DYNAMIC LANGUAGE MODELING FOR A DAILY BROADCAST NEWS TRANSCRIPTION SYSTEM

Ciro Martins+, António Teixeira*, João Neto+*

*Department Electronics, Telecommunications & Informatics/IEETA – Aveiro University, Portugal
+L2F – Spoken Language Systems Lab – INESC-ID/IST, Lisbon, Portugal
Ciro.Martins@l2f.inesc-id.pt, ajst@det.ua.pt, Joao.Neto@inesc-id.pt

ABSTRACT

When transcribing Broadcast News data in highly inflected languages, the vocabulary growth leads to high out-of-vocabulary rates. To address this problem, we propose a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and LM to the topic of the current news segment during a multi-pass speech recognition process. Based on texts daily available on the Web, a story-based vocabulary is selected using a morpho-syntatic technique. Using an Information Retrieval engine, relevant documents are extracted from a large corpus to generate a story-based LM.
Experiments were carried out for a European Portuguese BN transcription system. Preliminary results yield a relative reduction of 65.2% in OOV and 6.6% in WER.

***Index Terms*—** Speech recognition, Natural language interfaces

1. INTRODUCTION

Although the vocabularies of ASR systems are designed to achieve high coverage for the expected domain, out-of-vocabulary (OOV) words cannot be avoided. Particularly, for tasks like daily transcription of Broadcast News (BN) data in highly inflected languages, the rapid vocabulary growth leads to high OOV word rates [1]. This is the case of the Portuguese language, mainly due to the inflectional structure of verbs class.

This way, lexical coverage of a vocabulary should be as high as possible to minimize the side effects of OOV on system recognition performance. As stated in [2], vocabulary optimization is mainly dependent on the task, amount of training data used, source and recency of that data. Actually, assuming an open task like the BN data transcription, the topic changing over time leads to unlimited vocabulary. While a large vocabulary may be desirable from the point of view of lexical coverage, there is also the additional problem of increased acoustic confusability [3].

Recently researchers are using the Word Wide Web (WWW) as an additional resource of training data for language modeling adaptation procedures [4][5]. Different Information Retrieval (IR) techniques have been used for dynamic adaptation of vocabulary and/or language model to the topics present in a BN show using relevant documents obtained from a large general corpus or from the Web [2][6][7]. These multi-pass speech recognition approaches use the ASR hypotheses as queries to an IR system in order to select additional on-topic adaptation data. In [1] a different approach is suggested which uses a multi-pass recognition strategy to generate morphological variations of the list of all words in the lattice, thus dynamically adapting the recognition vocabulary.

In this paper, we propose a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and language model to the topic of the current news segment during a multi-pass speech recognition process. Based on texts daily available on the Web, a story-based vocabulary is selected using a morpho-syntatic technique derived in our previous work [8]. This technique uses part-of-speech (POS) word classification to compensate for word usage differences across the various training corpora. Using an Information Retrieval engine and the ASR hypotheses as query material, relevant documents are extracted from a dynamic and large-size dataset to generate a story-based language model. Since those hypotheses are quite small and may contain recognition errors, a relevance feedback method for automatic query expansion was used [9].

The proposed adaptation framework applied to a European Portuguese Broadcast News transcription system showed to be effective both in terms of OOV word rate reduction and recognition accuracy improvement (WER).

In section 2 we describe the baseline system and data sources used in our experiments. Section 3 and 4 provide a brief description of the vocabulary selection algorithm and information retrieval procedure. Section 5 describes the proposed multi-pass speech recognition framework and section 6 presents the multi-pass adaptation results, drawing in section 7 some conclusions.

2. BASELINE SYSTEM AND DATASETS

2.1. Baseline system

For the work presented in this paper, we used the system reported in [10]. This system is part of a closed-captioning system of live TV broadcasts in European Portuguese that is daily producing online captions for the main news show of one Portuguese Broadcaster.

