Oponentský posudek diplomové práce Radoslava Klíče

Automatické osvojení vzorů s minimální supervizí

Diplomová práce se zabývá metodami automatického a poloautomatického odvození flektivních vzorů na základě jazykového korpusu, s případným využitím ručně zpracovaného „zárodku“ (seed).

Práce je přehledně členěna do sedmi kapitol. Po úvodu obsahujícím motivaci a terminologii následuje kapitola **Previous Works**, ve které autor prokazuje, že se seznámil s relevantními pracemi týkajícími se různých přístupů ke komputační morfologii. Vše je dobře doloženo odkazy na příslušnou literaturu.

Připomínky:

- U vysvětlení vzorce na str. 9 dole jde zřejmě o „different tails“ („N_k denotes the number of all tails of length k“).
- Algoritmus popisující funkci nástroje Morfessor (str. 11, 2. odst. shora) je popsán příliš stručně. U zkratky MDL bych očekávala odkaz na místo, kde byla prvně použita (další možností je seznam zkratek na konci práce, ale vzhledem k malému množství použitých zkratek v celé práci to nebylo nutné). Dále se mluví o zmenšování velikosti korpusu, ale z popisu není jasné, jak takové zmenšování probíhá. Co znamená věta „randomly selected words are merged and resplit“?
- V kapitole 2.3.2 na str. 15 u výčtu možných transformací by bylo vhodné u každého typu uvést příklad. Celý další popis s řeckými písmeny by pak byl srozumitelnější. Odstavec začínající „As a whole“ na str. 16 mi není srozumitelný. Co je WordFrame v tomto kontextu a proč je uveden jako dvojice stejných řetězců? Opět by pomohl příklad.
- Příklad by byl vhodný i u stručného popisu přístupu v sekci 2.4 (str. 17).

Následuje důležitá kapitola 3, ve které autor popisuje fungování systému **Paramor**, jehož vylepšení je hlavním přínosem celé práce.

Připomínky:

- V sekci 3.1.1 postrádám popis inicializace svazu, takže není potom zcela jasný přechod mezi touto sekcí a sekcí 3.1.2 .
- Co znamená, že „scheme generates a set of types“ (str. 22, sekce 3.1.3)? Na první pohled by se zdálo, že je to množina slov, jejichž všechny suffixy jsou právě ty z daného schematu. Vzhledem k postupu vytváření schemat to tak ale zřejmě nebude. Rozhodně to neplatí o clustrech (cluster generates …). Chybí definice, příklady.
- Sekce 3.1.4 by rovněž zasluhovala rozvést. Např. kolik je „small number of types“?
- Kapitolka 3.1.5 mi není srozumitelná. Odkazuje na „remaining clusters“, ale není zcela zřejmé, z čeho clustry zbyly. Jsou to ty vyhozené z předchozího kroku? Uvedený anglický příklad se však zřejmě nevztahuje k těmto zbylým clustrům ..?

Kapitoly 4 a 5 představují hlavní přínos celé práce.
V kapitole 4 **Modifications of Paramor** se vysvětlují dvě úpravy Paramoru, které zlepší jeho fungování:

1. postup, který umožní využít ručně připravené zárodky pro lepší fungování celého systému a
2. zahrnutí alomorfů v kmenech slov.

Připomínky:

- V sekci 4.1 postrádám vysvětlení popisovaného formalismu. To je uvedeno až v příloze B, ale vzhledem k jeho využití v textu by bylo vhodnější tato pravidla uvést zde, přímo v textu.
- Překlad českého slova *letní* do angličtiny není *spring* (str. 25 dole). Uvedené hodnoty morfologických kategorií nejsou vyčerpávající.
- Poznámka pod čarou na str. 26 rozhodně nepatří pod čáru.
- Kapitolka 4.2.2 o automatickém odhalování alomorfů v kmenech slov je pěkná, jen popisek u tabulky 4.2 byl zřejmě zkopírován z tabulky 4.1 a neupraven (str. 29).

