Automatic Estimation of Language Model parameters for unseen Words using Morpho-syntactic Contextual Information

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

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

Abstract

Various information sources naturally contains new words that appear in a daily basis and which are not present in the vocabulary of the speech recognition system but are important for applications such as closed-captioning or information dissemination. To be recognized, those words need to be included in the vocabulary and the language model (LM) parameters updated. In this context, we propose a new method that allows including new words in the vocabulary even if no well suited training data is available, as is the case of archived documents, and without the need of LM retraining. It uses morpho-syntatic information about an in-domain corpus and part-of-speech word classes to define a new LM unigram distribution associated to the updated vocabulary.

Experiments were carried out for a European Portuguese Broadcast News transcription system. Results showed a relative reduction of 4% in word error rate, with 78% of the occurrences of those newly included words being correctly recognized.

Index Terms: morpho-syntactic analysis, POS tags, class-based language models, broadcast news, transcription systems

1. Introduction

Broadcast news (BN) naturally contains new words that appear in a daily basis and which are not present in the vocabulary of the speech recognition system but are important for applications such as closed-captioning or information dissemination. This way, lexical coverage of a vocabulary for this kind of applications should be as high as possible to minimize the side effects of out-of-vocabulary (OOV) words on system recognition performance. For these applications, it is important for the automatic speech recognition (ASR) component to recognize the new words accurately, since most of those words are names of people, location, organization, or other entities which play an important role in the final system performance.

In recent years, researchers are using the Word Wide Web (WWW) as an additional resource of training data for automatic vocabulary and language model adaptation procedures [1]. In [2][3] the authors implemented a rolling language model with an updated vocabulary by removing out-of-date words and adding new words found in newspapers articles to update the ASR component of broadcast transcription systems on a daily basis. In [4], we proposed a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and language model to the topic of the current news segment during a multi-pass speech recognition process, based on texts daily available on the Web.

But, language model adaptation procedures proposed on those works assume that well suited sources of data are available to estimate the language model parameters. However, sometimes we would like manually add new words to the system vocabulary which are likely to appear on certain broadcast shows, even if no well suited data is available at all, as is the case of archived BN documents [5], or just a small amount of data is available but not sufficient to apply those language model adaptation procedures. Thus, in this situation estimating the language model parameters for those words is problematic. For example, small amount of data like the anchor working scripts and other prior knowledge information, such as the speakers' names and show summary, can be available and used to extract new words with high probability to appear during the BN show, but not sufficient to extract language model parameters.

Thus, in this paper, we propose a new method that allows including new words in the system vocabulary without the need of additional adaptation data or language model retraining. This method uses morpho-syntatic information about an in-domain corpus and part-of-speech (POS) word classes to define a new language model unigram distribution associated to the updated system vocabulary.

The proposed framework applied to a European Portuguese Broadcast News transcription system showed to be effective in terms of recognition accuracy improvement (WER), with a relative reduction of 4%. Moreover, with this method we were able to correctly recognize about 78% of the occurrences of the newly introduced words.

In section 2 we briefly describe the baseline system and data sources used in our experiments. Section 3 provides a description of the proposed method for inclusion of new words in the system vocabulary and section 4 presents the experimental results, drawing in section 5 some conclusions and future research trends.

2. Baseline ASR system and datasets

2.1. Baseline system

For the work presented in this paper, we used the system reported in [6]. 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 since March 2003.

This system features a hybrid HMM/MLP system, using three MLPs, each of them associated with a different feature extraction process, where the MLPs are used to estimate the context independent posterior phone probabilities given the acoustic data at each frame. The phone probabilities generated at the output of the MLPs classifiers are combined using an appropriate algorithm [7]. The decoder used under this baseline system is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [8]. In this approach, the decoder search space is a large WFST that maps observation distributions to words.

The active system vocabulary and language model are dynamically adapt in a daily and unsupervised framework [4]. We defined a morpho-syntactic approach to select the target vocabulary by trading off between the OOV word rate and vocabulary size [9]. With this approach a 57K words vocabulary are selected based on two training corpora and texts daily available on the Web:

- NP-2003, a 604M word corpus of newspapers texts collected from the WWW since 1991 until the end of 2003 (out-of-domain dataset);
- BN-ALERT, a 531K word corpus corresponding to the broadcast news transcriptions of the acoustic training data (in-domain dataset).
- WEBNEWS, a corpus consisting of written news which are being collected from the online Portuguese newspapers Web editions in a daily basis. WEBNEWS corpus provides an average of 80K words per day. However, to construct a more homogeneous dataset to use on our daily adaptation framework, we merge the data from several consecutive days. Hence, for each day d, we use the texts from the current day and the 6 preceding days, which means, using one week of written news for daily adaptation proposes (a subset denoted as $O_7(d)$ - 7 days of Online written news).

