The Utility of Semantic-Pragmatic Information and Dialogue-State for Speech Recognition in Spoken Dialogue Systems

Georg Stemmer, Elmar Nöth, and Heinrich Niemann

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

Abstract. Information about the dialogue-state can be integrated into language models to improve performance of the speech recognizer in a dialogue system. A dialogue state is defined in this paper as the question, the user is replying to. One of the main problems in dialogue-state dependent language modelling is the limitation of training data. In order to obtain robust models, we use the method of rational interpolation to smooth between a dialogue-state dependent and a general language model. In contrast to linear interpolation methods, rational interpolation weights the different predictors according to their reliability. Semantic-pragmatic knowledge is used to enlargen the training data of the language models. Both methods reduce perplexity and word error rate significantly.

1 Introduction

Current spoken dialogue systems usually have a modular architecture. In most systems, the understanding of the user's utterances is done by several components, which usually form a processing pipeline: A speech recognizer generates one or several hypotheses of the spoken word sequence, which are transformed into a semantic-pragmatic representation by a language understanding unit. In combination with the actual dialogue-state, the semantic-pragmatic representation is used by the dialogue management for system-user interaction and communication with external systems, e.g. a database.

This paper addresses an approach to increase the performance of the speech recognizer of a spoken dialogue system by additional information sources. This is motivated by the observation, that errors during speech recognition often conflict with the current dialogue-state. E.g. for a train timetable information system, if the system asks for the city of departure, *yes* or *no* do not have a very high probability to occur in the spoken utterance. In literature, several approaches to use the dialogue-state as a predictor of the user response can be found. The training corpus is split into partitions, which correspond to the dialogue-states. Each partition is used to estimate a separate, dialogue-state dependent language model.

Because the partitioning reduces the amount of training data for each language model, interpolation or fall back strategies have to be applied in order to get robust language models. In [1], dialogue-states are generalized until the training data is sufficient. In the work of Riccardi et al. [3] the dialogue-state dependent language models are interpolated with a general language model. Wessel et al. [5] use interpolation to combine the dialogue-state dependent language models with each other. In the following, we basically propose two approaches in the field of dialogue-state dependent language modelling. First, we describe a new strategy for the interpolation between specialized and general language models, rational interpolation. In contrast to linear interpolation schemes, rational interpolation weights the different predictors with their reliability. Second, we show how to utilize semantic-pragmatic knowledge to find suitable backing-off data for the language models of rare dialogue-states.

2 Data

All evaluations were done on 20678 utterances, which have been recorded with the conversational train timetable information system EVAR, as it is described in [2]. Nearly all utterances are in German language. The total amount of data is 23 hours, the average length of an utterance is four seconds. 16767 utterances have been selected randomly for training and validation (15745 for training, 1022 for validation), the remainder of 3911 utterances is available for testing.

3 Semantic Attributes and Dialogue-States

The corpus is transcribed with a set of 58 semantic-pragmatic attributes, which give a formatted description of the meaning of the user's utterance in the context of train timetable information. The most important attributes are *time*, *date*, *sourcecity* (city of departure), *goalcity* (destination), *no* and *yes*.

As in the work of Eckert et al. [1], we define the dialogue-states 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.

4 Evaluation of the Baseline System

The baseline system has a word error rate of 21.2%. As integration of additional information will be used to increase the performance of semantic-pragmatic attribute recognition, we evaluate the semantic-pragmatic attribute accuracy of the baseline system. The attribute accuracy of the system is evaluated on 2000 utterances of the test data set, for which a manual transcription with semantic-pragmatic attributes is available. The attribute accuracy is defined similar to the word accuracy as the number of substituted, inserted or deleted attributes, subtracted from the number of attributes in the reference, relative to the number of attributes in the reference. Please note, that we do not compute a semantic concept accuracy here, i.e. if the recognizer intermixes two city names, there may be no influence on the attribute accuracy, because the correct attribute *goalcity* can still be generated by the language understanding component. If the language understanding component of the dialogue system is applied to the spoken word sequence, the attribute accuracy for the attributes *time*, *date*, *sourcecity*, *goalcity*, *yes*, and *no* is 90.8%. This value drops to 86.0%, if the language understanding component is applied to the output of the speech recognizer. The following experiments investigate, to which extent the loss in semantic-pragmatic attribute accuracy may be reduced by the integration of new information sources.

