ONLINE ADAPTATION FOR LANGUAGE MODELS IN SPOKEN DIALOGUE SYSTEMS

Bernd Souvignier, Andreas Kellner

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
{souvi,kellner}@pfa.research.philips.com

ABSTRACT

The robust estimation of language models for new applications of spoken dialogue systems often suffers from a shortcoming of training material. An alternative to training a language model is to improve an initial language model using material obtained while running the new system, thus adapting it to the new task.

In this paper we investigate different methods for online-adaptation of language models. Apart from the standard techniques of supervised and unsupervised adaptation, we look at two refined approaches: the first allows multiple hypotheses from N-best lists as adaptation material and the second uses confidence measures to exclude unreliably recognized sentences from adaptation.

We apply adaptation to both the language model used by the speech recognizer to focus the beam search and to the stochastic language understanding grammar. It turns out that the understanding grammar can be improved quite significantly using N-best lists or confidence measures, whereas unsupervised adaptation may even result in a deterioration of the system. The language model used by the speech recognizer is improved very satisfactorily by each of the chosen approaches.

1. INTRODUCTION

One of the crucial problems in developing new applications of spoken dialogue systems is the robust estimation of the parameters of the stochastic models. Collecting and transcribing large amounts of training data is both tedious and expensive and may even be impossible in certain applications because of privacy regulations. On the side of acoustic modeling, the benefits of adaptation are well-known, for example in channel or speaker adaptation (see e.g. [9], [11]). For language modeling, a standard approach to adaptation (cf. [3], [5]) is to interpolate a task-independent language model trained on a large background corpus with a task-specific model obtained from little task-specific material (e.g. a cache model). An alternative is exposed in [2], where the background model is used as a fill-up model for the task-specific model. A different approach that is especially powerful in topic adaptation is to adjust only the weights for the interpolation of several (typically topic specific) language models (cf. [3]).

All of the above methods for language model adaptation are well suited for biasing a system towards one of several known situations or to adjust it to a slight variation of a standard application. However, creating a language model for a new spoken dialogue system is a different problem for which these methods are not tailored. An advantage of dialogue systems is that structural information about the application is known a priori. This information (e.g. encoded in a context-free grammar) can be used to create initial language models as is described in [7] or [5]. The point we are addressing in this paper is to improve initial language models by exploiting material obtained during application of the new system.

In Section 2 we describe the set-up in which we apply adaptation to language models. Section 3 exposes how N-best lists and confidence measures can be used to obtain better adaptation material than by pure unsupervised adaptation. In Section 4 we investigate the effects of the different adaptation techniques on the language model used during recognition and on the stochastic language understanding grammar which is applied to interpret the recognition result. Experimental results documenting the improvements achieved with the different methods are displayed.

2. LANGUAGE MODEL SET-UP

The starting point for our investigations are *word graphs* obtained from a state-of-the-art HMM speech recognizer (see [1] for details). Such a word graph consists of arcs labeled by word hypotheses together with their acoustic likelihood and can be seen as a compact representation of multiple sentence hypotheses. We apply two methods to extract the best path from the word graph:

- Rescoring with an n-gram language model (LM) to observe the effects of adaptation on the language model used in the speech recognizer.
- Rescoring with a stochastic context-free grammar (SCFG) to investigate adaptation of the language understanding module (cf. [1, 8] or [6] for detailed information about SCFG in language modeling).

Since the probability $p(w_n|w_1,\dots,w_{n-1})$ of a word history in an n-gram language model is based on m-gram counts ($m \le n$),

This work was partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 701 C9. The responsibility for the contents of this study lies with the authors.

adaptation of such a language model can be performed by adding the m-gram counts from the adaptation material to the m-gram counts of the current model.

