Charles University in Prague
Faculty of Mathematics and Physics

Abstract of Doctoral Thesis

Mgr. Jan Dědek

Semantic Annotations

Department of Software Engineering

Supervisor: Prof. RNDr. Peter Vojtáš, DrSc.

2012

Doktorská práce byla vypracována v rámci doktorského studia, které uchazeč absolvoval na Katedře softwarového inženýrství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze v letech 2007–2012.

Uchazeč:
Mgr. Jan Dědek
Katedra softwarového inženýrství
MFF UK
Malostranské nám. 25, 118 00 Praha 1

Obor studia:
I-2 – Softwarové systémy

Předseda oborové rady:
Prof. Ing. František Plášil, DrSc.
Katedra softwarového inženýrství
MFF UK
Malostranské nám. 25, 118 00 Praha 1

Školitel:
Prof. RNDr. Peter Vojtáš, DrSc.
Katedra softwarového inženýrství
MFF UK
Malostranské nám. 25, 118 00 Praha 1

Oponenti:
Dr. Diana Maynard
Department of Computer Science
University of Sheffield, United Kingdom
Regent Court, 211 Portobello, Sheffield S1 4DP
e-mail: diana@dcs.shef.ac.uk

Doc. Ing. Filip Železný, Ph.D.
Katedra Kybernetiky, Fakulta elektrotechnická
České vysoké učení technické v Praze
Karlovo náměstí 13
121 35 Praha 2
e-mail: zelezny@fel.cvut.cz

Autoreferát byl rozeslán dne 24.8.2012.
Obhajoba se koná dne 24.9.2012 ve 13:00 hodin před komisí pro obhajoby disertačních prací v oboru I-2 – Softwarové systémy na Matematicko-fyzikální fakultě Univerzity Karlovy, Malostranské nám. 25, Praha 1, v místnosti S1.
S disertací je možné se seznámit na studijním oddělení MFF UK, Ke Karlovu 3, Praha 2.

1 Introduction

Four relatively separate topics are presented in the thesis and the discipline of Information Extraction is the central point of them. Each topic represents one particular aspect of the Information Extraction discipline.

The first two topics are focused on our information extraction methods based on deep language parsing. The first topic relates to how deep language parsing was used in our first method in combination with manually designed extraction rules.

The second topic deals with an alternative extraction method based on machine learning. An inductive procedure was developed based on Inductive Logic Programming (ILP), which allows automated learning of extraction rules from a learning collection.

The idea of the Semantic Web was the strongest motivation of our research from the very beginning. We wanted to exploit information extraction techniques to speed up the semantic web evolution. The third topic of the thesis presents even more than that. The core of the extraction method was experimentally reimplemented using semantic web technologies. Therefore not only the result of information extraction but also the extraction procedure itself is realized using semantic web technologies. The main advantage of this approach is the possibility to save the extraction rules in so called shareable extraction ontologies.

The last topic of the thesis is the most distant from the original information extraction topic. We have included it because it represents an important part of our research and considerable effort was spent on it. The topic deals with document classification and fuzzy logic. We are investigating the possibility of using information obtained by information extraction techniques to document classification. Our implementation of so called Fuzzy ILP Classifier was experimentally used for the purpose of document classification.

1.1 Motivation

The basic motivation of our research can be illustrated with three images or schemas that are presented in Figures 1, 2 and 3. The first two figures show some texts with several pieces of information decorated in it. If you show such images to a human, he or she will be shortly able to find such pieces of information in any other text of the same kind. But can this relatively simple task do a computer as well? Figure 3 represents our first ideas when we started to look for the answer. The figure shows a linguistic tree obtained by automated linguistic analysis of the last sentence of the second figure (Figure 2). It already contain lots of indications (decorated by corresponding labels) of where to look for the wanted piece of information, in this case, the amount of 8,000 Czech Crowns representing the total damage sought by the accident reported in the text.

