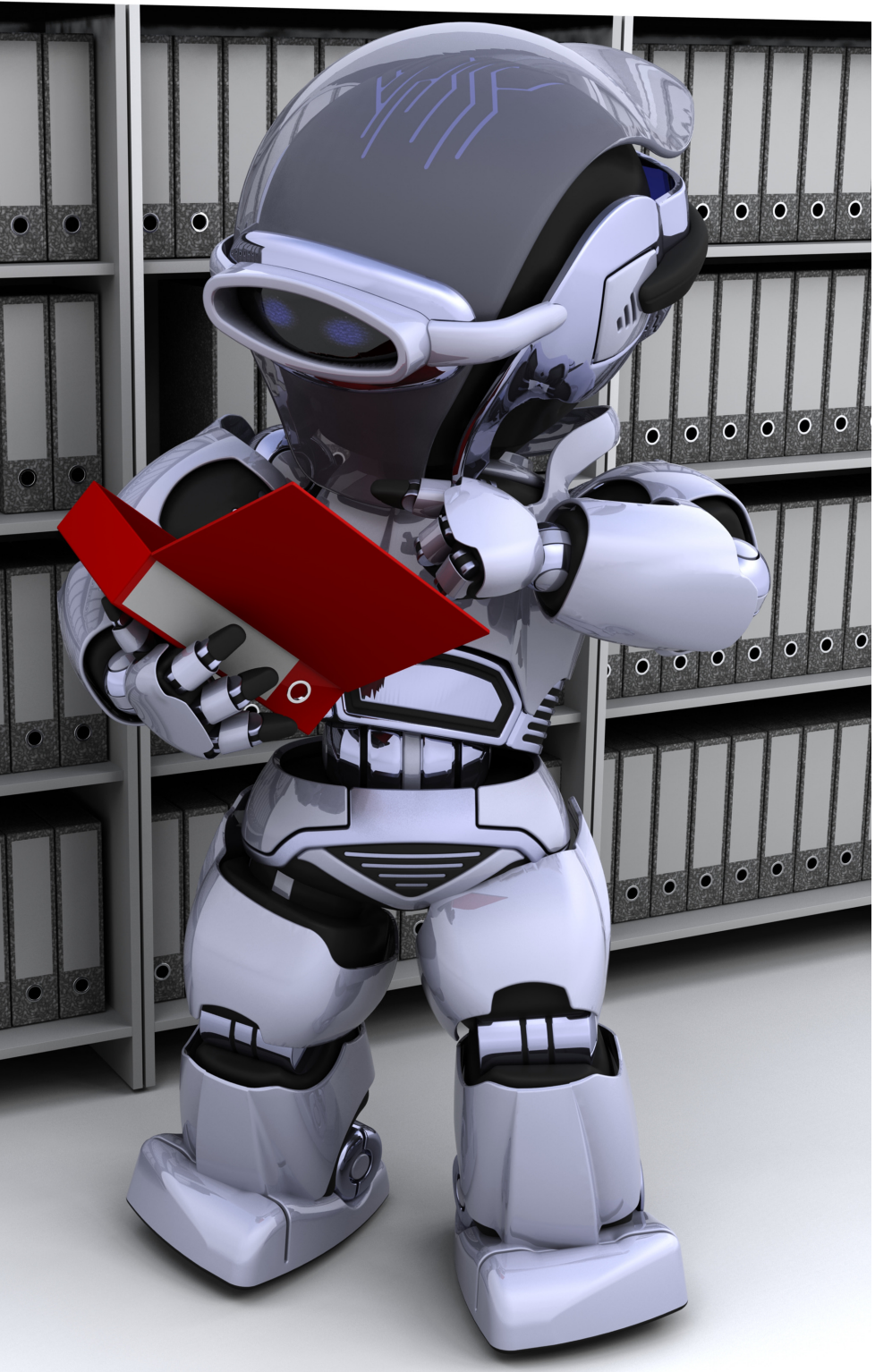

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

POSTAVLJANJE VEJIC V SLOVENŠČINI S POMOČJO STROJNEGA UČENJA IN IZBOLJŠANEGA KORPUSA ŠOLAR

ANJA KRAJNC
MARKO ROBNIK-ŠIKONJA

Pregled vsebine

- Zakaj vejice? Zakaj strojno učenje?
- Opis podatkovne množice
- Opis sprememb podatkovne množice
- Testiranje
- Rezultati
- Zaključek

Zakaj vejice?

