Vincent Kríž

Detecting Entity Relations in Texts

Intelligent library (INTLIB, TA02010182)

KEG Seminar, 2013-12-19

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

kriz@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/~kriz

Outline

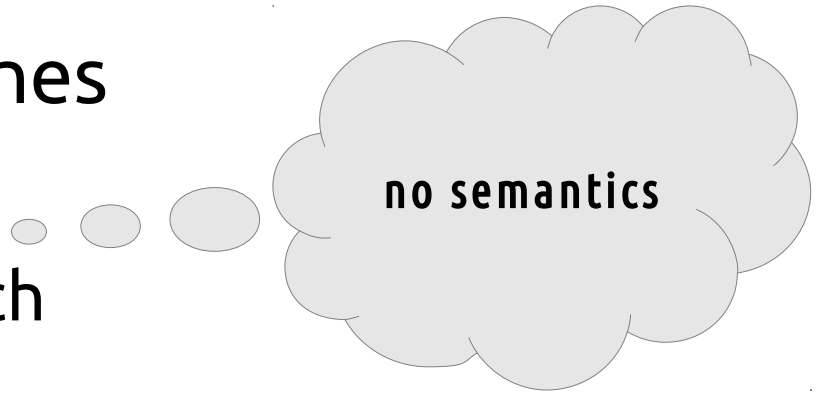
- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