The main motivation for creating our extraction methods was an attempt to use deep linguistic analysis for this task. Especially for the Czech language with free word order this seemed reasonable. It is much more straightforward to design extraction rules on the basis of linguistic dependency trees than to struggle with the surface structure of text. In a dependency tree, the position of a word is determined by its syntactic (analytical trees) or even semantic role (tectogrammatical trees). So the extraction rules might not be dramatically affected by minor variations of the word order.

Besides that information extraction and annotation is very interesting and challenging problem, it is also particularly useful. This period can be characterized by information overload and information extraction can provide partial answer to that. It provides fine grained indexing of documents, which supports precise search and document filtering. Navigation within individual documents can be faster and reading can be more effective. Other software programs can use the extracted information and perform additional computations resulting in summaries and answers integrated from different sources. The effort in this direction will hopefully culminate in the realization of the idea of the Semantic Web, when all the information will be available in a machine workable form and the whole (semantic) web could be used as a huge knowledge base.

2 Contribution

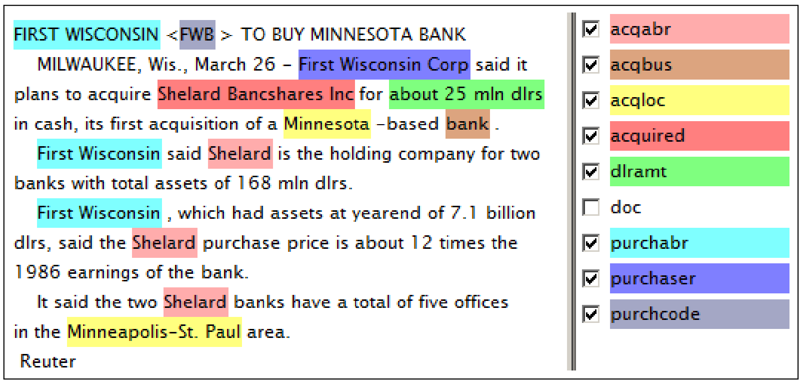
Let us summarize the main contributions of the present work in this section.

2.1 New Ideas, Models and Methods

Novel and original approaches or adaptations of existing ones are presented in this work.

2.1.1 Manual Design of Extraction Rules

The extraction method based on manually designed extraction rules is unique in the high expressiveness of extraction rules and the existence GUI (Graphical User Interface) for graphical design of these rules, both of these benefits are brought by the existence of the linguistic tool Netgraph [Mírovský, 2006], which was exploited as the core of the extraction method.



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.

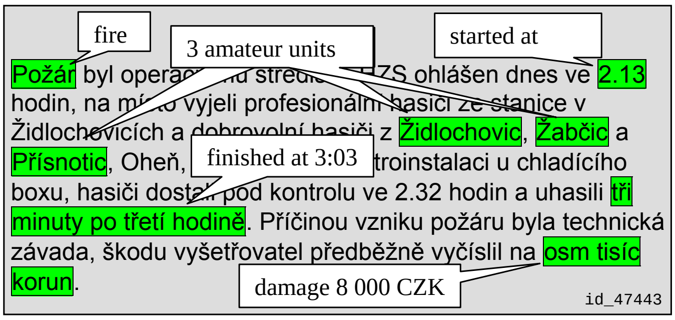
First Wisconsin , which had assets at yearend of 7.1 billion dlrs, said the Shelard purchase price is about 12 times the 1986 earnings of the bank.

