INTERSPEECH 2009 BRIGHTON

Using Dialogue-Based Dynamic Language Models for Improving Speech Recognition

Juan Manuel Lucas-Cuesta, Fernando Fernández, Javier Ferreiros

Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain

{juanmak, efhes, jfl}@die.upm.es

Abstract

We present a new approach to dynamically create and manage different language models to be used on a spoken dialogue system. We apply an interpolation based approach, using several measures obtained by the Dialogue Manager to decide what LM the system will interpolate and also to estimate the interpolation weights. We propose to use not only semantic information (the *concepts* extracted from each recognized utterance), but also information obtained by the dialogue manager module (DM), that is, the objectives or *goals* the user wants to fulfill, and the proper classification of those concepts according to the inferred goals. The experiments we have carried out show improvements over word error rate when using the parsed concepts and the inferred goals from a speech utterance for rescoring the same utterance.
Index Terms: spoken dialogue systems, dynamic language modeling, automatic speech recognition

1. Introduction

Statistical language model adaptation has become a current issue within the scope of Speech Technology. Its main goal consists of modifying the language model (LM) of which an automatic speech recognition system (ASR) makes use, in order to achieve better recognition rates. For instance we can modify a general LM to adapt it to a closed domain, trying to improve the overall response of a domain-dependent system in which the ASR is included.

There are several approaches to adapt language models, depending on the sources of the interpolating models ([1]). Perhaps the simplest one (and the best known and analyzed) consists of the linear interpolation between different LM ([2, 3, 4]). This approach tries to find out a good weight to combine a general LM (which we will refer to as *background model*), usually built with a high amount of data, with one or several adaptation LM, usually built with fewer but more specific data. In most cases interpolation approaches base their weight search on different algorithms (such as MAP or maximum entropy models) which minimize the perplexity of a test database.

An interesting point of view relies on developing the LM adaptation considering the evolution of natural language throughout a dialogue. This is based on the fact that people usually switch their way of expressing ideas depending on what they want to say, and who they are talking with. Thus a good way of building adaptive LM is to make time dependent dynamic language models that can evolve during dialogues ([5]).

State of the art spoken dialogue systems that use dialogue dependent language modeling usually take advantage of the information provided by the natural language understanding module (NLU). These systems base their LM estimation on the different labels or categories of each word or phrase ([6]). After the LM estimation stage, a rescoring task of the same utterance is done ([7]) in order to improve its recognition.

Our approach can estimate dynamic LM using not only NLU information (i.e. the semantic information or dialogue concepts), but also knowledge obtained by the dialogue manager about the objectives (i.e. the dialogue goals) that the user wants to fulfill.

Our final objective will be to modify the LM with dialogue-based information not just to improve the recognition of the current utterance using its own information, but also the dialogue concepts and goals of the preceding utterances. This way we will develop a more natural dialogue system.

The rest of the paper is organized as follows. Section 2 briefly presents our baseline dialogue system. The main details of our dynamic interpolation approach are discussed on Section 3. Next, Section 4 shows the experimental setup and our main results. Finally, Section 5 summarizes the conclusions we have come to, and presents some of our future research guidelines.

2. Baseline dialogue system

Here we briefly present the baseline dialogue system we have modified. A more detailed analysis can be found on [8] and [9].

We apply our dialogue system in the development of a conversational interface for controling a commercial Hi-Fi audio system using natural language sentences instead of a common infrared remote control.

We have designed a mixed-initiative spoken dialogue system based on the use of Bayesian Networks, BN, as the basis of our dialogue manager (DM). This approach can exploit the causal relationships between the semantics of an utterance (i.e. *dialogue concepts*) and the intention of the speaker (i.e. *dialogue goals*). We will refer to both concepts and goals as *dialogue items*. Dialogue items have been defined by hand using expert knowledge of the application domain.

2.1. Dialogue concepts and goals

We have set a concept dictionary trying to cover all the semantic categories in the application domain. We have defined three different concept subsets: *actions* to be executed over the system (e.g. to play), *parameters* that can be configured (e.g. equalization) and their corresponding *values* (e.g. a number). We have defined up to 50 different concepts, classified into 22 actions, 16 parameters and 20 values.