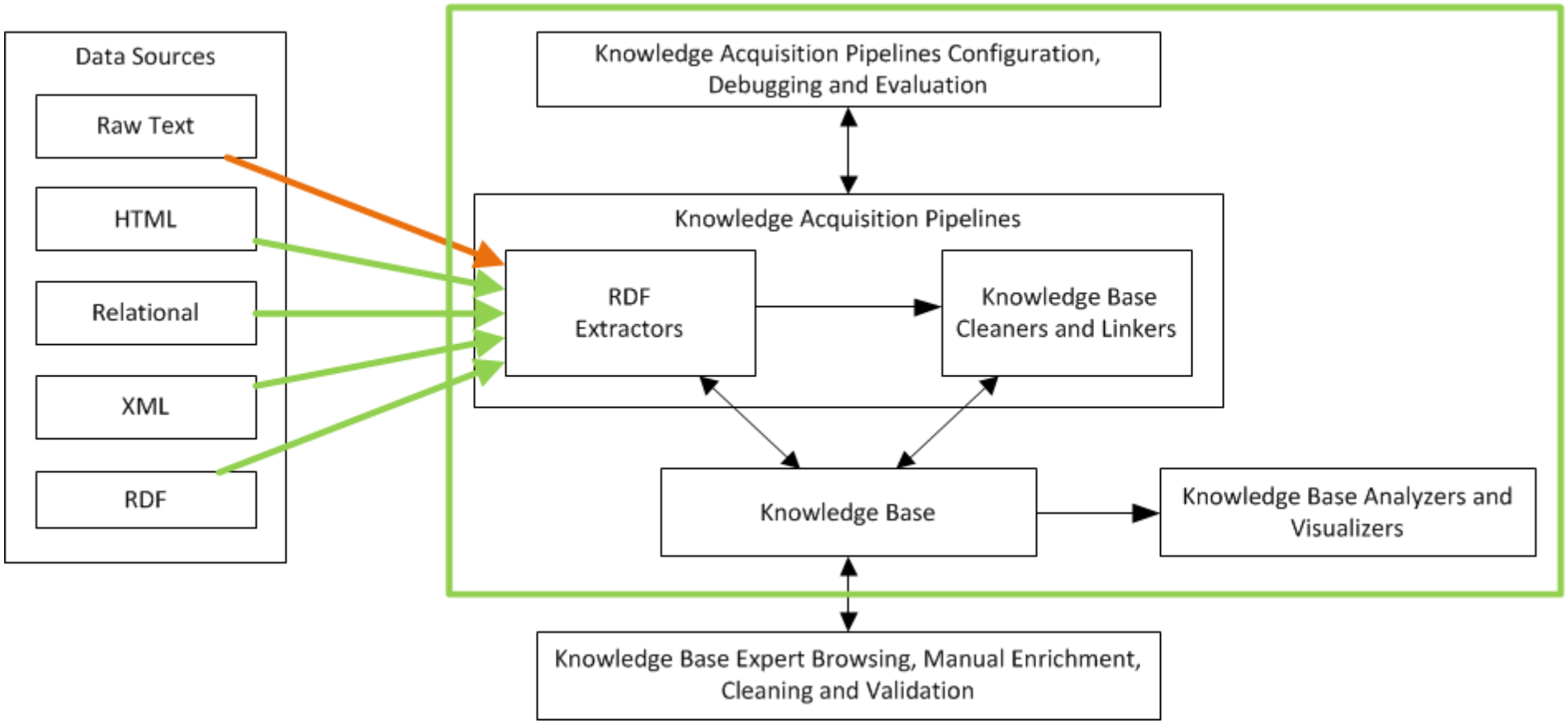
Motivation

- large collections of documents
- efficient browsing & querying
- typical approaches
  - full-text search
  - metadata search



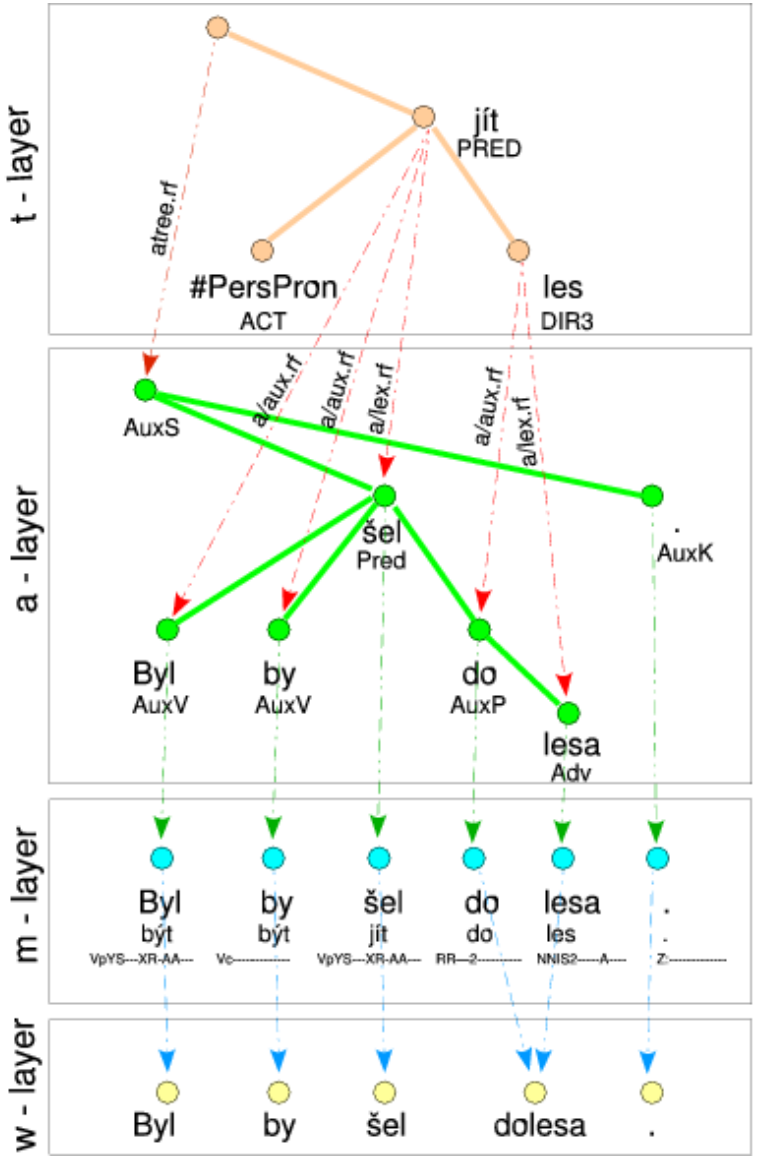
- semantics interpretation of documents → suitable DB & query language → user-friendly browsing & querying

INTLIB


Data Sources
Raw Text
HTML
Relational
XML
RDF
Knowledge Acquisition Pipelines Configuration, Debugging and Evaluation
Knowledge Acquisition Pipelines
RDF Extractors
Knowledge Base Cleaners and Linkers
Knowledge Base
Knowledge Base Analyzers and Visualizers
Knowledge Base Expert Browsing, Manual Enrichment, Cleaning and Validation