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

Reuter

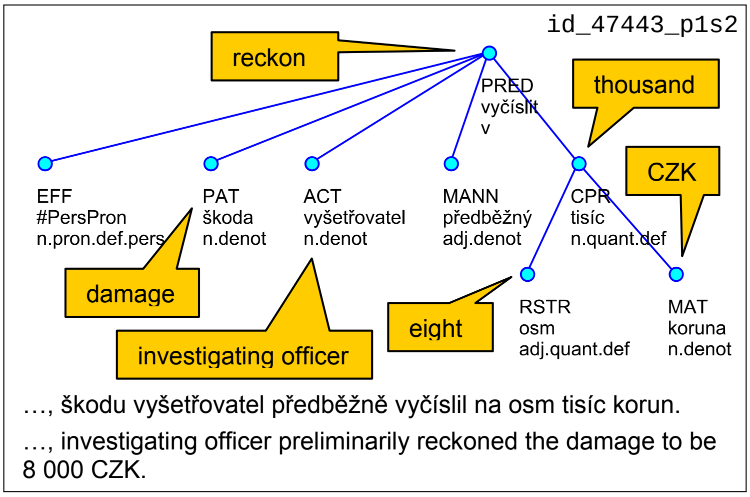
☑ acqabr
☑ acqbus
☑ acqloc
☑ acquired
☑ dlramt
□ doc
☑ purchabr
☑ purchaser
☑ purchcode

Figure 1: Annotations of Corporate Acquisition Events



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.

Figure 2: Annotations of Czech Fireman events



…, škodu vyšetřovatel předběžně vyčíslil na osm tisíc korun.

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

Figure 3: Example of a linguistic tree of one analyzed sentence.

2.1.2 Machine Learning of Extraction Rules

Very similar approaches to our extraction method based on ILP were already reported in literature [Ramakrishnan *et al.*, 2007; Aitken, 2002], but they were developed partly in parallel with our solution and they do not provide a publicly available and usable implementation.

The method also represents the first attempt to use PDT (Prague Dependency Treebank) resources (e.g. tectogrammatical trees [Mikulová *et al.*, 2006]) in the area of information extraction. Evaluation of the method on the language pair of Czech and English demonstrates its language independence.

2.1.3 Shareable Extraction Ontologies

The topic of shareable extraction ontologies introduces completely new paradigm to the design and usage of extraction ontologies [Embley *et al.*, 2002]. The usage of a semantic web reasoner as the interpreter of an extraction ontology has never been demonstrated before.

2.1.4 Fuzzy ILP Classification

Last but not least, the attempt to use information extracted from a document for document classification is also reported for the first time, although our attention is more focused on the implementation and evaluation of the Fuzzy ILP Classifier based on the sound theory of fuzzy logic [Hájek, 1998] and fuzzy ILP [Horváth and Vojtáš, 2007].

2.2 New Software

As a part of our work, new publicly available implementation of the described methods was created.

A simple and extensible API (Application Programming Interface) interface of the extraction method based on manually designed extraction rules is provided such that users can design extraction rules in the Netgraph GUI and evaluate them on the whole corpus using this interface.

The extraction method based on ILP is fully integrated in GATE (a widely used framework for text engineering [Cunningham *et al.*, 2002]) and it can be used as any other machine learning algorithm inside the framework. Moreover integration of TectoMT linguistic tools [Žabokrtský *et al.*, 2008] as well as the Netgraph tree viewer to the GATE framework was realized. Our implementation also provides utility functions for comparative information extraction experiments using the cross-validation technique and investigation of statistical significance.

The implementation of the case study with shareable extraction ontologies is not that general as the rest of the software part of our work but it can be easily followed and reproduced in similar experiments.

The implementation of the Fuzzy ILP Classifier is fully compatible with Weka (a widely used framework for machine learning experiments [Hall *et al.*, 2009]) and it can be used as any other Weka classifier, it also provides the possibility of custom integration of the classifier to an existing installation of the Weka system on the user's computer.

2.3 New Data

Several new datasets were established as a part of our work. We only list them here (see bellow), descriptions are available in the thesis.

- 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

2.4 Exploitation of Existing Resources

Work of other researchers was exploited, extended and/or evaluated in the present work. For example our extraction methods represent the first application of PDT resources (e.g. Netgraph, tectogrammatical trees, etc.) in the area of information extraction. Comprehensive experiments were performed e.g. with the PAUM extraction algorithm [Li *et al.*, 2002], with various semantic web reasoners (Jena[1] ,HermiT[2] ,Pellet[3] and FaCT++[4]), Weka classifiers (Multilayer Perceptron [Bishop, 1996], Support Vector Machine classifier SMO [Keerthi *et al.*, 2001], J48 decision tree [Quinlan, 1993], JRip rules [Cohen, 1995] and Additive logistic regression LogitBoost [Friedman *et al.*, 2000]), etc.

