A CLASS BASED APPROACH TO DOMAIN ADAPTATION AND CONSTRAINT INTEGRATION FOR EMPIRICAL M-GRAM MODELS

Klaus Ries

kries@ira.uka.de ries+@cs.cmu.edu

Interactive Systems Laboratories

University of Karlsruhe, Karlsruhe, Germany
Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

The first class based adaptation approaches [FGH$^+$97, Ueb97] take the use of classes in the construction of statistical m-gram models one significant step further than just using them as a smoothing technique: The m-gram of classes is trained on the large background corpus while the word likelihoods given the class are estimated on the small target corpus. To make full use of this technique a specialized clusteralgorithm has been developed [FGH$^+$97, Ueb97]. In this paper we extend class adaptation to make use of the m-gram distribution of the target domain. As a second independent contribution this paper introduces an efficient morphing algorithm, that tries to achieve adaptation by using a stochastic mapping of words between the vocabularies of the respective domains.

As a result we can show, that for small adaptation steps class based adaptation is a very useful technique. For larger adaptation steps the perplexity of the modified model is greatly improved, yet no improvement over the unadapted model was observed when used in linear interpolation. Whether this is due to the fact that we use class based adaptation or that we do just modify the unigram distribution is still unresolved, although the new stochastic mapping technique might help to give an answer to this question in the future.

1. INTRODUCTION

The adaptation of statistical mgrams of language to new domains is one of the big challenges in speech recognition. In the past a number of approaches like topic dependent (maximum entropy) models [PPMR92, LS95, RMR95] and linear interpolation have been proposed. Additionally topic mixture models [IO96] and backoff models have been suggested. The research question brought up here is whether there is a way to parameterize m-grams such that a large mgram can be modified with a relatively small number of parameters.

One of the most significant changes from one domain to another is the change of the monogram distribution. In information retrieval an estimate of the monogram distribution (bag of words) is frequently used as the input to the retrieval algorithm. Following this idea maximum entropy models have been proposed, which adapt the monogram distribution of higher order mgram models [PPMR92, LS95, RMR95]. These ideas did not have success on the Switchboard-domain, where [RMR95] shows in a contrastive experiment that a standard linear interpolation scheme works just as good or better.

In [FZM$^+$96] class-based mgrams are used in a linear interpolation setting to make maximal use of all knowledge sources. [FGH$^+$97, Ueb97] estimated the probability $\hat{p}(w|c)$ of word w given the class c is known based on the small corpus, whereas the mgram-model on the classes is estimated on a larger corpus. They showed, that a specialized clustering algorithm for that model is essential to get good performance (section 3). This paper takes this idea one step further and substitutes the above-mentioned estimate $\hat{p}(w|c)$ on the small corpus by an estimate of $\hat{p}(w|c,h)$ from the small corpus (section 2).

Without the modeling assumption of classes one can try to define a stochastic mapping function from one corpus to the other and estimate this mapping function directly without resorting to classes (section 4). Since, in contrast to the classical approach in statistical machine translation as exemplified as [BCP$^+$90], no alignment between the two corpora exists, the model is trained to estimate statistical coocurrence properties of the target domain using cooccurence statistics from the source domain. Some attempts to identify translation properties without alignments just based on coocurrence properties in bilingual corpora take a similar approach [Rei95, Fun95]. In our case the statistical property used is the bigram distribution, specifically we will define the stochastic mapping such that the image of $p(w_1, w_2)$ models the bigram distribution of the target corpus optimally.

The paper proceeds illuding the class based adaptation itself (section 2), presenting the word classification procedure (section 3) and the training of the stochastic mapping (section 4). The paper continues with experiments (Section 5) which are followed by the conclusions (section 6).

2. CLASS BASED ADAPTATION AND CONSTRAINT INTEGRATION

So far the most widely used adaptation technique is linear interpolation, yet the strength of the adaptation for a certain unit, be it syntactic, semantically or automatically defined word classes cannot be set globally, since we can

This research was performed at the University of Karlsruhe and in part at Carnegie Mellon University, Pittsburgh. The author of this paper was supported by the BMFT in the Verb-Mobil project. We gratefully acknowledge their support. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any organization mentioned above.

only change the strength of the adaptation for a certain context or history, not for the word that we are actually predicting. So if we say want to adapt heavily on certain content words and not so much on function words there is really no way of specifying that, since many contexts can be followed by both types of words.

