Vincent Kríž

Statistical Recognition of References in Court Decisions

Intelligent library (INTLIB, TA02010182)

Seminář strojového učení a modelování, 2014-02-27

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

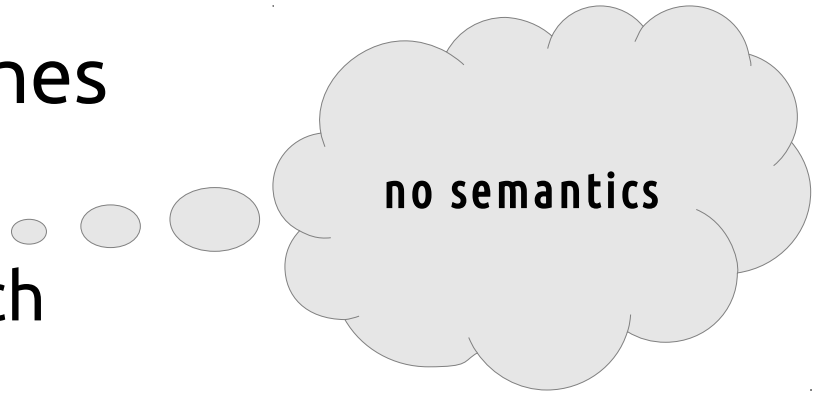
- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

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

Strategy

- **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

Legislation domain - semantics

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

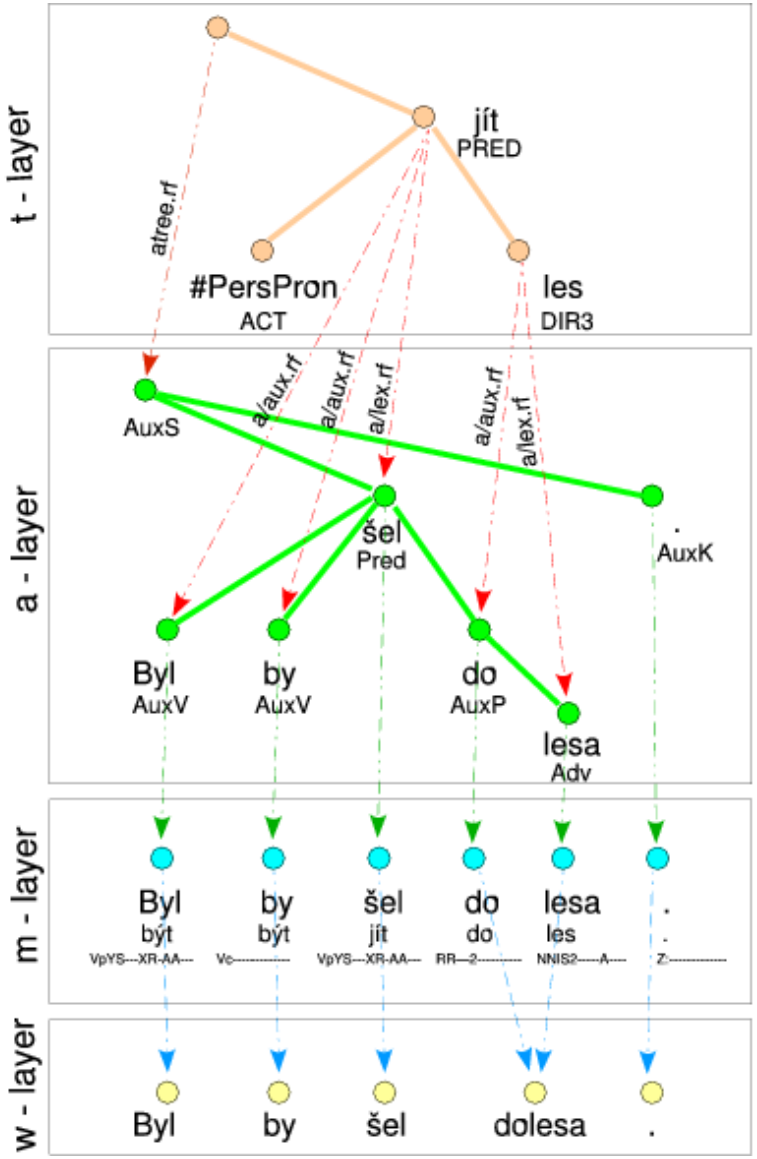
- court decisions
  - **Entities:** references, institutions, acts, dates
  - **Relations:** whole *story* of a case
- acts, regulations, ...
  - **Entities:** subjects, things, locations, ...
  - **Relations:** rights, obligations, ...

