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

Task: Unsupervised lexicon discovery

Given an entirely unlabeled recording of speech in any arbitrary language, identify repeated sounds, words, and phrases that make up the lexicon of 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; however, unsupervised models can.

This task is also linguistically interesting because human language learners face the same challenge of learning to segment a continuous speech stream into meaningful word units.

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 that of 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.

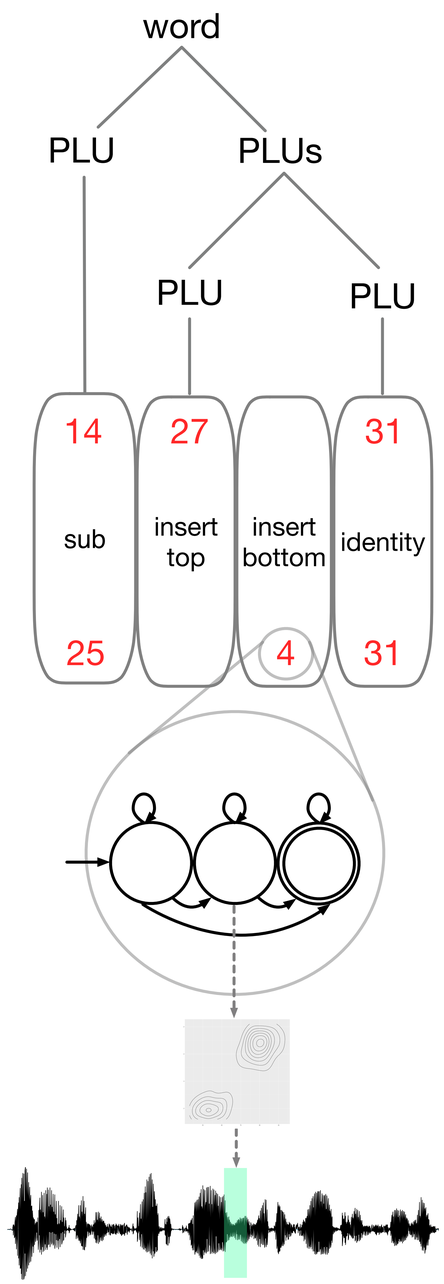
References

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.



Parsing sound units into words

We use **adaptor grammars**, introduced by Johnson et al. (2007), to learn parses of phone-like units (PLUs) into frequently-repeated structures representing morphemes and words.

- Adaptor grammars operate by dynamically adjusting probabilities of an underlying PCFG according to the data
- Uses the Pitman-Yor Process, a stochastic process
- Prefers re-use of rules for frequently observed sequences

We use the variational implementation of adaptor grammars by Zhai et al. (2014).

Noisy channel

The noisy channel model learns a sequence of substitute, insert, and delete operations between the PLU sequence parsed by the adaptor grammar and the sequence inferred from the audio input, thus mediating between the other two components.

Operation probabilities are conditioned on PLU context, approximating a phonological system. Probabilities are learned via application of the forward-backward algorithm.

Segmenting audio into sound units

We use a version of the **Dirichlet process hidden Markov model (DPHMM)** introduced by Lee et al. (2015) to learn a sequence of PLUs from audio input. In the generative model:

- A Dirichlet process generates a sequence of PLUs
- Each PLU is associated with an HMM which outputs a state sequence
- Each state outputs an audio sample according to a Gaussian

We use the variational DPHMM implementation by Ondel et al. (2016).

Future work

We plan to run lesioning experiments on different portions of the model, to investigate each component's contribution to the joint model; for example, Lee et al. (2015) tested performance when removing the noisy channel.

Given the model's language-independent nature, it also has applications to developing ASR resources, such as pronunciation dictionaries, for under-resourced languages.

Acknowledgements

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