Building language models for continuous speech recognition systems

Nuno Souto, Hugo Meinedo, João P. Neto

L^{2}F - Spoken Language Systems Laboratory, INESC ID Lisboa / IST
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://l2f.inesc-id.pt/

Abstract. This paper describes the work developed in the creation of language models for a continuous speech recognition system for the Portuguese language. First we discuss the process we use to create and update a text corpus based on newspaper editions collected from the Web from which we were able to generate N-gram language models. We also present the procedure we use to improve those models for a Broadcast News (BN) recognition task by interpolating them with a BN transcriptions based language model. Finally the paper details a method used to generate morpheme-based language models.

1. Introduction

In a continuous speech recognition system it is necessary to introduce linguistic restrictions through the use of a language model in the recognition system. Every language model we have been using are N-gram statistical models. These models use the previous N-1 words as its sole information source for predicting the current word, requiring large amounts of training data to obtain valid N-gram statistics. To obtain these large amounts of text, we have been collecting Portuguese newspaper editions available on the web during the last years.

The kind of texts that should be used to training a language model depends on the tasks where the model is going to be used. For example, if we want to create a language model to be used in a medical reports recognition task, it should be built using large amounts of medical reports or any other medicine related texts. If we want to create a more generic language model we should use more generic texts like newspaper texts. These generic models can also be used to generate task-adapted models when large amounts of related texts are not available, by interpolating a generic language model with a smaller model generated from specific text.

Section 2 gives a brief explanation on how a language model is used in a speech recognition system. Section 3 presents the process we use to create and update a large text corpus based on newspaper editions collected from the Web, from which we generate N-gram language models. Section 4 describes the use of interpolation as a technique to improve a language model for a specific task. In section 5 a method for generating morpheme-based language models is introduced. In Section 6 the results of some experiments using the language models described in the previous sections are shown. Section 7 presents some conclusions and further work to be done.

2. The use of Language Models in a Speech Recognition System

In general, the acoustic component of speech recognition systems produces a sequence of phonemes or phonetic segments that can be a set of hypotheses, which correspond to text recognized by the system. During the recognition, the sequence of symbols generated by the acoustic component is compared with the set of words present in the lexicon as to produce the optimal sequence of words that will compose the system's final output. It is important to introduce rules during this stage that can describe linguistic restrictions present in the language and can allow the reduction of the number of possible valid phoneme sequences. This is accomplished through the use of a language model in the system. A language model comprises two main components: the vocabulary which is a set of words that can be recognized by the system and the grammar which is a set of rules that regulate the way the words of the vocabulary can be arranged into groups and form sentences. The grammar can be made of formal linguistic rules or can be a stochastic model. The linguistic models introduce strong restrictions in allowable sequences of words but can become computational demanding when incorporating in a speech recognition system. They also have the problem of not allowing the appearance of grammatically incorrect sentences that are often present in spontaneous speech. This makes the stochastic models based on probabilities for sequences of words more attractive for use in speech recognition systems due to their robustness and simplicity. 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.

3. N-Gram Language Model creation

Presently the language models we are using in our speech recognition system have been created using text corpus obtained from the online newspaper editions collected daily from the web. The main reason for this is the facility offered by the Web for collecting the large amounts of texts required for building N-gram language models. The texts collected are newspaper texts since these are very general and closely related with the task where we presently develop our speech recognition system, which is a Broadcast News (BN) recognition task. Nevertheless they are mainly composed of written language, which does not reveal the spontaneous nature frequently present in spoken language.

Our previous language models were based on a text corpus built solely from "Público" newspaper editions available on the web and had about 46 million words. During the last years we have started collecting other Portuguese newspapers from the web, which allowed us to start building considerably larger text corpus. Recently a new corpus named CETEMPúblico [1] containing extracts from the "Público" newspaper editions from 1991 to 1998, with a total of 180 million words, was made available. This corpus was added to our own corpus, but as they both partially overlapped, special care was taken as to avoid inserting twice the same newspaper editions.