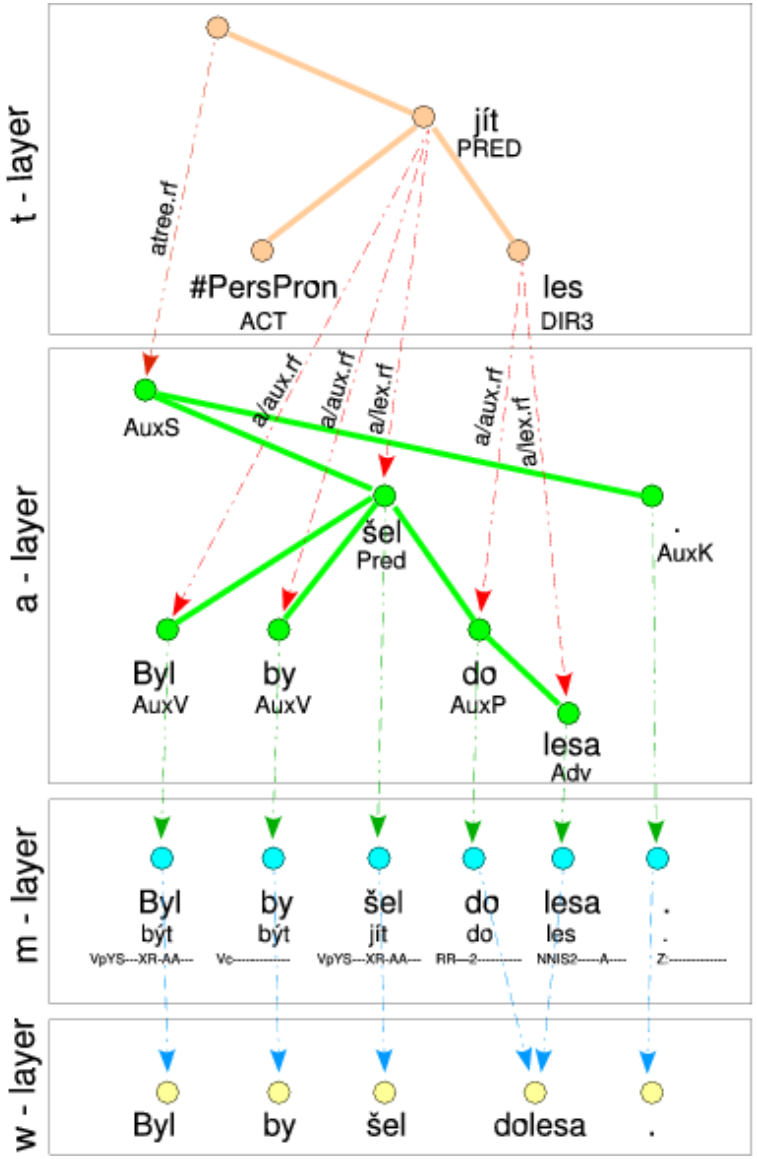
NLP Group

- **Documents**
  - semi-structured documents from some domain
  - legislative documents, project/medical documentation
- **Extractor**
  - NLP techniques
- **Data ›› Linked data**
  - automatically interconnected with other related data and with the original documents