The baseline language model combines a backoff 4-grams LM trained on NP-2003 corpus, a backoff 3-grams LM estimated on BN-ALERT corpus and a backoff 3-grams LM estimated on WEBNEWS corpus. The three models are combined by means of linear interpolation, generating a mixed model.

For each vocabulary, the lexical pronunciations are derived using a rule-based phonetizer.

This baseline system [4] is the state-of-the-art in terms of Broadcast News Transcription Systems for the European Portuguese language.

2.2. Evaluation dataset

To evaluate the proposed method we selected two broadcast news shows from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP, which have been daily recorded and automatically transcribed by the European Portuguese BN transcription system reported in [6]. The randomly selected shows had a total duration of about 2 hours of speech (about 16K words) and were collected on May 24th and 31st of 2007.

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

	May 24th	**May 31st**
Word tokens	8156	7744
Word types	2300	2295
%OOV tokens	0.82	0.77
%OOV types	2.48	2.22

Table 1 shows the statistics related to this evaluation corpus (RTP-07) with the OOV word rate for the baseline vocabulary of 57K words daily selected according to the adaptation framework briefly described in section 2.1. The results are shown for both word tokens, in which all occurrences of a word are counted, and word types, in which only unique words are counted. From our previous work [4], we could conclude that verbs make up for the largest portion of OOV words, accounting for almost 56% of them.

3. Proposed Method

From an ASR system point of view, adding a new word to its vocabulary implies the following tasks: deriving the possible phonetic transcription(s) associated to that word, and estimating its n-grams distributions within the language model.

Usually, the first task is accomplished by a rule-based phonetizer that automatically derives one or more lexical pronunciations using grapheme-to-phoneme rules. However, estimating the language model parameters for new words is more problematic, especially in cases where no data or insufficient relevant training data is available. As far as no additional training data is available, a new word is no more than an unseen event, which implies estimating n-grams distributions related to unseen words. In a standard approach, various classical smoothing techniques [10] exist which can be applied during language model parameters estimation. But, they treat unseen words in same way, not taking in consideration theirs types or linguistic roles.

In [11] special forms called back-off word classes (BOW) were used to introduce a word in the vocabulary without retraining the language model. Thus, during language model training one of these forms replaces one or more words which are not yet known, by discounting a mass of probability from the OOV words. Then, prior to decoding, new words can be added as alternate orthographic forms of these special classes. Words are linked with their lexical BOW according to their POS tag. However, the estimation of the probability of each word inside its BOW class relies on some additional adaptation data.

In this work, we propose a similar approach that allows including new words in the system vocabulary without the need of additional adaptation data or language model retraining. The basic idea is that if no training data is available for the new word, then we will take advantage of morpho-syntactic information related to words which have similar properties in terms of language modeling.

3.1. Updating unigram probabilities

For a standard back-off language model, the n-gram probabilities $P(w_0|h)$ related to unseen words w_0 and given the word history h are derived in a same way, using the unigram estimation $P(w_0)$. However, as no contextual information is available, classical smoothing techniques treat all those w_0 words in a similar basis. Thus, we propose to use classes of words as an alternative to better estimate those unigram probabilities. An additional advantage of classes is that we can gather statistics on the frequency of occurrence of words similar to the unseen ones. The idea is to build a unigram model that uses grammatical information to give a probability to words according to some predefined notion of similarity.

A class-based unigram model is used to implement this idea, where the classes are the parts-of-speech (POS). Therefore, a morpho-syntactic analyzer developed for the European Portuguese [12] is used to tag all the vocabulary words with their complete morpho-syntactic information. For example, for the Portuguese word “fala” (speech) that information consists of the following five possible tags: “Nc...sf...”, “V.ip3s=...”, “V.sp1s=...”, “V.sp3s=...” and “V.m=2s==..”, respectively referring to the feminine singular common noun, and four different flexions of the verb “falar” (to speak). Since keeping all type of morpho-syntactic information would result in too many tags and the training data would be insufficient, we focus only on the syntactic category of the tag to map words in their corresponding classes. In table 2 we list the final tag set used in this work and consisting of 11 grammatical categories. The “Others” category includes foreign words, abbreviations, acronyms and symbols. Hence, the word “fala” in our example, is classified in two different classes: class of nouns (N) and class of verbs (V).