We have also defined 15 different dialogue goals according to the intention of the user and the different actions the Hi-Fi system can perform.

2.2. Dialogue Management

Once the ASR has extracted the recognized text from the input utterance, and the natural language understanding (NLU) module has extracted the different concepts of that sentence, the DM has to identify the dialogue goals, using all the available information (i.e. dialogue concepts). Then, according to the inferred goals, the DM has to decide how the dialogue should continue.

Both tasks are based on a BN approach, by means of a *forward inference* procedure, FI. This algorithm estimates the posterior probability of each dialogue goal given the available evidences (i.e. the presence or absence of each concept). By comparing the resulting probabilities with several predefined thresholds θ_i the DM decides whether a goal is*present* or *absent*.

After the FI process, the DM assumes the inferred goals as new evidences, and it estimates similar posterior probabilities for the concepts. This is developed by a new Bayesian inference procedure, known as *backward inference*, BI. Again, the decision of assuming if a given concept should be present or not is taken by a comparison against different thresholds. The result of this decision process is used to develop the most suitable dialogue action (prompting the user for wrong or incomplete information, or developing the actions the user asked for).

3. Dynamic LM generation

We have included a dynamic grammar generator (DGG) module into the dialogue system presented on the previous section. This new subsystem will act as a feedback loop between the dialogue manager and the speech recognizer.

3.1. Offline LM estimation

First of all, we need to estimate the different language models of each dialogue item (concepts and goals). Our first approach is based on the estimation of a LM for each dialogue item. We will use each sentence of our database which makes reference to a certain dialogue item for estimating its LM. This way, a same sentence could be used into different LM if that sentence makes reference to different dialogue items.

3.2. Online LM selection

The dynamic LM estimation takes place online with the dialogue system working. Once a sentence has been recognized, and the DM has developed both forward and backward inferences, the posterior probabilities of both present or observed concepts and positively inferred goals are used to estimate the interpolation weights between the background (static) LM and the different dialogue item based LM.

Taking into account the baseline presented in the previous section, we can see that our system will not select those item-based LM related with absent concepts or with dialogue goals that the system classifies as not actives. That is, the dynamic LM will be built using only those dialogue items positively inferred by the dialogue manager.

3.3. Online LM interpolation

In this stage the system has to decide how to interpolate the selected LM with the background one. Instead of choosing every dialogue concept and goal inferred by the DM, we have defined different *relevance thresholds* for both dialogue items, Φ_C for concepts and Φ_G for goals.

It is important to emphasize that these new thresholds do not have to be the same than those thresholds predefined for the forward and backward inference procedures. We will study the influence of Φ_C and Φ_G on the system performance, by making different experiments to find out the best thresholds in terms of system performance.

If the posterior probability of a given dialogue item is over its corresponding threshold, the LM based on that item will be interpolated with the background one. As long as we are dealing with two different dialogue items (concepts and goals) we have defined two thresholds: Φ_C for concepts and Φ_G for goals.

If we start with the well-known interpolation equation between probabilistic LM, we will include the dynamic behavior in the form of a time dependency of the different interpolation weights. Thus the probability of a word *w* given its preceding words (its history) *h* in the interpolated model will be

$$p_T\left(w \mid h\right) = W_B\; p_B\left(w \mid h\right) + \left(1 - W_B\right)\; p_D\left(w \mid h\right) \quad (1)$$

being p_B the probability according with the background model, p_D the probability obtained dynamically with the item-based LM, and W_B the interpolation weight between both models.

3.3.1. Goal-based LM interpolation

First of all we have interpolated only goal-based LM with the background model. That is, the dynamic language model of equation 1, $p_D\left(w \mid h\right)$, will be equal to $p_G\left(w \mid h\right)$.

We have used the posterior probabilities of dialogue goals, given by the forward inference procedure, for defining the interpolation weigths of the dynamic model, taking into account the constraint that the interpolation weights must sum 1.

