FROM BROADCAST NEWS TO SPONTANEOUS DIALOGUE TRANSCRIPTION: PORTABILITY ISSUES

N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico and D. Giuliani

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo di Trento, Italy

ABSTRACT

This paper reports on experiments of porting the ITC-irst Italian broadcast news recognition system to two spontaneous dialogue domains. The trade-off between performance and the required amount of task specific data was investigated. Porting was experimented by applying supervised adaptation methods on acoustic and language models. By using two hours of manually transcribed speech, word error rates of 26.0% and 28.4% were achieved by the adapted systems. Two reference systems, developed on a larger training corpus, achieved word error rates of 22.6% and 21.2%, respectively.

1. INTRODUCTION

An interesting issue of speech recognition technology is the ability to port at low cost a system from one task to another. This paper[1] investigates portability issues encountered when adapting a large vocabulary Italian broadcast news recognizer to two spontaneous speech dialogue tasks.

Considering the introductory sentence, the first question that arises is how to measure the "cost" of a porting operation. Given that most of the development of a speech recognizer is data driven, i.e. acoustic and language modeling, it seems reasonable to relate the cost of porting to the amount of required data. In general, acoustic modeling requires speech recordings with accurate transcriptions, that include annotation of spontaneous speech phenomena. Language modeling usually requires task related transcripts.

The here considered approach is to apply acoustic model (AM) and language model (LM) adaptation techniques by using increasing amounts of supervised data from each task. Baselines of LMs and AMs of each task were available and taken as references. Two significant test sets were used to evaluate word error rates, LM perplexity and out-of-vocabulary word rates of each evaluated speech recognizer. Effectiveness of AM and LM adaptation was inspected by performing contrastive experiments, that keep either the AM or the LM fixed.

Experimental results showed that the manual annotation of spontaneous speech phenomena did not result relevant for the sake of AM adaptation, as they were sufficiently well modelled by the broadcast news AM. Moreover, by using up to two hours of supervised data for adapting the AM and LM of the broadcast news system, a 44.3% word error rate reduction was achieved on both tasks.

The paper is organized as follows. Section 2 quickly introduces the ITC-irst large vocabulary speech recognition system, and describes some of its features that were not covered by previous papers. Section 3 presents the broadcast news transcription baseline and the two spoken dialogue tasks with their respective reference baselines. Section 4 describes the methods applied to adapt from supervised data both the AM and LM of the broadcast news baseline. Section 5 presents the experimental results of this work. Section 6 ends the presentation by giving some conclusions.

[1]This work was partially financed by the European Commission under the project CORETEX (IST-1999-11876).

2. SYSTEM DESCRIPTION

The ITC-irst large vocabulary speech recognition system features a single pass beam-search decoder, context dependent HMMs, and a trigram LM. The acoustic front-end uses a sliding window of 20ms, with a step of 10ms, to compute 12 mel-scaled cepstral coefficients, the log-energy and their first and second time-derivatives. The following subsections will focus on two features of the system: the LM representation and some implementation solution exploiting parallelism. Complementary information about the system can be found in [1].

2.1. Shared-tail Representation of Trigram LMs

An interpolated trigram LM (see subsection 4.2) is mapped into a static network with a shared-tail topology. As described in [2], this topology allows to dramatically reduce the size of a static LM representation by exploiting both sparseness in the LM training data and redundancy in the tree-based representation. Since the publication of [2], the algorithm has been extended in two ways: first, support for multiple pronunciations of a single word was introduced, without requiring duplication of the successor tree; second, the compilation of trigram LMs was made possible.

In order to allow multiple pronunciations of words, during the construction of the network, a correspondence is kept between lexical entries, which depend on pronunciation, and words. Since the structure of trees depends on phonetic transcription, tree leafs are associated to lexical entries, while the root of the successor tree of a word depends only on the word identity. Therefore, different leafs corresponding to different pronunciations of a word are linked to the root of a common successor tree.

The generalization to trigram LMs extends definition and properties of the basic topology of bigram LMs presented in [2]. In particular, the same procedure of identifying and sharing linear tails in successor trees, that was applied to trees at the second level, is now applied to trees at the third level. Figure 1 depicts an example of a simple trigram set compiled into a network. The triangles denoted by $S(\cdot)$ are the successor trees of a context (a single word or a word pair), that is the support of the corresponding discounted relative frequency. The symbol $\lambda(\cdot)$ refers to the zero-frequency

probability of a context, as computed by the discounting method used in LM estimation (see subsection 4.2).

