Natančnost uvrščanja slovenskih besedil

Marko Hölbl, Petra Grm, Boštjan Brumen, Tatjana Welzer, Izidor Golob
Fakulteta za elektrotehniko, računalništvo in informatiko,
Univerza v Mariboru, Smetanova 17, 2000 Maribor
{marko.hölbl, petra.grm, brumen, welzer, izidor.golob}@uni-mb.si

Izvleček

V članku predstavljamo meritve natančnosti strojnega uvrščanja slovenskih besedil. Sorodne raziskave na tem področju so bile izvedene predvsem na angleško govorečem področju in na angleških besedilih, prav tako jih je bilo moč zaslediti tudi na ostalih govornih področjih, na primer na nemškem in kitajskem. Na slovenskem govornem področju pa podobnih raziskav ni bilo možno zaslediti. Namen raziskave je ugotavljanje učinkovitosti tujih orodij pri uvrščanju slovenskih besedil. V ta namen smo izbrali IBM-ovo orodje za uvrščanje besedil, imenovano "IBM Intelligent Miner for Text" [7]. Raziskava je potekala v 3 fazah. Najprej smo zbrali množico 270 slovenskih besedil iz dnevnega časopisja in jih razvrstili v 3 kategorije: politika, gospodarstvo in šport. Množico smo razdelili na učni del na testni del. Nato je sledila izgradnja uvrščevalnega modela s pomočjo učne množice. V tretji fazi smo testirali natančnost izgrajenega modela na podlagi testnega dela. Rezultati uvrščanja slovenskih besedil so se izkazali za presenetljivo dobre (natančnost nad 87 %), kljub temu, da smo uporabili orodje, ki ni namenjeno slovenskem govornem območju.

Abstract

Accuracy of Categorization of a Slovene Text

The article presents the topic of automatic text categorization of a Slovene text. Related articles mostly deal with the categorization of an English text; some address other languages like German or Chinese as well. A study of Slovene text categorization has not been carried out yet. We used a tool called "IBM Intelligent Miner for Text" [7] for this task. The research was conducted in three phases. In the first phase we collected 270 articles from daily newspapers and divided them into three categories: economics, politics and sports. We split the articles into a learning-group and a test-group. We created a categorization model in the second phase on the basis of the learning-group. Then in the third phase we tested the categorization model on the test-group. In spite of the fact that we used a foreign tool, which is not designed for the Slovene language environment, the accuracy of the tool was very good (average 87 %).

1 UVOD

Z razvojem sodobne informacijske družbe in s tem povezanim ogromnim naborom javno dostopnega pisnega gradiva v elektronski obliki se poraja potreba po samodejnem uvrščanju besedil po kategorijah. To omogoča hitrejše iskanje želenih informacij in hkrati prinaša večjo preglednost nad podatki. Prav zato je raziskovanje področja samodejnega uvrščanja besedil danes zelo aktualno.

Zato smo naredili raziskavo natančnosti uvrščanja slovenskih besedil. Ker pa slovenskega orodja ni na voljo, smo nalogo opravili s tujim. Izbrali smo IBM-ovo orodje za uvrščanje besedil "IBM Intelligent Miner for Text" [7]. Orodje nam omogoča uvrščanje besedil v kategorije s pomočjo poprejšnjega nadzorovanega učenja.

Članek je razdeljen na šest razdelkov. V razdelku 2 so navedene sorodne raziskave s tega področja, ki pa jih tudi v angleškem prostoru ni prav veliko. Razdelek 3 obravnava opis problema in metodologijo; 4. razdelek govori o postopku izvedbe uvrščanja besedil v kategorije. V razdelku 5 so podani dobljeni rezultati in obrazložitve le-teh; zadnji, 6. razdelek je zaključek.

2 SORODNE RAZISKAVE

Prve javno objavljene raziskave s področja avtomatskega uvrščanja besedil segajo v leto 1990. V tem času se je namreč povečala potreba po samodejnem uvrščanju tekstov.

