



Vincent Kríž, Barbora Hladká

# Statistical Recognition of References in Court Decisions

Aplikace NLP, 10.5.2016  
MFF UK

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

[kriz@ufal.mff.cuni.cz](mailto:kriz@ufal.mff.cuni.cz)  
<http://ufal.mff.cuni.cz/~kriz>


# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# Motivation

- large collections of documents
  - efficient browsing & querying
  - typical approaches
    - full-text search
    - metadata search
- 
- semantic interpretation of documents →  
suitable DB & query language →  
user-friendly browsing & querying

# Legal Domain

- identification and extraction of relevant information
  - open and general problem
  - special relevance in **legal domain**
- Czech court decisions
  - published on-line as single documents
  - no interlinking

# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# Named Entity Recognition (NER)

- NER systems detect entities in texts and classify their semantic type
  - e.g. names of people, places, organizations, products, dates
- reference recognition as the task of NER

# Methods for NER

- **Lookup**
- **Rule-based**
- **Statistical**



# Methods for NER

- **Lookup**

- use a list of entities and tag all mentions in texts
- may generate many false positives
- inflective languages need manipulation with several word forms per lemma
  - e.g. 7 cases × 2 numbers × 3 genders in Czech
- detect just entities in the list
  - names of new laws, new abbreviations, new acronyms

# Methods for NER

- **Rule-based systems**
  - a set of rules may recognize majority of entities and not produce many false positives
  - require manually annotated data and a large amount of effort
  - maintenance can be tricky
    - rules often intricate interdependencies

# Methods for NER

- **Statistical**

- a set of clues with their weights
- weights are combined, based on probability and statistical concepts
- alternative to rule-based systems
- require manually annotated data and a large amount of effort
- adding a new development data is straightforward

# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# Czech court decisions

- no unified style of court decisions
- no rules what to cite
  - other court decisions only
  - literature
  - everything
    - blogs, web sites, Bible, ...

# References to Court Decisions

- several formats
- identification numbers is not unique
  - two courts may publish document with the same name
- identification of publishing organization is essential

# References to Acts

- **Simple references**

- name (*Act on Customs Administration*)
- label & number (*§12, Head 1*)
- anaphors, acronyms, indirect references (*Charter*)
- reference to a particular version

- **Complex references**

- multi-valued (*§2, 3 and 14*)
- multi-layered (*section 1, §2, of the Charter*)

# Tagset

- **Entities**

- References on
  - court decisions
  - acts
- Effectiveness of Act
- Institutions

- **Relations**

- Publisher
  - Institution → Decision
- Abbreviation



# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# Manual annotation

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

The Constitutional Court states, first, that the identical legal issue addressed the position taken by the Plenum of the Constitutional Court on 28th April 2009 file no. Pl. US-st 27/09 (ST 27/53 SbNU 885; 136/2009 Coll.). Here said ... because from that date a unilateral increase rent allowed by § 3, paragraph 2 of Act No. 107/2006 Coll. Unilateral Increase of Rent and Amending Act No. 40/1964 Coll., the Civil Code, as amended.

Annotations shown in the image:

- Institution** (yellow box) above "The Constitutional Court" and "Plenum of the Constitutional Court".
- Act** (red box) above "136/2009 Coll.", "Act No. 107/2006 Coll.", and "Act No. 40/1964 Coll.".
- Effectiveness** (yellow box) above "as amended".
- publisher** (yellow box) above "file no. Pl. US-st 27/09" with an arrow pointing to "Decision".
- Decision** (green box) above "file no. Pl. US-st 27/09".

# 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
<b>Training set</b>	135	332,535	8,487	135	312,191	7,910
<b>Test set</b>	15	36,999	943	15	34,701	879
<b>Total</b>	150	369,534	9,430	150	346,892	8,789

# 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 length 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

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# 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 channel