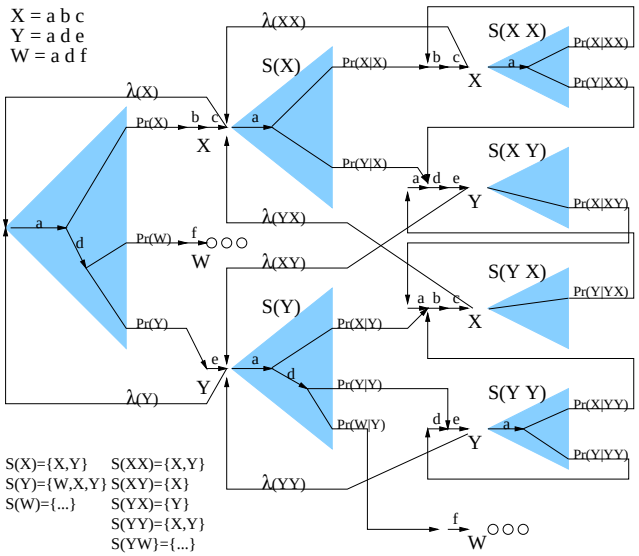

Fig. 1. Shared tail topology of a trigram LM.

As an example, the 61K-word trigram LM used for BN recognition includes about 20M parameters and is compiled into a network with about 11M states, 10M named transitions and 20M empty transitions.

2.2. Exploiting Parallelism

The necessity of processing large amount of data is by now an endured nuisance in the development and application of speech recognition systems. The relative inexpensiveness of medium class hardware permits to exploit parallelism by distributing the computational load among several CPUs. The following paragraphs describe the way coarse-grain parallelism was introduced in the ITC-irst system.

Distributed Processing. Most of the modules (feature extraction, decoding, etc.) run as "filters", reading data from standard input and writing results to standard output. A complete system is built by connecting modules by means of pipes. Parallelism is therefore introduced by inserting, at a certain point of the pipeline, a program that instantiates multiple copies of a module, running on the same machine or on different machines, and then dispatches incoming input records to the different instances. To avoid interleaving of results coming from different sources, the dispatcher also collects the output streams of the different processes and writes them on its standard output, preserving their structure. The dispatcher is implemented so as to balance the load among the different servers, and keep them all as busy as possible. In several experiments, a homogeneous distribution of processing time was actually observed, while the amount of processed data varied according to the relative speed of the servers.

Shared Memory Decoding. Memory sharing among processes running on the same machine can be applied in the speech decoding process. The ITC-irst system can store in shared memory the finite state networks used for representing the language model. For large language models, this allows to fit two decoding processes on a dual processor machine with a considerable saving in global memory space, e.g. two 700Mb speech decoders into 1Gb of memory.

Statistics Recombination. In HMM training, data are processed to collect global statistics in order to update the HMM parameters. If several processes are working on different portions of the training set, each of them will collect partial statistics. An additional module was developed that recombines different partial statistics into global ones. The additional step does not introduce a significant overhead.

3. BASELINES

Three baseline systems are involved in this work: an Italian broadcast news (BN) transcription system and two spontaneous dialogue recognition systems.
The AM of the BN speech recognizer was trained on recordings of radio and television news programs. The BN LM was trained over a large sample of newspapers, newswire, and news transcripts.
The domains of the two dialogue recognition systems are, respectively, appointment scheduling (SCHE) and tourist information (TOUR). A corpus of two-party task-oriented conversations was already available. Recordings were in studio quality and manually transcribed; spontaneous speech phenomena were accurately labeled. A single spontaneous dialogue (SD) AM was trained over the union of the SCHE and TOUR corpora, because both data sets present the same acoustic conditions. Task specific LMs were instead developed over the respective training transcript corpora.
Statistics regarding the AMs and LMs of the three baselines are reported in Table 1 and Table 2.

	BN	SD
duration	36h:34m	10h:56m
#triphones	6554	1956
#backoff models	2367	896
#gaussians	14087	8829

Table 1. Statistics of AM training sets and properties of models.

	BN	SCHE	TOUR
corpus size (#words)	215M	21.6K	61.7K
vocabulary size	61K	1.2K	2.2K
#trigrams	77M	14K	37.5K

Table 2. Statistics of LM training sets and properties of models.

4. ADAPTATION TECHNIQUES

4.1. AM Adaptation

Adaptation of the AM, that is based on HMMs having emission distributions modeled by mixtures of Gaussian densities with diagonal covariance matrices, is carried out through Maximum Likelihood Linear Regression (MLLR) [3, 4].
During adaptation a regression class tree is employed in order to determine regression classes according to the available adaptation

data. The regression class tree is generated by means of an agglomerative clustering procedure employing the likelihood measure. The regression class tree is built in two steps: first, for each phone-like unit Gaussian components are hierarchically clustered; second, the roots of trees obtained with the first step are clustered in their turn. Base regression classes are then determined by imposing a minimum number of Gaussian components (i.e. 32) per class.
Two MLLR iterations are performed to adapt means and variances. Mean vectors are adapted using full transformation matrices, while diagonal transformation matrices are used to adapt the variances.