This system features a hybrid HMM/MLP system, using 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 [11]. The decoder used under this baseline system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [12]. In this approach, the decoder search space is a large WFST that maps observation distributions to words.

The recognition vocabulary of the baseline system has been selected from two training corpora:

- NP-2003: 604M word corpus of newspapers texts collected from the WWW since 1991 until the end of 2003;
- BN-ALERT: 531K word corpus of broadcast news transcripts (in-domain dataset).

This vocabulary was created using words selected from the written news 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 acoustic training data were added to the vocabulary giving a total of 57,564 words (57K).

The baseline language model (LM) [13] combines a backoff 4-grams LM trained on NP-2003 corpus, and a backoff 3-grams LM estimated on BN-ALERT corpus. The two models were combined by means of linear interpolation, generating a mixed model.

This baseline system is the state-of-the-art in terms of Broadcast News Transcription Systems for the European Portuguese language.

2.2. Evaluation dataset

To evaluate the proposed framework we selected two BN shows from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP, which have been daily recorded and automatically transcribed by the European Portuguese BN transcription system reported in [10] since March 2003. The selected shows had a total duration of about 2 hours of speech (about 16K words) and were collected on May 24th and 31st of 2007.

Table 1 shows the statistics related to this evaluation corpus (RTP-07) with the OOV word rate for the baseline vocabulary of 57K.

Table 1. *Statistics for the RTP-07 dataset.*

	May 24th	**May 31st**
Word tokens	8188	7900
Word types	2303	2322
OOV tokens	110	116
%OOV tokens	1.34	1.47

The results are shown for both word tokens, in which all occurrences of a word are counted, and word types, in which only unique words are counted. All the OOV rates presented in this work are related to word tokens.

2.3. Adaptation dataset

For vocabulary and LM adaptation proposes the WEBNEWS corpus was used. This corpus consists of written news which are being collected from the online Portuguese newspapers web editions in a daily basis. Websites were selected for their content and reliability to better reflect the lexical and linguistic content of current news events. Due to the heterogeneous variety of sources, a normalization process is applied to the collected texts in order to clean errors due to common misspellings, expanding abbreviations and acronyms, processing ambiguous punctuation marks, and converting numbers to word sequences.

WEBNEWS corpus provides an average of 80K words per day. However, to construct a more homogeneous dataset to use on our daily adaptation framework, we merge the data from several consecutive days. Hence, for each day d, we use the texts from the current day and the 6 preceding days, which means, using one week of written news for daily adaptation proposes. In the following sections we denote this subset as $O_7(d)$.

In order to use the WEBNEWS corpus in our multi-pass speech recognition approach, the texts collected from the Web are provided in an article basis, being dynamically indexed and stored by the information retrieval engine implemented for this work and described in section 4.

3. VOCABULARY ADAPTATION ALGORITHM

An important problem in vocabulary design is to identify and rank the most relevant vocabulary words. Due to the large variety of topics discussed over time, this problem is even more serious for BN data. Even though the use of very large vocabularies in recognition systems can reduce the OOV word rates, in highly inflected languages or those with a high rate of word compounding, the OOV word rates still tend to be high [1]. The appropriate strategies to identify new words likely to be of interest will depend in some extend on the domain. The most common ones normally use the word frequency in each corpus as a criterion for the selection. Other approaches have been proposed where vocabulary adaptation is carried out by adding and removing words from the baseline vocabulary according to frequency and recency in contemporary written news [14][15]. In all these approaches, the baseline vocabulary is extended or modified by using all the new words appearing in the text corpus extracted from the Internet.

The method used in this work is a modified vocabulary selection technique which uses part-of-speech (POS) word classification to compensate for word usage differences across the various training and adaptation corpora. This approach is based on the hypothesis that the similarities between different domains can be characterized in terms of style (represented by the POS sequences). Hence, instead of simply add new words to the fixed baseline vocabulary, we use the statistical information related to the distribution of POS word classes on the in-domain corpus to dynamically select words from the various training corpora available, thus eliminating the need for human intervention. This algorithm was introduced in [9].

