Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 24 (2010) 750–773
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Dynamic language modeling for European Portuguese ☆

Ciro Martins [a,b], António Teixeira [a,*], João Neto [b]

[a] Department Electronics, Telecommunications & Informatics/IEETA – Aveiro University, Aveiro, Portugal
[b] L2F – Spoken Language Systems Lab – INESC-ID/IST, Lisbon, Portugal

Received 18 February 2009; received in revised form 14 February 2010; accepted 15 February 2010
Available online 19 February 2010

Abstract

This paper reports on the work done on vocabulary and language model daily adaptation for a European Portuguese broadcast news transcription system. The proposed adaptation framework takes into consideration European Portuguese language characteristics, such as its high level of inflection and complex verbal system.

A multi-pass speech recognition framework using contemporary written texts available daily on the Web is proposed. It uses morpho-syntactic knowledge (part-of-speech information) about an in-domain training corpus for daily selection of an optimal vocabulary. 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. When applied to a daily and live closed-captioning system of live TV broadcasts, it was shown to be effective, with a relative reduction of out-of-vocabulary word rate (69%) and WER (12.0%) when compared to the results obtained by the baseline system with the same vocabulary size.

Keywords: Vocabulary selection; Language modeling; Information retrieval techniques; Automatic speech recognition (ASR); Broadcast news transcription

1. Introduction

Up-to-date vocabulary and, consequently, language modeling, is a critical aspect for maintaining the level of performance for a speech recognizer over time for applications such as transcription of broadcast news (BN) and broadcast conversations (BC). In particular, the subtitling of broadcast news programs is starting to become a very interesting application due to the technological advances in ASR and associated technologies. However, for this challenging task, the ability to correctly address new words appearing in a daily basis is an important factor to take into account for performance. Due to the changes in common topics over time, which characterizes this kind of task, the appearance of new stories leads to high out-of-vocabulary (OOV)

☆ Parts of this study were presented in conferences (Martins et al., 2007a,b, 2008).

* Corresponding author. Address: Departamento de Electrónica Telec. & Informática/IEETA, Universidade de Aveiro, Campus Universitário de Santiago, 3810 193 AVEIRO, Portugal. Tel.: +351 234370500; fax: +351 234370545.
E-mail address: ajst@ua.pt (A. Teixeira).

0885-2308/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.02.003

word rates and consequently to a degradation of recognition performance. Moreover, for languages with complex morphology, i.e., productive word formation processes (inflection, derivation and compounding), which create a large number of possible word forms, the OOV word rates tend to be even higher (Geutner et al., 1998; Kirchhoff et al., 2006).

In this paper, we present the work done in terms of dynamic language modeling for European Portuguese, which is a morphologically complex language whose inflectional structure represents an additional problem to overcome, mainly the morpho-syntactic class of verbs (Martins, 1998; Martins et al., 2006). We describe a daily and unsupervised LM adaptation framework, which was applied to a live closed-captioning system of live European Portuguese TV broadcasts (Meinedo, 2008; Neto et al., 2008). This subtitling system runs daily at RTP (the Portuguese public broadcast company), successfully processing the 8 o'clock evening news and allowing subtitling to be created live and in real-time for the TV broadcaster.

Based on texts that are available daily on the Web, we proposed a morpho-syntactic algorithm to dynamically select the target vocabulary by trading off between the OOV word rate and vocabulary size (Martins et al., 2007a). Using an information retrieval (IR) engine (Strohman et al., 2005) and the automatic speech recognition (ASR) hypotheses as query material, relevant documents were extracted from a dynamic and large-size dataset to generate a story-based language model (LM) for the multi-pass speech recognition framework. Because the hypotheses are quite small and may contain recognition errors, a relevance feedback method for automatic query expansion was used (Lavrenko and Croft, 2001).

1.1. Linguistic properties of European Portuguese

The European Portuguese language shares its characteristics with many other inflectional languages, especially those of the Romance family. European Portuguese words often exhibit clearer morphological patterns in comparison to English words. A morpheme is the smallest part of a word with its own meaning. In order to form different morphological patterns (derivations, conjugations, gender, number inflections, etc.), two parts of a word are distinguished: stem and ending. The stem is the part of the inflected word that carries its meaning, while the ending specifically denotes categories of person, gender and number, or the final part of a word, regardless of its morphemic structure. The following example of two semantically equal sentences ("O meu amigo é professor" and "A minha amiga é professora") differing in subject gender (masculine and feminine, respectively) but identical in English ("My friend is a teacher"), outlines its linguistic characteristics.

European Portuguese language distinguishes between three types of gender: masculine, feminine and neuter, while English only has one form. All nouns, adjectives and verbs in European Portuguese have a gender. They present far more variant forms than their English counterparts. Words have augmentative, diminutive and superlative forms (e.g. "small house" = "casinha", where – inha is the suffix that indicates a diminutive). Moreover, European Portuguese is a very rich language in terms of verbal forms. While the regular verbs in English have only four variations (e.g. talk, talks, talked, talking), the European Portuguese regular verbs have over 50 different forms, with each one having its specific suffix (Orengo and Huyck, 2001). The verbs can vary according to gender, person, number, tense and mood. Three types for the grammatical category of person (1st, 2nd and 3rd person) reflect the relationship between communication participants. There are five tenses: present, past, past perfect, past imperfect, past pluperfect and future. Another grammatical category, mood, denotes the feeling of the speaker towards the act, which is defined by the verb. There are eight different types of mood in European Portuguese: indicative, subjunctive, imperative, conditional, infinitive, inflected infinitive, participle and gerund.

The rich morphology of European Portuguese causes a large number of possible word types, which in turn decreases the overall quality of the recognition process (high OOV rates). Fig. 1 shows a comparison of the vocabulary growth rates (the increase in the number of word types versus number of word tokens for a given text) for European Portuguese, English and, as an example of a language of the same family as Portuguese, Italian. Information for European Portuguese was calculated on the European Portuguese Broadcast News Speech corpus (ALERT-SR) used in this work. For English the 1997 English Broadcast News Speech corpus (HUB4) available from the Linguistic Data Consortium (LDC) was used. Italian information, using the IBNC II, was extracted from Federico et al. (2000).

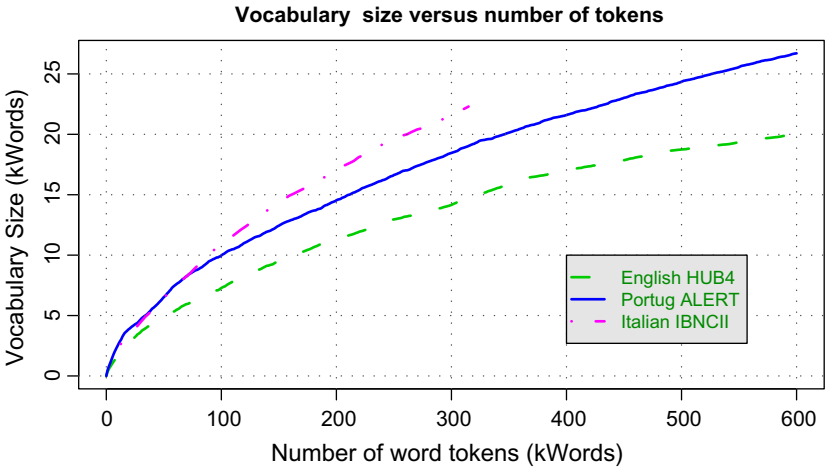

Fig. 1. Vocabulary growth rates for broadcast news data in European Portuguese (ALERT-SR dataset), Italian (IBNCII dataset) and English (HUB4 dataset).

As one can observe, the vocabulary growth rate of European Portuguese and Italian exceeds that of English significantly. For a corpus size of 600K word tokens, the HUB4 subset has a vocabulary size of 20,099 words (20K), while the vocabulary size for the ALERT-SR corpus is 26,704K (26K), i.e., about 32.9% more. This means European Portuguese, as well other languages like Italian, French, German, Spanish, etc., demand a larger vocabulary to obtain the same degree of coverage as English. In Lamel et al. (2004) the authors describe a work on broadcast and conversational speech transcription for multiple languages, where European Portuguese presents one of the smallest lexical coverages for a vocabulary size of 65K words (99.4% of coverage for English, 98.8% for French, 94.3% for Spanish and 94.0% for Portuguese).

1.2. Related work

Statistical language models have been successfully applied to many state-of-the-art ASR systems, with n-gram models being the dominant technology in language modeling. Usually, large corpora are used to estimate the LM parameters, and different smoothing techniques are applied to accurately estimate the LM probabilities (Xu and Jelinek, 2007). However, the collection of suitable training corpora is an expensive, time-consuming and sometimes unfeasible task. Therefore, the idea of language model adaptation is to use a small amount of domain specific data (in-domain data) to adjust the LM and reduce the impact of linguistic differences between the training and testing data over time. For that propose, several techniques have been developed by the research community (Bellegarda, 2004).

In terms of vocabulary selection, one approach is to choose the words in such a way as to reduce the OOV word rate by as much as possible – a strategy usually called vocabulary optimization. This optimization could either involve increasing the vocabulary size of the ASR component or selecting the words most representative for the target domain/task. The most common approaches are typically based on word frequency, including words from each training corpus that exceed some empirically defined threshold, which mainly depends on the relevance of the corpus to the target task (Gauvain et al., 2002). In Geutner et al. (1998) an approach targeted at reducing the OOV word rates for heavily inflected languages is suggested. Their work 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 for a second recognition pass. By applying this so called adaptation algorithm (HDLA-Hypothesis Driven Lexical Adaptation) both on Serbo-Croatian and German news data, OOV word rates were reduced by 35–45%. In Venkataraman and Wang (2003) three principled methods for selecting a single vocabulary from many corpora were evaluated, and they concluded that the maximum-likelihood-based approach is a robust way to select the vocabulary of a domain, especially when a reasonable amount of in-domain texts are available. A similar framework to the one presented in Geutner et al. (1998) was pro-

posed in Palmer and Ostendorf (2005), but it focused on names rather than morphological word differences. They proposed an approach for generating targeted name lists for candidate OOV words, which can be used in a second-pass of recognition. The approach involves offline generation of a large list of names and online pruning of that list by using phonetic distance to rank the items in a vocabulary list according to their similarity to the hypothesized word. Their reported experiments showed that OOV word coverage could be improved by nearly a factor or two with only a 10% increase in the vocabulary size. In Allauzen and Gauvain (2005b) a vectorial algorithm for vocabulary adaptation was used, which combines word frequency vectors estimated on adaptation corpora to directly maximize lexical coverage on a development corpus, thus eliminating the need for human intervention during the vocabulary selection process. The experiments reported by the authors showed a significant reduction of the OOV word rate compared with the baseline vocabulary: a relative decrease of 61% in French and 56% in English. The work presented in Oger et al. (2008) suggests that the local context of the OOV words contains relevant information about them. Using that information and the Web, different methods were proposed to build locally-augmented lexicons to be used in a final local decoding pass. This technique allowed the recovery of 7.6% of the significant OOV words, and the accuracy of the system was improved.

For broadcast news and conversational speech applications there have been various works using data from the Web as an additional source of training data for unsupervised language modeling adaptation over time, also referred to as dynamic LM adaptation. In Federico and Bertoldi (2004), a rolling language model with an updated vocabulary was implemented for an Italian broadcast news transcription system using a single-step adaptation framework. An open vocabulary language model was introduced by assuming a special OOV word class. Hence, the addition of new words to the vocabulary was done by extending the OOV word class and re-estimating its unigram distribution from the adaptation data. The baseline vocabulary of 62K words was extended by adding 60K new words selected from the contemporary written news and the baseline LM interpolated with a new language model estimated from the adaptation data. This approach allowed an average relative reduction of 58% in terms of OOV word rate and 3.4% in terms of word error rate (WER).

