Domain adaptation of a Broadcast News transcription system for the Portuguese Parliament

Luís Neves[1] , Ciro Martins[1,2], Hugo Meinedo[1], João Neto[1]

[1] L2F – Spoken Language Systems Lab – INESC-ID/IST
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[2] Department Electronics, Telecomunications & Informatics/IEETA
Aveiro University, Portugal
{Luis.Neves, Ciro.Martins, Hugo.Meinedo, Joao.Neto}@l2f.inesc-id.pt

Abstract. The main goal of this work is the adaptation of a broadcast news transcription system to a new domain, namely, the Portuguese Parliament plenary meetings. This paper describes the different domain adaptation steps that lowered our baseline absolute word error rate from 20.1% to 16.1%. These steps include the vocabulary selection, in order to include specific domain terms, language model adaptation, by interpolation of several different models, and acoustic model adaptation, using an unsupervised confidence based approach.

Keywords: vocabulary selection, model adaptation, domain adaptation, Portuguese Parliament, transcription systems.

1 Introduction

In the last decade Broadcast News (BN) transcription systems have been subject to a large effort of investigation and development by several international laboratories [1] [2] [3]. In our group we have been working specially on subtitling systems. This development allowed the construction of robust transcription systems, with high vocabulary coverage, low transcription word error rates and, in some cases, real time performance and online operation.

For the Portuguese language particular case, there have been great improvements, allowing the use of transcription systems in practical applications with diverse and complex context. An example of this is the large vocabulary transcription system currently being used to generate subtitles in the RTP1 evening news, working below real time on online mode.

There are several applications that could benefit from a large vocabulary automatic transcription system. The acknowledgment of this fact led us to adapt our BN transcription system to other domains of application, and in this particular work, for parliament meetings.

Automatic speech recognition of European Parliament Plenary Sessions has been one of the tasks of the TC-STAR project and one of the components involved in the translation process, with several participants like LIMSI [4], IBM [5], and NOKIA [6]

submitting their speech transcription systems for evaluation. This project was only focused in three different languages: European English, European Spanish, and Mandarin Chinese.

In the Portuguese Parliament plenary meetings there is the specific need to produce a journal that can be viewed by the general public, according to Parliament's Rules of Procedure. Our transcription system can be used to produce the journal entries, or to generate an initial transcription to be manually corrected, reducing the time demand of this process. The parliament plenary meetings are also broadcasted in television and online web video stream. Our transcription system can be used to produce subtitles allowing hearing impaired persons to follow these programs.

Section 2 summarizes the first task of the project, corpus collection, by retrieving previous plenary meetings available on the web, and recording video streams from the parliament's television channel. Section 3 describes our baseline Broadcast News transcription system and the corresponding results achieved without adaptation to the Parliament meetings task. Section 4 is dedicated to the adaptation of the transcription system's modules to the new domain. Finally we present some conclusions.

2 Corpora Collection

In this section we describe the corpora that were collected and processed to perform the work of domain adaptation.

2.1 Textual Corpora

In order to accomplish the speech transcription task in a given domain, it is necessary to obtain information about terms that are frequently used, and the way they appear in a sentence. This kind of information can be found in text material related to the target domain, and it is used to build the transcription system's vocabulary and language model. The system's performance is highly dependent of the quantity and quality of the text material. It was desirable to find manually transcribed plenary meetings, because they were most representative of the speech that would be recognized by our system. This text material was found in the Portuguese Parliament online site http://www.parlamento.pt, under The Journal of the Assembly - 1st Series section. Each document had the transcription of one parliament session; there were available 287 documents from the X Legislature and 281 from the IX Legislature, as shown in table 1. All of them were available as pdf files.

It was necessary to make the conversion from the pdf files to plain text. The format conversion was followed by a normalization process which eliminated punctuation, converted all text to lowercase, expanded abbreviations, spelled numbers and deleted speaker tags. This normalization was made with the same system used to normalize the text corpora from the broadcast news.

The X Legislature 3rd session was reserved as the text corpus development set, as shown in table 1. This set was required for the linear interpolation between language models, described in the domain adaptation section.

Table 1. Collected documents organization.

Legislature	Series	Time interval of the meetings	Number of documents	Set
X Legislature	1st series	2005-03-11 to 2006-09-08	149	Training
	2nd series	2006-09-16 to 2007-09-07	110	Training
	3rd series	2007-09-20 to 2007-12-14	26	Development
IX Legislature	1st series	2002-04-06 to 2003-09-04	146	Training
	2nd series	2003-09-18 to 2004-09-03	108	Training
	3rd series	2004-09-16 to 2005-01-27	24	Training

This way we had two different text corpora sets. The training set with 907,281 sentences and around 17M words, and the development set with 13,429 sentences and 205,795 words.