Največ raziskav s tega področja temelji, kot pričakovano, na angleških besedilih. Zasledili smo tudi raziskave z nemško govorečega področja, medtem ko je uvrščanje besedil v drugih jezikih manj raziskano.

Tiskovna agencija Reuters je eno prvih podjetij, ki se je začelo aktivno ukvarjati s tem problemom. Že leta 1990 je uporabljalo ekspertni sistem Construe [2] za avtomatsko uvrščanje besedil in z njim doseglo kar dobre rezultate. Učenje (izgradnja modela) je potekalo s pomočjo ročno vnesenih pravil, ki so jih pripravili domenski eksperti v podjetju. Razporejanje v kategorije je temeljilo na algoritmu k-tega najbližjega soseda (angl. k-Nearest Neighbor - kNN) [3]. Njihov sistem je na množici 750 testnih primerkov dosegel 90 % natančnost [2].

Prav tako smo zasledili članke, ki opisujejo primerjavo različnih algoritmov samodejnega uvrščanja tekstov. Članek [6] navaja različne algoritme oz. postopke, ki temeljijo na linearni algebri, teoriji verjetnosti in ostalih postopkih. Najbolj uporabljani med njimi so:

- Sistem k-tega najbližjega soseda (angl. k-Nearest Neighbor - kNN)
- Podporni vektorski stroji (angl. Support Vector Machines - SVM)
- Linearna aproksimacija s pomočjo najmanjših kvadratov (angl. Linear Least Squares Fit - LLSF)
- Nevronske mreže (angl. Neural Network – NNet)
- Metoda naivnega Bayesa (angl. Naive Bayes – NB)

Tudi ta članek obravnava samo primerjavo natančnosti posameznih algoritmov na angleških besedilih.

V članku [1] je bila izvedena raziskava na angleških in nemških besedilih. Tudi na nemških besedilih se je sistem Construe izkazal kot zelo perspektiven, čeprav je bila njegova uspešnost manjša kot pri uvrščanju angleških tekstov. To smo pričakovali, saj gre za angleško orodje. Sistem je uporabljal slovar, ki ga je zgradil v fazi učenja, hkrati s pravili, ki so jih določili uporabniki. Pri tej raziskavi je bila izvedena tudi primerjava ekspertnega sistema s samoučečim se postopkom, temelječim na odločitvenih drevesih in sistemu z Bayesovo verjetnostjo. Tudi tu se je za najučinkovitejšega izkazal Construe sistem.

Najbližje naši raziskavi je članek [4], kjer je skupina kitajskih raziskovalcev preučevala natančnost uvrščanja kitajskih besedil. Za uvrščanje so uporabljali naslednje metode:

- Sistem k-tega najbližjega soseda (angl. k - Nearest Neighbor - kNN)
- Podporni vektorski stroji (angl. Support Vector Machines - SVM)
- Adaptivna resonančna asociativna mreža (angl. Adaptive Resonance Associative Map – ARAM)

Postopki so se izkazali pri predpostavki, da je bila učna množica dovolj velika, kot relativno dobri, vendar avtorji ne navajajo konkretnega podatka o doseženi natančnosti.

3 OPIS PROBLEMA IN METODOLOGIJA

Samodejno uvrščanje se s pridom uporablja pri internetnih iskalnikih, za urejanje elektronske pošte in uvrščanje raznih člankov [8].

Cilj naše raziskave je bil ugotoviti natančnost uvrščanja slovenskih časopisnih člankov s pomočjo tujega orodja.

Na začetku raziskave smo bili skeptični glede natančnosti uvrščanja slovenskih besedil, saj smo uporabljali ameriško programsko orodje [7], ki ni namenjeno slovenskemu govornemu področju.

