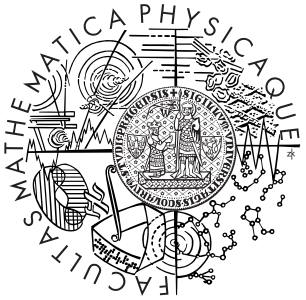
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jan Dědek

Sémantická anotace dat z webovských zdrojů

Katedra softwarového inženýrství

Vedoucí diplomové práce: Prof. RNDr. Peter Vojtáš, DrSc.
Studijní program: Informatika, I2 - Softwarové systémy

2007

Chtěl bych poděkovat vedoucímu Prof. RNDr. Peterovi Vojtášovi, DrSc. za motivující a inspirující vedení a kontrolu průběhu vzniku diplomové práce. Dále děkuji Ing. Zdeněku Žabokrtskému, Ph.D., RNDr. Václavu Klimešovi, Ph.D., doc. PhDr. Karelu Palovi, CSc za poskytnutí softwaru a přínosné konzultace.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 6. 8. 2007 Jan Dědek

Název práce: Sémantická anotace dat z webovských zdrojů
Autor: Jan Dědek
Katedra (ústav): Katedra softwarového inženýrství
Vedoucí diplomové práce: Prof. RNDr. Peter Vojtáš, DrSc.
e-mail vedoucího: Peter.Vojtas@mff.cuni.cz

Abstrakt: Tato práce se odráží od myšlenky sémantického webu. Stručně rozebírá možnosti formální reprezentace znalostí v deskripční logice a její paralelu v několika formalismech pro tvorbu ontologií. Ukazuje, jak lze využít ontologií při sémantické anotaci webovských zdrojů. Představuje sémantickou anotaci v praxi, v kontextu několika projektů z různých oblastí. V práci jsou rozebrány různé metody extrakce informací, které pomáhají sémantickou anotaci zautomatizovat. Podrobněji jsou v tomto ohledu popsány nástroje, které poskytuje současná česká počítačová lingvistika. Na teoretické úrovni se tato práce dotýká vztahu mezi lingvistickou anotací přirozeného jazyka a formální reprezentací znalostí v deskripční logice. V rámci této práce byl proveden experiment – zpracování českého přirozeného textu několika lingvistickými nástroji za účelem jeho sémantické anotace. Jsou popsány zkušenosti získané v tomto experimentu a doporučení, která z něj vyplývají pro extrakci informací z českých textů.

Klíčová slova: sémantický web, sémantická anotace, ontologie, zpracování přirozeného jazyka, extrakce informací

Title: Semantic annotation of data from web resources
Author: Jan Dědek
Department: Department of software engineering
Supervisor: Prof. RNDr. Peter Vojtáš, DrSc.
Supervisor's e-mail address: Peter.Vojtas@mff.cuni.cz

Abstract: This work starts with the idea of The Semantic Web. Then basic description logics is introduced with its parallel in a couple of ontology building formalisms. We show how the ontologies are employed in the semantic annotation process. We describe some projects that use semantic annotation in a practical way. Information extraction methods that help to automatize the semantic annotation process are mentioned. Tools for natural language processing of Czech are described in more detail. In a practical experiment it is shown how these tools can help with information extraction from plain text. Experiences and recommendations obtained in this experiment are presented in this work. This work also deals with the relationship of natural language processing and a formal representation of knowledge in description logics.

Keywords: semantic web, semantic annotation, ontology, natural language processing, NLP, information extraction

Obsah

Kapitola 1

Úvod

Sémantická anotace je .....

1.1 Motivace

1.2 Přínosy práce

Komu? Čím? Srozumitelné pro nelingvistu....

Mimo jiné: Tato práce se snaží přiblížit možnosti, jak využít dostupné lingvistické nástroje analyzující český text (4.4.3) především lidem, kteří se zabývají extrakcí informací z textu ale nejen jim. Práce na čtenáře neklade žádné nároky co se týká lingvistického vzdělání a sama základní znalosti z lingvistiky podněcuje (4). Těmito znalostmi se snaží pokrýt požadavky, které klade používání lingvistických nástrojů zde popisovaných. Zběžné znalosti zde poskytnuté jsou doplněny odkazy a referencemi na zdroje, kde se čtenář o dané problematice může dozvědět více.

1.2.1 Souhrn

Práce poskytuje základní souhrn v oblasti sémantické anotace, takže si čtenář může udělat představu o tom, kterými směry se sémantická anotace ubírá, jaké metody byly využity, s jakou úspěšností atp. 3.2.1

Doporučení autorům stánek (DRFa, HTML-A, přehled ontlogií) 3.1.

1.2.2 Průzkum

Součástí práce je praktický experiment (7) s lingvistickými nástroji: !!!!!!!!!!!!dopsat!!!!!!!!!!!. Čtenáři jsou poskytnuty zkušenosti z týkající se použití těchto nástrojů a z prací které s jejich použitím souvisely. Tyto zkušenosti se týkají především dostupnosti, zprovoznění, výkonnosti, přínosů a nedostatků těchto nástrojů. !!!!!!!!!!!!!!!!!!!!!!!!!!!!rozvést: navržený postup X další možnosti

Postup experimentu se v jednotlivých fázích snaží kopírovat skutečné akce, které by bylo nutné provést v opravdovém projektu zaměřeném na sémantickou anotaci. V práci tak vzniká jednoduchá základní analýza tohoto typu projektů. Ve skutečném projektu pak bude možné ji přinejmenším jako inspiraci využít.

1.2.3 Návrhy

V práci je navržena metodika, jak by se při extrakci informací pomocí nástrojů v rámci této práce testovaných dalo postupovat.

Je zde navržený jednoduchý dotazovací jazyk pro lingvistické anotace.
Stručný návrh indukce vzorů.
Zamyšlení nad možnostmi lingvistické anotace pro indexaci dokumentů.

1.2.4 Teorie

Pokus o teoretický přínos v oblasti sémantické anotace. 3.4

Zamyšlení mad možnými přínosy a vztahem mezi počítačovou lingvistikou a formální reprezentací znalostí. 5

1.3 Sémantický web

V roce 2001 napsal Tim Berners-Lee, tvůrce současného webu a ředitel Konsorcia W3C spolu s dalšími autory velmi známý článek The Semantic Web [1] (volný český překlad je k dispozici například v [5]). V tomto článku je popsána lákavá představa světa, kde všechny nepříjemné problémy spojené s zařizováním běžných životních problémů, jako je například návštěva lékaře, pomáhají vyřídit softwarový agenti. Pomáhají je vyřídit především

tím, že naleznou a zkombinují všechny důležité relevantní informace, které jsou potřeba. Tedy například najdou lékaře, který se specializuje na daný druh zdravotních potíží, adresu a otvírací dobu jeho ordinace, dopravní dostupnost tohoto místa atd.

Většina těchto informací je již dnes na webu dostupná, avšak i zkušeného uživatele internetu stojí jejich nalezení nezanedbatelný čas a energii. Navíc nalezení informací je teprve první část problému. To, jakým způsobem je zkombinovat a vyhodnotit, je část druhá. Avšak vyřešení tohoto druhého problému není pro současný software žádnou utopií. Například nalézt dopravní spojení na adresu lékařovy ordinace v dnešní době rozhodně nepovažujeme za programátorsky neřešitelný problém.

Softwarový agenti pravděpodobně nebudou v dohledné době tak „chytří“, aby se vyznali ve webových stránkách současného internetu a dokázali z nich vytěžit informace, které potřebují. Proto se v souvislosti s myšlenkou sémantického webu snažíme tyto stránky softwarovým agentům přiblížit, udělat je srozumitelnější respektive přístupnější pro strojové získávání informací z nich. Do stránek se vkládají takzvaná metadata (*data o datech*), která co možná nejpřesněji formálně zachycují obsah stránek jinak srozumitelný jen pro člověka. Tomuto obohacování stránek o metadata budeme říkat *sémantická anotace*.

Kapitola 2

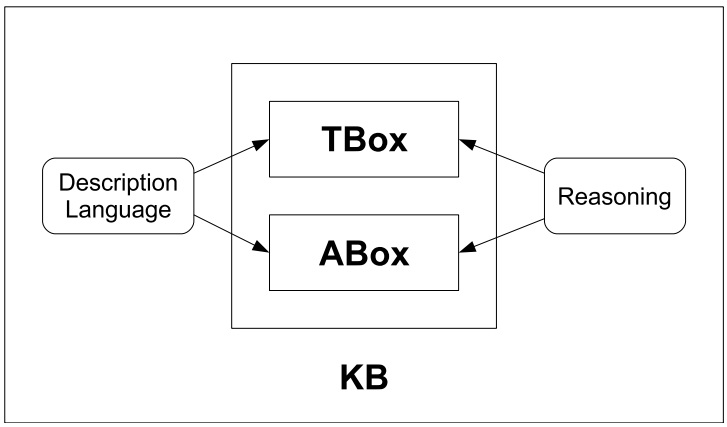
Reprezentace znalostí

Úvod podle: Matulik SemWeb [5]

2.1 Deskripční logika

Franz Baader, Werner Nutt. Basic Description Logics: [4]

Budeme zde popisovat deskripční jazyk $\mathcal{ALCN}$, který je z jazyků deskripční logiky nejbohatší.



