

Web Information Extraction for eEnvironment

We need help from domain experts!

Jan Dědek

Peter Vojtáš

Department of Software Engineering, Charles University in Prague
Institute of Computer Science, Czech Academy of Sciences

Outline

- Introduction – signal data versus human produced data
 - Information processing top down versus bottom up
 - Web and the environment, Web Information Extraction
- Our extraction framework
 - Web → Text → Linguistics → Structured Data
 - Extraction rules – Linguistic tree patterns
- Learning of our tool
 - **Human** learning and **Machine** learning
- Examples of extracted data
- Conclusion
 - We are interested in cooperation!

The Environment and Information on the Web

sensors/monitoring **versus** human created information

top down (egov) **versus** bottom up processing



World

Real existence of



Dangers to the Environment

Partially Reflected on the Web



WWW

Extraction



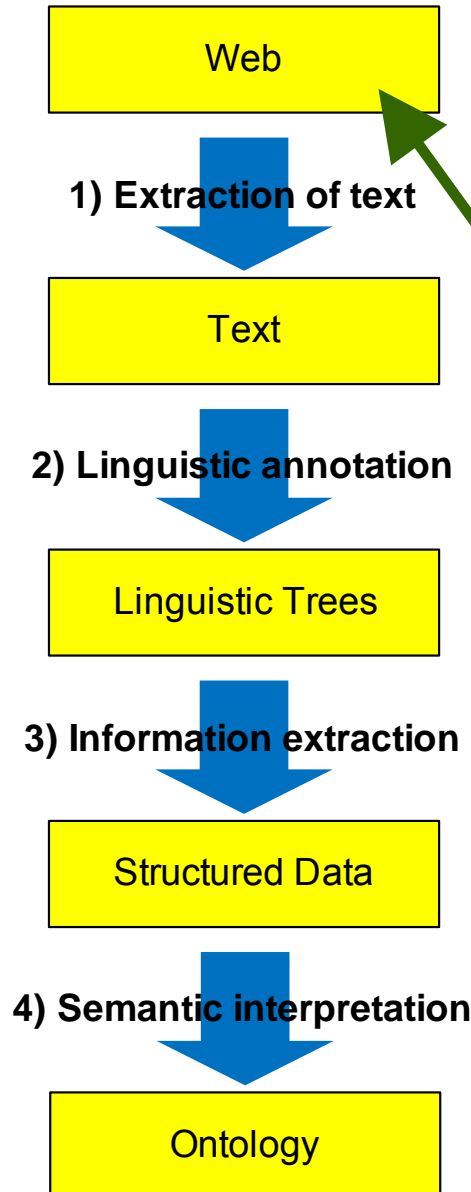
Web Information Extraction
Tool

Partial Evidence of
Dangers

Our Web Information Extraction Framework

- Motivated by the Semantic Web
 - Transformation of **human** understandable resources to **machine** understandable ones
- Information extraction from **texts** (from the Web)
 - Supported by **linguistic tools** (PDT, Netgraph, ...)
- Machine learning of the extraction procedure
 - Human annotated **training set** needed
- Structured (and semantic) **extraction output**
 - Structured output: XML or Ontology data structure
 - Automatically annotated web pages

The Data Flow



Domain expert has to:

- Select relevant pages
- **Support learning procedure**

Learning procedure produces rules for

- Data extraction
- Data interpretation

Example of processed web page

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 950 830 111,
<http://www.firebrno.cz>
Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé

K tragické nehodě dnes odpoledne hasiči vyjžděli 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 v Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 801. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní společnosti ze Zďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

Hasiči ud zadokumentovali zničený autobus a z kabiny vyprošťovali postupně od provozních kapalin. Naložit k od krátkce před 1

„Únik provozních kapalin nebyl zjištěn.“
“Outflow of Sealing liquid was not found out.”