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

NLP Group

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



Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Czech court decisions

- Unified style of court decisions
- No rules what to cite
  - Other court decisions only
  - Literature
  - Everything
    - Blogs, Web sites, Bible, …

Existing systems

- ASPI
  - No hyperlinks in texts

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Motivation

Ústavní soud rozhodl v senátu složeném z předsedy senátu Vojena Güttlera
a soudců Ivany Janů a Františka Duchoně o ústavní stížnosti stěžovatele
Ing. M. B., zastoupeného Mgr. Jaromírem Hladíkem, advokátem advokátní
kanceláře Hladík, Hladíková & Partneři se sídlem 17. listopadu 623, Pardubice,
proti výroku I. rozsudku Okresního soudu ve Svitavách ze dne 21. 02. 2006,
č. j. 0 Nc 426/2004-210, proti výroku I. rozsudku Okresního soudu ve Svitavách
ze dne 05. 12. 2007, č. j. 0 P 348/2006-297, proti výroku I. rozsudku Krajského
soudu v Hradci Králové - pobočka v Pardubicích ze dne 05. 08. 2008,
č. j. 22 Co 116/2008-516, proti usnesení Okresního soudu ve Svitavách ze
dne 13. 07. 2011, č. j. 0 P 348/2006-897, ve spojení s opravným usnesením
Okresního soudu ve Svitavách ze dne 20. 07. 2011, č. j. 0 P 348/2006-911,
a proti usnesení Krajského soudu v Hradci Králové - pobočka v Pardubicích
ze dne 17. 10. 2011, č. j. 22 Co 440/2011-953, a proti průtahům v řízení
vedeném u Okresního soudu ve Svitavách pod sp. zn. 0 P 348/2006, za účasti
Krajského soudu v Hradci Králové - pobočka Pardubice a Okresního soudu ve
Svitavách jako účastníka řízení a nezl. S. B. a M. B., jako vedlejších
účastníků řízení, zastoupených Městským úřadem v Moravské Třebové,
nám. T. G. Masaryka 29, Moravská Třebová (dále jen "opatrovník"), takto :

Motivation

Ústavní soud rozhodl v senátu složeném z předsedy senátu Vojena Güttlera
a soudců Ivany Janů a Františka Duchoně o ústavní stížnosti stěžovatele
Ing
kan
pro
č.
ze
sou
č.
dne
Okr
a p
ze
ved
Kra
Svi
úča
nám

Ústavní soud rozhodl v senátu složeném z předsedy senátu Vojena Güttlera a soudců Ivany Janů a Františka Duchoně o ústavní stížnosti stěžovatele Ing. M. B., zastoupeného Mgr. Jaromírem Hladíkem, advokátem advokátní kanceláře Hladík, Hladíková & Partneři se sídlem 17. listopadu 623, Pardubice, proti výroku I. rozsudku Okresního soudu ve Svitavách ze dne 21. 02. 2006, č. j. 0 Nc 426/2004-210, proti výroku I. rozsudku Okresního soudu ve Svitavách ze dne 05. 12. 2007, č. j. 0 P 348/2006-297, proti výroku I. rozsudku Krajského soudu v Hradci Králové - pobočka v Pardubicích ze dne 05. 08. 2008, č. j. 22 Co 116/2008-516,

Motivation

Včas podanou ústavní stížností, splňující i další formální
náležitosti podání dle zákona č. 182/1993 Sb.,
o Ústavním soudu, ve znění pozdějších předpisů
(dále jen "zákon o Ústavním soudu"),
brojí stěžovatelka proti výše citovaným rozhodnutím,
neboť má za to, že jimi bylo porušeno její ústavně zaručené
právo na spravedlivý proces garantované
čl. 36 Listiny základních práv a svobod
(dále jen "Listina") a čl. 90 Ústavy ČR (dále jen "Ústava").