Obrázek 2.1: Knowledge base systému pro reprezentaci znalostí

$$
\begin{array}{rll}
C, D \ \rightarrow & A \mid & \text{(atomický pojem)} \\
& \top \mid & \text{(univerzální pojem)} \\
& \bot \mid & \text{(prázdný pojem)} \\
& \neg C \mid & \text{(negace)} \\
& C \sqcap D \mid & \text{(průnik)} \\
& C \sqcup D \mid & \text{(sjednocení)} \\
& \forall R.C \mid & \text{(hodnotová restrikce)} \\
& \exists R.C \mid & \text{(existenční kvantifikátor)} \\
& \geqslant n\, R \mid & \text{(maximální kardinalita)} \\
& \leqslant n\, R & \text{(minimální kardinalita)}
\end{array}
$$

Obrázek 2.2: Syntax deskripční logiky

$$
\begin{aligned}
\top^{\mathcal{I}} &= \Delta^{\mathcal{I}} \\
\bot^{\mathcal{I}} &= \emptyset \\
(\neg C)^{\mathcal{I}} &= \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}} \\
(C \sqcap D)^{\mathcal{I}} &= C^{\mathcal{I}} \cap D^{\mathcal{I}} \\
(C \sqcup D)^{\mathcal{I}} &= C^{\mathcal{I}} \cup D^{\mathcal{I}} \\
(\forall R.C)^{\mathcal{I}} &= \{a \in \Delta^{\mathcal{I}} \ : \ \forall b.\ (a,b) \in R^{\mathcal{I}} \rightarrow b \in C^{\mathcal{I}}\} \\
(\exists R.C)^{\mathcal{I}} &= \{a \in \Delta^{\mathcal{I}} \ : \ \exists b.\ (a,b) \in R^{\mathcal{I}} \ \wedge \ b \in C^{\mathcal{I}}\} \\
(\geqslant n\, R)^{\mathcal{I}} &= \left\{a \in \Delta^{\mathcal{I}} \ : \ \left|\{b \ : \ (a,b) \in R^{\mathcal{I}}\}\right| \geq n\right\} \\
(\leqslant n\, R)^{\mathcal{I}} &= \left\{a \in \Delta^{\mathcal{I}} \ : \ \left|\{b \ : \ (a,b) \in R^{\mathcal{I}}\}\right| \leq n\right\}
\end{aligned}
$$

Obrázek 2.3: Interpretace deskripční logiky

$$
\begin{aligned}
\mathsf{Woman} &\equiv \mathsf{Person} \sqcap \mathsf{Female} \\
\mathsf{Man} &\equiv \mathsf{Person} \sqcap \neg\mathsf{Woman} \\
\mathsf{Mother} &\equiv \mathsf{Woman} \sqcap \exists\mathsf{hasChild}.\mathsf{Person} \\
\mathsf{Father} &\equiv \mathsf{Man} \sqcap \exists\mathsf{hasChild}.\mathsf{Person} \\
\mathsf{Parent} &\equiv \mathsf{Father} \sqcup \mathsf{Mother} \\
\mathsf{Grandmother} &\equiv \mathsf{Mother} \sqcap \exists\mathsf{hasChild}.\mathsf{Parent} \\
\mathsf{MotherWithManyChildren} &\equiv \mathsf{Mother} \sqcap \geqslant 3\,\mathsf{hasChild} \\
\mathsf{MotherWithoutDaughter} &\equiv \mathsf{Mother} \sqcap \forall\mathsf{hasChild}.\neg\mathsf{Woman} \\
\mathsf{Wife} &\equiv \mathsf{Woman} \sqcap \exists\mathsf{hasHusband}.\mathsf{Man}
\end{aligned}
$$

Obrázek 2.4: Terminologie (TBox) pro popis rodinných vztahů

$$
\begin{array}{ll}
\mathsf{MotherWithoutDaughter}(\mathsf{MARY}) & \mathsf{Father}(\mathsf{PETER}) \\
\mathsf{hasChild}(\mathsf{MARY}, \mathsf{PETER}) & \mathsf{hasChild}(\mathsf{PETER}, \mathsf{HARRY}) \\
\mathsf{hasChild}(\mathsf{MARY}, \mathsf{PAUL}) &
\end{array}
$$

Obrázek 2.5: Tvrzení o individuích (ABox) v terminologii rodinných vztahů

2.2 Ontologie

2.2.1 RDF - Resource Description Framework

2.2.2 OWL - Web Ontology Language

Svátek: Verze OWL

OWL Lite
– omezený z hlediska elementárních konstruktů; zejména neumožňuje definovat kardinalitu jinou než 0 nebo 1; výpočtově efektivní
OWL DL – „default“ verze
– stále ještě zachovává rozhodnutelnost hlavních odvozovacích úloh
– aktuálně vzniká obohacená verze OWL1.1
OWL Full
– stejné konstrukty jako OWL DL, ale méně omezení při jejich používání
– nezachovává oddělenost tříd, vlastností a instancí
– teprve OWL Full je nadjazykem RDF/S!

Kapitola 3

Sémantická anotace

Sémantická anotace není v současné době přesně vymezený termín. V této práci budu s tímto pojmem pracovat poměrně volně až na kapitolu o teoretických otázkách sémantické anotace 3.4, kde bych se o takové vymezení chtěl pokusit. V knize [2] se tomuto termínu vyhýbají opisem *annotation for the Semantic Web* tedy anotace pro sémantický web. V tomto smyslu budu užívat pojem sémantické anotace ve zbytku práce. Tedy sémantickou anotací budu rozumět takovou anotaci, která je určená pro sémantický web. Pro přesnost ještě doplňme, že anotací jako hotovým dílem rozumíme výsledek, výstup případně výstupní data procesu anotace.

Člověk by intuitivně očekával, že výsledkem sémantické anotace bude nějak označkovaný text, který je jejím vstupem. Případně se nemusí jednat o text ale o strukturovaná data - nejčastěji HTML případně XML. V našem způsobu chápání sémantické anotace, tedy jako anotace pro sémantický web, však jejím výstupem může být téměř libovolný datový celek. Příznačné pro sémantické anotace je to, že nějakým způsoben využívají ontologie a RDF, viz 2.2.

3.1 Autorům Web-stránek

Autorská anotace. RDFa, HTML-A

3.1.1 Často používané ontologie

3.2 Dodatečná anotace, strojová

3.2.1 Rozdělení

Po praktické stránce: dostupnost, upravitelnost (jiná doména), stabilita - náchylnost na změny v datech

Po teoretické stránce: čísla (úspěšnost), metody, podle [3] + doplnit.

```
*  Multistrategy
      o Pattern-based
            + Discovery (Seed expansion)
            + Rules (JAPE, Taxonomy label matching)
      o Machine Learning-based
            + Probabilistic (Hidden Markov Models, N-gram analysis)
            + Induction (Linguistic, Structural)
```

Platform	Method	Machine Learning	Manual Rules	Bootstrap Ontology
AeroDAML	Rule	N	Y	WordNet
Armadillo[1]	Pattern Discovery	N	Y	User
KIM	Rule	N	Y	KIMO
MnM[2]	Wrapper Induction	Y	N	KMi
MUSE[3]	Rule	N	Y	User
Ont-O-Mat: Amilcare[4]	Wrapper Induction	Y	N	User
Ont-O-Mat: PANKOW[5]	Pattern Discovery	N	N	User
SemTag[6]	Rule	N	N	TAP

3.3 Některé projekty

3.3.1 Semantic MediaWiki

Semantic MediaWiki (SMW)

3.3.2 Artequakt

3.3.3 WEESA

V článku !!!!doplnit!!!! je navržena přímočará deterministická metoda, jak XML data převádět na RDF data. K XML datům je nejprve nutné vytvořit jednoduchou ontologii pokrývající třídy a atributy, které chceme převést. Tato ontologie se pak automaticky naplní daty ze zdrojového XML. Metoda dobře funguje pro "dobře strukturované" XML data (XML elementy odpovídají třídám atp.) Problémy nastávají s daty tvaru:

```
<SearchDataElement name="Processor Speed" value="500MHZ">
```

...To že se jedná o vlastnost Processor Speed s hodnotou 500MHZ tato metoda nedokáže postihnout (záměrně – důraz byl kladen na jednoduchost a konzistenci návrhu).