Another straightforward approach for unsupervised adaptation is to use different information retrieval (IR) techniques for dynamic adaptation of vocabulary and/or LM to the topics presented in a BN show using relevant documents obtained from a large general corpus or from the Web. 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 Bigi et al. (2004) an approach of this type using the Kullback–Leibler symmetric distance to retrieve documents was implemented to select a dynamic vocabulary instead of a static one, obtaining an OOV word rate reduction of about 28% with the same vocabulary size as the baseline vocabulary. Moreover, a new topic LM was trained on the retrieved data and interpolated with the baseline LM, allowing for a relative improvement of 1.8% in terms of WER. A similar approach was implemented in Chen et al. (2004). In this case the vocabulary remained static, with only the language model being updated at each automatically detected story of the BN show, which meant estimating multiple story-based language models for each BN show. A relative WER reduction of 4.7% was obtained in English and a 5.6% relative character error rate reduction in Mandarin. In Boulianne et al. (2006) a quite different approach was implemented for closed-captioning a live TV broadcast in French, which used texts from a number of websites, newswire feeds and broadcaster's internal archives to adapt its eight topic-specific vocabularies, re-estimating the corresponding language models according to the minimum discrimination information (MDI) method. In Tam and Schultz (2006) the same MDI method was used in a similar LM adaptation approach, but it used the latent Dirichlet allocation (LDA) model (Blei and Jordan, 2003) to estimate the marginal unigram distribution based on the ASR hypotheses. Results computed on a Mandarin BN test set showed a relative character error rate reduction of 2% when compared to baseline LM. Recently, various language model adaptation strategies, including unsupervised LM adaptation from ASR hypotheses and ways to integrate supervised maximum a posteriori (MAP) and marginal adaptation within an unsupervised adaptation framework, were investigated (Wang and Stolcke, 2007). By combining these adaptation approaches on a multi-pass ASR system, a relative gain of 1.3% on the final recognition error rate in the BC genre was achieved. More recently, in Lecorvé et al. (2008) and Lecorvé et al. (2009) some improvements were obtained by using IR methods for unsupervised LM adaptation.

The work presented here can be seen as an advancement and adaptation of previous works to deal with the specific characteristics of the European Portuguese language in the following terms: the lexicon adaptation used morphological knowledge related to the language, the language model adaptation approach used both morphological knowledge and information retrieval analysis, both language model and lexicon were adapted during the first and second recognition pass, and the adaptation was applied on a daily frequency, which is very important for real applications like the live closed-caption broadcast news transcription systems.

1.3. Outline of this paper

The remainder of this paper is structured as follows: Section 2 provides a brief description of the baseline European Portuguese broadcast news transcription system and datasets used to evaluate the proposed adaptation framework. Section 3 describes the new vocabulary selection and language model adaptation procedures, with experimental results reported in Section 4. Finally, in Section 5 those results are discussed, drawing some conclusions and describing future research trends.

2. Baseline system and resources

Through participation in several projects, we acquired over the years vast experience in developing ASR systems for the English language. Since then, we have been using that experience to develop ASR systems for the European Portuguese language. Currently, our ASR core system, AUDIMUS, has been ported successfully to several different tasks like dictation, telephone, portable devices, and broadcast news. This section gives a brief overview of the BN transcription system we used, which we improved with the work reported here. Further details about the overall system can be found in Neto et al. (2008). Finally, it also describes the resources used for training and evaluation proposes.

2.1. Broadcast news transcription system

All experiments reported in this paper were done with the AUDIMUS.media ASR system (Meinedo et al., 2003). This system is part of a closed-captioning system of live TV broadcasts in European Portuguese that produces online captions daily for the main news show of the main Portuguese broadcaster – RTP (Neto et al., 2008). It features a hybrid HMM/MLP system using three processing streams, each of them associated with a different feature extraction process and an MLP that is specifically trained, where the MLPs are used to estimate the context independent posterior phone probabilities given the acoustic data at each frame (Fig. 2). The phone probabilities generated at the output of the MLP classifiers are combined using an appropriate algorithm (Meinedo and Neto, 2000). All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus silence. The training and development of this system was based on the European Portuguese ALERT BN database (Neto et al., 2003). The acoustic models used on this work were trained with over 47 h of manually transcribed speech, plus 378 h of automatically transcribed speech (used for unsuper-

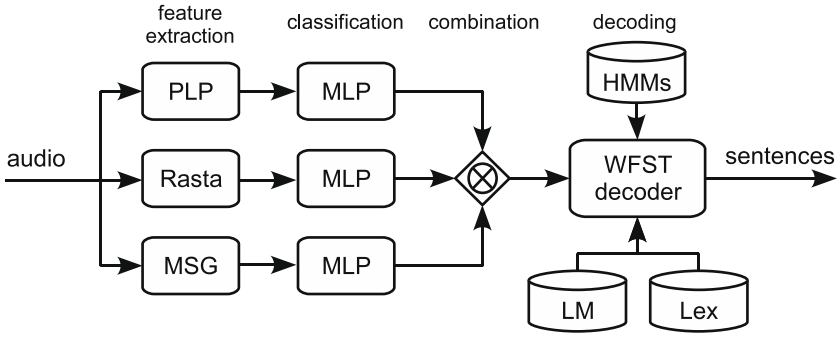

Fig. 2. AUDIMUS.media: a hybrid HMM/MLP recognition system. Extracted from Meinedo (2008).

vised training of the acoustic model). The decoder of this baseline system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition (Caseiro, 2003). In this approach, the decoder search space is a large WFST that maps observation distributions to words.

The recognition vocabulary of the baseline system was selected from two training corpora: a 604M word corpus of newspapers texts collected from the Web since 1991 until the end of 2003 (out-of-domain dataset – "NP.train"), and a 531K word corpus resulting from the manual transcription process of the 47-h training set (in-domain dataset – "BN.train"). This vocabulary was created using words selected from the written news set according to their weighted class frequencies of occurrence. All new different words present in the acoustic training data were added to the vocabulary, giving a total of 57,564 words (57K). This word list was then phonetically transcribed by a rule grapheme-to-phone system (Caseiro et al., 2002), generating an initial set of pronunciations. This automatically generated lexicon was then hand revised by a specialized linguist, generating a multi-pronunciation lexicon with 66,132 different pronunciations. The hand-held revisions and grammatical rules added to the generation of our phonetic pronunciations dictionary significantly contributed to the results obtained with the adaptation framework proposed here. The regular update of the pronunciation dictionary is especially useful in the case of new words such as proper names and foreign names. In fact, even if we are able to generate a better lexicon and an LM containing new names, if these have poor phonetic pronunciations, they will not be recognized in most cases.

The baseline LM (Martins et al., 2005) combines a backoff 4-g LM trained on the 604M word corpus and a backoff 3-g LM estimated on the 531K word corpus. The two models were combined by means of linear interpolation, generating a mixed model. This baseline system is state-of-the-art in terms of BN transcription systems for the European Portuguese language.

2.2. Evaluation datasets

To evaluate the proposed adaptation frameworks we selected two BN datasets consisting of BN shows collected from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP. The "BN.RTP-2004" dataset was drawn from the week starting on March 8th and ending on March 14th of 2004. Due to the unexpected and awful events occurring on March 11th of 2004 in Madrid (the Madrid train bombing), we would expect to cover a special situation of rich content and topic change over time. The "BN.RTP-2004" dataset has a total duration of about 5 h of speech and 52.5K word tokens. Later on, another evaluation dataset was collected, the "BN.RTP-2007", consisting of two BN shows that were randomly selected. Those two BN shows have a total duration of about 2 h of speech and 16K word tokens and were collected on May 24th and 31st of 2007.

Tables 1 and 2 present more detailed statistics related to these two evaluation datasets ("BN.RTP-2004" and "BN.RTP-2007", respectively). As one can observe, each BN show has an average size of about 8.3K tokens and only 2.1K different words (types). For day 9 of the "BN.RTP-2004" dataset, due to a technical problem, only half an hour of speech was collected. For that reason, this day has different statistics when considering the average values.

3. Dynamic language modeling for European Portuguese

This section addresses the problem of dynamically adapting over time both the vocabulary and language model of our European Portuguese broadcast news transcription system, using additional adaptation texts extracted from the Web. First, we present an analysis of the OOV words for the European Portuguese language using the evaluation datasets described in Section 2, which lead us to propose the multi-pass speech rec-

Table 1
BN.RTP-2004 dataset: text statistics.

	8th	9th	10th	11th	12th	13th	14th
#word tokens	8.7K	1.9K	8.4K	8.3K	8.8K	7.0K	9.4K
#word types	2.4K	0.7K	2.4K	2.0K	2.1K	1.8K	2.1K

Table 2
BN.RTP-2007 dataset: text statistics.

	May 24th	May 31st
#tokens	8.1K	7.9K
#types	2.3K	2.3K

ognition approach described in the following sub-sections. With respect to vocabulary selection, we describe the proposed algorithm and its integration in both single and multi-pass adaptation approaches.

3.1. OOV word analysis for European Portuguese

A major part of building a language model is to select the vocabulary of the ASR component that will have maximal coverage for the expected task/domain. Thus, the appearance of OOV words during the recognition process is closely related to the way the system vocabulary is chosen. Thus, an important aspect to take into account for vocabulary design is to identify and rank the most relevant vocabulary words in order to improve the coverage rate of that vocabulary on unseen data. Fig. 3 shows the coverage statistics related to the BN.RTP-2004 evaluation dataset for a baseline vocabulary of 57K words. For this vocabulary the OOV word rate averages 1.20%. To figure out what words contribute to these OOV rates, and which adaptation procedures we should pursue in order to better address this problem specific to highly inflected languages such as European Portuguese, we derived various analyses at the OOV word level.

First, we examined their classification into part-of-speech (POS) classes using a morpho-syntactic tagging system developed for the European Portuguese language (Ribeiro et al.,2003, 2004), presenting an overall success rate of 94.23% in identifying content words (proper and common names, verbs, adjectives and adverbs). In Table 3 we break down OOV words into three different categories using the morpho-syntactic tagging system: names (including proper names), adjectives and verbs. Other type of words, such as function words, are absent from the list shown in Table 3 because almost all those words are already in the 57K baseline vocabulary. We simply merged all of them together ("Others" category in Table 3).