In our applications, the SCFG consists of three parts: rules to parse the input and to extract the meaningful phrases (the *concepts*), a filler language model covering the not parseable parts of the input and a concept language model providing probabilities for the concept sequences. Typically, both the filler and the concept language models are n-gram models (where the concept language model's vocabulary is the list of concepts), hence adaptation of these language models can be realized by counting m-grams on the adaptation material. Finally, the rule probabilities are based on counting frequencies of rule applications and can thus also be updated by incrementing the counts of the rules used to parse the adaptation material.

3. MODES OF ADAPTATION

In this section we describe the different adaptation techniques that were investigated. Of course, supervised adaptation is not a real adaptation method, since a transcription of the actually spoken text is required. It serves as a baseline for the maximally achievable improvement of the system. The other extreme is unsupervised adaptation where the understood sentences are used as adaptation material. This is the simplest practical method, but it carries the potential of error reinforcement which may result in a deterioration of the system.

3.1. Adaptation using N-best lists

One of the disadvantages of unsupervised adaptation is that no information about the accuracy of the favoured hypothesis is used. In the context of adaptation of phoneme models an approach using N-best lists has been successfully investigated in [9]. In this section, we will demonstrate how N-best lists can also be exploited for language model adaptation. The motivation is that for correctly understood parts of a sentence it is very likely that these parts occur in most of the sentence hypotheses of an N-best list, whereas for misrecognized words there will usually be various alternatives (often including the correct word). Using the different hypotheses from an N-best list for adaptation will therefore put emphasis on reliably understood phrases and weaken the negative effects of recognition errors. Since weight is likely to be shifted from incorrectly recognized phrases to the actually spoken text, this can be seen as a step from unsupervised towards supervised adaptation.

In analogy to a method used in [10] to obtain confidence measures for semantic items we compute weights for the hypotheses in an N-best lists as follows: Denote by l_i the likelihood of the i-th hypothesis in the N-best list. Using a heuristic scaling factor λ, we define weights ω_i for the hypotheses by

$$\omega_i := \frac{l_i^\lambda}{\sum_{j=1}^N l_j^\lambda} = \left(\sum_{j=1}^N \left(\frac{l_j}{l_i} \right)^\lambda \right)^{-1}.$$

Obviously, the ω_i sum up to 1. The scaling factor λ determines how the weights are distributed over the N-best list. For $\lambda = 1$ the weights ω_i are the likelihoods renormalized such that the N-best list carries the full probability mass. For $\lambda = 0$ every hypothesis in the N-best list has the same weight $1/N$, and for $\lambda \gg 1$ the first-best sentence will accumulate the full weight.

Adaptation of the recognizer's LM and the SCFG is now performed by looping over the hypotheses in the N-best list, adding ω_i to the count of each rule used to parse the i-th hypothesis and to the count of each encountered m-gram in this hypothesis.

A remark about discounting for unseen events seems in place. A standard discounting method for bigram language models using integral counts is absolute discounting with a constant b. If $N(v)$ and $N(v, w)$ denote the counts of v and (v, w), respectively, and q is a less specific distribution, e.g. a unigram distribution, we have:

$$p(w|v) := \begin{cases} \frac{N(v,w)-b}{N(v)} + b \cdot q'(w|v) & \text{if } N(v,w) > 0 \\ b \cdot q'(w|v) & \text{if } N(v,w) = 0 \end{cases}$$

where q' is a rescaling of q such that the distribution p is normalized. An obvious generalization to fractional counts is to interpolate the above distribution linearly for counts between 0 and 1. This leads to:

$$p(w|v) := \begin{cases} \frac{N(v,w)-b}{N(v)} + b \cdot q''(w|v) & \text{if } N(v,w) \geq 1 \\ \frac{(1-b)N(v,w)}{N(v)} + b \cdot q''(w|v) & \text{if } N(v,w) < 1 \end{cases}$$

which replaces absolute discounting by linear discounting for counts between 0 and 1 (again with a suitable rescaling q'' of q).

