Hľadanie základného tvaru slovenského slova na základe spoločného konca slov *

Stanislav Krajči[1], Róbert Novotný[2]
[1]stanislav.krajci@upjs.sk
[2]robert.novotny@upjs.sk

[1,2]Ústav informatiky, Prírodovedecká fakulta, UPJŠ Košice

Abstrakt Na základe jednoduchého pozorovania, že na ohýbaní slovenského slova má najväčší vplyv jeho koniec, sme implementovali algoritmus, ktorého ambíciou je nájsť základný tvar daného tvaru slova. V prvej fáze sa spomedzi dopredu definovaných známych tvarov vybratých slov vyberie také, ktoré má s daným slovom čo najdlhší spoločný koniec, a stane sa tak jeho „predlohou“, v druhej fáze sa pomocou základného tvaru tohto už definovaného slova „podvojnou zámenou“ odvodí možný základný tvar daného slova. V (nepovinnej) tretej fáze sa skontroluje prítomnosť takto vzniknutého základného tvaru daného slova v zozname slov slovenského jazyka.

1 Východiská

V rámci projektu NAZOU (Nástroje na získavanie, organizovanie a udržiavanie znalostí v prostredí heterogénnych zdrojov) je pri spracúvaní nových ponúk do vektorového (resp. objektovo-atribútového) modelu, v ktorom sa eviduje viac či menej sofistikovaná štatistika výskytu termov (slov) v dokumente, užitočné združiť rôzne tvary toho istého slova do jednej skupiny a vybrať jej vhodného reprezentanta, či už ide o spoločný slovný koreň týchto tvarov alebo ich základný tvar.

V dvoch jazykoch, ktoré nás zaujímajú, – angličtine a slovenčine – je však tento proces diametrálne odlišný. V anglickom jazyku, ktorý pozná ohýbanie slov len vo veľmi prostej podobe, sa spoločný slovný základ dosiahne jednoduchšie, a to vzhľadom na pomerne dobre definovateľné pravidlá tvorby slov zo spoločného slovného základu (isteže, existujú aj výnimky (napr. nepravidelné slovesá), tých je však relatívne málo). Obvyklou metódou je tu tzv. stemming (podľa Porterovho algoritmu ([2])), ktorý spočíva v odstraňovaní niektorej z mála prípon (napríklad „-ed“, „-ing“, či „-s“).

Slovenčina však patrí k tzv. flexívnym jazykom, kde má väčšina slov niekoľko (desiatok) tvarov, ktoré sa vytvárajú storakými spôsobmi. Časová investícia do hľadania a aplikácie pravidiel ohýbania slov tak môže byť príliš veľká, ba väčšia

* Podporené štátnym projektom výskumu a vývoja Budovanie informačnej spoločnosti, Nástroje na získavanie, organizovanie a udržiavanie znalostí v prostredí heterogénnych zdrojov, 1025/04

než (časovo náročné, ale v podstate jednorazové) nájdenie všetkých tvarov týchto slov. (Tu spomeňme veľmi zaujímavú prácu E. Páleša ([1]), ktorá túto obavu nechtiac potvrdzuje.) Táto možnosť veľmi dlho neprichádzala do úvahy, veď celá oficiálna slovná zásoba – Slovník slovenského jazyka ([4]) má v papierovej podobe šesť hrubých zväzkov, pričom obsahuje v podstate len základné tvary slov. Ak si však uvedomíme, že fakticky ide asi „len“ o 150 000 slov, ktoré majú spolu pár miliónov tvarov, úloha uložiť ich všetky do databázy už prestáva byť v dnešnej dobe nepredstaviteľná.

S týmto cieľom prebehla na Ústave informatiky UPJŠ v Košiciach začiatkom tohto tisícročia elektronizácia spomínaného Slovníka slovenského jazyka, ale aj Veľkého slovníka cudzích slov ([3]). Po dôkladnom (hoci ešte stále neúplnom) vyčistení dát sme tak získali v podstate kompletný zoznam 150 000 doteraz oficiálnych slovenských slov doplnených o 60 000 slov cudzieho pôvodu. (Sme si, samozrejme, vedomí, že slovenský jazyk sa vyvíja a slovná zásoba sa od vydania Slovníka slovenského jazyka značne zmenila. Napriek tomu si však myslíme, že získaný zoznam slov je však určite dobrým východiskom pre náš ďalší výskum.)

