FAKULTA ELEKTROTECHNIKY A INFORMATIKY
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Ing. Ján Tóth

Autoreferát dizertačnej práce

Kontextová analýza textu

na získanie akademického titulu doktor (philosophiae doctor, PhD.)

v doktorandskom študijnom programe: 5.2.15 Telekomunikácie

Bratislava, jún 2013

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Ústave telekomunikácií FEI STU v Bratislave.

Predkladateľ: **Ing. Ján Tóth**
Ústav telekomunikácií FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Školiteľ: **doc. Ing. Gregor Rozinaj, PhD.**
Ústav telekomunikácií FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Oponenti: **doc. Ing. Jozef Juhár, PhD.**
Katedra elektroniky a multimediálnych telekomunikácií
Park Komenského 13, 041 20 Košice

Ing. Jozef Čepko, PhD.
CCW spol.s r.o.
Trenčianska 47, 821 09 Bratislava

Autoreferát bol rozoslaný: .....

Obhajoba dizertačnej práce sa koná .....22.8.2013.....**o** .....10:30.....**h.**

v zasadacej miestnosti dekana FEI STU v Bratislave, Ilkovičova 3, 812 19 Bratislava

.....
prof. RNDr. Gabriel Juhás, PhD.
dekan FEI STU Bratislava
Ilkovičova 3, 812 19 Bratislava

Obsah

1 Úvod

V súčasnosti máme možnosť vidieť ako počítače preberajú mnohé úlohy a práce, ktoré donedávna vykonával človek sám. Priemyselné odvetvia prechádzajú procesom automatizácie a tento proces sa nevyhol ani odvetviu telekomunikácií. Bežne v telekomunikáciách sa stretávame so syntézou ľudskej reči a keďže nastáva neustály pokrok aj v tejto oblasti, dokážeme s počítačmi komunikovať. Je nevyhnutné, aby informácie takýmto spôsobom sprostredkované a získané, boli zrozumiteľné, mali čo najvyššiu výpovednú hodnotu a boli pre človeka užitočné. Hlavná časť týchto informácií má formu prirodzeného jazyka.

Pod pojmom prirodzený jazyk rozumieme prostriedok, ktorý vymysleli ľudia, naši predkovia, za účelom dorozumievania sa medzi sebou. Prirodzený jazyk slúžil a aj slúži ľuďom na budovanie a rozvoj ľudskej spoločnosti. V evolúcii ľudstva môžeme zaznamenať rôzny spôsob komunikácie medzi ľuďmi a rôznu stavbu konkrétneho jazyka. Každý národ má svoj jazyk, ktorý najlepšie vystihuje ich spôsob komunikácie, vyjadrenie pocitov a myšlienok. Niet preto pochýb, že spracovanie a porozumenie iného jazyka, ako toho svojho, je pre človeka náročné.

Spracovanie prirodzeného jazyka (NLP) sa stalo jedným z hlavných problémov pri výmene informácií s počítačom. Rýchly rozvoj výpočtovej techniky výrazne urýchlil riešenie problémov spojené s týmto spracovaním. Počítačové systémy prirodzeného jazyka musia prevádzať vhodným spôsobom informácie z počítačových databáz do prirodzeného jazyka používaného ľuďmi, čo je veľmi náročný proces. Morfologický model reči musí byť schopný generovať gramaticky správne vety. Značné problémy spôsobuje flektívnosť jazyka (mnohotvarosť), ktorá sa do umelej inteligencie ťažko implementuje. Inteligentné spracovanie jazyka je založené na vede nazývanej výpočtová lingvistika.

Syntéza reči tiež predstavuje dôležitú oblasť výskumu problematiky spracovania rečového signálu. Syntéza ľudskej reči je proces, pri ktorom sa umelo vytvára reč, najčastejšie počítačom. Je viacero dôvodov prečo je syntéza ľudskej reči v dnešnej dobe potrebná:

- reč je najprirodzenejšou formou ľudskej komunikácie a v budúcnosti aj medzi človekom a počítačom
- vo viacerých odvetviach sa na zefektívnenie práce využívajú rečové syntetizátory
- nevyhnutná pomôcka pre ľudí so zrakovým postihnutím
- rôzne telekomunikačné služby, pre pristupovanie k údajom pomocou telefónneho prístroja, rôzne dialógové systémy

2 Prehľad súčasného stavu problematiky

2.1 Kontextová analýza

Kontextová analýza textu je jedným z dôležitých nástrojov na predspracovanie textu v procese syntézy reči. Kontextový rozbor zisťuje, aký vzťah majú frázy vo vete na veci, objekty zo skutočného sveta. Znamená to párovať a priraďovať frázy z jazyka k týmto objektom, resp. zlúčiť a zjednotiť tie frázy a časti jazyka, ktoré sa vzťahujú a určujú ten istý vonkajší objekt [13].V zásade sa v programoch používajú hlavne dva prístupy riešenia a to modely založené na pravdepodobnosti a deterministické modely.

2.1.1 Prístupy k analýze textu

Štatistický prístup k analýze textu

Pri štatistickom prístupe sa jazyková analýza kontextu textu chápe ako identifikácia, t.j. určenie identifikátora (vektora pojmov), obsahu dokumentu. Pri tomto prístupe možno formulovať nasledujúce hlavné body analýzy:

- **určenie neplnovýznamových slov**, tzv. stop-slová (angl. stop-words) - plnia len syntaktické funkcie. Tieto slová možno vylúčiť z ďalšej analýzy [3].
- **určenie synonymných slov** – rovnoznačných ale nerovnozvučných výrazov v texte. Slová s odlišnou formou a rovnakým alebo podobným významom (napr. *kniha, publikácia*) by sa mali reprezentovať tým istým termom. Na identifikáciu podobnosti významov nepostačuje skúmanie na úrovni morfológie ani syntaxe, potrebná je slovotvorná, sémantická a pragmatická analýza [7].
- **určenie homonymných slov** – rovnozvučných ale nerovnoznačných výrazov v texte. Slová s náhodne totožnou formou a s rôznym významom (napr. *jazyk* [reč] vs. *jazyk* [v ústach] vs. *jazyk* [na topánke], *akcia* [činnosť] vs. *akcia* [cenný papier], *kurz* [jazykový] vs. *kurz* [meny]) by sa mali podľa kontextu rozlíšiť a reprezentovať vzájomne rôznymi termami. Na rozlíšenie homonymných slov je treba uskutočniť syntaktickú a sémantickú analýzu textu [7].
- **hľadanie anaforických referencií**, odvolávok v texte pomocou zámen alebo presupozície. Určenie takýchto kontextových čŕt je pri spracovaní textu pomerne ťažké zachytiť [14]. V texte sa pomerne často vyskytujú konštrukcie obsahujúce odkazy na objekty z predchádzajúceho alebo nasledujúceho kontextu. Tieto odkazy sa väčšinou, hoci nie výlučne, realizujú pomocou zámen. Napríklad vo vete „*Znalosti programátorov by vo firme ostali aj po ich odchode*.“ zámeno *ich* odkazuje na *programátorov*. To však znamená, že vektor termov pre túto vetu by namiesto lemy zámena *ich*, teda tvaru *on*, mal obsahovať dvakrát substantívum *programátor*. Priradiť zámenu príslušnú frázu alebo objekt, na ktorý sa toto zámeno odkazuje, však vyžaduje komplexnú syntaktickú a sémantickú analýzu textu.
- **určenie kľúčového slova**, ktoré nie je obsiahnuté v spracovávanom texte, avšak je možné ho identifikovať pomocou istých deduktívnych a inferenčných nástrojov, patrí tiež medzi zložitú úlohu kontextovej analýzy.
- **Frázy, viacslovné ustálené pomenovania**. Slovné spojenia so špecifickým významom, napríklad *Technická univerzita v Košiciach*, *manažment znalostí*, *expertný systém*, a podobne. Majú charakter samostatných termínov. Ich celkový význam je iný ako významy slov, z ktorých sa tieto termíny skladajú. Bolo by teda vhodné, aby sa ustálené viacslovné pomenovania reprezentovali jedným viacslovným termom. Identifikácia viacslovných pomenovaní v texte, vrátane príslušných gramatických modifikácií, predpokladá integrovanú morfologickú a syntaktickú analýzu.

Štatistický prístup k jazykovej analýze si nezakladá na lingvisticky správnom opisovaní jazykových vzťahov medzi jednotlivými členmi v texte. Na druhej strane ponúka pomerne rýchlu operáciu a spracovanie veľkého objemu dokumentov, údajov a textov.

Kontextové spracovanie textu

Kontext možno definovať ako významovú súvislosť medzi formálne autonómnymi časťami textu, napríklad medzi práve skúmanou výpoveďou a výpoveďami, ktoré ju obklopujú. Možno to formulovať aj tak, že význam časti textu (povedzme jednej výpovede, môže to však byť aj slovo, slovné spojenie, odsek atď.) nie je apriórny, okrem lexikálneho významu závisí aj od jazykového okolia tejto časti. Kontextové spracovanie sa vyvíjalo paralelne s klasickým prístupom. Dlhú dobu sa nevyužívalo v praxi v takej miere ako klasický prístup [15].

Linguistický prístup k analýze textu

Tento prístup sa kvalitatívne líši od oboch predchádzajúcich. V [17] je popísaný ako prevod textového úseku (napr. výpovede, vety, odseku dokumentu a pod.) z povrchovej do hĺbkovej štruktúry pričom je využité maximálne množstvo poznatkov lingvistického opisu a so snahou o čo najúplnejšie modelovanie systému prirodzeného jazyka. Časové nároky nie sú rozhodujúce a prednostné. Dôraz je kladený na presnosť a úplnosť analýzy.

2.1.2 Nástroje a metodiky používané pri analýze textu

Metódy určovania kľúčových slov

Prvé postupy na spracovanie prirodzeného jazyka a extrakcie kontextu z textu mali prezentačnú štruktúru vhodnú pre príslušné klasifikačné algoritmy. Tieto sú najčastejšie založené na viacrozmernej analýze, ktorá na vstupe predpokladá údajovú *m*– rozmernú maticu pozorovaní na *n* objektoch [26]. Objektami sú v tomto prípade jednotlivé textové dokumenty v skúmanom súbore (*korpuse*). Pozorovania sú váhy jednotlivých kľúčových slov (resp. *termov*; angl. *keywords*, resp. *terms*) v texte každého z dokumentov. Vyjadrujú hodnoty týchto charakteristických znakov, resp. premenných, pozorovaných na textových dokumentoch [7].