According to findings reported in the literature, OOV words are mostly names. In Hetherington (1995), Bazzi (2002), Allauzen and Gauvain (2005a) a strong correlation between names and OOV words is reported. A similar conclusion is reported in Palmer and Ostendorf (2005), with names accounting for 43.66% of the OOV word types. Hence, as a first idea, we would expect to observe a similar behavior for the BN.RTP-2004 dataset, i.e., a strong correlation between names and OOV words, especially for this specific week with new and infrequent words appearing (train station names, terrorist names, journalist names, etc.). However, as one can observe from Table 3, verbs make up for the largest portion of OOV words. In fact, although verbs

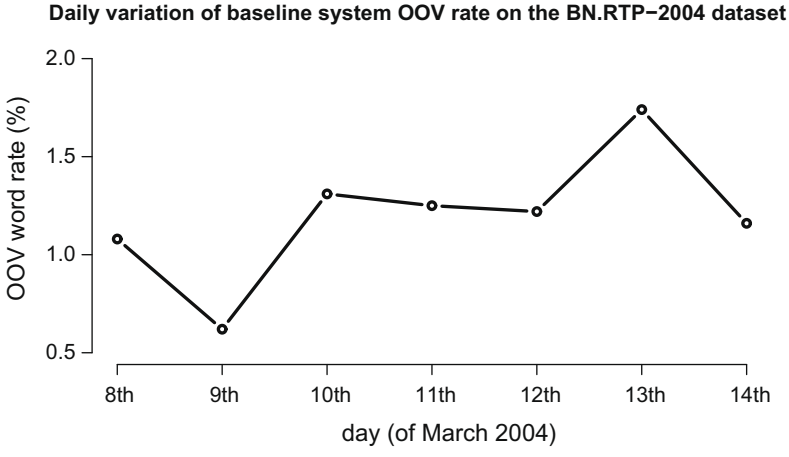

Fig. 3. OOV word rates (in%) measured on the seven news shows of the BN.RTP-2004 dataset with the baseline vocabulary of 57K words.

Table 3
Distribution (in%) of OOV words by POS-classes measured on the seven news shows of the BN.RTP-2004 dataset with the baseline vocabulary of 57K words. The last column shows the average values for this dataset.

Class	8th	9th	10th	11th	12th	13th	14th	Week
Names	27.0	15.4	18.9	20.4	28.2	42.9	34.7	27.9
Adjectives	10.1	7.7	18.2	16.9	16.4	8.9	9.8	13.2
Verbs	61.6	69.2	62.8	58.0	53.6	47.2	53.3	56.7
Others	1.3	7.7	0.1	4.7	1.8	1.0	2.2	2.2

represent only 17.6% (see Fig. 4) of the words in the BN.RTP-2004 dataset, they account for 56.7% of the OOV words.

From the 56.7% of verb OOVs, only 13.2% are due to new lemmas not already included in the lemma set of the 57K baseline vocabulary, i.e., the major part of verb OOVs are due to new inflection forms of lemmas already present in the vocabulary. Moreover, in this dataset, verbs are also very frequently the source of recognition errors, representing the largest portion of wrongly recognized words: 26.3% (see Fig. 5).

In a second analysis, and because our adaptation proposal was to take advantage of contemporary written news to dynamically adapt the system vocabulary, we examined the effect of augmenting the vocabulary with new words found in the same day of each tested BN show. Thus, taking into consideration these written text news collected for each day and the 57K word baseline vocabulary, an average of 5K new words were found in a daily basis, accounting for an upgraded vocabulary of 62K words (57K from the baseline vocabulary plus the average 5K new words).

By expanding the baseline vocabulary with those extra 5K words one can observed an OOV word reduction ranging between 15.4% (for March 9th) and 36.4% (for March 12th), with an average value around 28.4%. There are two main reasons for those variations: on day 9 of the BN.RTP-2004 dataset, due to a technical problem, only 12 min of speech was collected (in average one news show has about 45 min). For that reason, this day has different statistics in terms of the average values. The news show of day 13 has a different structure compared to the other days because it is composed mainly of a street interview in Madrid-Spain (outside studio with a large amount of spontaneous speech).

The graph in Fig. 6 gives us an overview of the kinds of words (POS-classes) we covered by adding the 5K extra new words. From that, we concluded that a significant OOV word reduction was obtained in the class of names. Some examples of OOV names recovered include important names related to the bombing (Alcalá, Atocha, El Pozo, Gregorio, Henares, Marañón, Rajoy, ...). Remarkably, on the news show of March 13th a reduction of 72% was achieved in the class of names.

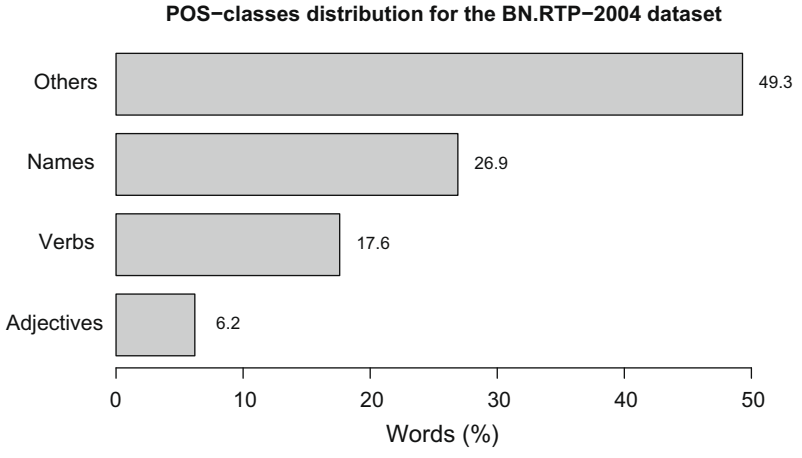

Fig. 4. Average distribution (in%) of words by POS-classes measured for the seven news shows of the BN.RTP-2004 dataset.

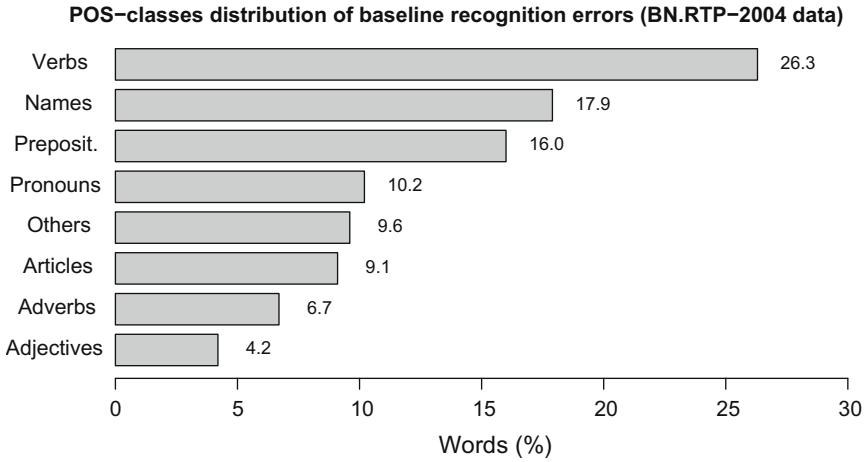

Fig. 5. Average distribution (in%) by POS-classes of the words recognized incorrectly in the seven news shows of the BN.RTP-2004 dataset. Recognition results obtained with the baseline system (vocabulary size of 57K words).

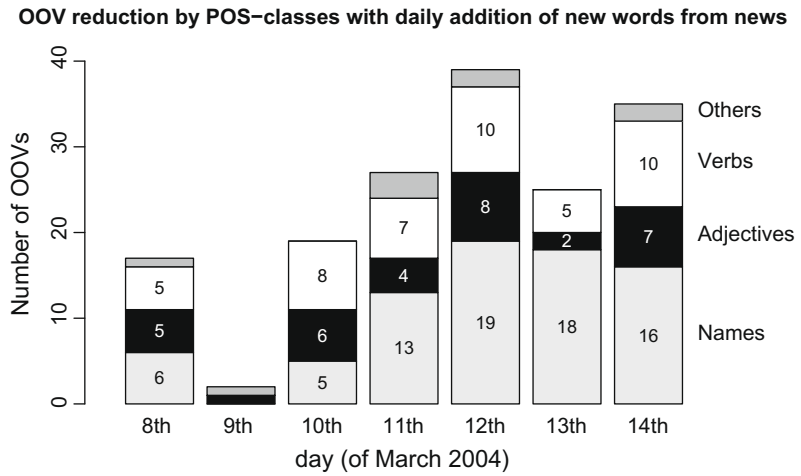

Fig. 6. OOV word rates reduction (in%) by POS-classes measured for the seven news shows of the BN.RTP-2004 dataset with the baseline vocabulary of 57K words augmented with new words found in the written text news of the same day of the news show tested.

Based on the above observations, we concluded that the strategy of using contemporary written text news to adapt the baseline vocabulary is useful, especially in covering new names that appear over time. However, even though verbs represent the largest portion of OOV words, the reduction for this class using contemporary written texts is not so significant. This is mainly due to the inflectional structure of verbs class for the European Portuguese language, which makes the use of verbs significantly different in written and spoken language. Moreover, the differences in terms of vocabulary growth and coverage for different domains and time periods makes it necessary to devise new vocabulary selection strategies that take into account those specific characteristics.

3.2. Vocabulary selection

In our preliminary works we tried to use a large vocabulary of 213K words selected according to their frequency of occurrence, obtaining an OOV word rate reduction of 79.2%, i.e., from 1.20% to 0.25%. However, this approach does not solve the problem of new words and infrequent words related to some important

events, which are critical and therefore need to be recognized accurately. Moreover, for applications such as the one we are addressing in our work, the processing time is a very important issue because we are doing online and real-time BN closed-caption transcription. Having larger vocabularies implies that the recognition process will require more time. Tables 2.1 and 2.2 in Section 2.2 show a maximum of only 2.4K word types occurring in a news show. Thus, defining a more rational approach to selecting the vocabulary other than by simple frequency of occurrence is needed.

In Martins et al. (2006) we proposed a procedure for dealing with the OOV problem by dynamically increasing the baseline system vocabulary, reducing the impact of linguistic differences over time. Our approach to compensate and reduce the OOV word rate related to verbs was supported by the fact that almost all the OOV verb tokens were inflections of verbs whose lemmas were already among the lemma set (L) of the words found in contemporary written news. Thus, the baseline system vocabulary is automatically extended with all the words observed in the language model training texts whose lemmas belong to L. Applying this adaptation approach on the seven news shows of the BN.RTP-2004 dataset, the baseline system vocabulary of 57K was expanded by an average of 43K new words each day, generating a significant improvement in terms of OOV word rate, which was reduced on average by 70.8%, i.e., from 1.20% to 0.35%. However, this procedure assumes an a priori selected static list of words – the baseline vocabulary – and adds new words in a daily basis. In this way, the system vocabulary is always extended, resulting in a vocabulary with an average size of 100K words. Thus, in Martins et al. (2007a) we derived a new algorithm for selection that allows for the definition of the size of the target vocabulary, selecting it from scratch.

Using the same morpho-syntactic analysis tool as before, we annotated both the in-domain and out-of-domain training corpora (BN.train and NP.train, respectively). In Fig. 7, we summarize the POS statistics obtained for both corpora by breaking down words into four main classes: names (including proper and common names), verbs, adjectives and adverbs. Other types of words, such as the functional words, are absent from the list shown in Fig. 7 because they represent almost closed grammatical classes in the European Portuguese language. These statistics are related to word types, i.e., only unique occurrences of a word/class are counted. As one can see, there is a significant difference in POS distribution when comparing in-domain and out-of-domain corpora, especially in terms of names and verbs. For in-domain data we observe a significant increment (from 30.5% to 36.9%) in the relative percentage of verbs when compared with the out-of-domain data, with the percentage of names decreasing from 45% to 40.6%.

