Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntatic Approach

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

Although the vocabularies of ASR systems are designed to achieve high coverage for the expected domain, out-of-vocabulary (OOV) words cannot be avoided. Particularly, for daily and real-time transcription of Broadcast News (BN) data in highly inflected languages, the rapid vocabulary growth leads to high OOV word rates. To overcome this problem, we present a new morpho-syntatic approach to dynamically select the target vocabulary for this particular domain by trading off between the OOV word rate and vocabulary size.

We evaluate this approach against the common selection strategy based on word frequency. Experiments have been carried out for a European Portuguese BN transcription system. Results computed on seven news shows, yields a relative reduction of 37.8% in OOV word rate against the baseline system and 5.5% when compared with the word frequency common approach.

Index Terms: morphology, vocabulary selection, broadcast news, transcription systems

1. Introduction

Although the vocabularies of ASR systems are designed to achieve high coverage for the expected domain, out-of-vocabulary (OOV) words cannot be avoided. Particularly, for tasks like daily and real-time transcription of Broadcast News (BN) data in highly inflected languages, the rapid vocabulary growth leads to high OOV word rates [1].

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

Usually, a number of training text corpora from various domains and time periods are available, with the amount of data available in the target domain being far less than in the other sources. Facing this situation, we would like to infer the target vocabulary from the individual training corpora of different origins, sizes and recencies, taking in consideration the observable portion of the domain text as a sample. The most common approaches to vocabulary selection and optimization are typically based on word frequency, including words from each corpus that exceed some empirically defined threshold, which mainly depends on the relevance of the corpus to the target task [4].

In [5] three principled methods for selecting a single vocabulary from many corpora were evaluated, concluding that the maximum-likelihood-based approach is a robust way to select a domain's vocabulary especially when reasonable amount of in-domain texts are available. Recently researchers are using the Word Wide Web (WWW) as an additional resource of training data for vocabulary and language model adaptation procedures [6]. In this paper, we propose a daily vocabulary adaptation framework for a European Portuguese broadcast news transcription system, using contemporary written texts available on the Internet.

In this framework, we introduce a modified vocabulary selection technique which uses part-of-speech (POS) word classification to compensate for word usage differences across the various training 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 [7] these similarities 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-gram model using part-of-speech (POS) labels that compensates for word usage differences across domains.

Results computed on seven broadcast news segments from March 2004, recorded during a period of one week, showed a relative reduction of 37.8% in OOV word rate against the baseline system and 5.5% when compared with the word frequency common approach.

In section 2 we describe the data sources used in our experiments. Section 3 provides a description of the problem and proposed approach and section 4 presents the vocabulary selection results, drawing in section 5 some conclusions.

2. Baseline system and datasets

2.1. Baseline system

For the work presented in this paper, we used the system reported in [8], an European Portuguese broadcast news transcription system. The recognition vocabulary of the baseline system has been selected from two training corpora:

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

This vocabulary was created using words selected from the written news set according to their weighted class frequencies

of occurrence. Different weights were used for each class of words. All new different words present in the acoustic training data were added to the vocabulary giving a total of 57,564 words (57K).

2.2. Evaluation datasets

To implement and evaluate the proposed approach we used two datasets defined in [9]. Those datasets were 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, we would expect to cover a typical situation of rich content and topic changing over time.

Thus, we collected two different datasets: the March11-N corpus containing the written news collected from the online Portuguese newspapers web editions in a daily basis (an average of 280K words per day); and the March11-B consisting of seven shows from the 8 o'clock pm (prime time) news from the main public Portuguese channel, RTP. These shows had a total duration of about 5 hours of speech (about 53.000 words).

Table 1 shows the statistics related to the March11-N written news corpus with the OOV word rate for the baseline vocabulary of 57K.

Table 1. *Statistics for the March11-N dataset.*

	8th	9th	10th	11th	12th	13th	14th	week
Word tokens	280K	270K	286K	232K	250K	319K	310K	2M
Word types	24K	23K	25K	24K	25K	27K	26K	66K
%OOV tokens	2.83	2.94	2.98	3.47	3.49	3.20	3.26	3.16

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. The OOV word rate for March11-B broadcast news transcriptions is presented in table 2.