NLP Group

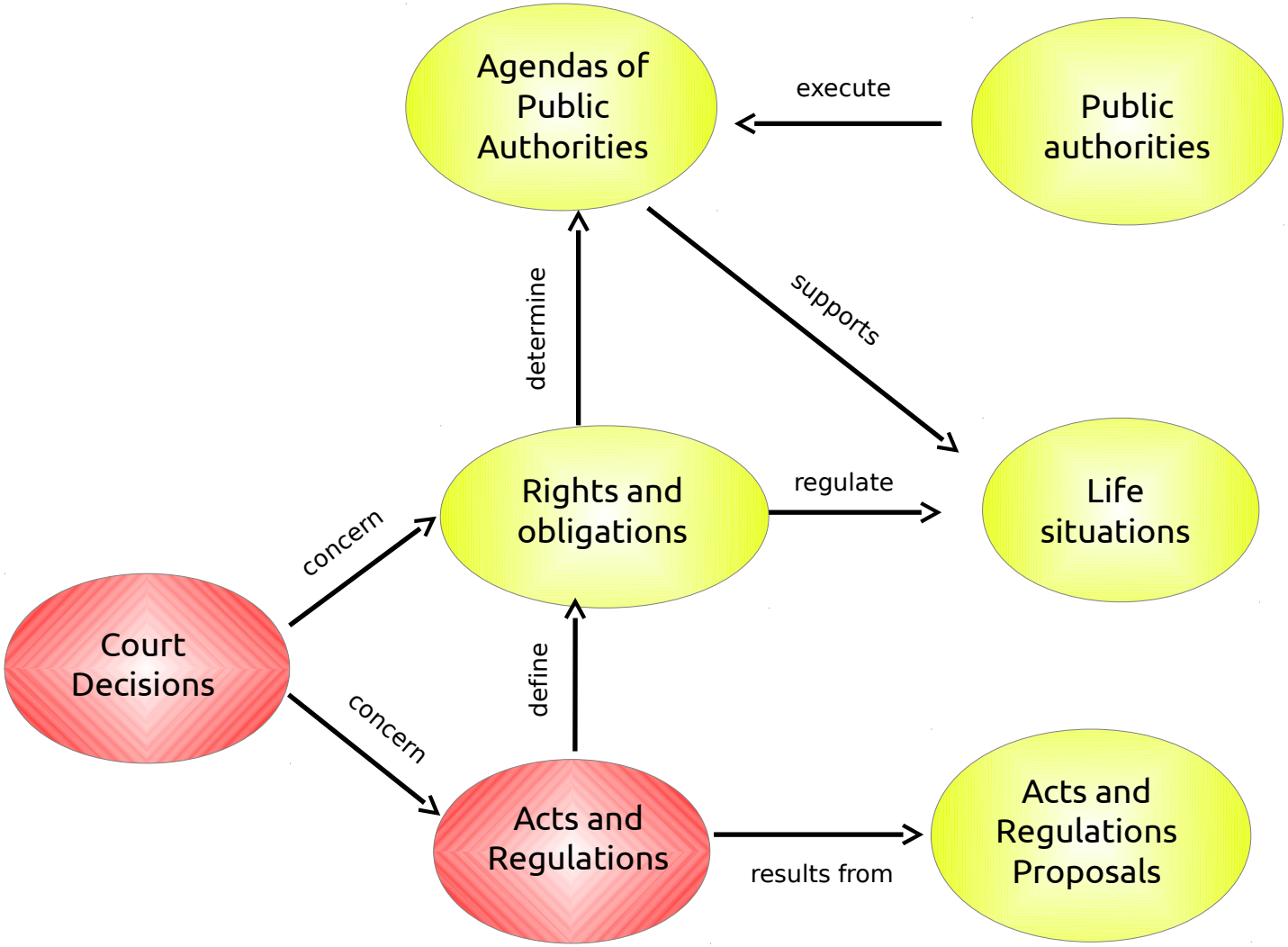
- **Tools**
  - segmentation & tokenization
  - lemmatization & morphology
  - syntactic parsing
  - deep syntactic parsing



Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - Enities
      - JTagger
    - Relations
      - Rextractor
      - JStories
  - Improving syntax

Legislation domain



Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

Legislation domain - structure

What we have done

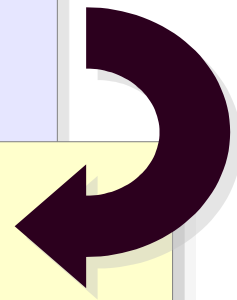
- ontology of legislative documents
- metadata and structure of acts, regulations and decrees represented as Linked Data
  - metadata about each version of each act, regulation and decree since 1945
  - structured content of versions of all acts, regulations and decrees valid in 2011, 2012

Legislation domain - structure

```
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy

Tato vyhláška zapracovává příslušné předpisy Evropské unie a upravuje:
   a) způsob vymezení hydrogeologických rajonů, vymezení útvarů podzemních vod,
   b) způsob hodnocení stavu podzemních vod a
   c) náležitosti programů zjišťování a hodnocení stavu podzemních vod.
```

