LESMEG.md

Grafem-til-fonem-modeller for norsk

Versjon

20200601: 1.0

Innledning

Denne ressursen innegolder *grafem-til-fonem*-modeller (GtP-modeller) for norsk som er tilpassa GtP-systemet Phonetisaurus. GtP-modellene brukes til å generere uttaleleksika fra ordlister. Mer informasjon om hvordan man gjør det, finner man på sidene til Phonetisaurus.

Modellene er trent på the norske uttaleleksikonet for talegjenkjenning som ble utvikla av det forenværende selskapet Nordisk språkteknologi (NST). Dette leksikonet distribueres nå av Nasjonalbiblioteket.

Vi har utvikla to modeller. Den første er trent på en full versjon av leksikonet, og inkluderer segmenter, markering av primær- og sekundærtrykk og tonemer. Den andre er trent på en forenkla versjon av leksikonet der toner og sekundærtrykk ikke er fjerna.

Innhold

- *train/*: inneholder modellene, samt andre filer Phonetisaurus bruker
  - *model-wtone-nob.fst* er trent på fullversjonen av leksikonet
  - *model-notone-nob.fst* er trent på en versjon av leksikonet uten toner og sekundærtrykk
- *lexica/*: inneholder leksika for trening og testing
  - *NST-total_train.dict* er treningssettet for *model-wtone-nob.fst*. Det inneholder 612 366 ortografi-transkripsjon-par (OTP), og utgjør 90% av alle unike OTP-er i NST-leksikonet
  - *NST-total_test.dict* er testsettet for *model-wtone-nob.fst*. Det inneholder de gjenstående 10% av OTP-ene i NST-leksikonet, og er tilfeldig plukka ut
  - *NST-total-notone_nosecstress_train.dict* er treningssettet for *model-notone-nob.fst*. Det er likt *NST-total_train.dict*, men markeringa av toner og sekundærtrykk er fjerna
  - *NST-total-notone_nosecstress_test.dict* er testsettet for *model-notone-nob.fst*. Det er likt *NST-total_test.dict*, men markeringa av toner og sekundærtrykk er fjerna
  - *NST-total_test_predicted.dict* er testsettet med transkripsjoner predikert av Phonetisaurus. Transkripsjonene har toner og sekundærstress
  - *NST-total_test_notone_predicted.dict* er testsettet med transkripsjoner predikert av Phonetisaurus. Transkripsjonene har ikke toner og sekundærstress *g2p_stats.py* er evalueringsskriptet som er brukt i evalueringa under.

Transkripsjonsstandard

Selv om den originale versjonen av NST-leksikonet bruker transkripsjonsstandarden X-SAMPA, har vi valgt å bruke en annen, ekvivalent standard som er lettere å lese og skrive for mennesker, *NoFAbet*. NoFAbet er til dels basert på ARPAbet og er laga av Nate Young på oppdrag for Nasjonalbiblioteket i forbindelse med utviklinga av *NoFA*, en norsk modell for forced alignment, som snart vil være tilgjengelig i Språkbankens ressurskatalog.

X-SAMPA-NoFAbet-ekvivalenstabell

X-SAMPA	NoFAbet	Eksempel
A:	AA0	b**a**d
{:	AE0	v**æ**r
{	AEH0	v**æ**rt

X-SAMPA	NoFAbet	Eksempel
{*I	AEJ0	s**ei**
E*u0	AEW0	s**au**
A	AH0	h**a**tt
A*I	AJ0	k**ai**
@	AX0	b**e**hage
b	B	**b**il
d	D	**d**ag
e:	EE0	l**e**k
E	EH0	p**e**nn
f	F	**f**in
g	G	**g**ul
h	H	**h**es
I	IH0	s**i**tt
i:	II0	v**i**n
j	J	**j**a
k	K	**k**ost
C	KJ	**k**ino
l	L	**l**and
l=	LX0	
m	M	**m**an
m=	MX0	
n	N	**n**ord
N	NG	e**ng**
n=	NX0	
o:	OA0	r**å**
O	OAH0	g**å**tt
2:	OE0	l**ø**k
9	OEH0	h**ø**st
9*Y	OEJ0	k**øy**e
U	OH0	f*o**r**t
O*Y	OJ0	konv**oy**
u:	OO0	b**o**d
@U	OU0	sh**ow**
p	P	**p**il
r	R	**r**ose
d`	RD	reko**rd**
l`	RL	pe**rl**e

X-SAMPA	NoFAbet	Eksempel
l`=	RLX0	
n`	RN	ba**rn**
n`=	RNX0	
s`	SJ	pe**rs**
t`	RT	sto**rt**
r=	RX0	
s	S	**s**il
S	SJ	**sj**u
s=	SX0	
t	T	**t**id
u0	UH0	r**u**ss
u0 j	UH0_J	Anh**ui**
}:	UU0	h**u**s
v	V	**v**ase
w	W	**W**ashington
Y	YH0	n**y**tt
y:	YY0	n**y**

Trykklette stavelser er merka med 0 etter stavelseskjernen. Stavelseskjernen er merka med *1* ved tonem 1 og *2* ved tonem 2. Sekundærstress merkes med *3*. I materialet uten toner og sekundærstress er alle 3-ere bytta ut med 0 og alle 2-ere med 1.

For å være kompatibel med NoFA er retrofleks *s* skrevet som *SJ*, ikke *RS*, noe som innebærer at det ikke er noen kontrast mellom retrofleks og postalveolar *s*-lyd i dette materialet.

Evaluering

Modell	Word Error Rate	Phoneme Error Rate
model-wtone-nob.fst	14.29	2.76
model-notone-nob.fst	10.44	2.00

Utregninga av Phoneme Error Rate er henta herfra.

Bruk

Modellene utvikla i dette prosjektet er offentlig eiendom med lisensen CC0. De kan brukes til et hvilket som helst formål. Se for øvrig lisensen til Phonetisaurus.