A New Technique Based on *Augmented Language Models* to Improve the Performance of Spoken Dialogue Systems

R. López-Cózar[1], *D. H. Milone*[2]
[1]Dpto. Electrónica y Tecnología de Computadores, Granada University, Spain, rlopezc@ugr.es
[2]Universidad Nacional de Entre Ríos, Argentine, d.milone@ieee.org
Tel.: +34-958-243271, Fax: +34-958-243230

Abstract

This paper presents a new technique that aims to improve the performance of spoken dialogue systems by using the so-called *augmented language models*. We define an augmented language model as a compound of a language model and a set of values concerning parameters that can influence the speech recognition when the language model is used. The diverse language models used by a dialogue system can be very different, in terms of perplexity for example. Then, the aim of the technique is to find and use the combination of values concerning the different parameters that leads to the best recognition results when the different language models are used by a dialogue system. The technique has been applied to a dialogue system for the fast food domain. The results show that when the augmented language models are used the system's performance is enhanced. In the experiments we have achieved a reduction of 9,33% in the word error rate and an increment of 11,26% in the sentence understanding.

1. Introduction

Recent advances in speech technology have made possible to develop spoken dialogue systems for a variety of applications [1]. These systems allow users to carry out some tasks and obtain information using the human language as interaction mode. In despite of the advances made in the last decade, the performance of such systems under real world conditions still presents severe limitations. On the one hand, these systems are difficult to specify, design, develop and maintain. The development requires expertise in multiple domains, mainly, speech recognition, natural language understanding and generation, dialogue management and speech synthesis [2]. Most developed dialogue systems are restricted to specific domains. They are based on many application-specific collected examples and are specialized to carry out determined tasks. Using the domain knowledge, the perplexity of the task to be performed by a dialogue system can be reduced and acceptable results can be obtained. On the other hand, some of these systems are difficult to use, specially for the non-experienced users. Finally, there are several drawbacks that difficult a better performance of these systems, such as the management of large-scale vocabularies in real time, the different pronunciations of words, the miscommunication between systems and users, the use of spontaneous speech, etc. Because of these problems, among others, several efforts have been made to develop tools for improving the technology employed [3].

This paper is organized as follows. In Section 2 we present the technique proposed in this paper. In Section 3 we introduce a brief description of the dialogue system used in the experiments. In Section 4 we describe the utterance corpus used. In Section 5 we present the experiments carried out to test the technique using the dialogue system. Finally, in Section 6 we present the conclusions and indicate possibilities for future work.

2. The new technique: The Augmented Language Models

We define an augmented language model (ALM) as a compound of a language model l and n values corresponding to different parameters, $v_j \in P_j$.

$$\text{ALM} = (l, v_1, v_2, ..., v_n)$$

The P_j represent parameters that can influence the speech recognition, as for example, pruning threshold, insertion penalty, grammar weight, etc. Each parameter P_j is defined as $P_j=\{p_1, p_2, ... , p_{Mj}\}$ where M_j is the number of test values in P_j. By preliminary experiments using a determined speech recognizer, a dialogue system designer can consider these values as appropriate candidates to provide good recognition results. The technique can be applied to any spoken dialogue system that uses a particular language model for every dialogue state, being the language model determined by the current dialogue system's prompt. When a system of this kind asks a user for a phone number, for example, only phone numbers -and some other utterance types for the dialogue management- can be recognized. We call this type of system a *state-based* dialogue system. The advantage of using these systems is that the recognition phase focuses only on the current context of the conversation. Then, the set of possible utterances that are considered at a given moment by the recognizer is much more reduced, leading to better recognition results. However, the drawback is that these systems impose a considerable restriction to the free interaction of users.

