Semantic Annotations

Jan Dědek

Department of Software Engineering
Faculty of Mathematics and Physics
Charles University in Prague

Defence of Doctoral Thesis
21st September
2012

Outline

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

1. Introduction
    - Information Extraction
    - Deep Language Parsing
    - Inductive Logic Programming
    - Organization of this Presentation

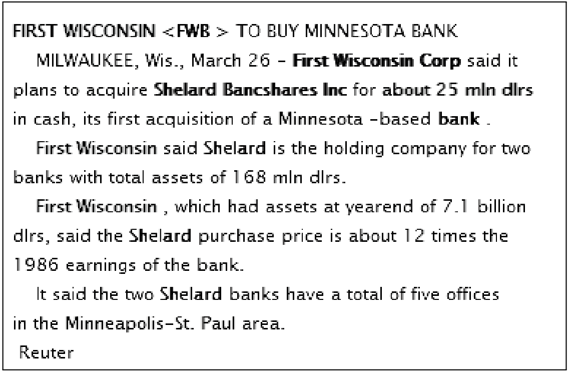
2. Contents
    - Manual Design of Extraction Rules
    - Induction of Extraction Rules
    - Shareable Extraction Ontologies
    - Fuzzy ILP Document Classification

3. Conclusion

4. Questions and Comments from Reviews
    - Review 1 (Filip Železný)
    - Review 2 (Diana Maynard)
    - PDT Annotations in GATE

Information Extraction (Problem)

- Let's have a text describing an acquisition event.



FIRST WISCONSIN <FWB > TO BUY MINNESOTA BANK

MILWAUKEE, Wis., March 26 - **First Wisconsin Corp** said it plans to acquire **Shelard Bancshares Inc** for **about 25 mln dlrs** in cash, its first acquisition of a Minnesota -based **bank** .

First Wisconsin said Shelard is the holding company for two banks with total assets of 168 mln dlrs.

First Wisconsin , which had assets at yearend of 7.1 billion dlrs, said the Shelard purchase price is about 12 times the 1986 earnings of the bank.

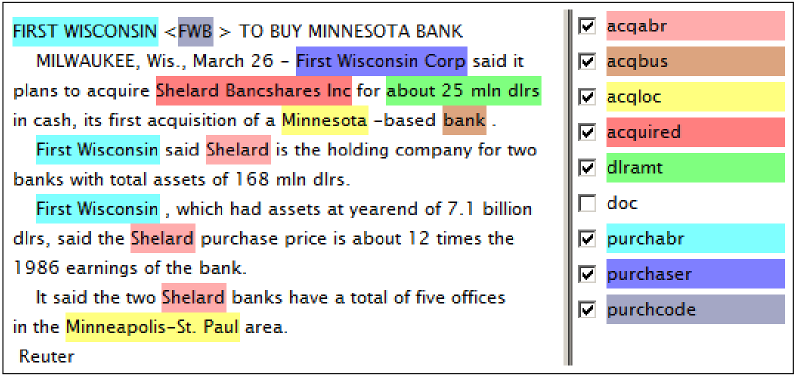
It said the two Shelard banks have a total of five offices in the Minneapolis-St. Paul area.

Reuter

- What was the object of the acquisition?
- Who was the buyer?
- What was the deal amount?

Information Extraction (Solution)

- Information Extraction (IE) tools can identify and extract such information.



FIRST WISCONSIN <FWB > TO BUY MINNESOTA BANK

MILWAUKEE, Wis., March 26 - First Wisconsin Corp said it plans to acquire Shelard Bancshares Inc for about 25 mln dlrs in cash, its first acquisition of a Minnesota -based bank .

First Wisconsin said Shelard is the holding company for two banks with total assets of 168 mln dlrs.

First Wisconsin , which had assets at yearend of 7.1 billion dlrs, said the Shelard purchase price is about 12 times the 1986 earnings of the bank.

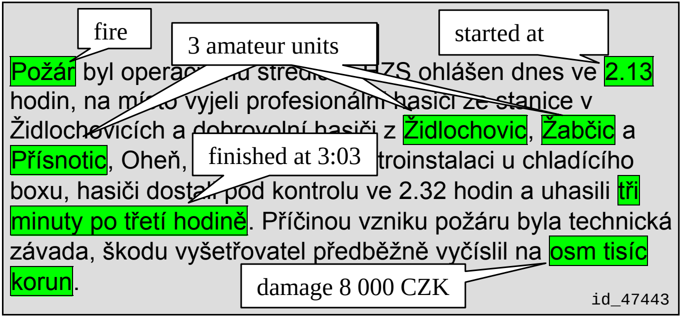
It said the two Shelard banks have a total of five offices in the Minneapolis-St. Paul area.

Reuter

☑ acqabr
☑ acqbus
☑ acqloc
☑ acquired
☑ dlramt
□ doc
☑ purchabr
☑ purchaser
☑ purchcode

Information Extraction (Czech Example)

- Information Extraction tools can identify and extract such information.



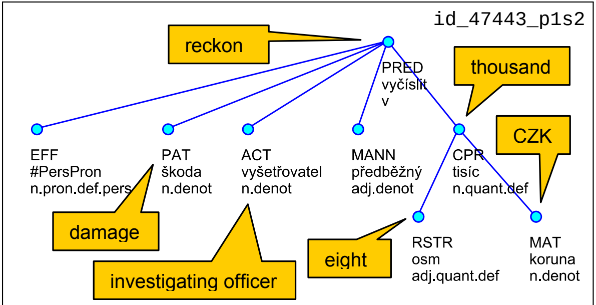
Požár byl operačnímu středisku HZS ohlášen dnes ve 2.13 hodin, na místo vyjeli profesionální hasiči ze stanice v Židlochovicích a dobrovolní hasiči z Židlochovic, Žabčic a Přísnotic. Oheň, [illegible] troinstalaci u chladícího boxu, hasiči dostali pod kontrolu ve 2.32 hodin a uhasili tři minuty po třetí hodině. Příčinou vzniku požáru byla technická závada, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

id_47443

1. Introduction
    - Information Extraction
    - **Deep Language Parsing**
    - Inductive Logic Programming
    - Organization of this Presentation
2. Contents
    - Manual Design of Extraction Rules
    - Induction of Extraction Rules
    - Shareable Extraction Ontologies
    - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
    - Review 1 (Filip Železný)
    - Review 2 (Diana Maynard)
    - PDT Annotations in GATE

Deep Language Parsing (Czech Example)

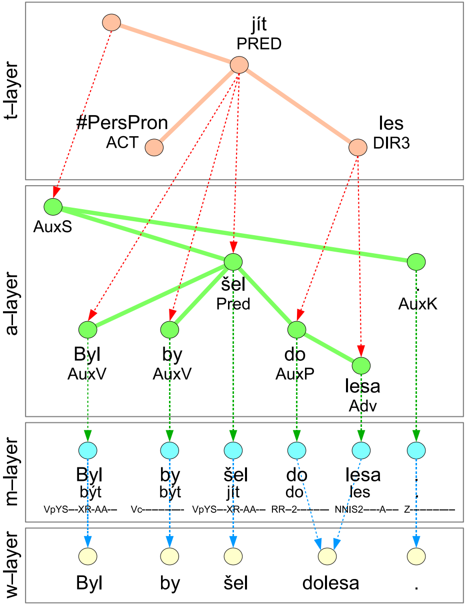
- Linguistic tools perform automated linguistic analysis.
- Producing so called *dependency trees*.



…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

…, investigating officer preliminarily reckoned the damage to be 8 000 CZK.

Layers of linguistic annotation in PDT



- Tectogrammatical layer
- Analytical layer
- Morphological layer

- PDT 2.0 on-line:

  `http://ufal.mff.cuni.cz/pdt2.0/`

Sentence:

Byl by šel dolesa.
He-was would went toforest.