Motivation

Včas podanou ústavní stížností, splňující i další formální náležitosti podání dle zákona č. 182/1993 Sb., o Ústavním soudu, ve znění pozdějších předpisů (dále jen "zákon o Ústavním soudu"), brojí stěžovatelka proti výše citovaným rozhodnutím, neboť má za to, že jimi bylo porušeno její ústavně zaručené právo na spravedlivý proces garantované čl. 36 Listiny základních práv a svobod (dále jen "Listina") a čl. 90 Ústavy ČR (dále jen "Ústava").

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Court decisions

- **References to whole documents**
  - Document names
    - *Ústava ČR*
    - *Listina základních práv a slobod*
  - Document numbers
    - *zákon č. 128/2008 Sb.*
  - Both the name & the number
    - *zákon č. 128/2008 Sb., o Ústavním soudu*
- **References to specific parts in a document**
  - *§ 128 čl. 3 odst 1*

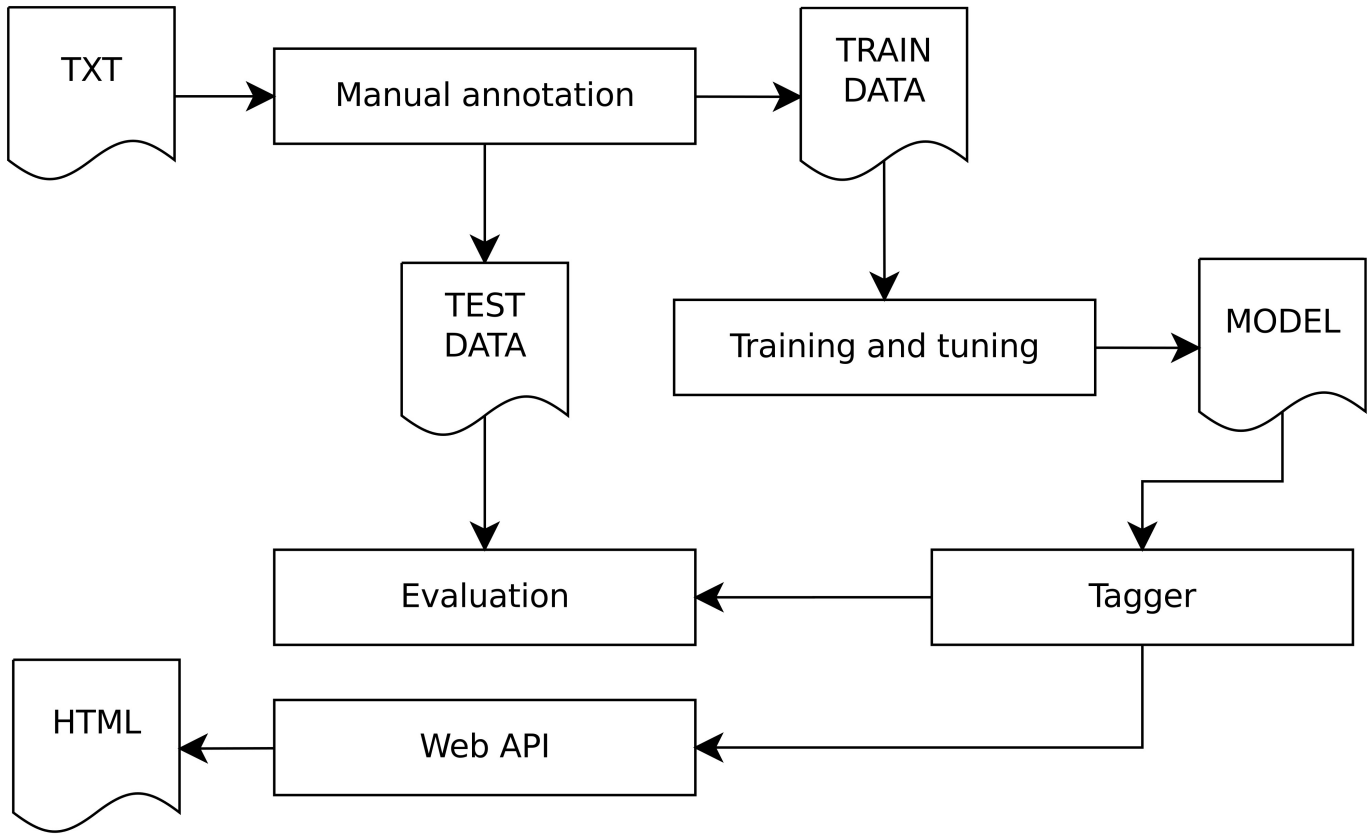
Court decisions

- **Abbreviations**
  - Acronyms
    - *Občanský soudní řád – OSŘ*
  - General words from the official institution's name
    - *Krajský soud v Pardubicích – krajský soud*
  - One word from the official act's name
    - *Listina základních práv a svobod – Listina*