To create a language model we need to have the text corpus in a simplified SGML format. In this format, there can be only one sentence per line and each sentence is delimited by the tags <s> e </s>. The collected web pages are written in HTML format and so they must be converted to SGML format using several tools. First all the HTML formatting tags are removed as well as all the text that does not belong to the newspaper articles. Whenever possible, a tag with the topic is added to the beginning of the article to keep them identified by subject. Repeated or very similar articles existing in the same newspaper edition or also existing in the edition from the previous day are removed. This first processing stage allows us to obtain the newspaper edition's articles in clean text and is executed automatically every day as soon as a newspaper is collected. Then it is necessary to convert the texts so that every number, date, money amount, time, ordinal number, web address and some abbreviations are written in full. All the text is converted to small caps, separated in one sentence per line, all the punctuation is removed or written in full and the delimiter tags are added. These stages are performed only when we want to create or update a text corpus, which we usually do every six months. In Table 1 we present some statistics of the latest updates to our newspaper based corpus.

Table 1. Number of words and sentences in the newspaper based corpus in the last updates.

	texts_2000	texts_2001a	texts_2001
total_words	335.7M	384.1M	434.4M
total_sentences	17.4M	20.7M	24.0M

With the text corpus built in this format it is possible to generate language models using either the CMU-Cambridge SLM Toolkit [2] or the SRI Language Modeling Toolkit. We have been using CMU-Cambridge Toolkit more often because of some memory problems we are having with SRI Toolkit.

Before we can start building language models we must create the vocabulary. Currently we are limiting our vocabulary size to 64k words. This vocabulary was first created using 56k different words selected from the text corpus according to their weighted class frequencies of occurrence. Different weights were used for each class of words. All new words, from a total of 12,812 different words, present in the transcripts of the training data of our BN database were added to the vocabulary giving a total of 57,564 words. The margin to the 64k is being kept to incorporate new words of the still unfinished training data transcripts.

From the vocabulary we were able to build the pronunciation lexicon. To obtain the pronunciations we used different lexica available in our lab. For the words not present in those lexica (mostly proper names, foreign names and some verbal forms) we used an automatic grapheme-phone system to generate corresponding pronunciations. Our present lexicon has a total of 65,895 different pronunciations. The lexicon has more pronunciations than the total number of words present in the vocabulary because many of the words have more than one pronunciation.

When a language model is created it is necessary to have some way of testing its quality. The most important method for doing this is to use the model in the application it was designed for and watch its impact in the overall performance. In the case of a language model designed for a speech recognition system, the best way of

testing its quality is to just evaluate the word error rate (WER) obtained when the model is used in the system. However this method is not very efficient, as it needs a lot of computer processing for reliably measuring the WER, being very time consuming. So alternative methods must be used instead. Perplexity is often used as a measure of the quality of a language model as it tests the capability of a model for predicting an unseen text which is a text not used in the model training. Perplexity of a model relative to a text with n words is defined as [3]:

$$\mathbf{PP=2^{LP}} \quad \text{where } LP = \left(\frac{1}{n}\right)\log P'(w_1 ... w_n) \qquad (\mathbf{1})$$

P' is the probability estimation of the sequence of n words given by the language model. Perplexity can be seen as the average size of the word set over which a word recognized by the system is chosen, and so the lower its value the better. It is a measure that depends on the model and on the text used for the test, and so to compare different models one should use the same texts. Perplexity does not take into account acoustic similarity between words, which means that lower perplexity values may not result in lower WER during recognition.

Even with large amounts of texts there are always sequences with N words that are not observed in the training texts but whose occurrence should be allowed. To enable this, some of the probabilistic mass of the observed events must be discounted and distributed as backoff values to the observed N-grams of order lower than N. The probability of occurrence of an unobserved N-gram is determined through a recursive method that uses the probabilities and backoff values of N-grams with order lower than N. There are various methods for the discounting redistribution of probabilistic mass, like linear discounting, absolute discounting, Good Touring discounting, etc.

In Table 2 we present the perplexity values obtained with several different language models computed on the transcriptions from the evaluation set of our BN database. These are backoff 4-grams models using absolute discounting, extracted from the newspaper text corpus during its latest updates: end of 2000, first semester of 2001 (2001a) and end of 2001. Some of the models were created with cutoff values of 2, 2, and 2 respectively for the 2-grams, 3-grams and 4-grams while others were created with cutoff values of 2, 3, and 4. A cutoff value means that only n-grams with a frequency of occurrence greater than the cutoff will be used in the model. It is clear that with small increments in the cutoff values we can create smaller models with just a little loss on the perplexity.

