Domain Adaptation of Maximum Entropy Language Models

Tanel Alumäe[*]
Adaptive Informatics Research Centre
School of Science and Technology
Aalto University
Helsinki, Finland
tanel@cis.hut.fi

Mikko Kurimo
Adaptive Informatics Research Centre
School of Science and Technology
Aalto University
Helsinki, Finland
Mikko.Kurimo@tkk.fi

Abstract

We investigate a recently proposed Bayesian adaptation method for building style-adapted maximum entropy language models for speech recognition, given a large corpus of written language data and a small corpus of speech transcripts. Experiments show that the method consistently outperforms linear interpolation which is typically used in such cases.

1 Introduction

In large vocabulary speech recognition, a language model (LM) is typically estimated from large amounts of written text data. However, recognition is typically applied to speech that is stylistically different from written language. For example, in an often-tried setting, speech recognition is applied to broadcast news, that includes introductory segments, conversations and spontaneous interviews. To decrease the mismatch between training and test data, often a small amount of speech data is human-transcribed. A LM is then built by interpolating the models estimated from large corpus of written language and the small corpus of transcribed data. However, in practice, different models might be of different importance depending on the word context. Global interpolation doesn't take such variability into account and all predictions are weighted across models identically, regardless of the context.

In this paper we investigate a recently proposed Bayesian adaptation approach (Daume III, 2007; Finkel and Manning, 2009) for adapting a conditional maximum entropy (ME) LM (Rosenfeld, 1996) to a new domain, given a large corpus of out-of-domain training data and a small corpus of in-domain data. The main contribution of this paper is that we show how the suggested hierarchical adaptation can be used with suitable priors and combined with the class-based speedup technique (Goodman, 2001) to adapt ME LMs in large-vocabulary speech recognition when the amount of target data is small. The results outperform the conventional linear interpolation of background and target models in both N-grams and ME models. It seems that with the adapted ME models, the same recognition accuracy for the target evaluation data can be obtained with 50% less adaptation data than in interpolated ME models.

2 Review of Conditional Maximum Entropy Language Models

Maximum entropy (ME) modeling is a framework that has been used in a wide area of natural language processing (NLP) tasks. A conditional ME model has the following form:

$$P(x|h) = \frac{e^{\sum_i \lambda_i f_i(x,h)}}{\sum_{x'} e^{\sum_j \lambda_j f_j(x',h)}} \qquad (1)$$

where x is an outcome (in case of a LM, a word), h is a context (the word history), and x' a set of all possible outcomes (words). The functions f_i are (typically binary) feature functions. During ME training, the optimal weights λ_i corresponding to features $f_i(x,h)$ are learned. More precisely, finding the ME model is equal to finding weights that maximize the log-likelihood $L(X;\Lambda)$ of the training data X. The weights are learned via improved iterative scaling algorithm or some of its modern fast counterparts (i.e., conjugate gradient descent).

Since LMs typically have a vocabulary of tens of thousands of words, the use of a normalization factor over all possible outcomes makes estimating a ME LM very memory and time consuming. Goodman (2001) proposed a class-based method that drastically reduces the resource requirements for training such models. The idea is to cluster

[*] Currently with Tallinn University of Technology, Estonia

Proceedings of the ACL 2010 Conference Short Papers, pages 301–306,
Uppsala, Sweden, 11-16 July 2010.

words in the vocabulary into classes (e.g., based on their distributional similarity). Then, we can decompose the prediction of a word given its history into prediction of its class given the history, and prediction of the word given the history and its class :

$$P(w|h) = P(C(w)|h) \cdot P(w|h, C(w)) \qquad (2)$$

Using such decomposition, we can create two ME models: one corresponding to $P(C(w)|h)$ and the other corresponding to $P(w|h, C(w))$. It is easy to see that computing the normalization factor of the first component model now requires only looping over all classes. It turns out that normalizing the second model is also easier: for a context $h, C(w)$, we only need to normalize over words that belong to class $C(w)$, since other words cannot occur in this context. This decomposition can be further extended by using hierarchical classes.