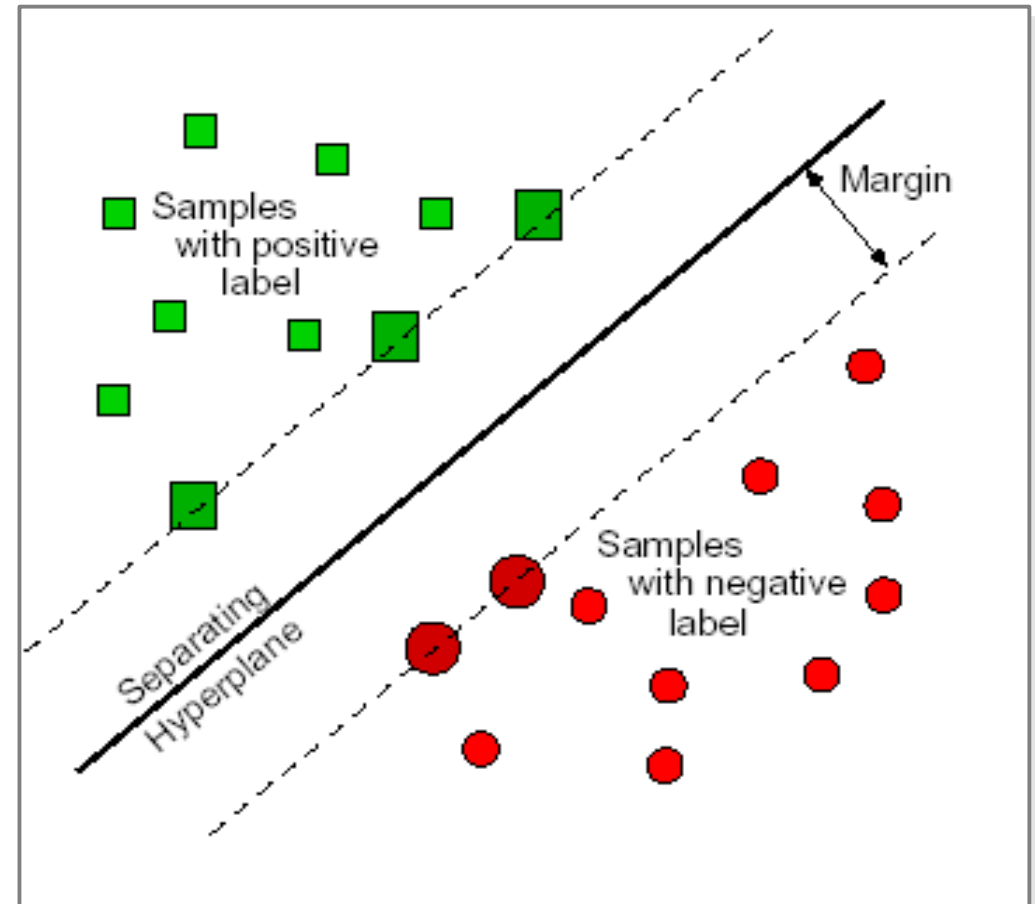


# Experiments

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

# Methods

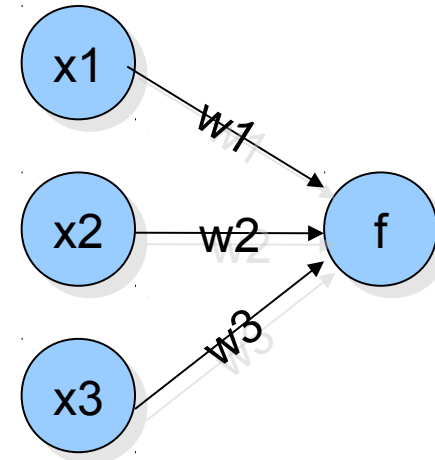
- **Support Vector Machines (SVM)**
  - attempt to find a hyperplane that separates data
  - goal: **maximize margin** separating two classes
  - wider margin = greater generalisation
  - kernel functions
  - simple extension for multiclass classifiers



Credit: GATE, The University of Sheffield

# Methods

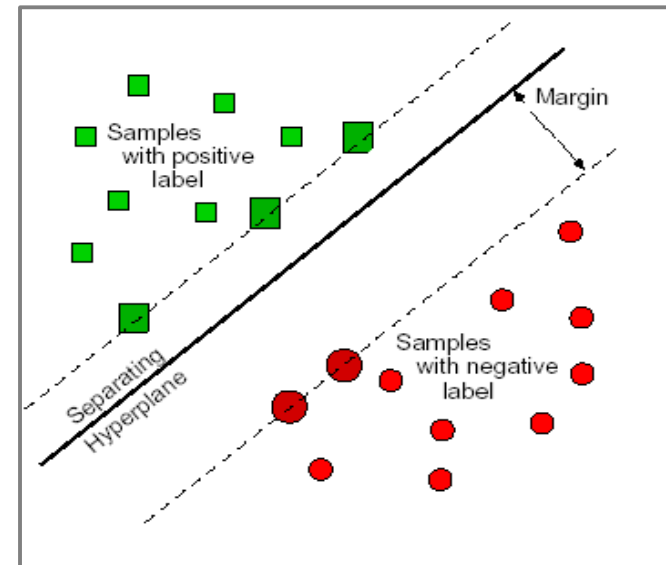
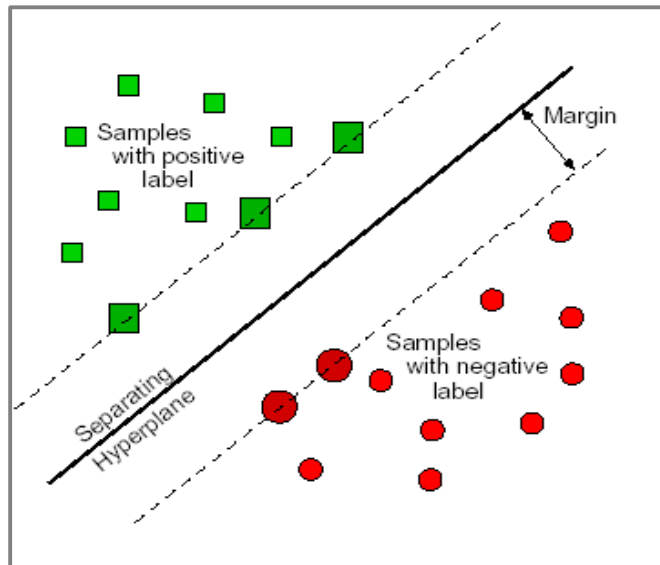
- **Perceptron**
  - oldest ML method
  - some similarities to SVM
  - checks the training examples one by one by predicting their labels
    - prediction is correct
      - » the example is passed
    - otherwise
      - » correct the model
  - stop when the model classifies all training examples correctly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



$$f(x) = wx + b$$

# 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 unbalanced
    - *instances of "Person"*



# Models

- Models
  - **HMM**
    - trigrams of word forms
  - **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

# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# 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 at the end of court names
  - *District Court for Prague 4*
- Names of foreign courts
  - *Land Court in Norimberg, Germany*

# Outline

- Introduction
- Named Entity Recognition
- Czech legal documents
- Data
- Methods & Models
- Evaluation & Error Analysis
- Conclusion

# JTagger

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

# Conclusion

- statistical-based system for reference recognition in Czech court decisions
- JTagger outperforms all results published so far
- JTagger published on-line