2.2 Audio Corpora

We have collected two video streams from the parliament channel's website, in 9 January and 10 January 2008.

For both video programs the audio stream was extracted to mp3 format, using open source tools. It was necessary to perform the conversion of the compressed audio to raw format at 16 KHz sampling rate, 16 bits per sample, which is currently one of the audio formats supported by our transcription system.

There were saturation levels in the 9 January program's audio, because the microphone recording level of the plenary session participants' was extremely variable. Usually this audio signal saturation increases the transcription system's error rate.

The total duration of the audio signal collected was 3 hours and 36 minutes. In order to evaluate the transcription system's performance, we selected a 21 minutes and 40 seconds audio segment, which was transcribed manually and used as the audio corpora evaluation set. This set has five male speakers and one female speaker, all of them with Lisbon accent, totalling 19 minutes and 12 seconds of net speech. The manual transcription of the evaluation set has 248 sentences with 2,850 words.

3 Baseline Transcription System

Our baseline large vocabulary transcription system was trained for Broadcast News in European Portuguese, entitled AUDIMUS [7]. It uses hybrid acoustic models that try to combine the temporal modeling capabilities of hidden Markov models with the pattern classification capabilities of MLPs (Multi-Layer Perceptrons).

The models have a topology where context independent phone posterior probabilities are estimated by three MLPs given the acoustic parameters at each frame. The MLPs were trained with different feature extraction methods: PLP (Perceptual Linear Prediction), Log-RASTA (log-RelAtive SpecTrAl) and MSG (Modulation SpectroGram). The two first referred above incorporate local acoustic

context via an input window of 13 frames, with the energy algorithm are extracted 12 coefficients and their first derivative, totaling a 26 elements vector. The last method uses an input window of 15 frames being extracted 14 coefficients. These are submitted to two filters (high-pass and band-pass), producing a 28 elements vector. The resulting network has two non-linear hidden layers with over 2000 units and 39 softmax output units (38 phones plus silence). The phone probabilities generated at the output of the MLPs classifiers are combined using an appropriate algorithm [8] to be used in the decoding process.

The decoder used in this system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [9]. In this approach, the decoder search space is a large WFST that maps observation distribution to words. The transcription system's full architecture is described in figure 1.

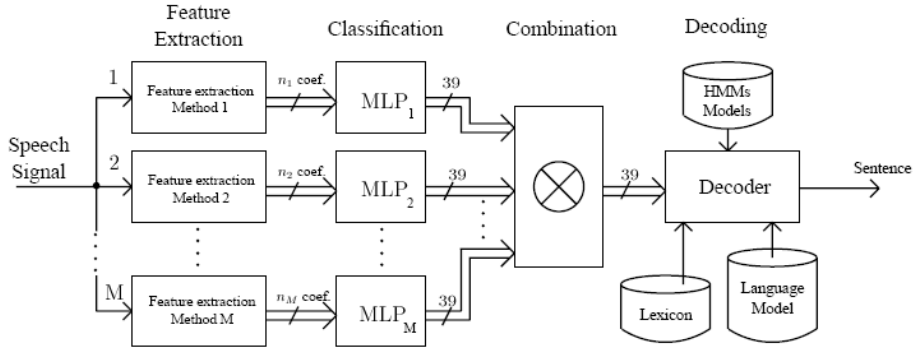

Fig. 1. Baseline transcription system's architecture [11]

The transcription system vocabulary and language model were built from two training corpora, newspapers texts collected from the WWW since 1991 until the end of 2003 with 604M words, and broadcast news transcripts with 531K words. The vocabulary was created selecting the 100,000 (100K) more frequent words from both corpora. The baseline language model (LM) [10] combines a backoff 4-grams LM trained on the newspapers corpus, and a backoff 3-grams LM estimated on the transcripts corpus. The two models were combined by means of linear interpolation, generating a mixed model.

The acoustic model used by the baseline system was trained with 46 hours of manual transcribed broadcast news programs, recorded during October 2000, and afterwards adapted using 332 hours of automatically transcribed material [11].

For the BN evaluation set corpus [11], the out-of-vocabulary (OOV) word rate is 1.4%, the average word error rate (WER) is 21.5% for all conditions and 10.5% for F0 conditions (read speech in studio).