[1] http://jena.sourceforge.net

[2] http://hermit-reasoner.com

[3] http://clarkparsia.com/pellet

[4] http://code.google.com/p/factplusplus

2.5 Evaluation Experiments

All approaches presented in the thesis were evaluated such that readers can obtain clear picture about the performance and usability of these approaches. Most evaluation experiments are detailed and comprehensive, investigating also the statistical significance of the results.

In following subsections, some examples of the evaluation results will be presented.

2.5.1 Examples of Learned Extraction Rules

In Figure 4, we present the most representative examples of extraction rules learned by ILP procedure from the whole dataset Czech Fireman Reports Manually Annotated. The rule with largest coverage for each extraction task was selected and it is provided in the figure. Each rule demonstrates a connection of the target token, annotated as '*mention(task_name)*', with other parts of the sentence through linguistic syntax structures.

For example the second rule (*damage* task, lines 6-10) connects the node A with its tectogrammatical counterpart – numeral B (*n.quant.def*) and with node D ('vyšetřovatel') representing the *investigating officer* who stated the mount of damage.

2.5.2 Czech Fireman Performance

Table 1 summarizes performance evaluation of our ILP based extraction method and its comparison with the PAUM based one [Li *et al.*, 2002] on the dataset Czech Fireman Reports Manually Annotated. The table shows results of the three main evaluation measures: strict precision, strict recall and strict F_1 for each extraction task as well as overall results. 8-fold cross validation was performed 8 times in the experiment. Average values and standard deviations are printed in the table and statistical significance is indicated. Root/subtree preprocessing/postprocessing was performed in the first three tasks: 'cars', 'damage' and 'end subtree'.

Although the precision of the ILP method was better in the majority of tasks and also its overall precision is statically better than the precision of the PAUM method, its recall was worse in all the measurements and also F_1 score is indicating better results of the PAUM method.

2.5.3 Evaluation of Fuzzy ILP Classification

In this section, evaluation experiments with the Fuzzy ILP Classifier will be presented.

```
% [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
```

Figure 4: Rules with largest coverage for each task learned from the whole dataset Czech Fireman Reports Manually Annotated.

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 ●
professional 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 ●

Strict Recall

Task	ILP	PAUM
cars	0.088 ± 0.129	0.353 ± 0.231 ○
damage	0.821 ± 0.261	0.933 ± 0.148 ○
end subtree	0.231 ± 0.203	0.601 ± 0.249 ○
start	0.908 ± 0.115	0.978 ± 0.058 ○
injuries	0.574 ± 0.309	0.814 ± 0.224 ○
fatalities	0.388 ± 0.449	0.536 ± 0.452 ○
professional unit	0.506 ± 0.191	0.811 ± 0.138 ○
amateur unit	0.886 ± 0.210	0.955 ± 0.096 ○
overall	0.550 ± 0.382	0.748 ± 0.312 ○

Strict F_1

Task	ILP	PAUM
cars	0.109 ± 0.147	0.335 ± 0.205 ○
damage	0.828 ± 0.217	0.876 ± 0.131 ○
end subtree	0.283 ± 0.219	0.525 ± 0.213 ○
start	0.912 ± 0.089	0.771 ± 0.111 ●
injuries	0.543 ± 0.280	0.498 ± 0.204
fatalities	0.306 ± 0.420	0.222 ± 0.308
professional unit	0.491 ± 0.200	0.730 ± 0.118 ○
amateur unit	0.827 ± 0.253	0.634 ± 0.296 ●
overall	0.537 ± 0.369	0.574 ± 0.295 ○

○, ● statistically significant improvement or degradation

Table 1: Evaluation on 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 increase or decrease

Legend:

Corr .... Percent correct
Inor .... Percent incorrect
Uncl .... Percent unclassified
Prec .... IR precision, weighted average from all classes
Rec .... IR recall, weighted average from all classes
F .... F measure, weighted average from all classes

Table 2: Evaluation of the methods on the Fireman dataset in 2 times 10-fold cross validation.

