Towards a Dynamic Adjustment of the Language Weight

Georg Stemmer, Viktor Zeissler, Elmar Nöth, and Heinrich Niemann

University of Erlangen-Nuremberg, Chair for Pattern Recognition, Martensstrasse 3, D-91058 Erlangen, Germany
stemmer@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

Abstract. Most speech recognition systems use a language weight to reduce the mismatch between the language model and the acoustic models. Usually a constant value of the language weight is chosen for the whole test set. In this paper, we evaluate the possibility to adapt the language weight dynamically to the state of the dialogue or to the current utterance. Our experiments show, that the gain in performance, that can be achieved with a dynamic adjustment of the language weight on our data is very limited. This result is independent of the information source that is used for the adaption of the language weight.

1 Introduction

1.1 Motivation

In most current speech recognition systems the decoding of a spoken utterance is done with the use of statistical models. This means, given an acoustic input $\boldsymbol{X}$, the recognizer performs a search for the best matching word chain $\hat{\boldsymbol{w}}$, which maximizes the a-posteriori probability $p(\hat{\boldsymbol{w}}|\boldsymbol{X})$. Because it is hard to model the corresponding density function directly, usually the value is determined using the Bayes formula. During the decoding phase of a speech recognizer, there is no interest in the real value of $p(\hat{\boldsymbol{w}}|\boldsymbol{X})$, only the best matching word chain $\hat{\boldsymbol{w}}$ is needed. Therefore, usually only the product $p(\boldsymbol{X}|\hat{\boldsymbol{w}}) \cdot P(\hat{\boldsymbol{w}})$ is evaluated for maximization. However, practical experience shows that there is a mismatch in the output scores of the acoustic models and the language model. Most speech recognizers introduce a language weight (also called linguistic weight) α to the score of the language model in order to make it fit better to the acoustic model. As a consequence, during the decoding phase of a speech recognizer, the algorithm searches for a word chain $\hat{\boldsymbol{w}}$ with

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\boldsymbol{w}) \cdot P(\boldsymbol{w})^{\alpha} \tag{1}$$

Usually, the value of α is constant for the whole test set. However, we can expect that using only one global language weight for the combination of acoustic and language model scores in the recognizer may not be optimal and it may be

better to adjust the weight dynamically. In this paper, we describe experiments that evaluate the influence of the language weight on the word error rate. We also investigate into the possibility to choose the language weight dependent on individual sentences. Because the performance of the language model depends on the dialogue-state, we evaluate, if the optimal value of the language weight is correlated to the actual dialogue-state. A more detailed description of the experiments can be found in [1].

1.2 Previous Work

To our knowledge, only a few approaches concerning the adjustment of the language weight have been published so far. In [2] the scores of the acoustic and the language model are seen as opinions of experts. The scores are weighted with a reliability factor and linear interpolation is used to combine them. The work in [3] utilizes word dependent language weights which are determined heuristically in order to improve the language model. The Unified Stochastic Engine as described in [4] improves performance by word dependent language weights, which are estimated jointly with other model parameters.

2 Data

All evaluations were done on 20074 utterances, which have been recorded with the conversational train timetable information system Evar, as it is described in [5]. Nearly all utterances are in German language. The total amount of data is ca. 23 hours, the average length of an utterance is four seconds. 15722 utterances have been selected randomly for training, 441 for the validation of the speech recognizer. The rest of 3911 utterances is available for testing. All utterances have been labeled with the current dialogue-state. As in the work of Eckert et al. [6], the dialogue-states are defined by the question the user is replying to. Examples for the questions are *what information do you need?* or *where do you want to go?*. We only take the six most frequent questions into account, and map the less frequent questions to the more frequent ones. Experiments for the global optimization of the language weight are also compared to the results gained on a different dataset which is a subset of the data collected in the Verbmobil project [7]. The Verbmobil database contains spontaneous speech utterances which are related to appointment scheduling. The subset used for the experiments which are described in this paper contains ca. 46 hours of data and 21147 utterances. 15647 utterances have been selected randomly for training, 500 for the validation of the speech recognizer, 5000 are used for testing.

The recording scenarios cause differences between the datasets in various aspects: The average utterance length of the Verbmobil data is 8 seconds (Evar: 4 seconds). The number of distinct words in Evar is 2158, in Verbmobil it is 8822. The perplexity of the corresponding 4-gram language model is much higher for the Verbmobil data (91.2) than for the Evar data (14.1).