The general direction proposed here is to change the trigram distribution of the background corpus in a manner, that preserves the essential properties of the trigrams of that corpus yet moves the m-gram distribution closer to the one found on the target corpus. Technically speaking we propose to choose the adapted model of the background corpus to be

$$\sum_{c} p(c_t|h) \cdot \hat{p}(w_t|c_t, \hat{h})$$

where c_t is the class of the word w_t. p is representing a distribution derived on the background corpus, $\hat{p}$ resp. on the target corpus. This model is preserving $p(c_t|h)$ of the background corpus in the adapted model yet within this constraint makes maximum use of the distribution $\hat{p}$ of the target corpus.

We will restrict ourselves to word classifications that partition the vocabulary and to specific choices of h and $\hat{h}$ such that the model can be stated as below. Please note that for $\hat{m} = 1$ we arrive at the model of the original work [FGH+97, Ueb97].

$$p(c_t|c_{t-1}, \ldots, c_{t-m+1}) \cdot \hat{p}(w_t|c_t, w_{t-1}, \ldots, w_{t-\hat{m}+1})$$

The partitioning of the words into classes as well as the size of the m-gram model of the corpus are ways of choosing the adaptation scheme. This way of combining constraints from different sources hasn't been used widely in the speech community, yet it could proof vital in similar settings as well such as integrating long distance constraints into an m-gram model.

If we have, for example, a shift in the domain and of two synonymous/similar distributed words one is used more often after the shift this model is adapting well. However the standard cluster algorithm itself is blind towards such generalizations and might well fail to find classes that facilitate this adaptation (see section 3). The grainsize of the classes is therefore indicative of the adaptation strength of this model.

If the classes are getting too coarse the predictive power of the model might get too low in the case of the original work [FGH+97, Ueb97]. By using higher order distributions of the target domain directly one can try to overcome the natural problem, that for coarse classes very little contextual information from the background corpus is used. Intuitively we try to balance between these two extreme cases:

one class the m-gram model of the target domain is used

all words in seperate classes standard m-gram model of the background corpus is used

3. ADAPTIVE CLUSTERING

For the purpose of adaptive clustering we consider only bigrams and clusters that partition the vocabulary. These restrictions are also commonly used in most practical approaches to word classification. As a common notion we will adopt that distributions or counts with a hat like $\hat{p}$ are on the target corpus, others like p are on the background corpus. To denote the first word resp. class in a bigram we write w_1 resp. c_1, to denote the second position we write w_2 resp. c_2. The log-likelihood of the standard class based model on testdata can be written as

$$L = \prod_{w_1, w_2} \log(p(w_1|c_1) \cdot p(c_1|c_2))^{\hat{N}(w_1, w_2)}$$

In the original [FGH+97, Ueb97] adapted scheme $p(w_1|c_1)$ from [KN93] was replaced by $\hat{p}(w_1|c_1)$, the new scheme proposed here replaces this by $\hat{p}(w_1|c_1, w_2)$ and the classes are found such that the likelihood is optimized on the testset. [KN93, FGH+97, Ueb97] use leaving one out estimates to accout for the fact, that the target corpus is used to estimate $\hat{p}$ from $\hat{N}(w_1, w_2)$. In principle it would not be necessary to make a leaving one out estimate for p, yet we decide to do so since in some situations we also mix the target and the background corpus [Ueb97].

The description of the algorithm will be done only for the [FGH+97, Ueb97] criterion. The new criterion using $\hat{p}(w_1|c_1, w_2)$ can be constructed fairly similar, though the presentation is slightly more complicated. Note that for the new scheme

$$\hat{p}(w_1|c_1, w_2) = \frac{\hat{p}(w_1, c_1, w_2)}{\hat{p}(c_1, w_2)} = \frac{\hat{p}(w_1, w_2)}{\hat{p}(c_1, w_2)}$$

holds and that $\hat{p}(w_1, w_2)$ can be eliminated if one wants to maximize L to find the classes. Taking logarithms, inserting maximum likelihood estimates and rearranging terms we get:

$$\begin{aligned} F'_{ML} &= \sum_{c_1, c_2} \hat{N}(c_1, c_2) \cdot \log(N(c_1, c_2)) \\ &- \sum_{c} \hat{N}(c) \cdot \log N(c) \\ &+ \sum_{w} \hat{N}(w) \cdot \log N(w) - \sum_{c} \hat{N}(c) \cdot \log \hat{N}(c) \end{aligned}$$

