A Pipeline for Multilingual Protest Event Selection & Annotation

Vera Danilova

Autonomous University of Barcelona, Spain
Russian Presidential Academy of National Economy and Public Administration, Russian Federation

September 1, 2015

Problem Definition

Social scientists are currently studying the benefits and drawbacks of the recently appeared open global event extraction systems/repositories, such as **GDELT** and **W-ICEWS**. Their main pitfalls for protest researchers are [Hanna, 2014]:

- non-coverage of important concepts by the existing ontologies (e.g., **CAMEO**): as a result, protest event types are not captured;
- limited event representation (event type, source actor, target actor, location): as a result, relevant protest event features like **Size, Claim, Violence use** and others are missing;
- dependency on machine translation quality.

Related Work

System	Wueest et al. (2013)	Hanna (2014)
Language Coverage	English	English
Training Set	”The Guardian”	”The New York Times”
Pre-processing toolkit	UIMA, HTM	–
Concept Hierarchy	–	DoCa
Postselection Algorithm	Active Learning	SVM (binary mode)
Coding Algorithm	NER & heuristics	SVM (multiclass)

Objective & Tasks

Enhance the quality of the analysis unit (protest scenario) by making it more informative
To this end:

- peform **event selection**: for a set of crawled multilingual news reports $H[h_1, h_2, ...h_n]$ identify a subset $H_p[h_{p1}, h_{p2}, ...h_{pn}]$ related to protest events;
- define and extract scenario slots: within H_p annotate protest-specific features: *Event_Type, Event_Location, Event_Reason, Event_Weight* (in this implementation).
- output scenarios in CSV.

Corpus

- a corpus of 14464 multilingual news lead sentences (from Bulgarian, French, Polish, Russian, Spanish, Swedish tabloids) has been collected with **Scrapy 1.0**;
- a subcorpus of 13710 lead sentences related to protest events has been formed using a **Python 2.7** script;

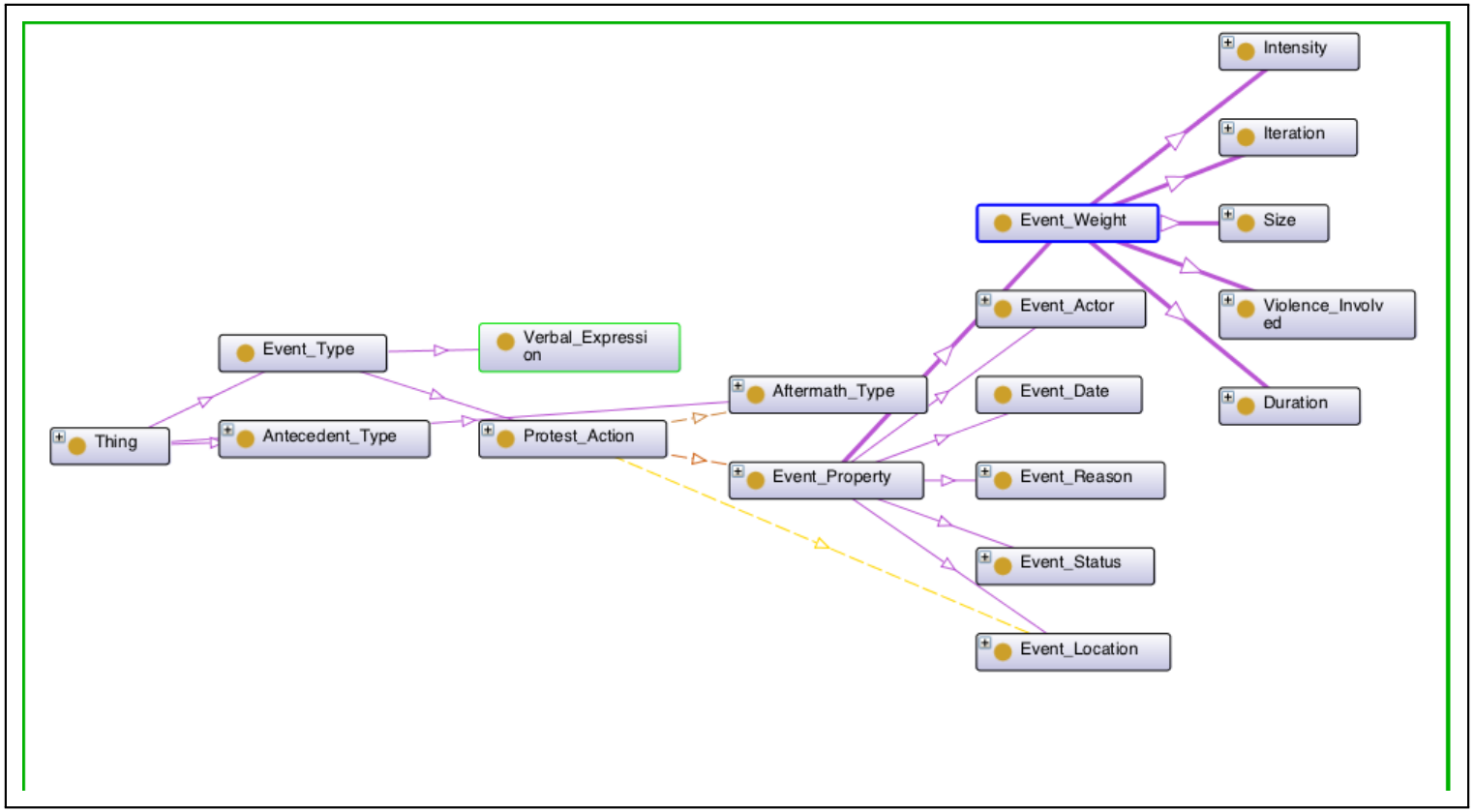
Why lead sentences?

