Posudek tezí doktorské disertační práce

# Vincent Kríž:
# Detecting Semantic Relations in Texts and their Integration with External Data Resources

Oponentka: Markéta Lopatková

Teze disertační práce předkládají zajímavý projekt detekování sémantických relací v právních textech; tento projekt je integrální součástí projektu INTLIB (Intelligent library, Nečaský, Hladká et al., 2012-2015).
V rámci šířeji definovaného úkolu získávání sémantické informace z nestrukturovaného textu se autor zaměřuje na problematiku extrakce dat a jejich interpretace (nikoli na jejich prohledávání a prezentaci uživateli). V práci definuje tři základní úkoly: identifikaci referencí v jednotlivých právních dokumentech, extrakci sémantických vztahů a parsing právních dokumentů.
Teze po poměrně široké motivaci a úvodu, který začleňuje projekt do systému zpracování nestrukturovaného textu využívajícího principů Linked Data (Nečaský et al., 2013), představují výsledky dosažené při řešení tří výše zmíněných úkolů, jejich vyhodnocení a stručné srovnání s výsledky obdobných projektů v zahraničí, včetně integrace do celého systému zpracování textů. V závěru tezí jsou stručně představeny další plány doktoranda směřující k dokončení práce na projektu a podání dizertace.

**Hodnocení a připomínky**
Navrhovaný projekt řeší velmi aktuální a zajímavý problém. Jeho výstupem je (i) funkční systém pro vyhledávání referencí v nestrukturovaném textu (JTagger) a (ii) systém RExtractor pro identifikaci entit a sémantických vztahů mezi nimi, který využívá existující NLP nástrojem (včetně vyhledávání ve stromech). Dále se autor soustřeďuje na (iii) adaptaci existujícího parsingu pro doménu legálních textů.

Zatímco systém JTagger je plně funkční a dosahuje velice dobrých výsledků (autor nepředpokládá jeho další zlepšování), systém RExtraktor bude nadále vyvíjen a zlepšován. Bohužel nefungují odkazy na ontologii reprezentující entity a relace mezi nimi, ani na ontologii s textovými shluky, z nichž RExtraktor získává jednotlivé koncepty a relace (http://purl.org/lex/ontology/concepts# , http://purl.org/lingv/ontology# ), což ztěžuje možnost vyhodnocení výsledků tohoto nástroje.

(1) Co se týká možnosti dotazů ve stromech (pomocí PML-TQ), ráda bych se zeptala, zda autor uvažuje/uvažoval o využití tektogramatické anotace (odstranila by poněkud krkolomné dotazy typu 7)?
(2) Uvažuje o využití Word Sketches (či nějakého nástroje pro klastrování) k rozšíření vzorku dotazů?

Posledním stanoveným úkolem je adaptace existujícího parseru pro zpracování právních textů. Autor se soustřeďuje na segmentaci a re-tokenizaci (vysoce relevantní pro složité a dlouhé souvětí v právnických textech); popisuje potřebné anotace a vyhodnocuje jejich vliv na úspěšnost parsingu. Tato sekce představuje předběžné výsledky a je základem pro další práci.

(3) Autor neuvádí, proč pro zpracování používá tzv. McDonaldův parser (McDonald et al., 2005); nejde o "deep syntactic parser" jak autor uvádí na str. 10.
K dispozici jsou i další nástroje, např. úspěšnější Malt Parser (Nivre, 2009).

(4) Anotátoři mají určovat reference, pomocí nichž lze ze segmentů zkompletovat závislostní strom – tvoří segmenty nutně souvislé podstromy? a mají segmenty nutně vždy jediného "rodiče" (u obecných textů to zřejmě neplatí)?
(5) Jak se může stát, že po re-segmentaci získáme více vět/segmentů s více než 90 tokeny (pokud správně interpretuji tabulku 8)?

V tezích není jasně vyznačeno, co dělal autor sám a co je společná práce většího týmu, což pokládám za zásadní nedostatek. Z textu vyplývá, že u tří definovaných úkolů jde zřejmě o samostatnou práci autora – je tedy škoda, že to není uvedeno explicitně.
V textu dizertace je nutné pečlivě vyčlenit a popsat samostatnou práci autora!

Text tezí je psán čtivou angličtinou, pečlivě a s naprostým minimem chyb či překlepů.
(Např. na s. 4 "a building of the network" → "building the network" (jde o děj, nikoli o výsledek či stav); s. 9 chybně čárka před "that" uvozujícím obsahovou větu, s. 9 zájmeno "he" odkazující ke "query" a některé další.)

**Shrnutí a závěr**
Navrhovaný projekt disertační práce považuji za velmi dobře připravený, s jasně stanovenými cíli a metodami zpracování. Teze i tři konferenční příspěvky dotýkající se daného projektu, které autor publikoval v r. 2014, stejně jako skutečnost, že systém RExtractor byl přijat jako demo na NAACL HLT 2015, svědčí o pokročilé práci na tomto tématu.
V tezích stanovené úkoly a cíle, přístup autora i plán práce a již dosažené výsledky jednoznačně považuji za dobré podklady pro brzké a úspěšné podání disertační práce.

Praha, 10.6.2015

doc. RNDr. Markéta Lopatková, Ph.D.
Ústav formální a aplikované lingvistiky
MFF UK

**Drobné připomínky**
- ad Figure 7
  V příslušné pasáži textu špatně srozumitelný popis "The query is designed to extract (*subject, predicate, object*) relations where the *subject* is the object in a sentence." → ?snad "The query is designed to extract (*subject, predicate, object*) relations where the *subject* of the relation is in the adverbial sentence position (Adv)."
- obsáhlejší popis k tabulkám:
  např. u tab. 2 (či v textu) postrádám vysvětlení zkratek pro systémy, není u ní uvedena použitá míra (F-measure, jak nastaven poměr precission – recall?)
- Figure 3, 6, 7 uvádějí anglický příklad, ale projekt je zaměřen na česká data?