To avoid overfitting, ME models are usually smoothed (regularized). The most widely used smoothing method for ME LMs is Gaussian priors (Chen and Rosenfeld, 2000): a zero-mean prior with a given variance is added to all feature weights, and the model optimization criteria becomes:

$$L'(X; \Lambda) = L(X; \Lambda) - \sum_{i=1}^{F} \frac{\lambda_i^2}{2\sigma_i^2} \qquad (3)$$

where F is the number of feature functions. Typically, a fixed hyperparameter $\sigma_i = \sigma$ is used for all parameters. The optimal variance is usually estimated on a development set. Intuitively, this method encourages feature weights to be smaller, by penalizing weights with big absolute values.

3 Domain Adaptation of Maximum Entropy Models

Recently, a hierarchical Bayesian adaptation method was proposed that can be applied to a large family of discriminative learning tasks (such as ME models, SVMs) (Daume III, 2007; Finkel and Manning, 2009). In NLP problems, data often comes from different sources (e.g., newspapers, web, textbooks, speech transcriptions). There are three classic approaches for building models from multiple sources. We can pool all training data and estimate a single model, and apply it for all tasks. The second approach is to "unpool" the data, i.e, only use training data from the test domain. The third and often the best performing approach is to train separate models for each data source, apply them to test data and interpolate the results.

The hierarchical Bayesian adaptation method is a generalization of the three approaches described above. The hierarchical model jointly optimizes global and domain-specific parameters, using parameters built from pooled data as priors for domain-specific parameters. In other words, 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, while a zero mean Gaussian prior is still applied for global parameters. For processing test data during runtime, the domain-specific model is applied. Intuitively, this approach can be described as follows: the domain-specific parameters are largely determined by global data, unless there is good domain-specific evidence that they should be different. The key to this approach is that the global and domain-specific parameters are learned jointly, not hierarchically. This allows domain-specific parameters to influence the global parameters, and vice versa. Formally, the joint optimization criteria becomes:

$$L_{hier}(X; \Lambda) = \sum_d \left(L_{orig}(X_d, \Lambda_d) - \sum_{i=1}^{F} \frac{(\lambda_{d,i} - \lambda_{*,i})^2}{2\sigma_d^2} \right) - \sum_{i=1}^{F} \frac{\lambda_{*,i}^2}{2\sigma_*^2} \qquad (4)$$

where X_d is data for domain d, $\lambda_{*,i}$ the global parameters, $\lambda_{d,i}$ the domain-specific parameters, σ_*^2 the global variance and σ_d^2 the domain-specific variances. The global and domain-specific variances are optimized on the heldout data. Usually, larger values are used for global parameters and for domains with more data, while for domains with less data, the variance is typically set to be smaller, encouraging the domain-specific parameters to be closer to global values.

This adaptation scheme is very similar to the approaches proposed by (Chelba and Acero, 2006) and (Chen, 2009b): both use a model estimated from background data as a prior when learning a model from in-domain data. The main difference is the fact that in this method, the models are estimated jointly while in the other works, background model has to be estimated before learning the in-domain model.

4 Experiments

In this section, we look at experimental results over two speech recognition tasks.

4.1 Tasks

Task 1: English Broadcast News. This recognition task consists of the English broadcast news section of the 2003 NIST Rich Transcription Evaluation Data. The data includes six news recordings from six different sources with a total length of 176 minutes.

As acoustic models, the CMU Sphinx open source triphone HUB4 models for wideband (16kHz) speech[1] were used. The models have been trained using 140 hours of speech.

For training the LMs, two sources were used: first 5M sentences from the Gigaword (2nd ed.) corpus (99.5M words), and broadcast news transcriptions from the TDT4 corpus (1.19M words). The latter was treated as in-domain data in the adaptation experiments. A vocabulary of 26K words was used. It is a subset of a bigger 60K vocabulary, and only includes words that occurred in the training data. The OOV rate against the test set was 2.4%.

The audio used for testing was segmented into parts of up to 20 seconds in length. Speaker diarization was applied using the LIUM_SpkDiarization toolkit (Deléglise et al., 2005). The CMU Sphinx 3.7 was used for decoding. A three-pass recognition strategy was applied: the first pass recognition hypotheses were used for calculating MLLR-adapted models for each speaker. In the second pass, the adapted acoustic models were used for generating a 5000-best list of hypotheses for each segment. In the third pass, the ME LM was used to re-rank the hypotheses and select the best one. During decoding, a trigram LM model was used. The trigram model was an interpolation of source-specific models which were estimated using Kneser-Ney discounting.