Legislation domain - structure

```
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy

Tato
   a
   b
   c
```



```
<head id="11" label="HLAVA I">
        <title>ÚVODNÍ USTANOVENÍ</title>
        <section id="12" label="§ 1">
                <title>Předmět úpravy</title>
                <text>Tato vyhláška zapracovává příslušné
                      předpisy Evropské unie a upravuje:</text>
                <section id="13" label="a)">
                        <text>způsob vymezení hydrogeologických rajonů,
                              vymezení útvarů podzemních vod,</text>
                </section>
                <section id="14" label="b)">
                        <text>způsob hodnocení stavu podzemních vod a</text>
                </section>
                <section id="15" label="c)">
                        <text>náležitosti programů zjišťování a
                              hodnocení stavu podzemních vod.</text>
                </section>
        </section>
</head>
```

Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - Jstories
    - Rextractor
  - Improving syntax

Legislation domain - semantics

Extracting **concepts** and **relationships** between them from documents

- court decisions
  - **Entities:** references, institutions, acts, dates
  - **Relations:** whole *story* of a case
- acts, regulations, ...
  - **Entities:** subjects, things, locations, ...
  - **Relations:** rights, obligations, ...

Outline

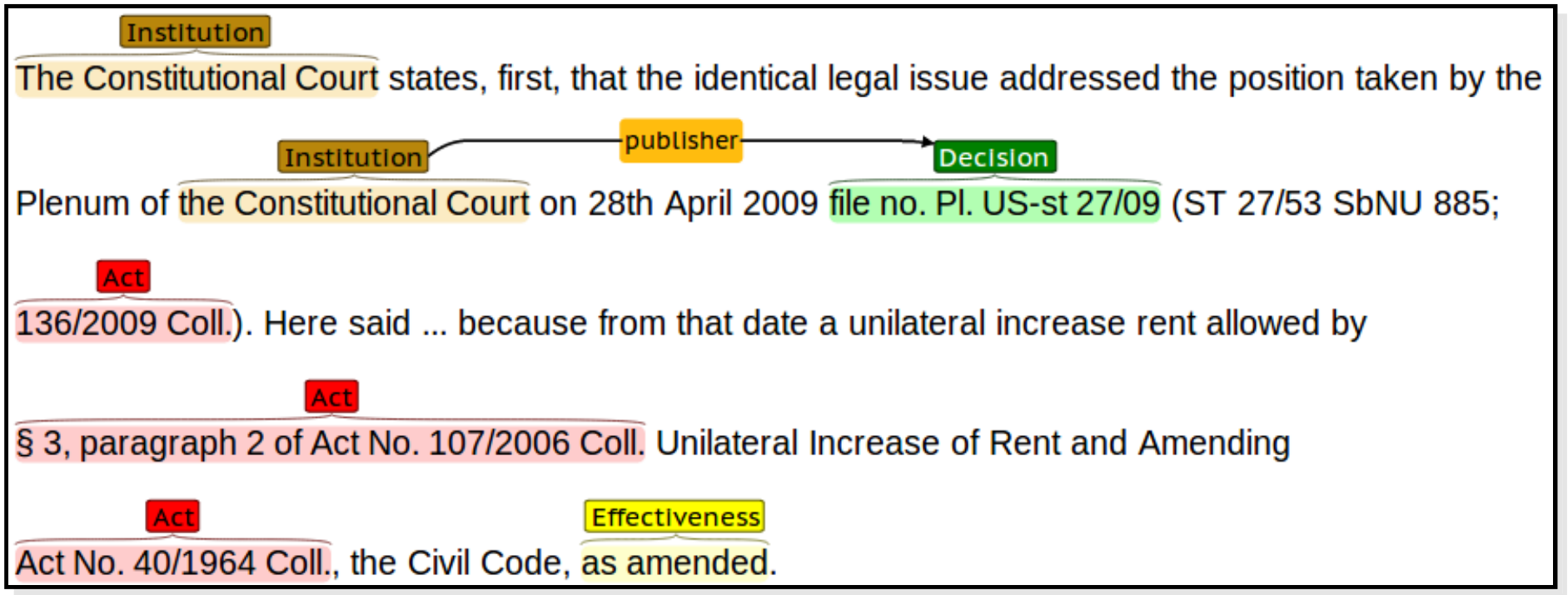
- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

JTagger

- **Entities**
  - References on
    - court decisions
    - acts
  - Effectiveness of Act
  - Institutions
- **Relations**
  - Publisher
    - Institution → Decision
  - Abbreviation

JTagger

- Annotation in **Brat** (http://brat.nlplab.org)



