Language Models in Automatic Speech Recognition

Ciro Martins, António Teixeira, João Neto[†]

Resumo - **O presente artigo descreve o trabalho desenvolvido com o objectivo de melhorar o desempenho da componente modelo de linguagem de um sistema de reconhecimento de fala contínua para a língua Portuguesa. Como modelo de base, utilizou-se um sistema de reconhecimento de fala de grandes vocabulários desenvolvido para uma tarefa de reconhecimento de notícias (Broadcast News). Foram analisadas duas metodologias diferentes com o objectivo de aumentar a eficácia do sistema: a utilização de maiores quantidades de dados para melhor estimação dos parâmetros associados ao modelo de linguagem, e a utilização de diferentes técnicas de "pruning" e "discounting" dos referidos parâmetros. Os resultados mostram que com a utilização de maiores quantidades de dados se obtiveram ligeiras melhorias a nível da taxa de eficácia de reconhecimento (cerca de 5%). Aplicando uma técnica de "pruning" baseada no conceito de entropia, obteve-se uma redução significativa da dimensão do modelo de linguagem (reduções de 30% ou mais), com um ligeiro incremento dos valores da perplexidade e taxa de erro ao nível da palavra.**

Abstract - **In this paper we describe the work done with the updating and improvement of the language model component of a continuous speech recognition system for the Portuguese language. As a baseline system we used a large vocabulary speech recognition system for the Portuguese language, developed for a Broadcast News (BN) recognition task. Two sources of performance improvement have been studied: the inclusion of more training data to better estimate the language model parameters, and the use of different discounting and pruning techniques. The results show that using more training data helped to achieve a small relative improvement in recognition accuracy (about 5%). Applying an entropy based pruning technique one can get up to more than 30% size reduction with a slightly increase on perplexity and WER.**

I. INTRODUCTION

Statistical language modeling has many applications in a large variety of areas, including speech recognition, optical character recognition, machine translation, spelling correction, etc. Despite all the research done on the last two decades, N-gram language models still dominate as the technology of choice for state-of-the-art speech recognizers.

Typically, N-gram models for large vocabulary speech recognizers are trained on hundred of millions or billions of word strings. In constructing such kind of models, we usually face two problems. First, the large amount of training data can lead to models too large for real applications. On other hand, to train a specific domain model, we must deal with the data sparseness problem, because large amount of domain specific data are not available.

To overcome this kind of problems, many different approaches have been suggested. Smoothing techniques are usually used to better estimate probabilities when there is insufficient data to estimate probabilities accurately [1]. In case where small amount of in-domain data is available, the use of mixture of language models by means of linear interpolation proved to increase the quality of language models [2]. On the other side, some form of model size reduction is critical for practical applications, especially when the model is trained with large amount of data. Many different pruning techniques have been proposed which leads to significant model size reduction without decreasing their performance [3] [4].

In the reminder of this paper we describe the work done with the updating and improvement of the language model component of a continuous speech recognition system for the Portuguese language. As a baseline system we used the work presented in [5] that we briefly describe in section 2. In section 3 we describe the datasets we have used to train and evaluate the new models we obtained applying some of the techniques referenced before. In section 4 we summarize some results in terms of perplexity and word error rate (WER), drawing in section 5 and 6 some conclusions and future work to be done.

II. BASELINE SYSTEM

As the starting point for the work presented in this paper, we used the same datasets and system reported in [5]. This is a large vocabulary speech recognition system for the Portuguese language, used for a Broadcast News (BN) recognition task.

[†] L2F – Spoken Language Systems Lab; INESC-ID/IST, Lisbon.

A. Acoustic Component

The baseline recognizer AUDIMUS [6], used in this task, is a hybrid HMM/MLP system. It uses three MLPs, each of them associated with a different feature extraction process, where the MLPs are used to estimate the context-independent posterior phone probabilities given the acoustic data at each frame. The phone probabilities generated at the output of the MLPs classifiers are combined using an appropriate algorithm [6]. All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus the silence and breath noises. The training and development of this system was based on the European Portuguese ALERT BN database (ALERT BN) [5]. The train of the recognition system was done using 45 hours of audio data. For system evaluation there exist two different sets: a development set comprising of approximately 6 hours and half, and an evaluation set comprising of 4 hours and half of audio data.

B. Language Model Component

The language model component of this system has been created using two different sources (table 1): a text corpus collected daily from the online Portuguese newspapers Web editions, and the training set transcriptions of ALERT BN database.