This algorithm can be briefly summarized as follows. Let $c_{i,j}$ be the counts from each of the available training corpora t_j ($j = 1,\ldots,n$), for the word w_i . Due to the differences in the amount of available data for each training corpus, we start by normalizing the counts according to their corpus length. The counts and normalization process were done using the SRILM toolkit [16] and Witten-Bell as the discounting strategy.

From these normalized counts ($\eta_{i,j}$) we want to estimate some kind of weighting η_i for each word w_i in order to select a vocabulary from the union of the vocabularies of t_1 through t_n that minimizes the OOV word rate for the in-domain task. In [17] this weighting is obtained by means of linear interpolation of the different counts, with the mixture coefficients λ_n calculated in order to maximize the probability of the in-domain corpus. Hence,

$$\eta_i = \sum_{j=1}^{n} \lambda_j \eta_{i,j} \qquad (1)$$

In our work we simply assigned identical values to all the mixture coefficients, i.e. $\lambda_j = \frac{1}{n}$. After η_i estimation, one generates a list W with all the words, sorted in descending order by the weighting factor η_i of each word w_i .

Assume V as the dimension of the target vocabulary. Hence, instead of simply selecting the first V words in W , we use statistical information related to the distribution of POS word classes on the in-domain corpus to dynamically select words. In our algorithm we used the following set of POS classes:

$$POSset = \{\text{names, verbs, adjectives, adverbs}\} \qquad (2)$$

Since the remaining POS classes (mainly containing functional words) represent almost closed grammatical classes in the European Portuguese language, their words were automatically added to the target vocabulary. In fact, in the training corpora used in this work we had only 468 words which POS class did not belong to $POSset$.

Hence, assuming $p \in POSset$ and $M(p)$ as the distribution of POS classes on the in-domain corpus, the number of words selected from W for each class p will be $V \times M(p)$. This way, for each class p the first $V \times M(p)$ words of W with the highest value η_i and belonging to class p are selected to be included in the target vocabulary.

In this work we used the in-domain BN-ALERT corpus to estimate the $M(p)$ distribution of POS classes. Using a morphological analysis tool developed for the European Portuguese language [18] the in-domain corpus was annotated. Table 2 shows the distribution of POS classes (in percentage) for the BN-ALERT corpus.

Table 2. *Distribution in percentage of POS classes for the BN-ALERT corpus.*

Names	Verbs	Adjectives	Adverbs
40.6	36.9	20.9	1.6

4. INFORMATION RETRIEVAL ENGINE

In this section, the information retrieval (IR) engine implemented within the multi-pass adaptation framework is described. The basic idea is as follows.

The initial ASR hypotheses (the result of the first decoding pass) which include texts on multiple topics are automatically segmented into individual stories, with each story ideally concerning a single topic. These segmentation boundaries are located by the audio partitioner [19] and topic segmentation procedure [20] currently implemented on the baseline system. The text of each segment can then be used as a query for the information retrieval engine to extract relevant documents from a general training dataset. This way a story-based corpus is extracted for each segment and used to dynamically build an adapted vocabulary and LM for each story present in the news show being recognized.

For our framework we looked for an information retrieval engine addressing the following requirements: the system architecture should support large-scale text databases, multiple databases, concurrent indexing and querying, fast indexing, different retrieval models, and relevance feedback models. According to these requirements we chose the Indri search engine [21] that is part of the Lemur Toolkit, an open-source toolkit for language modeling and information retrieval.

As the starting point the indexing of all training datasets (NP-2003 and BN-ALERT) has been done (about 1.5M articles). For the indexing phase we defined as term the concept of word. During the indexing/retrieval process we removed all the function words and the 500 most frequent words creating a stoplist of 800 words. The current IR dynamic database is now updated in a daily basis with documents collected for the WEBNEWS corpus.

