LANGUAGE MODEL ADAPTATION FOR CONVERSATIONAL SPEECH RECOGNITION USING AUTOMATICALLY TAGGED PSEUDO-MORPHOLOGICAL CLASSES

C. Crespo, D. Tapias, G. Escalada, J. Álvarez
Speech Technology Group
Telefónica Investigación y Desarrollo, S. A. Unipersonal
C/ Emilio Vargas, 6
28043 - Madrid (Spain)
crespo@craso.tid.es

ABSTRACT

Statistical language models provide a powerful tool to model natural spoken language. Nevertheless it is required a large set of training sentences to reliably estimate the model parameters. In this paper we present a method to estimate n-gram probabilities from sparse data. The proposed language modeling strategy allows to adapt a generic language model (LM) to a new semantic domain with just few hundreds of sentences. This reduced set of sentences is automatically tagged with eighty different pseudo-morphological labels, and then a word-bigram LM is derived from them. Finally, this target domain word-bigram LM is interpolated with a generic back-off word-bigram LM, which was estimated using a large text database. This strategy reduces a 27% the word error rate of the SPATIS (SPanish ATIS) task.

1. INTRODUCTION

It is well known the importance of the language model (LM) in the performance of continuous speech recognizers, hence, a well estimated LM is always convenient. This is achieved by using a large set of training sentences belonging to the task semantic domain and smoothing techniques to model poorly estimated n-grams.

The back-off method [1] smooths n-gram probabilities using (n-1)-gram probabilities. It basically, given an initial maximum likelihood estimation of a n-gram, smooths the model by discounting a certain probability mass from the seen events and then redistributes this mass on unseen n-grams according to their (n-1) order probability estimates.

The problem we are addressing is related with the difficulty of getting large enough textual corpus for a new semantic domain, this difficulty increases if we want to model a LM for person-machine dialogues. In this case, it is usual to have no more than some hundreds of sentences and the creation of a large enough text database takes a long time. For this reason, we are developing a new strategy for LM building, which tries to overcome this problem. The key idea is to build a generic LM using as many sentences as possible, and then adapting it to the target domain by using a small set of target domain sentences. The target domain sentences can be easily generated by implementing a “wizard of Oz” application.

We propose a method that combines a back-off word-bigram LM with a back-off class-bigram LM. It can be easily generalized to the trigram case, though all the experiments have been carried out with word-bigrams. The back-off word bigram LM is obtained with a large generic text database while the back-off class-bigram LM is obtained from some hundreds of tagged sentences belonging to the new semantic domain.

The word to class mapping is carried out by a tagger [2] developed at Telefónica Investigación y Desarrollo (TID) for the Spanish text to speech converter. The tagger assigns one out of eighty pseudo-morphological classes to each word of

the sentence, and can be used in any semantic domain. As there are many words that can perform several syntactical functions, the tagger is very useful to assign the proper tag to each word and to solve the ambiguity. Hence, the tag assignation depends on the function the word is performing in the sentence. The tag set used in the text to speech is designed to reflect the syntactic role of the words in the sentence. But for the language modeling task it was very useful to include number and gender information, in order to exploit the agreement among components of nominal sintagms.

This paper is organized as follows: In section 2 we describe the proposed method. Section 3 presents the application of this method for LM adaptation and shows a simplification of this procedure. Finally, section 4 presents the experiments, results and conclusions.

2. DESCRIPTION

Figure 1 shows the block diagram of the training procedure. The large text database is a collection of sentences gathered from as many different sources as possible. The training procedure follows two different branches:

(a) The left branch takes the database and produces a back-off word-bigram LM using the Statistical Language Model Toolkit (SLMT) developed at Carnegie Mellon University (CMU) [3][5]. This process is constrained by the dictionary of the continuous speech recognizer, so that words that do not belong to the dictionary are considered as "unknown".

(b) The right branch starts by tagging the large text database and producing the mapping frequencies of each word-class pair. The resulting tagged database is then processed by the SLMT in order to obtain a back-off class-bigram LM, which is composed of the unigram and bigram estimates of the eighty pseudo-morphological classes.

Later, it is built a word-bigram LM based on classes using both the mapping frequencies and the back-off class-bigram LM. The used formula to produce the word-bigram LM is given by the following expression:

$$p(\omega_2|\omega_1) = \sum_{i \in \Omega_1} \sum_{j \in \Omega_2} p(\omega_2|j) \cdot p(j|i) \cdot p(i|\omega_1) \qquad (1)$$

where:

Ω_n is the set of possible classes for ω_n

$p(j|i)$ is the class back-off probability for classes "i" and "j".