3.3.4 The Lixto Project

Software umožňující automaticky stahovat vybrané informace z webových stránek. Uživatelsky příjemné a propracované prostředí. Jedná se o nástroj, ve kterém si uživatel může naprogramovat svůj wrapper na míru šitý svým potřebám.

Lixto server

Lixto server umožňuje automaticky spouštět hotové Wrappery a získaná data dále zpracovat. Pro zpracování dat je k dispozici široká paleta možností. Zpracovaná data server umí doručit nejrůznějším aplikacím i zařízením.

Možnosti automatické sémantické anotace

Pro využití při automatizované sémantické anotaci je velkou překážkou nutnost pokaždé znovu ručně naprogramovat Lixto na každou stránku.

3.3.5 The KIM Platform: Knowledge & Information Management

3.3.6 Image Semantics without Annotations

Teoreticky bohatý článek !!!!doplnit!!!!, zabývá se indexací a vyhledáváním obrázků v internetu podobné databázi. Definuje internet jako graf provázaných dokumentů a obrázků. Navržena je algebra pro manipulaci s takovým grafem (operace jako: insert, delete, nodes, edges, union, ...) Velmi propracované je porovnávání obrázků pomocí (adaptivní) podobnostní míry, která se skládá ze třech složek:

- Linguistic Modality (popisky, okolní text)
- Closed Word Modality (ustálené, přesné pojmy uvnitř uzavřené komunity)
- Emergent Modality příp. User Modality (míra vzniklá z akcí a operací souvisejících s obrázky)

  Upravuje se např. pomocí dialogu ve kterém uživatel "přetahuje" podobné obrázky k sobě.

3.3.7 MUMIS - Multi-Media Indexing and Searching

NLP projekt MUMIS[7] využívající sémantickou anotaci (nejen) textů o fotbalových zápasech k (sémantické) indexaci videozáznamů těchto zápasů. Informace o zápasech jsou získávány z různých zdrojů (texty - vice i méně strukturované, zvukový záznam řeči) v různých jazycích (En, Ge, Nl - Dutch).

3.4 Teorie

Domnívám se, že nebylo mnoho teoretických otázek týkajících se sémantická anotace vůbec formulováno, natož uspokojivě vědecky vyřešeno.

[7] http://hmi.ewi.utwente.nl/Projects/mumis/

3.4.1 Otázky

V čem tkví sémantičnost anotace?

Co si představujeme pod ideální sémantickou anotací?

Jaký je vztah lingvistické a sémantické anotace?

Jsou mezi nimi hranice? Kde přibližně?

Dá se lingvistická anotace převést na sémantickou?

Dal by se z toho odvodit nijaký univerzálnější návod / algoritmus, jak od lingvistické anotace k té sémantické přejít?

Jak vypadá ideální anotace v ideálním případě?

Pokusit se stanovit podmínky které ideální anotaci braní, resp. předpoklady které by ji umožnily.

Jaky je rozdil mezi prirozenym jazykem a deskripcni logikou?

viz 5

Kapitola 4

Lingvistická anotace

Lingvistickou anotací budeme v této práci označovat činnost, při které se text přirozeného jazyka obohacuje o lingvistickou informaci o slovech, větách, vztazích mezi slovy, mezi větami, o typu a původu textu atp. Lingvistická anotace nebo též značkování korpusu je jedna z činností korpusové lingvistiky. Korpusem rozumíme soubor textů spolu s lingvistickou informací k nim dodanou. Korpusová lingvistika je poměrně novou disciplínou, jejíž vznik, stejně jako vznik celé počítačové lingvistiky vůbec, umožnil rozvoj výpočetní techniky posledních let. Korpusová lingvistika se zabývá zkoumáním a shromažďováním textů přirozeného jazyka (vytvářením korpusu). Texty se anotují za velké podpory počítače - například morfologická desambiguace (viz 4.4.3) by byla bez softwarové podpory nadlidský úkol. Avšak pro anotaci korpusu je stále nutná spousta lidské „ruční“ práce. Takto anotované texty představují velmi cenná data, ze kterých se především pomocí statistických metod dají vyvodit nové poznatky o jazyce. Díky ručně anotovaným korpusům vynikla a stále vyniká většina softwarových nástrojů pro počítačové zpracování přirozeného jazyka.

Korpusová lingvistika dnes nemalou mírou přispívá k jazykovému výzkumu. Na stránkách Českého národního korpusu[1] se dokonce uvádí, že přináší natolik nové poznatky o jazyce, že do dosavadního vývoje jazykovědy vnáší radikální převrat. Toto tvrzení nemusí působit překvapivě, pokud například srovnáme původní latinskou lingvistickou terminologii s tou, která vzniká v počítačové lingvistice v poslední době.

[1] http://ucnk.ff.cuni.cz/

Korpusy se v zásadě značkují třemi druhy značek. 1) Značky správní zachycují identifikační údaje o každém textu - informace o jeho původu a zdroji. 2) Značky strukturní zachycují hierarchickou strukturu textu tj rozdělení textu do kapitol, odstavců, vět, slov a interpunkčních znamének (tokenů). 3) Značky lingvistické jsou přiřazeny k jednotlivým slovům a nesou informaci o lingvistických kategoriích, které dané slovo nese.

Samostatnou kapitolou lingvistické anotace je potom zachycení gramatické stavby věty. K tomu se používají dva typy gramatik - složková a závislostní gramatika. Závislostní gramatika má dlouholetou tradici v popisu jazyků evropského kontinentu a zdá se, že má určité výhody i pro popis angličtiny, která bývá častěji zpracovávána gramatikou složkovou. Složkový popis je blíže Chomského pojetí jazyka. Věta je podle složkové gramatiky rekurzivně dělena do menších a menších složek. Postup začíná rozdělením věty na část podmětnou a přísudkovou a postupuje dělením těchto složek na podsložky až dojde k jednotlivým slovům. Závislostní přístup naproti tomu vezme jednotlivá slova a ta pospojuje závislostními hranami do takzvaného závislostního stromu. Velmi podrobně je problematika gramatik a větné syntaxe popsána v [7].

Kromě toho, že je možné strukturu věty přirozeného jazyka zapisovat pomocí různých typů gramatik, je též možné zapisovat tuto strukturu na různých významových rovinách. Lingvistický výzkum různých jazyků ukazuje vhodnost takového vícevrstvého popisu. Na různých rovinách je totiž možné přehledněji postihnout různé jazykové jevy. Tento problém je podrobněji rozveden v práci [6], dále je zde představena tradiční trojice rovin pro popis struktury věty. Jsou to 1) rovina morfématická (též morfologická) - tvarosloví 2) rovina povrchové syntaxe (též rovina mluvnické stavby věty či analytická rovina) 3) rovina tektogramatická neboli významová stavba věty nebo též hloubková syntax. Stejně jsou roviny lingvistické anotace zpracovány v projektu PDT (viz oddíl 4.2), z jehož popisu budeme dále vycházet.

4.1 Lingvistické značky

Lingvistické značky rozdělíme do tří kategorií. Podle roviny lingvistické anotace, tak jak jsou rozděleny v projektu PDT (viz oddíl 4.2). Značky morfologické roviny jsou nezávisle přiřazovány jednotlivým slovům. Naproti tomu značky analytické a tektogramatické roviny popisují strukturu věty a jejich

značky popisující vztahy mezi jednotlivými slovy se mohou týkat více slov najednou.

Následuje stručný popis jednotlivých značek každé roviny. Podrobnější popis lingvistických značek je možné najít například v [9], [10], [11].

4.1.1 Morfologická rovina

Slovní tvar

Tato značka obsahuje tvar, v jakém se dané slovo vyskytuje v původním textu včetně zápisu malých a velkých písmen. Od původního výskytu se liší se jen ve výjimečných případech, kdy například původní slovní tvar byl číslice s desetinnou čárkou (snaha o jednotný zápis čísel) nebo se jednalo o překlep.

Lemma

Lemma je takzvaný základní tvar slova. Jednoznačně slovo identifikuje. V tomto tvaru je dané slovo obvykle uváděno ve slovnících.

Morfologická značka

Morfologická značka v sobě spojuje informaci o morfologických kategoriích, které dané slovo nese. Z morfologické značky je možné zjistit slovní druh, jmenný rod, číslo, pád, osobu, čas, atd.

