Primož Jakopin
Ljubljana

UDK 821.163.6.09 Kosmač C. :004.6
UDK 821.163.6.09 Cankar I. :004.6

NIZKOENTROPIJSKI JEZIKOVNI MODEL NA BESEDILIH CIRILA KOSMAČA IN IVANA CANKARJA

V prispevku je bil jezikovni model, ki temelji na pogostnostih znakovnih n-terčkov (nizov znakov, tj. črk, presledkov, števk in ločil dolžine n), uporabljen na besedilih zbranih del Cirila Kosmača in Ivana Cankarja. Pri vsakem modelu je najprej treba napraviti Huffmanovo drevo iz vseh n-terčkov (n = 1 do 20, pogostnost vsaj 2) posamezne besedilne zbirke (400.000 oz. 2 milijona besed, 45.889.000 oz. 223.553.000 n-terčkov, 26.274.000 oz. 116.588.000 različnih n-terčkov), in izračunati ustrezne Huffmanove kode za vsak list v obeh drevesih. Pri uporabi modela na danem besedilu pa to besedilo razrežemo na n-terčke (1–20) tako, da je vsota dolžin Huffmanovih kod modela na danem besedilu najmanjša. Če model uporabimo na besedilu, iz katerega smo ga napravili, dobimo tudi najmanjšo entropijo besedila, ki je obenem tudi mera za njegovo informacijsko vsebnost. Dobljena entropija besedil Cirila Kosmača glede na njegov model je bila **2,26** bita na znak, entropija besedil Ivana Cankarja z njegovim modelom pa **2,27** bita na znak.

kvantitativno jezikoslovje, entropija, teorija informacij, jezikovni model, statistični opis besedila, C. Kosmač, Ivan Cankar

In the paper a language model based on probabilities of character n-grams has been applied to texts of collected works of two leading Slovenian twentieth-century writers, i.e., Ciril Kosmač and Ivan Cankar. During the construction of each model a Huffman tree is generated from all the n-grams (n=1 to 20, frequency of 2 or more) of each text corpus (0.4 and 2.0 million running words, 45,889,000 and 223,553,000 n-grams, 26,274,000 and 116,588,000 different n-grams), and appropriate Huffman codes are computed for every leaf in the tree. To apply the model to an arbitrary text, the text is cut into n-grams (1–20) in such a way that the sum of the lengths of the model Huffman codes for all the obtained n-grams of the new text is minimal. If the model is applied to the text from which it was generated, the resulting entropy is minimal; this entropy is also a measure of the information content of the text from the standpoint of information theory. When the model of Ciril Kosmač was applied to his texts, the entropy of **2.26** bits per character was obtained and **2.27** bits per character for the model and texts of Ivan Cankar.

quantitative linguistics, entropy, information theory, linguistic model, statistical description of text, C. Kosmač, Ivan Cankar

1 Uvod

V zadnjem času so nastali večji besedilni korpusi, zbirke elektronskih besedil, ki so javno dostopne in opremljene s spletnim vmesnikom. Primer za slovenski jezik je

Nova beseda, spletni naslov http://bos.zrc-sazu.si na Inštitutu za slovenski jezik ZRC SAZU (Jakopin 2001a), ki trenutno obsega 100 milijonov besed in v okviru katerega je tudi že nekaj zbranih del vidnih slovenskih pisateljev, med drugim Cirila Kosmača in Ivana Cankarja. Poleg tega pa smo v zadnjem času, pač v skladu z rastjo računalniških zmogljivosti, priče preporoda uporabe entropije za različne, z besedili povezane namene (npr. Nigam idr. 1999).

V prispevku je na primeru besedil obeh avtorjev predstavljen jezikovni model, ki sloni na pogostnostih znakovnih nizov v besedilu in ki ga je mogoče poleg bolj konkretnih namenov s področja teorije informacij, kot je npr. ocena informacijske vsebnosti ustreznega besedila uporabiti tudi za ugotavljanje razdalj med različnimi besedilnimi zbirkami.

2 Entropija

Pojem se je najprej pojavil pri znanstvenem obravnavanju toplotnih (takrat predvsem parnih) strojev sredi 19. stoletja. O entropiji recimo govori drugi zakon termodinamike, označena pa je kot mera za razpoložljivost energije. Entropija v smislu teorije informacij, njene matematične temelje je prvi opisal Claude E. Shannon (1948), je količina, ki je v premem sorazmerju z informacijo. Več kot je o nekem sistemu znanega, manjša je njegova entropija in manj kot o njem vemo, večja je njegova entropija. Entropija vesolja se s časom povečuje, vse od velikega poka pred milijardami let. Vse, kar je bilo še včeraj enostavno in bližje, je danes že malo manj jasno in dlje, ne glede na nenehni in neustavljivi napredek našega védenja in znanosti.