We have evaluated both ILP methods and compared them with five additional classifiers: Multilayer Perceptron [Bishop, 1996], Support Vector Machine classifier SMO [Keerthi *et al.*, 2001], J48 decision tree [Quinlan, 1993], JRip rules [Cohen, 1995] and Additive logistic regression LogitBoost [Friedman *et al.*, 2000].

We have evaluated all the methods two times by 10-fold cross validation. The obtained results (average values) are described in Table 2 (with standard deviations and marked statistically significant values).

There is no clear winner in our experiment. But the Fuzzy ILP classifier proved better results than majority of the methods on our data and the results are statistically significant in many cases. Very good results were also obtained using LogitBoost.

2.6 Publications and New Publicly Available Resources

Majority of topics presented in the thesis were already published (going through peer review process), presented and discussed with international audience. Moreover several citations can be found in the literature showing that the work is already contributing to the generally available knowledge. See the attached list of publications (page 17).

Also the software and data parts of our work are publically available on the web, see details on the project's home page[5].

[5] `http://czsem.berlios.de/`

3 Conclusion

Instead of focusing on a single topic and developing it in depth, the thesis has rather broader scope, even though many interesting ideas that came to our minds could not be investigated at all. For example the possibility to perform customized sentence clustering based on the syntactic tree structure could, supported by a handy GUI, result in a powerful tool for manual design of information extraction rules. We also could not develop any web information extraction approach exploiting the HTML structure, which, in combination with topological information about the placement of the elements on the screen, or even OCR, could bring new and interesting solution. We are pleased to share our best experience using the thesis, which includes:

- Rather practically oriented information extraction method based on deep language parsing and manually designed extraction rules.
- A challenge presented by the attempt to beat the performance of other information extraction system by our method based on Inductive Logic Programming. Let us note that the method touched the state-of-the-art in some cases but its time requirements are rather farther from it.
- Agitation brought by the presentation of something completely new, visionary and "unconventional", the idea of shareable extraction ontologies.
- Again a challenge connected with the evaluation of the Fuzzy ILP Classifier with the same result of touching the state-of-the-art (in some cases) and high time requirements; and also the experience with the implementation of formal mathematics with the result in a piece of working software (see the implementation details inside of the thesis.)

References

Stuart Aitken (2002), Learning Information Extraction Rules: An Inductive Logic Programming approach, in F. van Harmelen, editor, *Proceedings of the 15th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, URL http://www.aiai.ed.ac.uk/~stuart/AKT/ilp-ie.html.

Christopher M. Bishop (1996), *Neural Networks for Pattern Recognition*, Oxford University Press, 1 edition, ISBN 0198538642, URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0198538642.

William W. Cohen (1995), Fast Effective Rule Induction, in *In Proceedings of the Twelfth International Conference on Machine Learning*, pp. 115–123, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.8204.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan (2002), GATE: A framework and graphical development environment for robust NLP tools and applications, in *Proceedings of the 40th Anniversary Meeting of the ACL*.

David W. Embley, Cui Tao, and Stephen W. Liddle (2002), Automatically Extracting Ontologically Specified Data from HTML Tables of Unknown Structure, in Stefano Spaccapietra, Salvatore T. March, and Yahiko Kambayashi, editors, *ER*, volume 2503 of *Lecture Notes in Computer Science*, pp. 322–337, Springer, ISBN 3-540-44277-4.

J. Friedman, T. Hastie, and R. Tibshirani (2000), Additive logistic regression: a statistical view of boosting, *Annals of statistics*, 28(2):337–374.

Petr Hájek (1998), *Metamathematics of Fuzzy Logic*, Kluwer, ISBN 978-0-7923-5238-9.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009), The WEKA data mining software: an update, *SIGKDD Explor. Newsl.*, 11(1):10–18, ISSN 1931-0145, URL http://doi.acm.org/10.1145/1656274.1656278.

