IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Improving the Automatic Speech Recognition through the improvement of Laguage Models

Andrés Piñeiro-Martín, Carmen Garcia-Mateo, Laura Docio-Fernandez

atlanTTic Research Center, Multimedia Technologies Group, University of Vigo, Spain

apinheiro@gts.uvigo.es, carmen.garcia@uvigo.es, ldocio@gts.uvigo.es

Abstract

Language models are one of the pillars on which the performance of automatic speech recognition systems are based. Statistical language models that use word sequence probabilities (n-grams) are the most common, although deep neural networks are also now beginning to be applied here. This is possible due to the increases in computation power and improvements in algorithms. In this paper, the impact that language models have on the results of recognition is addressed in the following situations: 1) when they are adjusted to the work environment of the final application, and 2) when their complexity grows due to increases in the order of the n-gram models or by the application of deep neural networks. Specifically, an automatic speech recognition system with different language models is applied to audio recordings, these corresponding to three experimental frameworks: formal orality, talk on newscasts, and TED talks in Galician. Experimental results showed that improving the quality of language models yields improvements in recognition performance.

Index Terms: automatic speech recognition, language model, deep neural networks.

1. Introduction

Automatic Speech Recognition (ASR) is essential in applications where interaction with the user is through spoken communication, so that this can remain natural and effective. Such systems base their performance on acoustic models (AM) and language models (LM) that allow for the representation of the statistical properties of speech and language.

These days, with increases in computing power, the use of deep neural networks has spread to many fields. The current ASR systems are predominantly based on acoustic Hybrid Deep Neural Network–Hidden Markov Models (DNN-HMMs) [1] and the n-gram language model [2] [3]. However, in recent years, Neural Network Language Models (NNLM) have begun to be applied [4] [5] [6]. In these, words are embedded in a continuous space, in an attempt to map the semantic and grammatical information present in the training data, and in this way to achieve better generalization than n-gram models. The arrival of GPUs, multi-core GPUs and increases in computing power have made it possible to apply deep neural networks (DNNs) with multiple hidden layers, which are able to capture high-level, discriminating information about input features. The depth of the created network (the number of hidden layers), together with the ability to model a large number of context-dependent states, results in a reduction in Word Error Rate (WER). The type of neural networks most often used in language modelling are recurrent neural networks (RNNLMs). The recurrent connections present in these networks allow the modelling of long-range dependencies that improve the results obtained by n-gram models. In more recent work, recurrent network topologies such as LSTM (Long Short-Term Memory) [7] have also been applied. We should note that models of this type are far more complex than the n-gram model, and hence it takes more time to create them and they require more training data.

Furthermore, these models cannot be used straightforwardly in decoding, due to the large amount of computational resources needed, so the usual approach to their application is to perform a rescoring stage on a previously obtained word lattice using a n-gram language model. There are several algorithms that can implement this rescoring, such as those cited in [8], [9] and [10].

The current paper extends the work described in our previous study [11]. The impact on ASR performance in that study was analyzed when the quality of the text data corpora used for training the language models was increased and improved. The present paper looks specifically at the effect of enlarging the complexity of the models, increasing the order of the statistical models, and applying deep neural networks.

In addition, it should be noted that when working with minority languages such as Galician, obtaining the necessary data to train language models can itself be difficult. Therefore, this study also analyses how to use a modern system when working with languages for which the available data is limited.

The paper is organized as it follows: in Section 2 the ASR system is described. Section 3 then presents the experimental framework. Section 4 reports the experimental results. Section 5 provides a discussion of these, and finally Section 6 offers some final conclusions and suggestions for further lines of research.

2. Description of the ASR system

The ASR system was built using the Kaldi toolkit [12]. The acoustic models use a hybrid DNN-HMM modeling strategy with a neural network based on Dan Povey's implementation in Kaldi [13]. This implementation uses a multi-spliced TDNN (Time Delay Neural Network) feed-forward architecture to model long-term temporal dependencies and short-term voice characteristics. The inputs to the network are 40 Mel-frequency cepstral coefficients extracted in the Feature Extraction block with a sampling frequency of 16 KHz. In each frame, we aggregate a 100-dimensional iVector to a 40-dimensional MFCC input.

The topology of this network consists of an input layer followed by 5 hidden layers with 1024 neurons with RELU activation function. Asymmetric input contexts were used, with more context in the background, which reduces the latency of the neural network in on-line decoding, and also because it seems to be more efficient from a WER perspective. Asymmetric contexts of 13 frames were used in the past, and 9 frames in the future. Figure 1 shows the topology used and Table 1 the layerwise context specification corresponding to this TDNN.