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

Inductive Logic Programming

- Learning examples $E = P \cup N$ (Positive and Negative)
    - E.g. relevant and irrelevant pieces of text w.r.t. particular extraction task
- Background knowledge B
    - E.g. linguistic structure connecting individual words
- ILP task: To find logical program or hypothesis H such that all positive examples are covered and none negative

$$(\forall e \in P)(B \cup H \models e) \;\; \& \;\; (\forall n \in N)(B \cup H \not\models n).$$

- E.g. to find common pattern (in the linguistic structure) present around every relevant piece of text and none irrelevant.

- Main advantage: multirelational character (B can reside in several relational tables)

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation

2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification

3. Conclusion

4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

Four Main Topics

- **Manual Design of Extraction Rules**
- **Induction of Extraction Rules**
- **Shareable Extraction Ontologies**
- **Fuzzy ILP Document Classification**

Manual Design of Extraction Rules

Slides about the topic *Manual Design of Extraction Rules* will have **brown** headline background.

Induction of Extraction Rules

Slides about the topic *Induction of Extraction Rules* will have **green** headline background.

Shareable Extraction Ontologies

Slides about the topic *Shareable Extraction Ontologies* will have **cyan** headline background.

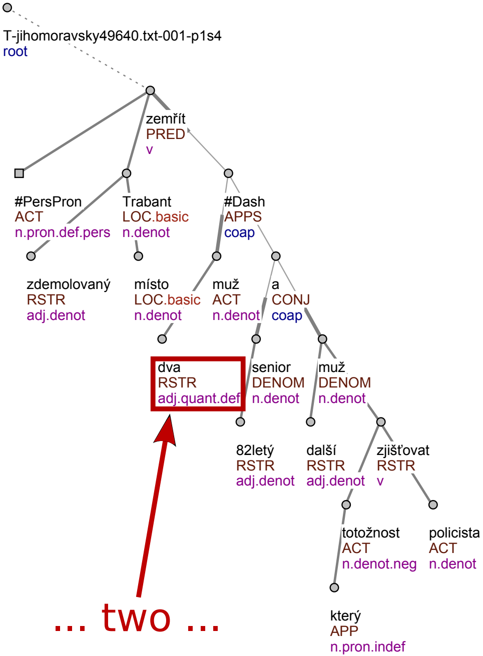
Fuzzy ILP Document Classification

Slides about the topic *Fuzzy ILP Document Classification* will have **magenta** headline background.

Ordinary

Manual Design of Extraction Rules

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE



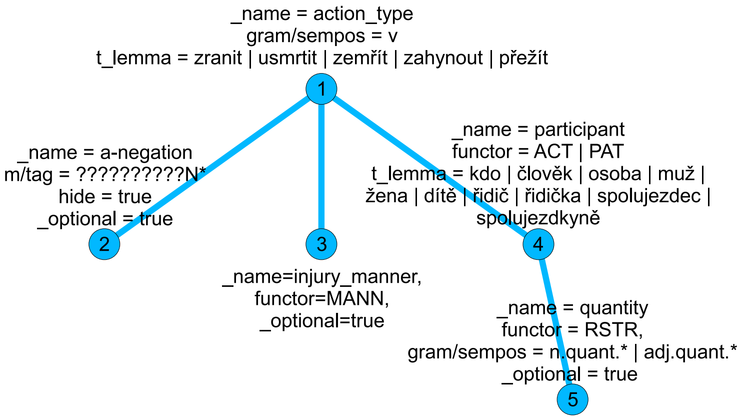
... two ...

- How to extract the information about two dead people?
- *Sentence:*

  Ve zdemolovaném trabantu na místě zemřeli dva muži -- 82letý senior a další muž, jehož totožnost zjišťují policisté.

  Two men died on the spot in demolished trabant -- …

Extraction rules -- Netgraph queries



_name = action_type
gram/sempos = v
t_lemma = zranit | usmrtit | zemřít | zahynout | přežít

_name = a-negation
m/tag = ??????????N*
hide = true
_optional = true

_name=injury_manner,
functor=MANN,
_optional=true

_name = participant
functor = ACT | PAT
t_lemma = kdo | člověk | osoba | muž |
žena | dítě | řidič | řidička | spolujezdec |
spolujezdkyně

_name = quantity
functor = RSTR,
gram/sempos = n.quant.* | adj.quant.*
_optional = true

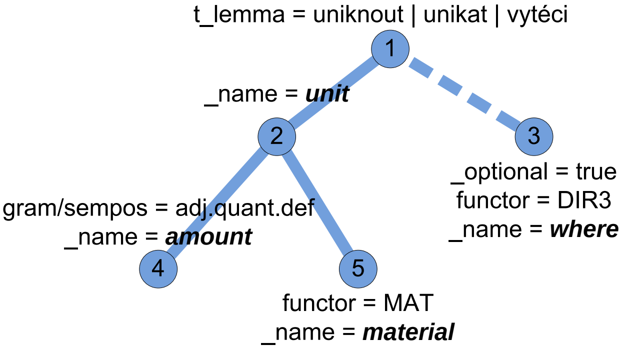
- Tree patterns on shape and nodes (on node attributes).
- Evaluation gives actual matches of particular nodes.
- Names of nodes allow use of references.

Raw data extraction output

```xml
<QueryMatches>
  <Match root_id="T-vysocina63466.txt-001-p1s4" match_string="2:0,7:3,8:4,11:2">
    <Sentence>
      Při požáru byla jedna osoba lehce zraněna - jednalo se
      o majitele domu, který si vykloubil rameno.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="injury_manner" attribute_name="t_lemma">lehký</Value>
      <Value variable_name="participant" attribute_name="t_lemma">osoba</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">jeden</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49640.txt-001-p1s4" match_string="1:0,13:3,14:4">
    <Sentence>
      Ve zdemolovaném trabantu na místě zemřeli dva muži - 82letý senior
      a další muž, jehož totožnost zjišťují policisté.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zemřít</Value>
      <Value variable_name="participant" attribute_name="t_lemma">muž</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">dva</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49736.txt-001-p4s3" match_string="1:0,3:3,7:1">
    <Sentence>Čtyřiatřicetiletý řidič nebyl zraněn.</Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="a-negation" attribute_name="m/tag">VpYS---XR-NA---
      </Value>
      <Value variable_name="participant" attribute_name="t_lemma">řidič</Value>
    </Data>
  </Match>
</QueryMatches>
```

SELECT action_type.t_lemma, a-negation.mtag, injury_manner.t_lemma, participant.t_lemma, quantity.t_lemma **FROM** ****extraction rule****