Za to orodje smo se odločili zaradi treh poglavitnih razlogov. Prvi razlog je bil ta, da slovenskega orodja za uvrščanje slovenskih besedil ni. Drugi pomemben razlog je dejstvo, da je orodje brezplačno za akademske ustanove. Tretji razlog: za že izdelano programsko orodje smo se odločili tudi zato, ker večina vsakdanjih uporabnikov ni dobro podkovana v poznavanju različnih algoritmov uvrščanja, ampak želijo le usluge programa. Zato smo nalogo uvrščanja opravili tako, da smo v orodje vnesli vhodne podatke in opazovali izhodne podatke. Princip, ki je znan kot princip črne škatle, je bil v našem vidiku zelo primeren, saj smo lahko z njim ugotovili dejansko učinkovitost oz. natančnost orodja, ki ga lahko uporablja tudi o podrobnostih nepoučeni uporabnik.

V naslednjem poglavju opisujemo izvedbo eksperimenta.

4 OPIS POSTOPKA

4.1 Priprava

Najprej smo se lotili iskanja primernih besedil za uvrščanje. Odločili smo se, da bomo za raziskavo uporabili članke iz dnevnega časopisja. Članke smo črpali iz internetnih strani časopisnih hiš in letne zgoščenke časopisa Večer [15], saj so le-ti laže dosegljivi. Izbrali smo internetne strani časopisov Delo [10], Večer [11], Dnevnik [12], Morel [14] in Primorski dnevnik [13]. Poiskali smo 270 člankov, ki smo jih razvrstili v tri kategorije. To so bili politika, gospodarstvo in šport. Skupna dolžina vseh člankov je znašala 1.027,3 kilobajtov, povprečna dolžina pa 3,8. Skupno je bilo v člankih 161.400, v povprečju pa 597,8 besed. V kategoriji šport

so bili v povprečju članki dolgi 474,0 besed oz. 2,9 kilobajtov, v kategoriji politika 796,3 besed oz. 5,1 kilobajtov in v kategoriji gospodarstvo 523,2 besed oz. 3,4 kilobajtov. Iz navedenega je razvidno, da so bili članki precej kratki (tabela 1).

	Povpr. dolžina (besed)	Povpr. dolžina (kb)
Šport	474,0	2,9
Politika	796,3	5,1
Gospodarstvo	523,2	3,4

Tabela 1: **Povprečne dolžine člankov**

	Število besed	σ
Šport	201,8	1,2
Politika	655,3	4,0
Gospodarstvo	250,7	1,6

Tabela 2: **Standardni odklon**

Za prvi dve kategoriji smo se odločili iz predpostavke, da je razpoznavanje tako podobnih področij, kot sta gospodarstvo in politika, še posebno zahtevno. Na drugi strani se je kategorija šport povsem razlikovala od prvih dveh omenjenih.

Članki in pripadajoče kategorije so na voljo na internetni strani http://lpt.uni-mb.si/public/ev/raz-slo-besedil.htm

Članke smo razvrstili v dve množici [9]:

- 2/3 člankov vsake kategorije je predstavljalo učno množico, na kateri se je orodje učilo,
- 1/3 člankov je predstavljalo testno množico člankov, ki smo jih razvrstili in z njimi preverili natančnost naučenega modela.

Razporejanje posameznih člankov v ti dve množici je bilo naključno.

Zaradi zahtev IBM-ovega orodja smo vse članke spremenili v tekstovne datoteke. Pri tem je bilo treba ustvariti še datoteko, ki je vsebovala seznam vseh člankov za posamezno kategorijo učne in testne množice.

S tem smo zaključili pripravo vhodnih podatkov.

