Speech Communication 93 (2017) 20–30

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Transcribing against time

Matthias Sperber[a,*], Graham Neubig[b], Jan Niehues[a], Satoshi Nakamura[c], Alex Waibel[a]

[a] Karlsruhe Institute of Technology, Germany
[b] Carnegie Mellon University, USA
[c] Nara Institute of Science and Technology, Japan

ARTICLE INFO

Article history:
Received 27 July 2016
Revised 20 July 2017
Accepted 31 July 2017
Available online 2 August 2017

Keywords:
Speech transcription
Error correction
Cost-sensitive annotation
User modeling

ABSTRACT

We investigate the problem of manually correcting errors from an automatic speech transcript in a cost-sensitive fashion. This is done by specifying a fixed time budget, and then automatically choosing location and size of segments for correction such that the number of corrected errors is maximized. The core components, as suggested by previous research (Sperber, 2014c), are a utility model that estimates the number of errors in a particular segment, and a cost model that estimates annotation effort for the segment. In this work we propose a dynamic updating framework that allows for the training of cost models during the ongoing transcription process. This removes the need for transcriber enrollment prior to the actual transcription, and improves correction efficiency by allowing highly transcriber-adaptive cost modeling. We first confirm and analyze the improvements afforded by this method in a simulated study. We then conduct a realistic user study, observing efficiency improvements of 15% relative on average, and 42% for the participants who deviated most strongly from our initial, transcriber-agnostic cost model. Moreover, we find that our updating framework can capture dynamically changing factors, such as transcriber fatigue and topic familiarity, which we observe to have a large influence on the transcriber's working behavior.

1. Introduction

High quality speech transcripts are required in many different tasks, including for example web-based lecture archives, training data for automatic speech recognition (ASR), and input to downstream applications such as translation. Unfortunately, in realistic settings automatically created transcripts often contain too many errors to be useful as-is, and human annotators must be employed to improve their quality. This manual transcription is costly and time-consuming.

Previous works have attempted to improve the efficiency of manual supervision for speech transcription by dividing the speech into small segments that are convenient to transcribe (Roy and Roy, 2009), and choosing low-confidence segments of an ASR transcript that are more likely to contain errors (Sanchez-Cortina et al., 2012; Sperber et al., 2013). Studies on cost-sensitive annotation have also shown that to maximize supervision efficiency, it is important to consider not only the number of errors that might be contained in a particular segment, but also the human supervision effort involved in correcting them (Settles et al., 2008; Tomanek and Hahn, 2010; Ramirez-Loaiza et al., 2014). Recent work has shown that supervision efficiency can be further increased by training models to estimate the transcriber effort and potential error reduction of each segment, and then explicitly optimizing the location of segment boundaries (Sperber et al., 2014c).

However, this use of models to predict transcriber effort has a downside: these models need to be trained. Thus, before starting the main transcription task, it is necessary to perform enrollment, where the transcriber annotates a certain amount of randomly selected enrollment data, which is then used to train the predictive cost models. However, enrollments are time-consuming, costly, and may be impractical; for example, in a crowd sourcing situation. In addition, these predictive models remain static and cannot account for dynamically changing factors such as the annotator's topic familiarity, fatigue, or increasing experience.

In this work, we introduce a framework that removes this need for enrollment by updating cost models and corresponding choice of data to annotate on-the-fly (during the ongoing transcription process).[1] This framework allows us to work within a fixed time limit for manual correction of a transcript or a series of transcripts.

* Corresponding author.
E-mail address: matthias.sperber@kit.edu (M. Sperber).

http://dx.doi.org/10.1016/j.specom.2017.07.006
0167-6393/

[1] We have previously presented the basic idea of this method and verified it through simulation in the proceedings of SLT2014 Sperber et al. (2014a). In this paper, we expand the description and add a full user study with 12 expert and non-expert participants.

We first start with a general, initial cost model used to compute a first selection of segments for correction. During the ongoing transcription process, the cost model is gradually improved and adapted toward the particular transcriber, and the choice of segments is updated to reflect both the updated cost model and the actual remaining time at a particular point. Locations and lengths of segments to annotate are chosen to optimize annotation efficiency, as proposed in previous work (Sperber et al., 2014c).

A number of challenges arise in this proposed dynamic annotation framework. (1) A suitable time limit must be chosen and incorporated into our framework as a stopping criterion. (2) Annotation effort is difficult to predict, because there are large differences not only between transcribers, but also for a single transcriber between different segments. Accurate cost modeling is important both for selecting suitable segments that are efficient to annotate, and for the amount of selected segments to be appropriate such that the time budget is not over- or underspent. (3) Choosing segments to annotate is a computationally difficult problem, but needs to be done quickly, as we desire to update the segmentation during the ongoing transcription process.