Eliminácia neplnýznamových slov

Po tokenizácii textu sú všetky identifikované tokeny kandidátmi na termy vektorového modelu, ktorý reprezentuje text. Termy (kľúčové slová), ktoré tvoria tento vektor, by mali čo najpresnejšie vyjadrovať obsah textu. Po procese tokenizácie je týchto termov veľa a pre ďalšie efektívne spracovanie je nutné napr. na základe klasifikačných algoritmov počet termov znížiť. Preto je nutné eliminovať termy s nízkym alebo žiadnym príspevkom k obsahu textu. Vo všeobecnosti sa predpokladá, že hlavnými nositeľmi obsahu textu sú plnovýznamové slová, predovšetkým substantíva a adjektíva. Lematizované tokeny z týchto slov sú vhodný kandidáti na reprezentáciu obsahu textu v podobe termov. Naopak, pri neplnovýznamových slovách sa nepredpokladá žiadny prínos k obsahu textu – tvoria ich spojky, predložky, zámená, častice. Sú to tzv. stop-slová, ktoré sa v texte vyskytujú s väčšou frekvenciou, avšak nedávajú žiadnu informáciu o výslednom obsahu textu. Tokeny, ktoré vznikli z týchto neplnovýznamových slov je nutné vylúčiť zo zoznamu termov [29], [30].

Zväčša sa pri odstraňovaní stop-slov používajú nasledujúce dve techniky:

- použitie *negatívneho slovníka* – ručne zostavený zoznam neplnovýznamových slov v podobe slovníka. Z tokenizovaného textu sa odstránia tokeny nachádzajúce sa v tomto slovníku a vo vektore termov textu už nebudú vystupovať. Efektivita a účinnosť tohto

postupu závisí od úplnosti a obsahu negatívneho slovníka. Nevýhodou použitia tohto slovníku je, že je závislý na jazyku a je možné ho aplikovať len na jeden jazyk

- *automatický spôsob* – na základe frekvencií výskytu tokenov v texte. Eliminujú sa iba tie tokeny, ktoré sa v spracovávanom texte nachádzajú s príliš veľkou a aj príliš malou frekvenciou. Podľa [31] majú najväčšiu váhu a prínos k obsahu slová, ktorých výskyt sa v celom korpuse textov nachádza medzi intervalom $\langle \frac{N}{100}, \frac{N}{10} \rangle$, kde N je počet textov v celom korpuse. Tento spôsob je použiteľný pre väčšinu jazykov, v praxi však vykazuje slabšie výsledky ako prvý prístup. Optimálne výsledky sa dosahujú kombináciou obidvoch prístupov [7].

Váhovanie a normovanie termov

Po identifikácii a lematizovaní termov v texte je potrebné ohodnotiť dôležitosť týchto termov, v rámci tohto textu, ale aj celého textového korpusu. Váhovacie a normovacie techniky umožňujú ohodnotenie termov a tým pádom zvyšujú efektívnosť na ďalšiu prácu s týmto textom, t.j. klasifikáciu a extrakciu ďalších informácií.

Pod váhovaním rozumieme úpravu frekvencie výskytu všetkých termov v celom spracovávanom textovom korpuse. Toto váhovanie podľa [32] prebieha v dvoch rovinách.

- *lokálne váhovanie* – váhovanie na základe frekvencie výskytov v danom texte $L(k_t, d_i)$, kde k_t je t-ty term v i-tom texte d_i.
- *globálne váhovanie* – váhovanie na základe frekvencie výskytov t-teho termu v celom korpuse $G(k_t)$

Potom váhovaná frekvencia výskytu t-teho termu k_t, v texte d_i je daná súčinom lokálnej a globálnej váhy:

$$w_{it} = L(k_t, d_i) \times G(k_t)$$

2.2 Kategorizácia textu

Úlohou kategorizácie textov je nájsť aproximáciu neznámej funkcie $\phi: D \times C \rightarrow \{true, false\}$, kde D je množina textov a $C = \{c_1, c_2, \ldots c_{|C|}\}$ je množina preddefinovaných kategórií. Funkcia ϕ nadobúda pre $\langle d_i, c_j \rangle$ hodnotu true ak text d_i patrí do kategórie c_j. Hľadaná funkcia $\phi: D \times C \rightarrow \{true, false\}$, ktorá aproximuje ϕ sa označuje ako klasifikátor. Medzi aplikácie kategorizácie textov patrí kategorizovanie nahovorených textov v kombinácii s rozpoznávaním reči, klasifikovanie multimediálnych dokumentov podľa textových titulkov, alebo identifikovanie autora a žánru literárnych textov. Kategorizáciu textov je ďalej možné využiť aj pri lingvistickom spracovaní prirodzeného jazyka (NLP) napr. pre určenie významu slov v texte, morfologické značkovanie, alebo identifikovanie fráz.

2.2.1 Metódy klasifikácie textov

Naivný Bayesov klasifikátor

Naivný Bayesov klasifikátor patrí medzi pravdepodobnostné modely, ktoré klasifikujú nové texty na základe podmienenej pravdepodobnosti $P(c_j|d)$, tj. pravdepodobnosti, že text d má byť zaradený do kategórie c_j. Pri učení klasifikátora je na vstupe trénovacia množina

dokumentov $D = \{d_1, \dots d_{|D|}\}$, ktoré sú klasifikované do množiny tried $C = \{c_1, \dots c_{|C|}\}$. Každý z textov d_i je reprezentovaný ako postupnosť termov, ktoré sa vyskytli v dokumente $d_i = \langle t_{i,1}, \dots, t_{i,|d|} \rangle$, kde $|d|$ je dĺžka textu d_i [36].

Na základe trénovacej množiny *D* nie je možné určiť odhad podmienenej pravdepodobnosti $P(c_j|d)$ pre neznámy dokument *d* priamo. Je však možné priamo určiť pravdepodobnosť kategórie P(c_j) a pravdepodobnosť výskytu termu *t* v dokumente za predpokladu, že dokument patrí do kategórie c_j, t.j. $P(t|c_j)$. Naivný Bayesov klasifikátor využíva odhad týchto pravdepodobností určených podľa trénovacej množiny dokumentov pre určenie pravdepodobnosti $P(c_j|d)$. V prvom kroku je potrebné skombinovať pravdepodobnosti $P(t|c_j)$ jednotlivých termov vyskytujúcich sa v dokumente tak, aby sa získal odhad pravdepodobnosti $P(d|c_j)$ pre celý dokument. Za predpokladu, že každý term sa vyskytuje v dokumente štatisticky nezávisle od ostatných termov je možné určiť $P(d|c_j)$ ako súčin:

$$P(d|c_j) = \prod_{i=1}^{|d|} P(t|c_j)$$

Pravdepodobnosť $P(c_j|d)$ je potom určená z $P(d|c_j)$ podľa Bayesovho pravidla:

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}$$

Ak má byť dokument klasifikovaný iba do jednej triedy, na základe Bayesovho pravidla sa vyberie trieda, pre ktorú je pravdepodobnosť $P(c_j|d)$ maximálna. Pravdepodobnosť $P(d)$ je možné v tomto prípade vynechať, keďže neovplyvňuje rozhodovanie podľa maximálnej hodnoty $P(c_j|d)$. Pre viacnásobnú klasifikáciu je dokument zaradený do triedy c_j, ak pravdepodobnosť $P(c_j|d)$ prekročí zvolenú prahovú hodnotu, napr. 0,5.

K-najbližších susedov

Učenie klasifikátorov založených na pravidle *k*–najbližších susedov je jednoduché – pri učení sa odpamätajú všetky trénovacie príklady z *D*. Pri klasifikácii nového dokumentu sa podľa metriky alebo funkcie podobnosti určí *k* najbližších (najpodobnejších) dokumentov z *D* a klasifikátor zaradí nový dokument do triedy určenej podľa klasifikácie susedov [37], [38].

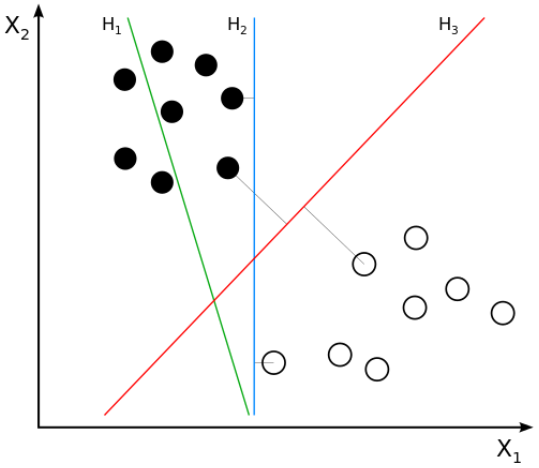
Najdôležitejšou časťou algoritmu je metrika na meranie vzdialenosti vo vektorovom priestore dokumentov, resp. funkcia podobnosti. Keďže dokumenty sú reprezentované ako reálne vektory, je možné použiť Euklidovskú metriku, ale významovej podobnosti dokumentov lepšie zodpovedá kosínusová funkcia podobnosti. Je možné použiť aj ďalšie podobné funkcie navrhnuté pre vyhľadávanie informácií, kde funkcia podobnosti určuje skóre dokumentu voči dopytu.

V prípade, že má byť príklad klasifikovaný iba do jednej triedy, vyberie sa najčastejšie sa vyskytujúca trieda. Okrem toho môže byť „hlasovanie“ jednotlivých susedov vážené podľa ich vzdialenosti voči klasifikovanému dokumentu, t.j. trieda najbližšieho suseda má najväčšiu váhu atď. Ak je potrebné klasifikovať dokumenty do viacerých tried, určí sa celkové skóre (suma pre všetkých susedov) pre každú triedu c_j a ak je skóre väčšie ako zvolená prahová hodnota príklad sa zaradí do c_j.

Support Vector Machines

SVM je metóda strojového učenia. Klasifikácia SVM hľadá rovinu, ktorá v priestore príznakov optimálne rozdeľuje trénované dáta. Optimálne navrhnutá rovina je taká, že body

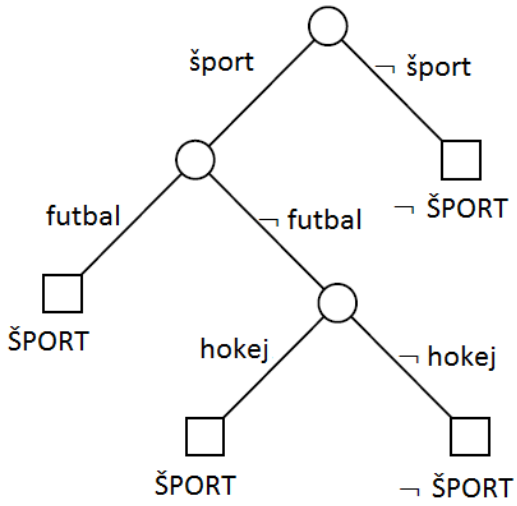
ležia v opačných polopriestoroch a hodnota najmenšej vzdialeností bodov od roviny je čo najväčšia – rovina H3 na obrázku nižšie. Okolo tejto roviny je čo najširší pruh bez bodov. Na popis roviny stačia len najbližšie body, ktorých je obvykle málo [39], [40]. Tieto body sa nazývajú podporné vektory, angl. support vectors, z nich pochádza názov metódy.