Tomáš Horváth and Peter Vojtáš (2007), Induction of Fuzzy and Annotated Logic Programs, *ILP: 16th International Conference, ILP 2006, Santiago de Compostela, Spain, August 24-27, 2006, Revised Selected Papers*, pp. 260–274, URL http://dx.doi.org/10.1007/978-3-540-73847-3_27.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy (2001), Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation*, 13(3):637–649.

Yaoyong Li, Hugo Zaragoza, Ralf Herbrich, John Shawe-Taylor, and Jaz S. Kandola (2002), The Perceptron Algorithm with Uneven Margins, in *ICML '02: Proceedings of the Nineteenth International Conference on*

Machine Learning, pp. 379–386, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-873-7.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský (2006), A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank, Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., URL `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html`.

Jiří Mírovský (2006), Netgraph: A Tool for Searching in Prague Dependency Treebank 2.0, in Jan Hajič and Joakim Nivre, editors, *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, 5, pp. 211–222, Prague, Czech rep., ISBN 80-239-8009-2.

J. Ross Quinlan (1993), *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN 1-55860-238-0.

Ganesh Ramakrishnan, Sachindra Joshi, Sreeram Balakrishnan, and Ashwin Srinivasan (2007), Using ILP to Construct Features for Information Extraction from Semi-structured Text, in Hendrik Blockeel, Jan Ramon, Jude W. Shavlik, and Prasad Tadepalli, editors, *ILP'07: Proceedings of the 17th International Conference on Inductive Logic Programming*, volume 4894 of *Lecture Notes in Computer Science*, pp. 211–224, Springer-Verlag, Berlin, Heidelberg, ISBN 3-540-78468-3, 978-3-540-78468-5, URL `http://dx.doi.org/10.1007/978-3-642-01891-6_5`.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas (2008), TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer, in *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pp. 167–170, ACL, Columbus, OH, USA, ISBN 978-1-932432-09-1.

Publications

Refereed (English)

- Jan Dědek, Peter Vojtáš, and Marta Vomlelová. Fuzzy ILP classification of web reports after linguistic text mining. *Information Processing & Management*, 48(3):438 – 450, 2012. Soft Approaches to IA on the Web. Available from: http://www.sciencedirect.com/science/article/pii/S0306457311000264.

- Jan Dědek. Towards semantic annotation supported by dependency linguistics and ILP. In *Proceedings of the 9th International Semantic Web Conference (ISWC2010), Part II*, volume 6497 of *Lecture Notes in Computer Science*, pages 297–304, Shanghai / China, 2010. Springer-Verlag Berlin Heidelberg. ISBN 978-3-642-17748-4. Available from: http://iswc2010.semanticweb.org/accepted-papers/219.

- Jan Dědek and Peter Vojtáš. Semantic annotation semantically: Using a shareable extraction ontology and a reasoner. In Pascal Lorenz and Eckhard Ammann, editors, *SEMAPRO 2011, The Fifth International Conference on Advances in Semantic Processing*, pages 29–34. (c) IARIA, XPS, 2011. ISBN 978-1-61208-175-5. Available from: http://www.thinkmind.org/index.php?view=article&articleid=semapro_2011_2_10_50013.

- Jan Dědek and Peter Vojtáš. Linguistic extraction for semantic annotation. In Costin Badica, Giuseppe Mangioni, Vincenza Carchiolo, and Dumitru Burdescu, editors, *2nd International Symposium on Intelligent Distributed Computing*, volume 162 of *Studies in Computational Intelligence*, pages 85–94, Catania, Italy, 2008. Springer-Verlag. ISBN 978-3-540-85256-8. Available from: http://www.springerlink.com/content/w7213j007t416132.