Let $p_f\left(g_i = 1 \mid e_{g_i}\right)$ be the posterior probability that the goal g_i is present on the utterance under analysis, given its evidence e_{g_i}, and let $p_{g_i}\left(w \mid h\right)$ be the probability of having the word w given its history h under the LM associated to the goal g_i. Using these definitions the goal-based dynamic LM is calculated as:

$$p_G\left(w \mid h\right) = \frac{1}{\sum_{g_i} p_f\left(g_i \mid e_{g_i}\right)} \sum_{g_i} \left[p_f\left(g_i \mid e_{g_i}\right)\; p_{g_i}\left(w \mid h\right)\right] \quad (2)$$

where both sums extend only to those goals g_i which posterior probability $p_f\left(g_i = 1 \mid e_{g_i}\right)$ is over the goal selection threshold Φ_G. Hence we give more relevance to the goals best scored by the DM, making that the sentences which make reference to those goals have more importance in the dynamic LM.

3.3.2. Concept-based LM interpolation

An equivalent analysis to the one presented before has been done to develop an interpolation using only concept information. On this new analysis, $p_D\left(w \mid h\right) = p_C\left(w \mid h\right)$.

We have used again the posterior probabilities of dialogue concepts, given by the backward inference procedure, for estimating the interpolation weights of the dynamic LM.

Now let $p_b\left(c_i \mid e_{c_i}\right)$ be the posterior probability assigned by the DM to the concept c_i given its evidence e_{c_i}, and $p_{c_i}\left(w \mid h\right)$ be the probability of the word w given its history h according to the LM associated to the concept c_i. The dynamic LM based only on concepts will thus take the form

$$p_C\left(w \mid h\right) = \frac{1}{\sum_{c_i} p_b\left(c_i \mid e_{c_i}\right)} \sum_{c_i} \left[p_b\left(c_i \mid e_{c_i}\right)\; p_{c_i}\left(w \mid h\right)\right] \quad (3)$$

Again, to give more relevance to the best scored concepts, both sums extend only to those slots c_i with posterior probability $p_b\left(c_i \mid e_{c_i}\right)$ over the concept selection threshold Φ_C.

3.3.3. Concept and goal merging

The objective of our final approach relies on estimating a dynamic LM based on both dialogue items (concepts and goals). In this last case $p_D\,(w \mid h)$ will be composed of an interpolation between a goal-based LM and a concept-based LM:

$$p_D\,(w \mid h) = \frac{1}{W_G + W_C}\,(W_G\; p_G\,(w \mid h) + W_C\; p_C\,(w \mid h)) \tag{4}$$

where p_G and p_C are the interpolated LM presented before, and W_G, W_C are the weights assigned to each of these models.

Instead of estimating both weights we directly obtain them from the posterior probabilities of slots and goals, according to

$$\begin{aligned} W_G &= \tfrac{1}{(1-\Phi_G)\;N_G} \textstyle\sum_{g_i} [p_f\,(g_i \mid e_{g_i}) - \Phi_G] \\ W_C &= \tfrac{1}{(1-\Phi_C)\;N_C} \textstyle\sum_{c_i} [p_b\,(c_i \mid e_{c_i}) - \Phi_C] \end{aligned} \tag{5}$$

In the previous equation Φ_G, Φ_C are the respective thresholds for considering the goal or concept LM to be interpolated; N_G, N_C are the total number of goals and concepts inferred from the input utterance, and $p_f\,(g_i = 1 \mid e)$, $p_b\,(c_i \mid e)$ are the posterior probabilities of each goal g_i and concept c_i of the utterance given its respective evidences, e_{g_i} and e_{c_i}.

As in the former approaches, sums in equations 5 take into account only those goals or concepts which a posteriori is over the corresponding threshold. That is, W_G will be estimated using only those goals g_i such that $p_f\,(g_i = 1 \mid e_{g_i}) > \Phi_G$, and the same way for W_C.

We have defined the previous equations to give more relevance to those dialogue items which have higher posterior probabilities, that is, those items in which the DM has more confidence to be in the input utterance. The chosen formulae also assures that both weights W_G and W_C takes values between 0 and 1 whatever the posterior probabilities are.

4. Experimental setup

This section presents the database we have used to evaluate the behaviour of the dynamically adapted LM, and the results of the different experiments we have carried out.

4.1. Baseline results

We have used a proprietary database called HIFI-MM1. This database is composed of 100 different sentences spoken by 13 different speakers (7 male, 6 female), giving thus a total of 1300 sentences related with the application domain.