Obrázok 2.1–Z hľadiska optimálnosti rôzne navrhnuté roviny polopriestorov pri metóde SVM

Rozhodovacie stromy a pravidlá

Príklad rozhodovacieho stromu je uvedený nižšie. Každý nelistový uzol stromu je označený testom, ktorý rozdeľuje dokumenty podľa výskytu jedného termu. Listové uzly sú označené priradením triedy. Klasifikácia prebieha rekurzívne od koreňového uzla, zvolením vetvy podľa testu až pokiaľ sa nedosiahne listový uzol, t.j. pre uvedený príklad sú dokumenty zaradené do triedy ŠPORT ak obsahujú term „šport" a zároveň obsahujú term „futbal" alebo „hokej". Rozhodovacie stromy sú určené metódou "zhora nadol" [41]. Počiatočný strom je tvorený iba jedným uzlom, ktorý pokrýva všetky trénovacie dokumenty. Ak nastane prípad, že všetky príklady pokryté daným uzlom majú rovnakú triedu, delenie množiny dokumentov nie je potrebné a uzol sa označí ako listový. Inak algoritmus priradí uzlu logický test, ktorý rozdelí príklady na disjunktné podmnožiny, pre ktoré sa vytvoria nové uzly stromu. Celý proces sa potom rekurzívne opakuje na nových potomkoch, až pokiaľ sa nedosiahne úplné oddelenie príkladov jednotlivých tried.



Obrázok 2.2–Rozhodovací strom a pravidlá – trieda ŠPORT

Štatistické metódy klasifikácie

Pri štatistických metódach klasifikácie prebieha na začiatku proces trénovania kategórií, ktorými budeme schopní v testovacom procese spracovávaný text klasifikovať. Pri procese trénovania prebieha generovanie a zber štatistických dát z trénovanej databázy textov, najčastejšie početnosti slov alebo ich častí, kľúčových slov, N-gramov atď. Pri procese kategorizácie sa z kategorizovaného textu extrahujú rovnaké štatistické dáta a tieto sa porovnávajú s už dosiahnutými z procesu trénovania [42], [43], [44]. Na základe tohto porovnania a vyhodnotenia vieme prehlásiť nakoľko sa kategorizovaný text svojimi parametrami podobá parametrom natrénovaných kategórií. Ďalšia časť tejto práce sa bude venovať práve tomuto typu klasifikácie a kategorizácie textových dokumentov.

3 Ciele práce

Z predchádzajúcej analýzy a opisu súčasného stavu boli vytýčené nasledovné ciele práce:

- Vytvorenie metodiky na tvorbu doménových profilov pre textovú kategorizáciu
- Vytvorenie metodiky na porovnávanie doménových profilov natrénovaním z dostupného korpusu založenej na normovaní početnosti
- Vytvorenie metodiky na porovnávanie doménových profilov natrénovaním z dostupného korpusu založenej na absolútnej početnosti
- Voľba vhodnej metodiky porovnania profilov na základe hranice dĺžky testovaných textov vzhľadom na maximalizovanie miery správneho určenia kategórie textu (CCR).

4 Kategorizácia textu

Súčasťou kvalitného systému syntézy reči je blok vyššej syntézy reči. Táto časť sa zaoberá syntetizovaným textom, spracúva ho a vykonáva sa jeho predspracovanie za účelom získania a vydolovania užitočnej informácie z tohto textu. Do tejto časti patrí aj oblasť kategorizácie textu, ktorou sa táto práca zaoberá. Tieto informácie majú spravidla slúžiť na zdokonalenie výslednej syntetizovanej reči na úrovni zrozumiteľnosti a aj prirodzenosti výsledného signálu. Na Ústave telekomunikácií sa taktiež vyvíja rečový syntetizátor, ktorého súčasťou je aj blok vyššej syntézy reči. Cieľom tejto práce je navrhnutie a vytvorenie systému kategorizácie textu, založenej na štatistických metódach klasifikácie a zdokonalenie už existujúcich metód.

Existuje viacero klasifikačných metód a hodnotiacich funkcií pre klasifikáciu textu. V našej práci sme sa zamerali na dva hlavné typy kategorizačných prístupov. Sú to:

- Tf – term frequency – metóda vážených početností N-gramov
- Distance-based algoritmy – metódy vzdialenostných metrík N-gramov

Proces tvorby kategórií prebieha trénovaním z vopred pripravených textov, ktorých kategória je už známa. Pri kategorizácií potom z už natrénovaných príznakov dokážeme určiť kategóriu už aj neznámemu textu.

4.1 Proces tvorby a trénovanie doménových profilov

Štatistický proces na kategorizáciu textu využíva podobnosť extrahovaných štatistických údajov textu so známou kategóriou (proces trénovania) so štatistickými údajmi textu

neznámej kategórie (proces kategorizácie). Pri štatistickom prístupe kategorizácie textu sa celý proces teda skladá z dvoch hlavných častí:

- Proces trénovania
- Proces kategorizácie (testovanie)

4.2 Metóda klasifikácie na základe extrakcie N-gramov

Pri spracovávaní N-gramov z textu sa extrahuje informácia o početnosti N-gramov vygenerovaných zo spracovávaného textu. V našom ponímaní sa za N-gram považuje postupnosť N znakov, pochádzajúcich zo slov a výrazov. Pre prípady keď N=1, 2 alebo 3, postupnosti nazývame unigramy, bigramy resp. trigramy. Pri spracovaní a extrakcii N-gramov je nutné v prvom kroku určiť hranice slov a v následnom kroku z týchto slov vygenerovať postupnosti – N-gramy. Kategorizácia na základe N-gramov môže byť použitá aj na klasifikáciu a kategorizáciu jazyka použitého v texte. V [48] využívali práve špecifickosť generovaných N-gramov v závislosti od použitého jazyka. Profily získavané pri procese trénovania sa radikálne odlišovali pri jednotlivých svetových jazykoch. Pri procese trénovania použili korpus jednotlivých jazykov národov vo veľkosti do 120kB. Boli použité texty správ. Dosahované vybrané výsledky klasifikácie som znázornil v nasledovnej tabuľke:

Veľkosť textu	≤300	≤300	≤300	≤300	≥300	≥300	≥300	≥300
Veľkosť profilu	100	200	300	400	100	200	300	400
Brazil	70	80	90	90	91,3	91,3	95,6	95,7
Canada	100	100	100	100	100	99,6	100	100
France	90	95	100	95	99,6	99,6	99,2	99,6
Germany	100	100	100	100	98,9	100	100	100
Mexico	90,6	100	100	100	94,8	99,1	100	99,5
Poland	93,3	93,3	100	100	100	100	100	100

Navrhnutý systém v prípade veľkosti profilu 400 N-gramov len v 7 prípadoch z 3478 textov vykonal zlú klasifikáciu čo predstavuje 99,8% úspešnosť klasifikácie. Tento výsledok sa dá považovať za vynikajúci a v prípade kategorizácie jazyka veľmi spoľahlivý.

V prípade doménovej a vecnej kategorizácie (Subject Classification), ktorej sa venujeme v tejto práci, už takú úspešnosť nedosahoval. Pri rovnakej konfigurácii dosahoval úspešnosť kategorizácie 80% (CCR – correct classification rate).

Metóda kategorizácie na báze N-gramov má viacero výhod: nízke nároky na pamäťové zdroje (doménové profily veľkosti cca 300kB), jednoduchý proces trénovania nových doménových profilov, rýchla práca s doménovými profilmi, robustnosť voči nečistému resp. gramaticky nesprávnemu textu (preklepy, chyby pri OCR, atď.) v procese trénovania nových doménových profilov. Ako príklad robustnosti voči chybám v texte uvádzam nasledovný prípad chyby v slove „football“. V prípade vnesenia chyby, napr. „footboll“, nestrácame informáciu o celom slove, iba na zopár postupnostiach:

Gramaticky správne slovo:

foo	oot	otb	tba	bal	all

Gramaticky nesprávne slovo:

foo	oot	otb	tbo	bol	oll

V prípade jednej vnesenej chyby nám namiesto znehodnotenia celého slova footboll vygenerovalo 3 chybné a 3 korektné trigramy zo 6 trigramov.

4.3 Získavanie a extrakcia trénovacích textov

Na proces trénovania doménových kategórií je potrebná textová databáza, korpus, ktorá bude obsahovať veľké množstvo textu so známou kategóriou. Vzhľadom na to, že našim cieľom bolo vytvorenie systému kategorizácie textu pre slovenský jazyk, bolo potrebné pracovať a zber štatistických dát vykonávať nad slovenským textom. Takouto vhodnou databázou a textovým korpusom v slovenčine je Slovenský národný korpus (SNK), ktorý sme sa rozhodli použiť pri procese trénovania doménových profilov. SNK je elektronická databáza obsahujúca slovenské texty z rôznych štýlov, žánrov, vecných oblastí, regiónov a pod. vybavená prídavnými jazykovými informáciami. Zahŕňa publicistické, literárne, vedecké a ďalšie štýly a žánre, ktoré Jazykovedný ústav SAV vedecky spracováva. Zozbierané sú aj rôzne slovníky, monografie, časopisy, zborníky a iné. Realizuje sa v oddelení Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied v Bratislave.

Dôležité parametre vytvorených a vygenerovaných korpusov, ktoré sme používali na trénovanie a testovanie doménových profilov uvádzam v nasledovnej tabuľke:

Kolekcia	SNK1-1	SNK1-2	SNK2-1	SNK2-2-test
Veľkosť (MB)	448,3 MB	381,1 MB	490 MB	476,4 MB
Počet znakov	371 987 764	311 918 906	398 958 587	394 393 078
Počet slov	56 606 603	48 100 730	60 768 066	59 095 293
Počet viet	3 420 707	3 184 667	4 169 306	3 684 260

Tabuľka 1–Údaje vytvorených kolekcií korpusov SNK

4.4 Spracovanie textu pre tvorbu doménových profilov

Za účelom tvorby doménových profilov pri procese trénovania je nutné vhodne spracovať texty z dostupného korpusu. Slovenský jazyk je vo svojej podstate flektívny jazyk, ktorého slová a výrazy majú rôznorodú podobu. Za účelom eliminácie rôznorodosti slov a výrazov avšak so zachovanou vecnou informáciou sa zaviedli rôzne normovania slov vo fáze predspracovania textu. V práci [17] autor zadefinoval viacero etáp, ktoré sa používajú pri procese štatistickej analýzy:

- Členenie textu na slová
- Eliminácia neplnovýznamových slov
- Izolácia koreňa
- Voľba termov

4.4.1 Členenie textu na slová

Členenie na slová je inicializačná textová operácia, pri ktorej sa vo vstupnom texte identifikujú jednotlivé slová. V reťazci znakov vstupného textu sa identifikujú alfabetické znaky, čísla a ostatné znaky – oddeľovače, interpunkčné znaky, atď. Problémom tohto kroku analýzy je, že prirodzený text obsahuje „špeciálne“ reťazce, ktoré tvoria oddeľovače slov, rôzne znaky časových, dátumových formátov a pod. Tieto reťazce treba správne identifikovať a aj na základe nich určiť správnu hranicu slov, ktoré spájali.

