Grafem-til-fonem-modeller for norsk bokmål

Versjon: 2.0 (2024-02-09)

Dette repoet inneholder G2P-modeller for norsk bokmål, som produserer fonemiske transkripsjoner for talenære uttalevarianter (som i vanlige samtaler: `spoken`) eller skriftnære uttalevarianter (som i tekstopplesing `written`) for 5 forskjellige dialektområder:

1. Østnorsk (`e`)
2. Sørvestnorsk (`sw`)
3. Vestnorsk (`w`)
4. Trøndersk (`t`)
5. Nordnorsk (`n`)

Oppsett

Følg installasjonsinstruksjoner fra Phonetisaurus. Du trenger kun å kjøre koden i disse avsnittene:

- "Next grab and install OpenFst-1.7.2"
- "Checkout the latest Phonetisaurus from master and compile without bindings"

Data

Uttaleleksikonene som ble brukt til å trene G2P-modellene er fritt tilgjengelige fra Språkbanken ressurskatalog: NB Uttale. De kan også genereres opp med data og kode fra Github-repoet Sprakbanken/nb_uttale.

Innhold

- `models/`: inneholder modellene, samt andre filer Phonetisaurus bruker
  - `nb_*.fst`: Modellene som kan kjøres med `phonetisaurus-apply`. Utfyllingen til `*` er en streng med dialekt og uttalevariant, f.eks. `e_spoken` eller `t_written`.
  - `nb_*.o8.arpa`: 8-gram-modeller for fonemsekvenser som Phonetisaurus bruker i treningsprosessen.
  - `nb_*.corpus`: Sammenstilte (eng. *aligned*) grafemer og fonemer fra leksikonene.
- `data/`: inneholder leksika for trening og testing, samt utdata fra modellene på testsettet.
  - `NB-uttale_*_train.dict` er treningssettene for `models/nb_*.fst`. De inneholder 543 495 ortografi-transkripsjon-par (OTP), og utgjør 80% av alle unike OTP-er i leksikonet.
  - `NB-uttale_*_test.dict` er testsettene for `models/nb_*.fst`. De inneholder de gjenstående 20% av OTP-ene i leksikonet, altså 135 787 OTP-er.
  - `predicted_nb_*.dict` er testsettet med transkripsjoner predikert av modellene.
  - `wordlist_test.txt` er ordlisten til testsettet, som modellen kjøres på og genererer transkripsjonsforslag for.
- `evaluate.py`: evalueringsskript som er implementert på nytt for å evaluere disse modellene.
- `g2p_stats.py`: evalueringsskript fra V1.0, som kan brukes for å sammenligne disse modellene med NST-modellene (med og uten markering av trykk og tone) i versjon 1.
- `LICENSE`: Lisensteksten for CC0, som denne ressursen distribueres med.

Bruk

```
phonetisaurus-apply --model models/nb_e_spoken.fst --word_list
data/wordlist_test.txt  -n 1  -v  > output.txt
```

- Inndata (`--word_list`) burde være en tekstfil med linjeseparerte ord. Se for eksempel `data/wordlist_test.txt`.
- De trente modellene er `.fst`-filer i `models`-mappen. I samme mappe ligger også `.corpus`-filer og 8-gram-modeller for fonemsekvenser (`.arpa`-filer) som kommer fra treningsprosessen til Phonetisauraus.
- `-n`-argumentet angir hvor mange av de mest sannsynlige transkripsjonsforslagene som skal skrives til utdata.

Evaluering

Det er to skript i repoet for å regne ut Word Error Rate (WER) og Phoneme Error Rate (PER), og som gir litt ulike svar pga. implementasjonen av utregningene.

`evaluate.py`

Regner statistikk for alle modellene dersom ingen argumenter er gitt.

```
python evaluate.py
```

Med `-l`-argumentet kan du få statistikk for spesifikke modeller, f.eks. `-l e_spoken`.

- **WER** er antall feil transkripsjoner fordelt på totalt antall ord i testdata. 1 feil = 1 ord.
- **PER** er antall feil fonemer for alle transkripsjonene, fordelt på totalt antall fonemer i testdata. 1 feil = 1 fonem.

Model	Word Error Rate	Phoneme Error Rate
nb_e_written.fst	13.661238654564869	1.9681178920293207
nb_e_spoken.fst	13.72501038144391	1.9832518286152074
nb_sw_written.fst	13.240048644480037	1.8396612197218096
nb_sw_spoken.fst	16.422702734768936	2.426312206336983
nb_w_written.fst	13.240048644480037	1.8396612197218096
nb_w_spoken.fst	16.892833837574894	2.5064155890730686
nb_t_written.fst	13.736133357062347	1.98774986044724
nb_t_spoken.fst	16.47992288013051	2.5809178688066843
nb_n_written.fst	13.736133357062347	1.98774986044724

Model	Word Error Rate	Phoneme Error Rate
nb_n_spoken.fst	17.22590930999963	2.8209779677747715

`g2p_stats.py`

Regner ut WER og PER for to inputfiler:

1. Referansefila /testdata, f.eks. `data/NB-uttale_e_spoken_test.dict`
2. Modellens transkripsjonsforslag, f.eks. `output.txt` fra kommandoen i Bruk, eller `data/predicted_nb_e_spoken.dict`.