10.21437/IberSPEECH.2018-8

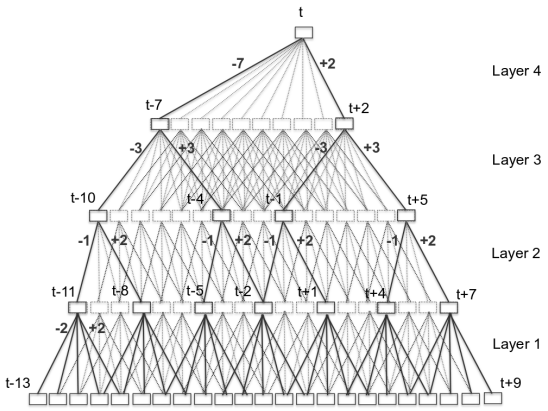

Figure 1: *TDNN used in acoustics models [13]*

Table 1: *Context specification of TDNN in Figure 1*

Layer	Input context
1	[-2,+2]
2	[-1,2]
3	[-3,3]
4	[-7,2]
5	{0}

The network has been trained with material corresponding to TC-STAR [14], with 79 hours of speaking in Spanish, and to Transcrigal [15], with 30 hours of speaking in Galician.

In terms of the language models, when working with n-gram LMs the model was trained using the SRI Language Modeling Toolkit. N-gram models of order 3 and 4 were used, that is, trigrams and tetragrams. A modified Kneser-Ney discounting of Chen and Goodman has also been applied, together with a weight interpolation with lower orders [16].

For training the RNNLMs, the Kaldi RNNLM [12] software was used. The neural network language model is based on a RNN with 5 hidden layers and 800 neurons, where TDNN layers with activation function RELU, and LSTM layers are combined. The training is performed using Stochastic Gradient Descent (SGD), and in several epochs (in our case, 20 epochs). All RNNLM models have been trained with the same material as in the case of n-gram statistical models. Figure 2 shows the topology of the network used.

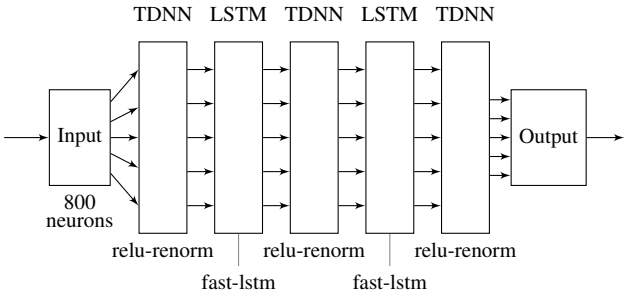

Figure 2: *RNNLM topology used in the ASR system*

3. Experimental framework

This section briefly describes both the text corpora used to train the LMs and the testing data used in the experiments.

3.1. Text corpora

Four text corpora were used (more detailed information in [11]):

- **DUVI:** texts from the DUVI (Diario da Universidade de Vigo). This is a small corpus, but it has very clean and representative text, including a large number of current words and expressions.
- **epub library:** a set of 5,000 Spanish novels translated into Galician using an automatic translator [17]. It should be noted that the results obtained using the LM when trained with this text may contain systematic errors due to these Spanish-Galician translations. One of the objectives in creating this corpus arises from the problem of finding large text corpora in minority languages. The results obtained illustrate the impact of using an LM with text translated from a language for which large quantities of text can be found.
- **Ghoxe, Eroski and Vieiros newspapers:** text from the Galicia Hoxe digital newspaper, published in Santiago de Compostela; text from the Eroski news page, which contains news text of a varied nature; and Vieiros, another digital newspaper published entirely in Galician. It is an extensive corpus, with clean and representative material, in that it contains news on a variety of themes.
- **CORGA:** texts from the Corpus de Referencia del Gallego Actual (CORGA), composed of different representative texts from books, newspapers, magazines, plays, audiovisual material and blogs. It is a medium-sized corpus with very carefully prepared and representative kinds of texts.

From the above text corpora, four LMs have been trained through a mixture of single models. The mixtures carried out were the following:

- **CLM1:** language model trained with the texts from DUVI, Ghoxe, Eroski and Vieiros newspapers, and CORGA.
- **CLM2:** directly combining the already-trained language models based on DUVI, Ghoxe, Eroski and Vieiros newspapers, and CORGA in a new language model. By doing this, we seek to analyze the differences in results between training new models with the whole text and mixing the already-trained models.
- **CLM3:** combining the single language models based on the epub library and CORGA.
- **CLM4:** combining the single language models based on DUVI, the epub library, Ghoxe, Eroski and Vieiros newspapers, and CORGA.