In our preliminary experiments we used the cosine as the similarity measure for the retrieval phase. All articles with a score exceeding an empirically determined threshold are extracted for each story. Since the number of words in the hypothesized transcript of each story is usually small and contains transcription errors, one uses a pseudo-relevance feedback mechanism for automatic query expansion [8]. This method uses the ASR hypotheses as an initial query, do some processing, and then return a list of expansion terms. The original query is then augmented with the expansion terms and rerun.

5. MULTI-PASS ADAPTATION APPROACH

Most ASR systems use static language models with vocabularies selected from large and fixed training corpora. However, an open task likes the BN data transcription, is characterized by rapid changes in topics, with corresponding changes of linguistic content and vocabulary words. To overcome this limitation, we proposed and implemented a multi-pass speech recognition approach which creates from scratch both vocabulary and language model components in a daily basis. Hence, for each day d this approach is performed according to the following steps:

- using the training corpora NP-2003, BN-ALERT and the adaptation subset $O_7(d)$ a new vocabulary V_0 with V words is selected according to the selection algorithm described in section 3;
- with V_0 three language models are estimated: a generic backoff 4-grams LM trained on NP-2003, an in-domain backoff 3-grams LM trained on BN-ALERT and an adaptation backoff 3-grams LM trained on $O_7(d)$;
- ASR transcriptions generated for the 21 preceding days are merged together. This dataset is denoted as T_{21};
- the three LM are linearly combined. The interpolation coefficients are estimated using the Expectation-Maximization (EM) algorithm to maximize the likelihood of T_{21} dataset. Finally, the mixed language model (MIX_0) is pruned to a reasonable size using entropy-based pruning [22];
- using V_0 and MIX_0 in a first decoding pass (1-PASS-POS) the initial set of ASR hypotheses (H_0) is generated;
- H_0 is automatically segmented into individual stories using the topic detection procedure;
- **for each story** S a topic-related dataset (D_S) is extracted from the IR dynamic database. All words found in this dataset are added to V_0 generating this way a story-specific vocabulary V_S . Note that for each word added the vocabulary size is kept constant by removing the word with the lowest frequency. With V_S an adaptation backoff 3-grams LM trained on D_S is estimated and linearly combined with MIX_0 to generate a story-specific LM (MIX_S);
- using V_S and MIX_S in a second decoding pass (2-PASS-POS-IR) the final set of ASR hypotheses is generated **for each story** S .

For each vocabulary, pronunciations are derived using a rule-based phonetizer augmented with a set of exceptions added by hand.

For the RTP-07 evaluation dataset the number of stories in each show was 33 and 31 for May 24th and May 31st respectively, with an average number of 251 words per story.

6. RESULTS

A range of experimental results are reported for the European Portuguese BN transcription system. We used RTP-07 dataset to compare the performance of our multi-pass adaptation approach, with the one obtained for the baseline system with a vocabulary size of V=57K words. The performance is accessed in terms of OOV word rate and recognition accuracy (WER). In addition, comparison of results using different vocabulary sizes is also described.

Table 3. *Comparison of OOV rates for the RTP-07 dataset (for a vocabulary size V=57K words).*

Approach	%OOV	%reduction
BASELINE	1,40	-
1-PASS-POS	0,74	47,0
2-PASS-POS-IR	0,49	65,2

As one can observe from table 3, the proposed two-pass speech recognition approach (2-PASS-POS-IR) using the morpho-syntatic algorithm for vocabulary adaptation (POS) and the Information Retrieval Engine (IR) for language model adaptation, yields a relative reduction of 65.2% in OOV word rate, i.e. from 1,40% to 0,49%, when compared to the results obtained for the baseline system. Moreover, this approach outperformed the one based on one single-pass (1-PASS-POS).