2 Hľadanie základného tvaru

Ako sme už naznačili, bohaté skúsenosti s hľadaním tvarov slova v angličtine sú do slovenčiny neprenosné, pre slovenský jazyk jednoducho pendant spomínaného Porterovho algoritmu neexistuje. Na vytvorenie základného tvaru daného slova použijeme (zámerne) *veľmi jednoduchú* metódu. Je založená na triviálnom pozorovaní, že *ohýbanie slova závisí od jeho konca, nie od začiatku.*

Vstupom nášho algoritmu je slovo v ľubovoľnom tvare (z technických dôvodov sa obmedzme len na podstatné mená, uvádzanú myšlienku však možno rovnako aplikovať i na ostatné (ohybné) slovné druhy), výstupom zoznam jeho možných základných tvarov. Predpokladáme pritom, že máme k dispozícii jednak spomínaný zoznam (základných tvarov) slovenských slov a jednak zoznam všetkých tvarov už vyskloňovaných podstatných mien. (V ideálnom prípade, o ktorom hovoríme v predchádzajúcej stati, by tento zoznam obsahoval všetky tvary všetkých postatných mien.) Algoritmus má tri fázy:

1. hľadanie zodpovedajúcich *predlôh*, t. j. zodpovedajúcich tvarov slov zo zoznamu vyskloňovaných podstatných mien
2. vytvorenie možných základných tvarov pomocou „podvojnej výmeny“
3. overenie prítomnosti možných základných tvarov v zozname všetkých základných tvarov (včítane kontroly rovnosti rodu s rodom príslušnej predlohy)

Ilustrujme tento algoritmus na konkrétnom príklade. Prepokladajme, že máme nájsť základný tvar slova „ponúk“, pričom sa toto slovo v zozname predlôh nenachádza, zato sa tam nachádza slovo „ruka“, a to včítane všetkých jeho tvarov. Zdôraznime, že predlohou nemusí byť žiaden dobre známy „školský“ vzor skloňovania („chlap“, „hrdina“, „dub“, „stroj“, atď.). Všimnime si tiež, že tu vôbec nerešpektujeme slovné základy, pri skloňovaní totiž nie sú dôležité.

– slovo: $X =$ „ponúk“
– jedna z nájdených predlôh: $Y =$ „rúk“
– spoločný (neprázdny) koniec: $K =$ „úk“
– začiatok predlohy: $Y' =$ „r“ (teda $Y = Y' + K$)
– začiatok slova: $X' =$ „pon“ (teda $X = X' + K$)
– základný tvar predlohy: $Y =$ „ruka“
– koniec základného tvaru ponuky: $K' =$ „uka“ (teda $Y = Y' + K'$)
– základný tvar slova: $X = X' + K' =$ „ponuka“
– overenie existencie slova X v zozname základných tvarov

Rovnako dobrou predlohou (za predpokladu, že by sa nachádzalo v ich zozname) by mohlo byť trebárs slovo „oblúk“, algoritmom v predposlednom kroku odvodený príslušný základný tvar „obluka“ by však neprešiel kontrolou v poslednom kroku, takýto základný tvar totiž neexistuje.

Ak je predlôh viac, uprednostníme tú, ktorá má s daným slovom najdlhší spoločný koniec. Tento postup, samozrejme, nevylučuje, že základných tvarov daného slova môže byť viacero. Ak sa však žiadny nenájde, dané slovo je vhodne vyskloňovať a doplniť ním zoznam predlôh.

3 Záver

V tomto článku sme sa pokúsili ilustrovať jednoduchú metódu na nájdenie základného tvaru slovenského slova. Pre potreby projektu NAZOU sa pokúsime rozšíriť jej funkčnosť o tieto črty:

– možnosť ignorovať diakritiku
– rozšírenie funkčnosti na slová mimo slovníka (napr. vlastné mená) (t. j. vynechanie tretej fázy algoritmu)
– rozšírenie na ostatné ohybné slovné druhy (hlavne prídavné mená a slovesá)

Ľahko si môžeme všimnúť, že uvedenú metódu možno prirodzeným spôsobom rozšíriť na hľadanie všetkých tvarov slova, teda nielen základného. (Aj pomocou takto upravenej metódy) získaný spomínaný zoznam všetkých tvarov slovenských slov by bol nielen užitočnou pomôckou pri kategorizovaní ponúk (a, samozrejme, aj iných textov), ale i dobrým východiskom pre automatickú syntaktickú analýzu slovenských viet.

Reference

1. Páleš, E.: Sapfo, parafrázovač slovenčiny, Veda, Bratislava, 1994. ISBN 80-224-0109-9.
2. Porter, M. F.: An algorithm for suffix-stripping, Program, 14 (3), pp. 130–137, 1980.
3. Šaling, S., Ivanová-Šalingová, M., Maníková, Z.: Veľký slovník cudzích slov, SAMO, 2003
4. kol.: Slovník slovenského jazyka, Vydavateľstvo SAV, Bratislava, 1959–1968