]”)

14
PERFORMANCE FOR SPANISH DATASETS 2/2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FactSpCIC</td>
<td>87%</td>
<td>70%</td>
</tr>
<tr>
<td>(grammatically correct)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw Web text</td>
<td>55%</td>
<td>49%</td>
</tr>
<tr>
<td>(noisy)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Precision** = \( \frac{\text{correct assertions}}{\text{all extracted assertions}} \)  

**Recall** = \( \frac{\text{correct assertions}}{\text{all possible assertions}} \)

- **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 DATASET

Gramatically correct dataset FactSpaCIC of 68 sentences was translated into English

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>correct extractions</th>
<th>found extractions</th>
<th>expected extractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExtrHech</td>
<td>87%</td>
<td>70%</td>
<td>99.5</td>
<td>115</td>
<td>137</td>
</tr>
<tr>
<td>(Spanish)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReVerb</td>
<td>76%</td>
<td>50%</td>
<td>71</td>
<td>93</td>
<td>139</td>
</tr>
<tr>
<td>(English)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• ReVerb turned out to be less robust:
  More unattempted sentences
REMINDER:
APPROACHES TO OPEN IE

1. Learning-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), ExtrHech (Zhila & Gelbukh, 2013)
   - **Shortcomings:** only verb–based relations
   **Advantages:** fast, easy to implement, accurate, efficient
## COMPARISON OF PERFORMANCE FOR VARIOUS OPEN IE SYSTEMS

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

• For English language system ReVerb:
  over the English translation of FactSpaCIC (68, gram. correct)
### CAUSES OF ERRORS FOR BOTH SYSTEMS 1/3

<table>
<thead>
<tr>
<th>Cause</th>
<th>ExtrHech ReVerb</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect coordinative conjunction resolution</td>
<td>43% 14%</td>
<td>The hypothalamus is responsible for certain body functions such as temperature control and receives the signal of sleep, hunger and thirst</td>
</tr>
<tr>
<td>N-ary relation</td>
<td>24% 41%</td>
<td>...crevices and folds that give it the appearance of a peeled walnut</td>
</tr>
</tbody>
</table>
# Causes of Errors for Both Systems 2/3

<table>
<thead>
<tr>
<th>Cause</th>
<th>ExtrHech %</th>
<th>ReVerb %</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect relative clause resolution</td>
<td>19%</td>
<td>9%</td>
<td>El lugar en el que florecieron las culturas</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;El lugar; florecieron; las culturas&gt;</td>
</tr>
<tr>
<td>Under-specified noun phrase</td>
<td>10%</td>
<td>9%</td>
<td>The data from the consulted sources must be registered in index cards.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;Arg1=the consulted sources&gt;</td>
</tr>
<tr>
<td>Incorrect POS-tagging</td>
<td>10%</td>
<td>5%</td>
<td>Archaeology uses new techniques to study the material remains and tracks and signs that man made in the past</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;the material; signs^V; that^PN man&gt;</td>
</tr>
</tbody>
</table>
CAUSES OF ERRORS FOR BOTH SYSTEMS 3/3

- Incorrect coord. conj.
- N-ary relation
- Incorrect relative clause
- Underspec. NP
- Incorrect POS-tagging
<table>
<thead>
<tr>
<th>Cause</th>
<th>Extr Hech</th>
<th>ReVerb</th>
<th>Example</th>
<th>Intuition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse word order</td>
<td>14%</td>
<td>–</td>
<td>De la médula espinal nacen los nervios periféricos.</td>
<td>Sp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;la médula espinal; nacen; los nervios periféricos&gt;</td>
<td></td>
</tr>
<tr>
<td>Non-contiguous relation</td>
<td>5%</td>
<td>–</td>
<td>bajo cuyo nombre pueden entrar los sextantes</td>
<td>Sp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;nombre; pueden entrar; los sextantes&gt;</td>
<td></td>
</tr>
<tr>
<td>Over-specified relation phrase</td>
<td>5%</td>
<td>–</td>
<td>La Botánica ha logrado analizar las características de la vegetación</td>
<td>sys</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;Rel = ha logrado analizar las características de&gt;</td>
<td></td>
</tr>
</tbody>
</table>
CAUSES OF ERRORS FOR **SPANISH SYSTEM 2/2**

- **Inverse word order**
- **Non-contiguous relation**
- **Over-specified relation phrase**

ExtrHech
## Causes of Errors for English System 1/2

<table>
<thead>
<tr>
<th>Cause</th>
<th>Extr Hech</th>
<th>ReVerb</th>
<th>Example</th>
<th>Intuition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinitive</td>
<td>–</td>
<td>9%</td>
<td>such as to interpret what the eyes see, think, and control many of the body's movements</td>
<td>Eng</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;the eyes; control many of; the body's movements&gt;</td>
<td></td>
</tr>
<tr>
<td>Under-specified relation</td>
<td>–</td>
<td>5%</td>
<td>a peaceful nation of navigators who was in contact with Egypt</td>
<td>sys</td>
</tr>
<tr>
<td>phrase</td>
<td></td>
<td></td>
<td>&lt;a peaceful nation of navigators; was in; contact&gt;</td>
<td></td>
</tr>
<tr>
<td>Over-specified noun phrase</td>
<td>–</td>
<td>5%</td>
<td>The mammoths migrated from Africa 3.5 million years ago</td>
<td>sys/Eng</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;Arg2 = Africa 3.5 million years&gt;</td>
<td></td>
</tr>
<tr>
<td>No extraction</td>
<td>–</td>
<td>23%</td>
<td>—</td>
<td>sys</td>
</tr>
</tbody>
</table>
CAUSES OF ERRORS FOR ENGLISH SYSTEM 2/2

- Infinitive
- Underspecified relation phrase
- Over-specified noun phrase
- No extraction

ReVerb
OUTLINE

Introduction
Open IE for Spanish
Experiments & Results
Error Analysis
Conclusions
CONCLUSIONS

- First cross-lingual comparative study of Open IE
- Open IE based on POS-tagged input & syntactic constraints
- We adapted this Open IE approach to Spanish
- Performance for Spanish is comparable with English
  - for system based on the same approach
  - though Spanish grammar is much more complicated
- Detailed analysis of errors
- POS-tagging accuracy of 95+% is sufficient for this task
- Language-specific errors exist and differ for each language
APPENDIX
REGEX EXAMPLES

Verb phrase:

\[ VREL \rightarrow (V \ W^*P) | (V) \]

\( W \) can be a noun, an adjective, an adverb, a pronoun, or an article

\[ W = \]
\[ r'(?:\s+\w+\w+\w+\w+\w+N.......) | (?:\s+\w+\w+\w+\w+\w+\w+A.....) | (?:\s+\w+\w+\w+\w+\w+\w+R.) | (?:\s+\w+\w+\w+\w+\w+\w+P.....) | (?:\s+\w+\w+\w+\w+\w+\w+\w+D.....) | (?:\s+\w+\w+\w+\w+\w+\w+\w+VMN....(?:\s+\w+\w+\w+\w+\w+\w+\w+PP...000)?)' \]
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 (Fausts; made a deal with; the devil)
APPROACHES TO OPEN IE 2/3

Syntactic-analysis based systems:


- 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