Delving deep into some UFAL data and tools

Ivana Lukšová

Supervisor: Barbora Hladká

May 18, 2014

Introduction

Master Thesis

- **Ontology Enrichment Based on Unstructured Text Data**
- supervised by Martin Nečaský, Department of Software Engineering

Dissertation Thesis

- **Exploiting linguistic knowledge for relation extraction from texts**

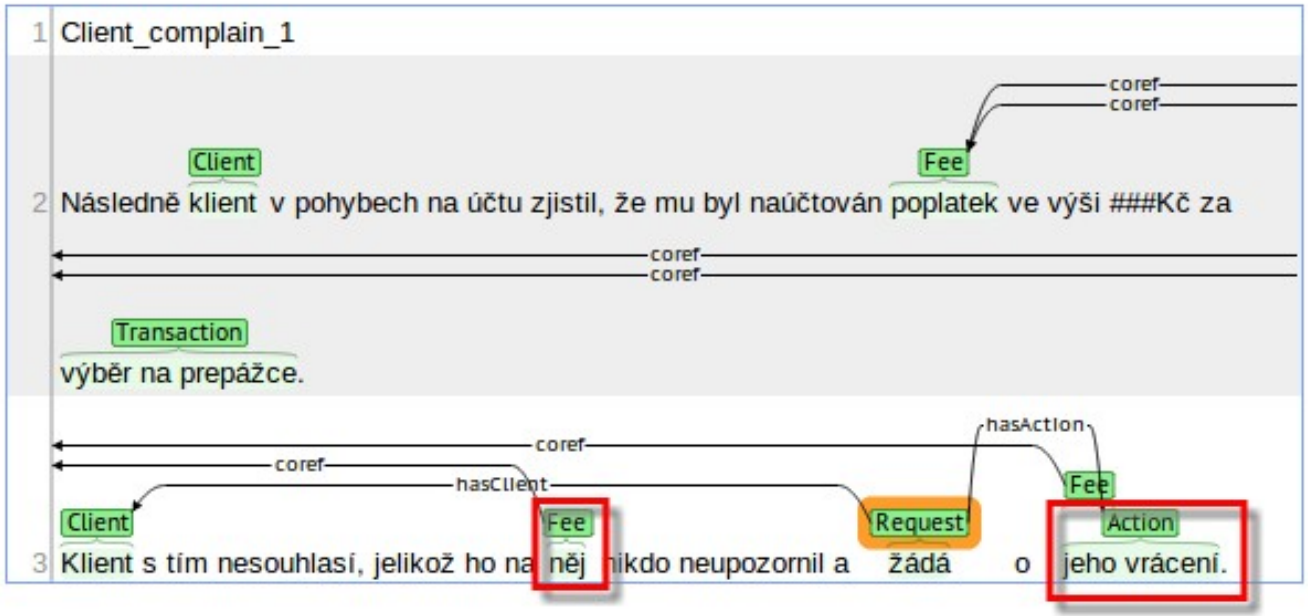
My very first year at UFAL

- Recent research on coreference resolution at UFAL
- Extraction of numerical expression in INTLIB project
- Sentence diagramming editor ČAPEK

Coreference - my field of interest

Coreference resolution within information extraction

- real-word data
  - medical domain
  - banking domain
- extraction of entities and relations
- coreference between entities



Coreference resolution

Grammatical coreference

- Nguy 2006
  - set of rules
  - over 90% F-measure

Textual coreference - pronouns

- Nguy and Žabokrtský 2007
  - rule based
  - 74,2% F-measure
- Nguy et al. 2009
  - C5.0 classifier - 76,3% F-measure
  - perceptron ranking - 79,4% F-measure

Coreference resolution, cont.

Textual coreference - name phrases

- Novák 2010
  - maximum entropy ranking
  - 39,4% F-measure
- Novák and Žabokrtský 2011
  - maximum entropy classifier - 42,3% F-measure
  - maximum entropy ranking - 44,3% F-measure
  - perceptron ranking - 44,4% F-measure

INTLIB - extraction of numerical expressions

`http://ufal.mff.cuni.cz/grants/intelligent-library`

Problem: extraction of structured information from unstructured text

- extraction of dictonary items
  - entities, units, attributes
- extraction of numerical expressions
- their linking

Provozní doba zařízení je 8 hod/den, tj. cca 1000-1120 hod/rok
Kapacita parkování v terminálu bude 507 vozidel.