4.4.2 Eliminácia neplnovýznamových slov

Za stop-slová sa pri počítačovom spracovaní prirodzeného jazyka označujú slová, ktoré sa v danom jazyku vyskytujú často avšak nenesú žiadnu významnú informáciu a majú spravidla len syntaktický význam. Bývajú to zväčša neohybné slovné druhy a neplnovýznamové slová, najmä spojky, predložky, častice, určité a neurčité členy, niektoré príslovky a zámená. Pre slovenské texty sú to napríklad a, v, na, ale, o, pri, atď. Mechanizmy na odstránenie stop-slov sú väčšinou implementované ako zoznamy týchto slov. Text spracovávaného dokumentu sa porovnáva so zoznamom a tie slová z textu, ktoré sa v zozname stop-slov nachádzajú, sa v ďalšom procese neberú do úvahy. Obyčajne sa stop-slová odstraňujú dvoma spôsobmi. Prvý používa zostavený zoznam neplnovýznamových slov, tzv. stop-list. Z textu sa pri identifikácii slova, ktoré sa nachádza v tomto zozname toto slovo eliminuje. Úspešnosť takéhoto postupu závisí od úplnosti definície stop-list zoznamu. Ďalšia nevýhoda je, že tento postup je závislý na použitom jazyku a nedá sa aplikovať na iné jazyky. Druhý spôsob je automatický, založený na odstraňovaní slov v dokumentoch vyskytujúcich sa s príliš malou a s príliš veľkou frekvenciou. Podľa [50] majú dostatočne silnú rozlišovaciu schopnosť tie slová, ktorých frekvencia dokumentov patrí do intervalu $\langle \frac{N}{100}, \frac{N}{10} \rangle$, kde N je počet dokumentov. Výhodou tohto prístupu je, že je použiteľný v praxi pre väčšinu jazykov, avšak v praxi dosahuje slabšie výsledky ako prvý prístup.

V ďalšej analýze sme zadefinovali skupinu stop-slov, ktoré boli eliminované a vyňaté z ďalšej analýzy. Sú to nasledovné slová: a, aby, aj, ako, ale, alebo, ani, asi, bez, by, byť, cez, čo, či, do, ho, i, iba, ja, je, jeho, jej, k, kam, každý, kde, kto, ktorý, ku, môcť, my, na, nad, niet, než, nič, o, od, on, po, pod, podľa, prečo, pred, preto, potom, pri, s, sa, si, so, svoj, tak, takže, teda, ten, tento, to, toto, tu, tvoj, ty, u, už, v, váš, viac, však, vy, z, za, že.

4.4.3 Izolácia koreňa – lematizácia a stemming

Lematizácia predstavuje jednu z možností normovania slov v jazyku. Lema reprezentuje všetky odvodené tvary slova v danom jazyku. V slovenčine existujú určité konvencie pri vytváraní lemy, napr. slovesá sú vždy v neurčitku, podstatné mená v nominatíve jednotného čísla, prídavné mená v predikatívnej pozícii a pod.

Stemming je proces redukcie flektívnych slov na ich koreň – slovotvorný základ. Napríklad pre slová školský, škole, školu, školník je slovotvorný základ „škol“.

Na naše účely sme vytvorili systém na lematizáciu slov, ktorý je založený na databáze Slovenského národného korpusu. Jedna z mnohých anotácií, ktorú SNK poskytuje je informácia o leme slov.

Nami navrhnutý systém lematizácie extrahoval informáciu o leme zo slov v korpuse SNK a zaznamenával do databázy. V takto získanej databáze sa nachádza 123206 flektívnych tvarov a výrazov, ktoré sú nahradené lemami s počtom 52811. Zásady lematizácie [6] v SNK som znázornil v nasledujúcej tabuľke:

Substantíva **Zámená** **Číslovky**	**Adjektíva** **Adjektiváliá**	**Slovesá**
príslušný rod singulár (ak existuje) nominatív (ak existuje)	maskulínum singulár nominatív pozitív	infinitív

Tabuľka 2–Zásady lematizácie v SNK

V nasledujúcej tabuľke uvádzam náhľad na nami vytvorenú databázu flektívnych tvarov a im prislúchajúcich lém vychádzajúc zo zásad lematizácie SNK.

Flektívny tvar	**Lema**
ázijská	Ázijský
ázijské	Ázijský
ázijského	Ázijský
ázijskú	Ázijský
ázijský	Ázijský
ázijských	Ázijský
ázijským	Ázijský
ázijskej	Ázijský
ázijskom	Ázijský

Tabuľka 3–Ukážka databázy flektívnych tvarov slov a im prislúchajúcich lém

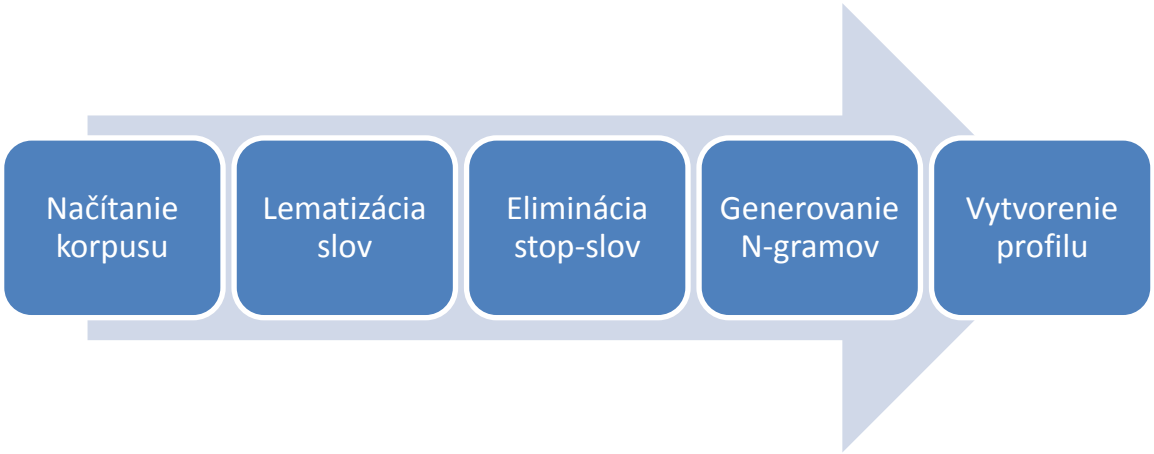
Takto navrhnutý lematizačný systém je síce spoľahlivý, avšak je obmedzený svojou databázou. Pokrýva a obsahuje najčastejšie používané tvary slov. Z tohto dôvodu je vhodné doplniť systém o pravidlá, ktoré by lemy generovali aj pre slová, ktoré sa v slovníku nenachádzajú.

4.5 Analýza získaných dát doménových profilov

V procese trénovania sme vychádzali z možností vecnej anotácie, ktorú ponúka SNK. V záujme overenia a testovania systému sme navrhli 3 vecné domény, na ktorých budeme testovať metódy klasifikácie textu:

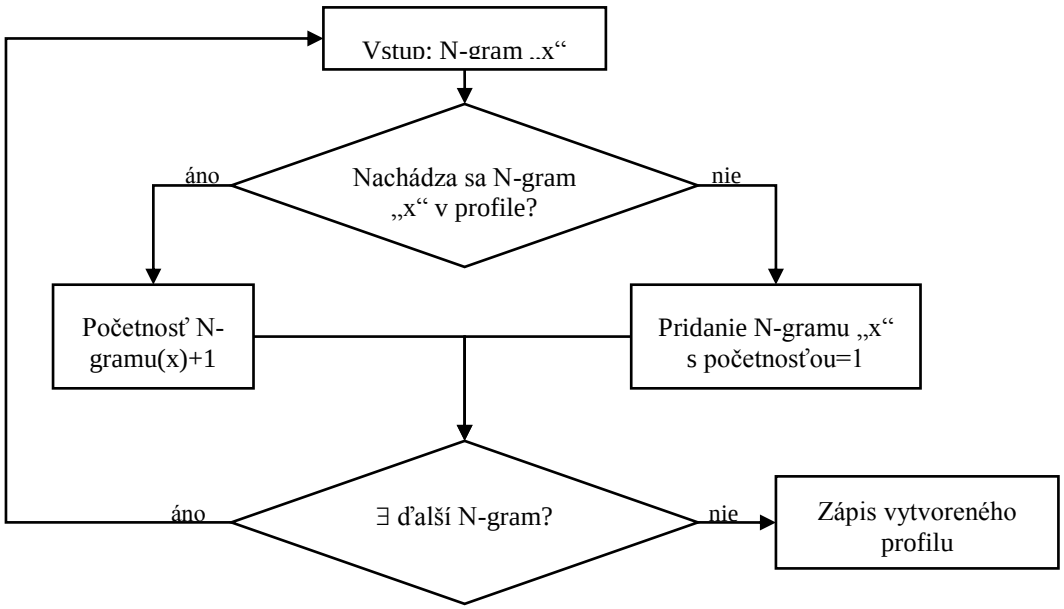
- Kultúra (CLT)
- Informatika (COM)
- Šport (SPO)

Napriek zvoleným trom kategóriám uvedené a použité postupy možno aplikovať na väčší počet kategórií. Za účelom získania, spracovania a vytvorenia jednotlivých doménových profilov sme vytvorili komplexný systém, ktorý všetky tieto funkcie vykonával na databáze z SNK. Celý proces tvorby doménového profilu je znázornený na nasledovnom obrázku:



Obrázok 4.1–Celý proces tvorby doménového profilu

V poslednej fáze tvorby doménového profilu, sa vygenerované N-gramy zahrnú do natrénovanej databázy profilu podľa nasledovného postupu:



Obrázok 4.2–Posledná fáza tvorby doménového profilu

Rozloženie početnosti N-gramov v doménových profiloch

Z rozloženia početností N-gramov v profile sme vychádzali pri samotnej kategorizácii textov. Predpokladali sme podobné rozloženie N-gramov testovaného textu s rozložením jemu prislúchajúceho profilu.