- lead sentences (title and subtitle of the articles) have been selected as the processing unit, because they proved their *informativeness* in multiple studies, plus their processing is much faster;

Event Extraction Tools & Evaluation

- protest event descriptions have been analyzed, and domain-specific concepts (protest events hierarchy, subevents and properties) have been structured and formalized in **Protégé - 4.3**;
- Patterns and gazetteers have been built according to **GATE 8.0** (General Architecture for Text Engineering) framework standards;
- A pipeline that uses default and external plug-ins (*PoS Taggers: Treetagger & The Stockholm Stagger (external) for noun phrases parts annotation, Jape Plus Extended, Extended Gazetteer, BWP Gazetteer*, etc.) has been geared;
- The performance has been tuned on the development set and tested on both development and test sets.

A Fragment of the Concept Hierarchy



JAPE Pattern/Rule Pairs

- JAPE - Java Based Annotations Patterns Engine
- JAPE patterns are regular expressions over typed feature structures
- The left-hand side (LHS) of a rule describes pattern constraints and the right-hand side (RHS) - annotation commands
- Currently, we use Jape Plus Extended plugin that allows the use of additional constraints

A Fragment of a Rule for ER Annotation

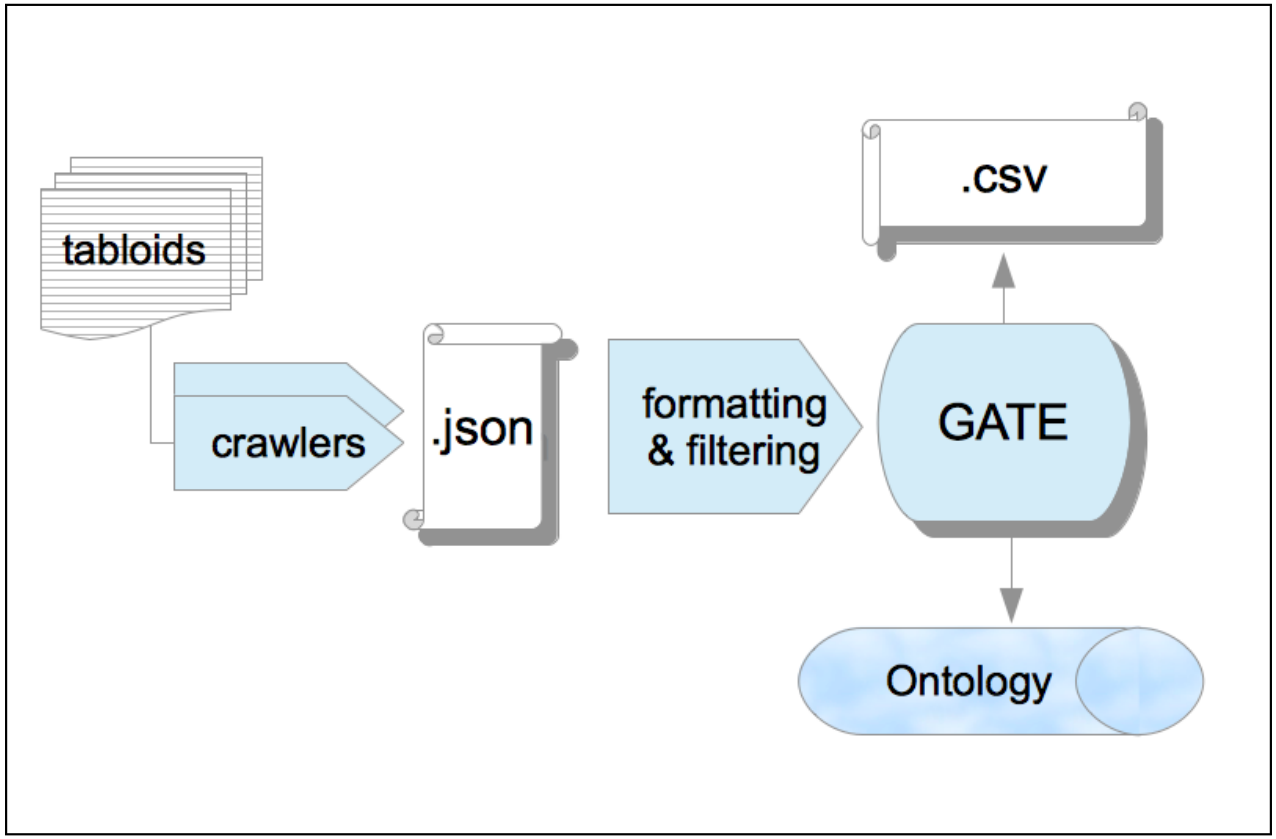
```
Rule: OntoLookup

({Event_Type})
({Token})[0,20]
({Lookup.majorType == Reason}):position
(CONTENT):issue
-->
:position{
 AnnotationSet pAS =
(AnnotationSet) bindings.get("position");
 AnnotationSet LookupAS = inputAS.get
("Lookup", pAS.firstNode().getOffset(),
pAS.lastNode().getOffset());
HashSet fNames = new HashSet();
    fNames.add("class");
    AnnotationSet ontoLookup =
LookupAS.get("Lookup", fNames);
...
```