JTagger

- **Data sets**
  - Corpus of manually annotated court decisions
    - The Supreme Court (150)
    - The Constitutional Court (150)

	SC			CC		
	# of docs	# of tokens	# of entities	# of docs	# of tokens	# of entities
Training set	135	332,535	8,487	135	312,191	7,910
Test set	15	36,999	943	15	34,701	879
Total	150	369,534	9,430	150	346,892	8,789

JTagger

- **Data sets**
  - Corpus of manually corrected court decisions
    - The Supreme Court (93)
    - The Constitutional Court (91)

	SC			CC		
	# of docs	# of tokens	# of entities	# of docs	# of tokens	# of entities
Total	93	120,856	6,047	91	100,464	4,945

JTagger

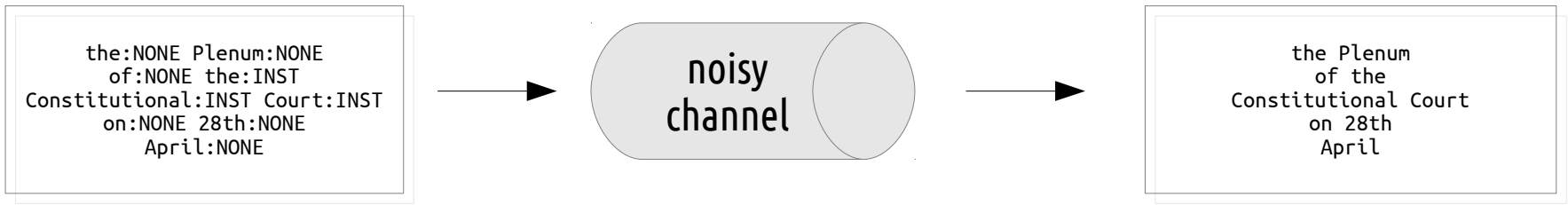
- **Machine Learning experiments**
  - Hidden Markov models (HMM)
  - Perceptron Algorithm with Uneven Margins (PAUM)

JTagger

- **Hidden Markov models (HMM)**
  - pattern recognition - speech, handwriting, gesture recognition, **part-of-speech tagging**, ...

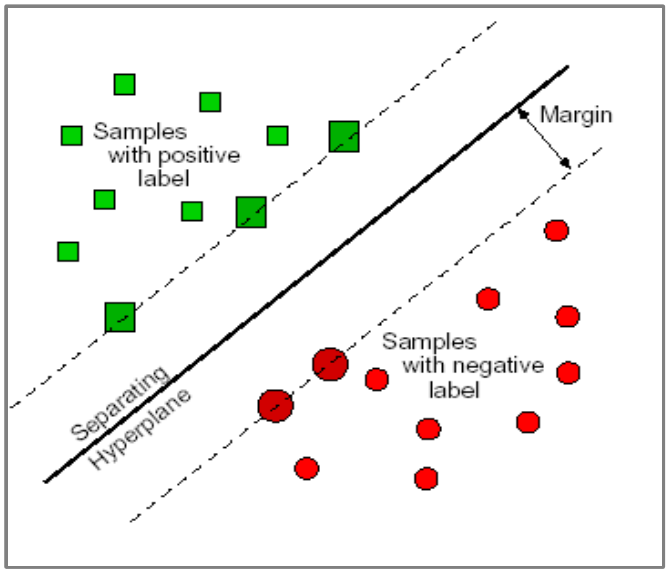
the	Plenum	of	the	Constitutional	Court	on	28th	April
DT	NNP	IN	DT	NNP	NNP	IN	JJ	NNP
NONE	NONE	NONE	INTS	INST	INST	NONE	NONE	NONE