Entropija kot mera nedoločenosti besedil (npr. Jakopin 2002) je bila včasih merjena s poskusi, ko so naključno izbranim bralcem pokazali veliko kratkih iztržkov iz besedil, do največ nekaj deset znakov, in so morali potem uganiti naslednjo črko, ki bi sledila koncu. Splošno znano je, da vsebuje besedilo manj informacije, kot bi jo naključno nametana zaporedja črk – če v besedilu izpustimo vsako drugo črko, ga še vedno lahko skoraj vedno pravilno prepoznamo. Spodaj so navedeni trije primeri nizov znakov, kjer je potrebno uganiti naslednjo črko, na začetku pikic:

1. *Ljublja*.....
2. *Ljubljan*.....
3. *Ljubljanski gra*.....

Entropija je v prvem primeru praktično enaka 0, saj nizu *Ljublja* skoraj vedno sledi črka *n*; podobno je v tretjem primeru, kjer za *Ljubljanski gra* navadno sledi *d*, čeprav bi bil v redkih primerih možen tudi *š* (*Ljubljanski graščaki*) ali *j* (*Ljubljanski grajski drevored*). Čisto drugače pa je v drugem primeru, kjer je možnih kandidatov za naslednjo črko dosti več, z znatnimi pričakovanimi pogostnostmi: *a*, *e*, *i*, *o*, *s*, *č*.

Znano je tudi, da je vsak znak ali simbol besedila (črka, številka, ločilo ...) v besedilnih računalniških datotekah navadno zapisan v enem bajtu, dolgem 8 bitov. Če bi bilo vseh 256 (2 na 8) možnih znakov v besedilih enako pogostih, bi bila njihova entropija največja, se pravi 8 bitov na znak. Temu seveda ni tako, črki *e* in *a* npr. sta veliko pogostejši od drugih, *e*-jev je skoraj 11 %, *a*-jev dobrih 10 %, *f*-jev pa stokrat manj, le 0,1 %. Še večje so razlike, če se od znakov preselimo k njihovim nizom – šesterček *se_je_* (presledek označen s podčrtajem) je npr. 20.000-krat pogostejši od bolj redkih šesterčkov, ki se pojavijo samo enkrat. Tako se je na 3 milijone besed dolgem vzorcu slovenskih besedil izkazalo (Jakopin 2002), da je njihova entropija 2,2 bita na znak, 3,6-krat manj, kot če bi bili znaki v besedilu naključno razmetani (8 bitov na znak).

3 Nizkoentropijski jezikovni model

Vzemimo, da bi želeli napraviti postopek, ki bi po žici ali brez nje digitalno prenašal besedila, zapisana v nekem jeziku, pa tako da bi bilo potrebno število bitov kar najmanjše. Vzemimo, da bi želeli poslati poved *Gori_na_gori_gori.* in jo zapisati kar najkrajše.

Tabela 1: Znaki iz povedi *Gori_na_gori_gori.* s frekvencami, z ASCII in s Shannon-Fanojevimi kodami

_	3	00100000	00
i	3	01101001	010
o	3	01101111	011
r	3	01110010	100
g	2	01100111	101
a	1	01100001	1100
G	1	01000111	1101
n	1	01101110	1110
.	1	00101110	1111

V tabeli 1 so v prvem stolpcu navedeni vsi simboli iz te povedi (presledek je označen s podčrtajem), v drugem njihove pogostnosti, v tretjem osembitne ASCII kode, s kakršnimi bi bili zapisani v računalniški datoteki s končnico TXT, v četrtem pa Shannon-Fanojeve kode teh simbolov. Te kode niso vse enako dolge, ampak imajo pogostejši simboli krajše kode, redkejši pa daljše. Dobimo jih tako, da najprej napravimo tabelo simbolov, padajoče urejeno po pogostnostih. Nato nadaljujemo po korakih, pri čemer na vsakem koraku seznam razdelimo v dve skupini, tako da sta vsoti pogostnosti simbolov v obeh skupinah kar najbolj enaki. Vsem simbolom zgornje skupine potem dodelimo binarno številko 0 kot prvo številko njihove kode, vsem simbolom spodnje skupine pa 1. Vsako skupino nato po istem načelu razdelimo naprej in njenim simbolom dodelimo dodatno binarno številko. Postopek

ponavljamo, dokler v vsaki skupini ne ostane samo po en simbol. Kot vidimo, bi bila poved, zapisana v ASCII kodi:

