Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 114 – 120

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016,
9-12 May 2016, Yogyakarta, Indonesia

Semi-Supervised Training of Language Model on Spanish Conversational Telephone Speech Data

Ekaterina Egorova[a,b,*], Jordi Luque Serrano[a]

[a]*Telefonica Research, Edificio Telefonica-Diagonal, Barcelona 08019, Spain*
[b]*Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno 61200, Czech Republic*

Abstract

This work addresses one of the common issues arising when building a speech recognition system within a low-resourced scenario - adapting the language model on unlabeled audio data. The proposed methodology makes use of such data by means of semi-supervised learning. Whilst it has been proven that adding system-generated labeled data for acoustic modeling yields good results, the benefits of adding system-generated sentence hypotheses to the language model are vaguer in the literature. This investigation focuses on the latter by exploring different criteria for picking valuable, well-transcribed sentences. These criteria range from confidence measures at word and sentence level to sentence duration metrics and grammatical structure frequencies.

The processing pipeline starts with training a seed speech recognizer using only twenty hours of Fisher Spanish phone call conversations corpus. The proposed procedure attempts to augment this initial system by supplementing it with transcriptions generated automatically from unlabeled data with the use of the seed system. After generating these transcriptions, it is estimated how likely they are, and only the ones with high scores are added to the training data.

Experimental results show improvements gained by the use of an augmented language model. Although these improvements are still lesser than those obtained from a system with only acoustic model augmentation, we consider the proposed system (with its low cost in terms of computational resources and the ability for task adaptation) an attractive technique worthy of further exploration.

Peer-review under responsibility of the Organizing Committee of SLTU 2016
Keywords: Speech recognition, language modeling, semi-supervised learning

1. Introduction

Manual transcription of training data is an expensive and time-consuming undertaking. Therefore, in case of sparse resources or limited time allowance, speech recognition system may suffer from undertraining because of insufficient training resources. Possible treatments of this issue include using data from other languages to enhance the system (also known as multilingual training [1]) or techniques which aim at dealing with non-labelled data in the target language in order to boost the system.

[*] Corresponding author. Tel.: +420-727-982-151.
E-mail address: iegorova@fit.vutbr.cz

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.038

This paper focuses on the latter, specifically on investigating the process of iterative addition of automatically transcribed data for improving the recognition system. This procedure is especially of high interest for task adaptation on low-resourced scenarios. The main focus of this work consists in studying methods for selection of new utterances, unknown from the training point of view, for system enhancement and retraining.

Different metrics are considered and compared, including the use of confidence score from the Automatic Speech Recognizer (ASR) at different levels, that is, word or sentence levels, and the use of grammatical analysis of the automatically transcribed sentences.

1.1. Previous work

Recent advances in unsupervised learning as applied to speech recognition task has ignited great interest in the speech community in recent years, fueled, among others, by IARPA BABEL[1] program, which aims at rapidly building speech recognition systems for under-resourced languages. A great number of works in the field of unsupervised learning have been focused on iterative retraining of acoustic models (AM)[2]. Indeed, it has been shown that AM retraining is more effective than language model (LM) retraining in terms of reducing Word Error Rate (WER) on test data[3]. Lightly-, semi-, and un-supervised AM training has been recently successfully used for broadcast data in several languages as part of the Babel program[4].

As for language model augmentation, the task of improving the LM and extending the vocabulary is most often approached by using different sources of texts on the Internet, such as blogs, news etc[5]. However, in a case of very specific tasks in which language presents tendency to peculiar grammatical constructions (e.g. call center data), ASR systems may benefit from adapting and expanding LM with the regular collection of further data. It may be especially useful if training data is scarce or deficient for covering an acceptable modeling of the language.

The few papers that do tackle in-domain LM retraining, concentrate on several frequent approaches. One of them involves detecting sentences decoded with very low confidence measures and marking them for manual annotation[6 7]. Semi-automatic approaches aim at adding high-confidence (in terms of decoding scores) sentences to the training data and then performing a system retraining with the use of new data. There are numerous ways of calculating confidence metrics in ASR systems, ranging from scores calculated on the phonetic level to the estimation of confidence scores at the utterance level[8]. For instance, in[7] this estimation is performed with the help of a confidence model, which is trained on a subset of training data. An even more creative approach, from our point of view, consists of picking ”well” decoded sentences using two independent ASR systems[9] in a voting scheme. It suggests training two separate ASR systems on two disjoint halves of the training data and decoding the untranscribed dataset with both of them. Only if the decoding obtained from two systems matches each other, the sentence is deemed well-transcribed and is subsequently added for further retraining.

