Massively Multilingual Neural Grapheme-to-Phoneme Conversion

Ben Peters

Saarland University/DFKI

OpenNMT Workshop Paris, March 2018

Outline

Outline

Grapheme-to-Phoneme Conversion

Convert orthography to pronunciation

- Graphemes: the abstract units of writing
- Phonemes: the abstract units of sound
- Figure out the relationship

Problem: scaling down to low resource languages

Figure: "Just sound it out..."

Grapheme-to-Phoneme Conversion

How do you build a system for a low resource language?

- Rule-based systems need language-specific expertise, which you don't have
- Statistical systems need annotated data, which you don't have

This seems intractable, but...

Idea

Train a single system for all languages

- A handful of scripts (Latin, Cyrillic, Arabic) cover most languages
- Spelling rules are similar cross-linguistically:
  English: <real> = /ɹiəl/
  Spanish: <real> = /real/
  German: <real> = /ʁeal/
  Brazilian Portuguese: <real> = /xeaw/
- Solve the data sparsity problem by letting low resource languages learn from high resource data

But traditional models can't take advantage of these similarities

- /ɹ/ and /ʁ/ are just different phonemes
- Solve it with a sequence-to-sequence model: details next section

Outline

Architecture

Fairly vanilla sequence-to-sequence model

- Basically OpenNMT (Klein et al., 2017) defaults, but with smaller embeddings and layers
- Global attention
- Bidirectional LSTM encoder
- Completely character-level
- Language indicated with token (Johnson et al., 2016) or language embedding (Östling & Tiedemann, 2017) on source side

The key point:

- Grapheme and phoneme embeddings are shared across all languages
- This captures our intuition: a grapheme usually mean similar things in different languages

Experiments

Dataset

- Deri & Knight (2016)'s corpus scraped from Wiktionary, cleaned to conform to Phoible (Moran et al., 2014).
- 311 languages to train, 507 to test
- We limit to 9000 words per language
- Evaluation on high and low resource subsets of test data

Training

- Stochastic Gradient Descent
- 64 words per mini-batch

Baseline

- Deri & Knight's system
- WFST models for high resource languages adapted for low resource languages
- Linguistic knowledge guides adaptation

Evaluation Metrics

- Phoneme Error Rate (PER): the Levenshtein distance between predicted and gold standard phoneme sequences, divided by the length of the gold length.
- Word Error Rate (WER): percentage of words in which the predicted and gold phoneme sequences do not match.
- Word Error Rate at 100 (WER 100): the percentage of words for which none of the first 100 guesses is correct.

Metrics are averaged across languages, weighting all equally

Results

Model	WER	WER 100	PER
wFST w/ adaptation	88.04	69.80	48.01
seq2seq w/ LangID feature	**71.94**	**40.69**	**35.38**
seq2seq w/ LangID token	74.10	43.23	37.85
seq2seq w/o LangID	83.65	47.13	51.87

Table: Results on the 229 language 'adapted' set, a mix of high- and low-resource languages. More details: (Peters et al., 2017)

But is there a benefit over a monolingually-trained network?