G	o	r	i	_	n	a	_	g
01000111	01101111	01110010	01101001	00100000	01101110	01100001	00100000	01100111

o	r	i	_	g	o	r	i	.
01101111	01110010	01101001	00100000	01100111	01101111	01110010	01101001	00101110

dolga 144 (18 krat 8) bitov, če jo zakodiramo s Shannon-Fanojevo metodo, ki da pri neskončnem številu simbolov v prenesenem sporočilu dokazano najkrajšo skupno kodo:

G	o	r	i	_	n	a	_	g	o	r	i	_	g	o	r	i	.
1101	011	100	010	00	1110	1100	00	101	011	100	010	00	101	011	100	010	1111

pa le 55 bitov, skoraj trikrat manj. Če bi zakodirano sporočilo res poslali z enega mesta na drugo, morata seveda oba, tako oddajnik kot sprejemnik, poznati kodni slovar – kode vseh simbolov, ki pridejo v poštev za prenos.

Pri nizkoentropijskem jezikovnem modelu (Jakopin 2001) postopek še posplošimo – namesto posameznih simbolov (črk, številk, ločil) v besedilu vzamemo vse možne nize simbolov (n-terčke) do določene dolžine, ki so se pojavili vsaj dvakrat, namesto Shannon-Fanojevega kodiranja pa Huffmanovo kodiranje, ki se ne razlikuje bistveno, da pa dokazano najkrajšo kodo že pri končnem številu simbolov v besedilu. Slovar vseh znakovnih n-terčkov z njihovimi (najkrajšimi) binarnimi kodami (kodni slovar) potem predstavlja nizkoentropijski model besedil nekega jezika. Kot primer navedimo najpogostejši šesterček v slovenskem jeziku, že omenjeni *se_je_*, ki ga na vrhu ne najdemo v nobenem drugem jeziku, tudi v nam najbližjih ne (v hrvaških besedilih najpogostejši šesterček *_da_je* je v slovenskih šele na 23. mestu, šesterčka *se_je_* pa pri njih med prvimi 1000 ne najdemo).

Tabela 2: Velikost zbranih del Cirila Kosmača (1910–1980) in Ivana Cankarja (1876–1918)

	Ciril Kosmač	Ivan Cankar
znakov	2.497.308	11.821.000
besed	407.938	1.991.000
povedi	37.459	157.000
vseh n-terčkov, n=1–20	45.889.000	223.553.000
različnih n-terčkov, n=1–20	26.274.000	116.588.000
različnih n-terčkov, n=1–20, frekvenca >=2	2.698.000	11.586.000

Kot vidimo v tabeli 2, kjer so navedeni osnovni podatki o obsegu opusov Cirila Kosmača in Ivana Cankarja, pa število n-terčkov zelo hitro narašča – pri Ivanu

Cankarju jih imamo že za *n* od 1 do 20 skoraj 224 milijonov, od tega skoraj 117 milijonov različnih. Količine, katerih obdelava tudi z računalniki, kakršne imamo na razpolago danes, zahteva zelo premišljen in dognan pristop. Kodna slovarja obeh nizkoentropijskih modelov, izpeljanih iz zbranih del Cirila Kosmača in Ivana Cankarja, sta seveda tudi ustrezno obsežna.

Tabela 3: Poved Cirila Kosmača, razrezana glede na model iz njegovih besedil

Tisti pomladni dan j | e bil lep | , svetel in zveneč, | kakor iz čistega sre | bra ulit.

Postopek izračuna entropije na znak za posamezno besedilo s posameznim modelom temelji najprej na razrezu besedila na zaporedje nizov, n-terčkov, katerih vsota dolžin Huffmanovih kod bo najmanjša. V tabeli 3 je naveden razrez znamenite povedi iz romana *Pomladni dan* Cirila Kosmača z modelom, izpeljanih iz njegovih besedil. Poved je razrezana samo 4 mestih.

Tabela 4: Poved Cirila Kosmača, razrezana glede na model iz besedil Ivana Cankarja

Tisti| pomladni dan| je bil lep, |svetel in |zveneč, |kakor iz |čistega s|rebr|a uli|t.

V tabeli 4 je ista poved, tokrat razrezana z modelom, dobljenim iz besedil Ivana Cankarja, ki v svojem opusu o svetlih in zvenečih, kakor iz čistega srebra ulitih dnevih ni veliko pisal. Delilnih mest je tokrat še enkrat več, 9. Oglejmo si še premislek pri ugotavljanju, kje deliti, da bo skupna vsota kod modela najmanjša.

