Raport de activitate
Proiect component TADARAV
etapa 3/2020

L. Georgescu, A. Caranica, C. Manolache, D. Oneață, G. Pop
H. Cucu, D. Burileanu, C. Burileanu

Laboratorul de cercetare Speech & Dialogue
Universitate POLITEHNICA din București

Activităţi

- Activitatea 3.9 - Analiza impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătățirii sistemelor de RAV
- Activitatea 3.10 - Îmbunătățirea soluției de filtrare și aliniere a transcrierilor aproximative cu semnalul de vorbire
- Activitatea 3.11 - Îmbunătățirea soluției pentru generarea de scoruri de încredere pentru RAV
- Activitatea 3.12 - Analiza impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV
- Activitatea 3.13 - Analiza impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV
- Activitatea 3.14 - Diseminare

Cuprins

- Activitatea 3.9
  - Sistemul de RAV bazat pe ESPNet
- Activitățile 3.10 și 3.12
  - Îmbunătățirea soluției de aliniere a transcrierilor aproximative cu semnalul de vorbire
  - Aplicarea metodei pe setul de date CoBiLiRo-raw
  - Aplicarea metodei pe setul de date Cdep-raw
- Activitățile 3.11 și 3.13
  - Scoruri de încredere pentru sisteme de RAV end-to-end
- Sistem de RAV actualizat în cadrul proiectului
- Activitatea 3.14: Publicații
- Activități 2021: cerere de brevet OSIM și articol de jurnal ISI

A3.9 Sistem de RAV ESPNet

- Motivație:
  - Sistem de RAV suplimentar -> informații complementare
  - Sistem de RAV end-to-end
- ESPNet – platformă de RAV modernă având la bază o arhitectură de tip Transformer
- Activități
  - Adaptarea arhitecturii pentru limba română
  - Calibrarea hiperparametrilor pentru modele acustice și lingvistice pentru limba română
  - Calibrarea parametrilor de decodare
  - Antrenarea unei rețele principale și a unui Transformer specific modelării lingvistice

A3.9 Sistem de RAV ESPNet

- Rezultate

Set de antrenare de vorbire adnotată	Model lingvistic	Parametri decodare		WER [%]		
		LMW	CTCW	RSC_eval	SSC_eval1	SSC_eval2
RSC-train + SSC-train1+2	Nu	n/a	n/a	13.3	25.1	78.0
	Da	0.5	0.5	8.8	**21.3**	**38.9**
	Da	0.6	0.4	**8.7**	21.6	46.5
	Da	0.7	0.3	9.0	23.0	64.3
	Da	0.8	0.2	11.2	25.8	84.6
RSC-train + SSC-train1+2 + SSC-train3-trans-v4 + SSC-train4-trans-v4	Da	0.6	0.4	3.4	15.3	23.5
Sistem RAV baseline bazat pe Kaldi (antrenat pe același set de date ca mai sus, folosind model de limbă pentru reevaluare lingvistică)				1.8	11.0	14.0

- Obs. și concluzii
  - Rezultate cu 50% mai slabe pe vorbire spontană
  - Rezultate de două ori mai slabe pe vorbire continuă
  - Antrenarea durează mult mai mult decât pentru sistemele Kaldi
  - Decodarea durează mult mai mult decât pentru sistemele Kaldi (xRT_{ESPNet}=3 vs. xRT_{Kaldi}=0.01)
  - Sistemul nu poate fi folosit pentru adnotare automată de vorbire

A3.10 Îmbunătăţirea soluţiei de aliniere transcrieri aproximative - vorbire

- Motivaţie: segmente scurte nealiniate între segmente lungi aliniate

Nr.	Secvenţa anterioară (aliniată)	Secvenţă nealiniată	Secvenţa următoare (aliniată)
1	… se îndreaptă către susţinătorii săi	o dronă	filmează evenimentul …
2	… în zona	Egiptului	unde există …
3	… a obţinut fondurile	administrator bloc	la doi ani …

- Propunere: utilizarea segmentelor nealiniate din transcrierea aproximativă
- Rezultat: au fost obţinute cu 10% mai multe materiale audio + transcrieri
- Utilizarea noului set de date pentru RAV:
  - Rezultate comparabile cu cele precedente

A3.10 Aplicare metodă pe CoBiLiRo-raw

- Setul de date CoBiLiRo-raw
  - 76 înregistrări audio + transcrieri aproximative
  - 70 ore de vorbire
- În urma aplicării metodei de aliniere, a fost aliniat:
  - 45% din materialul audio cu 37% din transcrierea aprox.
  - 31.5 ore de vorbire – 268k cuvinte
- Sistemul RAV reantrenat cu (225h + 31.5h)
  - WER pe RSC-eval: 1.9% -> 1.8%
  - WER pe SSC-eval1: 15% -> 14%

A3.10 Aplicare metodă pe CDep-raw

- Setul de date CDep-raw
  - +350k înregistrări audio + transcrieri aproximative
  - +3500 ore de vorbire de la +2500 vorbitori, +25M cuvinte
- În urma aplicării metodei de aliniere, a fost aliniat:
  - 25% din materialul audio cu 85% din transcrierea aprox.
  - 879 ore de vorbire – 21M cuvinte → situație ciudată!
- Sistemul RAV reantrenat cu (225h + 291h + 879h)
  - WER pe RSC-eval: 1.8% -> 1.7%
  - WER pe SSC-eval1: 11% -> 12.3%
  - WER pe SSC-eval2: 14% -> 15.4%
  - WER pe CDep-eval: 6.9% -> 14%