Each sentence of the database has been manually labeled with its appropriate concepts and dialogue goals. The following table shows the mean of the number of concepts and goals that each sentence makes reference to.

Table 1: *Statistical parameters of dialogue item distribution*

	μ
slots	4.31
goals	2.17

The relatively high number of inferred dialogue concepts and goals for each sentence enables a good accuracy of our approach (i.e. selecting the best scored concepts and dialogue goals when building the dynamic LM).

By means of a k-fold approach we have split the database into ten different folds (each one with 130 sentences picked up randomly from the database), with which we build three different sets: a *training* one, composed of eight folds (1040 sentences), and a *validation* and a *test* sets, each one with one fold (130 sentences).

Using round-robin we develop ten different experiments. On each one the training subset has been used to build the background LM, whilst the validation subset serves us to adjust the different parameters: LM weight (LMW), inter word penalty (IWP), interpolation weights and concept and goal thresholds.

Using the test subset to evaluate the performance of the ASR, the baseline results (without using dynamic LM interpolation) shows a word error rate of 5.33%.

4.2. On the use of goal information

Our first experiment consisted of taking into account only information about the objectives the user wants to fulfill, that is, the dialogue goals which the dialogue manager infers from the recognized concepts. We use the validation subset to estimate W_B and Φ_G, as we explained on section 3, as well as LMW and IWP.

Table 2 compares the average results obtained when applying the dynamic LM adaptation to the test subset, in terms of word error rate.

Table 2: *WER with goal-based LM interpolation.*

W_B	Φ_G	WER (%)	Baseline
0.9	0.43	4.73	5.33

As we can see, when we use only information about dialogue goals our system can improve its performance. The value of W_B indicates that a slight modification over the background model (keeping a 90% of the static LM) tends to reduce word error rate.

We can also see that Φ_G gets a value of 0.43. This implies that there is an important information even in goals that the forward inference assigns a relatively low posterior probability.

Finally, if we compare the average word error rate with the baseline result we can see a relative improvement of 11.24%. Despite the fact that the confidence intervals ($\pm$0.51%) are still slightly overlapped, our results are very promising.

4.3. On the use of concept information

We next evaluated the behaviour of the recognizer when interpolating concept dependent LM. This time the validation set helped us to estimate Φ_C as well as the rest of common parameters. Table 3 summarizes the recognition performance when recognizing the test subset.

Table 3: *WER with concept-based LM interpolation.*

W_B	Φ_C	WER (%)
0.87	0.53	4.78

If we compare the word error rate obtained when using only concept information with the baseline (5.33%, table 2) we see that using only semantic information our system achieves an improvement very close to the obtained with using goals.

The threshold Φ_C that we consider to decide if we use a certain concept based LM takes a value of 0.53. This implies that

despite the most outstanding information for building a dynamic concept-based LM relies on the concepts with high posterior probabilites, if we take also into account those concepts with middle values of posterior probabilities the recognition performance will improve.

The set of fully labeled sentences (with which we build the dialogue dependent LM) is reduced, the set of sentences for several concepts usually keeps below the set of sentences for certain goals. This is due to the fact that we have defined a set of concepts (58) larger than the set of goals (15), so the estimation of the LM is worse using concepts than using goals. Therefore the performance when using concepts is slightly worse than using goals.

We can also compare the F-measure of the forward inference (88.14%) and the backward inference (81.00%, [9]). Taking into account these values we can also say that the higher accuracy of the DM inferring goals than inferring concepts can also affect to the performance of the dynamic LM estimation, giving more relevance to the goal-based LM.

4.4. Merging both dialogue items

Our last experiment consists of using both information sources (concepts and goals together) to adapt the LM for each sentence. The average results of this experiment are shown on Table 4.

Table 4: *Word error rate with merging strategy.*

W_B	Φ_G	Φ_C	WER (%)
0.9	0.46	0.57	4.55

This last experiment yields the best results in terms of word error rate. Merging both concept and goal-based LM into a single dynamic LM takes advantage of both information sources, giving a relative reduction of WER of 14.63%. Despite the fact that the confidence intervals ($\pm$0.5%) still show a slight overlap with the baseline ones, the tendency of improving performance remains.