Tabela 5: N-terčki modela, izpeljanega iz besedil Ivana Cankarja, z dolžinami Huffmanovih kod

T 9	i 7	_ 6	p 9	o 7	m 9
Ti 12	i_ 9	_p 9	po 10	om 11	ml 14
Tis 14	i_p 13	_po 11	pom 14	oml 18	mla 15
Tist 14	i_po 14	_pom 15	poml 18	omla 19	mlad 15
Tisti 15	i_pom 19	_poml 19	pomla 18	omlad 19	mladn 21
Tisti_ 16	i_poml 22	_pomla 19	pomlad 18	omladn 22	mladni 23
Tisti_p 20	i_pomla 23	_pomlad 19	pomladn 21	omladni 23	mladni_ 23
Tisti_po 21	i_pomlad 23	_pomladn 21	pomladni 23	omladni_ 23	mladni_d 24
	i_pomladn 26	_pomladni 23	pomladni_ 23	omladni_d 24	mladni_da 25
		pomladni 23	pomladni_d 24	omladni_da 25	mladni_dan 25
		_pomladni_d 24	pomladni_da 25	omladni_dan 25	mladni_dan_ 26
		_pomladni_da 25	pomladni_dan 25	omladni_dan_ 26	
		_pomladni_dan 25	pomladni_dan_26		
		_pomladni_dan_ 26			

Pri ugotavljanju prvega delilnega mesta si pomagamo s podatki iz tabele 5. Na začetku povedi je bil najdaljši ugotovljeni niz dolg 8 znakov (*Tisti_po*), katerega Huffmanov kod je dolg 21 bitov. Niz *Tisti_pom* se v prvem vzorcu ni pojavil 2-krat

ali več. Začasno delilno mesto je torej postavljeno za osmim znakom. Najdaljši niz, ki sledi, je *mladni_dan_* z dolžino Huffm. kode 26 bitov. Skupaj 47 bitov za 19 znakov ali 2,47 bita na znak. Algoritem zdaj zahteva preverjanje toliko znakov nazaj, da dobimo slabši rezultat. En znak nazaj dobimo niza *Tisti_p* in *omladni_dan_* (niz *omladni_dan_j* ni imel frekvence 2 ali več), ki imata skupno dolžino Huffmanovih kod 20 + 26, se pravi 46 bitov, spet za 19 znakov, ali 2,42 bita na znak. Dva znaka nazaj opazujemo par *Tisti_* in *pomladni_dan_* (niz *pomladni_dan_j* se očitno tudi ni pojavil v seznamu), ki imata skupno dolžino Huffm. kod 16 + 26 = 42 bitov ali 2,21 bita na znak. Tri znake nazaj dobimo par *Tisti* in *_pomladni_dan_*, ki imata skupno dolžino Huffmanovih kod 15 + 26 = 41 bitov za 19 znakov, kar je 2,16 bita na znak, kar je še vedno boljše kot prej. Še eno mesto nazaj dobimo par *Tist* in *i_pomladn*, ki imata skupno dolžino Huffmanovih kod 14 + 26 = 40 bitov za 13 znakov, kar je 3,08 bita na znak ali slabše kot prej. Prvo delilno mesto torej pade za *Tisti*, s prirastkom 15 bitov (3,00 bita na znak).

Tabela 6: Entropija modelov iz besedil Cirila Kosmača in Ivana Cankarja, uporabljenih na obeh besedilih, v bitih na znak

	Ciril Kosmač	Ivan Cankar
Ciril Kosmač	2,26	2,96
Ivan Cankar	2,83	2,27

Oglejmo si na koncu v tabeli 6 rezultate uporabe obeh modelov, tistega, ki je izpeljan iz besedil Cirila Kosmača, in tistega, ki sloni na besedilih Ivana Cankarja, uporabljenih na obeh opusih. Izkaže se, da z uporabo modela iz Kosmačevih besedil na teh besedilih dobimo entropijo 2,26 bita na znak, s Cankarjevim modelom na njegovih besedilih pa 2,27 bita na znak. Vrednosti sta si zelo blizu, obe razmeroma ugodni (visoki), malenkost v prid Ivanu Cankarju, še posebej ker bi zaradi skoraj petkrat večje besedilne zbirke pričakovali pri njem nekoliko nižjo entropijo. Z uporabo Kosmačevega modela na Cankarjevih besedilih dobimo entropijo 2,96 bita na znak, z uporabo Cankarjevega modela na Kosmačevih besedilih pa 2,83 bita na znak. Razliko spet lahko razložimo predvsem z velikostjo Cankarjevega besedila.

