Univerzita
Pardubice
Fakulta elektrotechniky
a informatiky

Posudek oponenta diplomové práce

Jméno studenta: Bc. Ivo SNOZA

Téma práce: Lemmatizér českého jazyka

Cíle práce:
- Vytvořit systém pro automatickou lemmatizaci a anotaci slov českého jazyka,
- navrhnout datovou strukturu a databázi pro uložení informací,
- pro jednotlivá slova je potřebné určit následující vlastnosti: pád, slovní třídu a kongruenci,
- vstupy a výstupy budou v přesně definovaném formátu,
- systém je potřebné sestavit z modulů umožňujících jeho budoucí rozšíření při určování dalších vlastností,
- vytvořit editor umožňující definování vzorů ke slovům resp. kořenům, určení slovní třídy a další potřebné vlastnosti.

1. Uplatnění metod (příslušejících navazujícímu magisterskému studiu)

Student vypracoval dostatečný teoretický přehled lingvistických pojmů potřebných pro porozumění dalšímu textu.

2. Produkt vytvořený při vypracování DP

Byla navržena a implementována struktura databáze pro ukládání tvarů slov, program pro lemmatizaci s editorem tvarů a možností importu a exportu.

3. Prokázání správnosti navrženého řešení problému

V kapitole 5.1 na str. 39 je volena databáze Oracle XE s vyjmenovanými omezeními, ale student nezdůvodnil volbu právě tohoto databázového systému. Dá se očekávat nějaký rozbor požadavků na databázový systém a následně jeho volba. Zároveň student uvádí omezení databázového produktu, která již nejsou platná[1], aniž by upřesnil verzi produktu. Tato omezení považuje za závažná a doporučuje plnou verzi (opět bez uvedení podkladů pro své doporučení – tedy nároků aplikace na databázi).

V textu práce se nevyskytují důkazy o správnosti zvolených metod, nicméně je v teoretické části naznačeno, jak lze lemmatizér implementovat. Vzhledem ke komplexnosti takového systému se domnívám, že důkazy o správnosti zvolených algoritmů jsou nad rámec diplomové práce.

4. Naplnění uložených cílů

Cíle byly splněny.

1 Student uvádí omezení pouze 32 bitů, max. jeden procesor, max. 4 GB na disku, ale na webových stránkách společnosti Oracle se uvádí pro databázi 11g Express dostupnost i pro Linux na architekturu x86_64, max. 11 GB dat, lze instalovat na víceprocesorových systémech, ale použije se pouze jeden procesor.
Oracle Database Express Edition 11g Release 2. Oracle, 2011. [cit. 2011-09-05].
URL: <http://www.oracle.com/technetwork/database/express-edition/overview/>.

5. Kvalita textu z hlediska jeho struktury, srozumitelnosti, jazykové a typografické úrovně

Práce je logicky členěna do kapitol, které na sebe logicky navazují. V textu se nicméně občas vyskytují neobjasněné pojmy, aniž by u nich byl odkaz na definici (odkaz na jinou kapitolu nebo do glosáře) – např. samotný pojem „lemma“ je vysvětlen až na str. 25 v kap. 3.3.

Stylistika příliš neodpovídá odbornému stylu, student často používá první osoby. Nevyhnul se ani gramatickým chybám, zejména často chybějí čárky (obvykle za větou vloženou nebo před větou vedlejší); např. na str. 49 je význam sdělení nejasný: „Detailnější popis lemmatizovaného textu lze do souboru uložit ve formátu XML, který obsahuje u každého slova původní token a dále pokud byla určena lemma včetně detailů mluvnických kategorií a také původní slovo s těmito detaily.“

Občas se vyskytují chybné koncovky (shoda podmětu, přívlastku, pádové koncovky), např. na str. 45: „Všechny tyto lemmata se uloží do seznamu možných lemmat ve třídě Lemma, které jsou pak nabízeny uživateli k výběru. Pokud algoritmus nalezne pouze jednu lemma…“

Student nerozlišuje spojovníky a pomlčky, volně je v textu zaměňuje a také o znaku spojovník píše na str. 43 jako o pomlčce.

Často se vyskytuje též neobratné či redundantní vyjadřování jako např. na str. 19: „…a může nastat několik situací, kdy je potřeba token určit správně. Existuje několik situací, kdy je potřeba určit, zda je znak součástí tokenu a kdy není.“ Nebo na str. 20, kapitola 3.2: „Stemma jako bychom mohli lingvistice označit jako kořen slova.“ A tamtéž: „…neplatí komutativnost pořadí operandů (nelze měnit jejich pořadí)…“

Text obsahuje i chyby v odborných pojmech: „Slovesa vyjadřují děj… Jejich *skloňování* je…“, „druh neohebný, takže se *nepáduje* ani nečasuje“. Pro jednotku času je použit termín pro jednotku úhlové míry.

Na str. 41 je odkaz na ER-diagram v příloze A. Příloha A je uživatelská příručka, ERD se nachází pouze v příloze na CD. Popis se ale liší od diagramu (např. v textu se píše o tabulce Vyznamy, ale ERD obsahuje tabulku Vyskyty).

6. Nejasnosti v DP, které je třeba objasnit při obhajobě

V úvodním teoretickém rozboru chybí některé důležité lingvistické pojmy jako např. foném.

Není také zcela jasné, co je účelem vytvořeného lemmatizéru. Naplnit databázi tvary českých slov s mluvnickými kategoriemi? Vytvořit korpusy lemmatizovaných textů?

Doporučení k obhajobě: **ano**

Navržený klasifikační stupeň: **velmi dobře**

Posudek vypracoval(a):

Jméno, tituly: Mgr. Tomáš Hudec

Zaměstnavatel: UPa, FEI, KIT

V Pardubicích dne 5. 9. 2011 Podpis: .....