Table 2. Perplexity obtained with several newspaper based language models on the evaluation set transcriptions of the BN database.

Model	Evaluation	Dimension (.gz)
n4.2_2_2.2000	149.7	226 Mb
n4.2_2_2.2001a	145.9	255.6 Mb
n4.2_3_4.2001a	149.8	162.9 Mb
n4.2_2_2.2001	143.9	274 Mb
n4.2_3_4.2001	148.0	174 Mb

4. Language Model Interpolation

A language model generated from newspaper texts becomes too much adapted to the type of language used in those texts and when it is used in a continuous speech recognition system applied to a BN task it will not perform as good as one would expect because the sentences spoken in BN are not exactly like the sentences written in the newspaper. If we created a language model from BN manual transcriptions it would probably be more adequate for this kind of speech recognition task but we do not have enough BN transcriptions to generate a satisfactory language model. However we can adapt the language model to the BN task by interpolating a model created from the newspaper texts with a model created from BN transcriptions.

Linear interpolation is a way of combining *n* different information sources,

$$P_{\text{int}\,erpolated}(w \mid h) = \sum_{i=1}^{n} \gamma_i P_i(w \mid h) \qquad (\mathbf{2})$$

where $0 < \gamma_i \leq 1$ and $\sum_{i=1}^{n} \gamma_i = 1$.

The optimal interpolation weights are computed with regard to the BN transcriptions of the evaluation set, which means the resulting model will have the minimum perplexity possible for the evaluation set using a mixture of those two models [3]. These weights are not guaranteed to be optimal regarding the rest of the data where the model is going to be used but as long as the evaluation set is large enough and representative, the weights will be nearly optimal for the rest of the data.

We use the "interpolate" program of the CMU Cambridge toolkit to calculate the optimal weights using the Expectation Maximization algorithm (EM). To execute the interpolation for creating a new model we use the SRI-LM toolkit, which allows us to mix language models and specify the weights, we wish to use for each.

The interpolated models we have created are obtained mixing the newspaper based model from previous section with a backoff 3-gram model using absolute discounting based on the BN training set transcriptions.

In Table 3 we present the perplexities obtained from the interpolated models between the BN transcriptions and the n4.2_3_4.2001a and n4.2_3_4.2001 models together with the optimal weights applied to the interpolation. The perplexity values shown were computed over the transcriptions from the evaluation set and also over a small text set based on newspaper texts from the second semester of 2001, which could not be used to test the perplexity of the complete 2001 based models since those texts were used in their creation and would cause a very low perplexity value.

From the table it is clear that even using a very small model based in BN transcriptions we can obtain some improvement in the perplexity of the interpolated model.

Table 3. Perplexity obtained with interpolated language models and their components on the evaluation set transcriptions of the BN database and on a small newspaper text set.

Model	$\gamma_{Newspaper}$	$\gamma_{BNtranscriptions}$	Evaluation	Newspaper
BNtranscriptions_corpus.n3.0_0	-	-	551.6	1080.9
n4.2_3_4.2001a	-	-	149.8	138.3
interpolated_2_3_4.2001a	0.825	0.175	140.3	141
n4.2_3_4.2001	-	-	148.0	-
interpolated_2_3_4.2001	0.829	0.171	139.5	-

5. Morpheme-based Language Model

One of the main problems with speech recognition systems are the words that can be spoken during a speech recognition task but do not exist within the system's vocabulary. These are called out of vocabulary words (OOV's). The easiest solution would be to just extend the size of the vocabulary to push the OOV's rate beyond a reasonable limit but this is not very efficient, especially for highly inflectional languages like Portuguese, French and German. The problem in this kind of languages is the great vocabulary expansion caused by the large number of different words derived from basic words. Many words have different forms for singular and plural and may also change with the gender. This implies that we need to have vocabularies with much larger dimensions to obtain coverage similar to English. To get an idea on the differences between English and a highly inflectional languages like Portuguese, we compare the number of different forms of the verb "to sing" in Portuguese ("cantar") and English in the Simple Present, in Table 4.