4.2. LM Adaptation

Basic LMs are estimated through an interpolation scheme, i.e. the probability of an n-gram hw, where h represents the history of word w, is computed by:

$$Pr(w \mid h) = f^*(w \mid h) + \lambda(h) Pr(w \mid \bar{h}). \quad (1)$$

f^* is a discounted relative frequency that is smoothed with the zero frequency estimate $\lambda(h)$ weighted with the distribution of the lower order $(n-1)$-gram $\bar{h}w$.
Given k interpolated language models, one can define the following mixture discounted relative frequency:

$$f^*_{mix}(w \mid h) = \sum_{i=1}^{k} \mu_i f^*_i(w \mid h) \quad (2)$$

where μ_i are weights of a convex combination. From the definition of the zero frequency probability (see e.g. [5]), it follows that:

$$\lambda_{mix}(h) = \sum_{i=1}^{k} \mu_i \lambda_i(h) \quad (3)$$

which leads to the interpolation mixture model:

$$Pr_{mix}(w \mid h) = f^*_{mix}(w \mid h) + \lambda_{mix}(h) Pr_{mix}(w \mid \bar{h}) \quad (4)$$

An advantage of the proposed mixture model is that it preserves the basic interpolation scheme (1) and hence allows the efficient language model representation described in Section 2. Moreover, the mixture weights μ_i can be estimated by applying the EM algorithm.
Improvements in performance were obtained by letting the interpolation weights μ_i depend on the most recent word of the history h. Parameter tying was applied to cope with data sparseness.
In the here considered LM adaptation case, two component mixture LMs were taken, by combining the BN trigram LM with task specific trigram LMs estimated on relatively small adaptation texts. Both LMs applied a non-linear discounting method, i.e.:

$$f^*(w \mid h) = \frac{c(hw) - \beta}{c(h)} \quad \text{with} \quad 0 \leq \beta \leq 1 \quad (5)$$

with β estimated according to [5] on the adaptation data, and set to 1 for the BN LM. Moreover, trigram pruning was applied to the BN LM [1]. Finally, the EM estimation of the mixture parameters was carried out on the adaptation data sample, by applying a cross-validation technique.

5. PORTABILITY EXPERIMENTS

Several experiments on the portability of the BN system to the SCHE and TOUR domains were conducted, testing the adaptation algorithms described in Section 4. Performance evaluation were carried out on test sets with no speaker overlap with the training data. Statistics about the test samples are reported in Table 3.

	#turns	duration	#speakers	#words
SCHE	1007	1h:34m	24	10.8K
TOUR	1520	1h:46m	19	13.6K

Table 3. Test sets statistics of porting experiments.

5.1. Adaptation Corpora

In order to adapt BN AM and LM to the spontaneous dialogue tasks, increasing amounts of speech data from the task specific data were used; details are reported in Table 4.

	SCHE				TOUR			
	#trn	#spk	#wrd	voc.	#trn	#spk	#wrd	voc.
0.5h	386	14	3.5K	444	448	13	3.6K	557
1.0h	770	22	7.0K	634	913	24	7.7K	774
1.5h	1034	27	10.5K	790	1312	32	10.9K	863
2.0h	1356	31	14.2K	1020	1841	40	15.3K	967

Table 4. Statistics of supervised adaptation data.

	SCHE		TOUR	
	detailed	plain	detailed	plain
0.0h	31.8	31.8	32.4	32.4
0.5h	27.9	27.7	28.0	27.8
1.0h	25.4	25.5	26.2	26.4
1.5h	26.2	26.2	25.9	26.0
2.0h	25.8	25.8	25.0	25.3

Table 5. WER results after adapting the BN AM by using detailed or plain transcripts.

5.2. Results

Supervised AM adaptation was investigated by using manually produced verbatim speech transcripts either including (detailed) or not including (plain) the annotation of spoken language phenomena, i.e. hesitations, filled pauses, noises, mispronunciation of words, truncated words, etc. In both cases, the actual model sequence corresponding to a given speech signal was estimated through a decoding step using the BN AM and constrained by the supervision transcript, with the optional insertion of extra-linguistic phenomena between annotated events.
In Table 5 results are reported in terms of word error rate (WER), by using the baseline LM for each task. It comes out that AM adaptation is effective even if spontaneous speech phenomena annotation is not available. The detection of such phenomena achieved by BN AM proved to be sufficiently reliable. In terms of portability, this result can imply significant cost savings.
Hence, by exploiting plain transcripts both for AM and LM adaptation, a number of comparative experiments were performed on the SCHE and TOUR tasks, which are reported in Figures 2 and 3, respectively. The WER figures on the borders are to be interpreted

