ACOUSTIC DOMAIN ADAPTATION WITH ENROLLMENT DATA FROM CLINICAL TASKS

Maider Lehr and Izhak Shafran

The Center for Spoken Language Understanding (CSLU)
The Oregon Health & Science University, Portland, OR, USA
{maiderlehr,zakshafran}@gmail.com

ABSTRACT

The growing interest on applying NLP and speech processing techniques to facilitate the administration and making more robust the evaluation of clinical tasks bring the need of assessing the usefulness of the regular techniques on this domain. For example, most research on Acoustic Model (AM) adaptation reported in the literature has been framed around publicly available corpus such as Switchboard and Broadcast News. In this study, we examine the effectiveness of different AM adaptation strategies with data from an enrollment phase of a clinical task – supervised vs. unsupervised adaptation and MLLR vs. MAP adaptation. Then, we examine semi-supervised methods to bridge the gap between supervised and unsupervised adaptation. Our findings provide a few guidelines for best practices in future clinical data collections.

1. INTRODUCTION

The larger goal of our clinical study is to detect episodes of depression from everyday conversations. Depression or prolonged episodes of sadness that interferes with daily activities is a common but serious illness. Left untreated, the illness can be debilitating and lead to suicidal ideation. This disorder can however be treated. Despite the emotional and societal toll of this illness, current methods for diagnosis and monitoring are subjective. For example, one of the most common rating scale, the Hamilton Depression Rating Scale includes 21 questions with between 3 and 5 responses which increase in severity. The clinician must choose the possible responses to each question by interviewing the patient. Many of the questions attempt to measure the degree of disruption of daily activity and the degree of sadness and related emotional states. Errors due to extraneous factors such as poor memory recall cannot be easily avoided in these methods. Apart from the subjective nature of these rating scales, current methods are not practical for screening a large population.

In contrast, an objective mechanism for characterizing episodes of depression can not only alleviate the errors due to poor memory recall but also provide greater temporal and contextual details. Currently, methods for measuring physiological signals from patients are not practical for capturing depressive episodes in everyday life. One solution that has attracted considerable attention recently is to exploit spoken language as evidence for emotional state of the patients. For example, Mehl and Pennebaker recorded samples of everyday conversations, manually transcribed them, and found statistically significant differences in language use that may discriminate subject with depression from controls [3] . The need for manual transcription however poses a serious bottleneck in scaling their paradigm to examine large cohorts.

The motivation of our work is to fully automate the analysis of samples of everyday conversations. The problem we are tackling includes detecting speech robustly, distinguishing the subject from other speakers, measuring prosodic cues and recognizing the spoken words. Recordings of everyday conversations contain plethora of noisy environments making each step a challenging task. As a first step, in this paper, we focus our attention on automatically recognizing the spoken words, starting with recognizing the enrollment interviews. This requires adapting a recognizer trained on publicly available corpus (Broadcast News) to this new domain, where the speakers have significantly different speaking style (subjects are from Baltimore).

2. CORPUS

The subjects from our task, 110 elderly people from Baltimore went through a enrollment phase where they had to read 4 stories with a duration between 1-4 minutes each. In total about 15 hours of read speech was collected. This speech data is split in two sets, a training set, used for the domain adaptation of the recognizer, and a held out set. The training set and held out set contain 101 speakers with 13.5 hours, and 9 speakers with 1.5 hours respectively.
Currently, we have literal (verbatim) transcriptions of two stories, and non-literal (the text they should read) transcriptions of the other 2 stories, both are transcriptions at word level. The vocabulary size of each of the 4 stories is 134, 214, 80, and 267 words respectively, and the coverage of the 4 stories together 543 words. There is not any OOV in the non-literal transcripts of the stories.

3. PRELIMINARY EXPERIMENTS

Since we have limited amount of training data, we need to adapt a baseline recognizer trained from a publicly available corpus. Based on the performance in initial experiments with Switchboard and Broadcast News systems, we picked Broadcast News system trained on 430 hours as a baseline recognizer. The system gives a word error rate of 21.6% on the 2004 Rich Transcription benchmark by NIST [1] , which is comparable to state-of-the-art for equivalent amounts of acoustic training data.

Stories	Unsupervised (%)	Semi-supervised (%)
Story1	35.8	8.6
Story2	36.3	3.5
Story3	31.7	4.5
Story4	42.4	5.9

Table 1. WER of the training data in the unsupervised and semi-supervised scenario.

We use the read speech from the enrollment phase to adapt the out-domain speech recognizer into our domain task by first applying MLLR adaptation technique [2] . The idea is to first get a rough adaptation with the MLLR technique (due to the limited amount of data only some of the context dependent phones from the AMs occur frequently in the stories) and then, refine it by applying MAP adaptation on top of the MLLR adaptation.
We examine three different adaptation scenarios.

- Unsupervised adaptation. The transcripts for the adaptation are automatically generated with the out-domain ASR.
- Semi-supervised adaptation. The non-literal transcripts of the stories are used to create an in-domain Language Model (LM). The transcripts used for the adaptation are obtained after decoding the utterances with the out-domain AM and in-domain LM.
- Supervised adaptation. Literal or verbatim transcripts are used for the adaptation of the AMs.

Table 1 shows the Word Error Rate (WER) of the training data per story after decoding the utterances with the general LM (unsupervised scenario) and in-domain LM (semi-supervised scenario). The decoding is performed in three stages using three successively refined acoustic models – a context-dependent model, a vocal-tract normalized model and a speaker-adapted MLLR model.

In our work in progress we perform an empirical comparison of an unsupervised, semi-supervised, and supervised Acoustic Model (AM) adaptation of the out-domain speech recognizer into the target domain. The training or adaptation set comprises the 2 stories for which we have literal transcriptions; story1 and story3. The held out set containing one of the stories, story2, does not overlap with the stories and speakers contained in the training set.
Then, we increase the size of the training set to 3 stories, and 4 stories to analyze the impact of the size of the adaptation data. The adaptation is run several times with different number of transformation matrices to get the optimal number. After getting the optimal MLLR-adapted AMs, the MAP adaptation will be added.

4. DISCUSSION

Recent studies show the potential of speech and language processing techniques on the automatic evaluation of the data collected in clinical tasks. However, in order to take a full advantage of these tools, they need to be adapted and optimized to the target domain.
Specifically, our work in progress evaluates different scenarios on how we can use the speech from an enrollment phase collected in a clinical task for the adaptation of an out-domain speech recognizer. The adaptation experiments presented here just form a baseline for subsequent more sophisticated adaptation techniques we are planning to pursue. The adapted speech recognizer will be eventually used to extract relevant linguistic features from the spoken language of the patients to detect depression symptoms.

5. REFERENCES

[1] FISCUS, F., GAROFOLO, J., LE, A., MARTIN, A., SANDERS, G., PRZYBOCK, M., AND PALLETT, D. 2004 spring nist rich transcription (rt-04s) evaluation data.

[2] LEGETTER, C., AND WOODLAND, P. C. Maximum likelihood linear regression for speaker adaptatio of continuous density hmm. *Computer Speech and Language* (1995), 9:171–185.

[3] PENNEBAKER, J. W., CROW, M. D., DABBS, J., AND PRICE, J. H. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, and Computers* (2001), 33: 517–523.