odkazy

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

Ua:1509.cz

V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství
- RSS

Servis nejen pro novináře
Schengenská spolupráce
WebEditorial

Na našem serveru v jiných rubrikách

- Aktuality Národního archivu

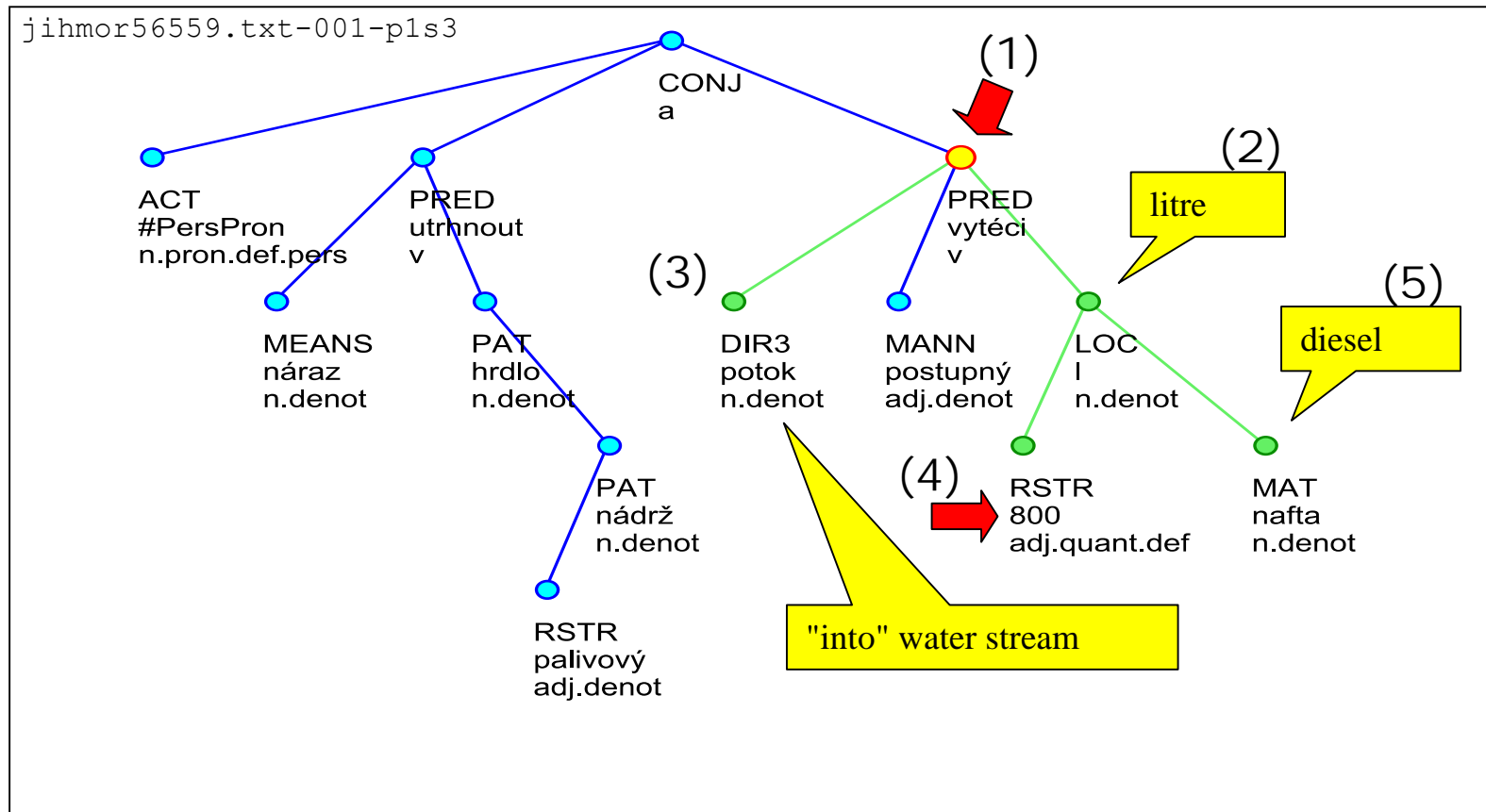
Relevant text

Information relevant to the environment

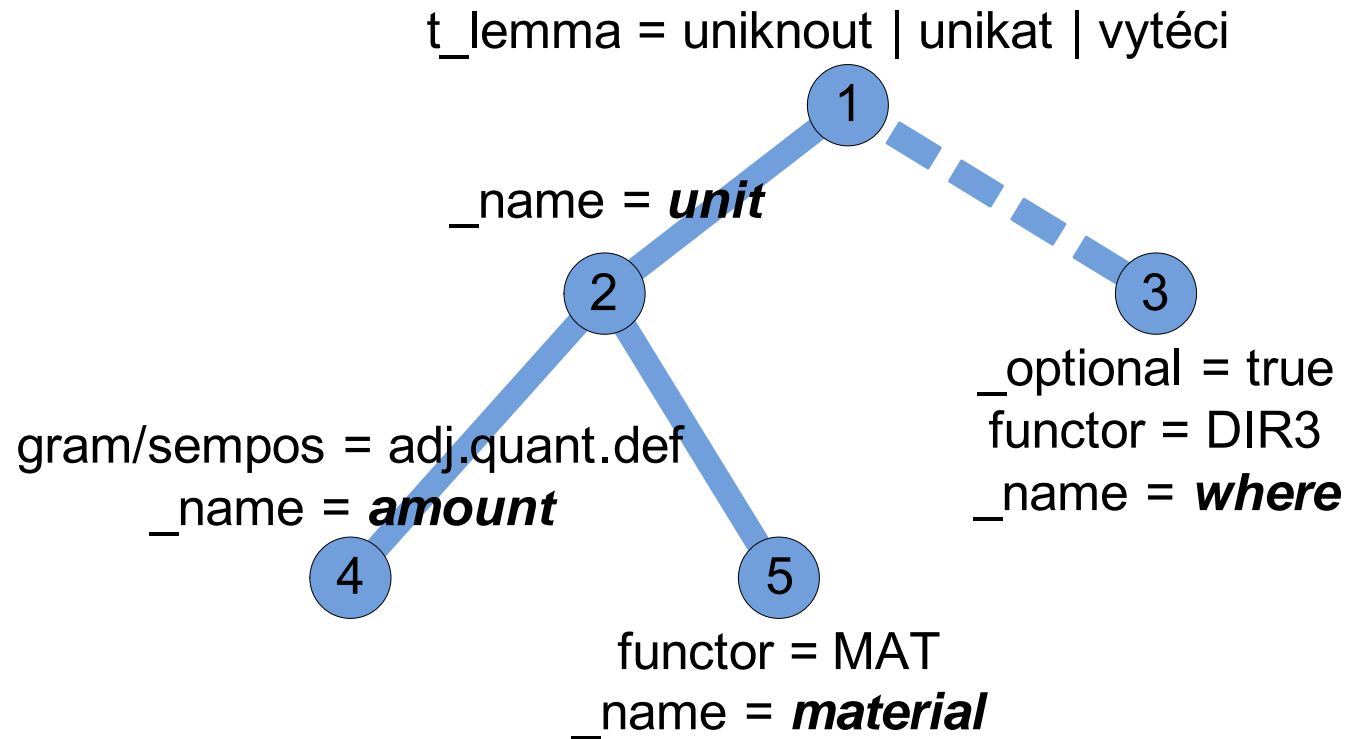
Example of a linguistic tree