Table 2 shows the main characteristics of these combined models.

3.2. Testing data

Tests were made on three audio corpora with different characteristics.

1. **First Corpus: Formal Orality**. Corpus with audio recordings of formal orality corresponding to literary

Table 2: *Characteristics of the combined Language Models*

	Nº INV words	Training text size (M of words)	OOV words
CLM1	720.000	118	2,5 %
CLM2	730.000	118	2,5 %
CLM3	630.000	352	2,8 %
CLM4	900.000	435	2,3 %

readings and orally produced and read speeches. It consists of 30 files with an average duration of 3:50 minutes per recording and a total duration of approximately 115 minutes (about 2 hours).

2. **Second Corpus: Speech in newscasts**. Speech in newscasts. A corpus with audio recordings of television newscasts from TVG (Televisión de Galicia). They present a mixture of spontaneous and planned speech or read speech, but with more contemporary themes and vocabulary than in the first corpus. It consists of 10 files with an average duration of 34 minutes per recording and a total duration of 340 minutes (5 hours and 40 minutes).
3. **Third Corpus: Speech in TED Talks**. A corpus with audio recordings from TED Talks in Galician [18]. They present planned speech but are not read, being of a spontaneous nature. It consists of 10 files with an average duration of 16 minutes per recording and a total duration of 163 minutes (2 hours and 43 minutes).

4. Experimental results

Three experiments were carried out to assess the impact of the different language models on ASR performance:

- **Experiment 1**: recognition using a single-pass decoding strategy using 3-gram LMs of Table 2.
- **Experiment 2**: rescoring with 4-gram language models.
- **Experiment 3**: rescoring with RNNLMs.

4.1. Experiment 1

The mixture of models described in Section 3.1 has been tested in each of the corpora described in Section 3.2. The results can be seen in Table 3. The **SLM** column shows the best result obtained by a single LM, that is, without mixing LMs.

Table 3: *Combined Language Models Results*

		SLM	CLM1	CLM2	CLM3	CLM4
First Corpus	WER%	**21.02**	17.61	**17.51**	17.55	18.14
	CI-95%	± 1.84	± 1.76	± 1.71	± 1.78	± 1.77
Second Corpus	WER%	**21.39**	21.56	**21.52**	23.46	22.86
	CI-95%	± 2.63	± 2.58	± 2.55	± 2.78	± 2.70
Third Corpus	WER%	**19.07**	18.77	**18.68**	19.48	19.18
	CI-95%	± 2.24	± 2.44	± 2.57	± 2.70	± 2.81

Table 3 shows that for the first and second corpus it is not possible to reduce the average WER with any of the LM combinations, but it is possible to reduce the confidence interval of the results. Only for the third corpus was the average WER obtained slightly reduced, improving the results of single LMs.

In view of these findings, it can be concluded that model combinations do not significantly reduce the averaged WER, but do lead to an improvement in the confidence interval. On average these combined models present more robust results than any of the single models.

It is interesting to compare the results obtained by CML1 and CML2. We recall that CML1 is trained by combining all the texts, whereas in CML2 the previously-trained models are mixed. Table 3 shows that the average WER values are lower for CML2, and therefore it is better to combine previously-trained single language models.

4.2. Experiment 2

In this Experiment a tetragram rescoring on the lattice obtained in the previous experiment is performed. Table 4 shows the average WER together with the 95% confidence interval of the rescoring results.

Table 4: *Tetragram Language Models Rescoring Results*

		SLM	CLM1	CLM2	CLM3	CLM4
First Corpus	WER%	17.60	17.51	17.52	**16.72**	17.01
	CI-95%	± 1.84	± 1.75	± 1.72	± 1.76	± 1.75
Second Corpus	WER%	22.80	21.46	**21.40**	22.64	21.79
	CI-95%	± 2.65	± 2.63	± 2.61	± 2.68	± 2.74
Third Corpus	WER%	19.45	18.72	18.52	18.45	**17.97**
	CI-95%	± 2.62	± 2.46	± 2.56	± 2.68	± 2.60
Average WER in analysis corpora		19.95	19.23	19.14	19.27	**18.92**

The average WER and the confidence interval for the three corpora analyzed is reduced. In the first corpus the average WER is reduced by approximately 1% (expressed in absolute terms), obtaining a value of 16.72%. The reduction is similar in the second corpus, going from 22.80% to 21.40% of WER. In the third corpus it goes from 19.45% to 17.97%.