3.2. Adaptation using confidence measures

Another possibility to avoid error reinforcement is to use confidence measures to exclude unreliably understood sentences from the adaptation material. One such approach is chosen in [4] where a recognition result is only accepted for adaptation if the likelihood ratio between the first-best and the second-best hypothesis is above a certain threshold. To focus on the effects of adapting a system to correctly recognized material, we decided to work with ideal confidence measures using the external knowledge whether a recognition result is correct or not.

It turned out that insisting on the full word sequence to be correctly understood is too restrictive, since it biases the adaptation material towards short sentences. A better approach is to accept those sentences, for which the sequence of concepts coincides with that of the actually spoken sentence. For example, if *"From Sydney to Adelaide"* is spoken, *"From Sydney to Alice Springs"* would be accepted, since both sentences have the concept sequence *[origin, destination]*, whereas *"From Sydney over night"* would be rejected, because its concept sequence is *[origin, time]*.

As a step towards real confidence measures we perturbed the ideal confidence measure by randomly changing a chosen amount of the correct/incorrect tagging. Looking at tagging error rates of 10% and 20% we observed that these random perturbations had very little influence on the quality of the adaptation material. The reason for this effect certainly lies in the randomness of the tagging errors, whereas for real confidence measures tagging errors tend to be more systematic. This was confirmed by some brief experiments performed with real confidence measures on sentence level. The results were much closer to those for unsupervised adaptation than to those obtained using (perturbed) ideal confidence measures.

4. RESULTS

The performance of the different methods for adaptation were evaluated on two different applications: the automatic train timetable information system TABA (see [1]) and the automatic telephone switchboard PADIS (see [8]).

All used n-gram language models are bigram models. For both applications, an initial system was trained on 100 sentences. In TABA, the vocabulary consists of 2847 words including 1180 station names, the concept language model has 34 concepts. The full adaptation corpus contains 12,000 sentences with 36,058 words, the evaluation corpus has 2278 sentences with 6972 words. PADIS has a vocabulary of 1942 words amongst which there are 715 last and 299 first names. The concept language model has 15 concepts. The adaptation corpus consists of 20,000 sentences with 58,735 words, the evaluation corpus of 5157 sentences with 14,976 words.

The random tagging error rate for the perturbed ideal confidence measure was chosen as 20% in both applications and the maximal length of the N-best lists as $N = 10$. Experiments with higher values of N did not further improve the results. In TABA, the scaling factor λ was set to $\lambda = 0.01$ which gives an almost equal weighting for the hypotheses in the N-best lists. In contrast to that, we found that for PADIS a value of $\lambda = 0.75$ gave the best results. The reason for this discrepancy is that in PADIS the N-best lists consist only of sentences that are consistent with a database (cf. [8]). Thus, the quality of the hypotheses in the N-best lists decreases much faster than for TABA and the higher value of λ shifts the weight towards the first hypotheses.

4.1. Language Model for the Recognizer

The effect of the adaptation methods on the language model used to focus the beam search in the recognizer is measured in terms of the perplexity (PP) of this LM and by computing the word error rate (WER) when the best path through the word graph is obtained by rescoring the hypotheses with this LM.

Table 1 displays results for the different modes of adaptation and varying sizes of the adaptation corpus on the PADIS application.

Here, perplexity is significantly reduced and adaptation using N-best lists gives slightly better results than the other methods. For the WER, supervised adaptation gives a relative improvement of 46.6%. Since this is the maximally achievable improvement, results for the other methods should be compared to this value. For adaptation using N-best lists the WER is reduced by 35.8% relative, which is 77% of what can be achieved. The effect of adaptation using a perturbed confidence measure is very similar (35.2% relative) and unsupervised adaptation gives a relative reduction of 33.9%, which is still 73% of what can be achieved by supervised adaptation.

Figure 1 summarizes the improvements of the recognizer's LM for the TABA application.