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

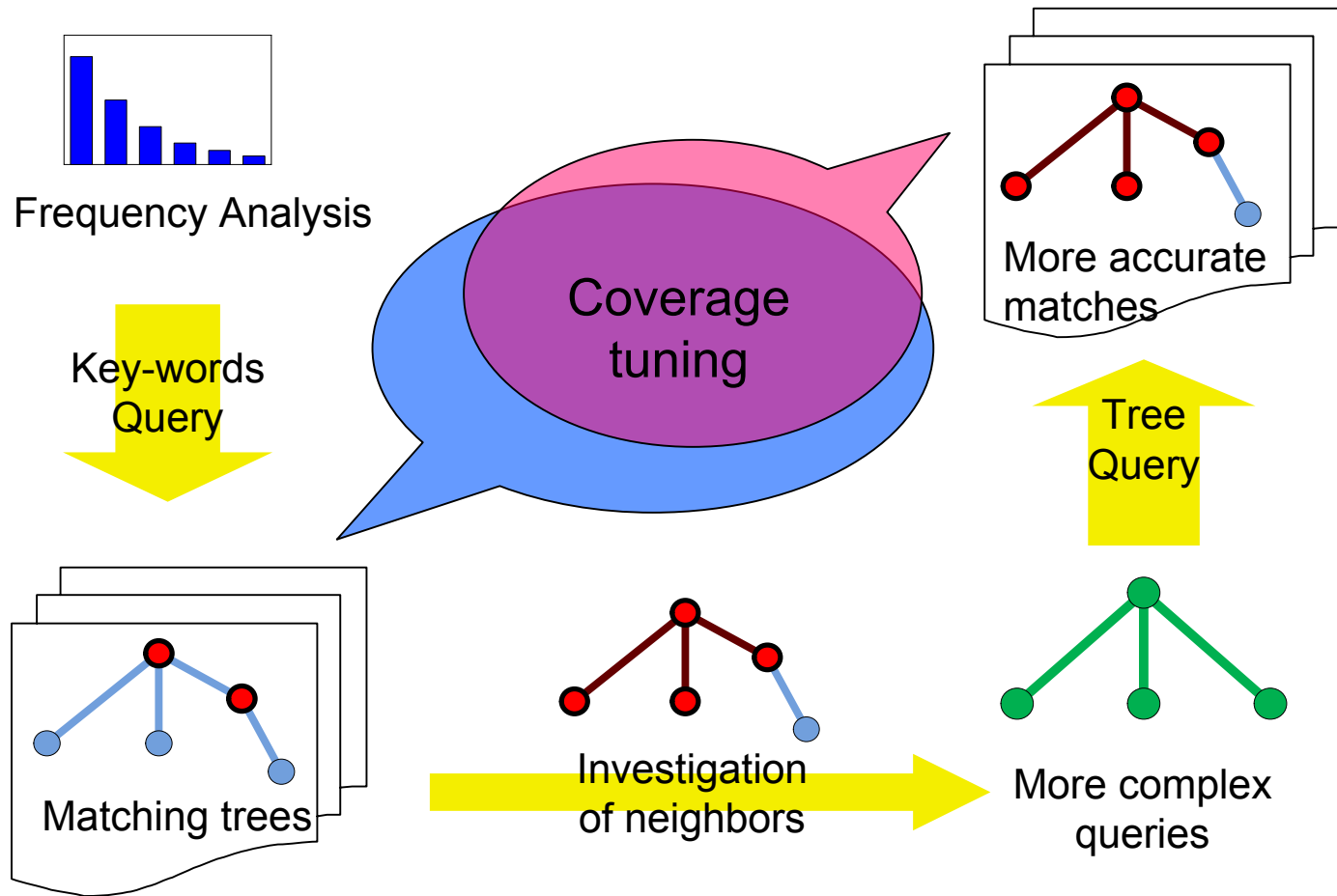
"Nárazem se utrhlo hrdlo palivové nádrže a do potoka postupně vyteklo na 800 litrů nafty."



Example of an extraction rule.