Extraction rules -- Environment Protection Use Case



t_lemma = uniknout | unikat | vytéci

1

_name = ***unit***

2

3

_optional = true
functor = DIR3
_name = ***where***

gram/sempos = adj.quant.def
_name = ***amount***

4

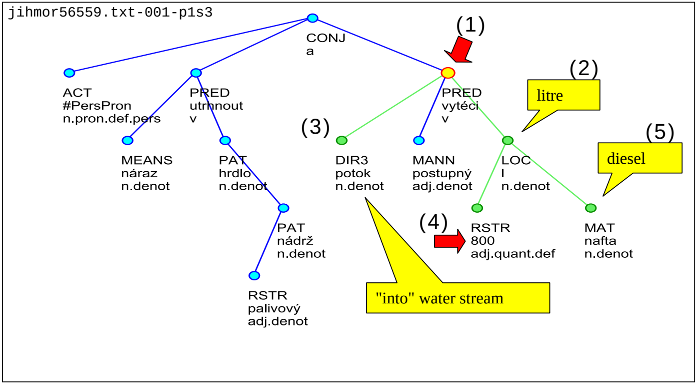
5

functor = MAT
_name = ***material***

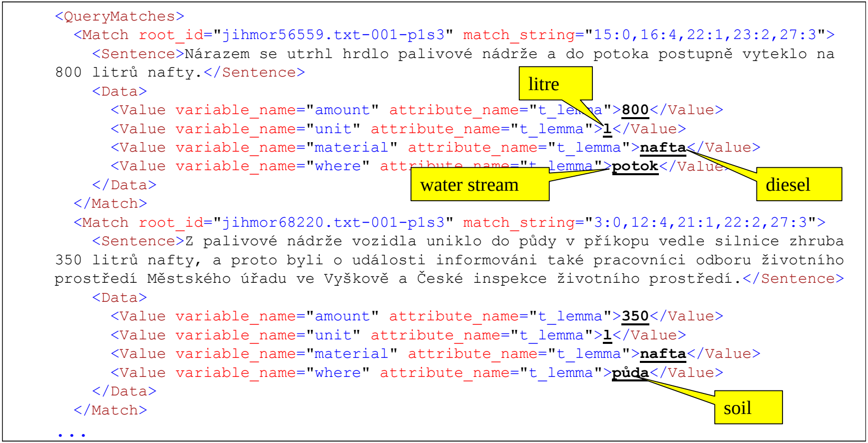
Matching Tree

"Due to the clash the throat of fuel tank tore off and 800 litres of oil (diesel) has run out to a stream."

“Nárazem se utrhl hrdlo palivové nádrže a do potoka postupně vyteklo na 800 litrů nafty.”



Raw data extraction output

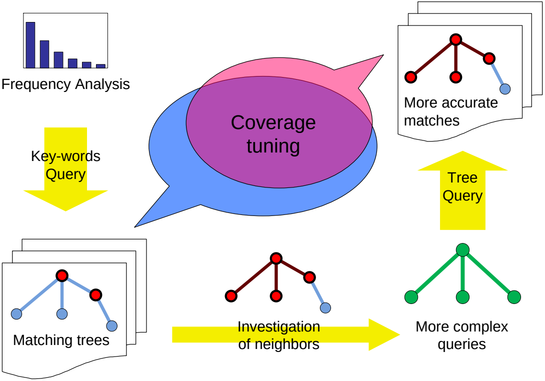


```
<QueryMatches>
 <Match root_id="jihmor56559.txt-001-p1s3" match_string="15:0,16:4,22:1,23:2,27:3">
  <Sentence>Nárazem se utrhl hrdlo palivové nádrže a do potoka postupně vyteklo na
800 litrů nafty.</Sentence>
  <Data>
   <Value variable_name="amount" attribute_name="t_lemma">800</Value>
   <Value variable_name="unit" attribute_name="t_lemma">l</Value>
   <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
   <Value variable_name="where" attribute_name="t_lemma">potok</Value>
  </Data>
 </Match>
 <Match root_id="jihmor68220.txt-001-p1s3" match_string="3:0,12:4,21:1,22:2,27:3">
  <Sentence>Z palivové nádrže vozidla uniklo do půdy v příkopu vedle silnice zhruba
350 litrů nafty, a proto byli o události informováni také pracovníci odboru životního
prostředí Městského úřadu ve Vyškově a České inspekce životního prostředí.</Sentence>
  <Data>
   <Value variable_name="amount" attribute_name="t_lemma">350</Value>
   <Value variable_name="unit" attribute_name="t_lemma">l</Value>
   <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
   <Value variable_name="where" attribute_name="t_lemma">půda</Value>
  </Data>
 </Match>
...
```

SELECT amount.t_lemma, unit.t_lemma, material.t_lemma, where.t_lemma

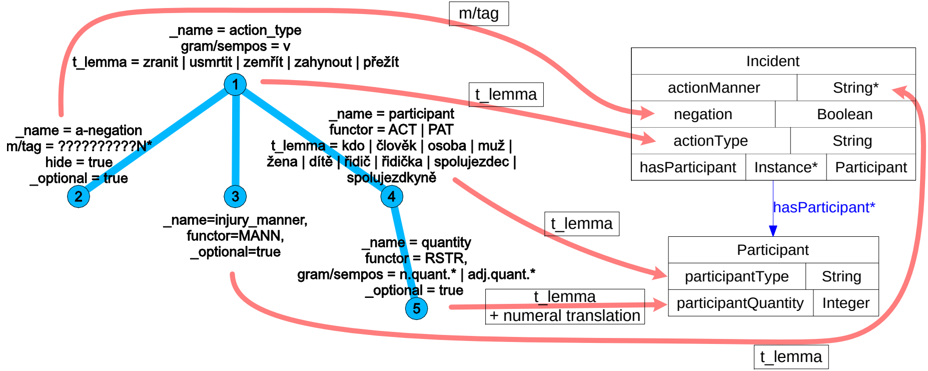
FROM ****extraction rule****

Design of extraction rules -- iterative process



1. Frequency analysis → representative key-words.
2. Investigating of matching trees → tuning of tree query.
3. Complexity of the query ≅ complexity of extracted data.

Semantic interpretation of extraction rules



1
_name = action_type
gram/sempos = v
t_lemma = zranit | usmrtit | zemřít | zahynout | přežít

2
_name = a-negation
m/tag = ??????????N*
hide = true
_optional = true

3
_name=injury_manner,
functor=MANN,
_optional=true

4
_name = participant
functor = ACT | PAT
t_lemma = kdo | člověk | osoba | muž |
žena | dítě | řidič | řidička | spolujezdec |
spolujezdkyně

5
_name = quantity
functor = RSTR,
gram/sempos = n.quant.* | adj.quant.*
_optional = true

Incident		
actionManner	String*	
negation	Boolean	
actionType	String	
hasParticipant	Instance*	Participant

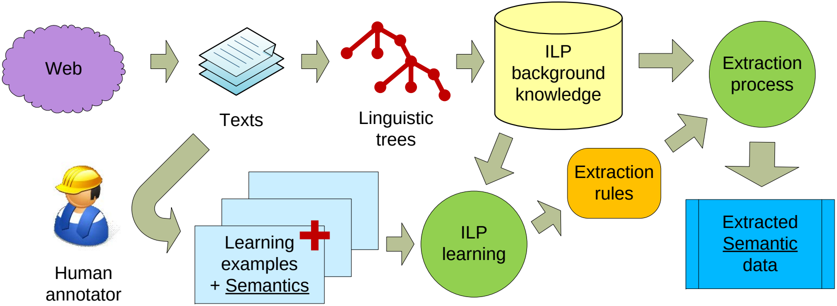
Participant	
participantType	String
participantQuantity	Integer

- Determines how particular values of attributes are used.
- Gives semantics to extraction rule.
- Gives semantics to extracted data.
- Only proposal

Induction of Extraction Rules

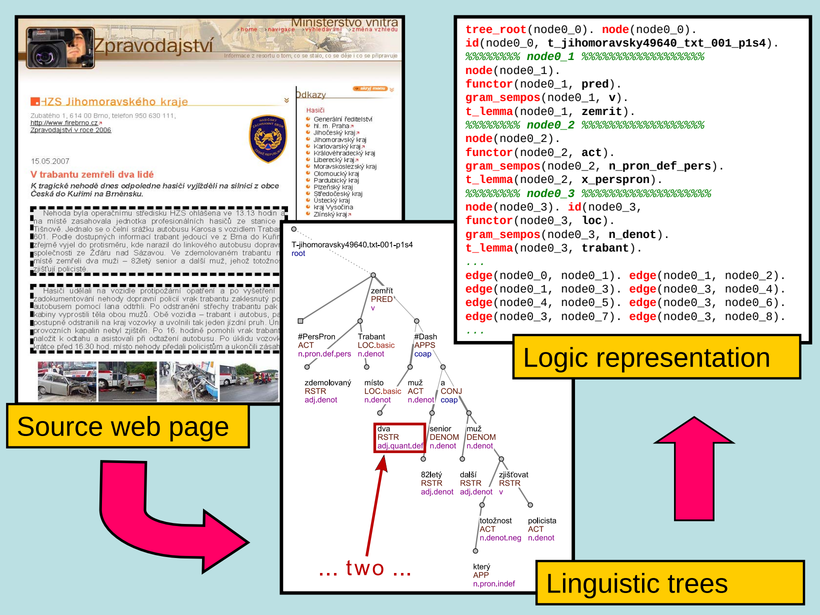
1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

Integration of ILP in our extraction process



- Main point: transformation of trees to logic representation.
- Human annotator does not need to be a linguistic expert.