Table 2. *Part-of-Speech for European Portuguese and their corresponding grammatical categories.*

Category	POS	Category	POS
Nouns	N	Prepositions	S
Adjectives	A	Conjunctions	C
Verbs	V	Numerals	M
Pronouns	P	Interjections	I
Articles	T	Others	X
Adverbs	R		

In the context of a class-based language model, an unseen word can be affected to one or more of these POS classes in order to inherit the contextual properties of the words belonging to these same classes. Thus, in this framework the unigram probabilities $P(w)$ are re-estimated as

$$P(w) = \sum_{c_i \in C(w)} P(w|c_i) P(c_i) \qquad (1)$$

where $C(w)$ represents the set of POS classes c_i assigned to word w. Therefore, after defining $C(w)$ for all the vocabulary words, the corresponding unigram distribution needs to be re-estimated. The next section describes the proposed method for its estimation that allows assigning non-zero probabilities for unseen words.

3.2. Parameters Estimation

In (1) the emission probability of a word given its class $P(w|c_i)$ and the class probability $P(c_i)$ are both computed through the maximum likelihood estimation (MLE) approach. For $P(c_i)$ estimation, only the in-domain dataset (BN-ALERT) was used. This decision was based on findings of our previous work presented in [9]. Analyzing both in-domain corpus (BN-ALERT) and out-of-domain corpus in terms of POS sequences, a significant difference was found, specially in terms of nouns and verbs. Hence, the BN-ALERT corpus was POS-tagged using a morpho-syntactic ambiguity resolver [13] which gives the POS of a word in its context. Table 3 presents an example of a tagged sentence.

Table 3. *Example of a POS-tagged sentence.*

	Sentence
original text	tenha um bom fim de semana
tagged text	tenha/**V** um/**T** bom/**A** fim/**N** de/**S** semana/**N**

The statistics of occurrence of POS classes in this in-domain corpus were then used to estimate $P(c_i)$, for $i = 1,\ldots,11$. In this estimation we used the Kneser-Ney discounting method [10] for smoothing proposes and to take into account unseen pairs of POS classes.

In addition to the BN-ALERT dataset, a sub-corpus of NP-2003 and the $O_7(d)$ subset of WEBNEWS were also POS-tagged, and their statistics used for maximum likelihood estimation of the emission probability of a word given its class as $P(w|c_i) = \dfrac{N(w/c_i)}{N(c_i)}$, with $N(w/c_i)$ being the count of occurrences of word w in the context of c_i class. However, the most problematic task is to estimate this probability distribution for new words since we assume that no additional data is available for training. To overcome this problem, we derived a heuristic approach to affect non-zero probabilities for those words by using the morpho-syntatic information of each word. Thus, considering w_0 as a unseen word to be introduced in system vocabulary, $\mathrm{M}(w_0)$ as its complete morpho-syntatic information, and $S = \{w : \mathrm{M}(w) = \mathrm{M}(w_0)\}$ as the set of all vocabulary words sharing the same morpho-syntatic information as w_0, then we define

$$N(w_0/c_i) = \max_{w \in S}\left(N(w/c_i)\right) \qquad (2)$$

Choosing the max function we heuristically assign a probabilistic mass to new words within classes, turning those words as probable as the ones with the highest probabilistic value in the same contextual position.

Finally, for the estimation of $P(w|c_i)$ we applied the Laplace smoothing method [10], which showed to be effective in our framework.

4. Evaluation Results

To evaluate and compare the performance of our framework, experimental results are reported according to two evaluation metrics: the word error rate (WER) and the percentage of new words introduced in the vocabulary and correctly recognized. As a preliminary evaluation, we performed an oracle experiment, i.e., considering the baseline vocabulary of 57K words, all the OOV words in the manual orthographic transcripts of the evaluation dataset were added to the vocabulary in a daily basis. An average of 55 new words were automatically added for each show.

The WER results are summarized in figure 1. As a reference, we present the WER obtained for the baseline system with a vocabulary size of 57K words (Baseline). As one can observe, applying the proposed LM updating framework (POS-based) for addition of new words to the baseline vocabulary, yields a relative reduction of 6.3% in terms of WER, from 19.0% to 17.8%. Moreover, this new approach clearly outperformed the standard one (Standard).

