Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 93–109

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Getting more from automatic transcripts for semi-supervised language modeling☆

Scott Novotney[a,*], Richard Schwartz[a], Sanjeev Khudanpur[b]

[a] *Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA, USA*
[b] *Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA*

Received 26 November 2014; received in revised form 13 July 2015; accepted 13 August 2015
Available online 3 September 2015

Abstract

Many under-resourced languages such as Arabic diglossia or Hindi sub-dialects do not have sufficient in-domain text to build strong language models for use with automatic speech recognition (ASR). Semi-supervised language modeling uses a speech-to-text system to produce automatic transcripts from a large amount of in-domain audio typically to augment a small amount of manual transcripts. In contrast to the success of semi-supervised acoustic modeling, conventional language modeling techniques have provided only modest gains. This paper first explains the limitations of back-off language models due to their dependence on long-span *n*-grams, which are difficult to accurately estimate from automatic transcripts. From this analysis, we motivate a more robust use of the automatic counts as a prior over the estimated parameters of a log-linear language model. We demonstrate consistent gains for semi-supervised language models across a range of low-resource conditions.

Keywords: Language modeling; Automatic speech recognition; LVCSR; Low-resource

1. Introduction

Most automatic speech recognition tasks take language modeling training data for granted. Voice search benefits from trillions of tokens of domain matched web queries. Broadcast news is well matched to newswire text and closed captions. However, for many other languages or domains, such as Arabic or Hindi diglossia, the only useful source of training data is expensive and time consuming in-domain manual transcription. While electronic resources may exist in the language, the wide gap in both vocabulary and word frequency limits the use of available newswire or web text for these conversational languages. Deploying a large vocabulary continuous speech recognition (LVCSR) system for an under-resourced language will require large investments in human labor and cost.

One hope for overcoming this deployment burden is semi-supervised estimation of the component models of an LVCSR system: a small amount of in-domain transcripts are used in conjunction with a (typically) large amount of unlabeled audio. We assume that for any task that requires automatic speech recognition, there must be an abundance

☆ This paper has been recommended for acceptance by Murat Saraclar.

* Corresponding author. Tel.: +1 443 452 8575.
E-mail addresses: snovotne@bbn.com (S. Novotney), schwartz@bbn.com (R. Schwartz), khudanpur@jhu.edu (S. Khudanpur).

http://dx.doi.org/10.1016/j.csl.2015.08.007
0885-2308/

of audio in need of transcription. This audio has the potential to usefully augment the small amount of in-domain transcripts. The success of semi-supervised acoustic modeling demonstrated that as little as 1 h of manual transcripts was sufficient to deploy an effective ASR system in a new domain (Ma and Schwartz, 2008). Yet the other half of the speech equation, language modeling, has not significantly benefited from semi-supervised methods. This article will explain the cause for limited previous successes and propose a framework for better exploiting available audio.

Ideally, training corpora for language model estimation should be **in-domain**, **copious** and **accurate**. When one of the three are absent, then we are tasked with low-resource language modeling. Initial language modeling research assumed a small amount of **in-domain** and **accurate** transcripts were available. For instance, the Brown corpus (Kucera and Francis, 1967) is "only" 1M tokens. These corpora led to improvements in smoothing of unseen events with techniques such as Kneser-Ney smoothing (Kneser and Ney, 1995) outperforming simple methods such as absolute discounting. When more electronic corpora became prevalent, language modeling work considered **copious**, **accurate**, but no longer in-domain corpora under the umbrella of domain adaptation. The final combination of the three desiderata is a **copious** amount of **in-domain**, but now inaccurate, transcripts.

This resource condition can arise when an automatic speech recognizer (or in-expert human transcriber (Novotney and Callison-burch, 2010)) produces transcripts with high error. Trained on a small amount of in-domain transcripts, the LVCSR system can quickly and inexpensively produce a large amount of n-gram counts from freely available in-domain audio. However, at higher error rates, the majority of these n-grams (as measured by type and token) are incorrect. As Section 3 will explore, the smoothing techniques and back-off language models standard in the research community are ill equipped to deal with such errors. The methods require accurate counts of the highest order n-grams (typically trigram) and are unable to adapt to accurate lower-order statistics. In low resource setting such as colloquial dialects, such resources may be scarce. Section 4 will motivate the use of a log-linear (*a.k.a.* maximum entropy) language model and a probabilistic use of automatic transcripts as a prior for adaptation.

In this article, we make the following conclusions: (I) Back-off language models are poorly suited for learning from noisy transcripts. (II) Accurate higher-order n-gram counts are critical for semi-supervised language modeling. (III) Using automatic transcripts as a prior for language model adaptation results in a robust model, which gives consistent reductions in WER even when the available corpus is small.

1.1. Prior work

Prior work on semi-supervised language model estimation initially considered only back-off language models. The first reference in the literature decoded 17 h of call center data with an initial model built from voicemail transcripts (Bacchiani and Roark, 2003). Unweighted n-gram counts from the automatic transcripts (at 20% WER) were used for MAP adaptation of a back-off language model. This resulted in a 4% absolute reduction in WER. However, self-adaptation of the test data did not result in further gains.

Other work also reported limited success for self-adaptation. Call center data was again the target domain for adaptation (Gretter and Riccardi, 2001). n-gram expected counts were estimated from lattice posteriors, with all posterior scores below a threshold mapped to a common "unknown" token. This form of count thresholding recovered 0.8% absolute WER of the 2.2% possible.

Semi-supervised language modeling was successfully combined with active learning (Nakano and Hazen, 2003). Experiments with an interactive dialog corpus used confidence estimates to select the most accurate utterances for inclusion in language modeling estimation. The least likely utterances were then manually transcribed, with a net savings in manual transcription cost and modest reduction in WER.

Work in *discriminative* language modeling has used ASR output to find confusion neighborhoods (Xu et al., 2009). Lattices from Broadcast news audio were collapsed into confusion bins, which were then used to create sets of word confusions. These were applied to newswire text to create artificial training for the discriminative model. Other work used unsupervised estimates of hypothesis error rates for discriminative updates (Dikici and Saraclar, 2014). This achieved half of the gain possible with fully supervised transcripts.

Log-linear language models were originally introduced to the NLP community as a method of self-adaptation of low-resource models (Della Pietra et al., 1992). The target model was adapted to unigram counts from one-best output for dictation. More commonly known in the speech literature as maximum entropy models (Rosenfeld, 1996), log-linear models are a flexible framework for incorporating a wide variety of features beyond n-grams. The closest work to this paper has applied Bayesian techniques to log-linear models for low-resource language modeling. Experiments with