Logic representation of linguistic trees



Ministerstvo vnitra

›home ›navigace ›vyhledávání ›změna vzhledu

Zpravodajství

Informace z resortu o tom, co se stalo, co se děje i co se připravuje

HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 950 630 111,
http://www.firebrno.cz
Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé

K tragické nehodě dnes odpoledne hasiči vyjížděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a
na místě zasahovala jednotka profesionálních hasičů ze stanice
Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabar
601. Podle dostupných informací trabant jedoucí ve z Brna do Kuřin
zřejmě vyjel do protisměru, kde narazil do linkového autobusu doprav
společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu n
místě zemřeli dva muži – 82letý senior a další muž, jehož totožnos
zjišťují policisté.

Hasiči udělali na vozidle protipožární opatření a po vyšetření
zadokumentování nehody dopravní policií vrak trabantu zaklesnutý po
autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak
kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pa
postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Úni
provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabant
naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovk
krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah

Odkazy

skryj menu

Hasiči

- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

```
tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_p1s4).
%%%%%%%% node0_1 %%%%%%%%%%%%%%%%%%
node(node0_1).
functor(node0_1, pred).
gram_sempos(node0_1, v).
t_lemma(node0_1, zemrit).
%%%%%%%% node0_2 %%%%%%%%%%%%%%%%%%
node(node0_2).
functor(node0_2, act).
gram_sempos(node0_2, n_pron_def_pers).
t_lemma(node0_2, x_perspron).
%%%%%%%% node0_3 %%%%%%%%%%%%%%%%%%
node(node0_3). id(node0_3,
functor(node0_3, loc).
gram_sempos(node0_3, n_denot).
t_lemma(node0_3, trabant).
...
edge(node0_0, node0_1). edge(node0_1, node0_2).
edge(node0_1, node0_3). edge(node0_3, node0_4).
edge(node0_4, node0_5). edge(node0_3, node0_6).
edge(node0_3, node0_7). edge(node0_3, node0_8).
...
```

Rules with largest coverage (Czech fireman dataset)

```
% [cars - Rule 3] [Pos cover = 5 Neg cover = 0]
mention(cars,A) :-
   'lex.rf'(B,A), sempos(B,'n.denot'), tDependency(C,B), t_lemma(C,vozidlo),
   functor(C,'ACT'), number(C,sg).   % vozidlo ~ vehicle
% [damage - Rule 1] [Pos cover = 14 Neg cover = 0]
mention(damage,A) :-
   'lex.rf'(B,A), sempos(B,'n.quant.def'), tDependency(C,B), tDependency(C,D),
   t_lemma(D,'vyšetřovatel').   % vyšetřovatel ~ investigating officer
% [end_subtree - Rule 7] [Pos cover = 6 Neg cover = 0]
mention(end_subtree,A) :-
   'lex.rf'(B,A), sempos(B,'n.quant.def'), tDependency(C,B), t_lemma(C,'ukončit').
       % ukončit ~ finish
% [start - Rule 2] [Pos cover = 15 Neg cover = 0]
mention(start,A) :-
   'lex.rf'(B,A), functor(B,'TWHEN'), tDependency(C,B), tDependency(C,D),
   t_lemma(D,ohlásit).   % ohlásit ~ report (e.g. a fire)
% [injuries - Rule 1] [Pos cover = 7 Neg cover = 0]
mention(injuries,A) :-
   'lex.rf'(B,A), functor(B,'PAT'), tDependency(B,C), t_lemma(C,'zraněný'),
   tDependency(D,B), aspect(D,cpl).   % zraněný ~ injured
% [fatalities - Rule 1] [Pos cover = 3 Neg cover = 0]
mention(fatalities,A) :-
   'lex.rf'(B,A), functor(B,'PAT'), tDependency(C,B), t_lemma(C,srazit).
       % srazit ~ knock down
% [professional_unit - Rule 1] [Pos cover = 17 Neg cover = 0]
mention(professional_unit,A) :-
   'lex.rf'(B,A), functor(B,'LOC'), gender(B,fem), tDependency(C,B),
   functor(C,'CONJ'), overlap_Lookup_tToken(D,B).
% [amateur_unit - Rule 1] [Pos cover = 19 Neg cover = 0]
mention(amateur_unit,A) :-
   'lex.rf'(B,A), tDependency(C,B), tDependency(D,C), tDependency(D,E),
   t_lemma(E,dobrovolný).   % dobrovolný ~ voluntary
```

Evaluation -- Czech fireman -- Precision (optimistic example)

Strict Precision

Task	ILP			PAUM		
cars	0.324	±	0.387	0.380	±	0.249
damage	0.901	±	0.178	0.860	±	0.176
end subtree	0.529	±	0.381	0.499	±	0.242
start	0.929	±	0.109	0.651	±	0.152 ●
injuries	0.667	±	0.291	0.398	±	0.205 ●
fatalities	0.814	±	0.379	0.307	±	0.390 ●
rofessional unit	0.500	±	0.241	0.677	±	0.138 ○
amateur unit	0.863	±	0.256	0.546	±	0.293 ●
overall	0.691	±	0.358	0.540	±	0.297 ●

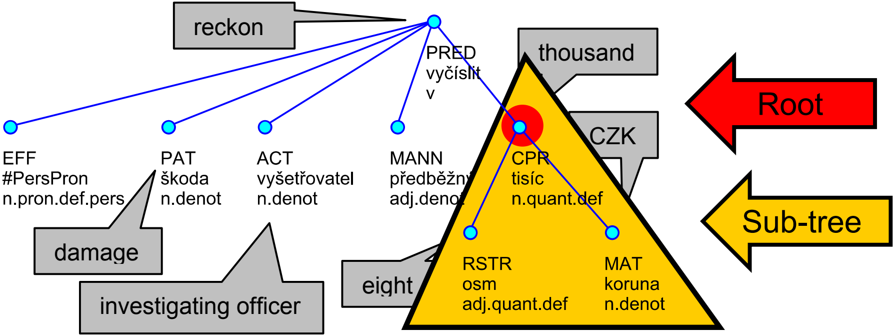
○, ● statistically significant improvement or degradation

Evaluation -- Corporate acquisitions -- Overall

	Annotations		Extraction Method					
Task	ver. A	ver. B	SRV	HMM	Elie	SVM+ILP	ILP	PAUM
acquired	**683**	651	38.5	*30.9*	43.5	41.8	31.3	**47.3**
acqabr	**1450**	1494	38.1	40.1	39.7	42.6	*25.8*	**45.6**
purchaser	**624**	594	45.1	48.1	46.2	45.4	*36.7*	**51.1**
purchabr	1263	**1347**	**48.5**	n/a	28.7	35.4	*17.2*	44.3
seller	267	**707**	23.4	n/a	*15.6*	**51.5**	17.0	23.2
sellerabr	431	**458**	**25.1**	n/a	13.4	21.7	*8.5*	20.2
dlramt	**283**	206	61.8	55.3	59.0	53.0	*28.0*	**65.9**
Total/Overall	5001	**5457**	41.1	n/a	33.5	40.8	*23.9*	**44.0**

- F_1 measure
- Two versions of the dataset (A - white / B - gray)
- Results taken form the literature (except ILP and PAUM)
- ``Baseline experiments'', see also the discussion slide (64) about future experimenting possibilities

Root/Subtree Preprocessing/Postprocessing



…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

…, investigating officer preliminarily reckoned the damage to be eight thousand Crowns (CZK).