4.1.2 Analytická rovina

Analytická rovina je první úroveň pro strukturní anotaci. Opouští se zde lineární anotace, kdy je každé slovo bráno samostatně bez ohledu na kontext, a do anotace textu se zavádí větná struktura. Všechna původní slova textu zůstávají zachována a dostávají ve výsledné struktuře svou funkci.

Na analytické rovině se vytváří stromová struktura věty (stromem rozumíme orientovaný acyklický graf s jedním kořenem). Uzly stromu jsou tvořeny jednotlivými slovy respektive tokeny. Hrany stromu reprezentují vztahy závislosti. Do kořene stromu je umístěno řídící sloveso věty, na toto sloveso

se pak zavěšují ostatní slova. V případě, že se jedná o souřadné souvětí, kořenem stromu je spojka případně čárka, která jednotlivé věty souvětí odděluje. Základním cílem je korektní zachycení struktury věty a označení typu závislosti. Typ závislosti je uložen uvnitř lingvistické značky *analytická funkce*.

Analytická funkce

Analytická funkce je poměrně dobře známý pojem, který se požívá na českých základních a středních školách při takzvaném větném rozboru. Tam se ale většinou neoznačuje jako analytická funkce ale jako *větný člen*.

V závislostním stromu analytické roviny anotace, se analytickou funkcí označí každá závislostní hrana. Analytická funkce označuje typ této závislosti.

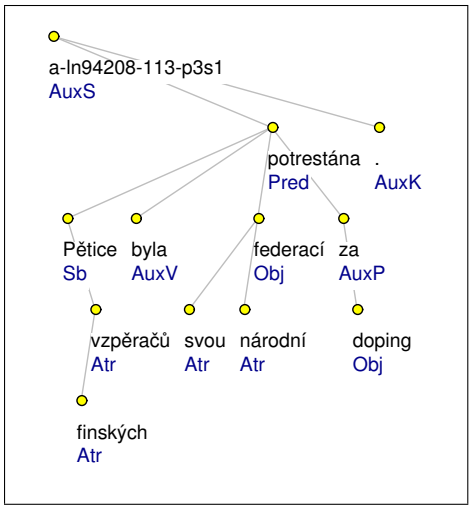
Příklady analytických funkcí:

- Predikát (přísudek)
- Subjekt (podmět)
- Objekt (předmět)
- Atribut (přívlastek)
- Adverbiale (příslovečné určení)
- ...

4.1.3 Tektogramatická rovina

Tektogramatická rovina je poměrně rozsáhlý koncept s hlubokou lingvistickou teorií v pozadí. Byla popsána už v roce 1961 v článku [8].

Tektogramatická rovina slouží k zachycení významové struktury věty. Struktura reprezentace zůstává stejná jako na analytické úrovni, avšak některé uzly se vypouští, některé se přidávají, a struktura věty může být obecně jiná, než na analytické úrovni. U vět které připouštějí více různých významů (víceznačné věty) je teoreticky možné vytvořit více tektogramatických stromů. V případě synonymie může naopak různým větám odpovídat tentýž tektogramatický strom. Tedy zatímco na morfologické rovině



Obrázek 4.1: Příklad notace na analytické rovině

jsou každému slovu věty přiřazeny jeho lema a tag (morfologická značka) a na analytické rovině každému slovu věty odpovídá uzel v analytickém stromě s příslušnou analytickou funkcí, tektogramatická rovina už tento těsný vztah k povrchovému zápisu věty nemá. Uzly tektogramatické roviny v sobě nesou informaci rozdělenou do několika atributů. Základními atributy uzlu tektogramatického stromu jsou tektogramatické lema, gramatémy a funktor. Vztah mezi uzly tektogramatické a analytické roviny (který je obecně typu M:N), je též zachycen v několika atributech uzlů tektogramatického stromu. Následuje podrobnější popis některých atributů

Tektogramatické lemma

Tektogramatické lemma (t-lemma) zachycuje lexikální význam uzlu. U jednoduchých uzlů odpovídá lemmatu, které bylo řídícímu slovu tektogramatického uzlu přiřazeno na morfologické rovině. Uzlům na tektogramatické rovině nově vytvořeným je přiřazeno zástupné t-lema speciálního tvaru.

Sémantický slovní druh a jeho podskupiny

Uzly tektogramatického stromu (respektive jejich řídící slova) se rozdělují do takzvaných *sémantických podskupin slovního druhu*. Toto dělení začíná roz-

dělením uzlů podle takzvaných *sémantických slovních druhů*. Z původních deseti slovních druhů, které v češtině rozlišujeme, vzniknou čtyři sémantické slovní druhy: sémantická substantiva, sémantická adjektiva, sémantická adverbia a sémantická slovesa. Tato se pak dále dělí do sémantických podskupin. Například sémantická substantiva se dělí na *pojmenovací*, *pronominální* a *kvantifikační*. Podrobně je celé toto rozdělení popsáno v [11].

Gramatémy

Gramatémy jsou tektogramatickým rozšířením morfologických značek. Gramatémy nalezneme pouze mezi atributy uzlů u kterých to má smysl, tedy u uzlů které se vztahují k nějakému významovému slovu věty. Navíc různým slovním druhům lze přidělit jen některé gramatémy (například nemá smysl určovat slovesný čas u podstatného jména). Podle toho, do které sémantické podskupiny slovního druhu daný uzel patří, je možné určit, které gramatémy pro něj mají smysl a které nikoli.

Funktor

Funktory jsou velkým přínosem tektogramatické roviny po praktické stránce. Funktory chápeme jako sémantické ohodnocení hran mezi uzly tektogramatického stromu. Tektogramatické funktory můžeme též chápat jako ekvivalent analytických funkcí. Rozdíl mezi tektogramatickými funktory a analytickými funkcemi je v tom, že funtoktory se snaží postihnout sémantiku vztahu, zatímco analytické funkce se zaměřují na jeho syntaktickou roli. Následuje popis některých důležitých funktorů vždy s několika příklady jejich výskytu ve větě. Vyčerpávající seznam je možné nalézt například v [11].

- Funktor ACT (actor) označuje původce děje, nositele děje nebo vlastnosti.
  - Její *manžel.*ACT tam však pracuje dál.
  - Ten *román.*ACT mě oslovil.
  - Českým *skokanům.*ACT se dařilo dobře.
  - Je *mi.*ACT smutno.

- Funktor ADDR (addressee) odpovídá roli příjemce děje.
  - Dal *dítěti.*ADDR hračku.
  - Učí *děti.*ADDR angličtinu.
  - Obrátil se na *soud.*ADDR s problémem.
- Funktor PAT (patiens) označuje předmět dějem zasažený.
  - Snědl *polévku.*PAT
  - Neubližujte *zvířatům.*PAT
  - Učil se *kominíkem.*PAT
  - Mít dost *peněz.*PAT
- Funktor MANN (manner) vyjadřuje, hodnotí způsob provedení děje.
  - Pracuje *pomalu.*MANN
  - *Nějak.*MANN to uděláme.
  - *Prudce.*MANN se zvýšily mezibankovní úrokové míry.
- Funktor TWHEN (temporal : when) vyjadřuje časové určení odpovídající na otázku "kdy?".
  - *Zítra.*TWHEN má být už hezky.
  - *Hned.*TWHEN se vrátím.
  - Součástka se *časem.*TWHEN opotřebuje.
- Funktor LOC (locative) označuje místo, do kterého je děj nebo stav vyjádřený řídícím slovem lokalizován.
  - Zůstaň *doma.*LOC
  - *Nalevo.*LOC stál pěkný dům.
  - *Místy.*LOC ležel v ulicích ještě sníh.
- Funktor DIR1 (directional: from) vyjadřuje určení místa odpovídající na otázku "odkud?".
  - Přijel z *Prahy.*DIR1

- Funktor DIR2 (directional: which way) vyjadřuje určení místa odpovídající na otázku "kudy?".

  - Jdou *lesem*.DIR2

- Funktor DIR3 (directional: to) vyjadřuje určení místa odpovídající na otázku "kam?".

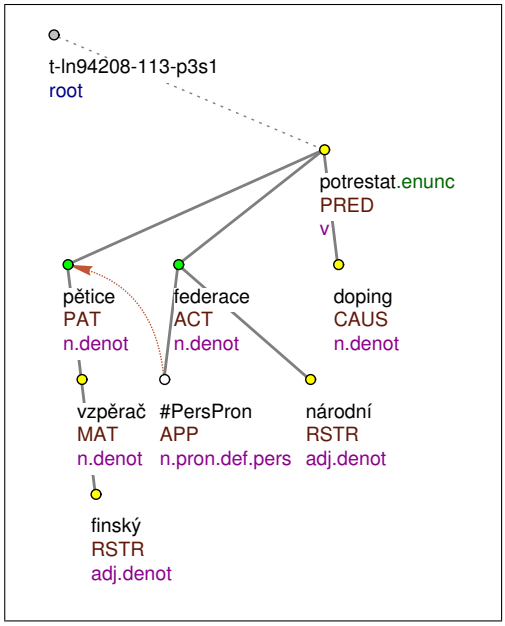
  - Přišel *domů*.DIR3

- Funktor RSTR volně doplňuje blíže specifikující řídící substantivum.