Table 2. *Statistics for the March11-N dataset.*

	8th	9th	10th	11th	12th	13th	14th	week
Word tokens	8.7K	1.9K	8.4K	8.3K	8.8K	7K	9.4K	53K
Word types	2.4K	0.7K	2.4K	2K	2.1K	1.8K	2.1K	13K
%OOV tokens	1.08	0.62	1.31	1.25	1.22	1.74	1.16	1.25

3. Vocabulary selection

An important problem in vocabulary design is to identify and rank the most relevant vocabulary words. Due to the large variety of topics discussed over time, this problem is even more serious for BN data. Even though the use of very large vocabularies in recognition systems can reduce the OOV word rates, in highly inflected languages or those with a high rate of word compounding, the OOV word rates still tend to be high [1]. The appropriate strategies to identify new words likely to be of interest will depend in some extend on the domain. The most common ones normally use the word frequency in each corpus as a criterion for the selection.

In [9] we presented a procedure for dealing with the OOV problem by dynamically increasing the baseline vocabulary over time. The idea of vocabulary adaptation was to use written news daily available on the Internet to adjust the vocabulary and reduce the impact of linguistic differences over time. Similar approaches have been proposed where vocabulary adaptation is carried out by adding and removing words from the baseline vocabulary according to frequency and recency in contemporary written news [10][11]. In all these approaches, the baseline vocabulary is extended or modified by using all the new words appearing in the text corpus extracted from the Internet.

For the approach proposed here we introduce a modified vocabulary selection technique which uses part-of-speech (POS) word classification to compensate for word usage differences across the various training corpora. This approach is based on the hypothesis that the similarities between different domains can be characterized in terms of style (represented by the POS sequences). Hence, instead of simply add new words to the fixed baseline vocabulary, we use the statistical information related to the distribution of POS word classes on the in-domain corpus to dynamically select words from the various training corpus available.

As in [7], we opted to use linguistic defined POS classes for representing style. Hence, using a morphological analysis tool developed for the European Portuguese [12], we annotated both in-domain corpus (BN-ALERT) and a segment of out-of-domain corpus (NP-2003) with a similar size (we randomly selected from NP-2003 a set of sentences equivalent in size to the BN-ALERT corpus). In table 3, we summarize the POS statistics obtained in both corpora by breaking down word types into four main classes: names (including proper and common names), verbs, adjectives and adverbs. Other type of words, such as the functional words, are absent from the list shown in table 3 because they represent almost closed grammatical classes in the Portuguese language.

Table 3. *Distribution in percentage of words by POS classes.*

	Names	Verbs	Adjectives	Adverbs
NP-2003	**45.0**	**30.5**	22.5	2.0
BN-ALERT	**40.6**	**36.9**	20.9	1.6

Table 3 clearly shows a significant difference in POS distribution when comparing in-domain and out-of-domain corpora, specially in terms of names and verbs.

Based on the above observations, we defined a new approach for vocabulary selection, which takes in consideration the differences in style across the various corpora. This approach can be briefly summarized as follows. Let $c_{i,j}$ be the counts from each of the available training corpus t_j ($j = 1,\ldots,n$), for the word w_i. Due to the differences in the amount of available data for each training corpus, we start by normalizing the counts according to their corpus length. The counts and normalization process were done using the SRILM toolkit [13] and Witten-Bell as the discounting strategy.

From these normalized counts ($\eta_{i,j}$) we want to estimate some kind of weighting η_i for each word w_i in order to select a vocabulary from the union of the vocabularies of t_1 through t_n that minimizes the OOV word rate for the in-domain task. In [5] 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. In our work we use a similar method

but simply assigning identical values to all the mixture coefficients. Hence,

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

Assuming $M(p), p \in POSset$ as the distribution of POS classes on the in-domain corpus, and V as the dimension of the target vocabulary, the number of words selected from each class p will be:

$$V \times M(p) \text{ words selected for each class } p \quad (2)$$

This way, for each p class the $V \times M(p)$ words with the highest weighting defined by η_i are selected to be included in the target vocabulary. In our implementation we defined

$$POSset = \{\text{names, verbs, adjectives, adverbs}\} \quad (3)$$

Finally, all the words belonging to the remaining POS classes (mainly functional words) are automatically added to the vocabulary. In the training corpora used in this work we have only 468 words which POS class doesn't belong to $POSset$.

4. Results

To evaluate this new vocabulary selection approach, we compared the OOV word rate obtained by four different methods over the seven broadcast news shows of March11-B dataset:

- BASELINE: baseline vocabulary consisting of 57K words.
- BASELINE+DAY: baseline vocabulary extended with the new words appearing on the newspapers texts extracted from the Internet in a daily basis (March11-N dataset). Applying this adaptation approach, the baseline vocabulary of 57K was expanded by an average of 5K new words each day, giving a final vocabulary size of 62K words per day.
- ADAPTED_WF: selecting a new vocabulary based on all the training corpora and using word frequency as the only selection criteria. In this case, the vocabulary is selected in a daily basis, using all the three training corpora: in-domain corpus (BN_ALERT), out-of-domain corpus (NP-2003) and written news collected day-by-day (March11-N).
- ADAPTED_POS: selecting a new vocabulary based on all the training corpora and using the new approach described in section 3. As in ADAPTED_WF approach, all the three training corpora have been used in the vocabulary selection process. To estimate the distribution of POS classes we used the in-domain corpus BN-ALERT. Hence,

$$\mathrm{M(p)} = \begin{cases} 40.6\text{ , class of names} \\ 36.9\text{ , class of verbs} \\ 20.9\text{ , class of adjectives} \\ 1.6\text{ , class of adverbs} \end{cases} \quad (4)$$

In both ADAPTED_WF and ADAPTED_POS approaches we used $V = 62K$ in order to make results comparisons with BASELINE+DAY approach.

As one can observe from table 4, the new proposed ADAPTED_POS approach yields a relative reduction of 37.8% in OOV word rate, when compared to the results obtained with the baseline vocabulary. Moreover, this approach outperformed all the other methods, being more effective than the common word frequency approach.

Table 4. *OOV rate applying different methods of vocabulary selection (V = 62K words).*

Approach	%OOV	%reduction
BASELINE (57K)	1.25	-
BASELINE+DAY (62K)	0.89	28.5
ADAPTED_WF (62K)	0.82	34.1
ADAPTED_POS (62K)	0.78	37.8

On table 5 we present the distribution (in percentage) of words by POS classes for the different vocabularies produced by each one of the selection procedures. One can clearly observe the differences in POS class distribution. In fact, the new approach selects words in a more balanced way, specially in case of names and verbs classes.

Table 5. *Distribution in percentage of words by POS classes for different vocabularies.*

	Names	Verbs	Adjectives	Adverbs
BASELINE+DAY	**61.6**	**22.5**	14.1	1,8
ADAPTED_WF	**56.3**	**26.9**	15.6	1,2
ADAPTED_POS	**40.6**	**36.9**	20.9	1.6

To better understand the performance of this new vocabulary selection procedure for different values of V (vocabulary size), we calculated the OOV word rate results for vocabularies of 5K, 25K, 50K, 100K, 150K and 200K words (table 6).

Table 6. *Word Frequency vs. POS approach results for different values of V (vocabulary size).*

V	Approach	%OOV	%reduction
5K	ADAPTED_WF	10.23	
	ADAPTED_POS	10.89	**-6.4**
25K	ADAPTED_WF	2.45	
	ADAPTED_POS	2.49	**-1.8**
50K	ADAPTED_WF	1.11	
	ADAPTED_POS	1.05	**5.8**
62K	ADAPTED_WF	0.82	
	ADAPTED_POS	0.78	**5.5**
100K	ADAPTED_WF	0.39	
	ADAPTED_POS	0.36	**6.8**
150K	ADAPTED_WF	0.22	
	ADAPTED_POS	0.20	**8.8**
200K	ADAPTED_WF	0.14	
	ADAPTED_POS	0.13	**7.9**

We compared the common word frequency based approach with the proposed POS class based approach. Results in table 6 show the relative good performance of ADAPTED_POS approach for selection of large-sized vocabularies.
Furthermore, as we would expect, for selection of small vocabularies better results are achieved by using the ADAPTED_WF method.

5. Conclusions

This paper described a dynamic vocabulary adaptation framework that tries to optimize the trade-off between the expected OOV word rate and the number of added words. It uses POS class information about an in-domain training corpus to select an optimal vocabulary for domain-specific language modeling tasks. When applied to a daily and real-time broadcast transcription task, this procedure showed to be effective in reducing the OOV word rate (more than 37%) when compared with the OOV word rate obtained with a baseline vocabulary. Moreover, this adaptation procedure is simple, extensible to any number of available training corpora and experimental results showed that when compared with the common word frequency based approach it gives better results, specially for selection of large-sized vocabularies.
To extend the effectiveness of this adaptation approach, we will study the use of additional in-domain corpus to estimate the mixture coefficients used to calculate the interpolation of normalized counts.

6. Acknowledgments

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

7. References

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

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

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

[4] Gauvain, J., Lamel, L. and Adda, G., “The LIMSI Broadcast News Transcription System”, Speech Communication, vol. 37, 2002.

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

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

[7] Iyer, R. and Ostendorf, M., “Transforming out-of-domain estimates to improve in-domain language models”, in Proc. of Eurospeech, 1997.

[8] 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.

[9] Martins, C., Teixeira, A., and Neto,J., “Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News Transcription System”, IEEE/ACL Workshop on Spoken Language Technology, December 2006.

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

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

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

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