- Multi-word expressions

Shareable Extraction Ontologies

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

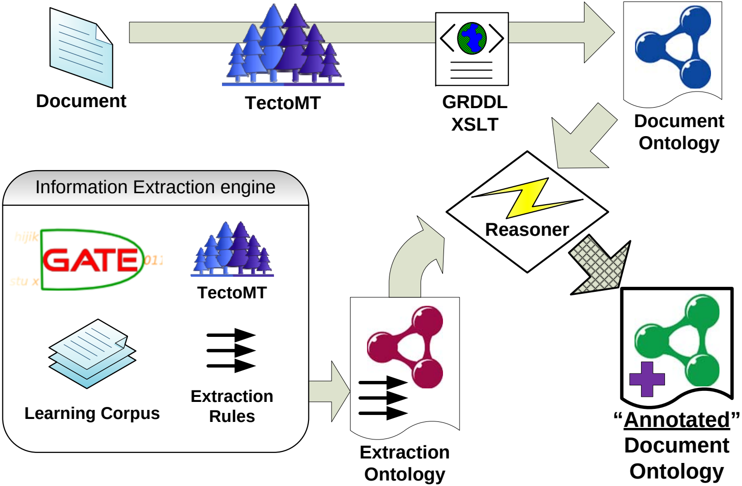
Extraction Ontology

- The knowledge (extraction model) used in the extraction process can itself be saved in an ontology.
  - So called Extraction Ontology

- D. W. Embley, ``Toward semantic understanding: an approach based on information extraction ontologies,'' in *ADC '04*. Darlinghurst: ACS, 2004, pp. 3--12.
- M. Labský et al., ``The Ex Project: Web Information Extraction Using Extraction Ontologies,'' in *Knowledge Discovery Enhanced with Semantic and Social Information*, ser. Studies in Comput. Intellig. Springer, 2009, vol. 220, pp. 71--88.

- But these Extraction Ontologies can only be used with the original tool.
- They are not shareable!

Extraction Rules Interpreted by OWL Reasoner



- Tool independent extraction ontologies

Extraction rules in OWL/XML syntax for Rules

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Ontology [
   <!ENTITY pml "http://ufal.mff.cuni.cz/pdt/pml/" >
]>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
   ontologyIRI="http://czsem.berlios.de/ontologies/...rules.owl">
   <DLSafeRule>
      <Body>
         <ObjectPropertyAtom> <ObjectProperty IRI="&pml;lex.rf"/>
            <Variable IRI="urn:swrl#b"/> <Variable IRI="urn:swrl#a"/>
         </ObjectPropertyAtom>
         <DataPropertyAtom> <DataProperty IRI="&pml;sempos"/>
            <Variable IRI="urn:swrl#b"/> <Literal>n.quant.def</Literal>
         </DataPropertyAtom>
         <ObjectPropertyAtom> <ObjectProperty IRI="&pml;tDependency"/>
            <Variable IRI="urn:swrl#c"/> <Variable IRI="urn:swrl#b"/>
         </ObjectPropertyAtom>
         <ObjectPropertyAtom> <ObjectProperty IRI="&pml;tDependency"/>
            <Variable IRI="urn:swrl#c"/> <Variable IRI="urn:swrl#d"/>
         </ObjectPropertyAtom>
         <DataPropertyAtom> <DataProperty IRI="&pml;t_lemma"/>
            <Variable IRI="urn:swrl#d"/> <Literal>vyšetřovatel</Literal>
         </DataPropertyAtom>
      </Body>
      <Head>
         <DataPropertyAtom> <DataProperty IRI="&pml;mention_root" />
            <Literal>damage</Literal> <Variable IRI="urn:swrl#a" />
         </DataPropertyAtom>
      </Head>
   </DLSafeRule>
</Ontology>
```

Extraction rules in Protégé

```
#[Rule 1]
lex.rf(?b, ?a), sempos(?b, "n.quant.def"), tDependency(?c, ?b),
tDependency(?c, ?d), t_lemma(?d, "vyšetřovatel") #investigator
      -> mention_root(?a, "damage")

#[Rule 2]
lex.rf(?b, ?a), functor(?b, "TOWH"), tDependency(?c, ?b),
tDependency(?c, ?d), t_lemma(?d, "škoda") #damage
      -> mention_root(?a, "damage")
```

Extraction rules in Jena

```
@prefix pml: <http://ufal.mff.cuni.cz/pdt/pml/>.
[rule-1:
        ( ?b pml:lex.rf ?a )
        ( ?b pml:sempos 'n.quant.def' )
        ( ?c pml:tDependency ?b )
        ( ?c pml:tDependency ?d )
        ( ?d pml:t_lemma 'vyšetřovatel' )
     ->
        ( ?a pml:mention_root 'damage' )
]
```

Performance Evaluation -- Datasets & Reasoners

dataset	domain	language	num of files	data size (MB)	num of rules
czech_fireman	accidents	Czech	50	16	2
acquisitions	finance	English	600	126	113

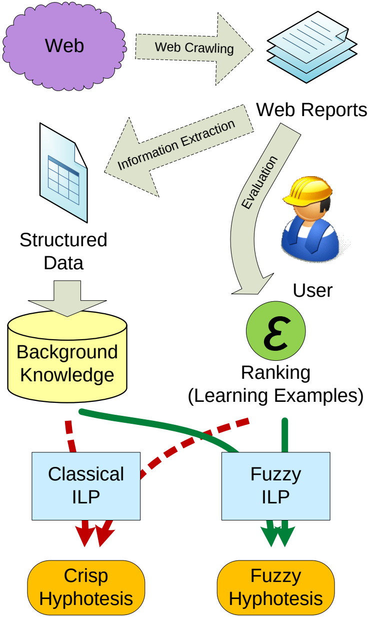
reasoner	**czech_fireman**	stdev	**acquisitions-v1.1**	stdev
Jena	161 s	0.226	1259 s	3.579
HermiT	219 s	1.636	≫ 13 hours	
Pellet	11 s	0.062	503 s	4.145
FaCT++	Does not support rules.			

- Poor performance...
- Because these tools are not optimized for these taks (yet?)

Fuzzy ILP Document Classification

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

Schema of the whole system



1. Web Crawling
2. Information Extraction and User Evaluation
3. Logic representation
   - Construction of background knowledge
   - Construction of learning examples
4. ILP Learning
   - Crisp
   - Fuzzy
5. Comparison of results

Essential difference between learning examples

Crisp learning examples

```
serious_2(id_47443). %positive

serious_0(id_47443). %negative
serious_1(id_47443). %negative
serious_3(id_47443). %negative
```

Monotonized learning examples

```
serious_atl_0(id_47443). %positive
serious_atl_1(id_47443). %positive
serious_atl_2(id_47443). %positive

serious_atl_3(id_47443). %negative
```

- For one evidence (occurrence, e.g. one accident)
- Crisp:
  Always one positive and three negative learning examples
- Monotonized:
  Up to the observed degree positive,
  the rest negative.

Monotonization of attributes

damage_atl ← damage

```
damage_atl(ID,N) :- damage(ID,N), not(integer(N)). %unknown values

damage_atl(ID,N) :- damage(ID,N2), integer(N2), %numeric values
                    damage(N), integer(N), N2>=N.
```

- We infer all lower values as sufficient.
- Treatment of unknown values.
- Negation as failure.

Rules for the whole Czech fireman dataset

```
% Crisp
serious_0(A) :- fatalities(A,0), injuries(A,0), cars(A,1),
                amateur_units(A,0), lather(A,0).
serious_0(A) :- fatalities(A,0), cars(A,0), amateur_units(A,0),
                professional_units(A,1).
serious_1(A) :- amateur_units(A,1).
serious_1(A) :- damage(A,300000).
serious_1(A) :- type(A,fire), amateur_units(A,0), pipes(A,2).
serious_1(A) :- type(A,car_accident), dur_minutes(A,unknown),
                fatalities(A,0), injuries(A,1).
serious_2(A) :- lather(A,unknown).
serious_2(A) :- cars(A,0), lather(A,0), aqualung(A,1), fan(A,0).
serious_2(A) :- amateur_units(A,2).
serious_3(A) :- fatalities(A,2).
serious_3(A) :- type(A,fire), dur_minutes(A,unknown), cars(A,0), fan(A,0).
serious_3(A) :- injuries(A,2), cars(A,2).
serious_3(A) :- fatalities(A,1).