- **WER** er antall feil transkripsjoner fordelt på totalt antall ord i modellens forslag. 1 feil = 1 ord.
- **PER** er summen av PER for hver transkripsjon, fordelt på totalt antall ord i modellens forslag. 1 feil = 1 fonem.

> **OBS**: Denne metoden tar ikke hensyn til ulik lengde på transkripsjonene. En transkripsjon med 2 fonemer der 1 er feil får en 50% PER, mens et ord på 10 fonemer med 1 feil får 10% PER, og gjennomsnittet blir 35%. Om man regner antall feil på totalt antall fonemer, blir gjennomsnittet for de to 16,7%.

```
python g2p_stats.py data/NB-uttale_e_spoken_test.dict
data/predicted_nb_e_spoken.dict
# WER: 14.049209423582523
# PER: 2.714882650391985
```

Model	Word Error Rate	Phoneme Error Rate
nb_e_written.fst	13.97114598599277	2.7038190765903214
nb_e_spoken.fst	14.049209423582523	2.714882650391985
nb_sw_written.fst	13.541060631724685	2.5423757844377284
nb_sw_spoken.fst	16.729141964989285	3.34063477772742
nb_w_written.fst	13.541060631724685	2.5423757844377284
nb_w_spoken.fst	17.186475877661337	3.4137304874392114
nb_t_written.fst	14.059519688924565	2.7190289235234104
nb_t_spoken.fst	--	--
nb_n_written.fst	14.059519688924565	2.7190289235234104
nb_n_spoken.fst	--	--

> **OBS**: *Modellforslagene fra t_spoken og n_spoken har ikke samme antall ord som testdata, som gjør at skriptet krasjer.*

Transkripsjonsstandard

G2P-modellene har blitt trent på data med transkripsjonsstandarden NoFAbet , som er lettere å lese for mennesker enn X-SAMPA. NoFAbet er delvis basert på ARPAbet og ble utviklet for Nasjonalbiblioteket av Nate Young i forbindelse med utviklingen av *NoFA*, en forced alignment-modell for norsk. Tabellen nedenfor viser ekvivalente symboler for notasjonene X-SAMPA, IPA og NoFAbet.

X-SAMPA-IPA-NoFAbet konverteringstabell

X-SAMPA	IPA	NoFAbet	Example
A:	ɑː	AA0	b**a**d
{:	æː	AE0	v**æ**r
{	æ	AEH0	v**æ**rt
{*I	æɪ	AEJ0	s**ei**
E*u0	æʉ	AEW0	s**au**
A	ɑ	AH0	h**a**tt
A*I	ɑɪ	AJ0	k**ai**
@	ə	AX0	b**e**hage
b	b	B	**b**il
d	d	D	**d**ag
e:	eː	EE0	l**e**k
E	ɛ	EH0	p**e**nn
f	f	F	**f**in
g	g	G	**g**ul
h	h	H	**h**es
I	ɪ	IH0	s**i**tt
i:	iː	II0	v**i**n
j	j	J	**j**a
k	k	K	**k**ost
C	ç	KJ	**k**ino
l	l	L	**l**and
l=	l̩	LX0	
m	m	M	**m**an
n	n	N	**n**ord
N	ŋ	NG	e**ng**

X-SAMPA	IPA	NoFAbet	Example
n=	ɳ	NX0	
o:	oː	OA0	r**å**
O	ɔ	OAH0	g**å**tt
2:	øː	OE0	l**ø**k
9	œ	OEH0	h**ø**st
9*Y	œy	OEJ0	k**øy**e
U	ʉ	OH0	f*__o__rt
O*Y	ɔy	OJ0	konv**oy**
u:	uː	OO0	b**o**d
@U	o~~ʉ~~	OU0	sh**ow**
p	p	P	**p**il
r	r	R	**r**ose
d`	ɖ	RD	reko**rd**
l`	ɭ	RL	pe**rl**e
l`=	ɭ̩	RLX0	
n`	ɳ	RN	ba**rn**
n`=	ɳ̩	RNX0	
s`	ʂ	RS	pe**rs**
t`	ʈ	RT	sto**rt**
s	s	S	**s**il
S	ʃ	SJ	**sj**u
t	t	T	**t**id
u0	~~ʉ~~	UH0	r**u**ss
}:	~~ʉ~~ː	UU0	h**u**s
v	ʋ	V	**v**ase
w	w	W	**W**ashington
Y	y	YH0	n**y**tt
y:	yː	YY0	n**y**

Trykklette stavelser er markert med 0 etter stavelseskjernen. Stavelseskjernen er markert med *1* ved tonem 1 og *2* ved tonem 2. Sekundærstress merkes med *3*.

Lisens

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