Dynamic Language Modeling for the European Portuguese

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

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

Abstract. Up-to-date language modeling is recognized to be a critical aspect of maintaining the level of performance for a speech recognizer over time for most applications. In particular for applications such as transcription of broadcast news and conversations where the occurrence of new words is very frequent, especially for highly inflected languages like the European Portuguese. An unsupervised adaptation approach, which dynamically adapts the active vocabulary and language model during a multi-pass speech recognition process, is presented. Experimental results confirmed the adequacy of the proposed approaches. Experiments were carried out for a European Portuguese Broadcast News transcription system with the best preliminary results yielding a relative reduction of 65.2% in OOV word rate and 6.6% in WER.

1 Introduction

Up-to-date language modeling is recognized to be a critical aspect of maintaining the level of performance for a speech recognizer over time for most applications. In particular for applications such as transcription of broadcast news (BN) and conversations where the occurrence of new words is very frequent, especially for highly inflected languages. This is the case of the European Portuguese language, where new names contain great deal of information and occur frequently in many domains as the BN one. Additionally, due to their inflectional structure, the verbs class represents another problem to overcome [1]. For a BN transcription system like the one used in this work, the ability to correctly address new words appearing in a daily basis, is an important factor to take in consideration for its performance.

In this paper, we present and compare two daily and unsupervised adaptation frameworks, which dynamically adapt the active system vocabulary and LM. Based on texts daily available on the Web, we defined two morpho-syntatic approaches to dynamically select the target vocabulary by trading off between the OOV word rate and vocabulary size [1][2]. Using an IR engine [3] and the ASR hypotheses as query material, relevant documents are extracted from a dynamic and large-size dataset to generate a story-based LM to the multi-pass speech recognition framework.

In section 2 we provide a brief description of the proposed vocabulary selection algorithms, LM adaptation procedures, and their integration into a multi-pass speech recognition framework. Section 3 describes some evaluation results.

2 Vocabulary Selection and Language Model Adaptation

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, those rates still tend to be high. In addition, just generically increasing the vocabulary size can improve the accuracy for many common words but degrades the recognition rate for less common words. Thus, defining a more rational approach to select/adapt the system vocabulary other than by simple word frequency is need.

In [1] we derived a procedure for dealing with the OOV problem by dynamically increasing the baseline system vocabulary. From the experiments derived, we observed that verbs make up for the largest portion of OOV words types, accounting for 56.2% of the OOV word types in a BN test dataset. Our approach to compensate and reduce the OOV word rate related with verbs was supported by the fact that almost all the OOV verb tokens were inflections of verbs whose lemmas were already among the lemmas set (L) of the words found in contemporary written news. Thus, the baseline vocabulary is automatically extended with all the words observed in the language model training texts and whose lemmas belong to L. Applying this adaptation approach, the baseline system vocabulary of 57K was expanded by an average of 43K new words each day. To apply this selection process, both training and adaptation word lists were morpho-syntactically classified and lemmatized using a morphological analysis tool developed for the European Portuguese [4].

In [2] we proposed another approach. It takes in consideration the differences in style across the various training corpora, especially in case of written versus spoken style. Using the same morphological analysis tool as before, we annotated both in-domain corpus and out-of-domain corpus, observing a significant difference in part-of-speech (POS) tags distribution, especially in terms of names and verbs. Hence, instead of simply adding new words to the fixed baseline system vocabulary, as the previously proposed approach, we use now 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.

For LM adaptation we proposed and implemented a multi-pass speech recognition approach which creates from scratch both vocabulary and LM components in a daily basis [5]. The first-pass is being used to produce online captions for a closed-captioning system of live TV broadcasts. Based on texts daily available on the Web and static training corpora, a new vocabulary V_0 is selected for each day d using the POS-based technique described in section 2. To construct a more homogeneous adaptation dataset, we merge Web data from the current day and the 6 preceding days ($O_7(d)$). Finally, with V_0, three LMs are estimated and linearly combined. The mixture coefficients are estimated using the Expectation-Maximization (EM) algorithm to maximize the likelihood of T_{21} dataset. This T_{21} held-out dataset consists of ASR transcriptions generated by the BN transcription system itself for the 21 preceding days. A confidence measure is used to select only the most accurately recognized transcription segments.