Based on the observations above, we proposed a new approach for vocabulary selection that uses the part-of-speech 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). In Iyer and Ostendorf (1997) these simi-

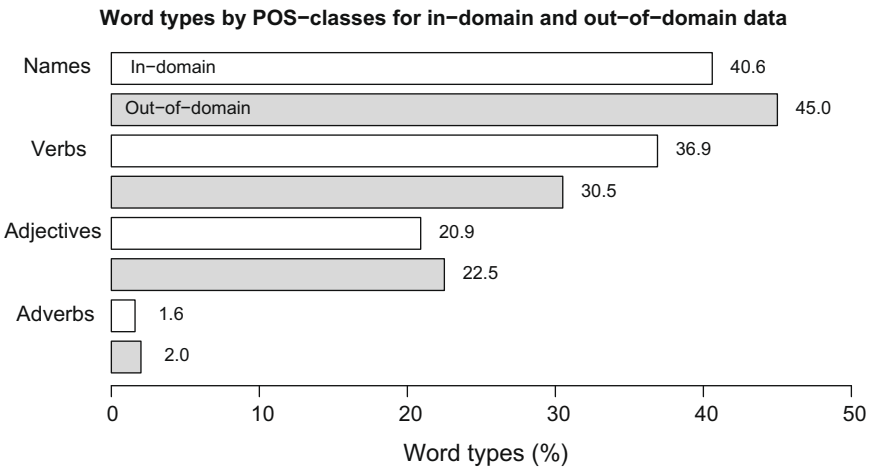

Fig. 7. Distribution (in%) of word types by POS-classes measured on two corpora with the same size: an in-domain corpus (broadcast news texts) and an out-of-domain corpus (written news texts).

larities have already been integrated to more effectively use out-of-domain data in sparse domains by introducing a modified representation of the standard word n-grams model using part-of-speech labels that compensate for word usage differences across domains. Thus, in this new approach, instead of simply adding new words to the fixed baseline system vocabulary, 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. Assume that we want to select a vocabulary V with $|V|$ words from n training corpora T_j, with $j = 1, \ldots, n$. The proposed approach can be described as follows:

3.2.1. Computation of normalized counts

Let $c_{i,j}$ be the counts from each one of the available training corpus T_j for the word w_i. In our case, we used three training corpora (i.e., $j = 3$): the written news text corpus, the broadcast news text corpus, and the written news texts collected from Internet on a daily basis. Due to the differences in the amount of available data for each training corpus, we started by normalizing the counts according to their respective corpus length, obtaining $\eta_{i,j}$ as the normalized counts. The Witten–Bell discounting strategy was used to ensure non-zero frequency words in the normalization process.

3.2.2. Estimation of a word weighting factor

From the normalized counts $\eta_{i,j}$ we wanted to estimate some kind of weighting factor 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 Venkataraman and Wang (2003) this weighting is obtained by means of linear interpolation of the different counts, with the mixture coefficients calculated in order to maximize the probability of the in-domain corpus. However, in our evaluation framework this technique performed slightly worst (in terms of OOV word rate) than simply applying the uniform distribution, maybe due to the relatively small size of our in-domain corpus. Hence, we simply assigned identical values to all the mixture coefficients,

$$\eta_i = \sum_{j=1}^{n} \lambda_j \eta_{i,j} \quad \text{with} \quad \lambda_j = \frac{1}{n} \tag{3.1}$$

3.2.3. Generation of an ordered word list W

All the words w_i were sorted in descending order according to the weighting factor η_i.

3.2.4. Definition of the POS-classes to use

In our implementation we used the following set of POS-classes:

$$POSset = \{\text{names}, \text{verbs}, \text{adjectives}, \text{adverbs}\} \tag{3.2}$$

All the remaining words (mainly functional words) were automatically added to the vocabulary. In fact, in the training corpus used in this work we obtained only 468 words whose POS-classes did not belong to *POSset*.

3.2.5. Estimation of POS distribution using an in-domain corpus

Using an in-domain dataset the distribution of words by POS-classes, $M(p)$ with $p \in POSset$, was computed through the maximum likelihood estimation (MLE).

3.2.6. Selection of $|V|$ words from the word list W

According to $M(p)$, the number of words selected from each class p will be $|V| \times M(p)$. Hence, for each class p, the first $|V| \times M(p)$ words of W belonging to class p were selected and included in the target vocabulary V. However, because a word can belong to more than one class, the first run of this process can produce a vocabulary list with less than $|V|$ words. In that case, the selection process is iterated until the target number of words is achieved.

For the proposed language model adaptation framework we used this new POS-based algorithm for vocabulary selection. In addition, we combined the POS-based method with our previous work based on lemmatization and inflection generation (Martins et al., 2006) to integrate them into the second-pass of the language model adaptation procedure as described in the next sub-section.

3.3. Multi-pass language model adaptation framework

As stated in Section 2.1, the baseline AUDIMUS.media ASR system is part of a closed-captioning system of live TV broadcasts, which is state-of-the-art in terms of broadcast news transcription systems for European Portuguese. However, the language modeling component of this baseline system uses a static vocabulary and language model, which is not able to cope with changes in vocabulary and linguistic content over time. To overcome that limitation we proposed and implemented an adaptation approach (Martins et al., 2007b), which creates from scratch both vocabulary and language model components on a daily basis using a multi-pass speech recognition process.

The first-pass was used to produce online captions for the closed-captioning system of live TV broadcasts (see Fig. 8). A new vocabulary V_0 was selected for each day d by applying the POS-based algorithm described in Section 3.2 and using three corpora as training data: the newspaper texts corpus NP.train, the broadcast news transcriptions corpus BN.train and the contemporary texts daily extracted from the Web (as adaptation data). The Web texts were selected from the online latest news of all major Portuguese newspapers, which included newspapers of different styles (daily newspapers covering all topics, weekly newspapers with a broad coverage of topics, economics newspapers and sports news). These newspapers were selected for their content and reliability to better reflect the lexical and linguistic content of current news events. However, only an average of 80K words was collected per day as adaptation data. Thus, to construct a more homogeneous adaptation dataset and collect enough n-grams containing new words, we merged Web data from several consecutive days. In our work we considered a heuristic time span of 7 days. Similar approaches were taken in Federico and Bertoldi (2004) and Allauzen and Gauvain (2005b). Hence, for each day d, we used the texts from the current day and the six preceding days (we denote this adaptation subset as $O_7(d)$ – 7 days of online written news). For the POS-based algorithm, we used BN.train as the in-domain corpus to estimate the POS distribution function. Using the selected vocabulary V_0, three language models were estimated: a generic backoff 4-g language model trained on NP.train, an in-domain backoff 3-g language model trained on BN.train and an adaptation backoff 3-g language model trained on $O_7(d)$. The generic language model was estimated using the modified Kneser–Ney smoothing (Ney et al., 1997), with absolute discounting used to estimate the other two language models. Finally, the three LMs were linearly combined. The mixture coefficients were estimated using the expectation–maximization (EM) algorithm to maximize the likelihood of a held-out dataset. For that purpose, we defined as our held-out dataset the set of ASR transcriptions generated by the broadcast news transcription system itself for the 21 preceding days (noted here as $T_{21}(d)$), i.e., 3 weeks of automatically

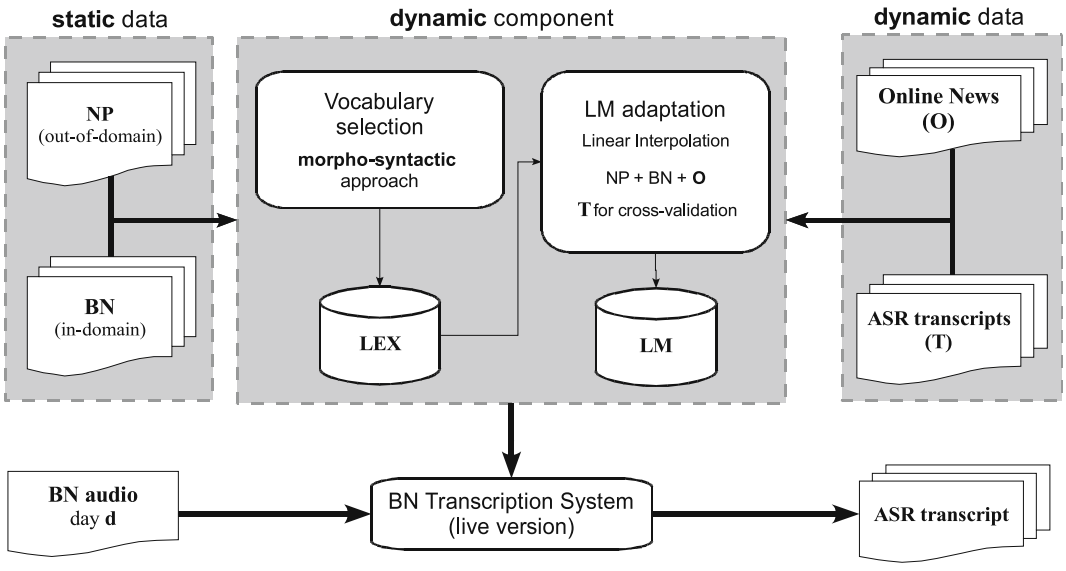

Fig. 8. Proposed multi-pass language model adaptation framework: first-pass (live version).

generated captions (generated by the second-pass). A confidence measure was used to select only the most accurately recognized transcription words. Thus, all the words w_i with a confidence value $P(correct|w_i)$ higher than 91.5% were included in the T_{21} (d) dataset. This is an important issue because recognition errors can skew the n-gram estimates and thus deteriorate the adapted language model. In fact, in Tam and Schultz (2006) and Wang and Stolcke (2007) a degradation of the recognition performance was reported when the baseline language model was adapted based on automatic transcriptions, with the authors postulating that this may have been caused by the recognition errors that were not smoothed properly. Finally, the mixed language model was pruned using entropy-based pruning (Stolcke, 1998). For a pruning threshold equal to 1e−09, we can obtain a language model that is about 30% smaller than the original one without significant degradation of WER (Martins et al., 2005).

In this multi-pass adaptation framework, a second-pass (Fig. 9) was used to produce improved offline transcriptions for each day using the initial set of ASR hypotheses generated during the live version. The basic idea is as follows.

The initial set of ASR hypotheses (the result of the first decoding pass), which include texts on multiple topics, is automatically segmented into individual stories with each story ideally concerning a single topic. These segmentation boundaries are located by the audio partitioner (Meinedo and Neto, 2005) and topic segmentation procedure (Amaral et al., 2006) currently implemented on the baseline system. The text of each segment is then used as a query for an information retrieval engine to extract relevant documents from a dynamic and large-size database. This way, a story-based dataset is extracted for each segment and used to dynamically build an adapted vocabulary and language model for each story present in the news show being recognized. For this framework we used the information retrieval engine INDRI – a language-based search engine (Strohman et al., 2005).

As the starting point, the indexing of all training datasets (NP.train and BN.train) was done, generating a total of about 1.5M articles indexed. For the indexing process 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 the contemporary texts, i.e., the texts used to generate the O_7 (d) dataset.