Tagset

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

Experiment pipeline

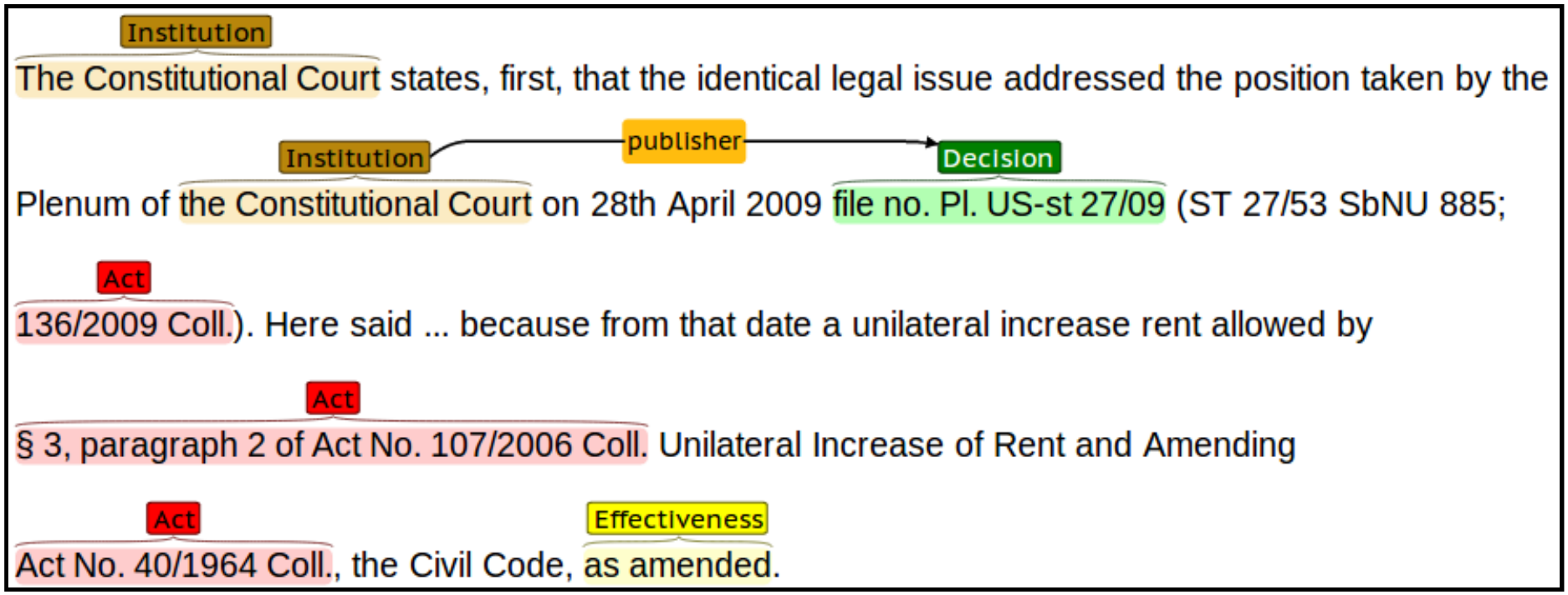


Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Manual annotation

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

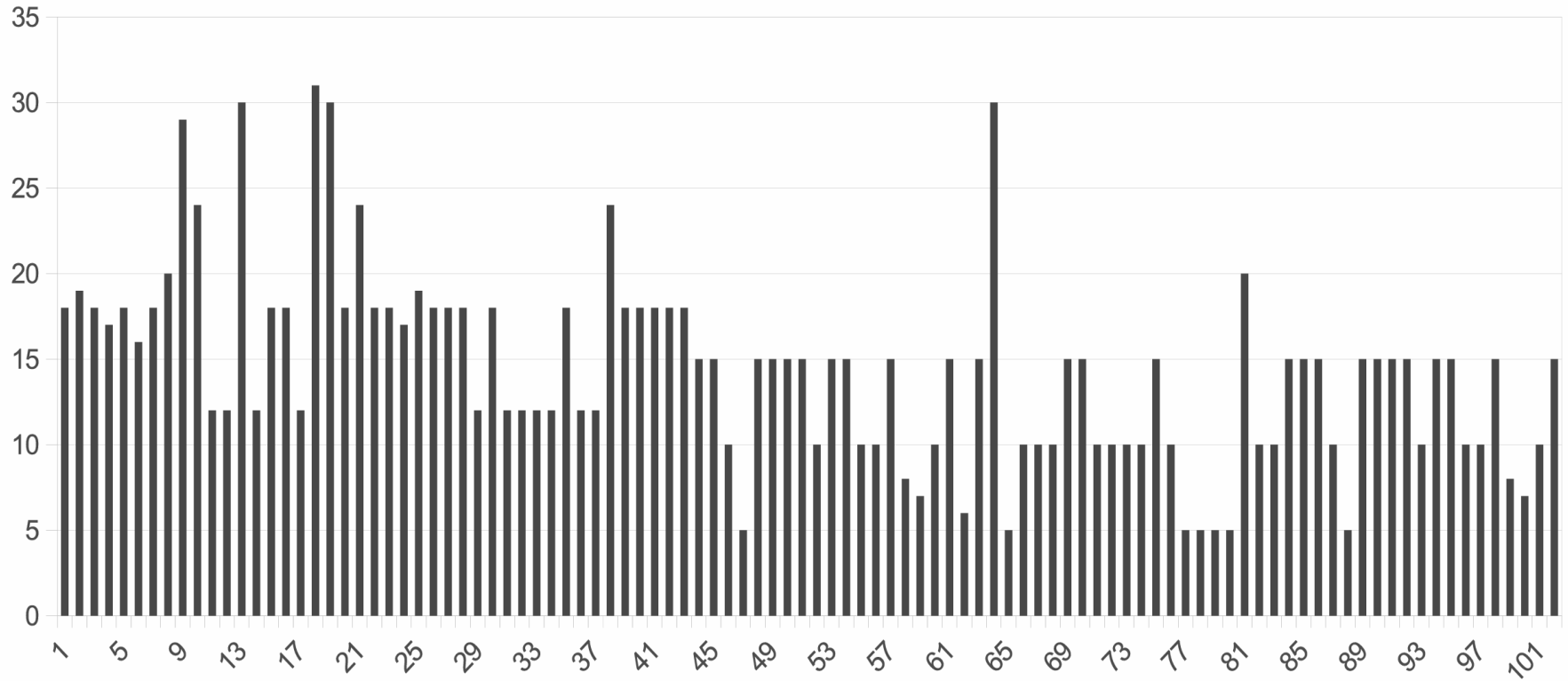


Manual annotation

- Annotate each occurrence of the entity
- One token can be annotated with more labels
  - *zákon o Ústatním soudu*
- Each document was annotated by one annotator

Manual annotation

- Minutes spend on the annotation (first 102 documents)



Manual annotation

- **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

Manual annotation

- **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

Manual annotation

- **Data sets**
  - Entity and token distribution in the training and test data averaged over 10 cross-validation folds

			Act		Decision		Effectiveness		Institution	
SC	# of Tokens	Training	43117	(89%)	11074	(86%)	1262	(83%)	12425	(90%)
		Test	5348	(11%)	1855	(14%)	265	(17%)	1450	(10%)
	# of Entities	Training	3949	(90%)	1304	(90%)	222	(90%)	2485	(90%)
		Test	439	(10%)	145	(10%)	25	(10%)	276	(10%)
CC	# of Tokens	Training	19675	(88%)	12780	(86%)	843	(89%)	14767	(89%)
		Test	2707	(12%)	2127	(14%)	102	(11%)	1743	(11%)
	# of Entities	Training	2338	(90%)	1481	(90%)	210	(90%)	3206	(90%)
		Test	260	(10%)	165	(10%)	23	(10%)	356	(10%)

Manual annotation

- **Data sets**
  - Average entity lengths in tokens averaged over 10 cross-validation folds