The term $\sum_w \hat{N}(w) \cdot \log N(w)$ can be dropped since it doesn't change when the classification is changed. Rearranging terms we get the criterion:

$$\begin{aligned} F_{ML} &= \sum_{c_1, c_2} \hat{N}(c_1, c_2) \cdot \log(N(c_1, c_2)) \\ &- \sum_{c} \hat{N}(c) \cdot (\log N(c) + \log \hat{N}(c)) \end{aligned}$$

Applying the leaving-one-out criterion the final criterion is received similar to [KN93]:

$$\begin{aligned} F_{ML}^{LOO} = \sum_{c_1, c_2, N(c_1, c_2) > 1} & \hat{N}(c_1, c_2) \cdot \log(N(c_1, c_2) - 1 - b) \\ & + \hat{n}_1 \cdot \log \frac{(n_+ - 1)b}{n_0 + 1} \\ & - \sum_{c} \hat{N}(c) \cdot (\log(N(c) - 1) + \log(\hat{N}(c) - 1)) \end{aligned}$$

where b is an absolute discounting factor, n_1 is the number of bigrams occuring exactly once in the large corpus, n_+ is the number of bigrams occuring at least once in the large corpus, n_0 is the number of bigrams not occuring in the large corpus and $\hat{n}_1 = \sum_{c_1, c_2, N(c_1, c_2) \leq 1} \hat{N}(c_1, c_2)$. This criterion corresponds exactly to [KN93] if the original corpus and the corpus that is being adapted on are identical. This criterion was integrated straightforward in the current fast implementation of [KN93], the criterion based on $\hat{p}(w_1|c_1, w_2)$ needed some more significant changes to the code.

4. DIRECT STOCHASTIC MAPPING

Assuming that we have a well estimated bigram probability distribution $p(w_1', w_2')$ on a background corpus we want to estimate a mapping $p(w_1|w_1')$ from the vocabulary of the background corpus to the vocabulary of the target corpus. In the sequel we will denote words on the target corpus without a prime, word like w' on the background corpus with a prime. The optimality criterion is, how well a bigramtable $c(w_1, w_2)$ of the target corpus is modeled by the resulting probability distribution $p(w_1, w_2) = \sum_{w_1', w_2'} p(w_1|w_1')p(w_2|w_2')p(w_1', w_2')$. We restate this problem as minimizing

$$L = \sum_{w_1,w_2} c(w_1, w_2) \cdot \log \sum_{w_1',w_2'} p(w_1|w_1')p(w_2|w_2')p(w_1', w_2')$$

We will freely write and make intuitive use of marginals and conditionals of the distribution

$$p(w_1, w_1', w_2, w_2') = p(w_1|w_1') \cdot p(w_2|w_2') \cdot p(w_1', w_2')$$

Within the EM framework the criterion of optimizing L can be reformulated as to find incremental estimates of $p^t(w_1|w_1')$ that obey $\forall w_1' : \sum_{w_1} p^t(w_1|w_1') = 1$, minimizing in each step

$$\sum_{w_1,w_2,w_1',w_2'} \begin{array}{l} p^{t-1}(w_1', w_2'|w_1, w_2) \cdot c(w_1, w_2) \\ \qquad \cdot \;\; \log\; p^t(w_1|w_1') \cdot p^t(w_2|w_2') \end{array}$$

Using Lagrange multipliers we derive the update formula, where $\lambda^t(w')$ is chosen to ensure that the distribution sums up to one again

$$\begin{aligned} \frac{\lambda^t(w') \cdot p^t(w|w')}{p^{t-1}(w|w')} &= \\ & \sum_x \frac{c(x,w)}{p(x,w)} \cdot \sum_{z'} p^{t-1}(x|z')p^{t-1}(z', w') \\ + & \sum_x \frac{c(w,x)}{p(w,x)} \cdot \sum_{z'} p^{t-1}(x|z')p^{t-1}(w', z') \end{aligned}$$

If we assume that most $p(w|w')$ are very small we can use the approximation $p(w|w') = \lambda\tilde{p}(w|w') + \frac{(1-\lambda)}{V}$, where V is the vocabularysize of the targetcorpus, λ a global smoothing constant and $\tilde{p}(w|w') \neq 0$ for only a very small number of w for each w'. The update formula shown above entails $p^{t-1}(w|w') = 0 \rightsquigarrow p^t(w|w') = 0$, such that the smoothing is crucial.