Figure 1 shows the procedure to create the augmented-language models for state-based dialogue systems. The procedure aims to find the combination of values concerning the different parameters that leads to the best system's performance in terms of speech recognition, and consequently, in terms of speech understanding. The procedure to find the augmented-language models for a state-based dialogue system is as follows. We must initially determine the language models l_i and the parameters P_j that will be used to create the ALMs. For example, the language models can be bigrams [4] associated to the different states of the dialogue, in such a way that a specific bigram is used at specific moment to recognize a user utterance. The parameters can be any of those mentioned above, i.e. pruning threshold, insertion penalty, grammar weight, or other. Let us define Σ as a combination of values of the different parameters, i.e., $\Sigma = (v_1, v_2, \dots, v_n)$ where $v_1 \in P_1, v_2 \in P_2 \dots v_n \in P_n$. The procedure proposes to evaluate the dialogue system's speech recognizer using l_i, with all the possible combinations Σ. If the best result for a language model l_i is achieved for a combination Σ_i, then the compound (l_i, Σ_i) is considered the augmented-language model ALM$_i$ for l_i. This augmented language model represents an optimization of the initial language model l_i. The goal of the procedure is to transform all the initial language models l_i used by a state-based dialogue system into the augmented language models ALM$_i$, and make the system use the ALM$_i$ instead of the initial language models to optimize its performance.

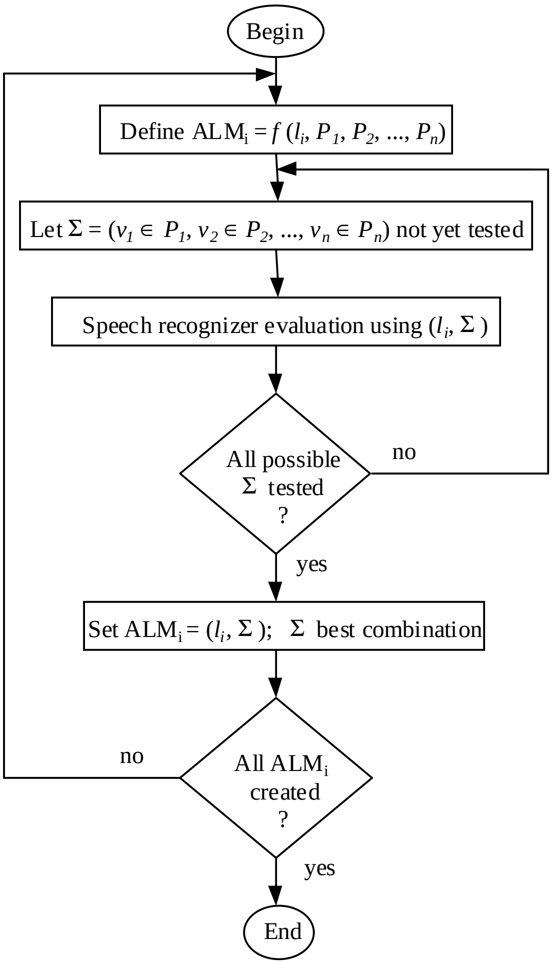
Begin
Define ALM$_i$ = f (l_i, P_1, P_2, ..., P_n)
Let $\Sigma = (v_1 \in P_1, v_2 \in P_2, \dots, v_n \in P_n)$ not yet tested
Speech recognizer evaluation using (l_i, Σ)
All possible Σ tested ?
no
yes
Set ALM$_i$ = (l_i, Σ); Σ best combination
no
All ALM$_i$ created ?
yes
End

Figure 1: Procedure to create the augmented-language models for state-based dialogue systems

3. Brief description of the dialogue system

We have tested the technique proposed in this paper using a state-based dialogue system, named SAPLEN, under development in our lab to deal with the product orders and queries of fast food restaurants' clients using the telephone [5]. Figure 2 shows the system structure. In order to optimize the speech recognition, the system's dialogue manager selects a language model (bigram) according to the state of the dialogue.

