Tvorba jednoduchého slovníka pre blízke jazyky*

Marek Grác

Fakulta informatiky, Masarykova univerzita, Brno

Rozmach digitalizácie informácií, vo všetkých podobách, nám umožňuje pristupovať k nepredstaviteľným množstvám údajov. Dostupné technológie nám v nich umožňujú vyhľadávať veľmi rýchlo a relatívne presne, ale musíme sa pohybovať v priestore používaného jazyka. V súčasnej dobe nie je prístupný žiaden vyhľadávač, ktorý by zvládol vyhľadávanie súčasne v českých i slovenských textoch. Aby sme boli schopní využiť niektorú z existujúcich aplikácií, potrebujeme dostatočne veľký, rozumne kvalitný, elektronický (*machine-readable*) slovník. Bohužiaľ česko-slovenský slovník, ktorý by spĺňal všetky tieto podmienky, nie je k dispozícii. Diferenčné slovníky nie sú dostatočne obsiahle, elektronické (napr. PC Translator) nie sú vhodné na strojové spracovanie. Aj veľmi jednoduchý slovník, ktorý obsahuje len prekladové dvojice, by nám veľmi pomohol. Preto sme sa rozhodli vytvoriť takýto slovník pre češtinu a slovenčinu.

Pri tvorbe nášho jednoduchého slovníka sme sa rozhodli vychádzať z predpokladu, že čeština a slovenčina sú príbuzné jazyky, a preto by mala byť prevažná väčšina slov podobná. Podobnosť je pojem, ktorý je možné s istými obtiažami algoritmizovať, a tak sme sa rozhodli pre automatickú extrakciu prekladových dvojích s ich následnou manuálnou korektúrou. Rozhodli sme sa, že slovník bude obsahovať len základné tvary slov a v prvej fáze sa vyhneme viacslovným výrazom. Aby sme mohli používať náš slovník na ľubovoľné účely, rozhodli sme sa využiť len voľne dostupné zdroje. To nám následne pomohlo aj pri experimentálnej tvorbe srbsko-slovinského slovníka. Medzi dostupné zdroje pre menšie jazyky patria zoznamy slov (opravy preklepov), wikipédia (prepojenia medzi jazykovými verziami) alebo frekvencie slov z národných korpusov.

Za základný zdroj dát sme sa rozhodli využiť zoznamy slov v základnom tvare z voľne dostupného korektora preklepov *ispell*. Zdroje tohoto typu sú dostupné pre väčšinu používanejších jazykov, pretože neobsahujú žiadne dodatočné informácie (napr. slovný druh) a tak sú tvorené bežnými používateľmi. Referenčným zdrojom dát je pre nás česko-slovenský slovník PC Translator, ktorý je dostupný v elektronickej podobe (pre naše potreby bol výrazne zmenšený tak, aby obsahoval len slová v základnom tvare).

Vzhľadom na podobnosť medzi oboma jazykmi sme sa rozhodli vytvoriť pravidlá, ktoré by popisovali prepisovanie znakov v jednotlivých slovách. Takto vytvorené slová sa následne pokúsime nájsť v cieľovom jazyku. Prvá sada pravidiel *B* neobsahuje žiadne pravidlá, preto získame len také slová, ktoré sa zapisujú rovnako v češtine a slovenčine. Druhá sada *G* obsahuje 15 pravidiel a sada *K* má 50 pravidiel. Rozdiely medzi sadami pravidiel K a G nie sú len v počte pravidiel, ale aj v samotnom prístupe. Sada K generuje priemerne 4x viac kandidátov na české slovo ako sada G. Každé vygenerované slovo, ktoré nájdeme v cieľovom jazyku, pridáme do slovníka. Týmto spôsobom zanášame do systému chyby pri tých slovách, ktoré majú odlišný význam a po aplikovaní pravidiel sa píšu rovnako (napr. (sk) kel -> (cz) kapusta, (sk) kapusta -> (cz) zelí). Nasledujúca tabuľka zobrazuje dosiahnuté výsledky pre tieto sady pravidiel.

