Fuzzy Classification of Web Reports with Linguistic Text Mining

Jan Dědek[1,2] Peter Vojtáš[1,2]

[1] Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

[2] Institute of Computer Science, Academy of Sciences of the Czech Republic

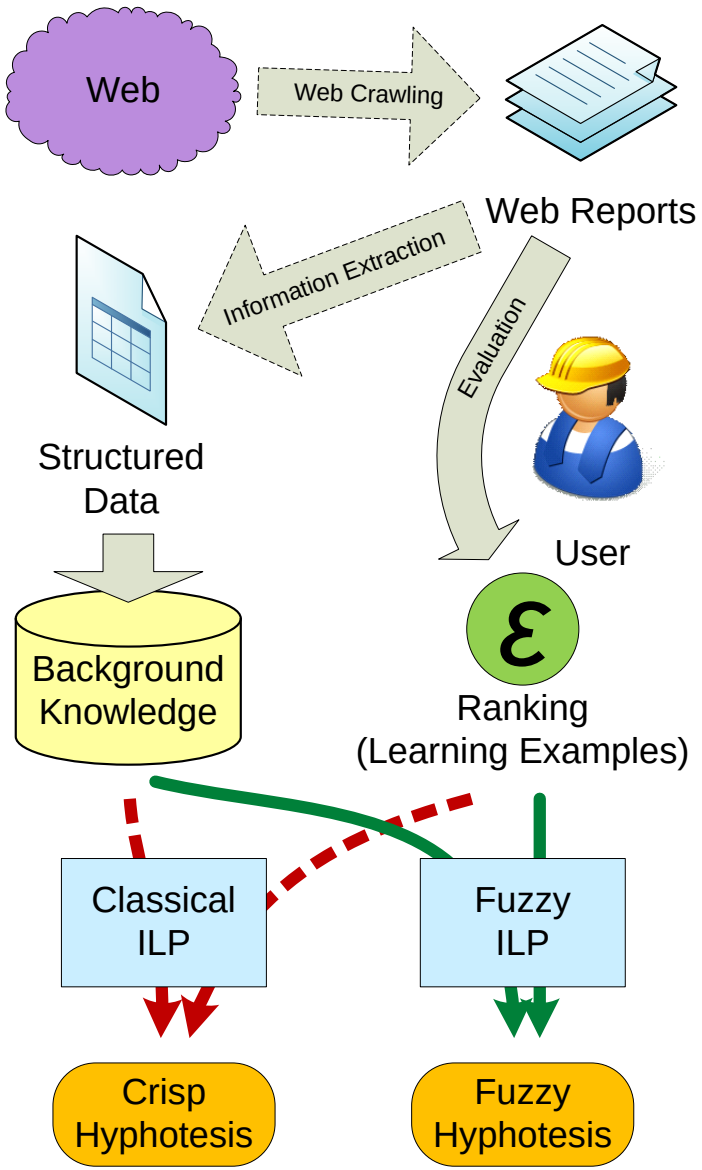
Soft approaches to information access on the Web,
Web Intelligence 2009, 15 – 18 September 2009
Università degli Studi di Milano Bicocca, Milano, Italy

Outline

Our work

- Extraction of semantic information form texts.
  - In Czech language.
  - Coming form web pages.
- Computing aggregations
  - From extracted semantic data.
- Using of Semantic Web ontologies.
  - RDF, OWL
- Exploiting of linguistic tools.
  - Mainly from the Prague Dependency Treebank project.
  - Experiments with the Czech WordNet.
- Rule based extraction method.
  - Extraction rules $\approx$ tree queries of Netgraph application

Schema of the wohle system



- Item1
- Item2

Sentence:

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

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

skrýt menu

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 v Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 601. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