Using our evaluation set, described in the audio corpora section, this baseline system achieved a WER of 20.1% for all acoustic conditions.

4 Domain Adaptation

The following subsections describe the adaptation stages to the Parliament's domain of the lexical, language and acoustic models.

4.1 Vocabulary and Lexical Model

In the adaptation to a new domain the vocabulary selection is extremely important. The specific frequent terms from the domain must be included, in order to the transcription system to recognize them.

To build the vocabulary and language model, we had available three different corpora. Two of them had been used training the broadcast news system, as described in section 3, and the third was our training textual corpora, described in section 2.1. The corpora collected for the broadcast news system, because of its size and generic characteristics, gave us the terms that were frequently used in the Portuguese language, while our textual corpora of manually transcribed plenary meetings gave us the specific terms of the domain.

We have decided to build our 100,000 words vocabulary based on word relative frequency. We have started by calculating the relative frequency value of each word in the three corpora, added these values for equal words, and selected the 100,000 words with the highest value. This extremely simple solution revealed itself effective, but there are other solutions for this problem, like morpho-syntactic analysis [12]. This selection method added 6,549 parliament transcriptions words that weren't in the initial broadcast news vocabulary. For our text development set the out-of-vocabulary (OOV) word rate was reduced from 2.0% to 1.1%.

The pronunciation lexicon was built by running the vocabulary through an automatic grapheme-to-phone conversion module for European Portuguese [13]. This module has the ability to produce multiple SAMPA pronunciations for each word, generating a pronunciation lexicon with 107,784 entries.

4.2 Language Model

It is important to introduce rules that can describe linguistic restrictions present in the language. This is accomplished through the use of a language model in the system. A language model represents a grammar which is a set of rules that regulate the way the words of the vocabulary can be arranged into groups and form sentences. Usually the grammar of a large vocabulary transcription system is a stochastic model based on probabilities for sequences of words. To create this kind of models it is required to use large amounts of training data as to obtain valid statistics that allow the construction of robust stochastic language models. This need for large amounts of training data lead us to build a mixed model, by linear interpolation of the broadcast news models and the model created with our manual transcriptions of plenary meetings, as shown in figure 2.

The process of creation, interpolation and quality analysis (perplexity calculation) of the language models was performed with the SRILM Toolkit [14].

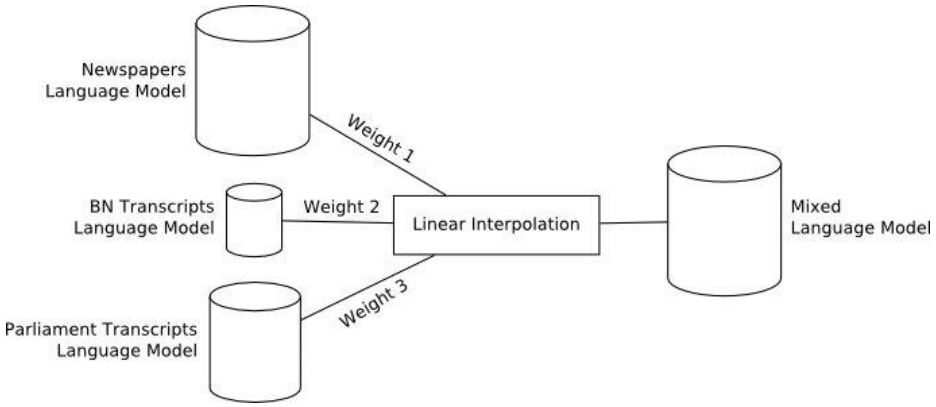

Fig. 2. Language Model linear Interpolation schematic

Our first step was to create a language model for each textual corpus available using our previous selected 100K vocabulary. To do this we selected the order and discount method that minimizes the perplexity of the model. This way we created a backoff 4-grams LM using absolute discounting trained with the newspapers texts, a backoff 3-grams LM using unmodified Kneser-Ney discounting trained with the broadcast news transcriptions, and a backoff 3-grams LM using unmodified Kneser-Ney discounting trained with the parliament transcriptions. The perplexity values, obtained for each one of these models for our development set, can be viewed in table 2.

Table 2. Language models parameters and perplexity.

