Razpoznavanje tekočega govora v slovenščini z bazo predavanj SI TEDx-UM

Andrej Žgank, Darinka Verdonik, Mirjam Sepesy Maučec

Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko

Laboratorij za digitalno procesiranje signalov

Uvod

- Avtomatsko razpoznavanje govora se vse pogosteje uporablja tudi izven laboratorijev – vendar ne enako uspešno za vse jezike.
- Predstavili bomo novi slovenski govorni vir SI TEDx-UM, nastal na osnovi posnetkov predavanj TEDx.
- Predavanja TEDx: aktualne tematike s področja tehnologije, izobraževanja, umetnosti in družbe. V SLO širše prisotna zadnjih 5 let.
- Področja uporabe vira: avtomatsko razpoznavanje govora, segmentacija in indeksiranje, odstranjevanje šuma, raziskave s področja diskurza.

Uvod

- Razlika v primerjavi z drugimi SLO govornimi bazami: uporabimo prosto dostopen govorni material, transkribiranje izvedemo v celoti avtomatsko.

- Takšen pristop izdelave govornih virov že v uporabi v tujini: TED-LIUM baza za angleški jezik v skupnem obsegu več kot 300 ur govora.

- Govorna baza SI TEDx-UM:

  – učni korpus: večina obsega, avtomatska segmentacija in transkribiranje z razpoznavalnikom govora za slovenski jezik.
  – testni korpus: transkripcije izdelane ročno, namenjen vrednotenju.

Baza SI TEDx-UM: zajem

- Posnetke predavanj TEDx smo zajeli s spletne strani YouTube.
- Na voljo je bilo več kot 300 predavanj v slovenskem jeziku.
- Posnetki so na voljo v različnih izgubnih kodekih (tipično za zvok: MPEG AAC, video: H.264). Vedno zajemali tistega z najvišjo kakovostjo zvoka.
- Zajete govorne posnetke smo pretvorili v format WAV s frekvenco vzorčenja 16 kHz s 16-bitno ločljivostjo. Format je tako skladen z večino ostalih SLO govornih virov.
- Video je služil zgolj kot pomoč pri transkribiranju.

Baza SI TEDx-UM: transkribiranje

- Transkribiranje je potekalo na osnovi rezultatov analize primerjave transkribiranja med bazo BNSI Broadcast News in korpusom GOS.
- Razlike v transkribiranju v primerjavi z BNSI Broadcast News: segmentiranje na zaključene izjave, način zapisa govora v dvotirni obliki (pogovorna in standardizirana).
- Razlike v transkribiranju v primerjavi s korpusom GOS: natančno označevanje akustičnega ozadja in akustičnih dogodkov.
- Uporabili smo orodje Transcriber AG. Precej težav pri njegovi uporabi.

Baza SI TEDx-UM: pregled

- V govorno bazo smo ročno izbrali 242 predavanja, ki ustrezajo kriterijem za učenje avtomatskih razpoznavalnikov govora.
- Glavni vzroki za izločanje predavanj: prekrivanje več govorcev, glasba kot akustično ozadje, slaba kakovost posnetkov,…
- Razmerjem med govorci: 66% moških, 34% žensk.
- Skupna dolžina: 54 ur, obdobje 6 let. Testni nabor vsebuje 13 predavanj v obsegu 3 ur.
- Transkripcije: 372k pojavnic, 32k različnih.

Avtomatsko transkribiranje

- Prvi korak: akustična segmentacija govor/tišina z uporabo Gaussovih modelov (GMM).
- Cilj: pridobiti akustično homogene segmente ustrezne dolžine za avtomatsko razpoznavanje govora.
- Drugi korak: avtomatsko transkribiranje z razpoznavalnikom govora UMB Broadcast News.
- Sistem ASR naučen na posnetkih dnevno-informativnih oddaj. Učni korpus dolžine 59 ur. Arhitektura prikritih modelov Markova (HMM).

Avtomatsko transkribiranje

UMB BN ASR	
Izloč. značilk	MFCC z normalizacijo
Karakteristike značilk	Okno 25 ms, korak 10 ms, MFCC 12 koef., energija, 1. in 2. odvod, 26 filtrov, normalizacija kepstra in energije
Akustični model	Medbesedni (Odell, 1995) trigrafemi
Kompleksnost AM	utežena vsota 16 Gaussovih porazdelitev verjetnosti na stanje
Združevanje AM	odločitveno drevo na osnovi grafemskih razredov (Žgank et al., 2005/2)
Jezik. modeli	Interpolirani trigrami
Vel. slovarja	64.000 besed

- Vrednotenje sistema UMB BN ASR na bazi BNSI Broadcast News (vsi f-razredi): napaka razpoznavanja besed (NRB) je 26,70%.

- Avtomatsko smo transkribirali vseh 242 predavanj v bazi SI-TEDx UM.

Jezikovna modela

- Za izdelavo avtomatskih transkripcij smo uporabili dva različna jezikovna modela. S tem smo delno vplivali na uporabljeno domeno.
- Gradnja jezikovnih modelov: 3-gramski statistični jezikovni model, Good-Turingovo glajenje in sestopanje po Katzu.
- Jezikovni model 1 - UMB Brodcast News: interpoliran model: FidaPLUS, BNSI-Speech, BNSI-Text, Večer.
- Jezikovni model 2 – FidaPLUS.
- Slovar obsega 64k besed, prilagojen domeni baze BNSI.

Jezikovna modela

Predavanje	Tematika	JM1 PP	JM2 PP	OOV
1	potovanja	409	431	21%
2	tehnologija	390	412	23%
3	družba	440	475	22%
4	tehnologija	379	405	28%
5	umetnost	481	506	26%
6	družba	491	491	26%
7	znanost	323	336	22%
8	znanost	242	234	20%
9	družba	429	451	27%
10	umetnost	400	399	24%
11	družba	428	451	19%
12	znanost	402	412	24%
13	družba	287	260	23%
vsa	različna	390	403	24%
BNSI eval	različna	247	387	4%

Tabela 1: Rezultati jezikovnih modelov UMB Broadcast News in FidaPLUS na testnih vzorcih SI TEDx-UM.

Rezultati

- Vrednotenje kakovosti avtomatsko tvorjenih transkripcij smo izvedli na testnem naboru 13 predavanj, za katere smo imeli na voljo tudi ročne transkripcije.

- Rezultati avtomatskega razpoznavanja govora so podani v obliki napake razpoznanih besed.

Rezultati

Predavanje	JM1 NRB(%)	JM2 NRB(%)
1	50,5	51,3
2	54,7	56,6
3	57,7	58,5
4	39,2	38,5
5	67,1	67,6
6	46,1	45,3
7	52,9	53,3
8	35,5	34,9
9	51,4	52,9
10	35,0	35,5
11	52,4	51,0
12	70,3	69,3
13	38,9	35,1
vsa	50,7	50,7
BNSI eval	26,6	26,6

Tabela 2: Napaka razpoznavanja govora na testnih vzorcih baze SI TEDx-UM za oba jezikovna modela.

Zaključek

- Govorna baza SI TEDx-UM predstavlja pomemben korak v načinu zbiranja materiala za govorne vire.

- Prvi rezultati kažejo na velik vpliv tematike predavanj na uspešnost razpoznavanja govora. Potrebne bodo ustrezne prilagoditve sistema za razpoznavanje govora.

- Govorna baza SI TEDx-UM je v skladu z licenco Creative Commons 3.0 prosto dostopna na spletni strani Inštituta za elektroniko in telekomunikacije UM FERI: http://ietk.feri.um.si/en/portfolio/sitedxumenglish