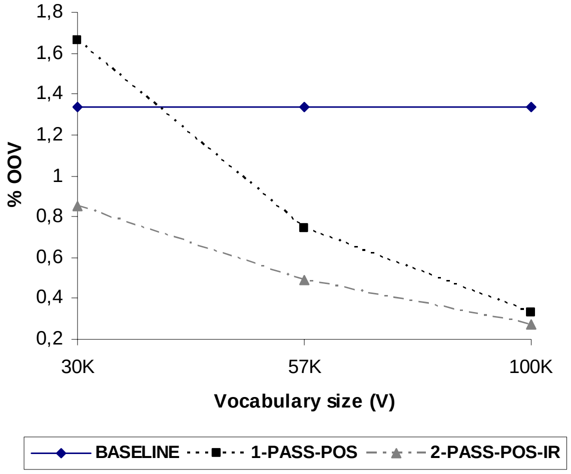
In terms of WER we got similar results, with the new approach (2-PASS-POS-IR) bringing a 6,6% relative gain (table 4). Analysis on the unknown words (OOV), which were found by our IR-based framework, showed that almost all the relevant terms like proper and common names (including names of persons, locations and organizations) were correctly recognized. This makes the framework especially useful since these words contain a great deal of information for novel applications where the use of automatic transcriptions is a major attribute.

Table 4. *Comparison of WER for the RTP-07 dataset (for a vocabulary size V=57K words).*

Approach	WER	%reduction
BASELINE	21,1	-
1-PASS-POS	19,9	5,7
2-PASS-POS-IR	19,7	6,6

To better understand the performance of this new vocabulary and language model adaptation procedure we calculated the OOV rate results for different vocabulary sizes (figure 1). Starting from the baseline vocabulary size of 57K words we defined two vocabulary sizes to test: a smaller one with about fifty percent of the size (30K) and another one with almost the double of the size (100K).

Figure 1. *OOV rate comparison for 3 different values of V (30K, 57K and 100K words).*



The graph in figure 1 shows the relative good performance of 1-PASS-POS and 2-PASS-POS-IR approaches for selection of large-sized vocabularies. Furthermore, as we would expect, for selection of small vocabularies better results are achieved by using the 2-PASS-POS-IR method. In fact, as one can see, with a vocabulary of 30K words we were able to get a better lexical coverage than the one obtained for the baseline system with a 57K words vocabulary.

In terms of accuracy we got a WER of 20,4% using a vocabulary of 30K against the 21,1% of the baseline system (see table 5). Even using a vocabulary with only 30K we were able to get a better WER with our adaptation framework than the one obtained for the baseline system.

Table 5. *WER comparison for 3 different values of V (30K, 57K and 100K words).*

Approach	30K	57K	100K
BASELINE	-	21,1	-
1-PASS-POS	21,0	19,9	19,5
2-PASS-POS-IR	20,4	19,7	19,3

Hence, implementing the proposed multi-pass adaptation approach and increasing the vocabulary size to 100K words we could obtain a relative gain of 8,5% in WER.

The baseline transcription system used in this work has already been updated according to this multi-pass speech recognition framework. Instead of using a static vocabulary and language model, the new system takes advantage from this dynamic procedure to better deal with new words appearing in BN shows in a daily basis. Hence, the first-pass is being used to produce online captions for the closed-captioning system of live TV broadcasts, while the second-pass is being used to produce offline captions.

7. CONCLUSIONS

This paper described a dynamic vocabulary and language model adaptation framework that tries to optimize the trade-off between the expected OOV word rate and the number of added words. We proposed a multi-pass speech recognition framework for a European Portuguese BN transcription system, using contemporary written texts available on the Internet and relevant documents extracted from a general corpus using an information retrieval engine. It uses POS class information about an in-domain training corpus to select an optimal vocabulary. When applied to a daily and real-time broadcast transcription task, this multi-pass approach showed to be effective in reducing the OOV word rate (more than 65%) and WER (about 6.6%) when compared to the results obtained by the baseline system with the same vocabulary size (V=57K).

To extend the effectiveness of this multi-pass adaptation framework, we will investigate the use of information retrieval techniques in the first-pass of speech recognition.