INTLIB - extraction of numerical expressions

Solution: extraction using the GATE tool

- dictionaries - generation of morphological forms with CzechMorpho
- annotation with GATE gazeteers & regular expressions
- tagging with Treex GATE plugin
- annotations linking with regular expressions over annotations (GATE JAPE plugin):

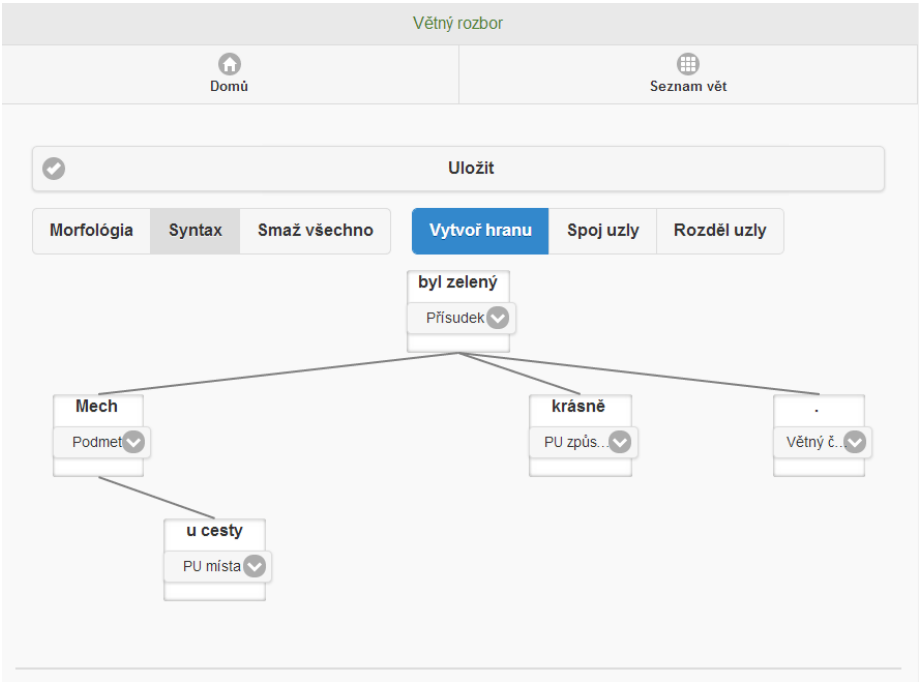
(Attribute) (Entity, Tag= "N...2") (Token){0,5} (Number) (Unit)

Provozní doba zařízení je 8 hod/den, tj. cca 1000-1120 hod/rok

ČAPEK

http://ufal.mff.cuni.cz/capek

- sentence diagramming editor for schoolchildren
- collect the diagrams and transform them to PDT
- based on the results of Hana and Hladká, 2012
- GAUK n. 1568314, duration 2014 - 2016



ČAPEK

Problems:

- how to combine sentence diagrams of multiple users to better one?
- how to measure user agreement on sentence diagrams?

Paper (Hana, Hladká, Lukšová, 2014) submitted to LAW VIII workshop, COLING 2014

ČAPEK - Tree Edit Distance

- measuring of user agreement on a sentence diagram
- adjusted Tree Edit Distance
- including 4 operations
  - split, join, create node, change parent, change label

$$TED(D_1, D_2, n) = (\#SPL + \#JOIN + \#INS + \#LINK + \#SLAB)/n$$

	(T1,T2)	(T1,S1)	(T1,S2)	(S1,S2)
# of sentences	101	91	101	91
$\overline{TED}$	0.26	0.49	0.56	0.69

Table: Average $\overline{TED}$ for pairs of teachers (T1, T2) and students (S1, S2)

ČAPEK - multiple diagram combination

- majority voting on token join/split
- diagram edge weighting based on user voting
- greedy algorithm to find edges and labels of a final sentence diagram

Experiments with 7 annotators:

	U1	U2	U3	U4	U5	U6	U7	their combination
$\overline{TED}$	0.78	0.63	0.56	0.76	0.38	0.62	1.21	0.40

Table: Average $\overline{TED}$ for pairs of teacher T1 and users U1,...,U7 and their combination MV

References I

[1] J. Hana and B. Hladká.
Getting more data - schoolkids as annotators.
In *LREC*, pages 4049–4054. European Language Resources Association (ELRA), 2012.

[2] G. L. Nguy.
Proposal of a set of rules for anaphora resolution in Czech.
Master's thesis, 2006.

[3] G. L. Nguy, V. Novák, and Z. Žabokrtský.
Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech.
In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285, London, UK, 2009. The Association for Computational Linguistics.

References II

[4] G. L. Nguy and Z. Žabokrtský.
Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data.
In A. Branco, T. McEnery, R. Mitkov, and F. Silva, editors, *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, pages 77–81, Lagos (Algarve), Portugal, 2007. University of Lisbon, Faculty of Sciences, CLUP-Center for Linguistics of the University of Oporto.

[5] M. Novák and Z. Žabokrtský.
Resolving Noun Phrase Coreference in Czech.
Lecture Notes in Computer Science, 7099:24–34, 2011.

[6] M. Novák.
Machine Learning Approach to Anaphora Resolution.
Master's thesis, 2010.