The different thresholds Φ_G and Φ_C takes values of 0.46 and 0.57, respectively. These values implies that our system gives more relevance to the goal-based LM. This is due to the different number of goals and concepts. As we have defined more concepts than goals, but the sentences for training the item-based LM are the same, the different goal-based models are estimated with higher amount of sentences than the concept-based LM. So the models based on goals are more reliable than the concept-based ones, but these LM still help to improve the overall performance.

5. Conclusions

We present a new approach of using dialogue information to estimate the interpolation weights of a dynamic LM interpolation. We have seen that an accurate estimation of concept and goal thresholds yields to better speech recognition rates.

The different blocks of a spoken dialogue system (natual language understanding, dialogue manager) provides information about what LM have to be used when building a dynamically interpolated LM. A proper selection of what dialogue concepts and dialogue goals scored by the DM will be part of the LM will imply an improvement of the system performance.

Concerning the usage of item-based language modeling, we can see that using a different LM for each dialogue goal the recognition rates improve up to 11.23% of WER relative reduction. However, using only dialogue concepts the recognition rate does not seem to improve as much as it does when using only dialogue goals. Still, the confidence intervals are so close that this difference is not significant. The best result takes place when merging both item-based LM into a single interpolation process, showing that our approach can improve a static LM.

The tendency of improvement shown when using dialogue goals relies on the different number of goals and concepts that we have defined for our system, and that we have used the same database for training every item-dependent language model. While we have defined only 15 goals, the number of concepts (58) is clearly higher. This implies that each concept-based LM is trained with a more reduced set of sentences than a goal-based one. We also have to take into account that the goals are inferred by using semantic information (i.e. concepts), so dialogue goals represent an integration of the available information, thus providing a more reliable source of knowledge. Therefore the estimation of the goal-based LM is more accurate than the concept-based one.

We are now working on different clustering approaches in such a way that different dialogue concepts build a single LM. This way we will have a reduced set of LM to interpolate with the background one, each of that trained with more examples.

In this work we have focused on rescoring each sentence with the semantic and goal knowledge of the same sentence. However, this will not be the actual behaviour of the dialogue system. Our final goal will be to dynamically modify the LM to better recognize the utterance that a speaker is saying to the system using the information of previous dialogue acts, improving the response of the dialogue system.

6. Acknowledgements

This work has been partially funded by the Spanish Ministry of Science and Innovation under contract TIN2008-06856-C05-05 (SD-TEAM UPM).

7. References

[1] Bellegarda, J.R., "Statistical language model adaptation: review and perspectives", Speech Communication 42:93-108, Elsevier, 2004.

[2] Kneser, R. and Steinbiss, V., "On the Dynamic Adaptation of Stochastic Language Models", Proc. ICASSP, II:586–589, 1993.

[3] Hsu, B-J., "Generalized Linear Interpolation of Language Models", Proc. ASRU, 136–140, 2007.

[4] Liu, X., Gales, M.J.F. and Woodland, P.C., "Context Dependent Language Model Adaptation", Proc. Interspeech, 837–840, 2008.

[5] Riccardi, G. and Gorin, A.L., "Stochastic Language Adaptation over Time and State in Natural Spoken Dialog Systems", IEEE Trans. Speech and Audio Proc., 8(1):3–10, 2000.

[6] Visweswariah, K. and Printz, H., "Language Models Conditioned on Dialog State", Proc. Eurospeech, 251–254, 2001.

[7] López-Cózar, R. and Callejas, Z., "Combining language models in the input interface of a spoken dialogue system", Computer, Speech and Language 20:420–440, Elsevier, 2006.

[8] Fernández, F., Ferreiros, J., Sama, V., Montero, J.M., San-Segundo, R. and Macías-Guarasa, J., "Speech interface for controlling a hi-fi audio system based on a bayesian belief networks approach for dialog modeling", Proc. Interspeech, 3421–3424, 2005.

[9] Fernández, F., Ferreiros, J., Córdoba, R., Montero, J.M., San-Segundo, R. and Pardo, J.M., "A Bayesian Networks Approach for Dialog Modeling: the Fusion BN", Proc. ICASSP, 4789–4792, 2009.