It is also interesting to compare the average WER for the three corpora shown in Table 4. The lowest average value is obtained by CLM4, the model that combines the greatest number of single LMs, and the one with the least out of vocabulary (OOV) words. It also shows how all the combinations of models obtain lower WER results than single models, that is, when applying the rescoring of tetragrams, the CLMs are clearly superior, being more robust in confidence interval and with a lower average WER.

4.3. Experiment 3

In this last experiment a rescoring with the RNNLM is applied to the lattices obtained by the decoding of experiment 2, that is, a rescoring RNNLM is applied to the lattice resulting from the rescoring of tetragrams. For this, RNNLMs have been trained using the same text as in the previous experiment. Table 5 shows the results.

In the first corpus all language models reduce the average WER obtained. An absolute reduction of up to 1.5% is achieved, reaching the value of 16.05% of average WER. However, in the second and third corpus, applying the rescoring with RNNLMs does not reduce the WER obtained for all LMs.

The average WER in the three analyzed corpora (final row in Table 5) is slightly reduced compared to the values obtained in experiment 3, achieving a value of 18.83% when using the

Table 5: *RNNLM Rescoring Results*

		SLM	CLM1	CLM2	CLM3	CLM4
First Corpus	WER%	**16.05**	16.82	16.63	16.52	16.57
	CI-95%	± 1.69	± 1.74	± 1.76	± 1.77	± 1.80
Second Corpus	WER%	22.98	21.61	**21.37**	23.44	21.39
	CI-95%	± 2.62	± 2.67	± 2.58	± 2.83	± 2.73
Third Corpus	WER%	19.27	19.07	18.73	18.82	**18.52**
	CI-95%	± 2.94	± 3.01	± 2.95	± 3.27	± 2.95
Average WER in analysis corpora		19.43	19.16	18.91	19.59	**18.83**

CLM4. We can conclude that the best strategy to reduce WER has been to combine the language models that have provided the best results in the first experiment. The combined models increase vocabulary size, provide more training texts and therefore reduce the OOV words, while also providing a greater robustness against variation in speech.

5. Discussion

This section offers a discussion of the WER results and the use of data in a modern system when working with a minority language.

5.1. WER results

The reduction of the average WER achieved in the first corpus is greater than that achieved in the second and third. Such a difference in behavior between the first corpus and the other two may be due to the different character of the linguistic samples [11]. The first corpus is composed mainly of read texts (written language), while the second corpus presents a high number of speakers, a heterogeneous mixture of speech types (read language, statements by different speakers, situations including noise, music, a mixture of Galician and Spanish, among others).

To see how such a heterogeneous mixture of speech affects the results, all interviews were removed from the recordings, leaving only the speech of presenters and reporters. The results show an absolute reduction of more than 7% compared to the best case for this corpus, obtaining an average WER of 13.88%. Therefore, the speech type of this corpus clearly does affect the results here.

A detailed analysis of the recognition errors can lead to a further reduction of the average WER. Yet it must be taken into account that some of the errors, at least in the oral corpus (not read), must be assumed to be inevitable. These errors reflect the doubts and errors in speakers' pronunciation, deviations in forms, etc. They might appear as errors in the transcription on which the calculation of the WER is based, but in which the recognition is in fact successful.

Finally, in order to check how far we can get with the current training data, a new RNNLM model was trained, introducing the transcripts of the analyzed corpora into the development text. With this, we sought to model the network so that it was specifically prepared to recognize the corpus on which it was going to be tested. Of course, applying this technique is only possible when the correct transcripts are available. The results show an absolute WER reduction of 0.2%, that is, a small and not significant improvement. This result leads us to conclude that it is difficult to continue reducing the WER with this training data and with these algorithms.

5.2. Use of limited data in a modern system

To train the models that the ASR systems uses, large corpora of audio (for the acoustic model) and text (for the language model) are necessary. In both cases, the type of data that is collected must be representative of the speech that one wants to recognize.

Obtaining a large corpus of audio recordings, with their corresponding transcriptions, and which are representative of the speech, is not easy when working with minority or less well-resourced languages. Large databases with this information are simply not available. Although there may be TV or radio stations that broadcast in the language, it is difficult to obtain such audio material with accurate transcriptions, and they often lack variety in terms of speech types.

