OPEN INFORMATION EXTRACTION FOR SPANISH AND ITS APPLICATION TO INFORMATIVENESS MEASURING FOR WEB DOCUMENTS

Alisa Zhila*, Christopher Horn°, Alexander Gelbukh*

***Natural Language Processing Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Mexico**
° Know-Center GmbH, Austria

Quality of Web documents

- Huge variety of texts on the Web:

News, encyclopedic articles, blogs, comments, computer-generated spam, documents created by copy and paste , etc.

- Decisions are made basing on the data obtained from the Internet
- Part of the data comes from sources of questionable reliability

Problem:

How to assess quality of texts on the Web?

- Assume that the purpose of a text document is to convey information
- *Informativeness*: the amount of useful information contained in a document
- **Suggestion:** informativeness is proportional to the number of "facts" in the document, or its *factual density*

Factual density *fd* of a text document *d*:

$$fd(d) = \frac{fc(d)}{size(d)} \ ,$$

where *fc(d)*: fact count in the document *d*,
size(d): size of the document in characters.

Now:

How to identify facts automatically?

- Open Information Extraction

Open Information Extraction

- **Open IE** is a task of extracting *facts* (*relational tuples*) from text without requiring a pre-specified vocabulary or manually tagged training corpora.
- **Fact** (relational tuple)
  (Argument 1; Relational phrase; Argument 2)

Example:

Text: *"Abraham Lincoln was the President of the United States"*
Fact:(*Abraham Lincoln; was the President of; the United States*)

ExtrHech: an Open IE system for Spanish

Input: Part-Of-Speech -tagged text → **Fact Extraction:** Syntactic constraints; Lexical constraints → **Output:** List of facts

- Syntactic constraints: via regular expressions
- Resolves coordinating conjunctions
- Correctly treats participle clauses
  "their beliefs related to the death" ⇒
  (their beliefs; are related to; the death)
- Filter out relative clauses
- Reflexive pronouns (specifically for Spanish)

Advantages:

- Fast
- Scalable to the Web
- Easy implementation
- Open vocabulary

Future work:

- Cope with free word order, via shallow parsing
- Treat co-references
- Cluster facts
- Extract discontinuous relational phrases

Evaluation

Ground Truth Ranking

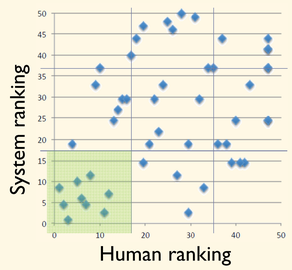
- 50 documents in Spanish randomly selected from the Internet
- 13 human annotators
- MaxDiff-type inquiry:
  *"Which document is the most and the least **informative**?"*

Thus obtained ground truth ranking for 50 documents

Automatic Ranking with ExtrHech

POS-tagging: Freeling-2.2 → **Fact Extraction:** ExtrHech → **Factual density** calculation → **Automatic Ranking**

Comparison



Correlation metric	Value	Significance Level
Ideal correlation	1	
Spearman's ρ	0.404	99.636%
Pearson's *r*	0.390	99.486%
Kendall's τ	0.293	99.653%
Random baseline	~0.018	

Correlation between rankings

- More correlation in the area of rankings 1-15, i.e. for more informative documents

Conclusions

- Statistically significant positive correlation between automatic and human annotator ranking
- Thus, Open IE is feasible for automatic assessment of quality of Web documents

Author Information: *Alisa Zhila (alisa.zhila@gmail.com) is a PhD student in the Center for Computing Research of Instituto Politécnico Nacional in Mexico.*
Christopher Horn (chorn@know-center.at) is a researcher at Know-Center GmbH, Austria
Prof. Alexander Gelbukh is Alisa Zhila's advisor.