$p(\omega_2|j)$, $p(i|\omega_1)$ represent the probability of a word given a class and viceversa, and are obtained from the mapping frequency table.

This formula applies the chain rule to provide the word-bigram probabilities based on the a posteriori and class-bigram probabilities.

Finally, the LM interpolation module plays an important role since it combines the left and right branch language models. The goal of the final interpolation step is to exploit the synergy of both language models: The left branch estimates accurately the most frequent bigrams while the right branch does it for the less frequent and unseen bigrams.

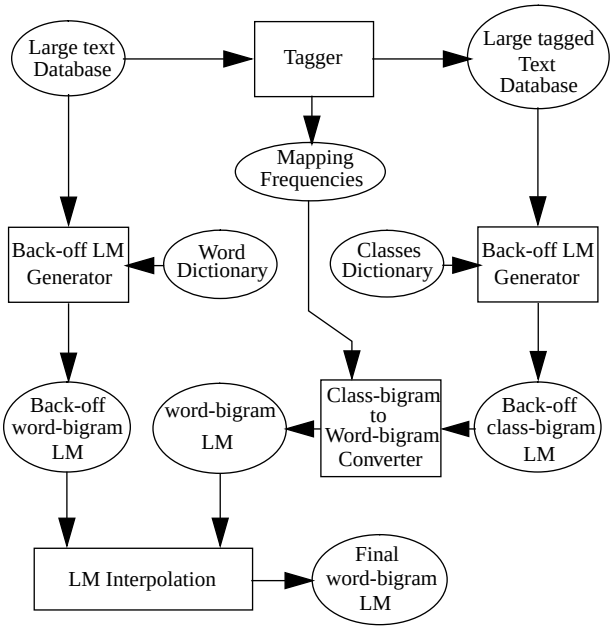

Figure 1: Block diagram of the LM generation process.

3. LANGUAGE MODEL ADAPTATION

A slight modification of the previously proposed method allows to adapt a general LM to a particular task. Figure 2 shows the block diagram of the modified method: The only difference between both methods is the right branch of the process. The set of sentences used to build the backoff class-bigram belongs to the semantic target domain, and can be gathered with little effort (just a few hundred sentences). This reduced set of sentences reflects the semantic structure of the domain we want to model.

The overall procedure can be summarize in four steps:

1.- A word-bigram backoff LM is built from a set of sentences of Spanish newspapers (left branch).

2.- A class-bigram backoff LM is built using the small set of sentences belonging to the target semantic domain (right branch). This set is composed of just a reduced set of sentences automatically tagged by the tagger developed at TID [2].

3.- A class-based word-bigram LM is built using the following expression for the bigram case (right branch), which can be simplified by considering only the most likely class for each word:

$$p(\omega_2|\omega_1) = p(\omega_2|j) \cdot p(j|i) \qquad (2)$$

where:

$p(j|i)$ is the class back-off probability for classes “i” and “j”.

$p(\omega_2|j)$ is obtained from the mapping frequency table, where “j” is the most likely class for word ω_2.

4.- Finally, an interpolated class-based LM is built. This is carried out by interpolating the LMs obtained in steps 1 and 3.

The right branch estimates accurately the less frequent and unseen bigrams and unigrams, which are mostly domain dependent and hence they help to adapt the general LM to the specific target semantic domain. The left branch just estimates a general LM that is constrained by the dictionary of the speech recognizer.

4. EXPERIMENTS AND RESULTS

The three experiments we present are devoted to show the improvement achieved by the proposed language modeling strategy when the available training sentences are not close to the target semantic domain.

Two text databases have been used to train the language models: (a) The Large Text Database (LTD) which is composed of 60000 sentences in Spanish (2 M words). They were picked up from the opinion section of Spanish newspapers. (b) The Task Dependent Database (TDD), which is composed of 300 sentences (20 K words). They were picked up from the SPATIS task.

