Variational Bayesian Inference for Unsupervised Lexicon Discovery

Emily Kellison-Linn, Elias Stengel-Eskin, Timothy O'Donnell

{emily.kellison-linn, elias.stengel-eskin}@mail.mcgill.ca, tim.odonnell@mcgill.ca

McGill
Department of Linguistics

mlmlab
Montreal Language Modeling Lab

Problem: Learning language from acoustic input

Given an entirely unlabeled recording of speech in any arbitrary language, identify repeated sounds, words, and phrases that make up that language.

There are thousands of languages in the world for which there exists only a small amount of recorded audio, all of it unlabeled. Supervised models cannot learn anything from these languages, but unsupervised models can.

This problem of learning to segment a continuous speech stream into meaningful units is also fundamental to the question of how children learn language.

The model

We implement a three-part hierarchical model based on linguistic theory which learns:

- a segmentation of audio into a sequence of reusable phone-like units
- a sequence of edit operations on that sequence, and
- a grouping into word- and phrase-like units

The model is based on Lee et. al. (2015). We address some of its limitations by:

- reimplementing the model using faster **variational inference**
- introducing a more sophisticated **noisy channel** representation

Variational inference

Training this system requires solving an intractable Bayesian inference problem, arising from computing the marginal probability of the data.

This problem is commonly solved via sampling. However, sampling approaches are difficult to scale to large datasets. Variational inference (VI) recasts this approximation as an optimization problem, using the lower-bound on the incalculable marginal probability as an objective function, written as

$$\log\ p(X \mid \Phi) \geq \mathbb{E}_q[\log\ p(Z, X \mid \Phi)] - \mathbb{E}_q[\log\ q(Z)]$$

VI converges faster than sampling and can yield better results. Most importantly, it lends itself well to parallelization, letting it scale to very large datasets.

Given the promising results of Lee et. al. (2015) on a fairly small dataset, implementing variational methods to allow faster testing with more data is a major step towards a scalable fully unsupervised algorithm for lexicon discovery.

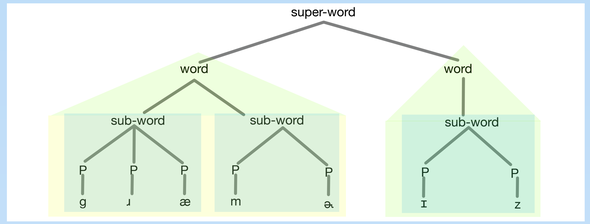
Simulations

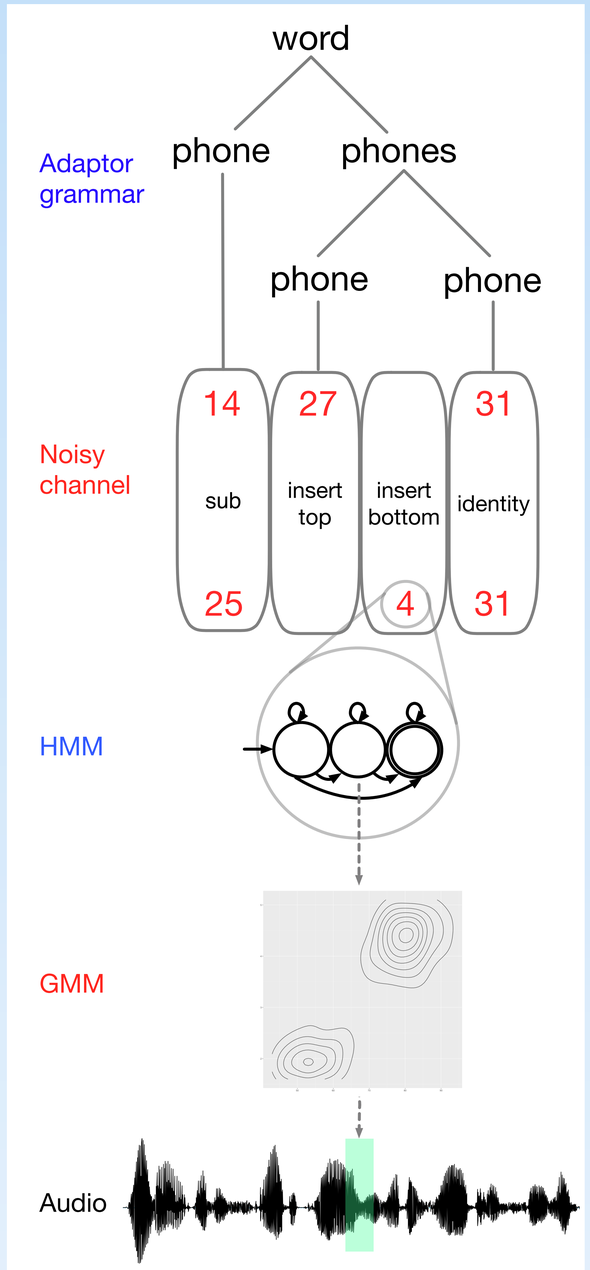
We are investigating how model performance changes under the following conditions:

- Varying the duration of training audio provided to the model
- Varying the structure of rules inferred by the adaptor grammar
- Removing the noisy channel component (lesioning experiment)

Acknowledgements

Thanks to Lucas Ondel, Jackie Lee, and Ke Zhai for sharing their code with us.





Parsing sound units into words

We use **adaptor grammars**, introduced by Johnson et al. (2007), to learn sentence, word, and sub-word structure.

- Adaptor grammars operate by dynamically adjusting probabilities of an underlying PCFG according to the data
- Parses sound-unit sequences into morphemes, words, and phrases
- Pitman-Yor Process prefers re-use of rules for frequent sequences

We base our implementation on the work of Zhai et al. (2014).

Through the noisy channel

The pronunciation of sounds and words in human speech is variable, so we need to account for this variation. This is the role of the **noisy channel model**.

- Learns a sequence of substitute, insert, and delete operations on the sound-unit sequence
- Operation probabilities conditioned on context, accounting for systematic pronunciation variation
- Probabilities learned via a novel algorithm similar to forward-backward algorithm

Segmenting audio into sound units

We must distinguish between distinct sounds in the audio stream while also identifying segments which are instances of the same sound. We do this with a version of the **Dirichlet Process hidden Markov model (DPHMM)** introduced by Lee et al. (2015). The DPHMM prefers reusing sound labels when two sounds are similar enough. In the generative model:

- A Dirichlet Process generates a sequence of sound types
- Each sound type is associated with an HMM which outputs a state sequence
- Each state outputs an audio sample drawn from a Gaussian mixture model

We base our work on the DPHMM implementation of Ondel et al. (2016).

Future work

- Extend the adaptor grammar module to handle more linguistically realistic formalisms
- Use deep-learning techniques to accelerate inference in the acoustic model
- Implement more realistic noisy channel assumptions
- Leverage language-independent nature of the model for developing ASR resources such as pronunciation dictionaries for under-resourced languages

References

Lee, C.-y., O'Donnell, T. J., and Glass, J. (2015). Unsupervised lexicon discovery from acoustic input. Transactions of the Association for Computational Linguistics, 3:389–403.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A frame- work for specifying compositional nonparametric Bayesian models. In Advances in neural information processing systems, pages 641–648.

Zhai, K., Boyd-Graber, J., and Cohen, S. B. (2014). Online adaptor grammars with hybrid inference. Transactions of the Association for Computational Linguistics, 2:465–476.

Ondel, L., Burget, L., and Černocký, J. (2016). Variational inference for acoustic unit discovery. Procedia Computer Science, 81:80–86.