In our experiments we have bound the maximal number of possible $\tilde{p} \neq 0$ to 10, a number much smaller then the vocabularysize that will be ignored in calculating the complexity of the algorithm. Even though we are using the smoothing mentioned above we can still for the most part ignore those $p(w|w')$ in the summations, where $\tilde{p}(w|w') = 0$. Checking the equations above we can then see, that calculating $p(w_1, w_2)$ is essentially possible in $O(V^3)$ and that for all w' the whole table $c(\cdot, \cdot)$ needs to be visited, which contains at max $O(V^2)$ entries and therefore each iteration has $O(V^3)$ steps. This EM algorithm can also be applied incrementally, meaning that we calculate new parameters for one $p(\cdot|w')$ and then update the model immediately. We are also calculating $p(w_1, w_2)$ incrementally since the implicit hidden constant of that calculation is fairly large otherwise.

5. EXPERIMENTS

In the original experiments [FGH+97] on the VerbMobil domain the only adaptation that needed to be done was a small change in the register of the language. It was demonstrated, that significant improvements over the straightforward and well know linear interpolation and over standard class based schemes are possible. These new experiments try to achieve a much higher goal, namely to make use of a large corpus of news shows (Broadcast News or BN) to improve the modeling of people carrying out spontaneous conversations over the phone about 70 topics, mostly with some political or social topic (Switchboard or SWB). First interpolation experiments indicated that the use of BN was far more effective than newspaper text, that was used previously.

For the class based adaptation we used three different partitions of the vocabulary in appr. 500 classes, one was derived with the standard word classification procedures (stand), one was derived using the adaptive word classification algorithm (adapt) [FGH+97, Ueb97]. For the class based adaptation we used both just $\hat{p}$ being a unigram-distribution (uni) as well as $\hat{p}$ being a trigram distribution (tri-adapt), p is always a trigram. If more than one model was used a context independent linear interpolation of the resulting models was used, where the weights were optimized on held out data of SWB. The perplexity is evaluated on a separate testset from SWB. Additional to that we used the proposal by [Ueb97] to use linear interpolation in the class-finding process.

	SWB		Broadcast News				
	direct	stand	direct	stand		adapt	
				uni	tri	uni	tri
PP	77	81	278	213	206	123	106
	Linear interpolated models						
73	●	●					
69	●	●	●				
69	●	●	●	●	●		
69	●	●				●	
69	●	●					●

These experiments seem to indicate, that class based adaptation, in contrast to the application to small shifts, seem to not apply towards adapting larger shifts. From the table we can see, that even tough the perplexity reductions before the interpolation are very substantial that this is not translating to a better performance after the interpolation.

To test our original hypothesis further we also modified the original adaptive clustering to make use of the bigram distribution of the target corpus instead of the unigram distribution as suggested in section 3 and also tested a number of different classsizes. In another experiment we tried to adapt the Switchboard corpus to the very similar American English CallHome corpus (CH), which consist of telephone conversation of relatives or close friends. Both experiments delivered the same pattern as show above, namely that straight interpolation is delivering about the same results. For the adaptation of SWB to CH we also tried word accuracy measurements with similar negative results. Eyeballing the classes generated by the different algorithms did not reveal any specific information about the reason for this behavior nor is it possible to eyeball the difference between normal classes and adaptively derived classes.

The experiments with the direct stochastic mapping technique are still very much in the beginning, but the results can be interpreted better by humans than the classes generated by class based adaptation. Specifically useful is an ordered list of the frequency, at which a word occurs as a target of a mapping and the following observations were made:

Adaptation on SWB to SWB Two types of mappings were discovered, that pointed to problems in the current use of the algorithm: The end-of-sentence symbol was mapped onto the begin-of-sentence symbol, which was due to an asymmtery in the estimation of the distribution. Additionally mappings onto KNOW was found. This word is fairly frequent yet almost solely occurs in the phrase YOU KNOW, so a well known misestimation of the underlying backoff-model was discovered.