5 Description of the Language Models

5.1 Rational Interpolation

The method of rational interpolation for language modelling has been proposed by Schukat-Talamazzini et al. [4] for the task of combining a set of conditional n-gram word probability predictors. The approach was to find a good approximation $\tilde{P}(w_t|v)$ of the conditional probability $P(w_t|v)$, that a word w_t occurs given a sentence history v. The approximation is based on the maximum likelihood estimates

$$\hat{P}_i(w_t|v) = P'(w_t|w_{t-i+1},..,w_{t-1}) = \frac{\#(w_{t-i+1},..,w_t)}{\#(w_{t-i+1},..,w_{t-1})} \tag{1}$$

where the function $\#(\cdot)$ counts the frequence of occurence of its argument. A set of predictors $\hat{P}_i(w_t|v)$, $i \in I$, results from assigning different values to i, i.e. looking at different portions of the sentence history v. All predictors $\hat{P}_i$ are combined in the following interpolation scheme:

$$\tilde{P}(w_t|v) = \frac{\sum_{i\in I} \lambda_i \cdot g_i(v) \cdot \hat{P}_i(w_t|v)}{\sum_{i\in I} \lambda_i \cdot g_i(v)} \tag{2}$$

The λ_i represent the interpolation factors, the function $g_i(\cdot)$ weights the predictor $\hat{P}_i$ with its reliability. The reliability is given by a hyperbolic weight function which is high for large values of $\#_i(v)$ and low for small values of $\#_i(v)$. The denominator is for normalization. The optimization of the λ_i is done by Newton iteration on a cross-validation data set. It is not possible to apply the EM-Algorithm here, because –in contrast to linear interpolation schemes– Eq. (2) cannot be interpreted as a stochastic process. In the following, the language model defined by Eq. (2) will be referred to as $\mathcal{L}_0$, it is a general model that is independent of the dialogue-state.

5.2 Rational Interpolation for Dialogue-State Dependent Models

Rational interpolation is basically a method for the combination of different predictors for the approximation of $P(w_t|v)$. Dialogue-state dependent modelling is based on the assumption, that the dialogue-state is a good predictor for what the user says. Instead of approximating $P(w_t|v)$, we want to find an estimation of $P_d(w_t|v)$, where d stands for the current dialogue-state. The first approach is to combine only the dialogue-state dependent predictors $\hat{P}_{i,d}(w_t|v)$ for the language model of one specific dialogue state d. In the following, the resulting model is called $\mathcal{L}_1$. Because the predictors $\hat{P}_{i,d}(w_t|v)$ are estimated only on a fraction of the original dialogue-state independent training data, the reliability of most predictors is much lower. Rational interpolation enables us to combine the predictors $\hat{P}_{i,d}(w_t|v)$ with the predictors $\hat{P}_i(w_t|v)$ to increase robustness of the language model. The resulting language model will be called $\mathcal{L}_2$:

$$\tilde{P}_{\mathcal{L}_2,d}(w_t|v) = \frac{\sum_{i\in I} \lambda_{i,d} \cdot g_{i,d}(v) \cdot \hat{P}_{i,d}(w_t|v) + \gamma \cdot \sum_{i\in I} \lambda_i \cdot g_i(v) \cdot \hat{P}_i(w_t|v)}{\sum_{i\in I} \lambda_{i,d} \cdot g_{i,d}(v) + \gamma \cdot \sum_{i\in I} \lambda_i \cdot g_i(v)} \tag{3}$$

The γ is the interpolation factor between the dialogue-state dependent and the general language model. For parameter adjustment, the $\lambda_{i,d}$ and γ are optimized jointly, the λ_i are the same parameters, that have been calculated for the general language model $\mathcal{L}_0$, and do not underlie any further optimization.

5.3 Incorporating Semantic-Pragmatic Knowledge

It is obvious, that the current dialogue-state primarily determines the meaning of the user's utterance. There is only a minor influence on the syntactical structure of the sentence. There is, for example, a high similarity between utterances, that contain information about the time of departure. This similarity is not influenced by the dialogue-state, in which a sentence was uttered. This observation can be expressed by the following simplified model of utterance production:

$$P_d(w_1, .., w_t) = \sum_s P_d(w_1, .., w_t|s) \cdot P_d(s) \approx \sum_s P(w_1, .., w_t|s) \cdot P_d(s) \tag{4}$$

The semantic-pragmatic content s of the utterance $w_1, .., w_t$ depends on the dialogue-state d, while the influence of the dialogue state on the syntactial realization of the utterance is neglected. A consequence of Eq. (4) is to integrate semantic-pragmatic predictors $\hat{P}_{i,s}(w_t|v)$ in the language model. These do not depend on a dialogue-state, the predictor $\hat{P}_{i,s}(w_t|v)$ is estimated on all sentences in the training corpus, that have the semantic transcription s, e.g. *goalcity*. The resulting language model will be referred to as $\mathcal{L}_3$:

$$pred_0 = \sum_{i\in I} \lambda_i \cdot g_i(v) \cdot \hat{P}_i(w_t|v) \tag{5}$$

$$pred_1(d) = \sum_{i\in I} \lambda_{i,d} \cdot g_{i,d}(v) \cdot \hat{P}_{i,d}(w_t|v) \tag{6}$$

$$pred_2(s) = \sum_{i\in I} \lambda_{i,s} \cdot g_{i,s}(v) \cdot \hat{P}_{i,s}(w_t|v) \tag{7}$$

$$\tilde{P}_{\mathcal{L}_{3,d}}(w_t|v) = \frac{\gamma_0 \cdot pred_0 + \gamma_1 \cdot pred_1(d) + \sum_{s\in S} \gamma_{2,s} \cdot pred_2(s)}{norm} \tag{8}$$