	SC		CC	
	Training set	Test set	Traininig set	Test set
Act	10.9	12.2	8.4	10.4
Decision	8.5	12.8	8.6	12.9
Effectiveness	5.7	10.7	4	4.4
Institution	5	5.3	4.6	4.9

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Experiments

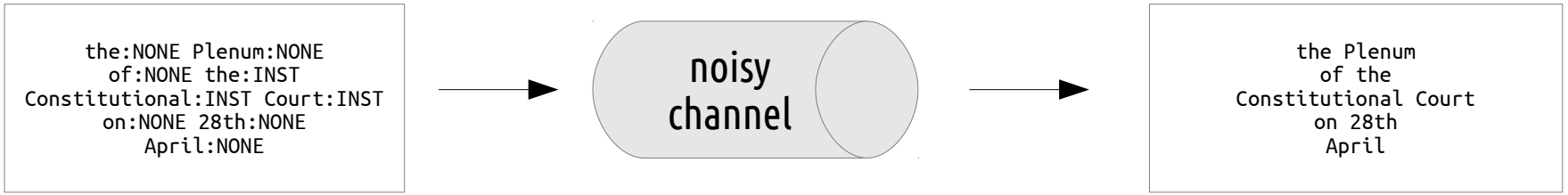
- **Machine Learning experiments**
  - Hidden Markov models (HMM)
  - Perceptron Algorithm with Uneven Margins (PAUM)

Methods

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

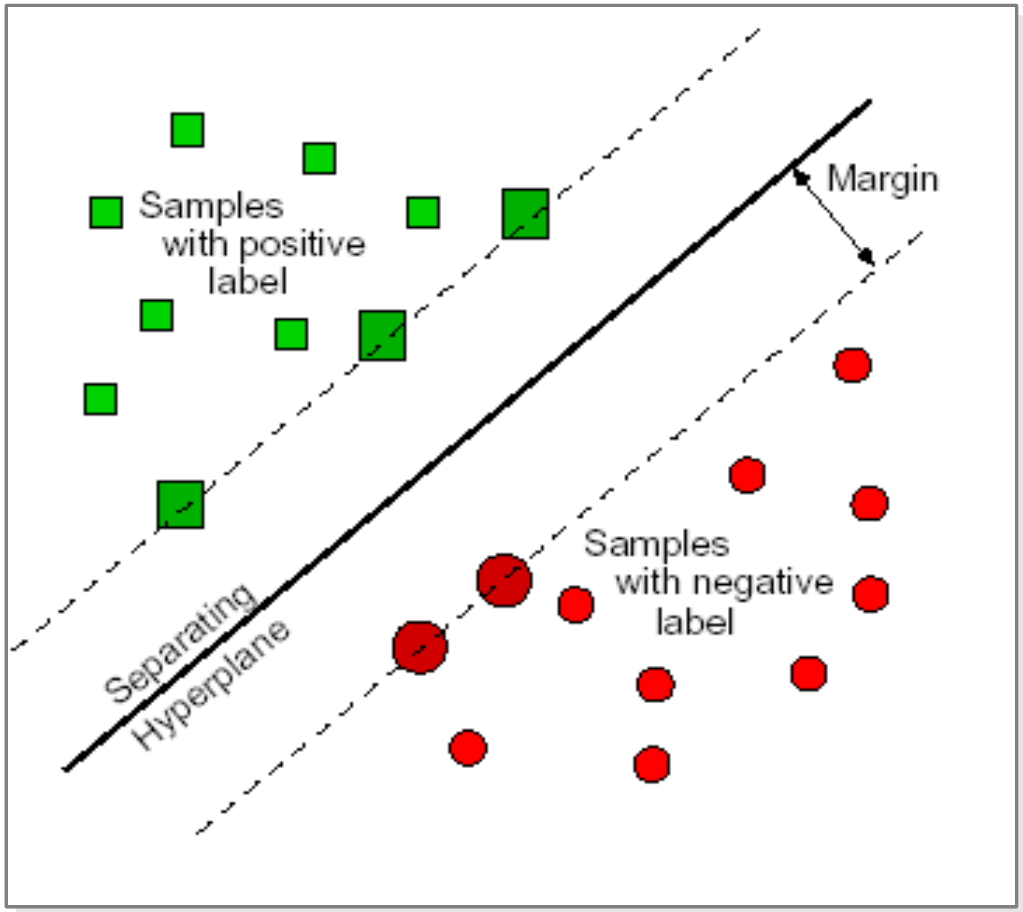
the	Plenum	of	the	Constitutional	Court	on	28th	April
DT	NNP	IN	DT	NNP	NNP	IN	JJ	NNP
NONE	NONE	NONE	INTS	INST	INST	NONE	NONE	NONE