This study has shown that one solution to deal with the shortage of resources in acoustic modeling is to use data from languages with similar phonetics. In our case, looking at Galician, the acoustic models have been trained using data from Spanish, multiplying by 4 the amount of information. Spanish, being a widely spoken language, has the necessary resources to obtain correctly transcribed audio corpus. Therefore, in our acoustic modeling, approximately 70% of data has been used in Spanish (over 79 hours of Spanish speaking). The other 30% corresponds to more than 30 hours of Galician.

For language models, a text corpus was created using data obtained from: 1) different magazines and newspapers published in the language; 2) downloading the information present in the Galician version of Wikipedia; 3) information obtained from small text corpora. In order to increase the amount of data available, another solution was to obtain a large corpus of data in another language, and translate it into Galician using a free automatic translation tool.

6. Conclusions

The results obtained for Galician ASR are promising. Improving the training text of the language models and applying RNNLM in decoding resulted in reducing the average WER obtained. However, it has also been shown that increasing the complexity of the system leads to more training data. The strategies applied to work with minority and less well- resourced languages have also contributed to the positive results in recognition.

As a future line of research, we plan to improve the acoustic and language models of the ASR, as well as to use more efficient algorithms in the decoding stage.

7. Acknowledgements

This work has received financial support from the Spanish Ministerio de Economía y Competitividad through project 'TraceThem' (TEC2015-65345-P), from the Xunta de Galicia (Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019, Galician Research Network TecAnDaLi ED431D 2016/011) and the European Union (European Regional Development Fund – ERDF). Our gratitude to the Ramon Piñeiro Institute of the Xunta de Galicia for allowing the use of the CORGA material and for its collaboration in the labeling of the second and third corpora.

8. References

[1] G. Hinton, L. Deng, D. Yu, and Y. Wang. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal

Processing Magazine, vol. 9, no. 3, pp. 82-97.

[2] J. T. Goodman. 2001. A bit of progress in language modeling. Computer Speech and Language, vol. 15, no. 4, pages 403–434.

[3] D. Jurafsky and J.H. Martin. 2008. Speech and Language Processing: An Introduction to Language Processing, Computational Linguistics, and Speech Recognition.

[4] T. Mikilov, S. Kombrink, A. Deoras, L. Bruget, and J. Cernicky. 2011. RNNLM-recurrent neural network language modeling toolkit, in Proc. of the 2011 ASRU Workshop, pages 196-201.

[5] E. Arisoy, T.N. Sainath, B. Kingsbury, and B. Ramabhadran. 2012. Deep Neural Network Language Models. In NAACL-HLT Workshop on the Future of Language Modeling for HLT, pages 20-28, Stroudsburg, PA, USA. Association for Computational LInguistics.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Juavi. 2003. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155.

[7] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. 2016. Exploring the limits of language modeling. In arXiv preprint arXiv:1602.02410.

[8] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur. 2018. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition, In ICASSP.

[9] M. Sundermeyer, Z. Tuske, R. Schluter, and H. Ney. 2014. Lattice decoding and rescoring with long-span neural network language models. En Fifteenth Annual Conference of the International Speech Communication Association.

[10] X. Chen, X. Liu, A. Ragni, Y. Wang and M. Gales. 2017. Future word contexts in neural network language models. ArXiv preprint arXiv:170805592.

[11] A. Piñeiro, C. García and L. Docío. 2018. Estudio sobre el impacto del corpus de entrenamiento del modelo de lenguaje en las prestaciones de un reconocedor de habla. Sociedad Española para el Procesamiento del Lenguaje Natural.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Quian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý. 2011. The Kaldi Speech Recognition Toolkit. In ASRU.

[13] V. Peddinti, D. Povey and S. Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of INTERSPEECH 2015.

[14] L. Docío, A. Cardenal and C. García. 2006. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. In Proc. Of TC-STAR Workshop on Speech-to-Speech Translation, ELRA, París, France.

[15] C. García, J. Tirado, L. Docío and A. Cardenal. 2004. Transcrigal: A bilingual system for automatic indexing of broadcast news. In IV International Conference on Language Resources and Evaluation.

[16] A. Stolcke. 2002. SRILM – An extensible language modeling toolkit. Proceedings of the International Conference on Statistical Language Processing, Denver, Colorado.

[17] I. Alegría, I. Arantzabal, M. Forcada, X. Gómez, L. Padró, J.R. Pichel, and J. Waliño. 2006. OpenTrad: Traducción automática de código abierto para las lenguas del estado Español. Procesamiento del Lenguaje Natural.

[18] TEDxGalicia. x=independent organized TED event. http://www.tedxgalicia.com/