Feature Selection

Out of the whole feature set, the following are selected:

- *Event_Type*: the *what* of the event, e.g.: rally, march, boycott, strike, picketing, hunger strike, riot, symbolic act, etc..
- *Event_Location*: the *where* of the event, e.g.: names of countries, cities and physical settings (**DBpedia** lookup).
- *Event_Reason*: the *why* of the event or **position** of a protesting group towards an **issue**: **for** *a cause (expressions of support incl. commemorations and demands) or* **against** *a cause*.
- *Event_Weight*: an attribute that defines the importance of an event and takes into account the values of the following slots: *Event_Duration, Event_Intensity, Event_Iteration, Event_Size, Violence_Use*.

The Resulting Workflow



Crawling, Filtering, Stopwords Removal

Language	Before Filtering	Total Duplicates	Stopwords
Bulgarian	4113 (528 kb)	1308	1306
French	8286 (615 kb)	1468	1242
Polish	5820 (591 kb)	1644	1561
Russian	7686 (673 kb)	4656	4654
Spanish	8678 (756 kb)	4683	4252
Swedish	3580 (180 kb)	705	695

Proof Checking

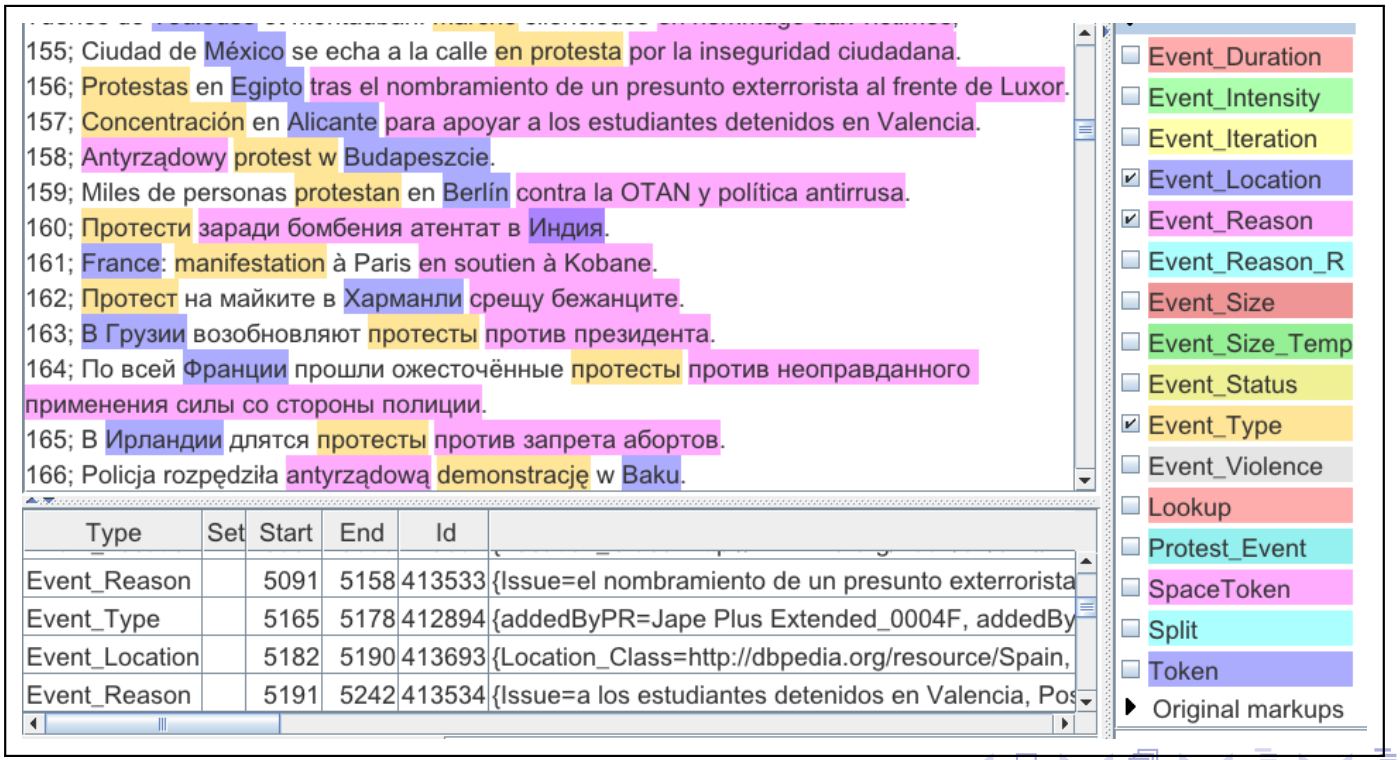
Language	Checked	True Negatives	True Positives
Bulgarian	700	8	99 %
French	700	159	78 %
Polish	700	144	84 %
Russian	700	55	92 %
Spanish	700	93	87 %
Swedish	700	22	97 %

A Fragment of a Rule for ER Annotation

```
Rule: OntoLookup

({Event_Type})
({Token})[0,20]
({Lookup.majorType == Reason}):position
(CONTENT):issue
-->
:position{
 AnnotationSet pAS =
(AnnotationSet) bindings.get("position");
 AnnotationSet LookupAS = inputAS.get
("Lookup", pAS.firstNode().getOffset(),
pAS.lastNode().getOffset());
HashSet fNames = new HashSet();
    fNames.add("class");
