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, \( \{h, l\} \), learn predictor \( f \) \( \rightarrow l \)
- Implement predictor as a neural network
- Represent text as sequence of characters: \( t := c_1, ..., c_l \)

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
- \( p_{\text{drop}}=0.2, p_{\text{reg}}=0.5 \), \( L=400 \), \( d_{\text{emb}}=50 \), \( d_{\text{hid}}=250 \)
- 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: 10 models trained on different 90%/10% splits

5. Results and Discussion
- Test Set A:
  - Run 1: 0.4487, Run 2: 0.8042, Run 3: 0.4485
  - Baseline: 0.083

- Test Set C:
  - Run 1: 0.4178, Run 2: 0.8249, Run 3: 0.4851
  - Baseline: 0.2279

6. Error Analysis
1. Competing features: AllInAnp, locA, ~A; verb-subject word order; AddEnt AInHep
2. Mixing: ~A HDel, bfy vs hl sqfd
3. Morphology: AllInHep
4. Word vs char: AllInHep, AInHep
5. ASR mistakes: bfy vs bgt/btr
6. Results: \( f_{\text{bc}} \) is byt/dim. But byt/ght common in NA

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.