  - *velký*.RSTR dům

- Funktor CONJ (conjuction) je kořen souřadné struktury (tektogramatického podstromu), která reprezentuje spojení dvou a více obsahů.

  - Jezte ovoce *a*.CONJ zeleninu.



Obrázek 4.2: Příklad notace na tektogramatické rovině

4.1.4 Příklady

Pro ilustraci předkládáme dva obrázky závislostních stromů, které vznikli ruční anotací věty:

> *Pětice finských vzpěračů byla svou národní federací potrestána za doping.*

Na obrázku 4.1 je strom analytické roviny a na obrázku 4.2 je strom tektogramatické roviny. Za povšimnutí stojí rozdíl v počtu uzlů obou stromů, který je na tektogramatické rovině o něco menší. Oba tyto příklady pocházejí z PDT 2.0 - sample data. Obrázky jsou vygenerované pomocí editoru TrEd (viz 4.4.2).

Kromě struktury jednotlivých stromů v této kapitole popisovaných jsou na obrázcích vidět i další především technické jevy spojené s konkrétní reprezentací lingvistické anotace. Například tetogramatické funktory a analytické funkce nejsou přiřazeny k hranám stromu ale k závislému uzlu, každý strom má navíc takzvaný *technický kořen*, na kterém teprve skutečný kořen lingvistického stromu visí. Technický kořen nese administrativní atributy stromu (například atribut *id* – jednoznačný identifikátor). U analytického stromu též stojí za povšimnutí umístění tečky na konci věty, která je potomkem umělého kořene. Uzly obou stromů jsou zleva doprava uspořádány podle pořadí, v jakém se slova odpovídající daným uzlům vyskytují v původní větě.

Schéma analytického stromu je jednodušší. Pod každým uzlem jsou vytištěny hodnoty dvou atributů (hodnoty dvou lingvistických značek). První atribut obsahuje původní *tvar slova* a v druhém je zkratka *analytické funkce* tohoto uzlu.

Schéma tektogramatického stromu je o něco složitější. Pod každým uzlem jsou vytištěny hodnoty tří atributů. První atribut obsahuje *lemma*, prostřední tektogramatický *funktor* a poslední vyjadřuje *sémantickou podskupinu slovního druhu*, do které řídicí slovo tohoto uzlu patří. Podrobnosti o těchto atributech a jejich hodnotách je možné nalézt v [11].

Na obrázku 4.2 (tektogramatického stromu) je vidět speciální zahnutá šipka od uzlu *#PersPron* k uzlu *pětice*. Konec této šipky ukazuje na cíl, na který odkazuje zájmeno uvnitř prvního (počátečního) uzlu.

4.2 The Prague Dependency Treebank

Pražský závislostní korpus (PDT) je probíhající projekt Centra počítačové lingvistiky Ústavu formální a aplikované lingvistiky (ÚFAL) v Praze[2]

Náplní projektu je především ruční lingvistická anotace velkého množství českých textů. Projekt se vyznačuje velkou hloubkou anotace, která sahá až po tektogramatickou rovinou. Kromě velkého množství anotovaných textů bylo v souvislosti s projektem vyvinuto i množství užitečných nástrojů pro práci s anotacemi a nástroje, které umožňují automatickou lingvistickou anotaci českého textu.

Historie projektu PDT začíná v roce 1995. Od té doby se korpus PDT rozšířil až na současnou (PDT 2.0) velikost 2 milióny slov s provázanými anotacemi na úrovni morfologie (2 milióny slov), povrchové syntaxe (1,5 mil. slov) a hloubkové syntaxe a sémantiky (0,8 mil. slov). Poprvé byl korpus PDT (verze 0.5) veřejně představen v roce 1998. V roce 2001 bylo publikováno CD-ROM PDT 1.0, které obsahovalo přibližně 1,5 mil. slovních jednotek anotovaných na analytické rovině.

V roce 2006 byla publikována poslední současná verze korpusu jako CD-ROM PDT 2.0 [12]. Korpus PDT je v této verzi rozšířen o tektogramaticky anotvaná data. CD-ROM PDT 2.0 dále obsahuje množství kvalitních lingvistických nástrojů (viz 4.4) a publikací, mezi které patří obsáhlý manuál (více než 1000 stran) pro tektogramatické značkování [11].

4.3 Jazyky pro zápis lingvistických anotací

Nyní krátce rozvedeme, v jakých formátech se lingvistické anotace uchovávají. Jedná se většinou o formáty vzniklé pro potřeby korpusu PDT a příbuzných nástrojů. Podrobnosti[3] o těchto formátech je možné nalézt v Průvodci PDT 2.0 [12].

[2] http://ufal.mff.cuni.cz/

[3] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/

4.3.1 CSTS - Czech Sentence Tree Structure

Formát zvaný CSTS, založený na SGML[4], byl hlavním formátem dat v PDT 1.0. Nyní (v PDT 2.0) je používán jen jako přechodný formát pro kompatibilitu se staršími nástroji pro zpracování jazyka (taggery, parsery, ...). CSTS může reprezentovat jen morfologickou a analytickou anotaci, není schopen plného popisu tektogramatické roviny.

4.3.2 PML - The Prague Markup Language

Hlavním formátem dat v PDT 2.0 je formát nazvaný PML. PML je založený na XML, je navržený pro reprezentaci bohaté lingvistické anotace textů, jako jsou morfologické značkování, závislostní stromy apod. V PML se mohou jednotlivé oddělené roviny anotace překrývat a mohou být konzistentně propojeny jak mezi sebou, tak i s dalšími zdroji dat. Každá rovina anotace je popsána v souboru PML schéma, který je jakousi formalizací abstraktního anotačního schématu pro tu konkrétní rovinu anotace.

Anotace PDT 2.0 je rozdělena do čtyř rovin, naskládaných jedna na druhou. Každá z těchto rovin má vlastní PML schéma a zpravidla se ukládá do zvláštního soubotu. Jedná se o tyto čtyři roviny: rovina slovní (soubory *.w*), rovina morfologická (soubory *.m*), rovina analytická (soubory *.a*) a rovina tektogramatická (soubory *.t*). Pro podrobnosti o rovinách lingvistické anotace viz 4.1.

Další informace je možné nalézt na stránkách PML projektu[5] případně v publikaci [17].

4.3.3 FS - Feature Structure

Formát FS („feature structure"“) je formát souborů pro reprezentaci stromů, jejichž uzly jsou struktury atribut-hodnota. Může být chápán jako "meta formát", podobně jako SGML nebo XML. Konkrétní použití tohoto formátu je plně specifikováno deklarací atributů v hlavičce FS souboru. Formát FS byl primárně vytvořen pro vyhledávací program Netgraph (viz 4.4.1).

[4] http://www.w3.org/MarkUp/SGML/
[5] http://ufal.mff.cuni.cz/jazz/PML/

4.3.4 PLS - Perl Storable Format

Čistě z důvodů optimalizace a časové úspory se při práci s nástroji TrEd a btred požívá formát PLS. Nástroje TrEd a btred jsou založeny na Perlu, při načítaní dat ve formátu PML a převodu PML dat do vnitřní paměťové reprezentace Perlu se spotřebuje mnoho času. Této časově náročné transformaci se lze vyhnout využitím formátu PLS (Perl Storable Format). Jde o binární datový formát, který přímo odráží vnitřní paměťovou reprezentaci dat v Perlu. Jeho ukládání a zpětné načítání nástroji TrEd a btred je tedy mnohem rychlejší.

4.3.5 Konverze mezi formáty PDT

Problém s konverzí mezi formáty lingvistické anotace je v tom, že všechny formáty nemohou nést přesně stejné množství informací. Přesto jsou v projektu PDT 2.0 zahrnuty skripty pro konverzi některých formátů:

- konverze analytické anotace typu PDT 1.0 do PML
- konverze a-dat PML do CSTS
- konverze m-dat PML do CSTS
- konverze dat PDT 2.0 do FS pro Netgraph
- konverze dat PDT 2.0 do PLS

4.4 Lingvistické nástroje

Nyní popíšeme některé lingvistické nástroje, které mohou pomoci při zpracování textů přirozeného jazyka a extrakci informací z nich. Většina těchto nástrojů je vyvíjena na Ústavu formální a aplikované lingvistiky (ÚFAL). Tyto nástroje je možné (kromě tektogramatického analyzátoru 4.4.4) získat jakou součást PDT 2.0 CD-ROM [12].