Table 4. Comparison of the forms of the verb "to sing" in the Simple Present, between Portuguese and English.

Portuguese	Canto, Cantas, Canta, Cantamos, Cantais, Cantam
English	Sing, Sings

As we can see, while in English there are only two different forms (sing, sings), in Portuguese there are six forms. Moreover, in Portuguese many verbs have the simple conjugation forms (Ex:"moveu"/ moved) but also pronominal conjugation forms (Ex: "moveu-a" / moved her), which can have a great number of different forms and may depend on the gender.

We have been working and developing a possible solution, which we already started discussing in [4,5], consisting on a decomposition method based on a partial morphological analysis of the words. On our previous studies, we classified morphologically all the words present in our previous newspaper "Público" based database and realized that about 35% of the words were verbal inflections. Actually the verb is the most variable word class in the Portuguese language being the inflectional derivation the main mechanism for generating the different forms of a

verb. In the inflectional derivation, words are formed through the combination of a root with a prefix, a suffix or both. This took us to conclude that a morphological decomposition of the regular verbs on their roots and suffixes would allow us to achieve a significant reduction on the vocabulary dimension.

Using specific tools and a hand made list containing all the possible regular verbs suffixes (813 different suffixes) we decompose the vocabulary and also words in the entire text corpus used to generate the usual word-based language models. This results in a whole new decomposed text corpus from which we can create morpheme-based language models. During this process a list, containing all of the words decomposed and respective roots and suffixes, is generated. To avoid ambiguity during the reconstruction of the words, the decomposition should only be performed when the root and suffix do not exist both as words. For example the word “como" (“eat”) can not be decomposed because it would result in a sequence of existing words “com o” (“with the”).

The decomposition must also be performed at the phonetic level, which means the lexicon must also be decomposed using a specific designed tool. This tool searches the original lexicon for the words that are supposed to be decomposed and then, for each word, tries to identify the phonemes of the word that belong to the root and those that belong to the suffix, using a set of pronunciation rules. The resulting lexicon keeps all the different pronunciations found for each root and suffix.

For the decomposition method to work, we need to have a post-processing step that joins back the morphemes, using the list with the decomposed words, so that we only have complete words at the system’s output. To avoid incoherence during recognition some restrictions must be introduced in the language model. To avoid the occurrence of a word or an incorrect suffix after a root that is not also a word we force all the n-grams ending with it to have zero for its backoff value. We also force the unigrams probabilities to be zero for all the suffixes that are not also words. This avoids the occurrence of a suffix after a word or an incorrect radical.

The method just described to create morpheme-based language models has an interesting aspect. Any word capable of being decomposed in a root and a suffix both existing in the new decomposed vocabulary, can be recognized by the system even if the complete word didn’t exist in the original vocabulary, as long as it was observed in the training text corpus. This happens because the morpheme-based language model can capture those decomposed words in the text corpus through the 2-grams, 3-grams, etc., containing their roots and suffixes and the rebuilding information is stored in the list of the decomposed words that is used in post-processing stage. In this way, although we are reducing the vocabulary dimension used in the recognition system, the total vocabulary admitted by the system is being expanded. Just to get an idea, our last decomposition produced a list of 114,432 decomposed words which is a much larger number than the total number of words in our original vocabulary (57,765 words). This capability for expanding the vocabulary could be further exploited to allow the recognition of words composed of a root and a suffix existing in the new decomposed vocabulary even if those words were not observed in the training texts.

6. Speech Recognition experiments

To make the speech recognition experiments we used our baseline recognizer AUDIMUS [6], which was originally developed with a corpus of read newspaper text. It is a hybrid system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multiplayer perceptrons (MLPs) [7]. In this hybrid HMM/MLP system a Markov process is used to model the basic temporal nature of the speech signal. Three MLPs, each associated with a different feature extraction process, are used for the acoustic modeling of AUDIMUS. A MLP estimates 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. The combination algorithm merges together the probabilities associated with the same phone. It was this system that served as baseline for the development of the BN system. To develop this system we have been collecting and building a Portuguese BN database comprising two main corpus [8]. The first corpus is comprised of approximately 80 hours to be used as training and test sets for the speech recognition systems. The second with more than 150 hours is aimed for the development of automatic topic detection algorithms. The collection of both corpus is completed. The first corpus was automatically transcribed using our baseline speech recognition system AUDIMUS and was manually corrected and annotated. We are currently using only 22 and half hours for the training of the system. For the system's evaluation we are using part of the development test set with a net duration of approximately 3 hours.