Most of the research on unsupervised learning has been concentrating on English language, and there have been a few experiments on Spanish data. In[10], the CallHome Spanish database is used to simulate an unsupervised learning scenario. The baseline system was trained on as few as 3 hours of data and then enhanced by 25 hours of untranscribed speech. The important requirement was that there were no unseen speakers in the untranscribed set.

The work presented in this paper is highly inspired by the above mentioned approaches. We suggest several strategies for ASR semi-supervised training and the results of their assessment on conversational Spanish telephone speech are reported.

2. Experimental setup

2.1. Methodology

The first requirement of an iterative system is training a reasonably good seed system. In our experiments, we used Kaldi toolkit[11] to build a single pass DNN system on top of filter-bank features, with GMM pre-training. The feed-forward DNN has 4 hidden layers (with 1024 neurons in each), not counting the output softmax layer.

[1] http://www.iarpa.gov/index.php/research-programs/babel

This seed system is then used to run the decoding on untranscribed data. The decoding chooses and outputs the best word sequence that the system can generate based on the word posteriors, calculated as a weighted linear combination of acoustic and LM scores of the lattice ark, given the path in the decoding graph (see Section 2.4).

Different metrics (see Section 2.4) may be then used to choose ”well-transcribed” sentences that finally will be added to the seed system for further training. At this stage, one can choose whether the new data should be employed uniquely for language model retraining, for acoustic model retraining or for both of them.

For LM retraining, the newly chosen sentences are added to the LM text pool and then a new LM is regenerated. The common approach to new sentence addition involves assigning them some weight on the interval from 0 to 1 since we are not so sure whether they are correctly transcribed or not. Such weight is found based on perplexity estimate. The number of sentences chosen at each step is regulated based on a threshold. For AM retraining, only words with high individual scores are added to the system.

2.2. Datasets

The experiments have been conducted on three Spanish databases. Note that first two of them are employed uniquely for seed ASR system training.

1. *SALA*: a phonetic database of separate words in different Latin American Spanish dialects recorded over fixed telephone network. It consists of more than 6000 speakers from 8 different areas of Latin America[12].
2. *TID*: 20 hours of phonetically rich telephone speech recorded in Telefónica Investigación y Desarollo. The database is composed of two sets: CEUDEX, the main set, with a corpus of 400 phonetically balanced sentences, and SPATIS, a task-oriented set which was inspired by ATIS (Air Travel Information System) standard application for English[13].
3. *Spanish Fisher Speech Corpus*, developed by the Linguistic Data Consortium, consists of 819 telephone conversations lasting around 10 to 12 minutes each, yielding roughly to 163 hours of telephone speech from 136 native Caribbean Spanish and non-Caribbean Spanish speakers. A broad set of topics is covered in the conversations ensuring speech variability. Speaker segmentation is done by analysing independently each conversation channel, which is supposed to correspond to one speaker[14]. Fisher corpus comprises a challenging, large vocabulary, spontaneous speech recognition dataset ideal for our purposes.

To experiment with iterative unsupervised learning, 80% of Fisher database was used as the ”new” untranscribed data, 10% was set out as test data and 10% was used for training the seed system. From the training data, 10% was set out as development set, to be used for perplexity estimation, etc. When subdividing the data, care was taken to separate the speakers. The seed system was also augmented by TID dataset for language modeling (LM) and both TID and SALA datasets for acoustic modeling (AM).

2.3. Baseline System Description

In order to assess how much information could be gained from adding the ”untranscribed” data to the system, a comparison is made between two systems, one using only 10% of Fisher database in the training and the other using 90% of Fisher for training. With the addition of the 80% of ”untranscribed” data to the system, WER was reduced from 59.5% to 43.3% (first and second rows in Table 1. It should be noted that by adding this new 80% we not only add more data for AM and LM retraining but also reduce the OOV (Out-Of-Vocabulary) rate in the test data by adding words from the new 80% to the lexicon. This improvement can be seen as the upper bound in the sense of WER improvement of the hypothetical best iterative system with respect to the baseline system.

Dealing with OOV words is out of the scope of this work. Thus, an experiment is made aiming at separating the percentage of errors on the test set due to OOV words and due to insufficient sentence statistics for language modeling. In the case the reference transcriptions from the test set are added to LM and all words from test set are also included in the lexicon, error rate goes down from 43.3% to 29.9% (fourth row in Table 1). And if sentence structures from the test set are added to LM, but the unseen words are not added to the dictionary and are substituted with the <garbage > token in the sentence, the error rate predictably falls in the middle between the two results, at 30.2% (see third row in Table 1). These experiments show us that growing LM by incorporating as much sentence structures as possible is beneficial as it would increase our chances to discover unseen sentence structures in the test set.