For the work presented in this study we used the *cosine* as the similarity measure for the retrieval phase. Thus, all articles with an IR score exceeding an empirically determined threshold were extracted for each story. To collect, for each topic, a dataset of texts similar in size to the BN.train and O_7 (d) datasets (531K and 560K words respectively), we extracted the first 1000 articles with the highest IR score for each one of the queries, which gave an average of 610K words. Because the number of words in the hypothesized tran-

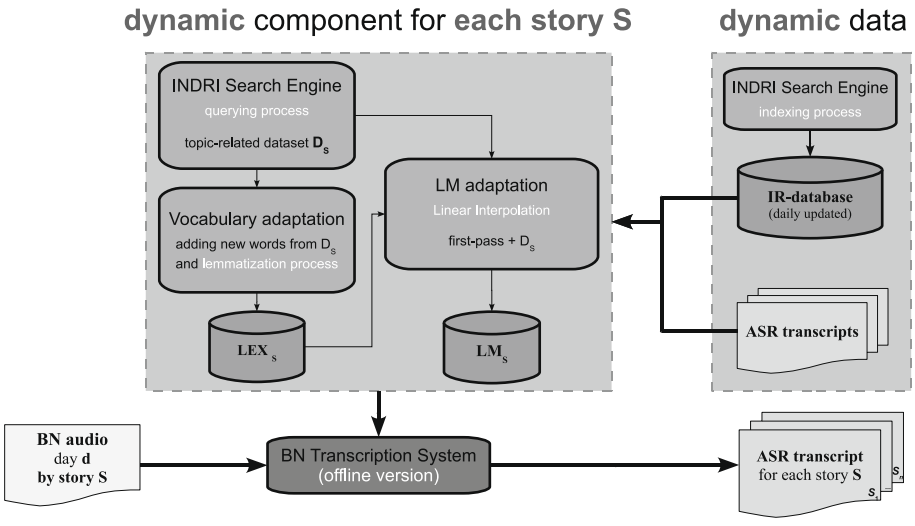

Fig. 9. Proposed multi-pass language model adaptation framework: second-pass (offline version).

script of each story was usually small and contained transcription errors, a pseudo-relevance feedback mechanism was used for automatic query expansion (Lavrenko and Croft, 2001). This method uses the ASR hypotheses as an initial query, does some processing and then returns a list of expansion terms. The original query is then augmented with the expansion terms and rerun. In our framework the 30 top retrieved and returned articles were used for query expansion. For training proposes only the 1000 articles with the highest IR score returned by the pseudo-relevance feedback procedure were used.

The functioning of the IR step was assessed by manual annotation regarding the relevance of each of the 1000 returned documents for three randomly topics/stories of BN.RTP-2007. Because IR evaluation is not the main focus of the present paper and topics were randomly selected and attributed to annotators, we consider the reduced number not to be a problem. Each topic was annotated by a different annotator. Despite preventing judge-to-judge correlation, it made it possible to have results for the three topics with a reasonable effort while solving the natural limitations of a unique annotator. In the annotations, a somewhat broad definition of the topic was adopted. As an example, for one of the selected topics, regarding losses due to bad weather, different weather problems (heavy rain, wind, etc.) were considered. Fig. 10 presents the accumulated number of relevant documents as a function of the number of documents. We considered this information more meaningful than presenting only the mean average precision.

For the three topics, the average number of relevant documents was 682 and the minimum 536. Mean average precision obtained was 0.79, which is the average precision for the individual topics of 0.74, 0.76 and 0.88. Therefore, the results show that the IR step is capable of returning a reasonable number of relevant documents.

For each story S a topic-related dataset D_s was extracted from the IR dynamic database, and all words found in D_s were added to the vocabulary V_0 selected on the first-pass, generating in this way a story-specific vocabulary V_S. Note that for each new word found in D_s we removed from V_0 the word with the lowest frequency, keeping the vocabulary size of vocabulary V_S equal to V_0. After this procedure, we augmented the V_S vocabulary by using the lemmatization and inflection generation method reported in our previous work (Martins et al., 2006). Hence, the V_S vocabulary was expanded with all the inflection forms of verbs whose lemmas were present in the text of each story S. Applying this lemmatization method resulted in an average of 800 new words added to the vocabulary V_S of each topic. Finally, with V_S, an adaptation backoff 3-g LM trained on D_s was estimated and linearly combined with the first-pass LM (MIX_0) 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 was generated for each story S.

According to the proposed framework, the following daily steps have been implemented in the current live system, which is generating closed-captions in real-time for the 8 o'clock evening news show of the Portuguese public broadcast company RTP:

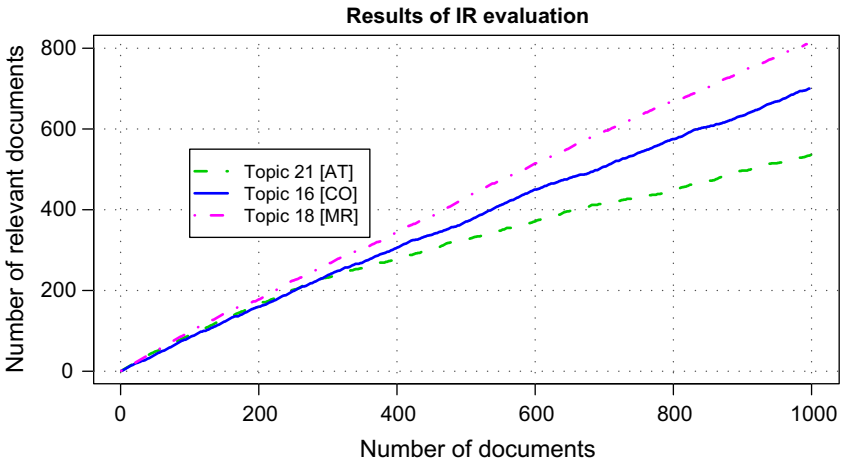

Fig. 10. Results of evaluation, regarding relevance, for the 1000 documents resulting from IR queries for three randomly selected topics of the BN.RTP-2007 evaluation set.

- Using the RSS News feeds services of six different Portuguese news channels, latest news are collected from the Web, normalized, stored and indexed by the IR engine (an average of 130 articles and 80K words collected per day). This process checks for more news blocks every hour. These six news channels were selected for their content and reliability to better reflect the lexical and linguistic content of current news events (news feeds with a broad coverage of topics, economics and sports news feeds). In particular, the news feed available at the RTP website and provided by the news agency of Portugal (LUSA – Agência de Notícias de Portugal) was selected because some of the stories presented at news shows are based on them.
- At 5 o'clock in the evening, and using the news texts collected up until then, a new vocabulary and LM are generated according the first-pass of the proposed adaptation approach and used by the ASR module at 8 o'clock to generate closed-captions for the TV news show.
- At the end of the TV news show, the second-pass is processed, generating improved BN transcriptions.

By applying this multi-phase adaptation approach we expect to improve the system performance over the first-pass. In the next section we will describe the experiments we have performed for evaluation purposes.

4. Evaluation results

To compare and evaluate the proposed adaptation approaches, a range of experimental results are reported using the baseline system described in Section 2.1. In these experiments we used two evaluation metrics: OOV word rate and WER over the two evaluation datasets (BN.RTP-2004 and BN.RTP-2007).

Comparison of results for both live and offline approaches using different vocabulary sizes is also described, allowing us to evaluate the effects of plain vocabulary growth on our adaptation framework. Hence, besides the baseline vocabulary size of 57K, we defined three more vocabulary sizes for evaluation purposes: a smaller one with about fifty percent of the baseline vocabulary size (30K words), another one with almost the double of the size (100K words) and a third with roughly three times the size (150K).

To better evaluate the relationship between the LM training data (quantity and recency) and the results obtained with our adaptation techniques, we performed another additional experiment by extended both NP.train and BN.train datasets and compared the results measured on the two news shows of the BN.RTP-2007 test dataset when applying the baseline system and the proposed multi-pass adaptation frameworks with the four vocabulary sizes (30K, 57K, 100K and 150K words). Additionally, we present some statistics comparing the amount of memory and processing time allocated to the ASR decoding process in each one of the various experiments.

When available for the two test sets, the daily values of OOVs and WER were subject to statistical analysis. Due to the violation of the normality or homogeneity of variances assumptions in several groups, non-parametric Friedman ANOVA for repeated measures was used, followed by pair-wise comparisons (Wilcoxon test).

4.1. OOV word rate results

We started our evaluation by fixing our vocabulary size to 57K, the value used by the baseline ASR used for this research. The average values for BN.RTP-2007 and the corresponding relative improvements are presented in Fig. 11.

The proposed second-pass speech recognition approach (2-PASS-POS-IR) using the morpho-syntactic algorithm (POS) plus the lemmatization method for vocabulary adaptation and the Information Retrieval Engine (IR) for language model adaptation yields a relative reduction of 69% in OOV word rate, i.e., from 1.40% to 0.44%, when compared to the results obtained for the baseline system. Moreover, this approach outperformed the one based on one single-pass (1-PASS-POS).

After the evaluation of our two methods with the same vocabulary size as our baseline system, and to better understand the performance of this new adaptation procedure, we calculated and compared the OOV word

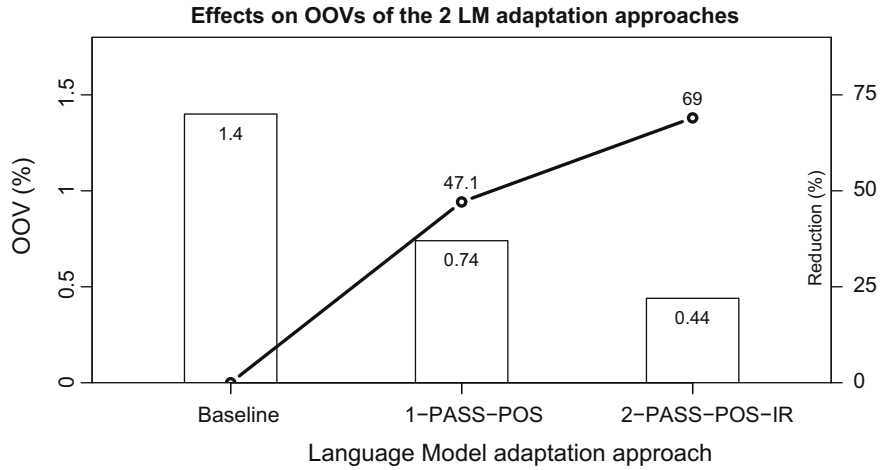

Fig. 11. Average OOV word rate (in%) measured for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed multi-pass LM adaptation frameworks (with a vocabulary size of 57K words).

rate results for four different vocabulary sizes (30K, 57K, 100K and 150K words) on the two evaluation datasets.

In Fig. 12 we present the average OOV word rate obtained from the daily values as a function of the evaluation dataset, method and vocabulary size. The graph shows the relatively good performance of 1-PASS-POS and 2-PASS-POS-IR approaches for the selection of large-sized vocabularies. We can observe that the proposed second-pass speech recognition approach (2-PASS-POS-IR) yields, for all vocabulary sizes and both test datasets, the lower OOV word rate. Moreover, this approach outperformed the one based on one single-pass (1-PASS-POS). The reduction of this metric with vocabulary size is also evident.