4.4.1 NetGraph

Netgraph je aplikace typu klient-server, která umožňuje prohledávat korpus podobný PDT (anotace mají strukturu závislostního stromu) současně několika uživateli, připojenými přes internet. Netgraph je navržený tak, aby prohledávání bylo co nejjednodušší a intuitivní, při zachování vysoké síly dotazovacího jazyka. Funkčnost aplikace je rozdělena na část, kterou vykonává klient, a na část, kterou vykonává server.

Netgraph klient je napsán v Javě a je nezávislý na platformě. Existuje ve dvou formách - jako samostatná Java aplikace a jako Java applet. Applet verze je oproti plné Java aplikaci ochuzena o některé funkce, přesto však poskytuje plnou vyhledávací sílu. Funkce klienta zahrnuje vytvoření (návrh) dotazu, jeho odeslání serveru a zobrazení, případně další zpracování výsledků vrácených serverem.

Netgraph server je napsán v C a C++ a běží v operačním systému Linux i dalších systémech - podrobnosti je možné nalézt v [14]. Umožňuje nastavit uživatelská konta s různými přístupovými právy. Korpus, určený k prohledávání Netgraphem, musí být ve formátu FS (viz 4.3.3). Funkce serveru spočívá ve vyhodnocování dotazů zasílaných klienty. Server prohledává korpus a stromy, které vyhovují dotazu vrací jako odpověď.

Dotazy v Netgraphu jsou definovány pomocí vlastního dotazovacího jazyka. Jedná se o jazyk formálně velmi jednoduchý avšak s vysokou expressivitou. Definovat dotaz v Netgraphu znamená definovat podstrom, který se má v prohledávaných stromech vyskytovat. Tedy v dotazu můžeme definovat požadovanou strukturu stromu. Navíc můžeme v každém uzlu dotazu vynutit hodnotu některých atributů tohoto uzlu.

Velmi jednoduchý dotaz, kdy chceme vyhledat všechny stromy obsahující slovo „hasič“ se skládá z jediného uzlu a restrikce na hodnotu atributu *lemma* (viz 4.1.1) na hodnotu *hasič* v tomto uzlu.

Podrobnosti o možnostech a syntaxi tohoto dotazovacího jazyka je možné nalézt například v [14]. Poznamenejme ještě, že dotazy mohou být dále rozšířeny tzv. meta atributy, které umožňují určení pozice dotazu v nalezených stromech, omezení velikosti nalezených stromů, určení vztahů mezi hodnotami atributů u různých uzlů v nalezených stromech, negaci a mnoho dalších podmínek.

Dotazy se v Netgraph klient vytvářejí v uživatelsky přívětivém grafickém prostředí. Uživatel si zde může „naklikat“ celý strom, který se má při vyhod-

nocování dotazu hledat. V grafickém rozhraní Netgraph klient má uživatel snadný přístup k možnostem dotazovacího jazyka, aniž by musel tento jazyk podrobně znát.

Další informace o aplikaci Netgraph je též možné nalézt na její domovské stránce[6].

4.4.2 Tree Editor TrEd

Tree Editor TrEd je velmi komplexní grafický editor, který umožňuje rychlé, pohodlné a flexibilní procházení, prohlížení a úpravu stromů v korpusech podobným PDT. TrEd prvotně sloužil jako hlavní anotační nástroj PDT, ale může být použit i k prohlížení dat a obsahuje několik druhů vyhledávacích funkcí. TrEd se vyznačuje svými možnostmi nastavení a přizpůsobení celé aplikace na míru potřebám nejrůznějších uživatelů. Silný nástroj představují uživatelská makra, která mohou být do aplikace kýmkoliv doprogramována v jazyce Perl. TrEd podporuje velké množství vstupních a výstupních formátů dat, jmenujme například FS, CSTS, PDT-PML (podrobnosti k jednotlivým formátům – viz 4.3).

TrEd je možné nainstalovat na většině v současné době používaných operačních systémů: Linux, UNIX (MacOS X, BSD, Solaris, ...) i Windows (funguje díky ActivePerl for Windows, který musí být v systému nainstalovaný). Na domovské stránce[7] editoru TrEd lze získat jednotlivé instalační balíčky i podrobné instrukce pro instalaci na konkrétní operační systém.

Ukázkové obrázky

Na obrázcích 4.1 a 4.2 jsou schémata stromů získaná přímo z editoru TrEd. Takto jsou při výchozím nastavení v TrEd editoru zobrazovány analytické a tektogramatické stromy.

btred / ntred

Součástí editoru TrEd jsou též dva nástroje - *btred* a *ntred*, které umožňují automatické (dávkové) zpracování stromů korpusu. Ntred je pouze rozšíře-

[6] http://quest.ms.mff.cuni.cz/netgraph/
[7] http://ufal.mff.cuni.cz/~pajas/tred/

ním nástroje btred o možnost zpracovávat korpus paralelně více počítači v síti najednou.

Tyto nástroje se spouštějí přímo z příkazové řádky, nemají grafické rozhraní. Činnost těchto nástrojů, je řízena uživatelským programem (makrem btred-u), které uživatel musí napsat v programovacím jazyce Perl. Při psaní toho makra má uživatel k dispozici velké množství specializovaných funkcí pro práci se strukturami korpusu: s jednotlivými stromy, s uzly stromů, s atributy uzlů atp.

Přehledný a dobře srozumitelný návod – „btred/ntred tutorial“[8], jak pracovat s nástroji btred a ntred je možné nalézt na domovských stránkách editoru TrEd.

4.4.3 Tools for machine annotation - PDT 2.0

Jedná o skupinu nástrojů, které provádějí plně automatickou lingvistickou analýzu českého textu. Ze surových českých vět vytvářejí závislostní stromy na analytické rovině. Proces anotace se skládá z následujících funkcí.

1. Rozpoznání slovních jednotek ve vstupním surovém textu a rozdělení textu na věty.
2. Morfologická analýza a tagging (morfologická desambiguace).
3. Závislostní parsing.
4. Přiřazení analytických (závislostních) funkcí všem uzlům zparsovaného stromu.

Tyto funkce jsou implementovány v celkem šesti oddělených nástrojích. Vstupem každého nástroje je vždy výstup předchozího s výjimkou prvního, jehož vstupem je prostý text. Nástroje jsou napsány z části v Perlu, zbytek tvoří přeložený kód (C++) pro Linux běžící na i386 architektuře.

Celý řetěz nástrojů se dá spustit jediným skriptem *run_all*.

Tyto nástroje a jejich podrobný popis[9] (včetně naměřené chybovosti) jsou k dispozici jako součást PDT 2.0 CD-ROM [12].

Následuje podrobnější popis jednotlivých nástrojů.

[8] http://ufal.mff.cuni.cz/~pajas/tred/bn-tutorial.html
[9] http://ufal.mff.cuni.cz/pdt2.0/doc/tools/machine-annotation/

Segmentation and tokenization

Provádí rozdělení textu na slova a interpunkční znaménka (tokenizace) a rozdělí tyto tokeny do vět (segmentace).

Morphological analysis

Pro každé slovo vyhledá všechna možná lemmata a morfologické značky, která by mu mohly odpovídat.

Morphological tagging

Ze všech možných alternativ získaných v předchozím kroku pro každé slovo vybere jedno lemma a morfologickou značku. Tento proces se často nazývá desambiguace. Tagging pro Češtinu je poměrně zajímavý vědecký problém, který je podrobně rozpracován v mnoha publikacích[10].

Parsing

Morfologicky označkovaná slova v každé větě uspořádá do závislostního stromu. Problém automatického závislostního parsingu[11] je stále poměrně živý. Aktuálně nejlepší parser [13] dosahuje přesnosti přibližně 86%

Analytical function assignment

Jednotlivým hranám závislostního stromu, které vznikly v předchozím kroku, přiřadí funktory analytické roviny. Nástroj pracuje jako klasifikátor založený na rozhodovacím stromu. Řídící rozhodovací strom byl vytvořen pomocí Quinlanova C5 klasifikátoru z dat PDT 1.0.

Conversion into PML

Zapíše výstup předchozího nástroje v PML jazyce. Pro podrobnější informace o PML viz oddíl 4.2.

[10] http://ufal.mff.cuni.cz/czech-tagging/
[11] http://ufal.mff.cuni.cz/czech-parsing/

4.4.4 Nástroj pro tektogramatickou analýzu češtiny

Jedná o nástroj, který provádí automatickou tektogramatickou lingvistickou anotaci. Jako vstup akceptuje na analytické rovině anotovaná data uložená ve formátu PML. Tedy dokáže výstup nástrojů výše (4.4.3 – Tools for machine annotation) povýšit na tektogramatickou rovinu.