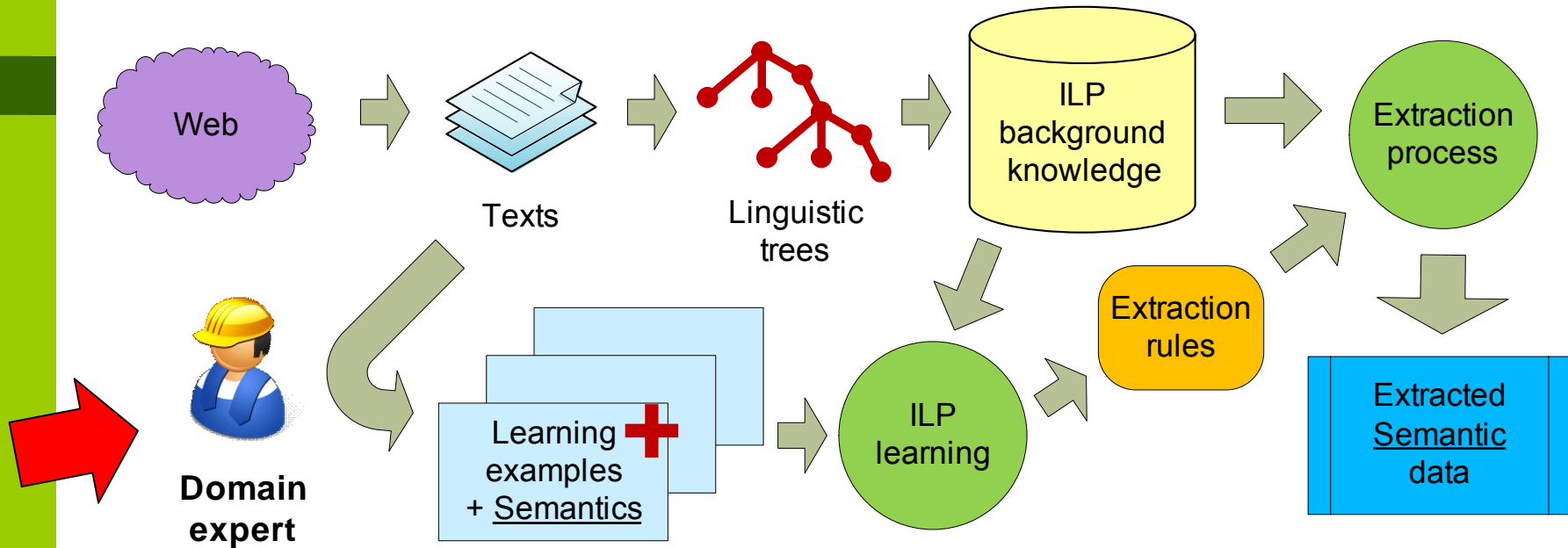


Manual learning of extraction rules



- Needs **expert** in both: in the **domain** and in **linguistics**

Machine learning of extraction rules



- Use of **Inductive Logic Programming**
- Learning examples
 - Marked directly in the text
 - Done by **domain expert**, no linguistic knowledge needed

Experimental results (1)

```
<QueryMatches>
  <Match root_id="jihmor56559.txt-001-pls3" match_string="15:0,16:4,22:1,23:2,27:3">
    <Sentence>Nárazem se utrhł 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">1</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-pls3" 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">1</Value>
      <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
      <Value variable_name="where" attribute_name="t_lemma">půda</Value>
    </Data>
  </Match>
```

litre

water stream

diesel

soil

Experimental results (2)

```
...
<Match root_id="kralovehrad54765.txt-001-p6s5" match_string="1:0,7:1,8:2,13:3">
  <Sentence>Z kamionu uniklo zhruba 20 litrů látky.</Sentence>
  <Data>
    <Value variable_name="amount" attribute_name="t_lemma">20</Value>
    <Value variable_name="unit" attribute_name="t_lemma">1</Value>
    <Value variable_name="material" attribute_name="t_lemma">látka</Value>
  </Data>
</Match>
<Match root_id="moravslez50487.txt-001-p4s1" match_string="43:0,49:1,50:2,55:3">
  <Sentence>Hasiči po likvidaci požáru trávy asi na 25 metrech čtverečních ještě
uklízeli společně s pracovníky Správy silnic Moravskoslezského kraje zhruba 15 metrů
silnice, na kterou vyteklo asi 40 litrů hydraulického oleje.</Sentence>
  <Data>
    <Value variable_name="amount" attribute_name="t_lemma">40</Value>
    <Value variable_name="unit" attribute_name="t_lemma">1</Value>
    <Value variable_name="material" attribute_name="t_lemma">olej</Value>
  </Data>
</Match>
</QueryMatches>
```

other material

gear oil

What is interesting on the Web?

- For **environment specialists**?
- What information from the Web can help with the **evidence, inspection** and **care** for the environment?
- Perhaps our method can provide it!
- We are interested in **cooperation**!

Concluding Appeal

- We need your help!
- Please **send us examples** of resources, content interesting for environmentalists!

Dedek@ksi.mff.cuni.cz

Vojtas@ksi.mff.cuni.cz

Thank you!