Our solutions to these challenges are as follows: (1) We propose to limit the transcription time budget to a fixed value, in order to reflect one's desired cost-quality tradeoff. How to choose a suitable time budget is task specific, but we argue that it is more practical than configuring a confidence threshold as was required in some previous works (Sanchez-Cortina et al., 2012; Sperber et al., 2013; Valor Miró et al., 2015). Moreover, by periodically updating segmentations (Section 4) we can recover from inaccurately predicted correction times, making sure that the budget is not over- or underspent. (2) We propose an approach of starting with an initial, general cost model and gradually adapting it toward the particular transcriber (Section 5). This has the potential to remove the need for enrollment, while being able to model transcriber-specific as well as dynamically changing characteristics. (3) We propose a new, more efficient algorithm for choosing which segments of the ASR transcript to annotate using the penalty method (Section 6). We demonstrate this algorithm to be fast enough to update segmentations during the transcription process, without degrading quality compared to the original, much slower algorithm (Sperber et al., 2014c).

We first conduct a partly simulated evaluation approach (Section 7). Results show that our method outperforms both cost-insensitive baselines and cost-sensitive baselines without updates. An analysis shows that efficiency gains are attributed to (1) increasingly accurate cost models, (2) adjusting to the actual remaining time budget with each update, and (3) choosing a sensible (although crude) initial cost model that includes cognitive overhead. Finally, we conduct a realistic user study (Section 8) in a typical scenario where several previously unknown transcribers conduct only a limited amount of work. We notice large differences between different transcribers, supporting our claim that transcriber-specific modeling is crucial. Moreover, we observe relative productivity gains of 15% on average, and 42% for those participants who deviated most from the initial cost model. The gains of using our updating framework were especially strong in the case where the initial ASR transcript was already of relatively high quality.

2. Related work

Adaptive user models have been studied in the human computer interaction community (Zigoris and Zhang, 2006), with very different requirements from ours. We are not aware of previous works on adaptive user modeling or choice of data to annotate in the context of cost-sensitive annotation or computational linguistics in general. A closely related work on quality estimation (de Souza et al., 2015) adapts an automatic quality estimator online and in a multi-task fashion, as more and more in-domain samples are annotated over time.

Efficient supervision strategies have been studied across a variety of NLP-related research areas, and received increasing attention in recent years. Examples include post editing for speech recognition (Sanchez-Cortina et al., 2012), interactive machine translation (González-Rubio et al., 2010), active learning for machine translation (Haffari et al., 2009; González-Rubio et al., 2011) and many other NLP tasks (Olsson, 2009), to name but a few studies. Most of these do not model annotation cost explicitly. However, it has been recognized that correcting only the instances of highest utility is often not optimal in terms of efficiency, since these parts tend to be the most difficult to manually annotate (Settles et al., 2008; Miura et al., 2016). As a solution, the idea of using an annotator cost model to predict the supervision effort has been developed (Settles et al., 2008; Tomanek et al., 2010; Specia, 2011; Cohn and Specia, 2013; Sperber et al., 2014c), which inspired our approach as well. Note that these previous works estimate static, annotator-specific cost models via enrollment, whereas our dynamic approach does not require enrollment.

Some studies have addressed the problem of balancing utility and cost in the context of active learning. A greedy approach to combine both into one measure is the "bang-for-the-buck" approach (Settles et al., 2008), where utility divided by effort is used as a per-instance efficiency measure. Such an approach can be effective for selecting isolated instances for annotation, but is problematic when selecting segments that can overlap and conflict with one another, as in our task. A more theoretically founded scalar optimization objective is the net benefit (utility minus costs) as proposed by Vijayanarasimhan and Grauman (2009), but unfortunately is restricted to applications where both can be expressed in terms of the same monetary unit. Vijayanarasimhan et al. (2010) and Donmez and Carbonell (2008) use a more practical approach that specifies a constrained optimization problem by allowing only a limited time budget for supervision, similar to our approach. Note that works on annotation apart from the active learning community have usually assumed stopping criterions based on confidence thresholds (Sanchez-Cortina et al., 2012; Sperber et al., 2013; Valor Miró et al., 2015), which measure only relative improvement and may not be intuitive to configure in practice.

3. Background: static segmentation

Our advocated "transcribing against time" paradigm builds upon the following three lines of work:

- Roy and Roy (2009) argue that human effort should be spent on transcription of speech, not its segmentation. Segmentation is time-consuming if performed manually, and can be done reliably in an automatic fashion. Segmentation is also a very important step, because suitable segment size leads to increased transcription efficiency.
- Rodriguez et al. (2007) propose the computer-assisted transcription (CAT) approach, in which a transcription is first created automatically, and then corrected manually where necessary. In their approach, a human checks the complete transcript and performs corrections whenever errors are found.[2]
- Sperber et al. (2014c) argue that the decision which segments to correct and which not, should also be done automatically to reduce human effort. They show how to select segment locations and sizes according to a desired utility-cost tradeoff,

[2] CAT also improves the automatic transcript on-the-fly using the transcriber's corrections. This is not the focus of our work, but could be integrated into our proposed updating framework.