Metóda	Pokrytie	Presnosť
B – presná zhoda	18,17 %	99,36 %
G – konzervatívna	37,91 %	98,98 %
K – liberálna	52,65 %	97,07 %

Na vytváranie prekladových dvojíc je možné využiť aj vzdialenosť medzi dvoma slovami. Existuje niekoľko metrík, ktoré pracujú na základe editačnej vzdialenosti. My sme si zvolili Levenshteinovu vzdialenosť, ktorá počíta koľko zmien (pridanie, zmazanie, výmena jedného znaku) je potrebné urobiť, aby sme získali druhé slovo (napr. l(kitten, sitting) = kitten – sitten – sittin – sitting = 3). Ďalšou jej dôležitou vlastnosťou je, že ju je možné použiť aj pre slová s rozličným počtom znakov. Po výpočte na referenčom slovníku sme zistili, že viac než 75 % prekladových dvojíc sa nachádza vo vzdialenosti 0 až 3. Vzdialenosť slov v prekladovej dvojici nie je výrazne ovplyvnená frekvenciou slova a rozdiely v distribúcii slov s rozličnou vzdialenosťou medzi 1 000, 5 000, 10 000 a 20 000 najfrekventovanejšími slovami sa pohybujú tesne nad úrovňou štatistickej chyby.

Popisované metódy ponúkajú rozličné kombinácie pokrytia a presnosti. Vzhľadom na relatívne nízke pokrytie najkvalitnejších (pravidlových) metód sa ponúka ich kombinácia s metódami založenými na vzdialenosti. Prišli sme na to, že použitie sady pravidiel zlepší výsledky, ale na samotnej kvalite (od istej úrovne) už nezáleží. Rozdiel pokrytia sadou pravidiel G a K sa po použití kombinovaných metód znížil z 18 percentuálnych bodov na 2 percentuálne body.

Tento projekt ukázal, že je možné vybudovať česko-slovenský slovník aj bez požiadaviek na náročné zdroje. Keďže sa pri slovníkoch očakáva veľmi vysoká presnosť, tak pomocou prezentovanej metódy nie je možné budovať slovník plne automaticky. Dosiahnutá presnosť však dáva nádej, že tento prístup bude možné použiť ako doplnok k existujúcim diferenčným slovníkom, ktoré by mali obsahovať všetky typy problémových slov. Takýto prístup k tvorbe prekladového slovníka zrejme nebude možné využiť pre rozdielnejšie jazyky, ale otvára sa jeho použitie napr. na tvorbu dialektologických slovníkov alebo v oblasti slovotvorby.

Literatúra:

BÉMOVA, Alla – KUBOŇ, Vladimír: Czech-to-russian transducing dictionary. In: Proceedings of the 13th conference on Computational linguistics. Morristown, NJ: Association for Computational Linguistics 1990, s. 314 – 316.

KOCEK, Jan – KOPŘIVOVÁ, Marie – KUČERA, Karel (eds.): Český národní korpus*:* Úvod a příručka uživatele. Praha: ÚČNK – FF UK 2000.

KOLÁŘ, Petr: Czech dictionary for ispell. <http://www.kai.vslib.cz/kolar/rpms.html> 2006.

LEVENSHTEIN, Vladimir: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady, 1966, roč. 10, s. 707 – 710.

Slovenský národný korpus – prim-2.0-public-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2005. Dostupný z WWW: <http://korpus.juls.savba.sk>.

PODOBNÝ, Zdeněk: Slovak dictionary for ispell. http://sk-spell.sk.cx (2006).
Wikipedia: <http://www.wikipedia.org> (2007).

* Táto práca bola čiastočne podporená Akadémiou vied ČR v rámci projektu T100300419, Ministerstvom školstva ČR v rámci Centra základného výskumu LC536 a Národným výskumným programom 2C06009 a Grantovou agentúrou ČR v rámci projektu 201/05/2781.