- odsevajo verodostojnost in strokovnost besedila
- ločijo stavke znotraj povedi
- različno postavljene vejice spremenijo pomen stavkov
- nakazujejo premor v govoru
- omogočajo enolično razumevanje stavkov
- napačno postavljene vejice zelo pogosta napaka piscev v slovenščini

Zakaj strojno učenje?

- **slovenščina ima zahtevno oblikoslovno podobo**, zato njena obdelava zahteva veliko napora
- **pravila za pisanje vejic v slovenščini so zahtevna za razumevanje,** njihova programska implementacija težko uresničljiva
- **strojno postavljanje vejic je del zahtevnejših jezikovnih tehnologij** (npr. računalniška prepoznava in obdelava govora), katerih cilj je vzdrževanje politike večjezičnosti

Opis podatkovne množice

- za analizo uporabimo korpus **Šolar**, ki je bil uporabljen v že obstoječi raziskavi (Holozan, 2012; Holozan, 2013)
- izhajamo iz te raziskave in uporabimo izboljšano in posodobljeno verzijo uporabljenega korpusa, ki jo je sestavil in nam jo posredoval Peter Holozan (2015)
- oblikoslovno označen in skladenjsko razčlenjen **korpus Šolar** - zbirka besedil, ki so jih napisali učenci in dijaki, skupaj z učiteljskimi popravki
- posodobljen in izboljšan **korpus Šolar2**

Opis podatkovne množice

- vsaka beseda z okoliškim oknom (5 besed spredaj in 5 besed zadaj), ki se pojavi v korpusu, pretvorjena v **seznam atributov**
- **dodan razred**, ki pove, ali besedi sledi vejica
- **67 atributov** za vsako besedo

Opis trenutne besede

STAVEK:

*"...tem mislim na bistvo te **zgodbe**, ki pa je to, da je..."*

zgodbe,zgodba,Sozer,0,0,0,
te,ta,Zk-zer,0,0,0,
bistvo,bistvo,Soset,0,0,0,
na,na,Dt,0,1,0,
mislim,misliti,Ggnspe,0,0,0,
tem,ta,Zk-seo,0,0,0,
ki,ki,Vd,1,0,1,
pa,pa,Vp,1,0,0,
je,biti,Gp-ste-n,0,0,0,
to,ta,Zk-sei,0,1,0,
da,da,Vd,0,0,1,
je-vejica

Spremembe podatkovne množice

- **dodani novi atributi**: 41 generiranih na podlagi pravil, ki jih za postavljanje vejic uporablja LanguageTool in nekaj dodatnih, ki povzročajo težave
- **izboljšave pravil pri generiranju atributov**: dodatni pogoj za členek "da" in večbesedne veznike
- **odstranjeni neinformativni atributi**: besede in leme (osnovne oblike besed)
- **preoblikovani atributi za zapis MSD kod** na dva načina

Opis trenutne besede

brez oblik in lem

zgodbe,**zgodba**,Sozer,0,0,0,
te,**ta**,Zk-zer,0,0,0,
bistvo,**bistvo**,Soset,0,0,0,
na,**na**,Dt,0,1,0,
mislim,**misliti**,Ggnspe,0,0,0,
tem,**ta**,Zk-seo,0,0,0,
ki,**ki**,Vd,1,0,1,
pa,**pa**,Vp,1,0,0,
je,**biti**,Gp-ste-n,0,0,0,
to,**ta**,Zk-sei,0,1,0,
da,**da**,Vd,0,0,1,
je-vejica

Sozer,0,0,0,
Zk-zer,0,0,0,
Soset,0,0,0,
Dt,0,1,0,
Ggnspe,0,0,0,
Zk-seo,0,0,0,
Vd,1,0,1,
Vp,1,0,0,
Gp-ste-n,0,0,0,
Zk-sei,0,1,0,
Vd,0,0,1,
je-vejica

Primer implementacije pravila za veznik 'ker'

Po pravilih, ki jih za postavljanje vejic uporablja orodje **LanguageTool**:
kadar naletimo na besedo **ker** in beseda pred njo ni eno izmed ločil ,(;-: ali ena izmed besed *in*, *ali*, *ter*, *a* in *temveč*, potem trenutni besedi verjetno sledi vejica

atribut za trenutni veznik zavzame vrednost 1

Opis trenutne besede po spremembah

pri opisu MSD kode z 11 atributi

S,0,0,0,
Z,0,0,0,
S,0,0,0,
D,0,1,0,
G,0,0,0,
Z,0,0,0,
Vd,1,0,1,
Vp,1,0,0,
G,0,0,0,
Z,0,1,0,
Vd,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,
je-vejica

trenutna beseda ustreza pravilom za vezniško besedo "**ki**"