4.6 Proces komparácie profilov

Proces komparácie profilov patrí pri kategorizácií textov medzi kľúčové operácie, kedy prebieha rozhodovanie o kategórii testovaného textu. V tejto časti je našim cieľom porovnať profily všetkých natrénovaných doménových profilov s profilom testovaného textu a určiť kategóriu, ktorej doménový profil sa najviac podobá profilu testovaného textu. Procesy porovnávania jednotlivých profilov a následné určenie podobnosti nazývame metódami komparácie profilov.

4.6.1 Návrh procesu komparácie profilov

Celý návrh komparácie profilov a procesu kategorizácie je rozdelený na viac fáz. V prvej fáze je testovanému textu vytvorený profil rovnakým spôsobom, za použitia rovnakých postupov, ako doménovým kategóriám v trénovacej fáze. V bloku „miera vzdialenosti profilov“ je vykonaná štatistická miera na báze danej použitej metódy (metódy sú popísane v nasledujúcej podkapitole). Ako posledný blok nasleduje rozhodovací článok, ktorý prehlási najpodobnejší doménový profil k profilu testovaného textu.

4.6.2 Metódy komparácie profilov

Existuje viacero metód na porovnanie profilov a určovanie ich podobností. V tejto časti práce sú opísané použitia dvoch hlavných prístupov: spracovanie poradia N-gramov na základe ich početnosti (tzv. distance-based) a spracovanie normovanej (váženej) početnosti N-gramov. V ďalšej časti práce je uvádzané ich porovnanie, závislosti úspešnosti kategorizácie textu od dĺžky testovaných textov obidvoch hlavných prístupov a optimálne použitie danej metódy pre danú dĺžku testovacieho textu.

Komparácia profilov na základe početnosti a pozície N-gramu

Pri metóde komparácie profilov na základe početnosti N-gramov sme pracovali s poradím N-gramov, ich umiestnením v zoradenom profile podľa početnosti N-gramov. Tieto metódy sa tiež označujú ako distance-based, metódy na základe vzdialenostných metrík. Pri tejto metóde sa testovaciemu textu vytvorí profil N-gramov rovnako ako každému natrénovanému doménovému profilu. Miera rozdielnosti profilov je definovaná súčtom rozdielov pozícií každého N-gramu z profilu testovacieho textu od N-gramu v profile porovnávanej kategórie:

$$X_k = \sum_{n=1}^{N} abs(A_{n\_pos} - B_{n\_pos});$$

kde X_K je miera rozdielnosti profilov pre kategóriu K, A_{n_POS} a B_{n_POS} predstavujú pozíciu n-tého N-gramu v testovanom a porovnávanom profile.

Z hore uvedeného vzťahu vyplýva, že ak chceme vykonať kategorizáciu do K natrénovaných doménových profilov, takýto výpočet musí prebehnúť s každým doménovým profilom, tzn. K krát. Ukážku výpočtu takejto metriky uvádzam na nasledujúcom obrázku.

Poradie	N-gram profilu kultúra	Poradie	N-gram profilu testovaného textu	**Vzdialenosť**
1.	pre	1.	ých	**2**
2.	tor	2.	tor	**0**
3.	ých	3.	pre	**2**
10.	kto	4.	str	**21**
12.	pri	5.	val	**15**
20.	val	6.	pri	**6**
25.	str	7.	kto	**3**
			Výsledná suma:	**49**

Obrázok 4.3–Výpočet miery rozdielnosti N-gramov testovaného profilu s doménovým profilom kategórie

Víťaznú kategóriu X a doménový profil, ktorý ju zastupuje určíme výberom najmenšej miery rozdielnosti vzdialenosti zo všetkých dosiahnutých metrík X_K:

$$X=\min(X_k)$$

Komparácia profilov na základe váženej početnosti

Pri predchádzajúcej metóde sme pracovali s pozíciami a vzdialenostnými metrikami N-gramov a s ich umiestneniami v profile. Zaujímal nás rozdiel týchto pozícií v porovnávaných profiloch. Pri druhej metóde budeme uvažovať váženú početnosť N-gramov, ktorú sme vyjadrili váženou mierou početností na základe nasledovného vzťahu:

$$Y_k = \frac{\sum_{n=1}^{N} P_n}{\sum_{m=1}^{M} P_m}$$

kde Y_K je vážená miera početností kategórie K, čitateľ reprezentuje sumu všetkých početností v doménovom profile kategórie K porovnávaných N-gramov z testovaného profilu a menovateľ reprezentuje súčet všetkých výskytov všetkých obsiahnutých N-gramov v doménovom profile kategórie K. Tak isto ako pri predchádzajúcom postupe musí výpočet prebehnúť nad všetkými kategóriami natrénovaných doménových profilov.

Víťaznú kategóriu Y a doménový profil, ktorý ju zastupuje určíme výberom najväčšej váženej miery početností zo všetkých dosiahnutých Y_k.

$$Y = \max(Y_k)$$

Miera úspešnosti kategorizácie – CCR

Ako hlavný ukazovateľ úspešnosti a vhodnosti použitia danej metódy na komparáciu profilov sme zadefinovali mieru úspešnosti kategorizácie (Correct Categorization Rate, skr. CCR). Je definovaná podielom počtu správne kategorizovaných testovacích textov k celkovému počtu všetkých kategorizovaných textov:

$$CCR = \frac{\#\ správne\ kategorizovaných\ textov}{\#\ nesprávne\ kategorizovaných\ textov} \times 100\ [\%]$$

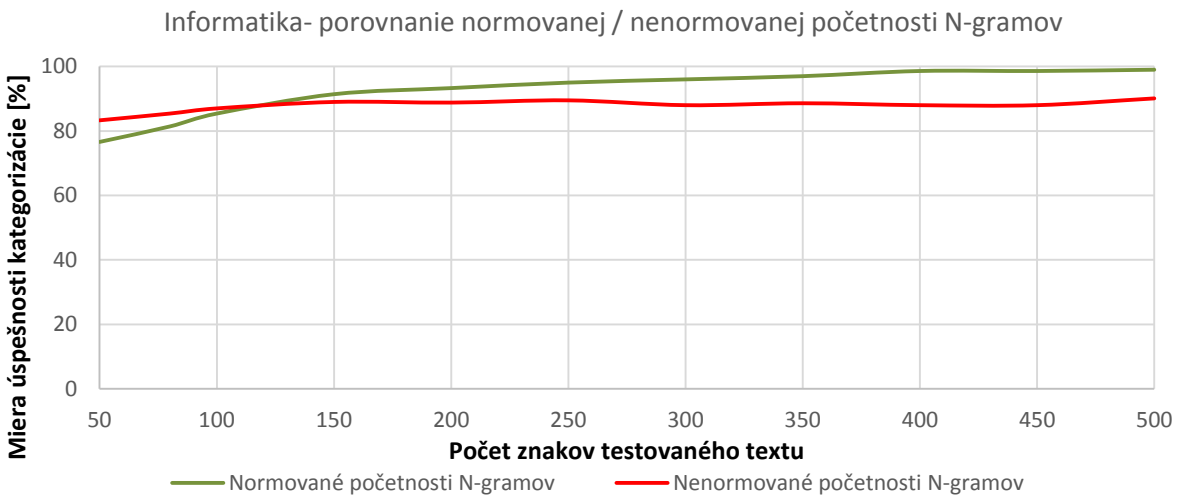
4.7 Testovanie a vyhodnotenie úspešnosti

Postup generovania testovacej kolekcie a následné testovanie prebieha nasledovne:

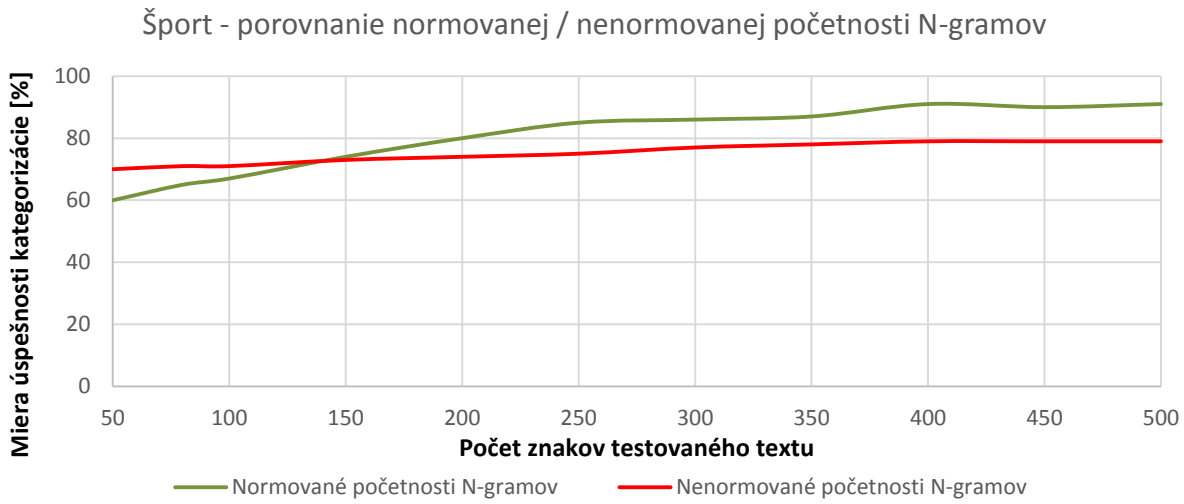
1. V prvom kroku sa zvolí jedna kategória z dostupných už natrénovaných, ktorej texty budú selektované a generované z testovacieho korpusu. Tento krok sme vykonali 3x (pre kultúru, informatiku a šport).
2. Zvolia sa parametre a konfigurácia testu t.j. dĺžka testovaných textov a počet testovaných textov.
3. Posledný krok tvorí vyhodnotenie testu. Výsledný štatistický súbor obsahuje úspešnosť testu, vyjadrenie percentuálnej úspešnosti kategorizácie vygenerovanej testovacej kolekcie.