Table 1. ”Oracle” gains from utilizing untranscribed data obtained by iteratively adding more Fisher data to the ASR system. The effect on WER error by both perfect LM and vocabulary are also reported

System training data	WER
10% Fisher	59.5%
90% Fisher	43.3%
90% Fisher + test LM	30.2%
90% Fisher + test LM + no OOVs	29.9%

2.4. Sentence picking strategies

In each iteration of unsupervised training, the main question that arises is how to pick sentences and how to add them to the new training set. Obviously, the system should be very sure about the hypothesised transcription but, in addition, discovered sentences should also bring something ”new” to the LM and not just reinforce the same n-gram constructions over and over again. This situation may lead to a bias in the LM and, therefore, produce the effect opposite to what we were looking for by skewing the LM estimation and consequently degrading the system performance. Thus different approaches have been explored:

1. **Sentence posteriors**
   Sentence posterior metric is a normalised sum of word posteriors, which is a linear combination of AM and LM posteriors of these words, given the path.

   $$p(w|x) = p(x|w)^{\alpha} p(w) \tag{1}$$

   where α is the acoustic weight, $p(x|w)$ is acoustic model probability and $p(w)$ is language model probability[15]. The bigger the posterior metric, the more confident is the system that the decoding is correct, hence the name ”confidence measure”[2].

2. **Minimum Bayes Risk scores**
   Another way to estimate how well a sentence is decoded is looking at Minimum Bayes Risk (MBR) scores.

   $$\delta_R(A) = \underset{W' \in \mathcal{W}_h}{\arg\min} \sum_{W \in \mathcal{W}} l(W, W') P(W|A) \tag{2}$$

   where A is an acoustic observation sequence, $P(W|A)$ is the probability of an utterance given the audio signal, and $l(W, W')$ is a real-valued loss function that describes the cost incurred when an utterance W belonging to language $\mathcal{W}$ is mistranscribed as $W' \in \mathcal{W}_h$[16].

3. **Length constraint**
   As our language model is based on a 3-gram, which means that the word probability depends on two previous words, it makes sense to filter out sentences consisting of less than three words, as they would not contribute to the retrained language model.

2.5. Word selection for acoustic retraining

Even if a sentence metric is good, single words constituting it may be decoded with a low probability which usually signifies that this word is an OOV word or a normal word but in an unexpected place or with an unexpected pronunciation. Thus, this word may not be useful for AM retraining and should probably not influence LM retraining. To find all these words, we look at per-word MBR scores, and all those words that get Bayes risk scores more than zero are substituted with the <garbage> token.

2.6. Results

As can be seen in Table 2, LM retraining using sentences picked with the help of posteriors confidence measure does not improve the WER, while MBR scores with length constraint can improve test set WER by about 0.1 percent. In case AM retraining is performed with the use of the newly picked words, WER is decreased by 0.4 absolute, much more than in the case of the sole LM retraining, which is consistent with previous analysis reported, for example, in[3].

Table 2. Results on semi-supervised training with Fisher data using the different ASR retraining strategies mentioned in the text. First row shows the baseline. Second and third row show systems with LM retraining. Second row displays results using confusion scores picking metric. Third row shows results of three iterations by using MBR scores and length constrain metrics. The final row shows WER improvement on AM retraining.

Unsupervised technique	relative WER
Baseline	59.5%
Confusion metric; LM only	+0.4%
MBR scores + len; LM only (1/2/3 iter)	59.5%/−0.1%/−0.1%
AM only	−0.4%

2.7. Analysis

For a more in-depth understanding of which sentences make the system better when added to the training set and which do not contribute anything, the following set of experiments was conducted: a bunch of 5000 sentences were picked randomly from the untranscribed set and included into the training data. This was repeated several times, giving us several systems differing only in the 5000 ”recovered” sentences. Each of the resulting systems was then evaluated on the same test data and the WER improvement (positive or negative) was added to contribution score of every sentence included into this bunch. Due to randomness, different sentences participated in a different number of bunches, so the contribution scores were normalised. The experiments have shown that there is no obvious correlation between the average WER improvement and the distribution of sentence MBR metrics in a bunch. This finding suggests that MBR may not be the best or the only metrics to guide us in sentence picking for further system retraining.