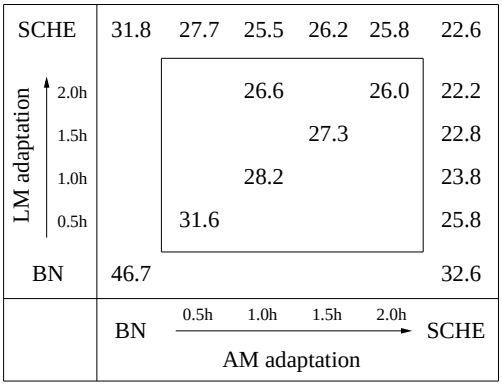

Fig. 2. WER results by porting the BN system to the SCHE domain.

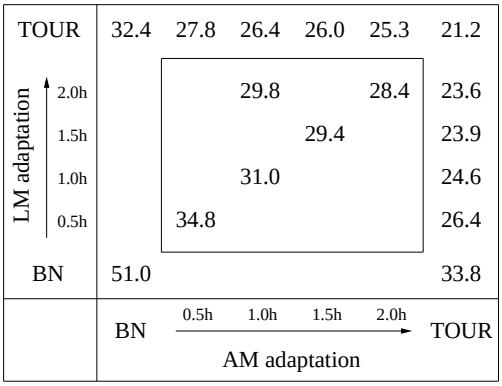

Fig. 3. WER results by porting the BN system to the TOUR domain.

as references, while those in the internal grid give the actual performance of different adapted recognizers. Reference results are instead obtained by using the AM and/or the LM of a baseline system.

Adaptation results show that two hours of adaptation material yields a relative error rate reduction of 44.3% both on the SCHE task (from 46.7% to 26.0%) and on the TOUR task (from 51.0% to 28.4%).

The WERs reported in the rightmost columns show that LM adaptation is already effective with little adaptation data, and that further improvements are obtained by enlarging the adaptation set. By comparing the LM adaptation speed against the AM one, in the uppermost row, it results that the difficulty in porting the BN system is mostly due to the acoustic mismatch.

Table 6 reports performance, on both tasks, of the reference and adapted LMs, in terms of perplexity (PP) and out-of-vocabulary word rate (OOV). It results that 90% of the gap between the BN LM and the task dependent LM is filled with just half an hour of adaptation material. The WER achieved on the SCHE task by adapting the BN LM with two hours of transcripts (22.2%) is even better than that of the baseline (22.6%). This is mainly due to the lower OOV rate achieved by the adapted BN LM.

It is also worth noticing that the potential increase in word confusability, introduced by the large BN LM vocabulary, does not seem to be an important source of errors.

	SCHE		TOUR	
	PP	OOV%	PP	OOV%
BN baseline	538.6	1.12	524.0	1.11
0.5h	91.9	0.67	106.9	0.67
1.0h	72.7	0.58	83.2	0.59
1.5h	65.2	0.53	74.3	0.53
2.0h	60.1	0.48	69.2	0.51
Task baseline	63.5	2.46	53.9	1.40

Table 6. LM adaptation results.

6. CONCLUSIONS

This work addressed the problem of porting a large-vocabulary broadcast news recognition system to two spontaneous dialogue domains: appointment scheduling and tourist information.

Since the development of a speech recognizer is data driven, the cost of porting was related to required amount of data and quality of the supervision. Recognition experiments were conducted by adapting the AM and LM with increasing amounts of task dependent supervised data. In particular, AM adaptation was evaluated versus two possible levels of supervision: verbatim transcripts or verbatim transcripts with the annotation of spontaneous speech phenomena.

Experimental results showed that manual annotation of spontaneous speech phenomena is not relevant for supervised AM adaptation, assuming that the baseline AM provides some coverage of these phenomena. By using up to two hours of transcribed speech, the WERs of the adapted systems were 26.0% and 28.4%, to be compared with 22.6% and 21.2% obtained by the task specific baselines.

In the future, the modeling of spontaneous speech phenomena will be investigated, which is known to have an impact on the system accuracy and was not taken into account in this work. In the reported porting experiments, the optional insertion between words of any noise or filled pause was in fact allowed with the same chance as in the BN system. Automatic estimation methods would be desirable that can automatically adapt existing models of spontaneous speech phenomena from unsupervised or lightly supervised data.

7. REFERENCES

[1] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "Advances in automatic transcription of broadcast news," in *Proceedings of ICSLP*, Beijing, China, 2000, pp. II:660–663.

[2] F. Brugnara and M. Cettolo, "Improvements in tree-based language model representation," in *Proceedings of EUROSPEECH*, Madrid, Spain, 1995, pp. 2075–2078.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[5] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.