4.7.1 Porovnávacie testy CCR oboch typov prístupov komparácie profilov

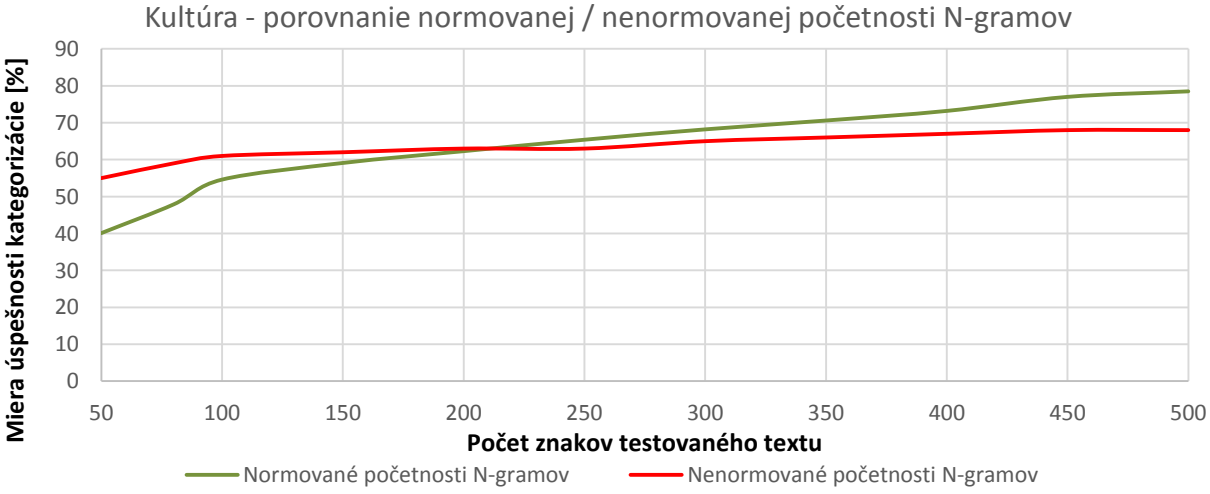
V tomto type testu sme mali za cieľ porovnať úspešnosť kategorizácie oboch typov prístupov metód komparácie profilov. Vykonali sme testy CCR s normovaním aj bez normovania početnosti N-gramov. Obe metódy boli testované na vzorke 1000 textov, veľkosti 50 až 500 znakov. Úspešnosti sú znázornené v tabuľke a vynesené do grafu. Tento postup sme aplikovali pre všetky 3 natrénované doménové kategórie.



Obrázok 4.4–Porovnávací test komparačných metód na doméne informatika



Obrázok 4.5–Porovnávací test komparačných metód na doméne šport



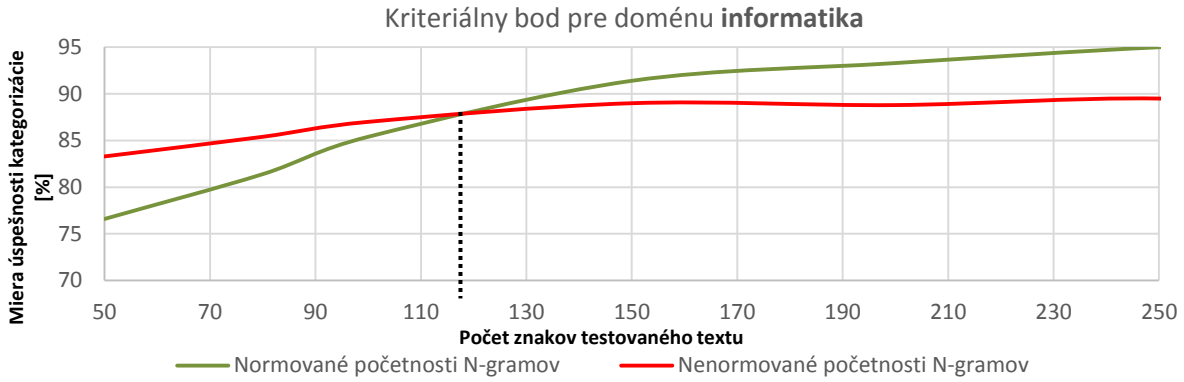
Obrázok 4.6–Porovnávací test komparačných metód na doméne kultúra

4.8 Voľba vhodnej metodiky komparácie profilov

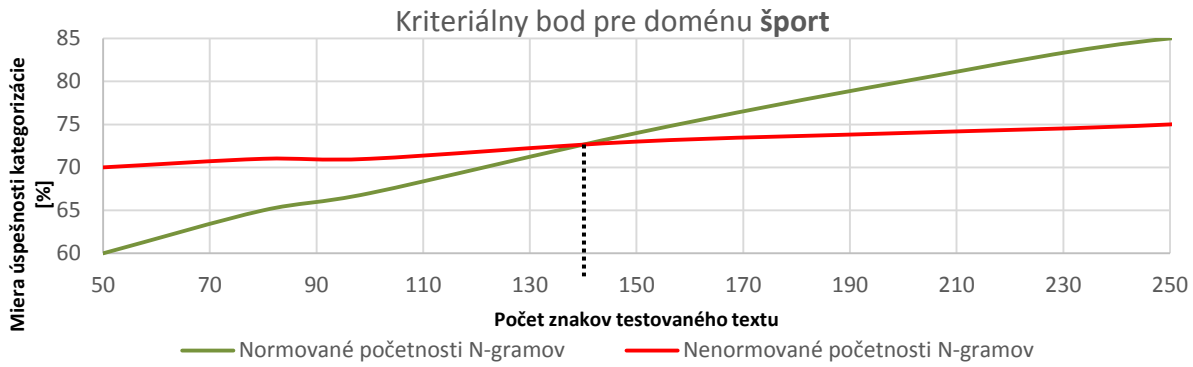
Z predchádzajúcich meraní a výsledkov testovania je možné pozorovať rôzny priebeh a závislosť úspešnosti kategorizácie na dĺžke kategorizovaných textov. Ako hlavné kritérium na voľbu vhodnej metodiky kategorizácie textov sme zvolili práve mieru úspešnosti kategorizácie CCR, ktorá priamo závisí od dĺžky kategorizovaných textov. Rôzna závislosť CCR od dĺžky kategorizovaného textu je spôsobená rôznou veľkosťou štatistického súboru N-gramov, ktoré systém kategorizácie generuje.

4.8.1 Závislosť CCR od veľkosti kategorizovaného textu

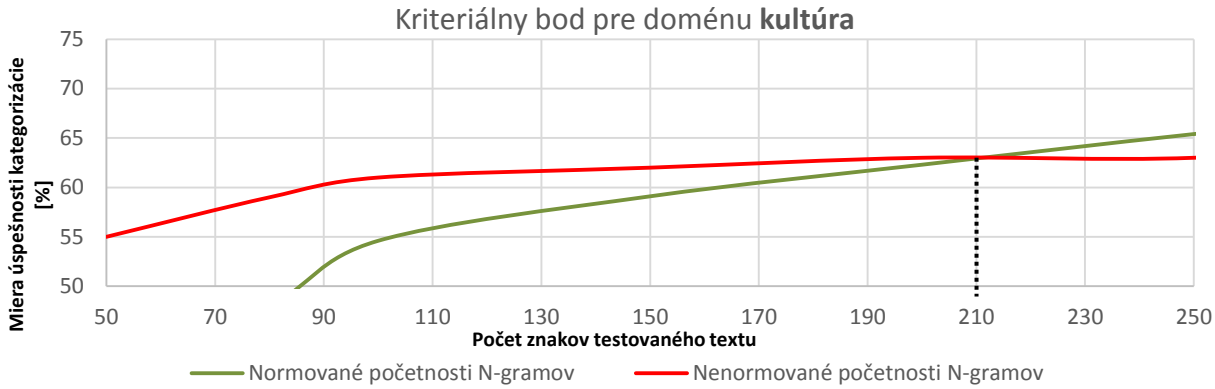
V tejto časti uvádzame detailný záber vykonaných meraní na priesečník závislosti úspešnosti kategorizácie dvoch rôznych použitých metód komparácie profilov. Nakoľko charakter použitých metód je rôzny, rôzna je aj závislosť CCR od veľkosti štatistického súboru vygenerovaných N-gramov. Z nasledujúcich závislostí badať, že metóda pracujúca s váhami jednotlivých N-gramov dosahuje spoľahlivejšie výsledky pri väčšom štatistickom súbore. Metóda pracujúca na princípe početností N-gramov a ich poradí, tzv. metóda distance-based, je úspešnejšia pri kratších textoch.



Obrázok 4.7–Detail kriteriálneho bodu pre doménu informatika



Obrázok 4.8–Detail kriteriálneho bodu pre doménu šport



Obrázok 4.9–Detail kriteriálneho bodu pre doménu kultúra

4.8.2 Optimálna hranica veľkosti kategorizovaného textu

Z predchádzajúcich grafov sme na základe pozorovaní odčítali kriteriálne body jednotlivých doménových kategórií. Tieto sú tvorené priesečníkmi funkčných závislostí úspešnosti kategorizácií od dĺžky textov jednotlivých komparačných metód. Dĺžky textov jednotlivých domén sú na základe nameraných kriteriálnych bodov naznačené v nasledovnej tabuľke:

Doména	Kultúra	Informatika	Šport
Dĺžka kategorizovaného	210	117	140

Tieto body určujú hranice medzi dvoma oblasťami a použitou metódou na komparáciu profilov pri kategorizácií textu. Pre kratšie texty, ktoré majú dĺžku textu pod danou hranicou je výhodnejšie použiť komparačnú techniku založenú na pozícií N-gramov na základe nenormovaných početností. Naopak, pre texty dlhšie ako kritériom určená hranica je výhodnejšie použiť komparáciu profilov založenú na váženej početnosti N-gramov, nakoľko tak dosahuje väčšiu úspešnosť kategorizácie tohto textu. Optimálny výber použitej metódy je znázornený v nasledovnej tabuľke:

Použitá komparačná metóda	Komparácia nenormovaných	Komparácia normovaných
Kultúra	≤ 210	> 210
Informatika	≤ 117	> 210
Šport	≤ 140	> 210

5 Výsledky dizertačnej práce

V tejto práci sa zaoberáme kontextovou analýzou a metódami kategorizácie textu slovenského jazyka. Napriek relatívne všeobecnému úvodu do spracovania textu pre rečový syntetizátor a kontextovej analýzy sa v práci podrobnejšie venujeme analýze hlavných prístupov metód ku kategorizácií textu a návrhu adaptabilnej metódy, ktorá ma za cieľ zvýšiť mieru úspešnosti kategorizácie textov v závislosti dĺžky kategorizovaných textov. Výsledkom práce je opis metodiky tvorby procesu kategorizácie textu slovenského jazyka. Navrhovaný postup je úpravou niektorých jazykovo závislých procesov adaptabilný a danú metodiku možno použiť aj pre kategorizáciu iných jazykov.

Práca je štruktúrovaná do šiestich hlavných kapitol a popisuje kontextovú analýzu a návrh procesov kategorizácie textov v slovenskom jazyku. V prvých troch kapitolách podávame prehľad o predspracovaní textu vo všeobecnosti a aj jeho využitie v aplikáciách rečového syntetizátora. Popisujeme tiež prístupy ku kontextovej analýze reči a používané metódy pri kategorizácií textu. Vo štvrtej kapitole uvádzame a vytyčujeme ciele tejto práce. V najrozšírenejšej piatej kapitole popisujeme použité a navrhnuté metodiky kategorizácie textu. V tejto kapitole uvádzame celý proces kategorizácie, tvorbu tohto systému, použité metodiky získavania trénovacieho a testovacieho korpusu, metodiky tvorby natrénovaných doménových profilov a popis metód komparácie profilov domén a testovacích textov. Ďalej venujeme pozornosť procesu testovania: tvorbe testovacej kolekcie dát a vyhodnoteniu testov. Z dostupných výsledkov a nameraných závislostí vidieť, že reakcia úspešnosti kategorizácie je pri porovnávaných metódach komparácie profilov rôzna v závislosti od dĺžky kategorizovaných textov. Za účelom zvýšenia úspešnosti kategorizácie a na základe nameraných závislostí sme stanovili pre každú doménu hranicu a limity dĺžky textov na použitie optimálnej metódy.