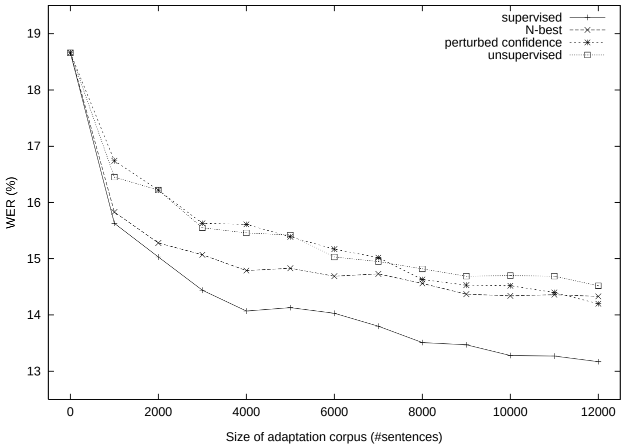

Figure 1: WER for different modes of adaptation on TABA

In this case, supervised adaptation gives a relative reduction of 29.4% in the WER. The other adaptation techniques reach comparable levels of performance, obtaining between 75% and 81% of the improvements from supervised adaptation. However, one observes that adaptation using N-best lists improves the LM faster than the other two methods.

4.2. Language Understanding

Since spoken dialogue systems have to derive the meaning of user utterances, the word error rate is not necessarily a good measure for their quality. A more important criterion is the *attribute error rate* (AER) measuring errors in the relevant information items. Especially for automatic inquiry systems the AER is highly significant, since correctness in the attributes determines whether the right database query is performed.

The effects of adaptation on the language understanding part of spoken dialogue systems (i.e. the SCFG, filler and concept language models) were only investigated for the TABA application. The database used by PADIS is not capable of adaptation and thus a crucial element of the language understanding module could not be improved. Even by supervised adaptation (i.e. training) of the SCFG on 20,000 sentences we could only obtain a relative improvement of 6.2% in the AER. One could therefore not expect any statistically significant effects from the other adaptation methods.

For the TABA application, Figure 2 shows the dependency of the AER on the size of the adaptation corpus for the different adaptation methods.

The results for supervised adaptation show that the overall achievable improvement is not very large (20.7% relative reduction of the AER), due to the fact that the initial system has a fairly high level of performance. This shows that the structural information contained in the SCFG is of great importance. However, performing adaptation using N-best lists, the AER is reduced from 17.80% to 15.38% which is a relative improvement of 13.6% and amounts to 66% of the reduction by supervised adaptation. Using a perturbed ideal confidence measure, we obtained slightly better results, the AER could be reduced to 14.92%, which is a relative improvement of 16.2% and amounts to 78% of the achievable im-

adapt. corpus # sentences	supervised		N-best		pert. confidence		unsupervised	
	PP	WER	PP	WER	PP	WER	PP	WER
0	74.83	28.72	74.83	28.72	74.83	28.72	74.83	28.72
1000	36.70	23.10	40.17	25.25	43.39	24.90	43.87	25.41
2000	29.40	21.23	37.25	24.31	40.46	24.21	40.74	24.35
4000	20.92	18.26	25.76	21.70	27.42	21.62	27.87	21.98
8000	17.13	16.92	20.76	19.72	22.33	19.92	22.22	20.37
20000	14.11	15.34	17.89	18.44	18.74	18.61	18.62	18.98

Table 1: Adaptation results (Perplexity and WER) for the recognizer's LM on PADIS