Furthermore, as we would expect, for the 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 obtain better lexical coverage than the one obtained for the baseline system with a vocabulary of 57K words. For both test datasets the best results were obtained for the 2-PASS-POS-IR and a vocabulary size of 150K, with a value of 0.17%. This OOV word rate corresponds to an improvement, relative to the 57K baseline, of 85.8%

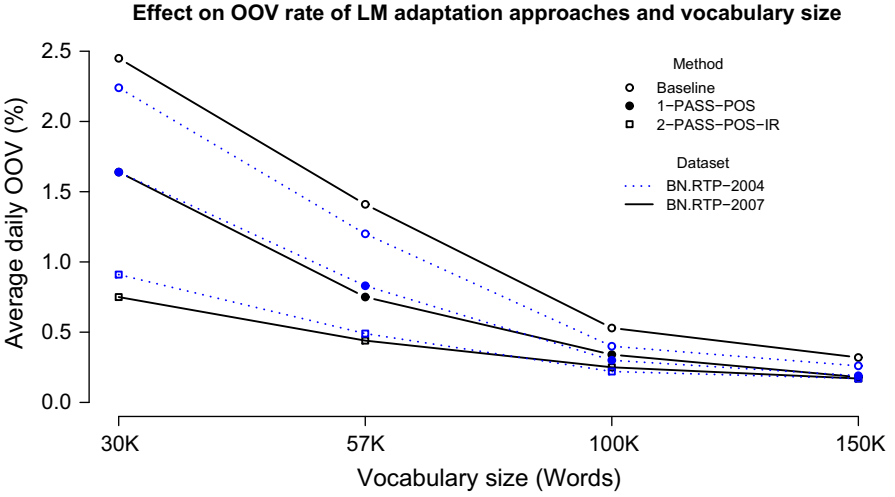

Fig. 12. Average OOV word rates (in%) measured for the two evaluation datasets (BN.RTP-2004 and BN.RTP-2007) when applying the baseline system and the proposed multi-pass adaptation frameworks with four different vocabulary sizes (30K, 57K, 100K and 150K words).

for the BN.RTP-2004 dataset (from 1.20% to 0.17%) and 87.9% for the BN.RTP-2007 dataset (from 1.40% to 0.17%). Even comparing the best results with the baseline for a 150K vocabulary (0.26% for BN.RTP-2004 and 0.32% for BN.RTP-2007) the improvement is at least 34.6%. The comparison of results obtained for both datasets allow us to conclude that even when the evaluation dataset is closer to the LM training corpora (BN.RTP-2004), the relative improvement is almost the same, showing that these adaptation approaches are useful even in that case. However, as one can observe, there is an inversion between the behavior of BN.RTP-2004 and BN.RTP-2007 when using the proposed adaptation techniques: for the baseline there are more OOV in the 2007 dataset, but for the first-pass adaptation, there is no difference, and by using the second-pass adaptation, there are even more OOV in the 2004 dataset. In fact, due to the gap between the 2007 dataset and the adaptation data, the proposed adaptation procedures were shown to perform well, in particular the second-pass adaptation using the IR framework for smaller vocabularies. For larger vocabularies, there is almost no gain between the first and the second-pass. In this case, the remaining OOV words are specific names or verbal lemmas not contained in the training/adaptation data, and the IR procedure is not able to recover them at all.

The graph of Fig. 13 shows the proportion of verbs in OOVs when we restrict or extend the vocabulary size and apply the proposed adaptation frameworks. As we can observe, for both evaluation datasets, there is a decrease in the proportion of verbs in OOVs for both adaptation approaches, with an average relative reduction of 57% for the 1-PASS-POS approach and 79% for the 2-PASS-POS-IR approach when compared to the baseline. In fact, the decrease obtained with the 2-PASS-POS-IR comprises the reduction due to the POS-based technique, applied in the first-pass, plus the IR and lemmatization processes, applied to each topic in the second-pass. We also see that the effect of the second-pass on the reduction of OOV verbs is larger on the BN.RTP-2004 dataset, while for the baseline and first-pass, the two evaluation datasets seem to behave identically. Observing the OOV verbs recovered, we could conclude this difference is mainly due to the fact that more verbal inflections were recovered by applying the IR procedure over the 2004 dataset, which can be justified by the fact that these BN shows comprise more street interviews with a large amount of spontaneous speech.

Moreover, by analyzing the OOV reduction by POS-classes, one can observe a similar behavior for both adaptation approaches: an average reduction of 31% for OOV names and 57% for OOV verbs when applying the first-pass, and an average reduction of 35% for OOV names and 52% for OOV verbs when applying the second-pass over the first one. This means that both approaches contribute in the same way to the type of OOV recovered.

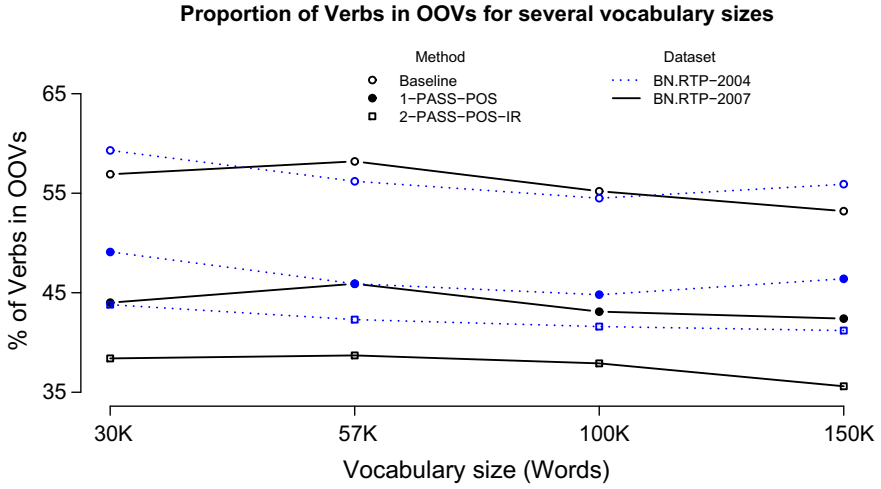

Fig. 13. Proportion of verbs in OOVs measured for the two evaluation datasets (BN.RTP-2004 and BN.RTP-2007) when applying the baseline system and the proposed multi-pass adaptation frameworks with four different vocabulary sizes (30K, 57K, 100K and 150K words).

Finally, to evaluate the benefit of our adaptation approaches even in the case of adding more LM training data, we extended both training corpora (NP.train and BN.train) with more data. In the case of NP.train, we added newspaper texts collected from 2004 until the end of 2006, going from the previous 604M words to approximately 740M words (NP.train extended dataset). In terms of broadcast news data, we added 5 h of BN in-domain texts to the previous 47 h (BN.train extended dataset).

The results of training with the extended training dataset and keeping the two factors (vocabulary size and adaptation method) constant are presented in Fig. 14. For comparison, the average values presented in previous figures are included. Results were only obtained for BN.RTP-2007 because the training material is more recent than our other evaluation dataset (BN.RTP-2004). From the graph we can conclude that the addition of more training data mainly produced better results in the case of the baseline vocabulary. In fact, for the proposed adaptation approaches there were only slight improvements comparing to the previous results. However, for all the vocabulary sizes, both the 1-PASS-POS and 2-PASS-POS-IR methods still produced lower OOV word rates than the baseline ones.

The Friedman test confirms as significant the effect of method [χ^2 (2) = 18.00, $p < 0.001$] and vocabulary size [χ^2 (3) = 27.00, $p < 0.001$]. The effect of test dataset did not prove to be significant ($p = 0.698$).

Regarding method, the pair-wise application of the Wilcoxon test with Bonferroni correction shows that OOV word rates for 2-PASS-POS-IR are significantly lower than for the 1-PASS-POS only method, ($p < 0.001$), and OOVs for 1-PASS-POS only are significantly lower than for the baseline ($p < 0.001$). Even with the reduced number of days on the test datasets and the variation in OOV word rates, the non-parametric test was capable of confirming that the 2-PASS-POS-IR method was significantly better than 1-PASS-POS and baseline methods.

For vocabulary size, the pair-wise application of the Wilcoxon test with Bonferroni correction showed a significant WER decrease between each pair of consecutive vocabulary sizes ($p < 0.001$).

4.2. WER results

In terms of WER, the average values for BN.RTP-2007 and the corresponding relative improvements are presented in Fig. 15. As we can observe, the new approach (1-PASS-POS) resulted in an 8% relative gain, from 19.9% to 18.25%, for a vocabulary size of 57K words. Moreover, the proposed second-pass approach yielded a relative reduction of 12% in WER when compared to the WER obtained for the baseline system, outperforming the 1-PASS-POS approach (a slight decrease in WER, from 18.25% to 17.55%).

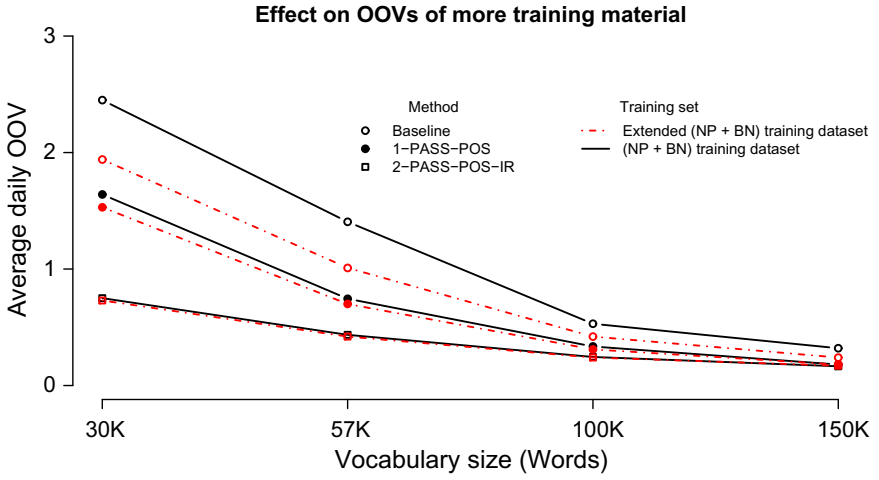

Fig. 14. Comparison of average OOV word rates (in%) measured for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed multi-pass adaptation frameworks with four different vocabulary sizes (30K, 57K, 100K and 150K words) and extending the training corpora (NP.train and BN.train) used for vocabulary and LM estimation.

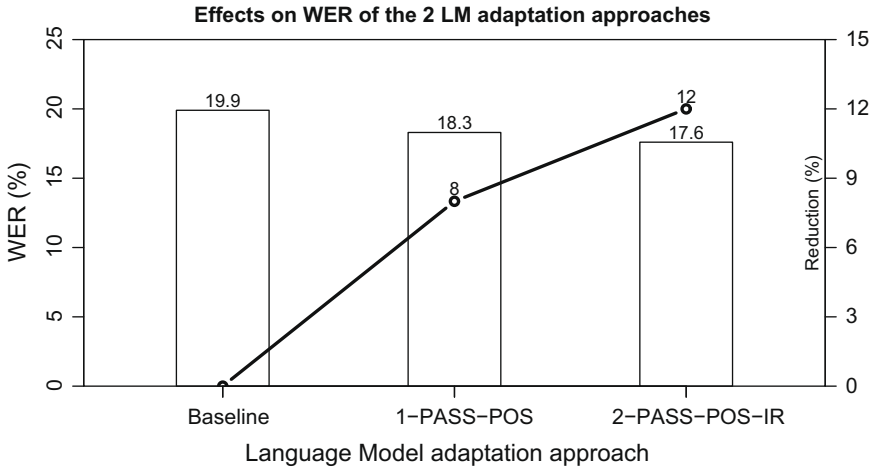

Fig. 15. Average WER measured for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed multi-pass adaptation frameworks (with a vocabulary size of 57K words).

To better evaluate the accuracy of our approach we performed a more detailed analysis of the WER obtained by the 2-PASS-POS-IR approach with a vocabulary of 57K words. For that analysis, we divided the adapted vocabulary V_S of each story S into two sets: the set of word types that were already present in the baseline vocabulary V_0 and the set of all new word types. From this last set (denoted by N_s), we removed all the word types except the ones occurring in the reference transcripts of the tested BN dataset (BN.RTP-2007). The number of word types in N_s was 105, with 182 occurrences in the reference transcripts. From these 182 occurrences, 133 were correctly recognized by the 2-PASS-POS-IR approach, which means 73.1% of new words found by our IR-based framework were correctly recognized.