- Jan Dědek and Peter Vojtáš. Computing aggregations from linguistic web resources: a case study in czech republic sector/traffic accidents. In Cosmin Dini, editor, *Second International Conference on Advanced Engineering Computing and Applications in Sciences*, pages 7–12. IEEE Computer Society, 2008. ISBN 978-0-7695-3369-8. Available from: http://dx.doi.org/10.1109/ADVCOMP.2008.17.

- Jan Dědek and Peter Vojtáš. Exploitation of linguistic tools in semantic extraction - a design. In Mieczysław Kłopotek, Adam Przepiórkowski, Sławomir Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent*

Information Systems XVI, pages 239–247, Zakopane, Poland, 2008. Academic Publishing House EXIT. ISBN 978-83-60434-44-4. Available from: `http://iis.ipipan.waw.pl/2008/proceedings/iis08-23.pdf`.

- Jan Dědek and Peter Vojtáš. Fuzzy classification of web reports with linguistic text mining. In Paolo Boldi, Giuseppe Vizzari, Gabriella Pasi, and Ricardo Baeza-Yates, editors, *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, volume 3, pages 167–170, Los Alamitos, CA, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3801-3. Available from: `http://dx.doi.org/10.1109/WI-IAT.2009.254`.

- Jan Dědek, Alan Eckhardt, and Peter Vojtáš. Experiments with czech linguistic data and ILP. In Filip Železný and Nada Lavrač, editors, *ILP 2008 - Inductive Logic Programming (Late Breaking Papers)*, pages 20–25, Prague, Czech Republic, 2008. Action M. ISBN 978-80-86742-26-7. Available from: `http://ida.felk.cvut.cz/ilp2008/ILP08_Late_Breaking_Papers.pdf`.

- Jan Dědek, Alan Eckhardt, Leo Galamboš, and Peter Vojtáš. Discussion on uncertainty ontology for annotation and reasoning (a position paper). In P. C. G. da Costa, editor, *URSW '08 Uncertainty Reasoning for the Semantic Web - Volume 4*, volume 423 of *CEUR Workshop Proceedings*, pages 128–129. The 7th International Semantic Web Conference, 2008. Available from: `http://c4i.gmu.edu/ursw/2008/papers/URSW2008_P2_DedekEtAl.pdf`.

- Jan Dědek. Web information extraction systems for web semantization. In Peter Vojtáš, editor, *ITAT 2009 Information Technologies - Applications and Theory*, pages 1–6, Seňa, Slovakia, 2009. PONT Slovakia. ISBN 978-80-970179-2-7. Available from: `http://ceur-ws.org/Vol-584/paper1.pdf`.

- Jan Dědek and Peter Vojtáš. Web information extraction for e-environment. In Jiri Hrebicek, Jiri Hradec, Emil Pelikan, Ondrej Mirovsky, Werner Pillmann, Ivan Holoubek, and Thomas Bandholtz, editors, *Proceedings of the European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT*, pages 138–143, Brno, Czech Republic, 2009. Masaryk University. ISBN 978-80-210-4824-9. Available from: `http://www.e-envi2009.org/?proceedings`.

Refereed (Czech and Slovak)

- Jan Dědek and Peter Vojtáš. Extrakce informací z textově orientovaných zdrojů webu. In Václav Snášel, editor, *Znalosti 2008*, pages 331–334, 2008. ISBN 978-80-227-2827-0. Available from: `http://znalosti2008.fiit.stuba.sk/download/articles/znalosti2008-Dedek.pdf`.

- Jan Dědek, Peter Vojtáš, and Marta Vomlelová. Evaluace fuzzy ILP klasifikátoru na datech o dopravních nehodách. In Pavel Smrž, editor, *Znalosti 2010*, pages 187–190, Jindřichův Hradec, 2010. VŠE v Praze, Oeconomica. ISBN 978-80-245-1636-3.

- Jan Dědek, Peter Vojtáš, and Juraj Vojtáš. Obsahuje web indikace blížící se krize? Umíme je rozeznat? In Jiří Voříšek, editor, *Systémová integrace 2010*, pages 166–175, Praha, 2010. VŠE v Praze, Oeconomica. ISBN 978-80-245-1660-8.