Source	# Sentences	# Words
newspapers set	24.0M	434.4M
ALERT BN training set	9K	142K

Table 1: Size of the corpuses used in language model training

At this time the newspapers set included texts collected from different Portuguese newspapers ("A Bola", "Diário de Notícias", "Diário Económico", "Expresso", "Expresso Diário", "Jornal de Notícias", "O Jogo", "O Independente", "O Público"), since 1991 until the end of 2001. The ALERT BN transcriptions used to train the language model component include only part of the ALERT BN training transcriptions available at that time. The complete and final set is the one described in the next section.

From this two sources and using the CMU Cambridge Toolkit [7], two different language models have been generated. From the newspapers set a backoff 4-grams LM has been generated using the absolute discounting method and applying cutoff values of 2, 3 and 4 respectively for 2-grams, 3-grams and 4-grams. From the ALERT BN training set a backoff 3-grams LM has been generated using the absolute discounting method without applying any kind of cutoffs. Finally, the two models were combined by means of linear interpolation, generating a mixed model referred here as MIX_BASELINE. The optimal interpolation weights obtained were 0.829 for the newspapers set component and 0.171 for the ALERT BN training set component.

C. Vocabulary and Lexicon

Currently the vocabulary size is limited to 57,564 words (referred here as 57k). This vocabulary was first created using 56k different words selected from the newspapers set according to their weighted class frequencies of occurrence. Different weights were used for each class of words. All new different words present in the training data transcriptions of ALERT BN database were added to the vocabulary giving a final vocabulary of 57,564 words.

The pronunciation lexicon was built from the vocabulary, giving a total of 65,585 different pronunciations.

D. Dynamic Decoder

The decoder used under this baseline system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [8]. In this approach, the decoder search space is a large WFST that maps observation distributions to words. This WFST consists of the composition of various transducers representing components such as: the acoustic model topology H; context dependency C; the lexicon L and the language model G. The search space is thus $H \circ C \circ L \circ G$, which is built "on-the-fly" in opposition to traditional approaches that compile it outside of the decoder, using it statically during the decoding process.

III. NEW DATASETS FOR LANGUAGE MODELING

To update and improve the language model component of the baseline system described before, we have collected more texts from the newspapers online editions until the end of 2003, and used the final ALERT BN training set that is now available. In table 2 we summarize the size of these new sets that have been used to generate and test the new language models we have developed. For the new ALERT BN training set which has 26,715 different words in a total of 531,757 words, the number of Out-Of-Vocabulary words (OOVs) using the 64k word vocabulary is 6138, representing an OOV word rate of 1,15%.

Source	# Sentences	# Words
newspapers set	38.8M	604.2M
ALERT BN training set	34K	531.7K

Table 2: Size of new training sets used in language model training

To evaluate the language models performance we used the ALERT BN evaluation set. To estimate some parameters like the ones necessary for the linear interpolation process we used the ALERT BN development set as a held-out corpus. In table 3 we describe these two corpuses.

Source	# Sentences	# Words	Duration (audio)
development set	4,194	66,495	6h 24m
evaluation set	3,125	47,473	4h 30m

Table 3: Size of ALERT BN development and evaluation sets

For the development set which has 8,538 different words in a total of 66,495 words, the number of OOVs words (OOVs) using the 57k word vocabulary is 879, representing an OOV word rate of 1.32%. The Evaluation set has 7,000 different words in a total of 47,473 words, having 675 OOVs, which represents an OOV word rate of 1.42%.

IV. EXPERIMENTAL RESULTS

The most common metric for evaluating a language model is perplexity. It is often used as a language model quality measure as it tests its capability to predict an unseen text, i.e., a text not used in model training. Formally, the word perplexity PP of a model relative to a text with n words is defined as:

$$PP = 2^{-\left(\frac{1}{n}\right)\log P(w_1 \ldots w_n)} \quad (1)$$

However, perplexity metric does not take into account the acoustic similarity between words. This means that lower perplexity values may not result in lower word error rate (WER) during the recognition process. For that reason, it is usual to use WER as another metric to evaluate the language model performance over all the recognition system.

For the experimental results we present in this paper we used both metrics to consistently evaluate and compare the relative language models performance. The reported results were conducted in the ALERT BN Evaluation set, using the ALERT BN Development set as a held-out set to estimate and optimize some parameters like the ones used by the interpolation process.

To generate the language models used in this work and evaluate their performance in terms of perplexity values, we used the SRI Language Modeling Toolkit (version 1.4) [9].