Kapitola 5 **Clustering Framework** je druhým vlastním autorovým výsledkem.
Zde se popisuje postup shlukování morfologicky příbuzných slov. V kontextu celé práce jde především o shlukování slov se stejným paradigmatem (stejným způsobem ohýbání), i když použití může být zřejmě širší.

Připomínky:

- Hned v úvodu se definují tři vzdálenosti mezi shluky. Zde obzvláště vyniká autorova nechuť psát definice formálně, neboť první dvě definice („nearest“ a „furthest member“) jsou přinejlepším dvojznačné, třetí definice („average distance“) je zcela nesrozumitelná.
- Termín „Euclidean combination“ použitý na str. 31 nahoře není všeobecně známý, ani jeho konkrétní zavedení dále na str. 32 neosvětluje, proč se autor rozhodl zrovna takovou kombinaci použít. Chybí tu odkaz na zdroj.

Kapitola 6 **Experiments and Results** popisuje výsledky experimentů.

Připomínky:

- Podkap. 6.3.1 je příliš stručná. Kde se hledá lexém? Jaké průniky se porovnávají? Kromě pečlivějšího vysvětlení by opět pomohl příklad.
- Očekávala bych delší komentáře k výsledkům uvedeným v tabulkách, případně pokus o zobecnění, jak které nastavení ovlivňuje výsledky v různých jazycích.

Appendix A – Software package
Popis programů v Příloze A je opět příliš stručný.
Formát výstupních dat je uveden jen formou příkladů, což považuji za nedostatečné.
Není uvedena struktura přiloženého DVD. Očekávala bych seznam a alespoň stručný popis jednotlivých adresářů a souborů.
Příklady konfiguračních souborů (setting files), které jsou nezbytné ke spuštění hlavních programů, jsou sice k nalezení, ale nelze je přímo použít, protože na DVD nejsou nahrány

soubory uvedené jako paradigmFile. Na DVD je skript, který je umí vyrobit a správně nazvat, ale ten není vůbec popsán.

Pro skutečné použití by bylo přínosem umožnit zadávat parametry ne pouze pomocí konfiguračního souboru, ale též z příkazové řádky.

Pro běžného uživatele by mělo význam i demo, které by se dalo spustit bez jakýchkoli příprav přímo na nainstalovaném software.

Appendix B

Zde mám připomínky pouze k příloze B.1, protože další jazyky neovládám.

- Pomocí uvedených pravidel lze utvořit dva špatné tvary: *maďařích* (správně *maďarech*) a *klecech* (správně *klecích*).

Všeobecné připomínky

- Jako zdroj dat pro všechny popisované systémy se uvádí (nějaký) korpus, přestože se vždy využívají jen jednotlivá slova. Podle mého názoru by bylo vhodnější mluvit o seznamu slov místo o korpusu. Na druhou stranu DVD obsahuje (bohužel nepopsané) skripty, které příslušné seznamy vyrobí.
- Odkazy uvnitř textu (např. na tabulky) by bylo kvůli snazší orientaci vhodné doplnit odkazy na příslušnou stránku.
- Termín „inflection“ je uveden na str. 5 v kapitole Terminology a popsán jako „process“. V textu se potom vyskytuje i v poněkud odlišných významech (např. str. 16, 24, 35), což je někdy matoucí.

Celkové hodnocení

Práce je psaná velmi dobrou angličtinou s minimem překlepů (např. *infected* místo *inflected* na str. 7 uprostřed) nebo vynechání členu, případně dalšími drobnostmi (*can be attached* místo *can attach* nahoře na str. 12). Typografická úprava je rovněž na vysoké úrovni, vytkla bych snad jen občasné opomenutí členu nebo krátké předložky na konci řádku (např. několikeré „*a*“ na konci řádků na str. 9).

Jak už jsem několikrát zmínila u konkrétních pasáží, stručnost textu je místy na hranici srozumitelnosti, případně způsobuje nejednoznačné chápání. Přesto si myslím, že autor odvedl pěknou práci, která může být aplikována na další jazyky, což se může hodit zejména u menšinových jazyků s omezenými vstupními zdroji (slovníky, gramatiky).

Proto doporučuji předloženou diplomovou práci k obhajobě.

Praha, 3. května 2012

Jaroslava Hlaváčová