Results

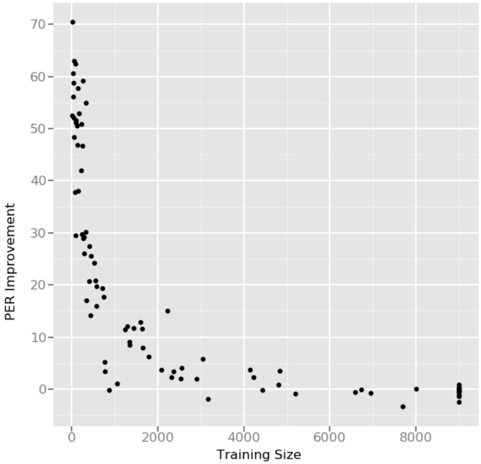
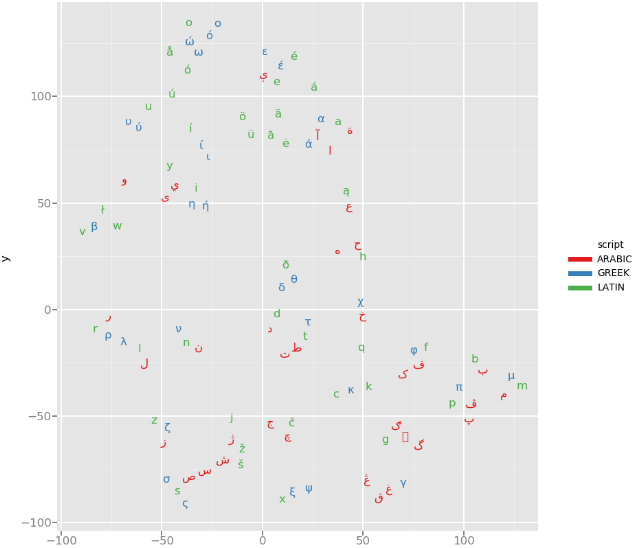
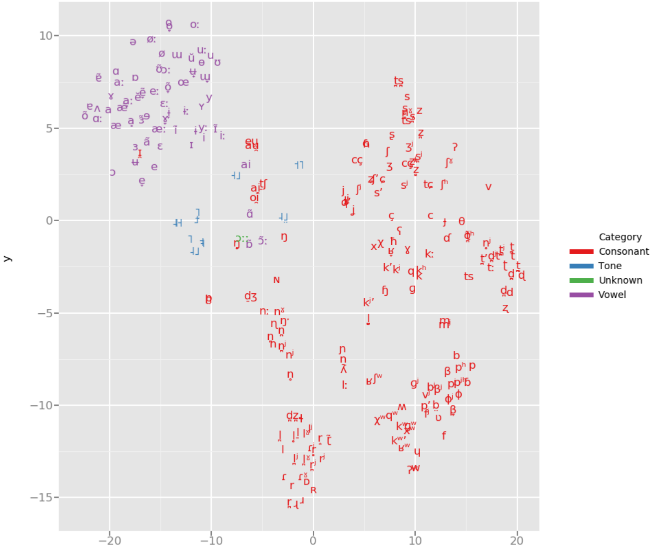


Figure: Training Size vs. Improvement over monolingual NN

Grapheme Embeddings



Phoneme Embeddings



Example Output

Identifying the language has a big effect on the system output

Language	**Pronunciation**
English	dʒuːæɪs
German	jʊtsə
Spanish	xwiθe̞
Italian	dʒuitʃe
Portuguese	ʒwisĩ
Turkish	ʒʊɪd̪ʒɛ
Arabic	juːis

Table: Pronunciations of ‘juice’ learned by the model

Outline

Data Cleaning

How do you clean multilingual data consistently?

- Different phonetic features matter in different languages
- Suprasegmentals matter
- How do you evaluate data cleaning (see Hixon et al., 2011)?

segment	consonantal	trill	anterior	distributed
ɹ	-	-	-	+
r	+	+	+	-

segment	consonantal	lateral	anterior	distributed
ɹ	-	-	-	+
l	+	+	+	-

Figure: ɹ is the same distance from l as from r

Linguistic Knowledge

How do we encode things we know that we can't get from the data?

- Great results for Serbo-Croat-Bosnian (tons of data in corpus)
- Much worse for Serbian, Croat, Bosnian (much less)

One possibility: directly use typology as extra input (Tsvetkov et al., 2016)

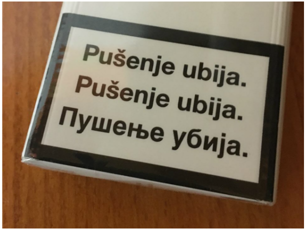


Figure: "Multilingual" labeling

Other Ideas

Multilinguality

- Output softmax layer gives probability distribution over union of all phoneme inventories
- For any language, most of these phonemes are a priori impossible
- Language-specific softmax?

Attention

- Little reordering in g2p, alignments close to one-to-one
- Local or monotonic attention?

Jointly learn phoneme-to-grapheme

- Doubles training data
- Enables zero-shot transliteration

References

- Deri, A. and Knight, K. (2016) Grapheme-to-phoneme models for (almost) any language. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* volume 1, pages 399 – 408.

- Hixon, B., Schneider, E., and Epstein, S. L. (2011) Phonemic similarity metrics to compare pronunciation methods. *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages pages 825 – 828.

- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016) Google's multilingual neural machine translation system: Enabling zero-shot translation. *ArXiv preprint*, 1611.04558.

- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017) OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv preprint*, 1701.02810.

- Moran, S., McCloy, D., and Wright, R., editors (2014) PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

References

- Östling, R. and Tiedemann, J. (2017) Continuous multilinguality with language vectors. *EACL 2017*, page 644.
- Peters, B., Dehdari, J., and van Genabith, J. (2017) Massively multilingual neural grapheme-to-phoneme conversion. *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 19-26.
- Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., and Dyer, C. (2016) Polyglot neural language models: A case study in cross-lingual phonetic representation learning. *NAACL 2016*.