The system uses a mixed-initiative dialogue management strategy [7]. Generally, the system takes the initiative during the interaction with users. However, users can take the control to correct errors and query for information whenever they want. The system uses both implicit and explicit confirmations [8]. Implicit confirmations are used to confirm the data extracted from the previous user utterance. For example, when a user orders for a product, the system repeats the order's data items in its next response and asks the user to confirm the data items. From this feedback the user can know whether the system understood the order correctly and can make a correction if necessary. Explicit confirmations are used at the end of the conversation, in order to confirm once more all the data previously extracted from the interaction with the user. The dialogue system's vocabulary is about 2,000 words, including restaurant-product names, numbers, names of streets, avenues, squares, etc. The system uses an implicit recovery strategy that, in some occasions, permits to obtain the correct semantic interpretation in spite of the fact that some words in the recognized utterance might have been wrongly recognized [9].

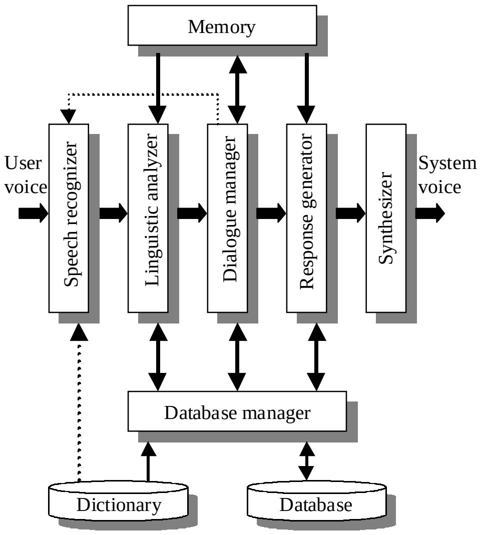

Figure 2: The SAPLEN dialogue system

4. The utterance corpus

We have considered seven utterance types in the fast food domain and have recorded 250 utterances for each one (see Table 1). So that, the utterance corpus contains 1750 utterances. Among them, 1050 have been used to create the ALMs and the remaining 700 utterances have been used for testing the SAPLEN system's speech recognizer using the ALMs created. The utterances have been recorded by 9 speakers. Four of them speak standard Spanish, four speak Spanish from southern Spain, and one speaker is a Japanese female who speaks Spanish. The utterances have been recorded under non-noisy lab conditions using a PC computer and 16 bits/sample at 8KHz. The dialogue system uses a specific language model to recognize utterance types 1, 3, 4, 5, 6 and 7. For example, when the system prompts for a telephone number, the recognizer uses the language model that corresponds to this utterance type. However, non specific language model has been defined for the corrections (utterance type 2) as users may correct system's errors anytime during a conversation. The utterance type 1 (Confirmation) is employed by users to explicitly confirm data (phone number, post code, address, ordered products, price payable and estimated delivery time). For this data, the system asks users for yes/no answers. For example, to confirm a telephone number, the system generates confirmation prompts such as *"Did you say 9, 5, 8, 1, 7, 1, 3, 2, 8? Please answer yes or no"*. Users can employ the utterance type 2 (Correction) to repair the recognition or understanding errors made by the dialogue system. Users can utter a variety of expressions to correct these errors, which makes the system to return to a previous state of the dialogue and ask the user again for the corresponding data items.

Type	**Utterance**
1	Confirmation
2	Correction
3	Post code
4	Product order
5	Telephone number
6	Address
7	Query

Table 1: Utterance types considered in the fast food domain

The utterance type 3 (Post code) consist of five digits. Users employ this utterance type to indicate the post code corresponding to their address. The product orders (type 4) are utterances employed by users to order for fast food products (foods and/or drinks). The telephone numbers (type 5) consist of nine digits. User can employ isolated digits or several combinations of digits, for example, *9, 5, 8, 17, 13, 28*. The utterance type 6 is concerned with the user address. Using this utterance type a user can inform the system about the street, building number, floor, etc. where (s)he lives in. Users can ask a variety of questions to the system (utterance type 7), for example, concerning available products *("What can I have to drink?")*, prices (*"How much is a ham and cheese sandwich?"*), ingredients *("What is a cantábrico sandwich?")*, etc.