Adaptating SWB to CH Additionally to the problems in selfadaptation some ideosyncratic properties of the CH database are reflected like citynames in foreign countries, travel related terms and terminology related to family affairs and "keeping in touch".

Adapting BN to SWB Two effects are dominant in the target of the mapping: Spontaneous speech disfluencies, backchannels etc. are topping the list, followed by letter sequences that have accidentially not been introduced in the BN corpus. Additional to that some effects of the topic specifity of the SWB corpus can be observed and as well as some placenames and some terms reflecting on hobbies, social activities etc.

6. CONCLUSION AND FUTURE RESEARCH

The adaptation of the monogram is an important step in the adaptation process of mgram models to small changes in a given domain, which was already demonstrated earlier [FGH$^+$97]. The question addressed by this paper was, whether the change in the unigram distribution is also an important matter when adapting from one domain to another. Though significant new developments have been made, the models themself do adapt and the new techniques add performance they do not add performance to a model that utilizes straight linear interpolation in the new adaptation scenario. One assumption could be, that the underlying class model is not appropriate in this case, since the adaptation forces the classes to be too small. We have therefore proposed a new EM algorithm, that can learn stochastic mappings from one vocabulary to another. This algorithm needs further testing and development, yet it seems to be a good tool to get some interpretable results about the data.

This result should not necessary lead to discouragement using this method, since most adaptation method do not even try to attempt across such different sets of data. The result should therefore be rather interpreted as indicating the range of applications that can be tackled with these methods. Specifically the improved methods have not been tested on the original VerbMobil domain and one could hope for improvements applying them in modeling the VerbMobil shift. Additionally the class based approach as shown here could also be used to adapt models within one domain dynamically. We have tried to use a variant of cache models on the Switchboard corpus based on that idea without success, yet one has to remark that cache models have not lead to success in modeling this corpus at all.

Future research will therefore be divided into four areas: First we want to check how well one can adapt within one domain. The second research question is how well the new stochastic mapping procedure works, especially in statistical machine translation or language modeling. The third research question is how to make the stochastic mapping a feasible technology for LVCSR. The fourth and general research question is, whether one can identify some structural knowledge that allows to modify mgram distributions other than retraining all their parameters.

7. REFERENCES

[BCP$^+$90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jellinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[FGH$^+$97] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The Karlsruhe-Verbmobil Speech Recogniton Engine. In *ICASSP*, 1997.

[Fun95] Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *The 3rd Annual Workshop on Very Large Corpora*, pages 173–183, Boston, Massachusettes, June 1995.

[FZM$^+$96] Michael Finke, Torsten Zeppenfeld, Martin Maier, Laura Mayfield, Klaus Ries, Puming Zhan, John Lafferty, and Alex Waibel. Switchboard evaluation report. In *Proceedings of LVCSR Hub 5 Workshop*, April 1996.

[IO96] Rukmini Iyer and Mari Ostendorf. Modeling long distance dependence in language, topic mixtures vs. dynamic cache models. In *ICSLP*, 1996.

[KN93] Reinhard Kneser and Herman Ney. Improved clustering techniques for class-based statistical language modeling. In *Eurospeech*, Berlin, Germany, 1993.

[LS95] John Lafferty and Bernhard Suhm. Efficient iterative scaling of a class of maximum entropy models. In *XV Workshop on Maximum Entropy and Bayesian Methods, Los Alamos*, 1995.

[PPMR92] S. Della Pietra, D. Della Pietra, R. L. Mercer, and S. Roukos. Adaptive language modeling using minimum discrimination information. In *ICASSP*, volume 1, pages 633–637, 1992.

[Rei95] Rapp Reinhard. Identifying word translations in nonparallel texts. In *ACL*, 1995.

[RMR95] P. Srinivasa Rao, Michael D. Momkowski, and Salim Rouskos. Language modeling adaption via minimum discrimination information. In *ICASSP*, 1995.

[Ueb97] Joerg P Ueberla. Domain adaptation with clustered language models. In *ICASSP*, 1997.