Nástroj je založený na strojovém učení, pro které byl použit nástroj *fnTBL toolkit*[12] [16]. Pro češtinu bylo učení realizováno na datech PDT 2.0. Podrobnosti o algoritmu a jeho úspěšnosti je možné nalézt v článku [15].

Autorem tohoto nástroje je Václav Klimeš[13]. U něho je možné tento nástroj získat společně s dalšími informacemi a instrukcemi pro instalaci. Poslední verze (rok 2007) byla určena pro operační systém Linux.

[12] http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html
[13] http://ufal.mff.cuni.cz/~klimes/

Kapitola 5

Lingvistika a znalostní inženýrství

5.1 Znalosti formulované v přirozeném jazyce

Pri hledani nejake vety k ukazkove anotaci jsem narazil na tuhle: (je to z prihlasky na DS)

Uchazec vyplni obor studia, vyzkumne tema, skalici pracoviste, zajisti podpis skolitele a podpis predsedy ...

Tahle veta je zvlastni v tom, ze je formulovana obecne: pro vsechny uchazece, pro libovolne tema, pracoviste, podpis libovolneho skolitele, predsedy. Avsak skolitel je pevne spojen s tematem prace a predseda je spojen s pracovistem.

Jak anotovat takovou vetu? Jaka je jeji semantika? Vznikne nejaky Abox? Nebo v anotaci pouzijeme nejake volne promenne pro individua. Nebo budeme anotovat pouze pomoci nazvu trid a instance nejakym zpusobem vynechame?

5.2 Rozdíl mezi konkrétním a abstraktním

Kapitola 6

WordNet

WordNet [18], [19] je lexikální databáze vybudovaná na základě psycholexikologického výzkumu o lidské lexikální paměti. Jazykové jednotky nejsou ve WordNetu uspořádány abecedně, ale podle jejich sémantických vztahů, tedy hierarchicky a shlukově. Tento typ lexikální databáze se často označuje jako *sémantická síť*.

WordNet jakožto sémantická síť obsahuje pouze slova, která nesou nějaký kognitivní význam, tedy podstatná jména, přídavná jména, příslovce a slovesa. Navíc jsou ve WordNetu obsažena i slovní spojení (sousloví), například „vysoká škola“. V dalším textu si však pro lepší přehlednost dovolíme zjednodušení a o všech těchto jazykových výrazech, které můžeme ve WordNetu nalézt budeme mluvit jako o slovech.

Lexikální matice

Základním formálním prostředkem pro zachycení významu slova je *lexikální matice*. Řádky této matice tvoří jednotlivé významy, sloupce jednotlivá slova. Záznam lexikální matice na souřadnicích $[i\ ,\ j]$ znamená, že slovo j nese význam i. Pokud se objeví dva záznamy na stejném řádku, znamená to, že odpovídající dvě slova mají stejný význam, jsou synonymní. Pokud se naopak objeví více záznamů v jednom sloupci, znamená to, že toto slovo nese více možných významů, je polysémické.

Synset

Záznamy ve WordNetu jsou organizovány podle významu, tedy podle řádků lexikální matice. Každý takový řádek ve WordNetu označujeme jako *synset* (množina synonym nebo též synonymická řada). Synset je pro WordNet tímtéž, čím je heslo pro obyčejný slovník.

Jelikož různé slovní druhy nemohou být synonymi v pravém slova smyslu, je sémantická síť WordNetu budovaná pro každý slovní druh zvlášť.

Číslování významů - literály

Jednotlivé prvky synsetů označujeme jako *literály*. Literál reprezentuje jeden záznam v lexikální matici, tedy dvojici slovo-význam. Literál budeme chápat jako slovo v daném významu.

Literály resp. významy daného slova se ve WordNetu číslují. Například anglické podstatné jméno *bank:1* označuje finanční instituci, *bank:2* – břeh.

Sémantické vazby

Wordnet obsahuje celou řadu sémantických vazeb mezi literály a zejména mezi synsety. Vzniká tak síť slov, tedy WordNet. Popišme nyní tyto vazby podrobněji.

- *Hyperonymie a hyponymie* jsou vztahy významové nadřazenosti a významové podřízenosti. Například *flanel* je druhem *textilie*. Vztahy hyperonymie a hyponymie vytvářejí základní hierarchickou strukturu WordNetu pro podstatná jména. V zásadě se jedná o stromovitou strukturu, kde blíže ke kořenu stromu znamená obecnější a blíže k listům znamená specifičtější. Tomuto vztahu se někdy též říká *lexikální dědičnost*. Příklad stromu lexikální dědičnosti je na obrázku 6.1.

- *Meronymie a holonymie* vyjadřují vztah mezi celkem a částí. Tedy například slovu *dům* je slovo *okno* meronymum a *město* holonymum.

- *Antonymie* vyjadřuje sémantickou protikladnost dvou synsetů. Například slova *mokrý* a *suchý* jsou antonymická.

6.1 Princeton WordNet

Duchovním otcem WordNetu je George A. Miller z univerzity v Princetonu. Zde je též pod Millerovým vedením stále vyvíjen a rozšiřován první a současně největší (americký) *Princeton WordNet*[1] (PWN). Současná verze WordNet 3.0 obsahuje 207 016 literálů (párů slovo-význam) v 117 597 synsetech. Data PWN jsou princentonským týmem poskytována volně.

6.2 EuroWordNet

Cílem projektu EuroWordNet[2] [20] bylo vytvořit WordNety pro další evropské jazyky a provázat je do multilingvální databáze.

Tento projet začal v roce 1997. V první fázi byly zpracovány jazyky: britská angličtina, holandština, italština a španělština, ve druhé pak čeština, estonština, francouzština a němčina. V roce 2001 byla tato činnost ještě rozšířena projektem BalkaNet[3] o dalších pět balkánských jazyků (bulharštinu, rumunštinu, řečtinu, srbštinu a turečtinu).

Velká snaha byla věnována co možná nejúplnějšímu provázání významů napříč různými jazyky. Společným podkladem všem novým slovníkům byl PWN 1.5. V něm každý synset dostal jednoznačný identifikátor. Tyto identifikátory sloužily pro vytváření ekvivalencí mezi synsety PWN a ostatních jazyků. Tak vznikl takzvaný mezijazykový index (Inter-Lingual Index, ILI).

Dalším vylepšením nových WordNetů bylo rozšíření počtu sémantických relací v rámci jednoho jazyka, tzv. Inter-Lingual Relations (ILR).

- Přibyla relace *near synonym*, která je určena k propojení literálů a synsetů, jejichž význam je sice blízký, avšak o úplná synonyma se nejedná.
- Dále vznikl soubor relací, které propojují synsety *napříč slovními druhy*. Tyto relace jsou užitečné pro zachycení slovotvorných vztahů. Vytvářejí se tak shluky slov odvozených od stejného slovního základu. Například *učit – učitel – učitelský*.

[1] http://wordnet.princeton.edu
[2] http://www.illc.uva.nl/EuroWordNet/
[3] http://www.ceid.upatras.gr/Balkanet/

6.3 Český WordNet

Český WordNet [21] začal pod vedením doc. Karla Paly vznikat v roce 1998 na Fakultě informatiky Masarykovy univerzity v Brně[4] v rámci druhé fáze projektu EuroWordNet. V současné době obsahuje český WordNet přibližně 30 000 synsetů.

Online interface k českému WordNetu je dostupný přes internet, přístupný po domluvě podmínek s vedoucím projektu doc. Karlem Palou[5]. K dispozici je webové rozhraní a jednoduché dobře dokumentované programátorské API[6], skrz které má programátor přístup ke všem funkcím online databáze WordNet.

V rámci projektu DEB II je vyvíjen i vizuální prohlížeč a editor online WordNetu DEBVisDic, který je volně k dispozici na stránkách[7] projektu.

6.3.1 SAFT - Semantic Analyzer of Free Text

Zajímavý experiment s českým WordNetem provedl Tomáš Čapek ve své práci [22]. V této práci představuje nástroj SAFT - Semantic Analyzer of Free Text. Vstupem tohoto nástroje je text přirozeného jazyka, k slovům vstupního textu SAFT vyhledává jejich významy ve WordNet databázi.

