A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects

Yonatan Belinkov, James Glass
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
{belinkov,glass}@mit.edu

1. Overview

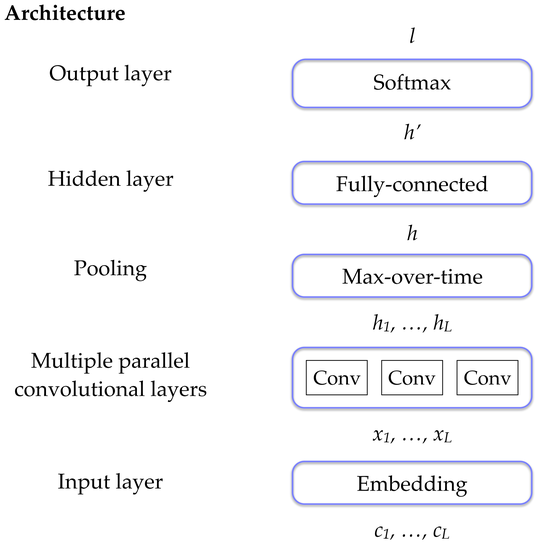
- Discriminating closely-related language varieties
- DSL shared-task with two sub-tasks:
  1. Similar languages, journalistic texts
  2. Arabic dialects, speech transcriptions
- Previous work mostly used sequences of characters and words, with simple machine learning algorithms (SVM, MaxEnt)
- We use a fully character-level convolutional neural network

2. Approach

Multi-class classification

- Given pairs of texts and labels, $\{t^{(i)}, l^{(i)}\}$, learn predictor $f : t \longrightarrow l$
- Implement predictor as a neural network
- Represent text as sequence of characters: $t := c_1, \ldots, c_L$

Architecture



3. Implementation Details

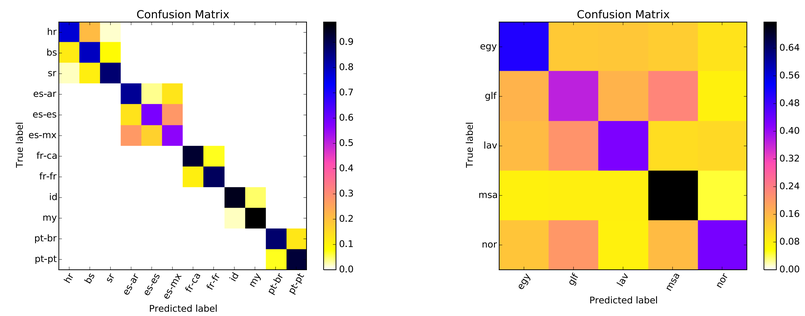
- Cross-entropy loss with mini-batches, Adam optimizer
- Early stopping on dev set with a 10 epoch patience
- Implemented in Keras with the TensorFlow backend
- Hyper-parameters tuned on 10% of the Arabic train set
  - ρ_{emb}=0.2, ρ_{fc}=0.5, L=400, d_{emb}=50, d_{fc}=250
  - Conv filters: *{1*50, 2*50, 3*100, 4*100, 5*100, 6*100, 7*100}*

4. Submitted Runs

- Sub-task 2 (Arabic dialects)
  - Run 1: 90% of train for training, 10% for development
  - Run 2: 100% of train for training, stop based on Run 1
  - Run 3: 10 models trained on different 90%/10% splits
- Sub-task 1 (languages): Run 1 similar; Run 2 more filters; Run 3 more hidden units and dropout in FC layer

5. Results and Discussion

Test Set	Track	Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
A	closed	Baseline	0.083			
A	closed	run1	0.8042	0.8042	0.8017	0.8017
A	closed	run2	0.825	0.825	0.8249	0.8249
A	closed	run3	**0.8307**	**0.8307**	**0.8299**	**0.8299**
A	closed	Best	*0.8938*			*0.8938*
C	closed	Baseline	0.2279			
C	closed	run1	0.4487	0.4487	0.4442	0.4449
C	closed	run2	0.4357	0.4357	0.4178	0.4181
C	closed	run3	**0.4851**	**0.4851**	**0.4807**	**0.4834**
C	closed	Best	*0.5117*			*0.5132*



(a) Sub-task 1, test set A, Run 3

(b) Sub-task 2, test set C, Run 3

Results

- 6/18 in sub-task 2; 2nd to last in sub-task 1
- Spanish most difficult, Malay/Indonesian easiest
- Gulf most confusing Arabic dialect, MSA easiest

Discussion of the Arabic task

- Transcribed texts, Buckwalter transliteration
- MSA confusion, news broadcasts
- Linguistic vs geographic proximity

6. Error Analysis

1. Competing features: *AllbnAnyp*, *hmA*, *-An*; verb-subject word order: *AndlEt AlHrb*
2. Mixing: *>h HDrp*, *bdy* vs *hl syHdv*
3. Morphology: *Alm$kwk fyh*
4. Word vs char: *AlmAlky*, *AlhA$my*
5. ASR mistakes: *byt>vr* vs *byt>tr*
6. Rare features: *<HnA wyAhm*. But: *bqyt* common in NA

	True	Predicted	Text
1	MSA	Levantine	AndlEt AlHrb AllbnAnyp EAm 1975 >Syb bxybp >ml whmA yrwyAn kyf ynhArwn wqthA
2	MSA	Egyptian	>h HDrp AlEmyd AlAHtkAk bdy dm$q AlEASmp AlsyAdyp AlEASmp AlsyAsyp fy fy >kvr mn mrp wbEmlyp nwEyp kbyrp jdA hl syHdv AlmnErj fy h*h AlmwAjhp
3	MSA	Gulf	>mA xrjt ElY tlfzywn Aldwlp fy Alywm AltAly lvwrp wqlt lh HAfZ ElY tAryx >ql AlwzArp Alywm Thr AlbrlmAn mn AlEDwyAt Alm$kwk fyh <dY msyrp Al<SlAH wbdA h*A qbl SlAp AljmEp
4	MSA	North African	>wlA Al Al Alsyd AlmAlky ytmnY mn TArq AlhA$my Alxrwj wlA yEwd
5	Egyptian	Gulf	>nA bEmrnA ftrp mn HyAty snp Al>xyrp mtEwd ElY wqf Altfrd AHtlAly tEtbr llmsjd gryb mn mjls AlwzrA' wmjls Al$Eb wAl$wrY wnqAbp AlmHAmyn wxlAfh fkAn >y wAlAHtjAjyp byt>vr bhA Almsjd b$>n Al>wDAE
6	North African	Gulf	$Ahd tglb wAjb \|xr mr Ebr EddA mn brnAmjh <HnA wyAhm lA ymnE xrwj bAlb$r ftzydh whlA Em lxrwqAt AlHq Al$yx xAld Hqq mEy lA yglq fyjb hdf AlnAtw bAlAxtyAr mn Altwqf lA tqAs bqyt Endk nsmH lkl $y' HtY tqrr trHb

Table 1: Example errors made by our system on the Arabic data set.

7. Future Work

- Combine word and char features
- Add word white-lists
- Combine acoustic and phonetic features

Acknowledgments: This work was supported by the Qatar Computing Research Institute (QCRI). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.