% Monotonized
serious_atl_0(A).
serious_atl_1(A) :- injuries_atl(A,1).
serious_atl_1(A) :- dur_minutes_atl(A,21), pipes_atl(A,1), aqualung_atl(A,0).
serious_atl_1(A) :- damage_atl(A,8000), amateur_units_atl(A,3).
serious_atl_1(A) :- dur_minutes_atl(A,197).
serious_atl_1(A) :- dur_minutes_atl(A,unknown).
serious_atl_2(A) :- dur_minutes_atl(A,50), pipes_atl(A,3).
serious_atl_2(A) :- size_atl(A,1364), injuries_atl(A,1).
serious_atl_2(A) :- fatalities_atl(A,1).
serious_atl_2(A) :- size_atl(A,1106), professional_units_atl(A,3).
serious_atl_3(A) :- fatalities_atl(A,1).
serious_atl_3(A) :- damage_atl(A,1500000).
```

Conversion of Results

serious_t ← serious_atl_t (selecting maximum)

```
serious_0(ID) :- serious_atl_0(ID),
                 not(serious_atl_1(ID)), not(serious_atl_2(ID)),
                 not(serious_atl_3(ID)).
serious_1(ID) :- serious_atl_1(ID),
                 not(serious_atl_2(ID)), not(serious_atl_3(ID)).
serious_2(ID) :- serious_atl_2(ID),
                 not(serious_atl_3(ID)).
serious_3(ID) :- serious_atl_3(ID).
```

Evaluation -- Czech fireman dataset

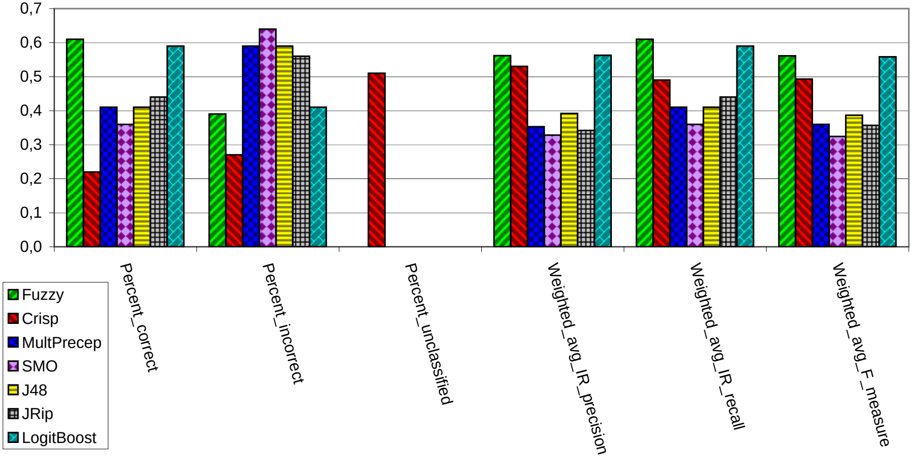
	Fuzzy	Crisp	MultPerc	SMO	J48	JRip	LBoost
Corr	0.61±.19	.22±.17 ●	.41±.19 ●	.36±.24 ●	.41±.22 ●	.44±.17 ●	.59±.26
Incor	.39±.19	.27±.24	.59±.19 ○	.64±.24 ○	.59±.22 ○	.56±.17 ○	.41±.26
Uncl	.00±.00	.51±.29 ○	.00±.00	.00±.00	.00±.00	.00±.00	.00±.00
Prec	.56±.24	.53±.37	.35±.20 ●	.33±.26	.39±.22	.34±.21 ●	.56±.28
Rec	.61±.19	.49±.32	.41±.19 ●	.36±.24 ●	.41±.22 ●	.44±.17 ●	.59±.26
F	.56±.20	.49±.33	.36±.19 ●	.32±.24 ●	.39±.21	.36±.19 ●	.56±.27

○, ● statistically significant improvement or degradation

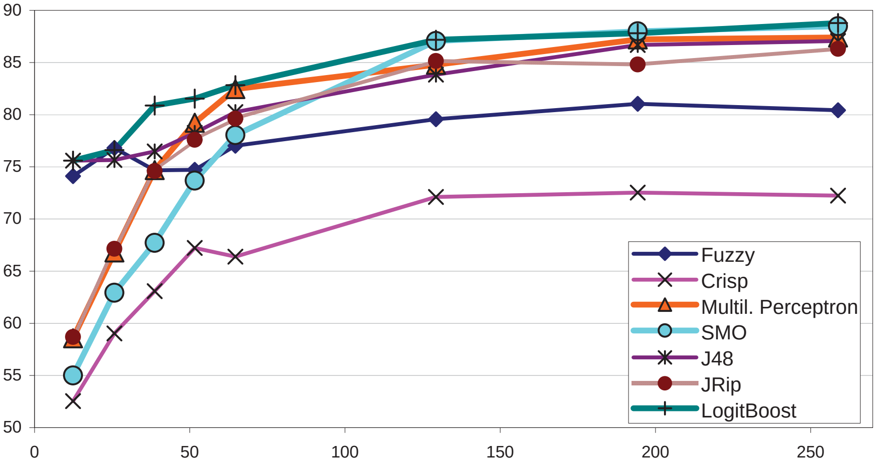
Fuzzy	czsem.ILP.FuzzyILPClassifier
Crisp	czsem.ILP.CrispILPClassifier
MultPerc	functions.MultilayerPerceptron
SMO	functions.SMO
J48	trees.J48
JRip	rules.JRip
LBoost	meta.LogitBoost

Corr	Percent correct
Inor	Percent incorrect
Uncl	Percent unclassified
Prec	Weighted avg IR precision
Rec	Weighted avg IR recall
F	Weighted avg F measure

Evaluation -- Czech fireman dataset



The impact of dataset size on classification performance



- `nursery' dataset from UCI ML Repository
- x-axis: number of training instances
- y-axis: percent of correctly classified instances
- average values from 10 repetitions

1. Introduction
   - Information Extraction
   - Deep Language Parsing
   - Inductive Logic Programming
   - Organization of this Presentation
2. Contents
   - Manual Design of Extraction Rules
   - Induction of Extraction Rules
   - Shareable Extraction Ontologies
   - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
   - Review 1 (Filip Železný)
   - Review 2 (Diana Maynard)
   - PDT Annotations in GATE

Conclusion

- Transparent Data Mining
  - Learned rules understandable & adjustable by humans
  - Tool Independence (Shareable Extraction Ontologies)
- Software
  - Nontrivial extensive implementation (c.f. slide 66)
  - Netgraph based IE API (declarative extraction rules)
  - Framework for IE experiments (based on GATE+Weka)
  - TectoMT (+ Treex) module for GATE
    - PML $\longrightarrow$ GATE
  - Netgraph Tree Viewer for GATE
  - Shareable Extraction Ontologies Prototype
    - PML $\longrightarrow$ RDF (XSLT $\approx$ GRDDL)
  - Fuzzy ILP Classifier for Weka
- Evaluation Experiments
  - Repeatable (software & data freely available)
  - Difficult (c.f. slide 64)
  - Direct comparison with the state-of-the-art

1. Introduction
    - Information Extraction
    - Deep Language Parsing
    - Inductive Logic Programming
    - Organization of this Presentation
2. Contents
    - Manual Design of Extraction Rules
    - Induction of Extraction Rules
    - Shareable Extraction Ontologies
    - Fuzzy ILP Document Classification
3. Conclusion
4. **Questions and Comments from Reviews**
    - Review 1 (Filip Železný)
    - Review 2 (Diana Maynard)
    - PDT Annotations in GATE

The Title

- Nobody is happy with the title!
  - Including the author himself …
- But it is quite difficult to find better one.
- ``Use of deep language parsing for generation of extraction rules''
  - The last two topics are not cowered
- ``How ILP and ontologies can help information extraction''
  - The first and the last topic is not cowered

1 Introduction

- Information Extraction
- Deep Language Parsing
- Inductive Logic Programming
- Organization of this Presentation

2 Contents

- Manual Design of Extraction Rules
- Induction of Extraction Rules
- Shareable Extraction Ontologies
- Fuzzy ILP Document Classification

3 Conclusion

4 Questions and Comments from Reviews

- Review 1 (Filip Železný)
- Review 2 (Diana Maynard)
- PDT Annotations in GATE

ILP Definitions

In 4.7.1.,.2 : it is not clear what really is B, in particular whether it is a set of only ground facts or it allows more general clauses.

- Answer: *B* allows also more general clauses.
- Unfortunately Definition 3 counts only with ground facts, which is sufficient for our purpose, but it is not explained there.
- More complex version of Definition 3 could probably be used for more general clauses.

The Classification Problem