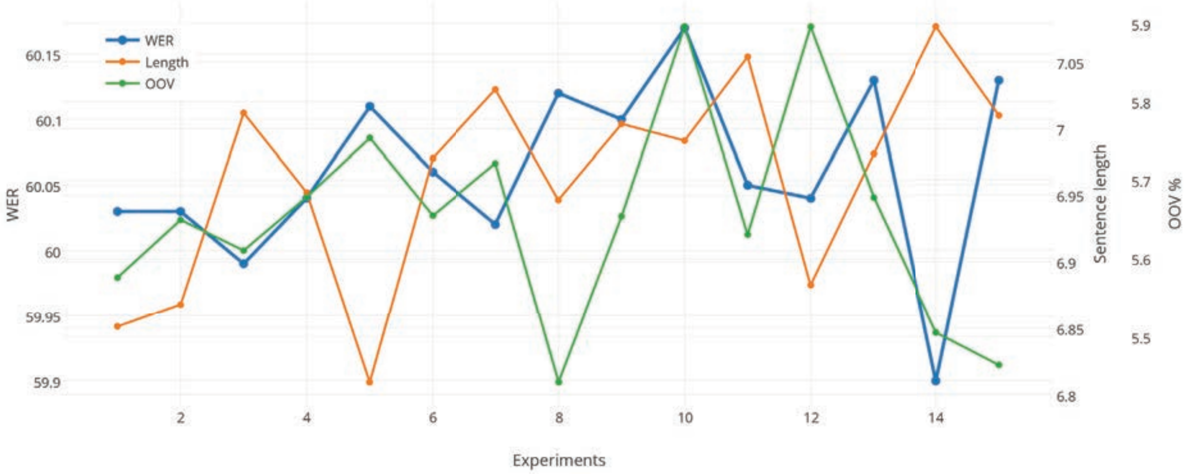

Fig. 1. Graph showing negative correlation (Pearson’s linear correlation coefficient $\rho = -0.22$) between WER score and average sentence length in the bunch and positive correlation (Pearson’s linear correlation coefficient $\rho = 0.23$) between WER scores and percentage of OOV words in a bunch.

As we have found out that MBR alone is inadequate for sentence picking, another metrics may be worth looking at. Figure 1 shows a correlation between WER scores using a current bunch of sentences, average sentence length

Table 3. Recurring grammatical structures (taken from Freeling grammar analyzer outputs) of the sentences. Table reports structures' contribution to the system when added to the LM. The numbers in the first and the third columns show the difference between the number of times the sentence structure occurred in ”good” bunches and the number of times it occurred in ”bad” bunches.

”Good” sentence structures		”Bad” sentence structures	
20	<garbage> I	-142	I
19	SPS00	-35	CS
17	RG <garbage>	-31	<garbage> RG
16	PR0CN000	-27	NCFS000
15	<garbage> <garbage>	-26	AQ0CN0
14	<garbage> CS	-20	I RG
13	VSIP3S0	-14	RG RN
12	<garbage> AQ0CN0	-13	<garbage> RG RG
11	SPS00 AQ0FS0 NCMS000	-13	RG NCMS000
11	PP1CSN00 RG	-12	PT0CN000 NCFS000

in the bunch and percentage of out of vocabulary (OOV) words in the current bunch (calculated from the reference transcriptions). It can be noted that there is a negative correlation (Pearson's linear correlation coefficient $\rho = -0.22$) between WER score and average sentence length in the bunch and positive correlation (Pearson's linear correlation coefficient $\rho = 0.23$) between WER scores and percentage of OOV words in a bunch. This basically means that 1) introducing sentences with a high number of OOVs to the LM do not improve the system and 2) longer sentences improve LM more than shorter sentences.

In order to investigate what it is that makes sentences useful for addition to the system, we analyzed sentences from ”good” and ”bad” bunches with Freeling grammar analyzer[2]. As the total number of random bunches participating in the experiments was 40, 10 bunches which addition resulted in the best WER improvement have been chosen as ”good” and 10 with the worst (even negative) WER improvement were chosen as ”bad”. Freeling analyzer may suggest multiple grammar tags for each word, but for our experiments, we take only the most likely variant. So after going through grammar analysis and parts of speech (PoS) tagging, each word in a sentence is substituted with its grammar tag. After that, the number of occurrences of each sentence structure in ”good” and ”bad” bunches was done. Table 3 reports the difference between the number of occurrences of a sentence structure in a good bunch compared to a bad one. Thus, the bigger the number, the more useful the structure is and the smaller the number, the more its addition to the LM degrades the overall system performance.

