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