Unsupervised Featureized HMM Modeling

Shuoyang Ding (sding8, dings@jhu.edu)
Adithya Renduchintala (arenduc1, adithya.renduchintala@jhu.edu)

November 17, 2014

1 Abstract

In this project we propose to implement a featurized version of baum-welch training of HMMs. Our implementation will closely follow [1]. The basic idea is to replace the probabilities in an HMM model with probabilities estimated using a log linear modeling technique. We would like to apply this model to the following common NLP tasks - Word Alignment, POS tagging and Word Segmentation.

2 Methods

In this section we will discuss how the training of such a model will work.
In the fully observed case, we can simply count the transitions and emissions and then maximize the probability of the data under the model using MLE. In the un-supervised case this process of counting becomes a process of getting expected counts. The probability of the data under the model is:

$$P(X \mid \theta) = \prod_{k,l \in \mathbf{S},\mathbf{S}} a(s_l \mid s_k)^{A_{kl}} \prod_{x_b, L \in \mathbf{V},\mathbf{S}} e(x_b \mid s_l)^{E_{lb}} \tag{1}$$

Where A_{kl} and E_{lb} are the expected counts of transitions and emissions. This can be expressed as a log likelihood shown below:

$$\begin{aligned} L(X,\theta) = & \sum_{s_k \rightarrow s_l} A_{kl}(\theta' \cdot f(s_k, s_l) - log \sum_{s'_l} exp(\theta' \cdot f(s_k, s'_l)))+ \\ & \sum_{s_l \rightarrow x_b} E_{lb}(\theta' \cdot f(x_b, s_l) - log \sum_{x'_b} exp(\theta' \cdot f(s_l, x'_b))) \end{aligned} \tag{2}$$

Here we have replaced the emission and transition probabilities using the log-linear form, given by:

$$a'(s_l \mid s_k) = \frac{exp(\theta' \cdot f(s_k, s_l))}{\sum_{s'_l} exp(\theta' \cdot f(s_k, s'_l))} \tag{3}$$

$$e'(x_b \mid s_l) = \frac{exp(\theta' \cdot f(s_l, x_b))}{\sum_{x'_b} exp(\theta' \cdot f(s_l, x'_b))} \tag{4}$$

Maximizing eq(2) with respect to log-linear feature weights θ is the optimization problem we want to solve.

$$\begin{aligned} \frac{\partial}{\partial \theta_j} L(X, \theta) = \sum_{s_k \rightarrow s_l} A_{kl}(f_j(s_k, s_l) - \sum_{s'_l} f_j(s_k, s'_l) \cdot a(s_l \mid a_k)) + \\ \sum_{s_l \rightarrow x_b} E_{lb}(f_j(s_k, s_l) - \sum_{x'_b} f_j(s_k, s'_l) \cdot e(x'_b \mid s_l)) \end{aligned} \tag{5}$$

3 Resources

We plan to apply our method to 3 tasks - word alignment , POS tagging and word segmentation. For the first 2 tasks we have access to large amounts of parallel sentences from the newscrawl corpus. Word segmentation is a preprocessing task needed when dealing with languages like Chinese. We plan to use PKU corpus to train and evaluate our performance on this task.

4 Milestones

4.1 Must achieve

As we have mentioned at the beginning of the proposal, we would like to implement a featurized version of baum-welch algorithm. We will incorporate the features that emulate IBM-Model1 word alignment and HMM word alignment and run word alignment test to prove that this implementation is correct. At this stage our implementation should give equivalent results compared to the baseline HMM word alignment model.

4.2 Expected to achieve

We would like to make use of the property of featurized model and incorporate more features that go beyond HMM word alignment. In this way we are able to achieve results somewhat higher than the baseline.

4.3 Would like to achieve

Beyond what we expect to achieve we would also like to explore the performance of this model on Chinese word segmentation and POS tagging.

5 Final Writeup

We would like to have have results showing the performance differences between Baum-Welch training and Featurized training described above. We would also like to explore different feature functions for each of the tasks and compare them. We hope to show accuracy gains, at the cost of training time. This is primarily because log-linear modeling requires a normalization term summing over all possible decisions. In our tasks this could be a very large space.

6 Bibliography

References

[1] Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. Association for Computational Linguistics, 2010.