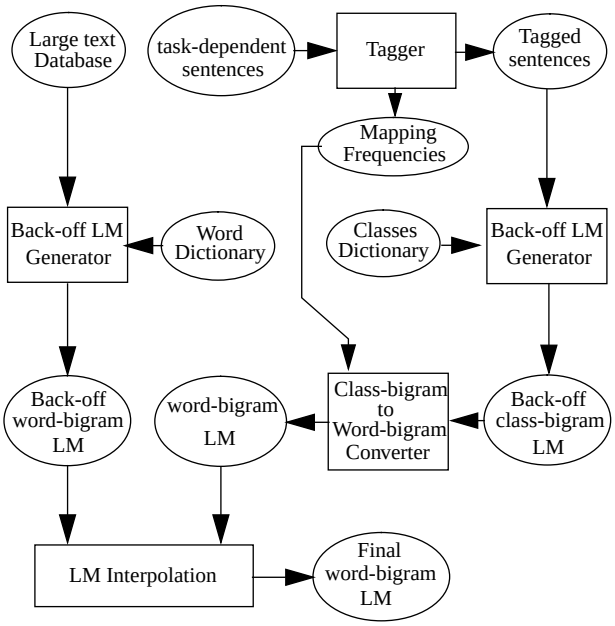

Figure 2: Block diagram of the LM adaptation process.

The speech recognizer has been tested using the SPATIS database [4]. The evaluation set is composed of 2300 read sentences. Three different language models have been tested:

(a) Baseline LM (BLM): It is a back-off word-bigram LM which has been trained using the LTD training set. As it is explained in section 2, the training process of this LM is constrained by the dictionary of the continuous speech recognizer. This constraint has demon-

strated to increase the word accuracy of the speech recognizer.

(b) Class-based trained LM (CTLM): It is a word-bigram LM that has been trained following the method described in section 2. It has been trained with the LTD training set.

(c) Adapted LM (ALM): It is a word-bigram LM which has been adapted using the procedure presented in section 3. It has been trained using the LTD and TDD training sets.

Table 1 presents the results obtained with the three different language models as well as the perplexity of each of them. As can be observed, our experiments show a 27% reduction in the word error rate when we use the adapted LM (ALM) instead of the baseline LM. Additionally, the CTLM reduces a 17.5% de word error rate of the BLM. As can be observed, the perplexity is appreciably reduced by the CTLM and ALM methods.

TABLE 1. Experimental Results

	BLM	CTLM	ALM
WER (%)	20.0	16.5	14.6
Perplexity	873	247	237

This is a very promising technique to adapt language models to different semantic domains, even though the results are still far from the 5% word error rate obtained with a LM trained with task-dependent sentences. We are currently introducing some improvements into the proposed method to take into account language knowledge during the recognition process. We are also studying the optimum number of adaptation sentences to reach a trade-off between the number of adaptation sentences and the word accuracy.

This method can also be implemented using a bootstrapping approach: A reduced set of sentences is used to adapt the general LM and facilitate a rapid prototyping of the natural language application. Then the LM can be incrementally adapted with more realistic data as more target domain sentences become available from the field trials.

ACKNOWLEDGEMENTS

We want to thank Ronald Rosenfeld (CMU) for providing us with the Statistical Language Modeling Toolkit and for his support. We are also grateful to the Text to Speech Conversion and the Speech Recognition Groups of Telefónica Investigación y Desarrollo for their contributions and suggestions in this work.

REFERENCES

[1] S. M. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer”, IEEE Trans. ASSP-35, 3, pp. 400-401, March 1987.

[2] M. A. Rodriguez, J. G. Escalada, A. Macarrón and L. Monzón, “AMIGO: Un Conversor Texto-Voz para el Español”, Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN’92, Vol. 13, pp. 389-400. Sept. 1992.

[3] R. Rosenfeld, “Adaptive Statistical Language Modeling: A Maximum Entropy Approach”. Doctoral Thesis. Carnegie Mellon University, April 1994.

[4] C. de la Torre, L. Hernández, D. Tapias, “CEUDEX: A Database Oriented to Context-Dependent Units Training in Spanish for Continuous Speech Recognition”, EUROSPEECH’95, pp.845-848, Sept. 1995.

[5] R. Rosenfeld, “The CMU Statistical Language Modeling Toolkit, and its Use in the 1994 ARPA CSR Evaluation”, In Proc. ARPA Spoken Language Technology Workshop, Austin, TX, January 1995.

[6] F. Jelinek, “Self Organized Language Modeling for Speech Recognition”, in Readings in Speech Recognition, Alex Waibel and Kai-Fu Lee (Eds.), Morgan Kaufmann, 1989.