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




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

Euroskop.cz

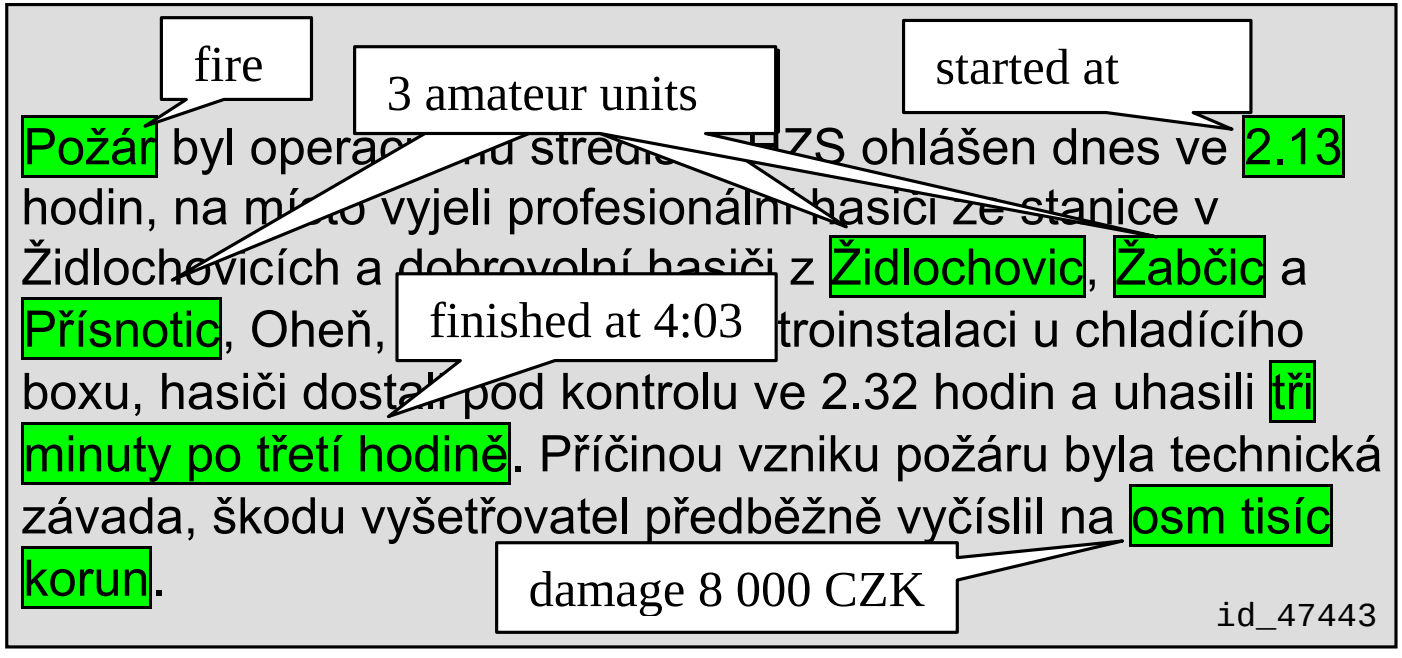
V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Policie
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