4.2 Izvedba meritve natančnosti

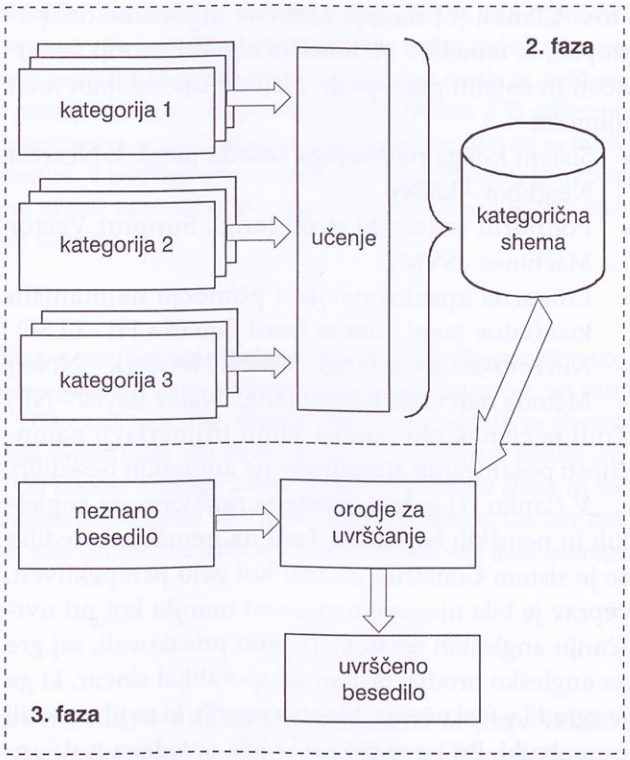
Meritve natančnosti uvrščanja besedil so bile izvedene v dveh korakih:

1. učenje na učni množici (izgradnja modela),
2. testiranje oz. uvrščanje besedil na testni množici.

S kazalčnimi datotekami učne množice smo izvedli učenje. Po opravljenem učenju na vseh treh učnih množicah – za vsako kategorijo ena učna množica – smo dobili kategorično shemo. Ta shema je dejansko slovar, ki hrani pomembne besedne statistike za vsako kategorijo. Te statistike so bile potem uporabljene za uvrščanje testnih primerov.

Podrobnosti principov delovanja algoritma, na katerem temelji učenje oz. izgradnja sheme, zaradi komercialne narave orodja ni poznana.

Omenjeni postopek prikazuje naslednja shema (slika 1):



Slika 1: **Prikaz postopka učenja in uvrščanja**

Po opravljenem učenju je sledila druga faza – testiranje oz. uvrščanje testne množice besedil.

Podana je bila kazalčna datoteka za testne primerke posamezne kategorije. Na podlagi kategorične sheme je orodje uvrstilo vse testne datoteke. Rezultat tega procesa je bila izhodna datoteka; primer takšne datoteke je prikazan na sliki 2. Vsebuje posamezni testni članek, za katerega so navedene točke ujemanja z določeno kategorijo. Članek pripada tisti kategoriji, pri kateri je dosegel največ točk.

Document Identifier: g02-d.TXT	
Category	*Score*
P1-trn	153.652
G1-trn	140.937
S1-trn	108.618

Slika 2: **Primer izhodne datoteke**

5 REZULTATI

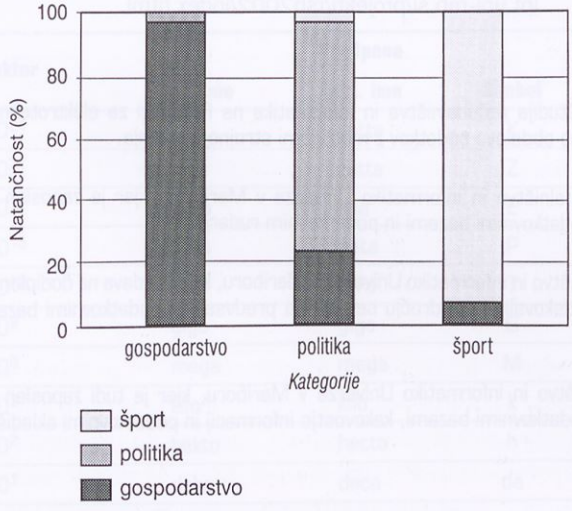
Iz tabele 3 je razvidno, da je testna množica za vsako kategorijo vsebovala 30 člankov. Iz kategorije gospodarstvo je orodje pravilno razvrstilo 29 člankov, enega pa je napačno razvrstilo v razred politika.

