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

Doktorandské dny Ústavu informatiky AV ČR, v. v. i.,
21. – 23. září 2009, Jizerka

[*] Originally presented at:
the SAIAW workshop of the Web Intelligence 2009 conference, Milano, Italy

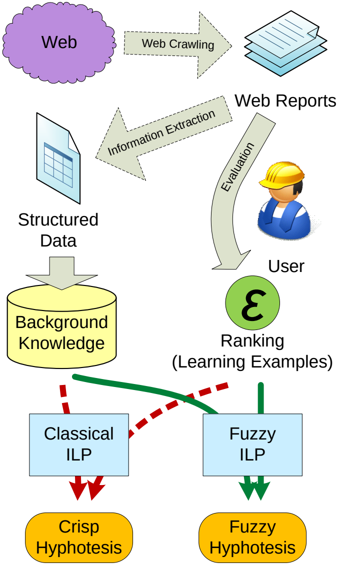
Outline

1. **Introduction**
   - Our Information Extraction System
   - Description of the extraction method
   - Fuzzy ILP
2. **Our Experiment**
   - Experiment Description
3. **Fuzzy ILP / GAP Implementation**
   - Monotonization
4. **Evaluation and Conclusion**
   - Learning Results
   - Evaluation
   - Conclusion

Our work

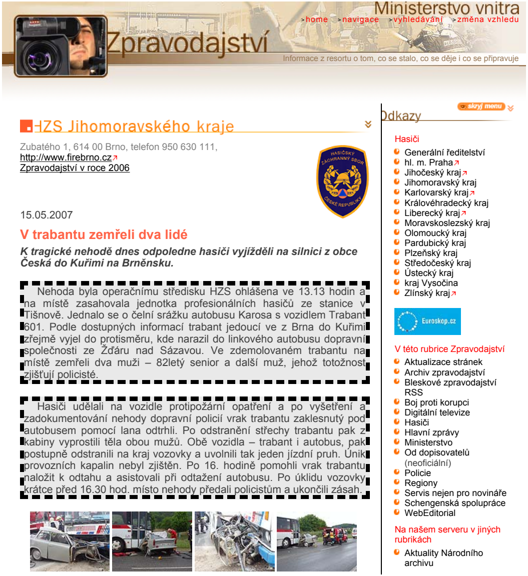
- Extraction of semantic information form texts.
  - In Czech language.
  - Coming form web pages.
- Using of Semantic Web ontologies.
  - RDF, OWL
- Exploiting of linguistic tools.
  - Mainly from the Prague Dependency Treebank project.
  - Experiments with the Czech WordNet.
- Rule based extraction method.
  - Extraction rules ≈ tree queries
  - ILP learning of extraction rules
- Fuzzy report classification
  - Application of Fuzzy ILP
  - Accident seriousness classification
  - Exploitation of extracted information

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

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
skryj 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

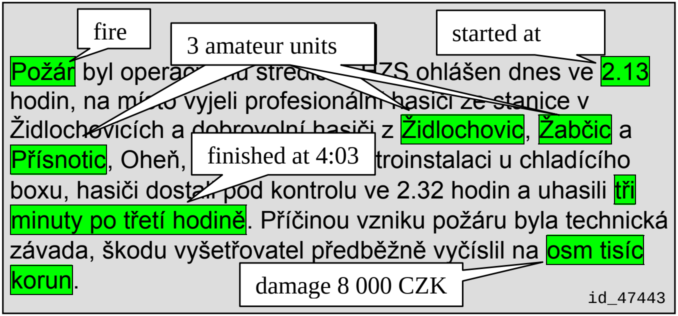
- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Policie
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

- Aktuality Národního archivu

- Fire and car accidents reports

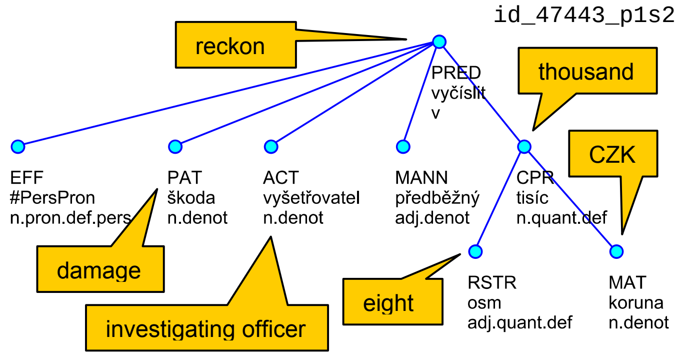
Example of processed text



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.

- Information to be extracted is decorated.
- See the last sentence on the next slide.