Language Model	Order	Discounting	Perplexity
Newspapers	4	Absolute	140.2
BN Transcripts	3	Kneser-Ney	436.6
Parliament Transcripts	3	Kneser-Ney	71.8

To perform the linear interpolation between the three models, it is necessary to compute the interpolation weights with regard to the development set. The resulting model will have the minimum perplexity possible for the development set using a mixture of those three models [15]. The interpolation weights were set to 0.190 0.002 0.808 for the newspapers, BN transcripts and Parliament Transcripts LM's respectively. The result was a single backoff 4-grams language model.

After interpolation the perplexity value decreased to 50.9 with an OOV rate of 1.1% for the development set. The absolute WER for our evaluation set using the new language model decreased to 16.3%, resulting in a relative WER reduction of 18.9%.

4.3 Acoustic Model

Usually the adaptation of the acoustic model requires a large amount of manual transcribed audio to adapt the Multi-Layer Perceptron (MLP) network weights effectively. Unfortunately, for this work in particular, the only manually transcribed material available was the evaluation set. To solve this problem we have used a multiple decoding stage approach.

In a first stage a first transcription of the entire audio corpora can be obtained with the base system acoustic model and the adapted language model. Then the transcribed targets from the first decoding stage are pruned according to the degree of confidence of the transcription obtained.

The transcription system can provide a confidence measure that compared to a threshold allow rejecting transcriptions potentially erroneous. To determine the value of the threshold, we have created the ROC curve based on the evaluation set, as shown in figure 3, and determined an appropriate working point. It was more important to have a low false alarm (erroneous transcription that was accepted) than a high detection (correct transcription that was accepted) percentage to assure the quality of the transcribed targets. This way we have selected a confidence threshold value of 0.915 which produced 12% false alarm and 66% detection.

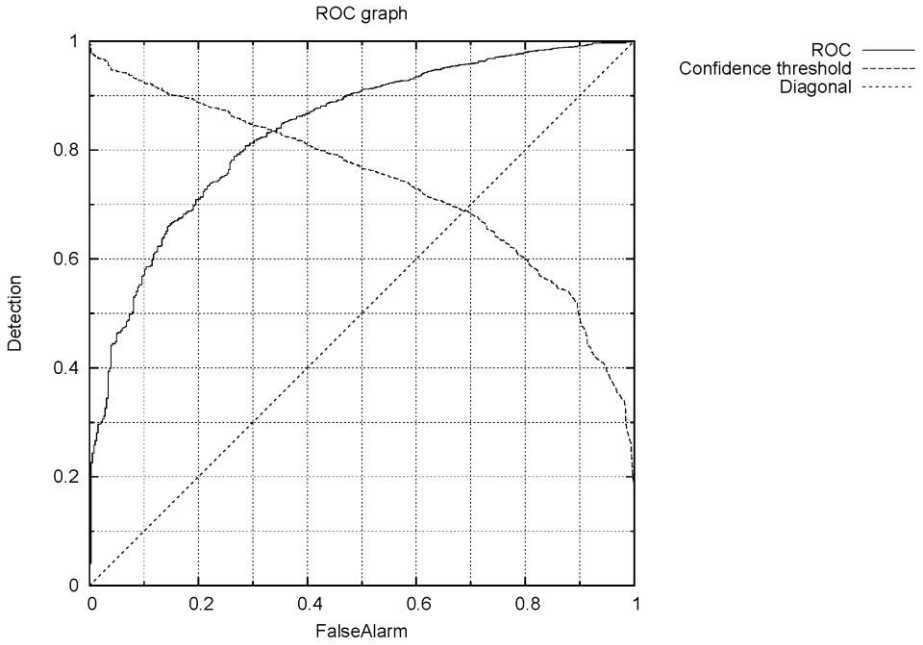

Fig. 3. ROC curve of the test set.

Using the pruned transcribed targets and the corresponding audio segments was possible to adapt the acoustic model to the new domain. We have conducted two different adaptations of the Multi-Layer Perceptron (MLP) network weights. In the

first one we used 10% of the selected targets as a cross-validation set and the rest of them as the training set. The cross-validation classification error defined the number of training iterations of the neural network. There were performed 6 training steps, producing a WER result of 16.7% in the evaluation set. In the second kind of adaptation we have used all the previous selected targets to adapt the weights. The results, available in table 3, showed that this was the best solution using all the selected targets to make just one neural network training iteration, to avoid over-adaptation.

Table 3. Transcription word error rate (WER) in the second stage decoding, without cross-validation set in the neural network adaptation.