In Table 4 we present the distribution of those 182 occurrences by grammatical category. In the "Names" category we generically included both proper and common names, even the foreign ones. The "Others" category included other foreign words, acronyms and abbreviations. As one can observe, more than 66% of those new words found by our algorithm belong to the verbs class. Moreover, the class of names is the one with the best recognition rate (80.2% of new names were correctly recognized), slightly outperforming the average value (73.1%). This shows that a significant number of relevant terms like proper and common names (including names of persons, locations and organizations) were correctly recognized, making the framework especially useful for novel applications like information dissemination, where those types of words contain a great deal of information.

In Fig. 16, we compared the accuracy of our two methods and the baseline for four different vocabulary sizes. One can observe that WER was reduced as the vocabulary size increased and was, for all vocabulary sizes, better with the 2-PASS-POS-IR method.

For the BN.RTP-2004 dataset, with the proposed multi-pass adaptation approach and increasing the vocabulary size to 150K words, we could obtain a relative gain of 10.4% in terms of WER, with a final WER of 19.81% against the 22.10% mean for the baseline system. For the BN.RTP-2007 test dataset, the corresponding gain was 13.1%, from 19.9% to 17.3%. Even using a vocabulary with only 30K words, we were able

Table 4
Distribution (in%) of new words by grammatical category and the percentage of them correctly recognized by the 2-PASS-POS-IR approach (a vocabulary size of 57K words).

POS	% of Occurrences	% Correctly recognized
Names	30.8	80.2
Adjectives	1.6	66.7
Verbs	66.5	58.9
Others	1.1	50.0

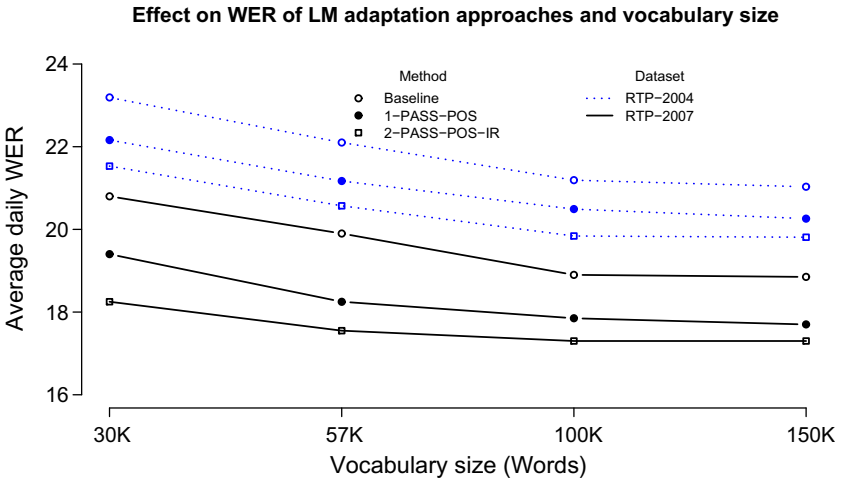

Fig. 16. Average WER measured for the two evaluation datasets (BN.RTP-2004 and BN.RTP-2007) when applying the baseline system and the proposed multi-pass adaptation frameworks with four different vocabulary sizes (30K, 57K, 100K and 150K words).

to get a better WER for both test datasets with our adaptation framework than the one obtained for the baseline system with a 57K word vocabulary.

Finally, in Fig. 17 we present the comparison of the average WER measured for the two news shows of the BN.RTP-2007 evaluation dataset when applying the baseline system and the proposed multi-pass adaptation frameworks with different vocabulary sizes and extending the training corpora. In this case, both the baseline and 1-PASS-POS produced similar improvements in terms of WER. For the 2-PASS-POS-IR only slight improvements were obtained, mainly in the case of smaller vocabularies. However, both approaches produced better WER results even when more recent data was added to the LM training corpora.

Separate application of Friedman's test showed that there were some statistically significant changes in the distribution of WERs over the three methods [χ^2 (2) = 18.00, $p < 0.001$] and over the four vocabulary sizes [χ^2 (3) = 24.71, $p < 0.001$]. A pair-wise application of the Wilcoxon test with Bonferroni correction levels showed that the WER for 2-PASS-POS-IR was significantly lower than for 1PASS-POS only method, ($p < 0.001$), and the WER for 1-PASS-POS only was significantly lower than for the baseline ($p < 0.001$). A similar application

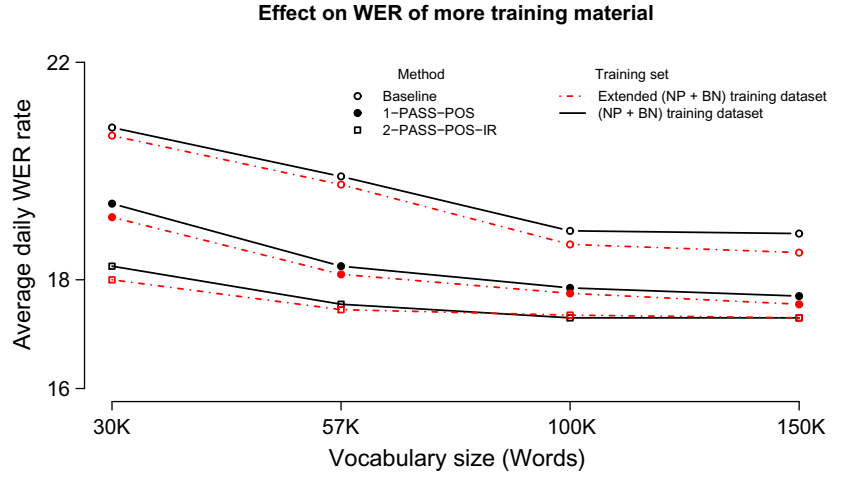

Fig. 17. Comparison of average WER measured for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed multi-pass adaptation frameworks with four different vocabulary sizes (30K, 57K, 100K and 150K words) and extending the training corpora (NP.train and BN.train) used for vocabulary and LM estimation.

of Wilcoxon test with correction for the vocabulary size factor showed that each vocabulary size had a WER significantly lower than that of all smaller sizes ($p < 0.001$).

4.3. Computational costs

To compare the baseline system and the proposed adaptation frameworks in terms of computational costs, we used the amount of memory and processing time allocated to the ASR decoding process. The averages presented were calculated for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed online adaptation framework (1-PASS-POS) for the four different vocabulary sizes and using the extended training corpora. Because computational costs are critical for our online task – the live and real-time subtitling of RTP news shows – we only compared the baseline with the 1-PASS-POS approach.

Memory requirements (in GBytes) and CPU time (in seconds) were measured on an Intel(R) Core(TM) 2 Quad CPU Q6600 @ 2.40 GHz equipped with 8 Gb of RAM. Results are presented in Fig. 18. We can observe that for the same vocabulary size, the two methods require similar memory and processing time, and both memory and CPU time increase with vocabulary size. This increase was more significant in the case of

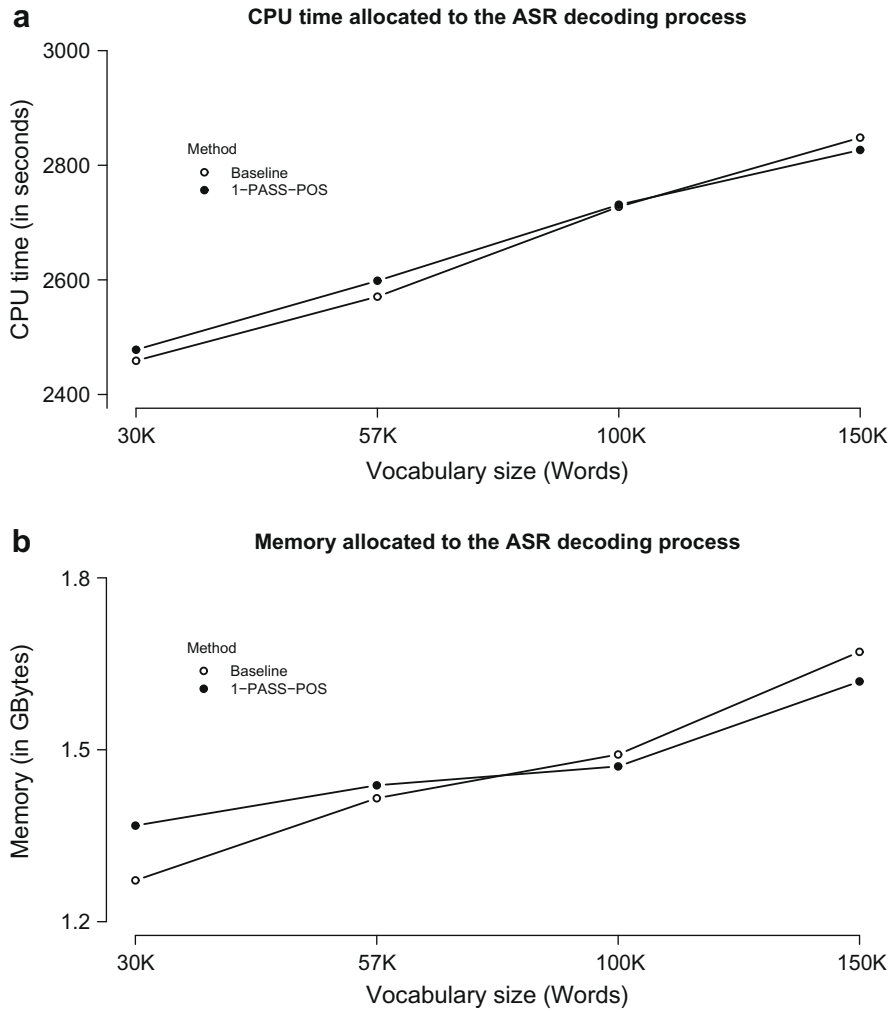

Fig. 18. CPU time (in seconds) and memory allocated to the ASR decoding process and measured for the two news shows of the BN.RTP-2007 dataset when applying the baseline system and the proposed 1-PASS-POS adaptation framework with four different vocabulary sizes and extending the training corpora.

CPU time. We can observe an increase of almost 10% when using a vocabulary size of 150K words instead of a 57K word vocabulary.

5. Discussion and future work

In this paper, we addressed the task of language modeling for the automatic transcription of European Portuguese BN speech, proposing a framework to dynamically address new words by adapting both the vocabulary and language model on a daily basis.

Analyzing the distribution of OOV words according to their grammatical properties in a given sentence (i.e., their part-of-speech), we found that more than 56% of them were verbs. This was an interesting result because from other findings published in the literature, OOV words are mostly names. This led us to study the use of specific linguistic knowledge (POS tagging) to improve the lexical coverage of a selected vocabulary.

In terms of LM adaptation, we proposed and implemented a multi-pass speech recognition approach that creates from scratch both vocabulary and LM components on a daily basis using adaptation texts extracted from the Internet and the new POS-based vocabulary selection algorithm. This framework was integrated into the fully functional system described in Section 2, which is being used to produce live captions for a closed-captioning system of live European Portuguese TV broadcasts. For the BN.RTP-2007 evaluation dataset, considering the same vocabulary size as the baseline one (57K words), a relative gain of 12% in the WER and a relative reduction of 69% in OOV word rate were observed. Moreover, by applying the proposed adaptation framework and increasing the vocabulary size to 150K words we obtained a relative gain of 13.1% in terms of WER with a relative OOV word rate reduction of 87.9%.