Celkovo možno vlastné dosiahnuté výsledky a prvky dizertácie zhrnúť do nasledujúcich bodov:

- Navrhnutá metodika na získavanie trénovacích dát, vytvorený vlastný postup predspracovania a normalizácie textu, vytvorená metodika na tvorbu doménových profilov pre textovú kategorizáciu a analýza vytvorených dátových súborov.
- Vytvorená metodika na porovnávanie doménových profilov z dostupných korpusov. Navrhnutý a implementovaný proces komparácie profilov založenej na normovaní početností N-gramov. Testovanie implementovaného procesu komparácie profilov.
- Navrhnutý a implementovaný proces komparácie profilov založenej na nenormovaní početností N-gramov. Testovanie implementovaného procesu komparácie profilov.
- Vytvorená metodika optimálneho výberu komparačnej metódy za účelom maximalizácie miery úspešnosti kategorizácie textu.

6 Konkrétne závery pre ďalší vedecký rozvoj

Jednou z hlavných motivácií pre realizáciu systému kategorizácie textu je jeho implementácia a využitie v systéme syntézy ľudskej reči. Takto aplikovaný systém kategorizácie by mohol výrazne ovplyvniť prirodzenosť generovanej reči. V tomto zmysle je samozrejme nutné prispôsobiť proces s ohľadom na systém syntézy: návrh doménových kategórií, komunikáciu atď.

Samotný systém na kategorizáciu sa skladá z viacerých procesov. Tieto procesy majú priestor na zlepšovanie, najmä v prostredí slovenského jazyka, pretože výskum v tejto oblasti nie je tak intenzívny ako v iných jazykoch. Ide najmä o metódy spracovania textu, extrakcie, normovanie a lematizácia slov, ktorými sme sa v práci zaoberali.

Ďalším námetom na výskum sú samotné metódy klasifikácie textov do kategórií, použité metriky a procesy komparácie profilov.

7 Riešené výskumné projekty autora

[1] Algorithms and Methods of Multimedia Signal Processing for Human Machine Interface, Rozinaj Gregor, VEGA 1/0718/09, (2009-2011)

[2] ASIMD - Audio-Speech Interface for Mobile Devices, Rozinaj Gregor, DAAD, (2010-2011)

[3] HBB-Next - Next-Generation Hybrid Broadcast Broadband (http://www.hbb-next.eu), Rozinaj Gregor, Small or medium-scale focused research project (STREP) proposal ICT Call 7, FP7-ICT-2011-7 - 287848, (FEI No: 5828) (2011-2014)

[4] IMUROSA - Integration of Multimedia Signal Processing Methods into Multimodal Interface and Network Applications (Integrácia metód spracovania MUltimediálnych signálov do multimodálneho ROzhrania a Sieťových Aplikácií), Rozinaj Gregor, VEGA 1/0708/13, (FEI No: 1494/115722) (2013-2015)

[5] Optimalizácia efektívnosti kódovania videa pre prenos a záznam, VEGA-1/0602/11, (2011-2013), prof. Ing. Jaroslav Polec, PhD.

[6] Pokročilé algoritmy spracovania obrazov na efektívne vyhľadávanie a kódovanie ľudských tvárí, VEGA 1/0961/11, (2011-2013), doc. Ing. Jarmila Pavlovičová, PhD.

8 Ocenenia autora

[1] 1. Miesto vo fakultnej súťaži ŠVOČ, Odbor: Telekomunikácie, 2008

[2] Ocenenie dekana fakulty za výborne vypracovanú diplomovú prácu, 2010.

9 Publikácie

Kvalifikačné práce

[1] TÓTH, J.: Inteligentné rozhranie pre komunikáciu s počítačom v aplikácií čítanie rss správ, [Bakalárska práca], FEI STU, Katedra telekomunikácií, Bratislava, 2008.

[2] TÓTH, J.: Fonetická transkripcia skratiek pri syntéze reči, [Diplomová práca], FEI STU, Katedra telekomunikácií, Bratislava, 2010.

[3] TÓTH, J.: Kontextová analýza textu, [Písomná správa k dizertačnej skúške], FEI STU, Katedra telekomunikácií, Bratislava, 2010.

Domáce konferencie

[4] TÓTH, J., VALENT, M.: Inteligentné rečové komunikačné rozhranie pre komunikáciu s počítačom v aplikácii Čítanie RSS správ a Program kín, STOČ, UTB Zlín, 25.4.2008.

[5] TÓTH, J., ROZINAJ, G.: Aplikácia čítanie RSS správ pre Inteligentné rečové komunikačné rozhranie. In: ŠVOC, Katedra telekomunikácií, FEI STU, Bratislava, apríl 2008

Medzinárodné konferencie

[6] TÓTH, J., VALENT, M.: Intelligent Speech Communication Interfacefor Communication with Computer in RSS News Article Reading Application, In: Redzur International Workshop on Speech and Signal Processing, Bratislava, May 2008.

[7] TÓTH, J., KONDELOVÁ, A., GONŠOR, J., VALENT, M.: Intelligent Interface for Communication in applications Reading RSS, Cinema program, Timetable buses and trains. In: Redzur International Workshop on Speech and Signal Processing, Bratislava, September 2009.

[8] TÓTH,J.: Phonetic Abbreviation Transcription in Speech Synthesis, In: Redzur International Workshop on Speech and Signal Processing, Bratislava, May 2010.

[9] KONDELOVÁ, A., TÓTH, J., ROZINAJ, G.: Analysis of Prosody Features in Slovak. In: Proceedings ELMAR-2010 : 52nd International Symposium ELMAR-2010. Zadar, Croatia, 15.-17.9.2010. Zadar : Croatian Society Electronics in Marine, 2010. ISBN 978-953-7044-11-4. s. 371-374.

[10] KONDELOVÁ, A., TÓTH, J., DROZD, I., HORVÁTH, T., SEMBER, M.: Modular Speech Synthesizer, In: 5th International Workshop on Multimedia and Signal Processing, Bratislava, 12. May 2011, ISBN 978-80-227-3506-3.

[11] TÓTH, J., KONDELOVÁ, A., GUZMICKÝ, P.: Simulation of Prosody Contours with Embedded Signal Generator, In: IWSSIP- International Conference on Systems, Signals and Image Processing. 16 -18 June, 2011,Sarajevo, Bosna and Hercegovina.

[12] TÓTH, J., KONDELOVÁ, A., ROZINAJ, G.: Natural Language Processing of Abbreviations. In: Proceedings ELMAR-2011 : 53rd International Symposium ELMAR 2011, 14-16 September 2011, Zadar, Croatia. Zadar : Croatian Society Electronics in Marine, 2011. ISBN 978-953-7044-12-1. s. 225-228

[13] KONDELOVÁ, A., TÓTH, J., VASEK, M. ROZINAJ, G.: Introduction to Speech Synthesis Management Tools. In: IWSSIP 2012 : 19th International Conference on Systems, Signals & Image Processing. Vienna, Austria, April 11-13, 2012. Vienna : Technical University, 2012. ISBN 978-3-200-02588-2. s. 643-646

[14] MOLČAN, Ľ., VANČO, M., TÓTH, J.: Touchless Computer Control with Hands. In: Proceedings Redžúr 2012 : 6th International Workshop on Multimedia and Signal Processing. April 11, 2012, Vienna, Austria. Bratislava : Nakladateľstvo STU, 2012. ISBN 978-80-227-3686-2. s. 77-80

[15] ŠTOFAŇÁK, M., TÓTH, J., KONDELOVÁ, A.: Computer Control with Eyes Pupil Localization. In: Proceedings Redžúr 2012 : 6th International Workshop on Multimedia and Signal Processing. April 11, 2012, Vienna, Austria. Bratislava : Nakladateľstvo STU, 2012. ISBN 978-80-227-3686-2. s. 81-84

[16] KONDELOVÁ, A., TÓTH, J., SEMBER, M., ROZINAJ, G.: Prosody Modification by using Sinusoidal Models. In: Proceedings Redžúr 2013 : 7th International Workshop on Multimedia and Signal Processing. 1. máj, 2013, Smolenice, Slovensko. Bratislava : Nakladateľstvo STU, 2013, ISBN 978-80-227-3921-4, s. 9-13.

[17] TÓTH, J., KONDELOVÁ, A., ROZINAJ, G.: N-gram-Based Text Categorization. In: Proceedings Redžúr 2013 : 7th International Workshop on Multimedia and Signal Processing. 1. máj, 2013, Smolenice, Slovensko. Bratislava : Nakladateľstvo STU, 2013, ISBN 978-80-227-3921-4, s. 23-26.

[18] TÓTH, J., DROZD, I., ROZINAJ, G.: Text Categorization Implementation into the Modular Speech Synthesizer. In: Proceedings ELMAR-2013 : 55rd International Symposium ELMAR 2013, 25-27 September 2013, Zadar, Croatia. Zadar : Croatian Society Electronics in Marine, 2013. – Paper accepted1

Publikácie v zahraničných vedeckých časopisoch

[19] TÓTH, J., KONDELOVÁ, A., ROZINAJ, G.: Advanced Text Categorization Methods with Statistical Approach. In: Elektrorevue, Vol.4, No. 2, 20 June 2013. Brno: ISES (International Science and Engineering Society), 2013. ISSN 1213-1539. S. 40-44.

[20] TÓTH, J., KONDELOVÁ, A., ROZINAJ, G.: Statistical Approach for Prosody Contour Modeling based on Sentence Classification. In: Elektrorevue, Vol. 4, No. 2, 20 June 2013. Brno: ISES (International Science and Engineering Society), 2013. ISSN 1213-1539. S. 34-39.

10 Zoznam použitej literatúry

[1] Psutka, J. Mluvíme s počítačem česky. Praha : Akadémia Praha, 2006. ISBN 80-200-1309-1.

[2] Rybárová, R. Metódy učenia pre syntézu reči. [Teoretická príprava k dizertačnej práci]. Bratislava : FEI STU, Katedra telekomunikácií, 2008.

[3] Sproat, R. W. Multilingual Text To Speech Synthesis. : Kluwer Academic Publishers Norwell, 1997. ISBN 0792380274.