Task 2: Estonian Broadcast Conversations. The second recognition task consists of four recordings from different live talk programs from three Estonian radio stations. Their format consists of hosts and invited guests, spontaneously discussing current affairs. There are 40 minutes of transcriptions, with 11 different speakers.

The acoustic models were trained on various wideband Estonian speech corpora: the BABEL speech database (9h), transcriptions of Estonian broadcast news (7.5h) and transcriptions of radio live talk programs (10h). The models are triphone HMMs, using MFCC features.

For training the LMs, two sources were used: about 10M sentences from various Estonian newspapers, and manual transcriptions of 10 hours of live talk programs from three Estonian radio stations. The latter is identical in style to the test data, although it originates from a different time period and covers a wider variety of programs, and was treated as in-domain data.

As Estonian is a highly inflective language, morphemes are used as basic units in the LM. We use a morphological analyzer (Kaalep and Vaino, 2001) for splitting the words into morphemes. After such processing, the newspaper corpus includes of 185M tokens, and the transcribed data 104K tokens. A vocabulary of 30K tokens was used for this task, with an OOV rate of 1.7% against the test data. After recognition, morphemes were concatenated back to words.

As with English data, a three-pass recognition strategy involving MLLR adaptation was applied.

4.2 Results

For both tasks, we rescored the N-best lists in two different ways: (1) using linear interpolation of source-specific ME models and (2) using hierarchically domain-adapted ME model (as described in previous chapter). The English ME models had a three-level and Estonian models a four-level class hierarchy. The classes were derived using the word exchange algorithm (Kneser and Ney, 1993). The number of classes at each level was determined experimentally so as to optimize the resource requirements for training ME models (specifically, the number of classes was 150, 1000 and 5000 for the English models and 20, 150, 1000 and 6000 for the Estonian models). We used unigram, bigram and trigram features that occurred at least twice in the training data. The feature cut-off was applied in order to accommodate the memory requirements. The feature set was identical for interpolated and adapted models.

[1] http://www.speech.cs.cmu.edu/sphinx/models/

	Interp. models		Adapted models		
Adaptation data (No of words)	σ^2_{OD}	σ^2_{ID}	σ^2_{*}	σ^2_{OD}	σ^2_{ID}
English Broadcast News					
147K	$2e8$	$3e5$	$5e7$	$2e7$	$2e6$
292K	$2e8$	$5e5$	$5e7$	$2e7$	$2e6$
591K	$2e8$	$1e6$	$5e7$	$2e7$	$2e6$
1119K	$2e8$	$2e6$	$5e7$	$2e7$	$5e6$
Estonian Broadcast Conversations					
104K	$5e8$	$3e5$	$5e7$	$1e7$	$2e6$

Table 1: The unnormalized values of Gaussian prior variances for interpolated out-of-domain (OD) and in-domain (ID) ME models, and hierarchically adapted global (*), out-of-odomain (OD) and in-domain (ID) models that were used in the experiments.

For the English task, we also explored the efficiency of these two approaches with varying size of adaptation data: we repeated the experiments when using one eighth, one quarter, half and all of the TDT4 transcription data for interpolation/adaptation. The amount of used Gigaword data was not changed. In all cases, interpolation weights were re-optimized and new Gaussian variance values were heuristically determined.

The TADM toolkit[2] was used for estimating ME models, utilizing its implementation of the conjugate gradient algorithm.

The models were regularized using Gaussian priors. The variance parameters were chosen heuristically based on light tuning on development set perplexity. For the source-specific ME models, the variance was fixed on per-model basis. For the adapted model, that jointly models global and domain-specific data, the Gaussian priors were fixed for each hierarchy node (i.e., the variance was fixed across global, out-of-domain, and in-domain parameters). Table 1 lists values for the variances of Gaussian priors (as in equations 3 and 4) that we used in the experiments. In other publications, the variance values are often normalized to the size of the data. We chose not to normalize the values, since in the hierarchical adaptation scheme, also data from other domains have impact on the learned model parameters, thus it's not possible to simply normalize the variances.

The experimental results are presented in Table 2. Perplexity and word error rate (WER) results of the interpolated and adapted models are compared. For the Estonian task, letter error rate (LER) is also reported, since it tends to be a more indicative measure of speech recognition quality for highly inflected languages. In all experiments, using the adapted models resulted in lower perplexity and lower error rate. Improvements in the English experiment were less evident than in the Estonian system, with under 10% improvement in perplexity and 1-3% in WER, against 15% and 4% for the Estonian experiment. In most cases, there was a significant improvement in WER when using the adapted ME model (according to the Wilcoxon test), with and exception of the English experiments on the 292K and 591K data sets.