In this standard approach, the vocabulary was extended and the LM re-estimated in a standard way, using the absolute discounting technique for smoothing proposes. Thus, applying the POS-based approach we could get an improvement of about 4% over the standard one.

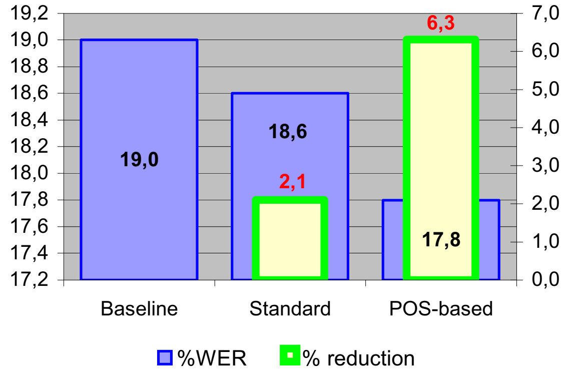

Figure 1: *WER results obtained for the baseline system and with a new vocabulary applying two different LM strategies: standard and POS-based.*

The relative percentage of new words introduced in the vocabulary and correctly recognized, is another important metric to measure the performance of the proposed framework. In table 4 we present these statistics, by evaluation show, for both LM updating strategies. While only 31.8% of new words were correctly recognized when applying the standard LM approach, a significant improvement has been observed when we used the POS-based one, with 78% of those words being correctly recognized. This result showed us the effectiveness of this new approach for LM re-estimation when new words need to be added to the ASR system in an easy and automatic way, even if no adaptation data is available for that.

Table 4. *Percentage of new words correctly recognized with both LM updating strategies: standard and POS-based.*

	May 24th	May 31st	average
Standard	25.9	37.5	31.8
POS-based	72.2	83.5	78.0

After analyzing the ASR results we could observe that some new words were wrongly recognized since their automatic transcriptions were not correctly derived. This occurs mainly in case of foreign words, whose automatic transcriptions are not reliable. Moreover, 62.5% of those wrongly recognized words were verbs. This last observation suggests us that special focus should be given to this syntactic category of words in our future research trends.

5. Conclusions and Future Work

In this paper we presented a method for including new words in the system vocabulary without the need of additional adaptation data or LM retraining. We proposed a heuristic approach to assign non-zero probabilities for those words by using the morpho-syntactic information of each word and POS classes. This way, the LM unigram distribution associated to the updated system vocabulary is re-estimated.

An oracle recognition experiment was carried out for a daily European Portuguese BN transcription task using two BN shows to evaluate and compare the performance of the proposed approach. It showed to be effective in terms of recognition accuracy improvement (WER), with a relative reduction of 4% when compared to a standard LM updating approach. Moreover, with this method we were able to correctly recognize 78% of the occurrences of the newly introduced words.

To extend the effectiveness of this LM updating framework, we will investigate its application as a smoothing method for n-grams of higher order. Moreover, we will try to redefine the word classes using more details from the morpho-syntactic information available for each word, with special focus on verbs.

6. Acknowledgements

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). Special thanks to our colleague Hugo Meinedo for his efforts to make available all the manual transcriptions for the evaluation dataset we used.

7. References

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

[2] Federico, M. and Bertoldi, N. (2004). Broadcast news LM adaptation over time. Computer Speech and Language, vol. 18, 2004.

[3] Ohtsuki, K. and Nguyen, L. (2007). Incremental Language Modeling For Broadcast News. In Proc. of ICASSP, 2007.

[4] Martins, C., Teixeira, A., and Neto, J. (2007). Dynamic Language Modeling for a daily Broadcast News Transcription System. In Proc. of ASRU, 2007.

[5] Barras, C., Allauzen, C., Lamel. L. and Gauvain, J. (2002). Transcribing audio-video archives. In Proc. of ICASSP, 2002.

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

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

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

[9] 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.

[10] Chen, S. and Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Computer Science Group - Harvard University, Cambridge, Massachusetts TR-10-98, 1998.

[11] Allauzen , A. and Gauvain, J. (2005). Open Vocabulary ASR for Audiovisual Document Indexation. In Proc. of ICASSP, 2005.

[12] 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.

[13] Ribeiro, R. (2003). Anotação morfossintáctica desambiguada do português. Master's thesis, Instituto Superior Técnico, Lisbon, Portugal, 2003.