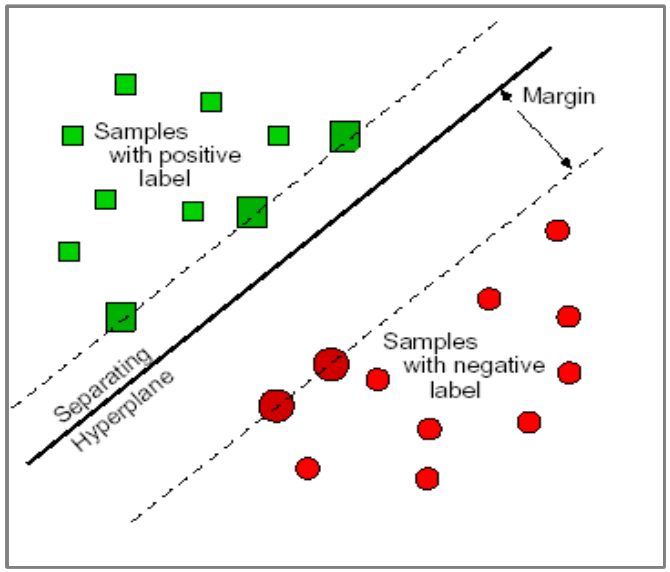
  - noisy chanel



JTagger

- **Perceptron Algorithm with Uneven Margins (PAUM)**
  - Implementation in the GATE framework
  - Perceptron and Support Vector Machines
- PAUM doesn't position the separator right between the points, but over one side





Credit: GATE, The University of Sheffield

JTagger

- Models
  - **HMM**
  - **PM Small**
    - trigrams of word forms
  - **PM**
    - 5-grams of word forms
  - **PM POS**
    - 5-grams of lemmas and POS-tags
  - **PM POS EXT**
    - PM POS + orthography features

JTagger

Strict F1 on entities

	Entity	HMM	PM pos ext	PM pos	PM	PM small
SC	A	0,75±0.02	0,91±0,02	0,91±0,03	0,89±0,03	0,88±0,03
	D	0,82±0,08	0,97±0,02	0,96±0,02	0,95±0,03	0,94±0,02
	E	0,89±0,04	0,90±0,05	0,89±0,05	0,88±0,08	0,82±0,1
	I	0,92±0,03	0,96±0,02	0,96±0,02	0,95±0,02	0,96±0,02
CC	A	0,63±0,05	0,87±0,02	0,86±0,02	0,84±0,03	0,78±0,03
	D	0,83±0,05	0,95±0,03	0,95±0,03	0,93±0,03	0,92±0,03
	E	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03
	I	0,91±0,02	0,93±0,02	0,93±0,02	0,92±0,01	0,92±0,01

JTagger

Lenient F1 on entities

	Entity	HMM	PM pos ext	PM pos	PM	PM small
SC	A	0,93±0,02	0,96±0,01	0,96±0,01	0,95±0,01	0,95±0,02
	D	0,91±0,03	0,98±0,01	0,97±0,02	0,96±0,02	0,95±0,02
	E	0,94±0,04	0,91±0,05	0,90±0,05	0,90±0,06	0,83±0,1
	I	0,97±0,01	0,98±0,00	0,98±0,01	0,97±0,01	0,97±0,01
CC	A	0,89±0,02	0,94±0,01	0,94±0,01	0,94±0,01	0,93±0,02
	D	0,93±0,03	0,97±0,02	0,97±0,02	0,96±0,02	0,95±0,03
	E	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03	0,96±0,03
	I	0,97±0,01	0,98±0,01	0,98±0,01	0,97±0,01	0,97±0,01

JTagger

- On-line DEMO
  - **http://ufal.mff.cuni.cz/jtagger**
- Open data
  - JTagger as a component of ODCleanStore
    - **http://sourceforge.net/projects/odcleanstore/**
  - daily, fully automatic
  - processing and publication of the new decisions