All the experiments were done using the same 57,564 word closed-vocabulary. End-of-sentence symbols were included in perplexity computations, but out-of-vocabulary words were not. Related with recognition results, we used all the evaluation set, which means the results take into account the effect of OOVs during the recognition process, i.e., since we are using a closed-vocabulary all the OOVs are misrecognized by the system.

A. Perplexity Results

First of all, we started by computing the perplexity value for the baseline language model (MIX_BASELINE) using the SRI Toolkit. For this model we obtained a perplexity value of 117.5. This was a reference value, being used to make performance comparison to the new language models. In [5] one can realize a different perplexity value of 139.5 for the same baseline language model, which is due to the fact that we are now using SRILM Toolkit instead of CMU Toolkit. These toolkits treat sentences clues in a different way when measuring text perplexities.

To evaluate the effect of using more data we trained the language models (the newspapers model and the ALERT BN model) based on the new training sets. This train was done using the same conditions as the baseline ones, i.e., we used the same discounting method (absolute discounting), the same model order (4-grams for the newspapers model and 3-grams for the ALERT BN model) and the same cutoff values.

For the newspapers language model we generated two different versions: one using the newspapers data available until the end of 2001 (referred as NP_2001; line 1 of table 1) and another one using all the newspapers data available until the end of 2003 (referred as NP_2003; line 1 of table 2). For the ALERT BN language model, and since we didn't have the partial training set used to generate the baseline ALERT BN model, we only generated one version using the final ALERT BN training set (referred as ALERT_BN_ALL; line 2 of table 2). Finally we generated two mixed models: one using NP_2001 and ALERT_BN_ALL (referred as MIX_2001) and another one using NP_2003 and ALERT_BN_ALL (referred as MIX_2003).

From table 4 we can realize a small decrease (less than 2%) on the perplexity value when we used more training data. Comparing baseline LM perplexity to the new models perplexity we conclude that the biggest improvement (almost 4.5%) is due to the use of more training data related with the domain. In fact, the final ALERT BN training set is 4 times bigger that the one used to generate the baseline model (MIX_BASELINE). The column "Param" gives the number of stored N-grams (only the last order).

Models	Interpol. weights		PP	Param
	α_{NP}	α_{BN}		
MIX_BASELINE	0.829	0.171	117.5	6,731,820
NP_2001	-	-	122.5	6,741,258
NP_2003	-	-	121.0	9,904,128
ALERT_BN_ALL	-	-	335.2	364,004
MIX_2001	0.816	0.184	114.1	6,741,258
MIX_2003	0.814	0.186	112.3	9,904,128

Table 4: Comparison of Baseline LM perplexity vs. New LMs perplexities

Taking into account the experiments described in [10] we decided to investigate the results we would achieve by applying a modified interpolated form of Kneser-Ney discounting method [1]. Kneser-Ney smoothing uses a modified backoff distribution based on the number of

contexts each word occurs in, rather than the number of occurrences of the word. In [10] it is showed that a modified interpolated form of Kneser-Ney smoothing outperformed other smoothing techniques.

Models	PP	
	Absolute disc.	Kneser-Ney disc.
ALERT_BN_ALL	335.2	299.1
NP_2003	121.0	122.7

Table 5: Kneser-Ney discounting method vs. absolute discounting method

Table 5 shows that in case where a small quantity of data is available we can get better results by applying the interpolated Kneser-Ney discounting method (for ALERT BN model we obtained a perplexity reduction of about 11%). However, for large data training sets, as the newspapers one, we didn't get advantage in applying Kneser-Ney method. Mixing newspapers model obtained with absolute discount and ALERT BN model obtained with Kneser-Ney discount we get an interpolated model (referred here as MIX_2003_BEST) with a perplexity value of 111.4, our best result. The optimal interpolation weights obtained were 0.796 for the newspapers set component and 0.204 for the ALERT BN training set component.

Finally, we investigated the effects of pruning language models using an entropy-based pruning technique [3], i.e., pruning all n-grams that would increase the relative perplexity by less than a given threshold. Simultaneously, we pruned all the n-grams having probabilities lower than the corresponding backed-off estimates. This last pruning is useful to generate models that can be correctly converted into probabilistic finite-state grammars.

For this experiment we used the best mixed language model (MIX_2003_BEST). Table 6 shows model size and perplexity results obtained with various pruning thresholds. As shown, pruning is highly effective. For a threshold of 1e-09 we obtain a model that is about 30% the size of the original model without significant degradation of perplexity. On the following point we present the results in terms of WER.