Taking into consideration the application framework we are using, the processing time is crucial. In fact, generating the live captions as quickly as possible and with negligible delay is a very important issue for us. Hence, improving the system accuracy without compromising its response speed is clearly beneficial for this kind of application. Combining results for WER and CPU time, the benefit of using our online adaptation approach (1-PASS-POS) is clear. Using the 1-PASS-POS method with a vocabulary of 57K words results in a relative decrease of 12.4% in terms of CPU time compared to the baseline system with a vocabulary of 150K without compromising system accuracy. Even better, 1-PASS-POS for 57K words presents a slight better accuracy than the 150K baseline (18.1% versus 18.5%). The accuracy gain of the proposed method allows us to use more reduced vocabulary sizes to comply with the live application demands. In fact, even if we can take advantage of more powerful machines both in terms of memory and CPU capacity, in our opinion it is useful to restrict the vocabulary size for the application type we are using. Using large-sized vocabularies may be desirable from the point of view of lexical coverage. However, there is always the additional problem of increased acoustic confusability. Moreover, as we reported in Section 2.2, each BN show comprises an average of 8300 word tokens and only 2200 word types. Thus, the majority of the vocabulary words are irrelevant when adapting to a single BN show. Therefore, future research should focus on methods to better constrain vocabulary growth while preserving adaptation performance.

All the work described on this paper was motivated and influenced by the finding that verbs constitute the largest portion of OOV. Hence, both vocabulary optimization and language model adaptation approaches were based on the integration of different knowledge sources and techniques (language modeling, information retrieval and morphological knowledge). However, even if we were able to improve the overall system performance, we think it is worthwhile to investigate it further in future work, especially to find solutions that can directly solve the verb OOV problem.

While our focus was on European Portuguese broadcast news, where morphological variants such as inflectional verb endings have been shown to be an important problem to address, we believe the framework proposed here would likely lead to improved performance for other inflectional languages and/or applications, especially the POS-based vocabulary selection procedure.

In terms of future work, there are other directions that, in our opinion, can be investigated to enhance the global performance of the language model component. Considering the vocabulary selection, we believe that better results can be achieved by exploring more deeply the linguistic knowledge of the in-domain corpus. We could use not only the grammatical property of words (POS), but also their morphological information (gender, number, conjugation, etc.). To follow this research trend, two resource constraints must be overcome:

more in-domain data and a morphological analyzer for European Portuguese that could give us that level of morphological information with an acceptable accuracy. Moreover, we can extend the effectiveness of our multi-pass adaptation framework by using the IR techniques in the first-pass to select and cluster data, reducing its redundancy and improving the generic language model estimation. In fact, entropy-based clustering and data selection have been used with significant gains for defining topic-specific subsets and pruning less useful documents from training data, both for acoustic and language model adaptation (Hwang et al., 2007; Ramabhadran et al., 2007; Wu et al., 2007).

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 was sponsored by an FCT scholarship (SFRH/BD/23360/2005). Special thanks to our colleague Hugo Meinedo for his efforts to make available all the manual transcriptions for the evaluation datasets we used. We also thank the 3 annotators involved in IR evaluation.

References

Allauzen, A., Gauvain, J., 2005. Open vocabulary ASR for audiovisual document indexation. In: Proceedings of ICASSP, 2005.

Allauzen, A., Gauvain, J., 2005a. Diachronic vocabulary adaptation for broadcast news transcription. In: Proceedings of Interspeech, 2005.

Amaral, R., Meinedo, H., Caseiro, D., Trancoso, I., Neto, J., 2006. Automatic vs. manual topic segmentation and indexation in broadcast news. In: IV Jornadas en Tecnologia del Habla, November 2006, pp. 123–128.

Bazzi, I., 2002. Modeling out-of-vocabulary words for robust speech recognition. Ph.D. Thesis, Massachusetts Institute of Technology.

Bellegarda, J., 2004. Statistical language model adaptation: review and perspectives. Speech Communication 42.

Bigi, B., Huang, Y., Mori, R., 2004. Vocabulary and language model adaptation using information retrieval. In: Proceedings of ICSLP, 2004.

Blei, A., Jordan, M., 2003. Latent dirichlet allocation. Journal of Machine Learning Research.

Boulianne, G., Beaumont, J., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., et al., 2006. Computer-assisted closedcaptioning of live TV broadcast in French. In: Proceedings of Interspeech, 2006.

Caseiro, D., 2003. Finite-state methods in automatic speech recognition. Ph.D. Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Caseiro, D., Trancoso, I., Oliveira, L., Ribeiro, M., 2002. Grapheme-to-phone using finite-state transducers. In: 2002 IEEE Workshop on Speech Synthesis.

Chen, L., Gauvain, J., Lamel, L., Adda, G., 2004. Dynamic language modeling for broadcast news. In: Proceedings of ICSLP, 2004.

Federico, M., Bertoldi, N., 2004. Broadcast news LM adaptation over time. Computer Speech and Language 18 (4), 417–435.

Federico, M., Giordani, D., Coletti, P., 2000. Development and evaluation of an Italian broadcast news corpus. In: Proc. LREC, Athens, Greece, 2000.

Gauvain, J., Lamel, L., Adda, G., 2002. The LIMSI broadcast news transcription system. Speech Communication 37, 89–108.

Geutner, P., Finke, M., Sheytt, P., Waibel, A., Wactlar, H., 1998. Transcribing multilingual broadcast news using hypothesis driven lexical adaptation. In: Proceedings of ICASSP, 1998.

Hetherington, I., 1995. A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding. Ph.D. Thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science.

Hwang, M., Peng, G., Wang, W., Faria, A., Heidel, A., Ostendorf, M., 2007. Building a highly accurate mandarin speech recognizer. In: Proceedings of ASRU, 2007.

Iyer, R., Ostendorf, M., 1997. Transforming out-of-domain estimates to improve in-domain language models. In: Proceedings of Eurospeech, 1997.

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., Stolcke, A., 2006. Morphology-based language modeling for conversational Arabic speech recognition. Computer Speech and Language 20 (4), 589–608.

Lamel, L., Gauvain, J., Adda, G., Adda-Decker, M., Canseco, L., Chen, L., et al., 2004. Speech transcription in multiple languages. In: Proceedings of ICASSP, 2004.

Lavrenko, V., Croft, W., 2001. Relevance-based language models. In: Proceedings of SIGIR'01, 2001.

Lecorvé, G., Gravier, G., Sébillot, P., 2008. An unsupervised web-based topic language model adaptation method. In: Proceedings of ICASSP, 2008.

Lecorvé, G., Gravier, G., Sébillot, P., 2009. Constraint selection for topic-based MDI adaptation of language models. In: Proceedings of InterSpeech, 2009.

Martins, C., 1998. Modelos de Linguagem no reconhecimento de Fala Contínua. Master's Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Martins, C., Teixeira, A., Neto, J., 2005. Language models in automatic speech recognition. Revista Electrónica e Telecomunicações, Departamento de Electrónica e Telecomunicações, Universidade de Aveiro, Aveiro, vol. 4, no. 4, 2005.

Martins, C., Teixeira, A., Neto, J., 2006. Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system. In: Proceedings of IEEE/ACL Workshop on Spoken Language Technology, 2006.

Martins, C., Teixeira, A., Neto, J., 2007a. Vocabulary selection for a broadcast news transcription system using a morpho-syntactic approach. In: Proceedings of Interspeech, 2007.

Martins, C., Teixeira, A., Neto, J., 2007b. Dynamic language modeling for a daily broadcast news transcription system. In: Proceedings of ASRU, 2007.

Martins, C., Teixeira, A., Neto, J., 2008. Dynamic language modeling for the European Portuguese. In: Proceedings of PROPOR 2008, Curia, Portugal.

Meinedo, H., 2008. Audio pre-processing and speech recognition for broadcast news. Ph.D. Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Meinedo, H., Neto, J., 2000. Combination of acoustic models in continuous speech recognition hybrid systems. In: Proceedings of ICSLP 2000, China.

Meinedo, H., Neto, J., 2005. A stream-based audio segmentation, classification and clustering preprocessing system for broadcast news using ANN models. In: Proceedings of Interspeech, 2005.

Meinedo, H., Caseiro, D., Neto, J., Trancoso, I., 2003. AUDIMUS.MEDIA: a broadcast news speech recognition system for the European Portuguese language. In: Proceedings of PROPOR 2003, Portugal.

Neto, J., Meinedo, H., Amaral, R., Trancoso, I., 2003. The development of an automatic system for selective dissemination of multimedia information. In: Proceedings of Third International Workshop on Content-based Multimedia Indexing – CBMI 2003.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D., 2008. Broadcast news subtitling system in Portuguese. In: Proceedings of ICASSP, 2008.

Ney, H., Martin, S., Wessel, F., 1997. Statistical language modeling using leaving-one-out. In: Young, S., Bloothooft, G. (Eds.), Corpus-based Methods in Language and Speech Processing. Kluwer Academic Publishers, Dordrecht, pp. 174–207 (Chapter 6).

Oger, S., Linares, G., Bechet, F., Nocera, P., 2008. On-demand new word learning using world wide web. In: Proceedings of ICASSP, 2008.

Orengo, V., Huyck, C., 2001. A stemming algorithm for the Portuguese language. In: Proceedings of the Eighth International Symposium on String Processing and Information Retrieval, 2001.

Palmer, D., Ostendorf, M., 2005. Improving out-of-vocabulary name resolution. Computer Speech and Language 19, 107–128.

Ramabhadran, B., Siohan, O., Sethy, A., 2007. The IBM 2007 speech transcription system for European parliamentary speeches. In: Proceedings of ASRU 2007.

Ribeiro, R., Oliveira, L., Trancoso, I., 2003. Using morphossyntactic information in TTS systems: comparing strategies for European PortugueseComputational Processing of the Portuguese Language. Lecture Notes in Computer Science (Subseries LNAI), vol. 2721. Springer, pp. 143–150.

Ribeiro, R., Mamede, N., Trancoso, I., 2004. Morpho-syntactic Tagging: a Case Study of Linguistic Resources ReuseLanguage Technology for Portuguese: Shallow Processing Tools and Resources. Edições Colibri, Lisbon, Portugal.

Stolcke, A., 1998. Entropy-based pruning of backoff language models. In: Proceedings of DARPA News Transcription and Understanding Workshop, 1998.

Strohman, T., Metzler, D., Turtle, H., Croft, W.B., 2005. Indri: a language-model based search engine for complex queries (extended version). CIIR Technical Report, 2005.

Tam, Y., Schultz, T., 2006. Unsupervised language model adaptation using latent semantic marginals. In: Proceedings of Interspeech, 2006.

Venkataraman, A., Wang, W., 2003. Techniques for effective vocabulary selection. In: Proceedings of Eurospeech, 2003.

Wang, W., Stolcke, A., 2007. Integrating MAP, marginals, and unsupervised language model adaptation. In: Proceedings of Interspeech, 2007.

Wu, Y., Zhang, R., Rudnicky, A., 2007. Data selection for speech recognition. In: Proceedings of ASRU, 2007.

Xu, P., Jelinek, F., 2007. Random forests and the data sparseness problem in language modeling. Computer Speech and Language 21 (1), 105–152.