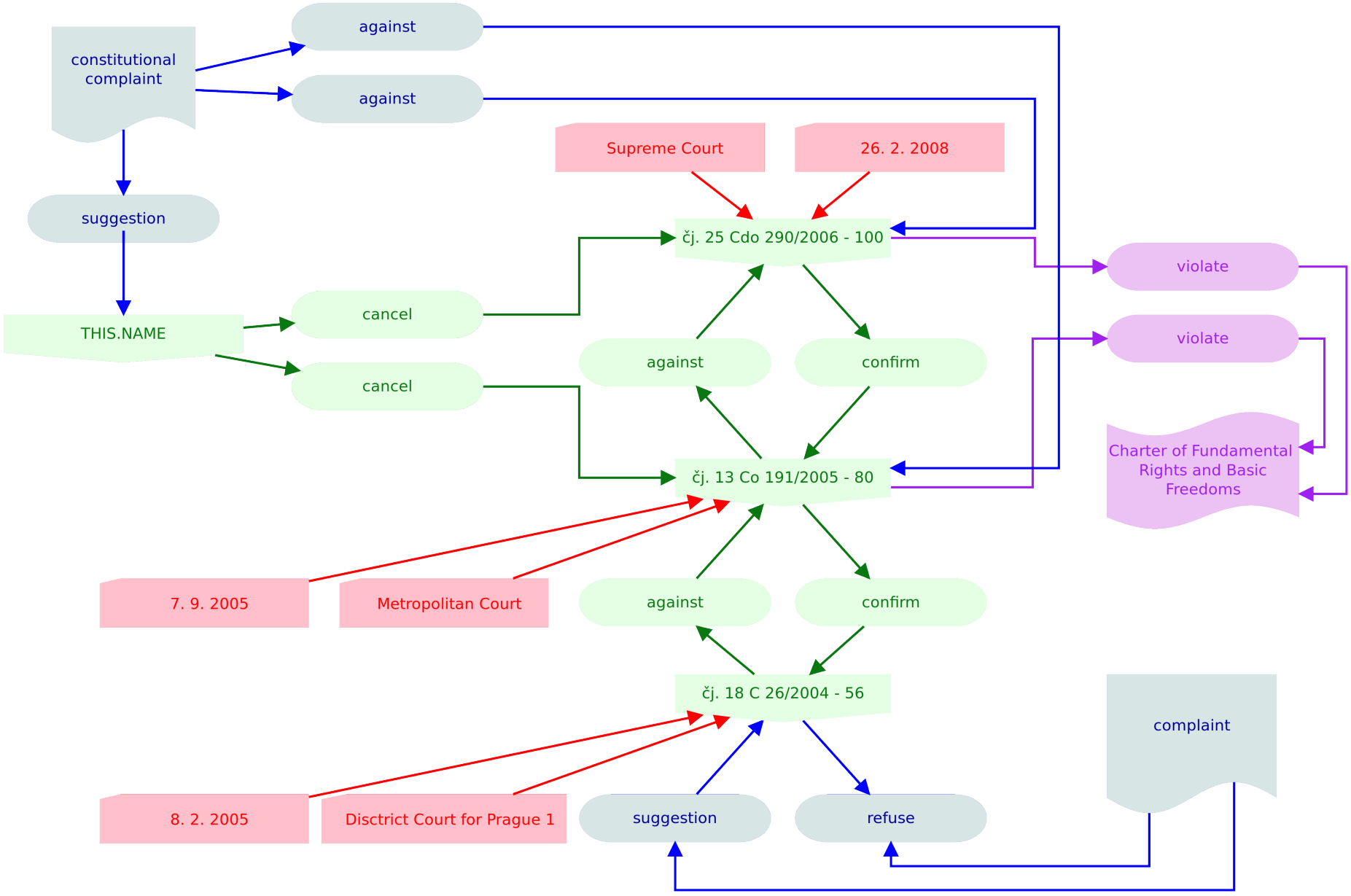
Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

JStories

- Legal case retrospection
  - a story
  - begins when an accuser files a complaint to a court
  - ends when a final decision is rendered

JStories


against
constitutional complaint
against
Supreme Court
26. 2. 2008
suggestion
čj. 25 Cdo 290/2006 - 100
violate
cancel
THIS.NAME
violate
against
confirm
cancel
Charter of Fundamental Rights and Basic Freedoms
čj. 13 Co 191/2005 - 80
7. 9. 2005
Metropolitan Court
against
confirm
čj. 18 C 26/2004 - 56
complaint
8. 2. 2005
Disctrict Court for Prague 1
suggestion
refuse

Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

RExtractor

- General framework for information extraction from texts
- Entities detection and identifying relationships between them
- Pipeline:
  - syntactic analysis of the text
  - entities detection (in syntactic trees)
  - PML-TQ for relations

http://prezi.com/jvrkcl1ui-ug/?utm_campaign=share&utm_medium=copy&rc=ex0share

Outline

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax

Improving syntactic parsing

- Parsers trained on PDT data
  - How they work on legal texts?
  - Too many nodes make automatic parsing almost impossible
- Manual syntactic analysis of Czech legal texts

http://prezi.com/m4r9r1nhzi7n/?utm_campaign=share&utm_medium=copy&rc=ex0share

Conclusion

- INTLIB
- Legislative documents
  - Structure
  - Semantics
    - JTagger
    - JStories
    - Rextractor
  - Improving syntax