  - noisy chanel



Methods

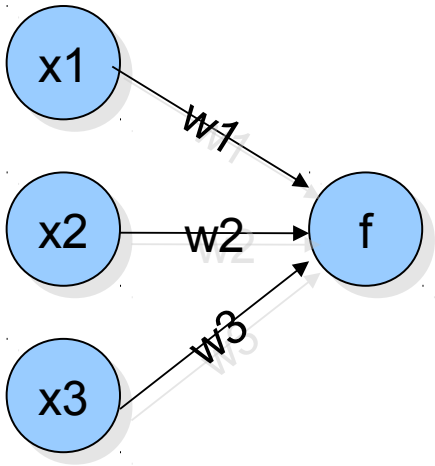
- **Support Vector Machines (SVM)**
  - attempt to find a hyperplane that separates data
  - goal: **maximize margin** separating two classes
  - wider margin = greater generalisation
  - kernel functions
  - simple extension for multiclass classifiers



Credit: GATE, The University of Sheffield

Methods

- **Perceptron**
  - oldest ML method
  - some similarities to SVM
  - checks the training examples one by one by predicting their labels
    - prediction is correct » the example is passed
    - otherwise » correct the model
  - stop when the model classifies all training examples correctly
  - **no-margin constraint**

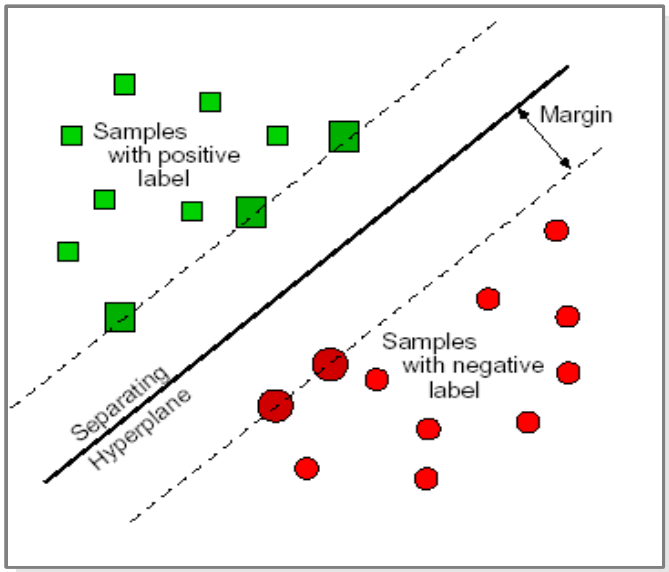


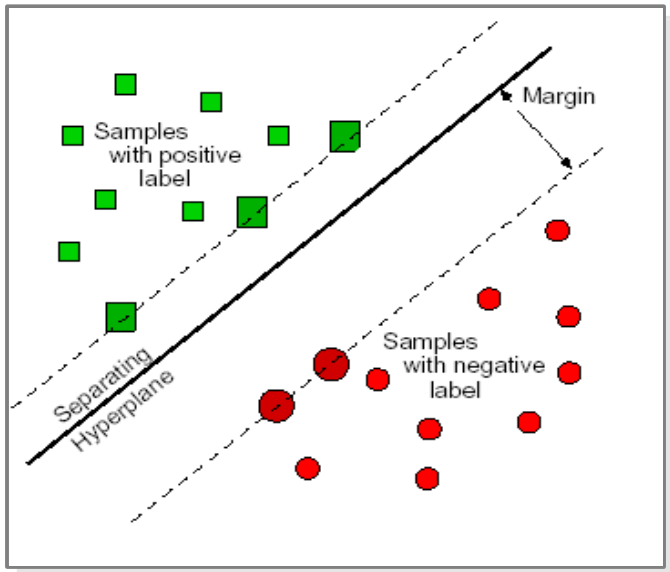
$$f(x) = \mathbf{w}x + b$$

Credit: GATE, The University of Sheffield

Methods

- **Perceptron Algorithm with Uneven Margins (PAUM)**
  - PAUM doesn't position the separator right between the points, but over one side
  - In NLP the datasets are often very imbalanced
    - *instances of "Person"*