3 Short Description of the Baseline System

The baseline system which has been used for the experiments, is a speaker independent continuous speech recognizer. It is based on semi-continuous HMMs, the output densities of the HMMs are full-covariance Gaussian. The recognition process is done in two steps. First, a beam search is applied, which generates a word graph. The beam search uses a bigram language model. In the second phase, the best matching word chain is determined by an $A^\star$-search, which rescores the graph with a category based 4-gram language model. In both phases, a different language weight is used. For the experiments, which are described below only the language weight that is used during the $A^\star$-search is varied. A more detailed description of the speech recognizer can be found in [5].

4 Influence of the Language Weight on the Error Rate

At first, we attempt to get baseline results concerning the relationship between word error rate and the language weight. Therefore, we varied the language weight in a large interval, and computed the corresponding word error rate on the test set. As a result we get the word error rate depending on the value of a global language weight.

5 Dialogue-State Dependent Language Weight

The next step is to perform a dialogue-state dependent optimization of the language weight. As mentioned before, the motivation for this approach is the observation, that the performance of the language model depends on the dialogue-state [8]. We examine, if we can achieve a better adjustment between acoustic and language model by choosing a dialogue-state dependent value of the language weight.

6 Utterance Dependent Language Weight

A further refinement in the adjustment of the language weight is possible by adapting its value to the current utterance. However, it is not clear, which information sources could be utilized in order to determine the optimal value during recognition. One possibility among others are confidence measures, which estimate the reliability of the language model or of the acoustic model. Other information sources could be used also, for example the current speaking rate. In this paper, we measure the maximum reduction in word error rate that could be achieved by such an approach. That means, we compute the optimal value of the language weight for each utterance in the test data individually. The word error rate on the test set is evaluated with the optimal language weights.

7 Experimental Results

7.1 Influence of the Language Weight on the Error Rate

For the global optimization of the language weight four different validation sets are selected from the EVAR training data. Each of the subsets contains 500 utterances. In Table 1 the word error rates which have been computed on the

Table 1. Global optimization of the language weight on the EVAR dataset. For each validation set language weight and word error rates on the validation and test set are shown. Please note that each validation set was excluded from the training data for the acoustic and language models, so the rows of the table correspond to recognizers, which have been trained differently.

validation set	optimal language weight	WER [%] on validation set	WER [%] on test set
1	6.0	20.5	26.9
2	4.8	25.3	26.5
3	7.1	24.0	26.8
4	7.5	23.5	26.1
test set	8.0	-	26.1

EVAR test set are shown. Depending on the particular validation dataset, word error rates between 26.9% and 26.1% have been achieved. The word error rate on the test set is 26.9% at the value of the language weight which is optimal on the validation data (6.0). In order to evaluate the maximum performance that could be reached, we also use the test set for optimization. This experiment results in a word error rate of 26.1%, the corresponding language weight is 8.0.

The curve in Fig. 1 depicts the progression of the word error rate depending

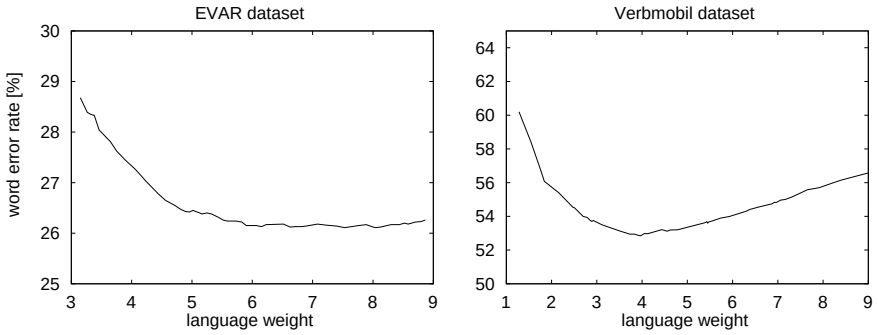

Fig. 1. Relationship between word error rate and the language weight for the test subsets of the EVAR and the VERBMOBIL data.

on the language weight for the test sets of the EVAR and VERBMOBIL databases. On the EVAR test set, the word error rate is nearly constant in the range from 5 to 9. The curve for the VERBMOBIL data has a minimum at a language weight of 4, the word error rate rises slowly for larger weights. While both curves are relatively smooth their most significant difference is the position of the global minimum, which seems to be data-dependent (see Section 2).