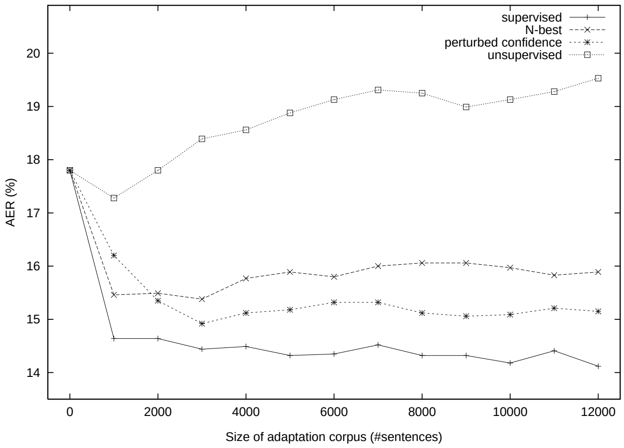

Figure 2: AER for different modes of adaptation (TABA)

provements. An important aspect is that using N-best lists almost the full improvement of the system is obtained on the first 1000 sentences adaptation material, which is similar to the behaviour of supervised adaptation. For adaptation using confidence measures, it takes longer (about 3000 sentences) to reach the final level of performance.

Finally, one observes that unsupervised adaptation suffers from error reinforcement and leads to a deterioration of the system. The AER rises from 17.80% to 19.53% after 12,000 sentences adaptation material which is a relative increase of 9.7%.

5. CONCLUSION

We have reported on methods for online-adaptation of language models for spoken dialogue systems. In particular, the benefits of using N-best lists and confidence measures was demonstrated. The results show that the language model used during recognition can be improved very impressively by these adaptation methods and that the resulting language models almost reach the quality of a trained language model.

In language understanding, experiments showed that for unsupervised adaptation error reinforcement can in fact lead to a deterioration of a system. In contrast to that, adaptation using N-best lists or confidence measures lead to very satisfactory results, as a big portion of the error reduction by supervised adaptation could be obtained.

Noticing that the improvements from adaptation using N-best lists are obtained very rapidly, the results of this paper indicate that online-adaptation allows to produce good language models for new applications of spoken dialogue systems without collecting and transcribing task-specific training material.

6. ACKNOWLEDGEMENT

We would like to thank Stefan Besling, Dietrich Klakow, Jochen Peters, Bernd Rueber and Eric Thelen for interesting discussions and fruitful suggestions.

7. REFERENCES

1. Aust, H., Oerder, M., Seide, F., and Steinbiss, V., *The Philips Automatic Train Timetable Information System,* Speech Communication 17, 249-262, 1995.
2. Besling, S., and Meyer, H.-G., *Language Model Speaker Adaptation,* Proc. EUROSPEECH'95, Madrid, Spain, 1755-1758, 1995.
3. Clarkson, P.R., and Robinson, A.J., *Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache,* Proc. ICASSP'97, Munich, Germany, 799-802, 1997.
4. Homma, S., Aikawa, K., and Sagayama, S.,*Improved Estimation of Supervision in Unsupervised Speaker Adaptation,* Proc. ICASSP'97, Munich, Germany, 1023-1026, 1997.
5. Issar, S., *Estimation of Language Models for New Spoken Language Applications,* Proc. ICSLP'96, Philadelphia, USA, 869-872, 1996.
6. Jurafsky, D. et al., *Using a Stochastic Context-Free Grammar as a Language Model for Speech Recognition,* Proc. ICASSP'95, Detroit, USA, 189-192, 1995.
7. Kellner, A., *Initial Language Models for Spoken Dialogue Systems,* Proc. ICASSP'98, Seattle, USA, 185-188, 1998.
8. Kellner, A., Rueber, B., Seide, F., and Tran, B.-H., *PADIS - An automatic telephone switchboard and directory information system,* Speech Communication 23, 95-111, 1997.
9. Matsui, T., Matsuoka, T., and Furui, S., *Smoothed N-best-based Speaker Adaptation for Speech Recognition,* Proc. ICASSP'97, Munich, Germany, 1015-1018, 1997.
10. Rueber, B., *Obtaining Confidence Measures from Sentence Probabilities,* Proc. EUROSPEECH'97, Rhodes, Greece, 739-742, 1997.
11. Thelen, E., Aubert, X., and Beyerlein, P., *Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition,* Proc. ICASSP'97, Munich, Germany, 1035-1038, 1997.