Credit: GATE, The University of Sheffield

Models

- 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

Models

- **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

- **HMM**
  - Sequence of token + label

```
the:NONE
Plenum:NONE
of:NONE
the:INST
Constitutional:INST
Court:INST
on:NONE
28th:NONE
April:NONE
```

Models

- **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

- **PM Small**
  - trigrams of word forms

```
[the, Plenum, of]:NONE
[Plenum, of, the]:NONE
[of, the, Constitutional]:NONE
[the, Constitutional, Court]:INST_S
[Constitutional, Court, on]:NONE
[Court, on, 28th]:INST_E
[on, 28th, April]:NONE
```

Models

- **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

- **PM**
  - 5-grams of word forms

```
[the, Plenum, of, the, Constitutional]:NONE
[Plenum, of, the, Constitutional, Court]:NONE
[of, the, Constitutional, Court, on]:NONE
[the, Constitutional, Court, on, 28th]:INST_S
[Constitutional, Court, on, 28th, April]:NONE
```

Models

- **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

- **PM POS**
  - 5-grams of lemmas and POS-tags

```
[the, Plenum, of, the, Constitutional]:NONE
[DT, NN, RR, DT, AD]:NONE
[Plenum, of, the, Constitutional, Court]:NONE
[NN, RR, DT, AD, NN]:NONE
[of, the, Constitutional, Court, on]:NONE
[RR, DT, AD, NN, RR]:NONE
[the, Constitutional, Court, on, 28th]:INST_S
[DT, AD, NN, RR, CR]:INST_S
```

Models

- **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

- **PM POS EXT**
  - PM POS + orthography features

```
[FIRST, the, Plenum, of, the]:NONE
[XX, DT, NN, RR, DT]:NONE
[XX, LC, UI, LC, LC]:NONE
[the, Plenum, of, the, Constitutional]:NONE
[DT, NN, RR, DT, AD]:NONE
[LC, UI, LC, LC, UI]:NONE
```

Outline

- **INTLIB**
- **ÚFAL NLP World**
- **Legislation domain**
- **JTagger**
  - Motivation
  - Court decisions
  - Manual annotation
  - Experiments
  - Evaluation & Error analysis

Evaluation

- Standart evaluation measures
  - Accuracy $= tp + tn / tp + fp + fn + tn$
  - Precision $= tp / tp + fp$
  - Recall $= tp / tp + fn$
  - F-measure $= 1 / (a/P + 1-a/R)$
- Confusion matrix

		Goldstandard	
		True	False
Recognizer	Positive	tp	fp
	Negative	fn	tn

Evaluation

- Multi-token entities » overlapping matches
- **Evaluation**
  - on tokens
  - on entities
    - strict
    - lenient
- **Statistical significance**
  - 10-fold cross-validation
  - Confidence intervals (t-Test)

Evaluation

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

Evaluation

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

Evaluation

F1 on tokens

	Entity	HMM	PM pos ext	PM pos	PM	PM small
SC	A	0,96±0,01	0,96±0,01	0,96±0,01	0,96±0,02	0,95±0,02
	D	0,95±0,02	0,98±0,01	0,98±0,02	0,97±0,02	0,96±0,02
	E	0,94±0,03	0,89±0,06	0,88±0,06	0,88±0,06	0,79±0,12
	I	0,96±0,01	0,97±0,01	0,97±0,01	0,97±0,01	0,96±0,02
CC	A	0,94±0,01	0,94±0,01	0,93±0,01	0,93±0,02	0,89±0,02
	D	0,95±0,02	0,96±0,02	0,96±0,01	0,96±0,02	0,94±0,02
	E	0,96±0,03	0,96±0,04	0,96±0,04	0,96±0,04	0,96±0,04
	I	0,95±0,01	0,95±0,01	0,95±0,02	0,95±0,01	0,94±0,01

Error analysis

- References labeled with two separate tags instead of one tag
  - *file no. 7 To 346/2011*
- Numbers in the ends of court names
  - *District Court for Prague 4*
- Names of foreign courts
  - *Land Court in Norimberg, Germany*

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