Threshold	PP	Param	Size (.gz)
no pruning	111.4	9,904,128	303.1 Mb
1e-09	112.9	4,887,956	210.6 Mb
1e-08	119.7	1,511,364	102.1 Mb
1e-07	150.4	102,878	18.4 Mb

Table 6: Perplexity as a function of pruning threshold and language model size

B. Speech Recognition Results

Speech recognition experiments were conducted in a Pentium IV 2.8GHz computer with 2Gb RAM running Linux. The experiments were done under the same conditions, only varying the language model used. For these experiments we used the absolute discount version of NP_2003 model and the Kneser-Ney discount version of ALERT_BN_ALL. Table 7 summarizes the word error rate (WER) obtained for the different language models using the baseline system described in section 2. In this work we used the parametric conditions defined in line 6 of table 3 presented in [5]. However, we can not directly compare the baseline WER (26.5%) obtained in [5] since it was based on the development test set. For that reason, we evaluated the baseline language model over the evaluation test set, getting a WER of 28.2%, as expressed in line 1 of table 7.

Models	%WER	xRT
MIX_BASELINE	28.2	2.4
NP_2003	27.8	2.3
ALERT_BN_ALL	37.2	0.8
MIX_2003_BEST	26.9	2.4

Table 7: Speech recognition results as a function of language model

From line 4 of table 7 we can realize a small WER relative improvement of about 5% when comparing to the mixed baseline model. Finally, we evaluated the pruning effect on the WER. For this propose we used again the best mixed language model (MIX_2003_BEST). The results are summarized in table 8.

Pruning Threshold	%WER	xRT
no pruning	26.9	2.4
1e-09	26.7	2.0
1e-08	27.3	1.6
1e-07	29.4	1.1

Table 8: Pruning effect in speech recognition results

The results show that language models can be reduced up to 30% of its original size without significantly affecting the recognition accuracy. In this case we were able to get real-time decoding performance with only a small increase in word error rate, generating a language model 94% times smaller than the unpruned one.

V. CONCLUSIONS

From the results obtained in our experiments we concluded that in case where small data corpuses are available one get better results using a modified interpolated form of Kneser-Ney discounting method instead of absolute discounting.

In this work we increased the general domain data training set from 434.4 million words to 604.2 million words and we were able to obtain only a relative recognition error

decrease of about 5%. This suggests us that one should try to investigate other kind of approaches to improve the language model component, especially the ones related to domain adaptation of language models.

The applied entropy based pruning algorithm is highly effective. For a pruning threshold equal to 1e-09, we obtained a model that is 30% smaller than the original one without degradation in recognition performance (a slightly decrease in WER and a speed-up of about 17% in decoding time).

V. FUTURE WORK

As future work we will investigate the use of different clustering techniques applied to the Portuguese language, using class-dependent language modeling. Using these techniques we hope to get improvements not only in WER but mainly in language model size reduction, which will permit us to increase the system vocabulary size without compromise its practical level.

REFERENCES

[1] Chen, S. and Goodman, J., "An Empirical Study of Smoothing Techniques for Language Modeling", Computer Science Group - Harvard University, Cambridge, Massachusetts TR-10-98, 1998.

[2] Rosenfeld, R., "Adaptive Statistical Language Modeling: A Maximum Entropy Approach", PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1994.

[3] Stolcke, A., "Entropy-based Pruning of Backoff Language Models", in Proc. DARPA News Transcription and Understanding Workshop, Lansdowne, VA, 1998.

[4] Goodman, J. and Gao, J., "Language Model Size Reduction by Pruning and Clustering", in Proc. ICSLP 2000, Beijing, China, 2000.

[5] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language", presented at PROPOR 2003 - VI Encontro para o Processamento Computacional do Português Escrito e Falado, Faro, Portugal, 2003.

[6] Meinedo, H. and Neto, J., "Combination of Acoustic Models in Continuous Speech Recognition Hybrid Systems", in Proc. ICSLP 2000, Beijing, China, 2000.

[7] Clarkson, P. and Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", in Proc. EUROSPEECH 97, Rhodes, Greece, 1997.

[8] Caseiro, D., "Finite-State Methods in Automatic Speech Recognition", Lisbon, Portugal: Instituto Superior Técnico, Universidade Técnica de Lisboa, 2003.

[9] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. International Conference on Spoken Language Processing, Denver, USA, 2002.

[10] Goodman, J., "A Bit of Progress in Language Modeling - Extended Version", Machine Learning and Applied Statistics Group - Microsoft Research, Redmond, WA MSR-TR-2001-72, 2001.