In Table 5 we present the word error rate (WER) obtained when using some of the language models created within the speech recognition system used to recognize the evaluation data of the BN database. From the results we can see that as expected the interpolated models achieved the lowest WER values. This confirms the previsions we made based on the best perplexity values obtained with these models in the BN evaluation transcriptions.

Table 5. WER obtained with the speech recognition system in a BN recognition task using different language models. F0 refers to sentences of prepared speech, with low background noise and good quality speech signal. All F refers to all possible focus conditions (including F0) [8].

Model	F0	All F
n4.2_2_2.2001a	17.2%	32.9%
n4.2_3_4.2001a	16.9%	32.9%
BNtranscriptions_corpus.n3.0_0	34.3%	49.2%
interpolated_2_3_4.2001a	16.3%	32.8%
interpolated_2_3_4.2001	16.3%	32.6%

In Table 6 it is presented the WER obtained with a morpheme based language model used in the recognition of the evaluation data of the BN database. It is also

presented the same results obtained with a similar word-based language model. These results were obtained with a different trained and aligned acoustic model. We can see that a small degradation resulted when using the morpheme based language model.

Table 6. WER obtained with a morpheme based language model used in the recognition of the evaluation data of the BN database.

Model	F0	All F
n4.2_2_2.2001a	17.2%	32.9%
n4.dec.2_2_2.2001a	20.3%	36.8%

7. Conclusions

This paper reported our work on developing language models for a continuous speech recognition system for the Portuguese language. We started by showing the method we have been using to create and update a text corpus with large dimensions using collected newspaper editions from the Web. This text corpus is used to generate N-gram language models. It was shown that through interpolation it is possible to improve a newspaper texts based language model using a small BN manual transcriptions based model, decreasing the WER of a speech recognition system applied to a BN recognition task. This proved to be a very effective method for creating language models more adapted to specific tasks when there are not enough amounts of related texts available. In the future, the availability of more training data on our BN database will result on the upgrade of the transcriptions based language model improving even further the interpolated model.

We presented a method for vocabulary decomposition based on morphological analysis of words. This method allowed us to reduce the vocabulary dimension with a small degradation in the WER obtained. We believe that in the future a more sophisticated morpheme-based language model can be implemented.

8. Acknowledgments

This work was partially funded by IST-HLT European program project ALERT and by FCT project POSI/33846/PLP/2000. Two of the authors, Nuno Souto and Hugo Meinedo, were supported by scholarships in the scope of the FCT project POSI/33846/PLP/2000. INESC ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

References

1. Paulo Rocha and Diana Santos, “CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa”, in Proceedings PROPOR’2000, Brasil, 2000 (Portuguese text) [http://cgi.portugues.mct.pt/cetempublico/].
2. P. Clarkson and R. Rosenfeld, “Statistical Language Modeling Using the CMU-Cambridge Toolkit”, in Proceedings of EUROSPEECH 97, Rhodes, Greece, 1997.
3. Ronald Rosenfeld, “Adaptive Statistical Language Modeling: A Maximum Entropy Approach”, PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1994.
4. Ciro Martins, “Modelos de Linguagem no Reconhecimento de Fala Contínua”, Tese de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, 1998 (Portuguese text).
5. Ciro Martins, João P. Neto, Luís B. Almeida, “Using Partial Morphological Analysis in Language Modeling Estimation for Large Vocabulary Portuguese Speech Recognition”, in Proceedings of Eurospeech 1999, Budapest, Hungary, 1999.
6. H. Meinedo and J. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems", in Proceedings ICSLP 2000, Beijing, China, 2000.
7. H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994
8. H. Meinedo, N. Souto and J. Neto, "Broadcast News speech recognition for the Portuguese language", in Proceedings ASRU, Italy, 2001.