A3.11 Scoruri de încredere pentru sisteme de RAV end-to-end

- Am propus o serie de noi scoruri de încredere la nivel de cuvânt
  - Bazate pe trăsături precum
    - Probabilitatea token-urilor
    - Entropia calculată peste vocabularul de token-uri
  - Folosind mai multe metode de agregare a scorurilor la nivel de cuvânt
- Am analizat care scoruri sunt mai bune pe diverse baze de date
  - În limba engleză: TED-LIUM, librispeech, Common-Voice
  - În limba română: RSC-eval, SSC-eval
- Detalii în (Oneață, SLT 2021)
- Nu au putut fi aplicate pentru adnotarea de date pentru că decodarea cu ESPNet durează foarte mult

Sistem de RAV actualizat în ReTeRom

- Îmbunătățiri aduse sistemului RAV al SpeeD
  - Îmbunătățiri ale platformelor software (Kaldi, ESPNet)
  - Îmbunătățiri ale algoritmilor de procesare de text
  - Creșterea seturilor de date de vorbire și text
  - Actualizarea modelelor acustice și de limbă

- Acronime seturi de date de vorbire
  - BAS (RSC-train + SSC-train1 + SSC-train2), 225h
  - SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4), 292h
  - COB (CoBiLiRo-trans-v4), 31h
  - COR (CoRoLa), 84h
  - CDP (cdep-trans-v4), 879h

Sistem de RAV actualizat în ReTeRom

- Rezultate

Cod model acustic	Set de antrenare					Set de evaluare (WER[%])			
	BAS 225h	SSC 292h	COB 31h	COR 84h	CDP 879h	RSC-eval	SSC-eval1	SSC-eval2	CDep-eval
train-base	x					1.9	15.0	20.0	10.8
train3-v4	x	x				1.8	11.0	14.0	6.9
train14	x	x	x			1.6	11.3	14.4	7.4
train11	x	x	x	x		1.6	10.3	12.2	6.1
train11*	x	x	x	x		1.9	9.4	11.4	5.4

* Actualizare modele de limbă

- Observaţii şi concluzii
  - Îmbunătăţire relativă semnificativă (40%) pe vorbire spontană
  - Nu se obţin îmbunătăţiri simultane şi pe vorbire continuă şi pe vorbire spontană

Sistem de RAV actualizat în ReTeRom

- Contribuția diverselor seturi de date create în proiect

Cod model acustic	Set de antrenare					Set de evaluare (WER[%])			
	BAS 225h	SSC 292h	COB 31h	COR 84h	CDP 879h	RSC-eval	SSC-eval1	SSC-eval2	CDep-eval
train-base	x					1.9	15.0	20.0	10.8
train3-v4	x	x				1.8	11.0	14.0	6.9
train15	x		x			1.8	14.0	21.1	13.1
train17	x			x		1.8	11.9	15.4	8.0

- Observații și concluzii
  - SSC > COR > COB, atât dpdv dimensiune, cât și dpdv contribuție la performanță
  - CDP nu a fost utilizat în acest experiment

Sistem de RAV actualizat în ReTeRom

- Contribuția seturilor de date adăugate în 2020

Cod model acustic	Set de antrenare					Set de evaluare (WER[%])			
	BAS 225h	SSC 292h	COB 31h	COR 84h	CDP 879h	RSC-eval	SSC-eval1	SSC-eval2	CDep-eval
train-base	x					1.9	15.0	20.0	10.8
train3-v4	x	x				1.8	11.0	14.0	6.9
train14	x	x	x			1.6	11.3	14.4	7.4
train18	x	x		x		1.8	10.5	12.5	6.4
train13	x	x			x	1.7	12.3	15.4	14.0

- Observații și concluzii
  - COB adăugat peste BAS+SSC nu ajută pe vorbire spontană
  - COR adăugat peste BAS+SSC nu ajută pe vorbire continuă
  - CDP adăugat peste BAS+SSC strică cel mai mult

Sistem de RAV actualizat în ReTeRom

- Observații și concluzii finale
  - Pentru fiecare sarcină de RAV este nevoie de materiale de vorbire de antrenare adaptate sarcinii
  - Creșterea setului de date de antrenare fără legătură directă cu sarcina de RAV nu conduce neapărat la creșterea performanței
  - Cazul CDep mai trebuie studiat
  - Per ansamblu sistemul RAV al SpeeD a fost îmbunătățit:
    - WER pe RSC-eval: 1.9% -> 1.6%
    - WER pe SSC-eval1: 15% -> 9.4%
    - WER pe SSC-eval2: 20% -> 11.4%
    - WER pe CDep-eval: 10.8% -> 5.4%

Activitatea 3.14 Diseminare

- A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, “RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition,” LREC 2020, pp. 6606-6612.
- C. Manolache, A.-L. Georgescu, A. Caranica, H. Cucu, “Automatic Annotation of Speech Corpora using Approximate Transcripts,” TSP 2020.
- D. Oneață, A.-L. Georgescu, H. Cucu, D. Burileanu, C. Burileanu, “Revisiting SincNet: An Evaluation of Feature and Network Hyperparameters for Speaker Recognition,” EUSIPCO 2020.
- G. Pop, H. Cucu, D. Burileanu, C. Burileanu, “Cough Sound Recognition in Respiratory Disease Epidemics,” in ROMJIST, vol. 23, no. S, pp. S77–S89, 2020, ISI IF 0.661.
- A.-L. Georgescu,, C. Manolache, D. Oneață, H. Cucu, C. Burileanu, “Data-filtering methods for self-training of automatic speech recognition systems,” IEEE SLT 2021.
- D. Oneață, A. Caranica, A. Stan, H. Cucu, “An evaluation of word-level confidence estimation for end-to-end automatic speech recognition,” IEEE SLT 2021.

Activităţi 2021

- Management și diseminare:
  - Cerere de brevet depusă la OSIM
  - Articol ştiinţific în jurnal ISI

Mulțumesc!