Model $\mathcal{L}_3$ integrates the dialogue-state dependent, general and semantic-pragmatic predictors. *norm* is the normalization factor given by the sum of all interpolation and weighting factors. S is a set of semantic-pragmatic attributes, which have a high probability to occur in d. For our experiments, $\mathcal{L}_3$ only uses semantic predictors for the model of dialogue-state three, which is the state with the lowest amount of training data. S contains only one semantic-pragmatic attribute here, *goalcity*, which is the most frequent attribute in dialogue-state three. All other dialogue-states are represented as in language model $\mathcal{L}_2$. To keep the number of parameters that are subject to optimization low, only the γ factors are optimized jointly, all λ factors can be optimized with separate (dialogue-state or attribute dependent) language models.

6 Experimental Results

For evaluation, we have measured the perplexities and the word error rates on the test sentences. Table 1 shows the amount of training and test sentences, that were available for the different language models and the perplexities. As can be seen from columns four and five, dialogue-state dependent modeling without interpolation can increase perplexities, if the training data is not sufficient. The same effect influences also model $\mathcal{L}_2$, because the optimization algorithm does not allow the weights to become zero. In model $\mathcal{L}_3$ the perplexity in all dialogue-states is reduced, even in dialogue-state three, mainly because additional training data becomes available for the model of state three (3131 sentences from the training corpus, which are labelled with *goalcity*). For the application of the models in our dialogue system, we integrated the new interpolation methods into the bigram model of the forward search of the recognizer and into the 4-gram model that is used for the rescoring of the word lattice during the $A^\star$ search. Table 2 shows the word error rates for the different models. The best word error rate reduction is 3.8% relative, while the attribute accuracy increases by 0.8% absolute.

Table 1. Number of sentences for each dialogue-state in the training and test corpus and perplexities for the different language models on the test corpus

dialogue-state	train	test	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$
0	8550	1957	13.0	12.4	12.3	12.3
1	1975	566	13.5	11.9	11.3	11.3
2	1397	379	16.5	13.8	12.1	12.1
3	763	219	17.1	44.2	21.2	14.3
4	1321	373	18.1	13.6	12.3	12.3
5	1739	417	21.8	16.3	14.7	14.7

Table 2. Word error rate (WER) and attribute accuracy (AA) for the different language models

model	WER [%]	AA [%]
$\mathcal{L}_0$	21.2	86.0
$\mathcal{L}_1$	20.9	86.1
$\mathcal{L}_2$	20.5	86.8
$\mathcal{L}_3$	20.4	86.8

7 Conclusion

Errors of the word recognizer cause a loss in speech understanding performance. We presented two new models which apply the method of rational interpolation to the problem of smoothing between language models. Additional training data becomes available by the incorporation of semantic-pragmatic knowledge. Sentences which have a similar meaning can be used for the estimation of backing-off statistics of a model for a dialogue-state. We showed that the integration of our dialogue-state dependent language models into the speech recognizer improves performance: The word error rate is reduced by 3.8% relative, there is also an increase in speech understanding performance. Further experiments will investigate into a more sophisticated definition of the dialogue-state.

References

1. 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, USA (1996) 423–426
2. F. Gallwitz and M. Aretoulaki and M. Boros and J. Haas and S. Harbeck and R. Huber and H. Niemann and E. Nöth: The Erlangen Spoken Dialogue System EVAR: A State–of–the–Art Information Retrieval System. Proceedings of 1998 International Symposium on Spoken Dialogue, Sydney, Australia (1998) 19–26
3. G. Riccardi and A. L. Gorin: Stochastic Language Adaptation Over Time and State in a Natural Spoken Dialog System. IEEE Trans. on Speech and Audio Proc., Vol. 8, No. 1, January 2000 3–10
4. E.G. Schukat-Talamazzini and F. Gallwitz and S. Harbeck and V. Warnke: Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. Proc. European Conf. on Speech Communication and Technology, Rhodes, Greece, (1997) 2731–2734
5. F. Wessel and A. Baader: Robust Dialogue–State Dependent Language Modeling Using Leaving–One–Out. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Phoenix, USA (1999) 741–744