5. Experiments

In the experiments, we have only used one parameter that affects speech recognition. This parameter is the pruning threshold (PT). So that, the ALMs we have created are as follow:

$$\mathrm{ALM}_i = (l_i, v_i)$$

where l_i represents a language model (a bigram in the case of the SAPLEN system) and v_i represents a value of the pruning threshold PT. This value represents a trade-off between recognition time and word error rate (WER).

We have carried out two types of evaluation. Firstly, we have evaluated the performance of the dialogue system without using ALMs. In this case, the recognizer used a fixed threshold that was independent of the language model selected by the dialogue manager. The performance of the recognizer has been measured in terms of recognition time and WER, and the performance of the system's linguistic analyzer has been measured in terms of sentence understanding (SU) [10]. Figure 3 shows the results obtained concerning: (a) recognition time, (b) WER and (c) SU. The results have been obtained using the six values of the pruning threshold (PT) considered good candidates to provide acceptable results (10, 20, 30, 40, 50 and 60).

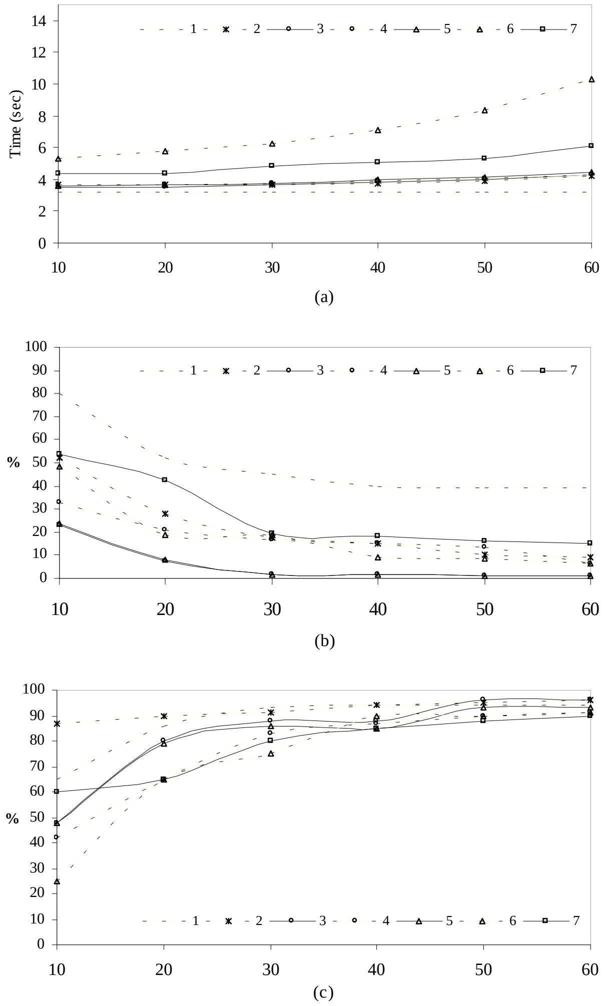

Figure 3: Evaluation results for several PT. (a) Recognition time. (b) WER. (c) SU.

To carry out this initial evaluation, we have used the 1750 utterances in the whole utterance corpus. Table 2 sets out the average results obtained.