In this multi-pass adaptation framework, a second-pass is being used to produce offline transcripts for each day using the initial set of ASR hypotheses generated during the live version and automatically segmented into individual stories, with each story ideally concerning a single topic. Using an Information Retrieval engine [3] and the text of each story segment as query material, relevant documents are extracted from a dynamic and large-size database to generate a story-based vocabulary and LM. Since those text story segments can be quite small and may contain recognition errors, a relevance feedback method for automatic query expansion was used [6]. Thus, for each story S a topic-related dataset D_S is extracted from the IR dynamic database and all words found in D_S are added to the vocabulary V_0 selected on the first-pass, 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 LM trained on D_S is estimated and linearly combined with the first-pass LM to generate a story-specific LM (MIX_S). Using V_S and MIX_S in a second decoding pass the final set of ASR hypotheses is generated for each story S.

3 Evaluation Results

All experiments reported in this work were done with the AUDIMUS.media ASR system [7]. 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 - RTP.

To evaluate the proposed framework we selected a BN dataset (RTP-07) consisting of BN shows collected from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP. The RTP-07 BN shows were collected on May 24th and 31st of 2007, having a total duration of about 2 hours of speech and 16.1K words.

Table 1. Comparison of OOV word rates for the **RTP-07** dataset.

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 1, the proposed second-pass speech recognition approach (2-PASS-POS-IR) using the POS-based algorithm for vocabulary adaptation and the Information Retrieval Engine (IR) for LM adaptation, yields a relative reduction of 65.2% in OOV word rate (from 1.40% to 0.49%), when compared to the results obtained for the baseline system with a vocabulary of 57K words. Moreover, this approach outperformed the one based on one single-pass (1-PASS-POS).

In terms of WER (figure 1), the new approach (2-PASS-POS-IR) resulted in a 6.6% relative gain. Even using a vocabulary with only 30K we were able to get a WER better than the one obtained for the baseline system with a 57K words vocabulary. Thus, 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 terms of WER.

Analysis on the OOV words, which were found by our IR-based framework, showed that almost all the relevant terms like proper and common names were correctly recognized. This makes the proposed framework especially useful, since these words contain a great deal of information for systems where the use of automatic transcriptions is a major attribute, as is the case of our BN transcription system.

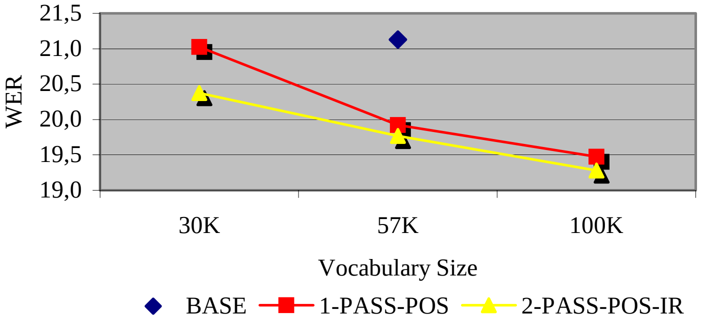

Fig. 1. WER comparison for 3 different vocabulary sizes (30K, 57K and 100K words).

4 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).

5 References

1. Martins, C., Teixeira, A., and Neto,J. (2006). Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News Transcription System. IEEE/ACL Workshop on Spoken Language Technology, December 2006.
2. Martins, C., Teixeira, A., and Neto, J. (2007). Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntatic Approach. In Proc. of Interspeech, 2007.
3. Strohman, T., Metzler, D., Turtle, H., and Croft, W.B. (2005). Indri: A language-model based search engine for complex queries (extended version). CIIR Technical Report, 2005.
4. Ribeiro, R., Mamede, N. and Trancoso, I. (2004). 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.
5. Martins, C., Teixeira, A., and Neto, J. (2007). Dynamic Language Modeling for a daily Broadcast News Transcription System. In Proc. of ASRU, 2007.
6. Lavrenko, V., and Croft, W. (2001). Relevance-Based Language Models. In Proc. of SIGIR’01, 2001.
7. Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I. (2003). AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language. In Proc. of PROPOR 2003, Portugal, 2003.