Uvrščen kot \ Dejanski	Gospodarstvo	Politika	Šport
Gospodarstvo	29	7	2
Politika	1	22	0
Šport	0	1	28

Tabela 3: **Razvrstitev člankov v kategorije**

Pri politiki je bila natančnost uvrščanja nižja, saj je bilo le 22 od 30 člankov pravilno uvrščenih. V drugi dve kategoriji, gospodarstvo in šport, je bilo uvrščenih skupno osem člankov s področja politike, in sicer sedem v kategorijo gospodarstvo ter eden v kategorijo šport.

Članki s športno tematiko so bili uvrščeni bolje kot politični članki. Od 30 testnih primerkov je bilo pravilno umeščenih 28, dva pa sta bila uvrščena kot gospodarstvo.



Slika 3: **Učinkovitost uvrščanja po temah**

Največje težave pri uvrščanju so se pojavile pri politiki, saj se politika meša z gospodarstvom, kot na primer v članku številka g47-d.txt [16].

Problem izvira iz dejstva, da je politika tematika, ki je v dnevnem časopisju nastopala na različnih področjih, med drugim tudi pri gospodarstvu. Velikokrat sta se gospodarstvo in politika prepletala. Kljub težavnemu ločevanju med tema dvema kategorijama se je orodje izkazalo kot uspešno – 73, 3 %.

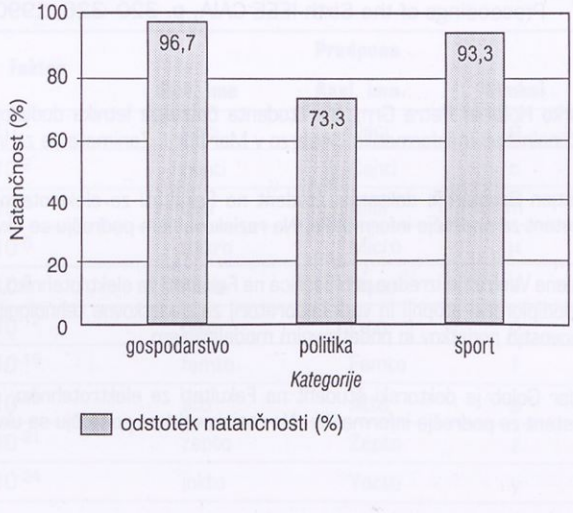
Osnova razpoznavanja je bila kategorična shema, ki temelji na že prej omenjenih besednih statistikah za vsako posamezno kategorijo. Problem se je pojavil, ker so lahko članki različnih kategorij vsebovali enake besede oz. besedne zveze. Tem bolj je bila tematika kategorij različna, tem manjša je bila verjetnost, da nastopajo v člankih različnih kategorij iste besede.

V nadaljevanju podajamo tabelo (tabela 4) in graf natančnosti (slika 4) v odstotkih za posamezno kategorijo.

Kategorija	**Odstotek natančnosti [%]**
Gospodarstvo	96,7
Politika	73,3
Šport	93,3
Skupaj	87.8

Tabela 4: **Natančnost uvrščanja**

Kot je razvidno iz grafa (slika 4), je natančnost uvrščanja dobra, v povprečju 87,8 %. Tako je bilo orodje uspešno v vseh kategorijah, ki so služile kot testne in učne množice za uvrščanje.



Slika 4: **Natančnosti uvrščanja**

6 SKLEP

Namen raziskave je bil ugotoviti natančnost uvrščanja slovenskih besedil. Takšna raziskava še ni bila izvedena na slovenskem govornem področju. Raziskovali smo s pomočjo ameriškega orodja, saj v Sloveniji trenutno ni na voljo ustrezne programske opreme. Kljub temu, da smo razpoznavanje izvedli s pomočjo tujega orodja, ki ne podpira slovenskega jezika, smo prišli do spoznanja, da lahko natančno uvrščamo slovenska besedila s pomočjo omenjenega orodja. Dano orodje uspešno ločuje besedila, ki so si po tematiki podobna. Če bi želeli uvrščati tekste, ki bi se po tematiki zelo razlikovali, bi dobili uspešnost nad 90 %, ki je primerljiva z natančnostjo uvrščanja angleških besedil [1], [2]. Že sedanja natančnost uvrščanja je boljša kot natančnost, ki jo je dobil Reuters [2].