This way we will expect to take advantage of these techniques even for the online version of our broadcast news transcription system. Furthermore, we will explore our vocabulary adaptation algorithm in order to reduce the number of vocabulary words during the second-pass, improving this way the overall system performance.

8. ACKNOWLEDGMENTS

This work was partially funded by PRIME National Project TECNOVOZ number 03/165 and by the FCT project POSC/PLP/58697/2004. Ciro Martins is sponsored by a FCT scholarship (SFRH/BD/23360/2005).

9. REFERENCES

[1] Geutner, P., Finke, M., Sheytt, P., Waibel, A. and Wactlar, H., “Transcribing Multilingual Broadcast News using Hypothesis Driven Lexical Adaptation”, in Proc. of ICASSP, 1998.

[2] Bigi, B., Huang, Y. and Mori, R., “Vocabulary and Language Model Adaptation using Information Retrieval”, in Proc. of ICSLP, 2004.

[3] Rosenfeld, R., “Optimizing Lexical and n-gram Coverage via judicious use of Linguistic Data”, in Proc. Eurospeech, vol. 2, 1995.

[4] Schwarm, S., Bulyko, I. and Ostendorf, M., “Adaptive Language Modeling with Varied Sources to Cover New Vocabulary Items”, IEEE Transactions on Speech and Audio Processing, vol. 12, n. 3, May 2004.

[5] Allauzen, A., and Gauvain, J., “Diachronic vocabulary adaptation for broadcast news transcription”, in Proc. of Interspeech, 2005.

[6] Boulianne, G., et al., “Computer-assisted closed-captioning of live TV broadcast in French”, in Proc. of Interspeech, 2006.

[7] Chen, L., Gauvain, J., Lamel, L., and Adda G., “Dynamic Language Modeling for Broadcast News”, in Proc. of ICSLP, 2004.

[8] Martins, C., Teixeira, A., and Neto,J., “Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntatic Approach”, Proc. of Interspeech, 2007.

[9] Lavrenko, V., and Croft, W., “Relevance-Based Language Models”, SIGIR’01, 2001.

[10] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., "AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language", in Proc. of PROPOR 2003, Portugal, 2003.

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

[12] Caseiro, D., "Finite-State Methods in Automatic Speech Recognition". PhD Thesis, IST Technical University of Lisbon, Portugal, 2003.

[13] Martins, C., Teixeira, A. and Neto, J., “Language Models in Automatic Speech Recognition”, in Magazine of DET-UA. Aveiro, vol. 4, nº 4, 2005.

[14] Federico, M. and Bertoldi, N., “Broadcast news LM adaptation over time”, Computer Speech and Language, vol. 18, 2004.

[15] Auzanne, C., Garofolo, J. S. Fiscus, J. and Fisher, W., “Automatic Language Model Adaptation for Spoken Document Retrieval”, in Proc. of RIAO Content-Based Multimedia Information Access, France, 2000.

[16] Stolcke, A., “SRILM – An extensible language modeling toolkit”, in Proc. of ICSLP 2002, Colorado, 2002.

[17] Venkataraman, A. and Wang, W., “Techniques for Effective Vocabulary Selection”, in Proc. of Eurospeech, 2003.

[18] Ribeiro, R., Mamede, N. and Trancoso, I., “Morpho-syntactic Tagging: a Case Study of Linguistic Resources Reuse”, chapter of the book “Language Technology for Portuguese: shallow processing tools and resources”, Edições Colibri, Lisbon, Portugal, 2004.

[19] Meinedo, H., and Neto, J., “A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models”, in Proc. of Interspeech, 2005.

[20] Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I. and Neto, J., “Automatic vs. Manual Topic Segmentation and Indexation in Broadcast News”, in IV Jornadas en Tecnologia del Habla, pages 123--128, November 2006.

[21] Strohman, T., Metzler, D., Turtle, H., and Croft, W.B., “Indri: A language-model based search engine for complex queries (extended version)”, CIIR Technical Report, 2005.

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