Is the problem of classification of accident severity really relational?

- Answer: No, it is not!
- All the attributes are saved in a single table.
- And all evaluated classification algorithms (mainly propositional) were given the same data.

ILP Performance

> *I was not happy with the simple conclusion that ILP does not perform very well in the extraction rule learning task.*

- Answer: not just ILP, but ILP with PDT trees

> *... a deeper investigation should be done ... The tDependency predicate is often employed in the learned rules in a chain-like manner. Would it help to define its transitive closure in the background knowledge?*

- Answer: That is a reasonable comment, thank you for suggestions.
- We experimented only with ``optional edge'' predicates, but did not register much improvement.
- For sure, additional experiments would be beneficial, but the experiments are demanding and there are other three topics in the thesis...
- C.f. slides 64 and 65 about *future experimenting possibilities* and *available resources criterion*

ILP Performance

There are many possible sources of the low performance, for example the way numeric attributes are handled. For example the learned condition damage_atl(A,1500000) does not seem to make too much sense since it clearly overfits to the one specific value. Were predicates such as greater_than(X,Y) available to the learner at all?

- Answer: Yes, see slide 45.

ILP Performance

> *Finally, Aleph is a general purpose system with a large set of parameters (proof depth, clause length, search method, ...), which influence the final performance.*

- Question: Did you try to change them at all?.
- Answer: Yes, see e.g. page 97:

  > *Learning settings are the same as before (Table 8.12) except for settings of both ILP classifiers, which performed a little bit better with modified settings.*

  See the learning configuration files for extraction rules induction.

- Question: Did you tune their best values through internal validation?
- Answer: No, it would take quite a lot of additional work...

1 Introduction

- Information Extraction
- Deep Language Parsing
- Inductive Logic Programming
- Organization of this Presentation

2 Contents

- Manual Design of Extraction Rules
- Induction of Extraction Rules
- Shareable Extraction Ontologies
- Fuzzy ILP Document Classification

3 Conclusion

4 Questions and Comments from Reviews

- Review 1 (Filip Železný)
- Review 2 (Diana Maynard)
- PDT Annotations in GATE

Topics / Questions / State-of-the-art

- The thesis consists of four equally important topics
- We cannot concentrate on a single one (e.g. **Induction of Extraction Rules)**

- There are many questions in the review
  - Let us go through them, they all can be resolved!

- It was really had work to prepare and evaluate the baseline method
- And make it directly comparable with the state-of-the-art solutions
- The framework is now open to everybody
- Almost infinite amount of additional experiments can be performed
  - See also the discussion slide (64) about future experimenting possibilities and their complexity

The Task of Information Extraction

> *The task of IE should be clearly defined, making clear what it involves, why it is needed, and why it is hard. ... Sections on IE components ... are much too short. ... I would expect these issues to be covered and discussed in depth, with appropriate references (these issues have all been discussed many times in the literature), ...*

- Is it really necessary to discuss all these topics?
    - It concerns only 2 of 4 main topics of the thesis.
    - We are not creating a text book, are we?
    - What is the contribution in it?
- What reference is actually missing?
- Is the used terminology lacking something important?

Dědek	MUC-6 1995, Appelt & Israel 1999
Entity Recognition	Named Entity Recognition
Relation Extraction	Template Element Construction
Event Extraction	Template Relation Construction
Event Extr. Encoded as Ent. Rec.	Template Unification
Instance Resolution	Scenario Template Production

Experimenting Possibilities and Experiment Complexity

- All the possibilities form one wide Cartesian product of the factors:
  - Used evaluation measures
  - Number of datasets and extraction tasks
  - Chosen number of training instances
  - Preprocessing / Postprocessing
    - Linguistic representation (PDT-m/a/t, Stanford, CoNLL’X, trimming)
    - Used tagger, lemmatization, parser, POS generalization (of Czech tags)
    - Named entity recognition (Stanford, different settings of ANNIE)
    - Additional gazetteers, parsers for dates, money amounts, ...
    - Multi-word (root/subtree, merging, begging/end tokens)
  - Learning settings
    - ILP tool (Progol, Aleph, ReLF) or even Constraint Programming, ILP settings,
    - ILP predicates: transitive closures, optional edges
    - Position based features (window size)
  - Additional resources (e.g. WordNet, domain ontology)
  - Coreference resolution
  - Propositionalization

Available Resources Criterion: Time, Effort, Allocated Capabilities

- One common answer to comments like:
  - ``Chapter, section, etc. is too short.''
  - ``Problem, solution, etc. should be more discussed.''
  - ``The techniques could easily be described and motivated in much more detail.''
  - ``More examples should be given.''
  - ``Evaluation dataset is rather too small.''
- The answer is:
  - Yes, that is reasonable comment, but there were no more available resources for it.
  - Is the work as a whole too short?
  - Are there parts that should have been omitted?

  - We did our best to include the most important and relevant things.
  - But then, oops, the time was up!
- Let's look at this in more detail on the next slides...

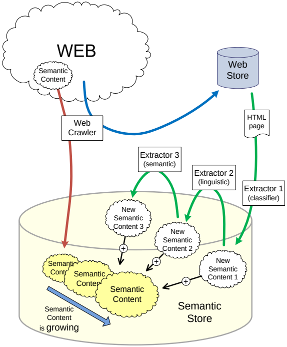
Work Performed -- Implementation

- Nontrivial extensive implementation
- Use and integration of following tools and technologies:
  - Linguistics
    - PDT 2.0 analysis tools + TectoAnalysis by Václav Klimeš
    - TectoMT (Treex currently also supported)
    - Perl/brted programming of first *procedural* extraction rules
    - Netgraph by Jiří Mírovský, *declarative* extraction rules
    - GATE
  - Semantic Web
    - OWL API + Pellet, HermiT and FaCT++
    - Jena (including Jena Rules)
    - SweetRules
    - PML → RDF (OWL) transformation (XSLT ≈ GRDDL)
    - ILP Extraction rules → SWRL transformation
  - Data Mining
    - **ILP** (Progol, Prolog + Aleph):
      Integration with GATE (IE Rules Induction) and
      Weka (Fuzzy ILP Classifier)
    - **Weka:** Fuzzy ILP Classifier and Statistical significance of GATE experiments
  - XML RPC (Perl server, Java client)

Work Performed -- Other

- Construction (or contribution) of new datasets:
    - Czech Fireman Reports without Annotations
    - Czech Fireman Reports Manually Annotated
    - RDF Dataset Based on Czech Fireman Reports
    - RDF Dataset Based on Corporate Acquisition Events
    - Classification Dataset Based on Czech Fireman Reports

- Evaluation experiments
    - **Direct** comparison with state-of-the-art
- Publications:
    - Including E-Environment and Economics (Crisis prediction)
- Development of the idea of **Web Semantization**
    - Finally not included in the thesis
    - But published in selected papers

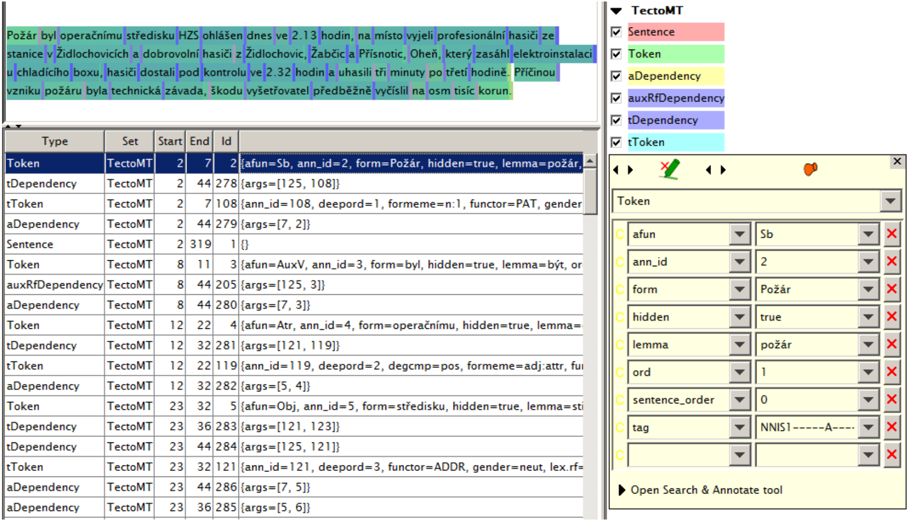


1. Introduction
    - Information Extraction
    - Deep Language Parsing
    - Inductive Logic Programming
    - Organization of this Presentation
2. Contents
    - Manual Design of Extraction Rules
    - Induction of Extraction Rules
    - Shareable Extraction Ontologies
    - Fuzzy ILP Document Classification
3. Conclusion
4. Questions and Comments from Reviews
    - Review 1 (Filip Železný)
    - Review 2 (Diana Maynard)
    - PDT Annotations in GATE

Integration of PDT in GATE

- GATE: General Architecture for Text Engineering
- The University of Sheffield
- `http://gate.ac.uk/`
- Implemented TectoMT (+ Treex) modules for GATE
  - Transformation of PDT annotations to GATE
- Netgraph used as a tree viewer
  - Works also for Standford Dependencies

PDT Annotations in GATE



Požár byl operačnímu středisku HZS ohlášen dnes ve 2.13 hodin, na místo vyjeli profesionální hasiči ze stanice v Židlochovicích a dobrovolní hasiči z Židlochovic, Žabčic a Přísnotic. Oheň, který zasáhl elektroinstalaci u chladícího boxu, hasiči dostali pod kontrolu ve 2.32 hodin a uhasili tři minuty po třetí hodině. Příčinou vzniku požáru byla technická závada, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

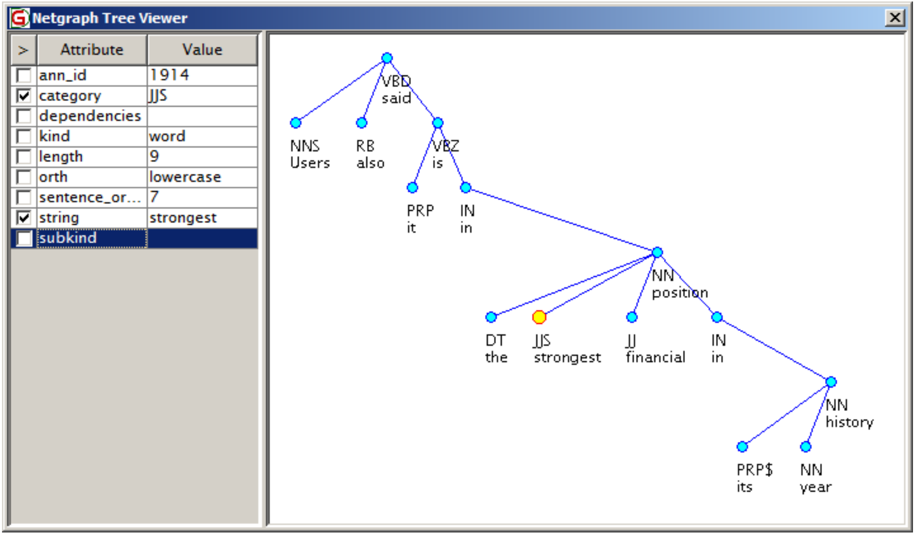
Type	Set	Start	End	Id	
Token	TectoMT	2	7	2	{afun=Sb, ann_id=2, form=Požár, hidden=true, lemma=požár,
tDependency	TectoMT	2	44	278	{args=[125, 108]}
tToken	TectoMT	2	7	108	{ann_id=108, deepord=1, formeme=n:1, functor=PAT, gender
aDependency	TectoMT	2	44	279	{args=[7, 2]}
Sentence	TectoMT	2	319	1	{}
Token	TectoMT	8	11	3	{afun=AuxV, ann_id=3, form=byl, hidden=true, lemma=být, or
auxRfDependency	TectoMT	8	44	205	{args=[125, 3]}
aDependency	TectoMT	8	44	280	{args=[7, 3]}
Token	TectoMT	12	22	4	{afun=Atr, ann_id=4, form=operačnímu, hidden=true, lemma=
tDependency	TectoMT	12	32	281	{args=[121, 119]}
tToken	TectoMT	12	22	119	{ann_id=119, deepord=2, degcmp=pos, formeme=adj:attr, fu
aDependency	TectoMT	12	32	282	{args=[5, 4]}
Token	TectoMT	23	32	5	{afun=Obj, ann_id=5, form=středisku, hidden=true, lemma=stř
tDependency	TectoMT	23	36	283	{args=[121, 123]}
tDependency	TectoMT	23	44	284	{args=[125, 121]}
tToken	TectoMT	23	32	121	{ann_id=121, deepord=3, functor=ADDR, gender=neut, lex.rf=
aDependency	TectoMT	23	44	286	{args=[7, 5]}
aDependency	TectoMT	23	36	285	{args=[5, 6]}

- TectoMT
  - ☑ Sentence
  - ☑ Token
  - ☑ aDependency
  - ☑ auxRfDependency
  - ☑ tDependency
  - ☑ tToken

Token

afun	Sb
ann_id	2
form	Požár
hidden	true
lemma	požár
ord	1
sentence_order	0
tag	NNIS1-----A----

Netgraph Tree Viewer in GATE (for Stanford Dependencies)


Netgraph Tree Viewer

>	Attribute	Value
☐	ann_id	1914
☑	category	JJS
☐	dependencies	
☐	kind	word
☐	length	9
☐	orth	lowercase
☐	sentence_or...	7
☑	string	strongest
☐	subkind	

Sentence: Users also said it is in the strongest financial position in its 24-year history.