7.2 Dialogue-State Dependent Language Weight

For dialogue-state dependent optimization of the language weight we use three different validation sets, each contains 2000 utterances. The results can be seen in Table 2. When compared to the global optimization of the language weight,

Table 2. Dialogue-state dependent optimization of the language weight. For each dialogue-state (DS0, .., DS5) the language weight is chosen separately. The optimal weights and the word error rates for the test set are shown.

validation set	opt. language weights						WER [%]
	DS0	DS1	DS2	DS3	DS4	DS5	
1	6.8	7.8	7.1	5.7	5.8	6.1	26.2
2	8.6	5.1	7.3	4.9	7.1	8.4	27.2
3	7.5	6.1	6.2	5.5	7.9	6.1	26.4

only slight improvements of the word error rate can be achieved. Depending on the validation set that has been used for the optimization, the word error rate on the test set lies between 27.2% and 26.2%. The optimal values of the language weights have a very high variance. In some cases, the variance for the same dialogue-state across the validation data sets is even higher than the variance across different dialogue-states in the same validation data set. The high variance may be caused by an insufficient size of the data sets. Another explanation is that there is no or only little correlation between the optimal value of the language weight and the current dialogue-state.

7.3 Utterance Dependent Language Weight

Because we want to measure the maximum possible reduction in word error rate, we do not use a separate validation set for this experiment. The language weight is optimized for each utterance of the test set. The approach results in a word error rate of 25.4% on the test set, which corresponds to a reduction of 0.7 percent points or 2.8% relative. The slight improvement is the best possible reduction of the word error rate that can be achieved with an utterance dependent language weight. For this result, it is irrelevant, which information source or confidence measure can be utilized for the adaption. In order to understand the results better, we analyzed the change in the word error rate on individual sentences

dependent on the language weight. We discovered, that the error rate on 1046 (27%) of the sentences in the EVAR test set is completely independent of the language weight. In 2991 sentences (77%), the error rate does not change, when the language weight is varied in a range from 1 to 9. Utterances which have an error rate that does not depend on the language weight, are significantly shorter than the average.

8 Conclusion and Outlook

In most speech recognition systems, the score of the language model is manipulated with a language weight. Usually a constant value of the language weight is chosen for the whole test set. In this paper, we evaluated the possibility to adapt the language weight dynamically to the state of the dialogue or to the current utterance. Our experiments show, that a dynamic adjustment of the language weight does not necessarily result in a better performance. This may be caused by the great portion of relatively short utterances in our data, which is typical for dialogues in an information retrieval system.

Further experiments will investigate into the VERBMOBIL data set, which contains utterances with a larger duration. We also plan to apply alternative methods (e.g. [9]) which are not necessarily based on a weighting factor to reduce the mismatch between acoustic model and language model.

References

1. V. Zeissler: Verbesserte Linguistische Gewichtung in einem Spracherkenner. Master thesis (in German), Chair for Pattern Recognition, University of Erlangen-Nuremberg, Erlangen (2001)
2. H. Bourlard and H. Hermansky and N. Morgan: Towards Increasing Speech Recognition Error Rates. Speech Communication, vol. 18 (1996), 205-231
3. Ramesh R. Sarukkai and Dana H. Ballard: Word Set Probability Boosting for Improved Spontaneous Dialogue Recognition: The AB/TAB Algorithm. University of Rochester, Rochester (1995)
4. X. Huang and M. Belin and F. Alleva and M. Hwang: Unified Stochastic Engine (USE) for Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Minneapolis (1993) 636-639
5. F. Gallwitz: Integrated Stochastic Models for Spontaneous Speech Recognition. Dissertation, University of Erlangen-Nuremberg, Erlangen (to appear)
6. W. Eckert and F. Gallwitz and H. Niemann: Combining Stochastic and Linguistic Language Models for Recognition of Spontaneous Speech. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta (1996) 423–426
7. W. Wahlster, Verbmobil: Foundations of Speech-to-Speech Translation. Springer, New York, Berlin (2000)
8. G. Stemmer and E. Nöth and H. Niemann: The Utility of Semantic-Pragmatic Information and Dialogue-State for Speech Recognition in Spoken Dialogue Systems. Proc. of the Third Workshop on Text, Speech, Dialogue, Brno (2000) 439-444
9. V. Fischer and S.J. Kunzmann: Acoustic Language Model Classes for a Large Vocabulary Continuous Speech Recognizer. Proc. Int. Conf. on Spoken Language Processing, Bejing (2000) 810–813