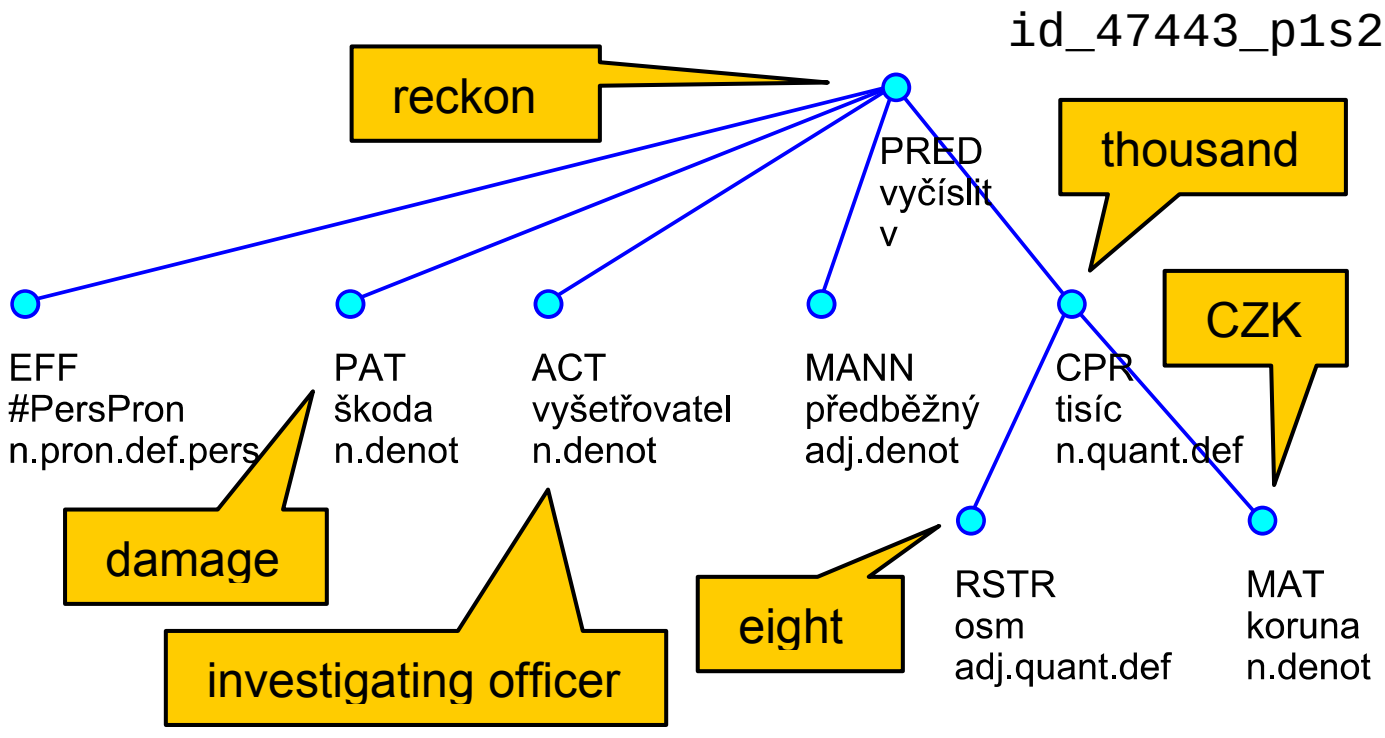
- Aktuality Národního archivu

Example of processed text



- See the last snetence on the next slide.

Example of a linguistic tree



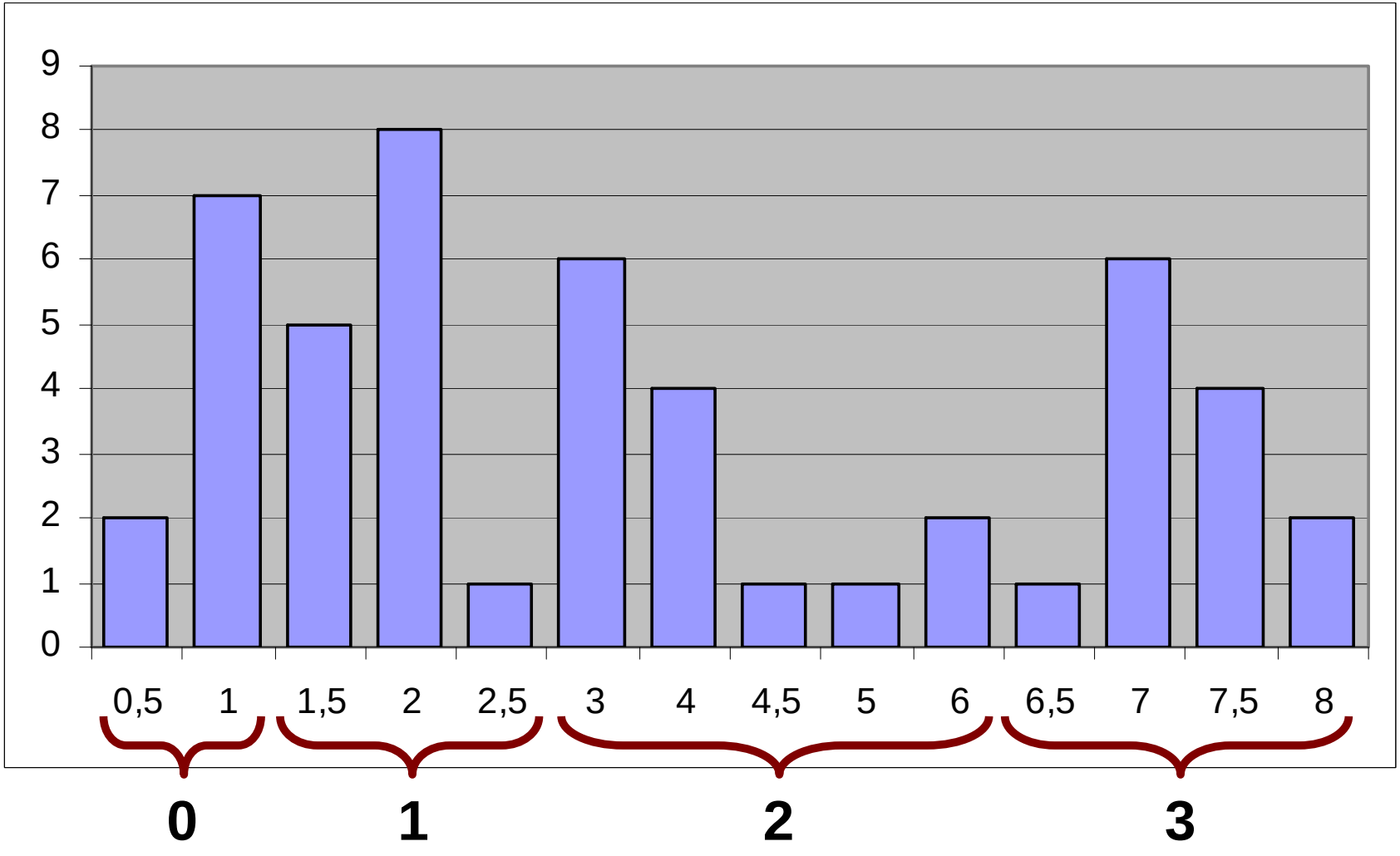
…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

