Masarykova univerzita
Fakulta informatiky

Posudek oponenta bakalářské práce

Autor práce: Dávid Hubač
Název práce: Generovanie slovníka pomocou pivotného jazyka
Jméno oponenta: Zuzana Nevěřilová

Cílem bakalářské práce Dávida Hubače je generování překladového slovníku mezi dvěma jazyky pomocí třetího (pivotního) jazyka. Výsledkem práce je generátor česko-japonského a japonsko-českého slovníku o velikosti 3 271 hesel. Slovník samotný není součástí odevzdané práce. Práce má dvě části: implementační a písemnou.

V implementační části autor vypracoval generátor, který podle metod popsaných v [2] a [4] vygeneruje překladový slovník. Implementace je funkční, i když její efektivita (jak je naznačeno na str. 14 a 25) je otázkou do diskuse.

Písemnou část práce tvoří text o šesti kapitolách a 29 stranách (vč. příloh), seznam literatury obsahuje 8 položek. Sekce 1.3 a 1.4 v podstatě kopírují zdroj [2]. Na tom by nebylo nic špatného, protože autor přiznává, že použil stejnou metodu jako [2] nebo [4], jen bohužel neuvádí příliš jasně, že obě sekce jsou v podstatě překladem [2] resp. [4], pouze místo maďarštiny je uvedena čeština. S tím také souvisí nedostatek, že volba koeficientů je převzata z [2], kde je však uvedeno, že jde o jazykově závislé koeficienty.

Text práce nemá příliš jasnou strukturu (např. v kap. *3 Návrh aplikace* je pouze návrh uživatelského rozhraní). „Výpočet“ synonym je zbytečné popisovat (str. 17), víme-li, že *synset* znamená synonymická množina (řada), tj. členy synsetu jsou jen synonyma. Vůbec není jasné, proč autor používá slovníkové definice místo ID jednotlivých synsetů (která jsou jistě jednoznačná).

V sekci 5.1.1 autor uvádí, že struktura *Dictionary* obsahuje pro každou sémantickou skupinu anglické slovo jako klíč a „množinu informací“ jako hodnoty, v 5.2 je ovšem uvedeno, že hodnotou je „list překladů k tomuto slovu“.

V textu jsou drobné nepřesnosti či neobratnosti (např. „angličtina je nejrozšířenější jazyk“, str. 1, „u slovesných hyponym je to trochu choulostivější“, str. 5). Idea, že „až bude [...] Wiktionary dokončený a pro všechny jazyky bude obsahovat všechna data“, je vzhledem k neustálému vývoji jazyka poněkud naivní. Text o WordNetu v sekci 2.2.1 je přinejmenším nejasný, což je způsobeno faktem, že jde o nekvalitní (a nevyznačený!) doslovný překlad popisu této lexikální databáze.

Na konci kapitoly 5 autor uvádí, že z původních 4 950 kandidátů na překlad získal 3 271 „správných překladů“. Jakým způsobem zjistil, že překlady jsou „správné“? Výsledek práce není vůbec vyhodnocen. V textu není uvedeno, které slovníky konkrétně autor použil z `dicts.info` a jaký byl jejich rozsah.

Text obsahuje poměrně málo pravopisných chyb (většinou chybějící čárka) a překlepů, na str. 9 je chybné vyznačení kurzívou. Vážnější výhrady mám k bibliografickým záznamům: WordNet není citován tak, jak si autoři WordNetu žádají (tj. Miller, Fellbaum nebo About

WordNet), zdroj [3] je z konference EMNLP'09, nikoli z A Digital Archive of Research Papers in Computational Linguistics. U žádného bib. záznamu není místo konání konference.

Jak plyne z výše uvedeného, práce splňuje požadavky na rozsah i obsah. Práci doporučuji k obhajobě a navrhuji známku C až D podle výsledku obhajoby.

9. června 2014
Zuzana Nevěřilová