Obsah

Kapitola 1

Úvod

Sémantická anotace je .....

Sémantický web

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 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 poskytuje. 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 dovědět vice.

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.

XXXXX

Základní přehled ontlogií.

Doporučení autorům stánek (DRFA, HTML-A)

1.2.2 Průzkum

Součástí práce je praktický experiment s lingvistickými nástroji. Čtenáři jsou poskytnuty zkušenosti z týkající se použití těchto nástrojů a z prací které s jejích použitím souvisely. Tyto zkušenosti se týkají především dostupnosti, zprovoznění, výkonnosti, přínosů a nedostatků těchto nástrojů.

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.

Kapitola 2

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 2.2, kde bych se o takové vymezení chtěl pokusit. V knize [1] 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 !!!!!!!!!!!!!!doplnit!!!!!!!!!!!!!!!!!!!!!!!!.

Autorská anotace.

Dodatečná anotace, strojová.

2.1 Rozdělení

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

Po teoretické stránce: čísla (úspěšnost), metody

Labský (pouze extrakce informací, praktický, granty)

2.2 Teorie

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

XXX

rozdíl mezi konkrétním a abstraktním

Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Zbytek rozpracováno

Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

V cem tkvi semanticnost anotace?

Co si predstavujeme pod idealni semantickou anotaci?

Jaky je vztah lingvisticke a semanticke anotace? Jsou mezi nimi hranice? Kde priblizne?

Dal by se z toho odvodit nejaky univerzalnejsi navod/algoritmus, jak od lingvisticke anotace k te semanticke prejit?

Pokusit se stanovit podminky ktere idealni anotaci brani, resp. predpoklady ktere by ji umoznily.

Jaky je rozdil mezi prirozenym jazykem a deskripcni logikou?

Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

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?

Kapitola 3

Lingvistická anotace

Lingvistickou anotací budu ve své práci označovat činnost při které se text přirozeného jazyka obohacuje o lingvistickou informaci o slovech, větách, vztazích mezi slovy, mezi větami apod. Lingvistická anotace nebo též značkování korpusu je jedna z činností korpusové lingvistiky. Korpusem rozumíme soubor textů. Korpusová lingvistika se zabývá zkoumáním a shromažďováním textů přirozeného jazyka (zkoumáním a vytvářením korpusu).

Korpusová lingvistika je podobně novou disciplínou, jejíž vznik, stejně jako celé počítačové lingvistiky vůbec, umožnil rozvoj výpočetní techniky posledních let. Tato činnost 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.

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 slovo nese.

XXxx

Xxxx rozebrat závislostní X složkovou lingvistickou anotaci, Rozepsat se o ÚFALu a Praze, jejich vlastní cestě. Odkaz na PDT?

Xxxx složkovou lingvistickou anotace: http://citeseer.ist.psu.edu/149407.html

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

Xxxx

3.1 Lingvistické značky

Lingvistické značky rozdělím do tří kategorií. Podle roviny lingvistické anotace, tak jak se jsou rozděleny v projektu PDT (viz oddíl 3.3). 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 [2], [3], [4].

3.1.1 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.

3.1.2 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 věty je umístěno řídící sloveso věty, na toto sloveso se pak zavěšují ostatní slova. Základním cílem anotace je správná struktura 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:

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

3.1.3 Tektogramatická rovina

Přidat referenci !!, je to velmi rozsáhlý koncept s hlubokou lingvistickou teorií v pozadí.

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 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.

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.

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 [4].

- 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.

XXX

XXX

XXX doplnit obrázky stromů

XXX

XXX

XXX

3.2 Jazyky pro zápis lingvistických anotací

3.2.1 PML

http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml/index.html

3.3 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.

dopsat

PDT 1.0

PDT 2.0

3.4 WordNet

3.5 Lingvistické nástroje

3.5.1 NetGraph

3.5.2 Tred

3.5.3 Tools for machine annotation - PDT 2.0

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

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

1. Rozpoznání slovních jednotek ve vstupním surovém textu a rozdělení textu na věty.

2. Morfologická analýza a tagging (morfologická disambiguace).

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[3] (včetně naměřené chybovosti) jsou k dispozici jako součást PDT 2.0 CD-ROM.

Následuje podrobnější popis jednotlivých nástrojů.

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 disambiguace. Tagging pro Češtinu je poměrně zajímavý vědecký problém, který je podrobně rozpracován v mnoha publikacích[4].

[3] http://ufal.mff.cuni.cz/pdt2.0/doc/tools/machine-annotation/
[4] http://ufal.mff.cuni.cz/czech-tagging/

Parsing

Morfologicky označkovaná slova v každé větě uspořádá do závislostního stromu. Problém automatického závislostního parsingu[5] je stále poměrně živý. Aktuálně nejlepší parser [5] 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 3.3.

[5] http://ufal.mff.cuni.cz/czech-parsing/

Kapitola 4

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

4.1 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.

4.1.1 Hasiči

4.1.2 Úpadci

4.2 Výstupní data

Vzhledem k tomu, že pojem sémantické anotace, jak ho zmiňuji v kapitole 2, je velmi široký, není ani přesně určeno, jaká data by při procesu sémantické anotace měla vzniknout.

4.3 Software

4.3.1 Modul XY

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

Literatura

[1] 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.

[2] Dan 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.

[3] 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.

[4] 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.

[5] Daniel Zeman, Zdeněk Žabokrtský. Improving Parsing Accuracy by Combining Diverse Dependency Parsers. In: Proceedings of the Inter-

national Workshop on Parsing Technologies (IWPT 2005). Association for Computational Linguistics, Vancouver, British Columbia.