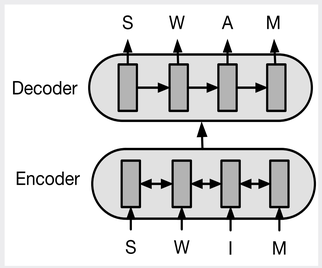
Recurrent Neural Networks as a Strong Domain-General Baseline for Morpho-Phonological Learning

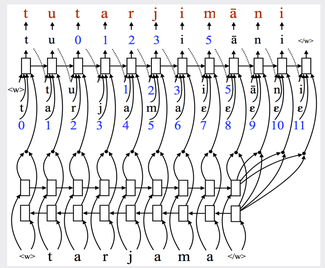
Christo Kirov
Johns Hopkins University Center for Language and Speech Processing, ckirov at gmail dot com

Introduction

- A standard strategy for comparing competing theories of learning in morpho-phonology is to compare the fit of models that embody each theory to data.
- This strategy can be used to study the role of bias in learning, comparing models with UG bias to domain-general models.
- It is important to use the strongest available domain-general models as the baseline for comparison.
- Furthermore, as better domain-general learning mechanisms are discovered, old model comparisons should be re-examined.

Encoder/Decoder Networks





Older connectionist models such as the classic wicklephone-based Rumelhart & McClelland (1986) past-tense network had many limitations. Modern recurrent encoder-decoder architectures (Sutskever et al. 2014) meet Albright & Hayes's (2002) criteria for a complete morpho-phonological model:

- Able to generate complete output forms, rather than classify inputs into categories such as 'regular' or 'irregular.'
- Able to make multiple output guesses, and assign numerical ratings to each.
- Able to generalize to unseen data.

Additional benefits of encoder-decoder architectures include:

- State-of-the-art performance in string-to-string transduction tasks, such as morphological (re-inflection).
  - ∼95% of inflected forms predicted correctly across multiple languages (including templatic morphology in Arabic and Maltese, and long-distance harmony patterns in Hungarian and Turkish) during SIGMORPHON 2016 competition.
- Highly domain-general by default. Input is a series of abstract symbols presented without phonological features. Network must learn distributed representation for each phoneme.

*All models described in this paper were trained for a maximum 15 epochs using a variant of the LSTM-based architecture in (Aharoni et al. 2016) with 2-layer decoder and encoder of 100 hidden units each.

Case Study: Romanian Plurals

- Romanian has four plural suffixes: (-i -e -uri -ale). Predicting which suffix pluralizes a given neuter or feminine stem is difficult.
- Previous best performance achieved by hand-tuned Conditional Maximum Entropy Grammar from Grosu & Wilson (2016):

$$p(\text{suffix}|\text{stem}) \propto exp(-\sum_{c,w \in C} wc(\text{suffix}, \text{stem}))$$

- Network directly predicts plural form given singular input: (f: fata → fete), (n: scaun → scaune)

Given random %80/%20 train/test split of the 39,500 singular/plural pairs collected by Grosu & Wilson:

Romanian	Network	MaxEnt
Feminine	**95.8%**	92.2%
Neuters	**86.2%**	80.3%

Case Study: English Past

- Humans can generalize sub-regularities in irregular past tense forms (swim/swam/swum ∼ spring/sprang/sprung)
- Albright & Hayes (2002) trained their Minimal Generalization Learner to predict past forms of 4253 CELEX stems.
- They used the model to predict human human production probabilities of past tense forms given 'wug' present stems.

Network trained on the same CELEX data correlates with human production probabilities better than MGL:

English	Network	MGL
Regular (rife ∼ rifed, n=58)	**.735**	.619
Irregular (rife ∼ rofe, n=74)	**.711**	.143

Case Study: Turkish Laryngeals

- Some Turkish noun stems undergo a laryngeal alternation when a vowel-initial suffix is added (ret ∼ reddim).
- Becker et al. (2011) claim UG learning bias against using stem vowel cues to predict alternation.
  - Fit logistic regression models to a lexicon of 3,002 nominal stems.
  - Models without vowel factors perform better at predicting human forced-choice performance on 72 'wug' nominal stems.

But, a network trained on the same lexicon performs as well the UG-biased regression model on wugs:

Model	D_{xy}
place*size (regression without vowel cues)	.360
place*size + place*high + place*back (unbiased regression)	.349
network (unbiased)	**.368**

Conclusions and Outstanding Issues

Modern recurrent neural networks provide a strong domain-general morpho-phonological learning baseline. This makes them a promising tool for studying the role cognitive and linguistic biases play in learning:

- Are there attested patterns that networks have a hard time learning (e.g., as shown by slower convergence during training)?
  - Of particular interest here would be reduplication and metathesis, behaviors that can't be easily represented by simple finite-state transducers without extensive restrictions.
  - If such patterns exist, is it possible to bias the network to make learning easier (e.g., by initializing weights at a specific starting point).
- Similarly, are there *un*attested or difficult patterns, as found either by typological surveys or artificial grammar learning experiments, that networks learn too easily?
  - If so, what kind of bias or regularization can be built into the networks to limit learnability?
- What representations and mechanisms do networks actually learn?
  - Do hidden units show large changes in activation when particularly informative input symbols are reached (Kirov et al. 2011; Kadar et al. 2016)? If so, network analysis can help us understand and describe phonological patterns by highlighting important parts of the input that we had not considered to be relevant cues.
  - Can learned representations be transformed into feature vectors comprised of traditional linguistic features (e.g., +/- vocalic, +/- labial)?

References

1. Aharoni, R; Goldberg, Y; and Belinkov, Y. 2016. *Improving Sequence to Sequence Learning for Morphological Inflection Generation: The BIU-MIT Systems for the SIGMORPHON 2016 Shared Task for Morphological Reinflection.* Proceedings of the 2016 Meeting of SIGMORPHON, Belin, Germany: Association for Computational Linguistics. https://github.com/roeeaharoni/morphological-reinflection.
2. Albright, A; and Hayes, B. 2002. *Modeling English Past Tense Intuitions with Minimal Generalization.* Proceedings of the 2002 Workshop on Morphological Learning, Philadelphia, PA: Association for Computational Linguistics.
3. Becker, M; Ketrez, N; and Nevins, A. 2011. *The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations.* Language 87:1, pp. 84?125.
4. Cotterell, R; Kirov, C; Sylak-Glassman, J; Yarowsky, D; Eisner, J; and Hulden, M. 2016. *The SIGMORPHON 2016 Shared Task—Morphological Reinflection.* Proceedings of the 2016 Meeting of SIGMORPHON, Berlin, Germany: Association for Computational Linguistics. http://ryancotterell.github.io/sigmorphon2016/.
5. Grosu, I; and Wilson, C. 2016. *Experimental Evidence for Stem Ending and Size Factors in Romanian Plural Formation.* The 2016 Annual Meeting of the Linguistic Society of America, Washington, D.C.
6. Kadar, A; Chrupala, G; and Alishahi, A. 2016. *Representation of Linguistic Form and Function in Recurrent Neural Networks.* TACL.
7. Kirov, C; and Frank, B. 2011. *Processing of Nested and Cross-Serial Dependencies: an Automaton Perspective on SRN Behavior.* Connection Science. Volume 24. Issue 1. pp. 1-24.
8. Rumelhart, D; and McClelland, J. 1986. *On Learning the Past Tenses of English Verbs.* Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 2, Cambridge, MA: MIT Press, pp. 216-271.
9. Sutskever, I; Vinyals, O; and V. Le., Q. 2014. *Sequence to Sequence Learning with Neural Networks.* CoRR, abs/1409.3215.