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

Accident attributes

attribute name	distinct values	missing values	monotonic
size (of file)	49	0	yes
type (of accident)	3	0	no
damage	18	30	yes
dur_minutes	30	17	yes
fatalities	4	0	yes
injuries	5	0	yes
cars	5	0	yes
amateur_units	7	1	yes
profesional_units	6	1	yes
pipes	7	8	yes
lather	3	2	yes
aqualung	3	3	yes
fan	3	2	yes
ranking	14	0	yes

Histogram of ranking attribute



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
```

- Item1
- Item2

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
```

Sentence:

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

Monotonization of attributes

- Item1
- Item2

damage → damage_atl

```
damage_atl(ID,N) :- %unknown values
        damage(ID,N), not(integer(N)).
damage_atl(ID,N) :- %numeric values
        damage(ID,N2), integer(N2),
        damage(N), integer(N), N2>=N.
```

Sentence:

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

```
serious_0(A):-dur_minutes(A,8).
serious_0(A):-type(A,fire),pipes(A,0).
serious_0(A):-fatalities(A,0),pipes(A,1),lather(A,0).
serious_1(A):-amateur_units(A,1).
serious_1(A):-amateur_units(A,0),pipes(A,2),aqualung(A,1).
serious_1(A):-damage(A,300000).
serious_1(A):-damage(A,unknown),type(A,fire),prof_units(A,1).
serious_1(A):-dur_minutes(A,unknown), fatalities(A,0), cars(A,1).
serious_2(A):-lather(A,unknown).
serious_2(A):-lather(A,0), aqualung(A,1), fan(A,0).
serious_2(A):-amateur_units(A,2),prof_units(A,2).
serious_2(A):-dur_minutes(A,unknown),injuries(A,2).
serious_3(A):-fatalities(A,1).
serious_3(A):-fatalities(A,2).
serious_3(A):-injuries(A,2), cars(A,2).
serious_3(A):-pipes(A,4).

serious_atl_0(A).
serious_atl_1(A):-injuries_atl(A,1).
serious_atl_1(A):-lather_atl(A,1).
serious_atl_1(A):-pipes_atl(A,3).
serious_atl_1(A):-dur_minutes_atl(A,unknown).
serious_atl_1(A):-size_atl(A,764),pipes_atl(A,1).
serious_atl_1(A):-damage_atl(A,8000),amateur_units_atl(A,3).
serious_atl_1(A):-type(A,car_accident).
serious_atl_1(A):-pipes_atl(A,unknown), randomized_order_atl(A,35).
serious_atl_2(A):-pipes_atl(A,3), aqualung_atl(A,1).
serious_atl_2(A):-type(A,car_accident), cars_atl(A,2),prof_units_atl(A,2).
serious_atl_2(A):-injuries_atl(A,1),prof_units_atl(A,3),fan_atl(A,0).
serious_atl_2(A):-type(A,other), aqualung_atl(A,1).
serious_atl_2(A):-dur_minutes_atl(A,59), pipes_atl(A,3).
serious_atl_2(A):-injuries_atl(A,2),cars_atl(A,2).
serious_atl_2(A):-fatalities_atl(A,1).
serious_atl_3(A):-fatalities_atl(A,1).
serious_atl_3(A):-dur_minutes_atl(A,unknown),pipes_atl(A,3).
```

Crisp & monotonized hypothesis

- Item1
- Item2

Evaluation results

		Raw ILP	Monot. ILP
Monot. test set	TP:	42	57
positive: 64	FP:	7	6
negative: 36	Precision:	0,857	0,905
sum: 100	Recall:	0,656	0,891
	F-measure:	0,743	0,898
Crisp test set	TP:	12	15
positive: 25	FP:	13	10
negative: 75	Precision:	0,480	0,600
sum: 100	Recall:	0,480	0,600
	F-measure:	0,480	0,600

Conversion of results

crisp → monotone

```
serious_2(ID) :- serious_atl_2(ID),
                 not(serious_atl_3(ID)).
```

monotone → crisp

```
serious_atl_0(ID) :- serious_2(ID).
serious_atl_1(ID) :- serious_2(ID).
serious_atl_2(ID) :- serious_2(ID).
```