PT	**TIME**	**WER**	**SU**
10	3,9	44,91	53,57
20	3,98	25,45	75,71
30	4,17	17,25	85,14
40	4,41	14,32	89
50	4,72	12,8	92,29
60	5,25	11,15	93

Table 2: Average results obtained without using ALMs

Secondly, we have created the six ALMs that correspond to the six language models previously defined to deal with the seven utterance types described in Table 1. To do so, we have followed the procedure shown in Figure 1. We have used 175 utterances for creating each ALM, which makes a total of 1050 utterances. The result of the procedure is six pairs (*language model*, *value*). The *value* is the most suited pruning threshold for the *language model*. It represents a trade-off between recognition time and word error rate when the *language model* is used. In order to create the ALMs, we have taken into account the least recognition time that provides good results in terms of word accuracy. Table 3 sets outs the ALMs obtained. Every ALM can be used to cope with a particular utterance type uttered at a determined state of the dialogue. The possible corrections of users (utterance type 2) can be recognized in any state of the dialogue.

Dialogue state	ALM
Confirmation	(l_1,50)
Post code	(l_2,50)
Product order	(l_3,60)
Telephone number	(l_4,50)
Address	(l_5,50)
Query	(l_6,60)

Table 3: ALMs obtained

Finally, we have evaluated the system using the ALMs instead of the initial language models. In this evaluation the recognizer did not use a fixed threshold independent of the language model selected by the dialogue manager. Instead, it used the threshold associated to each language model, considering the six ALMs previously created (see Table 3). We have used the 700 test utterances, different from the 1050 utterances used to create the ALMs. Table 4 shows the average results obtained concerning recognition time, WER and SU, using and not using the ALMs.

	TIME	WER	SU
Not using ALMs	4,41	20,98	81,45
Using ALMs	4,87	11,65	92,71

Table 4: Average results using and not using the ALMs

6. Conclusions and future work

As can be observed in Table 4, the experimental results show that when the ALMs are used the SAPLEN dialogue system's performance in terms of WER and SU is enhanced. When the ALMs are used, the recognition time is increased in 0,46 sec., the WER is reduced in 9.33% and the SU is increased in 11,26%. The ALMs are optimizations of the initial language models. In the experiments presented in this paper, the ALMs have been obtained using the most suited pruning threshold for every initial language model.

Several aspects concerning the technique proposed in this paper must be improved. On the one hand, we have used a relatively reduced number of utterances to create the ALMs and test the dialogue system. Additionally, these utterances have been recorded under non-noisy lab conditions. In order to improve the performance of the system before it is set up into the real world, we must use more utterances for refining the ALMs, particularly, utterances recorded under noisy conditions should be taken into account.

7. References

[1] N. O. Bernsen et al., "Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems", Proc. of Eurospeech 99, pp. 1147-1150

[2] N. Dahlbäck, A. Jönsson, "Knowledge Sources in Spoken Dialog Systems", Eurospeech '99, pp. 1523-1526

[3] M. F. McTear, "Software to Support Research and Development of Spoken Dialogue Systems", Proc. of Eurospeech 99, pp. 339-342

[4] Siu M., Ostendorf M., "Variable N-Grams and Extensions for Conversational Speech Language Modeling", IEEE Trans. on Speech and Audio Proccessing, January 2000, vol. 8, pp. 63-75

[5] López-Cózar R. et al., "Evaluation of a Dialogue System Based on a Generic Model that Combines Robust Speech Understanding and Mixed-Initiative Control", Proc. of LREC '2000, Athens (Greece) pp. 743-748

[6] Rabiner L., Juang B. H., Fundamentals of Speech Recognition, Prentice-Hall, 1993

[7] Relaño Gil J. et al., "Flexible Mixed-Initiative Dialogue for Telephone Services", Eurospeech '99, pp. 1179-1182

[8] Lavelle C. A. et al., "Confirmations Strategies to Improve Correction Rates in Telephonic Inquiry Dialogue System", Eurospeech '99, pp. 1399-1402

[9] Albasano D. et al., "DIALOGOS: A Robust System for Human-Machine Spoken Dialogue on the Telephone", Icassp '97, pp. 1147-1150

[10] C. Müller, K. Schröder, "Standardised Speech Interfaces -Key for Objective Evaluation of Recognition Accuracy", Proc. of Eurospeech 99, pp. 931-934