Tabela 7: Nizkoentropijski jezikovni model na besedilih Platonove *Države* v 16 jezikih

	Jezik	Prevod	Leto	Prenos v el. obliko	Besed	Znakov	b/z
-	slovenski	Jože Košar	1976	Primož Jakopin	92.741	565.604	2,37
1.	srbohrvaški	A. Vilhar, B. Pavlović	1983	Duško Vitas	107.506	613.082	3,77
2.	hrvaški	Damir Salopek	1976	Marko Tadić	92.870	532.497	3,84
3.	bolgarski	-	-	Patrice Bonhomme	112.676	678.131	3,96
4.	češki	Radislav Hošek	1993	František Čermák	110.466	636.201	4,10
5.	poljski	Wladyslaw Witwicki	1991	-	107.559	645.532	4,32
6.	ruski	-	-	-	99.503	649.102	4,46
7.	slovaški	Július Španár	1990	Alexandra Jarošová	99.661	622.463	4,46
8.	latvijski	Gustavs Lukstinš	1982	Andrejs Spektors	45.238	290.508	4,74
9.	litvanski	Jonas Dumčius	1981	-	85.144	584.318	4,94
10.	angleški	Paul Shorey	-	-	129.331	692.058	5,40
11.	francoski	-	1993	-	142.624	817.658	5,67
12.	nemški	Karl Vretska	1982	Joachim Hohwieler	104.876	641.333	5,69
13.	romunski	Andrei Cornea	1986	Dan Tufis	131.064	658.804	5,76
14.	finski	Marja Itkonen-Kaila	-	Anna Mauranen	75.800	582.522	6,11
15.	madžarski	Szabó Miklós	1984	Tamás Váradi	105.538	728.501	6,47

Razdaljo med jezikoma obeh avtorjev, kot je bila izmerjena z obema modeloma na besedilih drugega, nam pomagajo dodatno osvetliti vrednosti iz tabele 7 (Jakopin 2002: 111). V njej so navedeni rezultati uporabe modela na istem besedilu, prevodu *Platonove Države*, v 16 evropskih jezikih. Tabela je urejena padajoče po povprečnem številu bitov na znak, ki jih je pri kodiranju posameznega besedila dosegel model. V vsaki vrstici je najprej navedena zaporedna številka jezika, nato njegovo ime, prevajalec, leto izdaje dela, ki je služilo za prenos v elektronsko obliko (navadno se ujema z letom prevoda), vodjo prenosa v elektronsko obliko (le prvega, kadar jih je bilo več), število besed v delu, število znakov in na koncu še entropija, povprečno število bitov na znak, ki jih je potreboval model. Entropija modela, dobljenega iz treh milijonov besed slovenskega leposlovja, na slovenskem besedilu *Platonove Države*, znaša 2,37 bita na znak, na besedilu istega dela v našemu najbližjem jeziku pa 3,77 bita na znak.

4 Zaključek

V prispevku je prikazan nizkoentropijski jezikovni model, kvantitativno orodje, s katerim je mogoče s stališča informacijske vsebnosti ovrednotiti dano besedilo, pa tudi meriti oddaljenost med posameznimi besedili oz. jeziki. Zbrana dela dveh vidnih slovenskih pisateljev 20. stoletja so bila podlaga za konstruiranje dveh modelov, ki sta pokazala, da so vrednosti z obema modeloma dobljenih entropij na znak zelo blizu, da pa sta oba opusa, če si vsakega ogledamo z modelom drugega avtorja, jezikovno vseeno jasno razmejena.

Literatura

Primož Jakopin, 2002: *Entropija v slovenskih leposlovnih besedilih*. Ljubljana: ZRC SAZU.

– – 2001a: Distance beetween languages as measured by the minimal-entropy model; Plato's republic – Slovenian versus 15 other translations. *International journal of corpus linguistics* 6, special issue. [43]–53.

– –2001b: Beseda: a Slovenian text corpus. *Digital Evidence: selected papers from DRH2000, Digital Resources for the Humanities Conference, University of Sheffield, September 2000*. Ur. M. Fraser, N. Williamson, M. Deegan. London: Office for Humanities Communication. 229–241.

Kamal Nigam, John Lafferty, Andrew McCallum, 1999: Using Maximum Entropy for Text Classication. *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*. 61–67.

Claude Eugene Shannon, 1948: A Mathematical Theory of Communication. *Bell System Technical Journal* 27. 379–423, 623–656.