Domain Adaptation of Maximum Entropy Language Models

Tanel Alumäe[a], Mikko Kurimo
Adaptive Informatics Research Centre, School of Science and Technology
Aalto University, Helsinki

[a]Currently with Tallinn University of Technology, Estonia

Summary

In large vocabulary speech recognition, a language model (LM) is often estimated from large amounts of written text data (e.g., newspapers) and small amount of human-transcribed speech data. How to build the best model from such combination of out-of-domain and in-domain training data?
We investigate a recently proposed Bayesian adaptation approach [1, 2] for adapting a conditional **maximum entropy** (MaxEnt) LM to a new domain, given a large corpus of out-of-domain training data and a small corpus of in-domain data. Experiments show that the method consistently outperforms linear interpolation which is typically used in such cases.

Maximum Entropy LM

A conditional MaxEnt model has the following form:

$$P(x|h) = \frac{e^{\sum_i \theta_i f_i(w,h)}}{\sum_{x'} e^{\sum_j \theta_j f_j(x',h)}}$$

where

- x is a word
- h is a context (the word history)
- x' is a set of all possible words
- f_i are (typically binary) feature functions (in our case, N-gram features)
- feature weights θ are learned via gradient descent
- log conditional likelihood of the data $\mathscr{L}(X;\theta)$ is maximized
- smoothing by adding a zero-mean Gaussian prior:

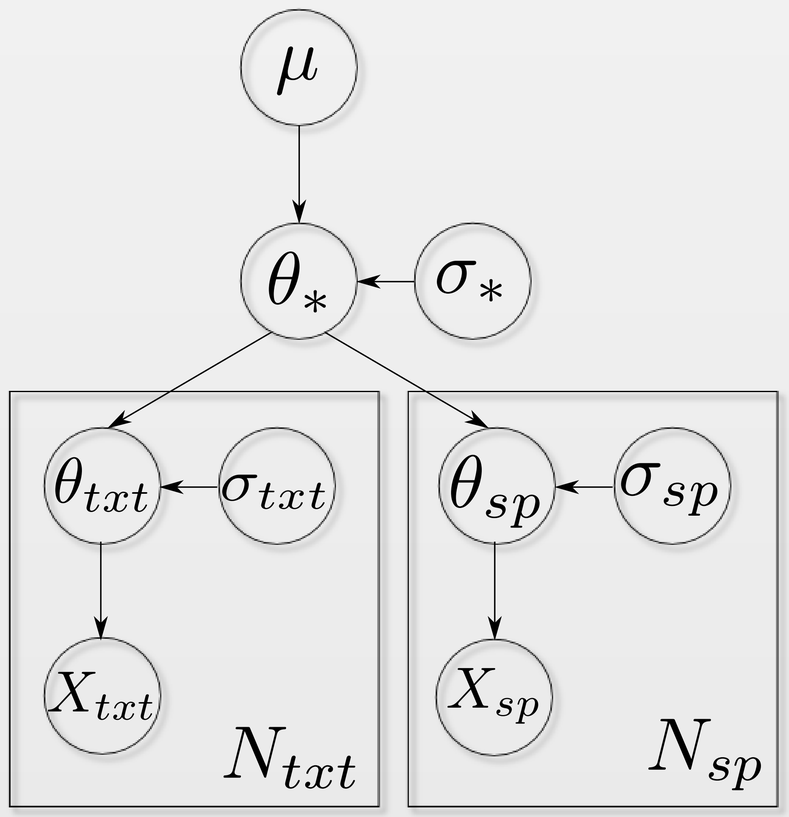
$$\arg\max_\theta \left(\mathscr{L}(X;\theta) - \sum_{i=1}^{F} \frac{\theta_i^2}{2\sigma_i^2} \right)$$

- fixed hyperparameter $\sigma_i = \sigma$
- optimization encourages feature weights with small absolute values

References

[1] H. Daume III. Frustratingly easy domain adaptation. In *Proceedings of ACL*, 2007.
[2] J. R. Finkel and C. Manning. Hierarchical Bayesian domain adaptation. In *Proceedings of HLT-NAACL*, 2009.

Adaptation



Adaptation is done using a hierarchical model [1, 2]:

- **jointly** optimizes global parameters θ_*, out-of-domain parameters θ_{txt} characterizing textual data X_{txt} and in-domain parameters θ_{sp} for speech data X_{sp}
- Gaussian prior μ (usually $\mu = 0$) is applied for global parameters (with variance σ_*)
- global parameters are used as **priors** for domain-specific parameters (with variance σ_{txt} and σ_{sp})
- instead of using smoothing to encourage parameters to be closer to zero, it encourages domain-specific model parameters to be **closer to the corresponding global parameters**
- global and domain-specific parameters are learned jointly. This allows domain-specific parameters to influence the global parameters, and vice versa.

Formally, the joint optimization criteria becomes:

$$\arg\max_\theta \left[\sum_{d \in \{txt,sp\}} \left(\mathscr{L}(X_d,\theta_d) - \sum_{i=1}^{F} \frac{(\theta_{d,i} - \theta_{*,i})^2}{2\sigma_d^2} \right) - \sum_{i=1}^{F} \frac{\theta_{*,i}^2}{2\sigma_*^2} \right]$$

Experiments

We experimented with two speech recognition tasks:

- **English Broadcast News** (2003 NIST Rich Transcription Evaluation Data). For training LM, we used 5M sentences from the Gigaword (2nd ed.) corpus (99.5M words, out-of-domain), and broadcast news transcriptions from the TDT4 corpus (1.19M words, in-domain).
- **Estonian Broadcast Conversations** (40 minutes of live talk programs from Estonian radio). We trained a morpheme-based LM from two sources: about 10M sentences from newspapers (185M morphemes), and transcriptions of 10 hours of radio live talk programs (104K morphemes).

We built Kneser-Ney smoothed trigram models and MaxEnt models with trigram features and used them to evaluate test set perplexity and to rescore N-best lists. Results:

	Perplexity				**WER**		
Adaptation data (No. of words)	Pooled N-gram	Interpol. N-gram	Interpol. ME	Adapted ME	Interpol. N-gram	Interpol. ME	Adapted ME
English Broadcast News							
147K	290	255	243	**230**	27.2	26.3	**25.9**
292K	286	250	236	**223**	26.7	25.8	**25.6**
591K	280	243	228	**215**	26.6	25.9	**25.6**
1119K	272	232	217	**204**	26.2	25.6	**24.9**
Estonian Broadcast Conversations							
104K	237	197	200	**169**	40.5	38.9	**37.4**

Observations and conclusion

- Adaptation using hierarchical models outperforms linear interpolation
- With adapted models, the same performance can be obtained using **50-75% less adaptation data** than with interpolated ME models
- Choice of hyperparameters σ_{txt} and σ_{sp} have a strong impact on the performance of the hierarchical model
- Hierarchical model not limited to simple two-level hierarchies: it is possible to add additional levels, or even use multiple parents or a grid structure (e.g., build models adapted to both style and topic).
- Shortcomings: difficult to find good hyperparameters, training requires more memory