    AnnotationSet ontoLookup =
LookupAS.get("Lookup", fNames);
...
```

Annotation Sample

- A sample annotation of the multilingual devset (*Event_Type*, *Event_Reason*, *Event_Location*) in GATE 8.0 GUI



Metrics

$Precision = \frac{|G \bigcap C|}{|G|}$, $Recall = \frac{|G \bigcap C|}{|C|}$, where G is the number of annotations extracted from all the documents (lead sentences) for a given information slot, C is the amount of documents for which a given annotation corresponds to an expert annotation of the same slot.

Test Set Annotation Evaluation: Precision

Language	No. of annots	Precision			
		ET	**ER**	**EL**	**EW**
Bulgarian	1053	0.99	0.98	0.95	0.87
French	1093	0.92	0.92	0.91	0.66
Polish	893	0.91	0.96	0.90	0.90
Russian	1031	0.93	0.95	0.97	0.67
Spanish	1023	0.96	0.93	0.90	0.90
Swedish	906	0.99	0.97	0.95	0.80

Table : Feature extraction performance evaluation on the test set (500 documents per language)

Test Set Annotation Evaluation: Recall

Language	No. of annots	Recall			
		ET	**ER**	**EL**	**EW**
Bulgarian	1053	1	0.97	0.87	0.86
French	1093	0.99	0.90	0.91	0.62
Polish	893	1	0.84	0.84	0.67
Russian	1031	0.99	0.90	0.89	0.75
Spanish	1023	0.99	0.88	0.93	0.83
Swedish	906	1	0.95	0.91	0.54

Table : Feature extraction performance evaluation on the test set (500 documents per language)

Conclusion Note

The current version uses a knowledge-driven approach and aims at assisting human annotators in creating feature-rich multilingual protest event datasets. The nearest plans are:

- to make the feature set more complete and useful,
- to enhance extraction using machine learning and multilingual knowledge bases,
- to become a part of a larger project.