The comparison between N-gram models and ME models is not entirely fair since ME models are actually class-based. Such transformation introduces additional smoothing into the model and can improve model perplexity, as also noticed by Goodman (2001).

5 Discussion

In this paper we have tested a hierarchical adaptation method (Daume III, 2007; Finkel and Manning, 2009) on building style-adapted LMs for speech recognition. We showed that the method achieves consistently lower error rates than when using linear interpolation which is typically used in such scenarios.

The tested method is ideally suited for language modeling in speech recognition: we almost always have access to large amounts of data from written sources but commonly the speech to be recognized is stylistically noticeably different. The hierarchical adaptation method enables to use even a small amount of in-domain data to modify the parameters estimated from out-of-domain data, if there is enough evidence.

As Finkel and Manning (2009) point out, the hierarchical nature of the method makes it possible to estimate highly specific models: we could draw style-specific models from general high-level priors, and topic-and-style specific models from style-specific priors. Furthermore, the models don't have to be hierarchical: it is easy to generalize the method to general multilevel approach where a model is drawn from multiple priors. For

[2] http://tadm.sourceforge.net/

	Perplexity				WER			LER		
Adaptation data (No. of words)	Pooled N-gram	Interp. N-gram	Interp. ME	Adapted ME	Interp. N-gram	Interp. ME	Adapted ME	Interp. N-gram	Interp. 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	17.7	17.3	16.6

Table 2: Perplexity, WER and LER results comparing pooled and interpolated N-gram models and interpolated and adapted ME models, with changing amount of available in-domain data.

instance, we could build a model for recognizing computer science lectures, given data from textbooks, including those about computer science, and transcripts of lectures on various topics (which don't even need to include lectures about computer science).

The method has some considerable shortcomings from the practical perspective. First, training ME LMs in general has much higher resource requirements than training N-gram models which are typically used in speech recognition. Moreover, training hierarchical ME models requires even more memory than training simple ME models, proportional to the number of nodes in the hierarchy. However, it should be possible to alleviate this problem by profiting from the hierarchical nature of n-gram features, as proposed in (Wu and Khudanpur, 2002). It is also difficult to determine good variance values σ_i^2 for the global and domain-specific priors. While good variance values for simple ME models can be chosen quite reliably based on the size of the training data (Chen, 2009a), we have found that it is more demanding to find good hyperparameters for hierarchical models since weights for the same feature in different nodes in the hierarchy are all related to each other. We plan to investigate this problem in the future since the choice of hyperparameters has a strong impact on the performance of the model.

Acknowledgments

This research was partly funded by the Academy of Finland in the project Adaptive Informatics, by the target-financed theme No. 0322709s06 of the Estonian Ministry of Education and Research and by the National Programme for Estonian Language Technology.

References

Ciprian Chelba and Alex Acero. 2006. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, October.

S. F. Chen and R. Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50.

S. F. Chen. 2009a. Performance prediction for exponential language models. In *Proceedings of HLT-NAACL*, pages 450–458, Boulder, Colorado.

Stanley F. Chen. 2009b. Shrinking exponential language models. In *Proceedings of HLT-NAACL*, pages 468–476, Boulder, Colorado.

H. Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263.

P. Deléglise, Y. Estéve, S. Meignier, and T. Merlin. 2005. The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. In *Proceedings of Interspeech*, Lisboa, Portugal.

J. R. Finkel and Ch. Manning. 2009. Hierarchical Bayesian domain adaptation. In *Proceedings of HLT-NAACL*, pages 602–610, Boulder, Colorado.

J. Goodman. 2001. Classes for fast maximum entropy training. In *Proceedings of ICASSP*, Utah, USA.

H.-J. Kaalep and T. Vaino. 2001. Complete morphological analysis in the linguist's toolbox. In *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pages 9–16, Tartu, Estonia.

R. Kneser and H. Ney. 1993. Improved clustering techniques for class-based statistical language modelling. In *Proceedings of the European Conference* *on Speech Communication and Technology*, pages 973–976.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228.

J. Wu and S. Khudanpur. 2002. Building a topic-dependent maximum entropy model for very large corpora. In *Proceedings of ICASSP*, Orlando, Florida, USA.