Opis trenutne besede po spremembah

pri opisu MSD kode z 99 atributi

S,o,z,e,r,*,*,*,*,0,0,0,
Z,k,-,z,e,r,*,*,*,0,0,0,
S,o,s,e,t,*,*,*,*,0,0,0,
D,t,*,*,*,*,*,*,*,0,1,0,
G,g,n,s,p,e,*,*,*,0,0,0,
Z,k,-,s,e,o,*,*,*,0,0,0,
V,d,*,*,*,*,*,*,*,1,0,1,
V,p,*,*,*,*,*,*,*,1,0,0,
G,p,-,s,t,e,-,n,*,0,0,0,
Z,k,-,s,e,i,*,*,*,0,1,0,
V,d,*,*,*,*,*,*,*,0,0,1,

trenutna beseda ustreza pravilom za vezniško besedo "**ki**"

0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,
je-vejica

Testiranje

- **testiranje opravimo z različnimi algoritmi**: naivni Bayesov klasifikator, RBF mreža, alternirajoče odločitveno drevo, AdaBoostM1, odločitvena tabela, metoda podpornih vektorjev in naključni gozdovi
- **implementiramo učenje**, s katerim želimo združiti prečno preverjanje in podvzorčenje
- **z mero ReliefF** ocenimo atribute in izberemo podmnožico atributov

Podatkovne množice

- Šolar1 - osnovni
- Šolar1 - MSD 11
- Šolar1 - MSD 11 - uravnoteženo
- Šolar1 - MSD 99
- Šolar1 - MSD 99 - uravnoteženo

- **Šolar2 - MSD 11**
- **Šolar2 - MSD 11 - uravnoteženo**
- **Šolar2 - MSD 11 - uravnoteženo z obdržanim razmerjem**
- **Šolar2 - MSD 99**
- **Šolar2 - MSD 99 - uravnoteženo**
- **Šolar2 - MSD 99 - uravnoteženo z obdržanim razmerjem**

Rezultati

		Ni vejice			Je vejica			Točnost [%]	AUC
		Natančnost	Priklic	F1	Natančnost	Priklic	F1		
Šolar2-99	NaiveBayes	0,979	0,840	0,904	0,333	0,813	0,473	83,746	0,905
	RBFNetwork	0,927	0,985	0,955	0,591	0,217	0,318	91,634	0,868
	ADTree	0,951	0,995	0,972	0,898	0,482	0,638	94,866	0,925
	AdaBoostM1	0,951	0,970	0,961	0,620	0,489	0,547	92,735	0,882
	RandomForest	0,956	0,996	0,976	0,925	0,536	0,579	95,455	0,973
Šolar2-11	NaiveBayes	0,983	0,769	0,863	0,269	0,861	0,410	77,754	0,903
	RBFNetwork	0,922	0,992	0,956	0,654	0,147	0,240	91,655	0,870
	ADTree	0,946	0,996	0,971	0,916	0,426	0,581	94,502	0,913
	AdaBoostM1	0,952	0,970	0,961	0,621	0,499	0,553	92,775	0,867
	RandomForest	0,957	0,995	0,975	0,913	0,542	0,680	95,428	0,943
1/10	DecisionTable	0,958	0,995	0,976	0,918	0,653	0,698	95,589	0,939
Šolar2-99	SMO-linearno	0,955	0,996	0,975	0,928	0,529	0,674	95,366	0,762
	SMO-kvadratno	0,927	1,000	0,962	0,996	0,210	0,347	92,848	0,605
1/10	DecisionTable	0,959	0,995	0,977	0,920	0,577	0,709	95,720	0,940
Šolar2-11	SMO-linearno	0,954	0,997	0,975	0,942	0,520	0,670	95,367	0,758
	SMO-kvadratno	0,943	0,992	0,967	0,834	0,392	0,534	93,799	0,692

Zaključek

- **osnova je dober korpus**: kvaliteten, po možnosti homogen korpus, sestavljen iz dobrih in večkrat lektoriranih besedil s strani strokovnjakov za jezik
- izjemno **pomembne so jezikovne tehnologije**, kot so lematizator, označevalnik in skladenjski razčlenjevalnik
- **bolje definirana pravila, ki bi bila enostavna za implementacijo:** dodali bi lahko še več (idealno vse!) atributov, generiranih na podlagi teh pravil
- **preizkusiti tudi druge ideje za opis atributov z informacijo o MSD oznaki:** opis MSD oznake s po 38 atributi