Most of the ”good” sentence structures are full of prepositions, relative and personal pronouns, subordinate conjunctions, etc., while ”bad” sentences tend to have much more noun phrases. It seems that the system benefits most from the addition of the sentences with words from closed classes, which makes sense in the light of the presence of OOV words in the test set.

Moreover, when distributions of phrase structures from ”good” and ”bad” bunches are compared to the distribution of phrase structures estimated through Freeling analysis of the train set, it is found out that on average ”good” bunches have 6% more new structures, which did not occur in the training set. The latter suggests that the gain is bigger when grammatically new sentences are added to the train set at each iteration. It is worth to note that previous result suggests an innovative sentence picking strategy. Nevertheless, it still needs further experimental validation that we hope to investigate in future works.

3. Conclusions

Experiments have shown that adding the best system-generated labels for untranscribed data to the training data can help improve the system performance. The gain from retraining acoustic models is more profound, but language model also benefits from the method.

[2] http://nlp.lsi.upc.edu/freeling/

The key to getting good performance of the semi-supervised system is to combine different metrics for picking the well-transcribed sentences, including MBR scores, sentence length, OOV rates and grammatical analysis of a sentence.

Further investigation may concern itself with finding the optimal ratio of the different metrics presented in the paper for making the best choice which sentences to add to the training data in which iteration. Various techniques may be tested for automatically setting the thresholds and preventing over-training. It is also planned to extend this method to other databases in the same domain to prove that the method is reproducible. Yet another valid idea would be to try incorporating grammar tags into the decoding procedure and eventually use them to enhance the training.

Acknowledgements

This project has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement No.645323 BISON. This text reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains. This work has been also partly supported by Technology Agency of the Czech Republic project No.TA04011311 ”MINT”.

While working on this project, Ekaterina Egorova was doing an internship at Telefonica Research, Barcelona.

References

1. Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., et al. Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models. *Proceedings of ICASSP* 2010.
2. Wessel, F., Ney, H.. Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing* 2005;**13**(1):23–31.
3. Novotney, S., Schwartz, R., Ma, J.. Unsupervised acoustic and language model training with small amounts of labelled data. *Proceedings of ICASSP* 2009.
4. Knill, K., Gales, M., Ragni, A., Rath, S.. Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting. *Proceedings of INTERSPEECH* 2014.
5. Mendels, G., Cooper, E., Soto, V., Hirschberg, J., Gales, M., Knill, K., et al. Improving Speech Recognition and Keyword Search for Low Resource Languages Using Web Data. *Proceedings of INTERSPEECH* 2015.
6. Yu, K., Gales, M., Wang, L., Woodland, P.. Unsupervised training and directed manual transcription for LVCSR. *Speech Communication* 2010;**52**:652–663.
7. Nakano, M., Hazen, T.. Using Untranscribed User Utterances for Improving Language Models based on Confidence Scoring. *LDC2010S01 DVD Philadelphia: Linguistic Data Consortium* 2003.
8. Hazen, T., Seneff, S., Polifroni, J.. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language* 2002.
9. Cucu, H., Buzo, A., Besacier, L., Burileanu, C.. Enhancing ASR Systems for Under-Resourced Languages through a Novel Unsupervised Acoustic Model Training Technique. *Advances in Electrical and Computer Engineering* 2015.
10. Zavaliagkos, G., Colthurst, T.. Utilizing Untranscribed Training Data to Improve Performance. *DARPA Broadcast News Transcription and Understanding Workshop* 1998.
11. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. The Kaldi Speech Recognition Toolkit. *Proceedings of ASRU* 2011.
12. Moreno, A.. SALA: SpeechDat Across Latin America. *Proceedings of LREC* 2000.
13. Torre, C., Hernández-Gómez, L., Tapias, D.. CEUDEX: A Data Base oriented to Context-Dependent Units Training in Spanish for Continuous Speech Recognition. *Eurospeech* 1995.
14. Graff, , David, , et al, . Fisher Spanish Speech. *LDC2010S01 DVD Philadelphia: Linguistic Data Consortium* 2010.
15. Wessel, F.. *Word Posterior Probabilities for Large Vocabulary Continuous Speech Recognition*. Ph.D. thesis; Technical University of Aachen; 2002.
16. Goel, V., Kumar, S., Byrne, W.. Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition. *IEEE Transactions on Speech and Audio Processing* 2004;**12**(3):234–249.