[4] Kondelová, A. Analýza prozodických vlastností slovenskéj reči. [Diplomová práca]. Bratislava : FEI STU, Katedra telekomunikácií, máj 2010.

[5] Ján, T. Fonetická transkripcia skratiek pri syntéze reči. [Diplomová práca]. Bratislava : FEI STU, Katedra telekomunikácií, 2010.

[6] Garabík, R.; Gianitsová, L.; Horák, A.; Šimková, M. Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. 2004.

[7] Paralič, J.; Furdík, K.; Tutoky, G.; Bednár, P.; Sarnovský, M.; Butka, P.; Babič, F. Dolovanie znalostí z textov. Košice : s.n., 2010. ISBN 978-80-89284-62-7.

[8] Lužný, M. Normalizácia vstupného textu pre rečový syntetizátor. [Diplomová práca]. Bratislava, 2009.

[9] Laclavík, M. a Šeleng, M. Vyhľadávanie informácií. Bratislava : Slovenská technická univerzita v Bratislave, Vydavateľstvo STU, 2012. ISBN 978-80-227-3829-3.

[10] Sičáková, Ľ. Slovenská morfológia v teorii a v praxi. [Diplomová práca]. Prešov.

[11] Valent, M. Automatická detekcia slovných druhov v slovenskej vete. [Diplomová práca]. Bratislava : FEI STU, Katedra telekomunikácií, 2010.

[12] Turi Nagy, M. Využitie sínusoidálneho modelu pre spracovanie audio signálu. [Dizertačná práca]. Bratislava : FEI STU, Katedra telekomunikácií, 2006.

[13] Páleš E. SAPFO – Parafrázovač slovenčiny. 1. vyd. Bratislava: VEDA, 1994, 305 s. [Online]

[14] Sgall, P., Hajičová, E., Buráňová, E. Aktuální členění věty v češtině. Academia, nakladatelství ČSAV, Praha, 1980.

[15] Cigánik, M. Automatizovaný systém spracovania textových informácií. Alfa, Bratislava, 1989.

[16] Barrody, A.J., Dewitt, D.J. Object-oriented approach to database system implementation. ACM Transactions on Database Systems, 6, p.581-601, 1981.
[17] Furdík, K. Získavanie informácií v prirodzenom jazyku s použitím hypertextových štruktúr. [Doktorandská dizertačná práca]. Košice : FEI TU, 2003.
[18] Sparck-Jones K., Willett P. Readings in Information Retrieval. Academic Press/Morgan Kaufmann, 1997. [Online]
[19] Šimková, M. a Rehm, G. Slovenský jazyk v digitálnom veku. Berlín : Springer, 2012, Vol. 5, s. 30-36. ISBN 987-3-642-30370-8.
[20] Jurafsky, D.; James, H. M.; Kehler, A. Speech and Language processing: an introduction to natural language processing. Comutational linguistics and speech recognition. 1999, Vol. 2.
[21] Krajči, S.; Novotný, R.; Turlíková, L. Použitie lematizácie vo fulltextovom vyhľadávaní v slovenských dokumentoch. 2nd Workshop on Intelligent and Knowledge oriented Technologies. 2007.
[22] Porter M. F. An algorithm for suffix stripping. In: V. Program 14(3), 1980, s.130-137.
[23] Porter, M. [Online] http://tartarus.org/~martin/PorterStemmer/.
[24] Galamboš L. Lemmatizer for Document Information Retrieval Systems in JAVA. In: SOFSEM 2001: Theory and Practice of Informatics. Proceedings. LNCS 2234, Springer Berlin / Heidelberg, 2001. [Online]
[25] Krajči S., Laclavík M., Novotný R., Turlíková L. The tool Morphonary/ Tvaroslovnik: Using of word lemmatization in processing of documents in Slovak. In: P. Návrat, D. Chudá (eds.), Proceedings of the 8th annual conference Znalosti (Knowledge) 2009,. Brno, 4.-6. február 2009. Vydavateľstvo STU, Bratislava, 2009, s. 119-130. [Online]
[26] Řezanková H., Húsek D., Snášel V. Shluková analýza dat. Professional Publishing, Praha, 196 s, 2007.
[27] Yi, G., Zhiqing, S. a Hua, N. Content-Oriented Automatic Text Categorization with the Cognitive Situation Models. International Symposium on Computer Science and Computational Technology ISCSCT '08. 22-28 December 2008, s. 512-516.
[28] Wang, M. Y. a Liu, T. Method of Chinese Text Categorization Based on Variable Precision Rough Set. Third International Symposium on Intelligent Information Technology Application Workshops, 2009. IITAW '09. 21-22. November 2009, s. 26-29.
[29] Lili, H.; Lizhu, H. Automatic Identification of Stop Words in Chinese Text Classification. International Conference on Computer Science and Software Engineering, 2008. December 12.-14., 2008, Vol. 1, pp. 718-722.
[30] Han, L.; Yu, Z.; Deng, J.; Zhang, Ch. The effects of domain knowledge relations on domain text classification. 27th Chinese Control Conference, 2008. CCC 2008. 16.-18. Júl 2008, s. 460-463.
[31] Salton G., Wong A., Yang C. S. A vector space model for automatic indexing. In: Communications of the ACM, Vol. 18, Issue 11, 1975, s. 613-620. [Online]
[32] Dumais S. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, and Computers, 23(2):229–236, 1991. [Online]
[33] Larkey L. S., Croft W. B. Combining classifiers in text categorization. Proceedings of SIGIR-96, 19th ACM International conference on research and development in informationa retriaval, Zürich, 289-297, 1996. [Online]
[34] Meadow C. Text information retrieval systems. Academic Press Inc., San Diego, California, 1992.
[35] Farrow J. F. A cognitive process model of document indexing. Journal of Documentation, 47(2): 149-166.

[36] Sang-Bum, K.; Kyoung-Soo, H.; Hae-Chang, R.; Sung Hyon, M. Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering. November 2006, Vol. 18, pp. 1457-1466.
[37] Liang Jun, L.; Bin, Z.; Yuanyuan, Ch.; Ming, Y. KD-KNN Text Categorization Method Based On Improvement TFI-DF. International Conference on Information Engineering and Computer Science, 2009. ICIECS 2009. 19.-20.. December 2009, s. 1-5.
[38] Xinghua, F.; Hongge, H. A New Model for Chinese Short-text Classification Considering Feature Extension. International Conference on Artificial Intelligence and Computational Intelligence (AICI), 2010. 23.-24.. Október 2010, s. 7-11.
[39] Suzuki, M.; Yamagishi, N.; Yi-Ching, T.; Ishida, T. English and Taiwanese text categorization using N-gram based on Vector Space Model. International Symposium on Information Theory and its Applications (ISITA), 2010. 17-20. Október 2010, s. 106-111.
[40] Suzuki, M.; Yamagashi, N.; Ishida, T.; Goto, M. On a new model for automatic text categorization based on Vector Space Model. IEEE International Conference on Systems Man and Cybernetics (SMC), 2010. 10.-13.. Október 2010, s. 3152-3159.
[41] Zheng, Z.; Gang, B.; Jinhua, L.; Yi, Z. An intelligent representing system of search results. International Conference on Computer and Information Application (ICCIA), 2010. 3.-5.. December 2010, s. 297-300.
[42] Bo, Ch.; Hui, H.; Jun, G. Language Feature Mining for Document Subjectivity Analysis. The First International Symposium on Data, Privacy, and E-Commerce, 2007. ISDPE 2007. 1.-3.. November 2007, s. 62-67.
[43] Zhen, Y.; Xiangfei, N.; Weiran, X.; Jun, G. Application of the Character-Level Statistical Method in Text Categorization. International Conference on Computational Intelligence and Security, 2006. 3.-6.. November 2006, Zv. Vol. 2, s. 1412-1417.
[44] Zhou, F.; Zhang, F.; Yang, B.; Xingang, Y. Research on Short Text Classification Algorithm Based on Statistics and Rules. Third International Symposium on Electronic Commerce and Security (ISECS), 2010. Júl 29.-31., 2010, pp. 3-7.
[45] Mohammed, F. S.; Zakaria, L.; Omar, N.; Albared, M. Y. Automatic Kurdish Sorani text categorization using N-gram based model. International Conference on Computer & Information Science (ICCIS). 12-14. Jún 2012, s. 392-395.
[46] Lakshmi, K. a Mukherjee, S. Profile Extraction from Mean Profile for Automatic Text Categorization. International Conference on Computational Intelligence for Modelling, Control and Automation. 28-30. November 2005, s. 384-389.
[47] Špilka, M. Určenie témy textu s využitím pravdepodobnosti výskytu slov v jazyku. [Bakalárska práca]. Bratislava : FEI STU, Katedra telekomunikácií, 2013.
[48] Cavnar, W. B.; Trenkle, J. M. N-gram-based text categorization. Ann Arbor MI : s.n., 1994.
[49] Jazykovedný ústav Ľ. Štúra SAV. Slovenský národný korpus – r-mak-3.0. [http://korpus.juls.savba.sk] Bratislava : s.n., 2009.
[50] Salton, G., Yang, C. a Yu, C. A theory of term importance in automatic text. Journal of the American Society for Information Science. 1975, s. 237-253.
[51] Toman, M.; Tesar, R.; Ježek, K. Vliv normalizace slov na klasifikaci textu. Plzeň : Katedra informatiky a výpočetní techniky, FAV.
[52] Li Y.; Sheng Y.; Luan, L. A Text Classification Method with an Effective Feature Extraction Based on Category Analysis. 2009, Fuzzy Systems and Knowledge Discovery, s. 95-99.

11 Resumé

In this work we deal with contextual analysis and text categorization methods in Slovak language. The scope of this thesis is focused on a design of a text categorization system. Improvements of already known categorization methods and designing new categorization methodology for achieving better classification rate. The work also includes procedure of complex categorization process, the creation of this system, the methodology of obtaining training and testing corpus from Slovak National Corpus database. In other part the domain profile creation is described and methods of comparison each profiles are designed. Available results and the measured dependence implies that the success of categorization process used with the method of comparison of profiles varies depending on the length of categorized texts. In order to increase correct categorization rate on the basis of the measured dependence was determined for each domain boundary and limits of texts to use optimal methods.

This work had delivered following results:

- Proposed methodology for obtaining training data, create a custom procedure preprocessing and normalization of text, created the methodology for the creation of domain profiles
- Created methodology for comparing the domain profiles from the available corpus. Designed, implemented and tested a process of comparison of profiles based on normalized N-grams frequencies.
- Designed, implemented and tested a process of comparison of profiles based on non-normalized N-grams frequencies and their positions.
- Based on the analysis of the results was developed optimal methodology choice of the comparative method in order to maximize the success rate of text categorization

Poznámky

Poznámky

Poznámky