Training Iteration	Without Cross-validation
1	16.1%
2	16.4%
3	16.6%
4	18.6%
5	17.3%

The final result from the second stage transcription of the evaluation set, with the adapted acoustic model, was 16.1% WER, resulting in a relative WER reduction of 19.9% to the baseline system.

5 Conclusions

This paper reported our work on adapting a broadcast news transcription system to a new domain of application, the Portuguese Parliament plenary meetings. This work involved several steps in order to adapt the vocabulary, language model and acoustic model used.

Our first impressions of this work, suggested that the greater difference that existed between the two domains lied in the vocabulary and consequently in language model used. This was later confirmed during our work, because the greater WER reduction (18.9%) was achieved with the vocabulary and language model adaptations.

The correct adaptation of the acoustic model is directly related with the amount of training audio corpora used to adapt the neural network weights. The small gain in the WER obtained with our model adaptation can be justified with the small amount of audio used for this task (around 4h) when compared to the amount used to train the baseline model (around 348h). Besides, the baseline model was already expected to perform well in the new domain since it was trained with a wide range of acoustic conditions and speakers.

Probably a slightly better result could be achieved with the creation of manual transcriptions for the training audio corpora, because in this case there was no transcription error in the targets used in the neural network adaptation, but usually this is the most time demanding task. Our adaptation process allows us to eliminate this

problem, thus, reducing drastically the time needed to deploy our transcription system in a new domain.

6 Acknowledgements

The authors would like to thank Alberto Abad, for many helpful discussions.
This work was funded by PRIME National Project TECNOVOZ number 03/165.

7 References

1. Gales, M., Kim, D., Woodland, P., Mrva, D., Sinha, R., Tranter, S.: Progress in the CU-HTK Broadcast News Transcription System. In: IEEE Transactions on Audio Speech and Language Processing (2006).
2. Sinha, R., Gales, M., Kim, D., Liu, X., Sim, K., Woodland, P.: The CU-HTK Mandarin Broadcast New Transcription System. In: Proceedings ICASSP (2006).
3. Nguyen, L., Abdou, S., Afify, M., Makhoul, J., Matsoukas, S., Schwartz, R., Xiang, B., Lamel, L., Gauvain, J., Adda, G., Schwenk, H., and Lefevre, F.: The 2004 BBN/LIMSI 10xRT English Broadcast News Transcription System. In: Proceedings DARPA RT04, Palisades NY, November (2004).
4. Lamel, L., Gauvain, J., Adda, G., Barras, C., Bilinski, E., Galibert, O., Pujol, A., Schwenk, H., and Zhu, X.: The LIMSI 2006 TC-STAR EPPS Transcription Systems. In: Proceedings of ICASSP, pages 997-1000, Honolulu, Hawaii, April (2007).
5. Ramabhadran, B., Siohan, O., Mangu, L., Zweig, G., Westphal, M., Schulz, H., Soneiro, A.: The IBM 2006 Speech Transcription System for European Parliamentary Speeches. In: ICSLP, September (2006).
6. Kiss, I., Leppanen, J., and Sivadas, S.: Nokia's system for TC-STAR EPPS English ASR evaluation task. In: Proceedings of TC-STAR Speech-to-Speech Translation Workshop, Barcelona, Spain, June (2006).
7. Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I.: AUDIMUS.media: a Broadcast News speech recognition system for the European Portuguese language. In: Proceedings of PROPOR 2003, Portugal (2003).
8. Meinedo, H., Neto, J.: Combination of acoustic models in continuous speech recognition. In: Proceedings ICSLP 2000, Beijing, China (2000).
9. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. In: ASR 2000 Workshop. (2000).
10. Martins, C., Teixeira, A. and Neto, J.: Language models in automatic speech recognition. In: Magazine of DET-UA. Aveiro, vol. 4, n.º 4 (2005).
11. Meinedo, H.: Audio pre-processing and speech recognition for Broadcast News. In: PhD thesis, IST (2008).
12. Martins, C., Teixeira, A., Neto, J.: Dynamic Broadcast News transcription system. In: ASRU 2007 (2007).

13. Caseiro, D., Trancoso, I., Oliveira, L., Viana, C.: Grapheme-to-phone using finite state transducers. In: Proc. 2002 IEEE Workshop on Speech Synthesis, Santa Monica, CA, USA (2002).
14. Stolcke, A.: Srlim - an extensible language modeling toolkit. In: Proc. ICSLP'2002, Denver, USA (2002).
15. Souto, N., Meinedo, H., Neto, J.: Building language models for continuous speech recognition systems. In: PorTAL 2002 (2002).