Example of a linguistic tree

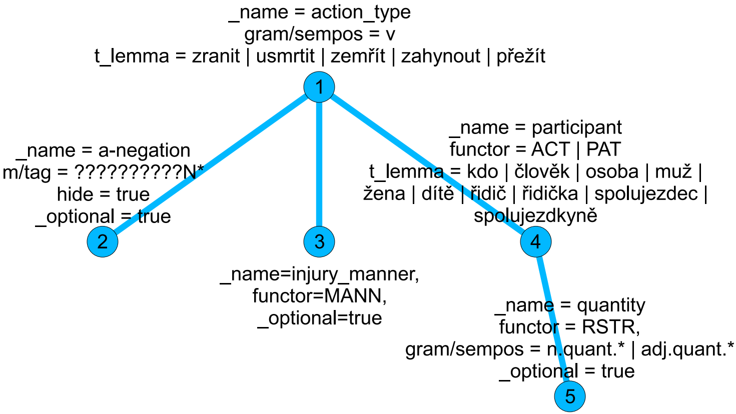


…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

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

- Our IE method uses tree queries (tree patterns)

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

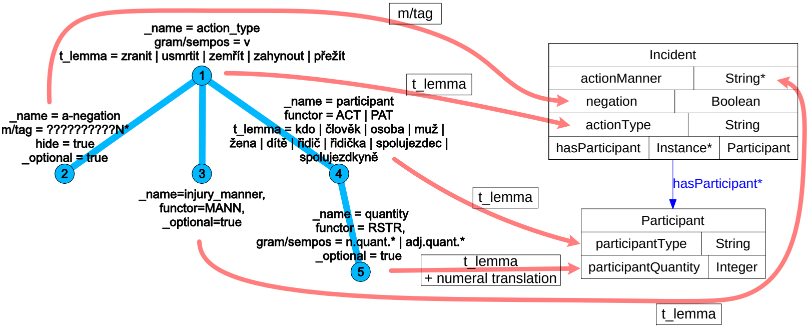
_name=injury_manner,
functor=MANN,
_optional=true

_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.

Semantic interpretation of extraction rules



m/tag

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

1

t_lemma

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

2

3

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

4

_name=injury_manner,
functor=MANN,
_optional=true

t_lemma

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

5

t_lemma
+ numeral translation

t_lemma

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

hasParticipant*

Participant	
participantType	String
participantQuantity	Integer

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

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

- Information that we can/could extract from a report.
- Not everything is always mentioned.

Classical ILP and Fuzzy ILP principles

- Learning examples $E = P \cup N$ (Positive and Negative)
- Background knowledge B
- ILP task – to find hypothesis H such that:

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

- Fuzzy learning examples $\mathcal{E} : E \longrightarrow [0, 1]$
- Fuzzy background knowledge $\mathcal{B} : B \longrightarrow [0, 1]$
- Fuzzy ILP task – to find hyp. $\mathcal{H} : H \longrightarrow [0, 1]$ such that:

$$(\forall e_1, e_2 \in E)(\forall \mathcal{M})(\mathcal{M} \models_f \mathcal{B}\cup\mathcal{H}) \; : \; \mathcal{E}(e_1) > \mathcal{E}(e_2) \Rightarrow \|e_1\|_{\mathcal{M}} \geq \|e_2\|_{\mathcal{M}}$$

Generalized Annotated Programs

- Fuzzy ILP is equivalent to Induction of Generalized Annotated Programs[1]
- For implementation we use GAP or strictly speaking: *Definite Logic Programs with monotonicity axioms* (also equivalent)
- Basic paradigm: deal with values as with degrees.
  - We don't have to normalize values, they order is enough.
- For example with monotonicity axioms we can use rule:

  `serious(A, 4)` ← `fatalities(A, 10).`

  and form the fact `fatalities(id_123, 1000)` deduce

  `serious_alt(id_123, 4).`

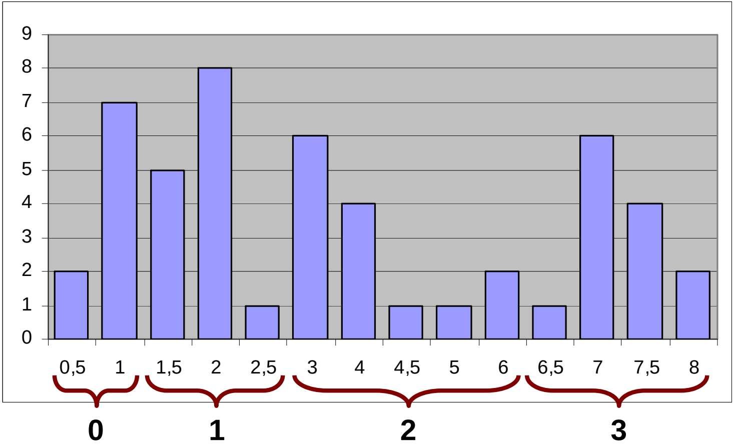
[1] See in S. Krajci, R. Lencses and P. Vojtas: "A comparison of fuzzy and annotated logic programming", Fuzzy Sets and Systems, vol.144, pp.173–192, 2004.

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

- Almost all attributes are numeric.
  - So monotonic
  - This will be used for “fuzzyfication”
- Artificial target attribute seriousness ranking.

Histogram of the seriousness ranking attribute



- 14 different values, range 0.5 – 8
- Divided into four approximately equipotent groups.

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):

- 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) :- %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.
```

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

- Crisp hypothesis

```
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).
```

- Monotonized hypothesis
  - Monotonicity axioms
  - Monotonized learning examples

```
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).
```

Evaluation and Comparison of 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

- Rules evaluated on both testing sets.
    - By use of conversion predicates (next slide)
- Monotonized rules better in both cases.

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).
```

Conclusion

- We used Fuzzy/GAP ILP in an experiment closely connect with WIE.
- Showed basic principles and implementation of Fuzzy/GAP ILP.
- Compared results of Fuzzy/GAP ILP and Classical ILP.
- Observed much better results in the Fuzzy case.

- Future work:
  - Improvement of the extraction method
  - Other languages, other domains
  - Finer “approximatization” of target attribute (not only “four degrees”).