Experiment spočíval ve spuštění tohoto nástroje na část českého korpusu DESAM (vyvinutého na Fakultě informatiky MU Brno). Ukázalo se, že přibližně 50% slov není v českém WordNetu zastoupeno vůbec, avšak uvědomíme-li si, že WordNet pokrývá pouze výrazy nesoucí kognitivní význam (tedy podstatná jména, přídavná jména, příslovce a slovesa), není tento výsledek tak špatný. Autor dokonce tvrdí, že většina nenalezených slov patří právě do kategorie slov bez kognitivní sémantiky.

[4] http://www.fi.muni.cz
[5] http://www.muni.cz/fi/people/Karel.Pala
[6] http://nlp.fi.muni.cz/trac/deb2/wiki/WordNetApi
[7] http://nlp.fi.muni.cz/projekty/deb2/

```
entita:1
  objekt:1
    celek:1
      artefakt:1, výtvor:2, výrobek:2
        vybavení:2
          přepravní prostředek:1, transportní prostředek:1
            veřejná doprava:1
               autobus:1, autokar:1
            dopravní prostředek:1
               kolové vozidlo:1
                 samohybné vozidlo:1, vozidlo s vlastním pohonem:1
                   motorové vozidlo:1
                     nákladní automobil:1
                       kamion:1
```

Obrázek 6.1: Příklad stromu lexikální dědičnosti v českém WordNetu pro slova kamion a autobus

6.4 Kritika WordNetu

WordNet bývá často kritizován ([5], [23]) z různých důvodů. Tradiční lexikografové vidí mnoho problémů v nejasně definované (a v čase se měnící) koncepci tvorby hesel (synsetů), ve kterých panuje značný nepořádek (synonyma nejsou přesnými synonymy, hyponyma jsou nestejnorodá, klasifikace nenavazují na běžné oborové klasifikace).

V našem experimentu (viz kapitola 7) s českým WordNetem jsme se potýkali s nedostatečným pokrytím české slovní zásoby. Provázání synsetů sémantickými hranami je v českém WordNetu též poměrně řídké. Například nejbližší společné hyperonymum pro slova *autobus* a *kamion* není synset *motorové vozidlo:1* ani *dopravní prostředek:1* ale až synset *přepravní prostředek:1, transportní prostředek:1*. Na obrázku 6.1 je vidět stromu lexikální dědičnosti tato dvě slova.

Kapitola 7

Experiment

Experiment provedený v rámci této práce spočíval v otestování některých dostupných nástrojů pro lingvistickou anotaci českých textů. Jmenovitě se jednalo o tyto nástroje: !!!!!!!!!!!!doplnit!!!!!!!!!!!!!!!!. Byly prozkoumány možnosti vyžití těchto nástrojů pro extrakci informací a sémantickou anotaci.

!!!!!!!Pzor následující odstavec jsem okopíroval i do kapitoly přínosy - vyřešit

Postup experimentu se v jednotlivých fázích snaží kopírovat skutečné akce, které by bylo nutné provést v opravdovém projektu zaměřeném na sémantickou anotaci. V práci tak vzniká jednoduchá základní analýza tohoto typu projektů. Ve skutečném projektu pak bude možné ji přinejmenším jako inspiraci využít.

Pro experiment byla vybrána a použita data ze dvou poměrně odlyšných zdrojů.

Extrakce a čištění (zamyšlení nad různými formáty zdroje PDF, DOC, HTML, XML, částečné řešení v GATE softu)

Lingvistická nanotace

Extrkace informací

Sémantické anotace - víceméně jen teoreticky

7.1 Vstupní data

Volba zdrojových textů

Proč hasiči?

Proč ne korpus PDT? — Důraz byl kladen na co možná největší se přiblížení k podmínkám a problémům skutečného projektu. V takovém případě bychom se těžko mohli opřít o to, že by nám data která chceme analyzovat někdo ručně lingvisticky anotoval.

+ pokusy s PDT sample data. Nicméně pokusy nad ručními lingvistickými anotacemi dat PDT[1] proběhly.

7.1.1 Hasiči

7.1.2 Úpadci

7.2 Výstupní data

Vzhledem k tomu, že pojem sémantické anotace, jak ho zmiňuji v kapitole 3, je velmi široký, není ani přesně určeno, jaká data by při procesu sémantické anotace měla vzniknout.

7.3 Software

7.3.1 Modul XY

[1] Jedná se o *sample data* PDT 2.0, http://ufal.mff.cuni.cz/pdt2.0/data/sample/

Seznam obrázků

Literatura

[1] Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American, 2001.

[2] S. Handschuh, S. Staab (edited by). Annotation for the Semantic Web. Volume 96 Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam, The Netherlands, 2003. ISBN 1-58603-345-x.

[3] Lawrence Reeve, Hyoil Han. Survey of Semantic Annotation Platforms. SAC '05: Proceedings of the 2005 ACM symposium on Applied computing, ACM Press, New York, USA, 2005. 1634–1638
URL http://dx.doi.org/10.1145/1066677.1067049

[4] Franz Baader, Werner Nutt. Basic Description Logics. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider, editors, The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2003.
URL http://citeseer.ist.psu.edu/baader03basic.html

[5] P. Matulík, T. Pitner, P. Smrž. Sémantický web a jeho technologie (1,2,3). Zpravodaj ÚVT MU. ISSN 1212-0901, 2004, roč. XIV, č. 3,4,5. 15–17, 9–13, 14–16.
URL http://www.ics.muni.cz/zpravodaj/issues/serials.html#2

[6] E. Hajíčová, M. Plátek, P. Sgall: Komunikace s počítačem v češtině, Sborník referátov seminára SOFSEM 81, Výzkumné výpočtové stredisko Bratislava, 1981. 85–114.

[7] P. Sgall a kolektiv: Úvod do syntaxe a sémantiky, Academia, Praha, 1986.

[8] Curry, H. B.: Some logical aspects of grammatical structure, Structure of Language and Its Mathematical Aspects (red. R. Jakobson), Proceedings of Symposia in Applied Mathematics 12. American Math. Society, Providence, RI 1961.

[9] Daniel Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Barbora Hladká, Emil Jeřábek. A Manual for Morphological Annotation, 2nd edition. Technical Report 27, ÚFAL MFF UK, Prague, Czech Republic, 2005. URL http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/pdf/m-man-en.pdf.

[10] Eva Hajíčová, Zdeněk Kirschner, Petr Sgall. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL, MFF UK, Prague, Czech Republic, 1999. URL http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf.

[11] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářova-Řezníčková, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský. Anotace Prazžského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory [A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank]. Technical report, ÚFAL, MFF UK, Prague, Czech Republic, 2005. URL http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/pdf/t-man-cz.pdf.

[12] Jan Hajič, Eva Hajičová, Jaroslava Hlaváčová, Václav Klimeš, Jiří Mírovský, Petr Pajas, Jan Štěpánek, Barbora Vidová Hladká, Zdeněk Žabokrtský. Prague Dependency Treebank 2.0, CDROM, Linguistic Data Consortium, 2006. In press. URL http://ufal.mff.cuni.cz/pdt2.0/.

[13] Daniel Zeman, Zdeněk Žabokrtský. Improving Parsing Accuracy by Combining Diverse Dependency Parsers. Proceedings of the International Workshop on Parsing Technologies (IWPT 2005). Association for Computational Linguistics, Vancouver, British Columbia.

[14] Jiří Mírovský. Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0. Proceedings of The Fifth International Treebanks and Linguistic Theories conference, Prague, Czech Republic, 2006. 211–222.

[15] Václav Klimeš. Transformation-Based Tectogrammatical Analysis of Czech. Proceedings of Text, Speech and Dialogue 2006, Springer-Verlag, Berlin Heidelberg, 2006. ISSN 0302-9743.

[16] Grace Ngai, Radu Florian. TransformationBased Learning in the Fast Lane. Proceedings of NAACL 2001, Pittsburgh, PA, 2001. 40–47.

[17] Petr Pajas, Jan Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information, ELRA, Genoa, Italy, 2006.

[18] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller. Five papers on wordnet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[19] Christiane Fellbaum (editor). WordNet: An Electronic Lexical Database. Bradford Books, The MIT Press, 1998. ISBN 0-262-06197-x.

[20] Vossen P. EuroWordNet: a multilingual database for information retrieval. In Proceedings of DELOS workshop on Cross-language Information Retrieval, 1997.

[21] Pala K., Ševeček P. The Czech WordNet, final report. Technical report, Masarykova univerzita, Brno, 1999.

[22] Tomáš Čapek. Systém pro částečné sémantické značkování volného textu. Diplomová práce, Fakulta informatiky, Masarykova univerzita, Brno, 2006.

[23] Martin Ph. Correction and Extension of WordNet 1.7. ICCS 2003, 11th International Conference on Conceptual Structures, Springer Verlag, Dresden, Germany, 2003. 160–173 URL http://www.webkb.org/doc/papers/iccs03/