Vendar omenjeno ne izključuje možnosti razvoja sorodnega slovenskega orodja.

V naši raziskavi smo ugotovili, da je v primeru zahteve po večji natančnosti, potrebno razviti posebno orodje.

V nadaljnjih raziskavah bomo proučili uspešnost orodja z večjim številom kategorij. Prav tako bomo za boljšo oceno natančnosti uvrščanja uporabili metodo navzkrižne validacije.

7 LITERATURA

[1] C. Apte, F. Damerau, S. M. Weiss,
Towards Language Independent Automated Learning of Text Categorization Models, In Proceedings of the ACM SIGIR Conference, 1994.

[2] P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt. TCS:
A Shell for Content-Based Text Categorization. In Proceedings of the Sixth IEEE CAIA, p. 320–326, 1990.

[3] T. Joachims,
Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In Proceedings of the European Conference on Machine Learning, Springer, 1998.

[4] Ji He, Ah-Hwee Tan, Chew-lim Tan,
A Comparative Study on Chinese Text Categorization Methods, PRICAI Workshop on Text and Web Mining, 2000.

[5] Ah-Hwee Tan, Fon-Lin Lai,
Text Categorization, Supervised Learning, and Domain Knowledge Integration, In proceedings, KDD'2000 International Workshop on Text Mining, Boston, pp. 113-114, August 2000.

[6] Yiming Yang, Xin Liu,
A re-examination of text categorization methods, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, p. 42–49), 1999.

[7] IBM Intelligent Miner for Text Version 2 Release 3 Text Analysis Tools, Documentation on CD, Third edition, December 1998.

[8] IBM Intelligent Miner for Text Version 2 Release 3, http://www-3.ibm.com/software/data/iminer/fortext/, zadnji obisk 15. 3. 2003.

[9] P. R. Cohen,
Empirical Methods for Artificial Intelligence, The MIT Press, Cambridge, USA, 1995.

[10] Delo, 2002, http://www.delo.si, zadnji obisk 15. 3. 2003.

[11] Večer, 2002, http://www.vecer.com, zadnji obisk 15. 3. 2003.

[12] Dnevnik, 2002, http://www.dnevnik.si, zadnji obisk 15. 3. 2003.

[13] Primorski dnevnik, 2002, http://www.primorski.it, zadnji obisk15. 3. 2003.

[14] Elektronski časopis Morel, 2002, http://www.morel.si/, zadnji obisk 15. 3. 2003.

[15] CD Večer, ČZP Večer, 2001.

[16] Naslov strani s testno in učno množico člankov, http://lpt.uni-mb.si/projekti/rsb2002/index.html.

Marko Hölbl in Petra Grm sta študenta četrtega letnika dodiplomskega študija računalništva in informatike na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Zanimata se za inteligentno obdelavo podatkov z metodami strojnega učenja.

Boštjan Brumen je doktorski študent na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer je zaposlen kot asistent za področje informatike. Na raziskovalnem področju se ukvarja s podatkovnimi bazami in podatkovnim rudarjenjem.

Tatjana Welzer je izredna profesorica na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer predava na dodiplomski in podiplomski stopnji in vodi Laboratorij za podatkovne tehnologije. Na raziskovalnem področju se ukvarja predvsem s podatkovnimi bazami, kakovostjo podatkov in podatkovnim modeliranjem.

Izidor Golob je doktorski študent na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, kjer je tudi zaposlen kot asistent za področje informatike. Na raziskovalnem področju se ukvarja s podatkovnimi bazami, kakovostjo informacij in podatkovnimi skladišči.