Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 45 – 52

Procedia Computer Science

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

Developing Speech Resources from Parliamentary Data for South African English

Febe de Wet*, Jaco Badenhorst, Thipe Modipa

Human Language Technology Research Group, CSIR Meraka, South Africa

Abstract

The official languages of South Africa can still be classified as under-resourced with respect to the speech resources that are required for technology development. Harvesting speech data from existing sources is one means to create additional resources. The aim of the study reported on in this paper was to improve the harvesting and transcription accuracy of a corpus derived from parliamentary data. This aim was achieved by improving on the text normalisation process and pronunciation modelling as well as by iteratively training more accurate in-domain acoustic models. In this manner, more data could be harvested with higher confidence than using baseline pronunciation dictionaries and out-of-domain speech data.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Under-resourced languages; speech data; South African English; automatic alignment.

1. Introduction

In terms of the data that is required to develop Human Language Technology (HLT), the official languages of South Africa can still be classified as being highly under-resourced. Although the situation has been addressed to some extent, much remains to be done in terms of both language and speech resource and technology development.

Previous attempts to compile speech resources for the development of automatic speech recognition (ASR) technology involved extensive data collection efforts [1,2]. Participants were recruited and their speech was recorded during dedicated data collection campaigns. While these projects resulted in new corpora being developed, data collection was an extremely resource-intensive process. Subsequent speech data collection efforts therefore tried to exploit existing sources of recorded speech to create new resources [3,4].

The study reported on in this paper elaborates on a previous investigation on the possibility to automatically align speech data from South Africa's National Parliament (SANP) with the associated transcriptions [4]. Initial alignment

* Corresponding author. Tel.: +27-12-841-4303 ; fax: +27-12-841-4720.
E-mail address: fdwet@csir.co.za

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.028

was performed using models that were derived from an extended version of NCHLT English corpus[5]. An alignment procedure similar to the one described by Moreno *et al.*[6] was implemented to identify the audio segments that could be aligned most accurately. This system and the corresponding alignments will be referred to as the *Baseline* in the rest of this paper.

The current work aimed to improve the accuracy of the alignments and to compile a bigger usable corpus from the source data. This aim was achieved by using the initial alignments provided by the *Baseline* system to develop in-domain acoustic models. Acoustic modelling was enhanced by improving the quality of the text normalisation process and by creating more accurate pronunciation dictionaries.

2. Background

Prior to the development of the *Baseline* system, a comparative study on how HLTs are being used to provide support to the language units in the parliaments of other countries was compiled. The countries included in the study were India, Australia, Canada, Czech Republic, Denmark, Isle of Man, Japan, Europe and United Kingdom. Speech-to-speech translation[7], machine translation (MT), automatic transcription and manual transcriptions were investigated as examples of HLTs. The survey found that ASR is the most widely used HLT in parliamentary applications and that ASR works particularly well in monolingual or bilingual countries[8,9,10,11,12]. In a few instances, MT or a combination of MT and ASR is used. However, operational applications of MT are limited to text translation, no speech-to-speech translation systems have been deployed yet.

Given these observations, it was decided to start by developing resources that could be used to create ASR technology to support the transcription unit at the SANP. Since all the 11 official languages of South Africa are spoken at National Parliament, the initial aim of the project was to compile a multi-lingual speech corpus and to develop ASR systems for all 11 languages. However, the speech data from the National Assembly used in this study revealed that South African English (SAE) is predominantly spoken in the debates. The other official languages are only used incidentally. The project was subsequently re-scoped to create the resources required to develop ASR technology for SAE.

Technically, the work reported on here elaborates on a previous study by Kleynhans & de Wet[4] which, in turn, was based on different aspects of the data harvesting techniques proposed by Moreno *et al.*[6] and Davel *et al.*[3]. An important difference between the current and previous work is that the alignments produced during the previous study[4] were available to fast track the development of in-domain acoustic models. This study could therefore focus on identifying the most reliably aligned segments and using them in combination with improved transcriptions and pronunciation dictionaries to enhance the quality of the alignments.

The next section introduces the speech and text data that was used during the investigation. The automatic alignment system is presented in Section 4. Data selection, pronunciation dictionaries and in-domain acoustic models are described in Section 4.1. The system configurations that were experimented with are introduced in Section 4.2 and the evaluation procedure that was used to assess their performance is presented in Section 4.3. Results are presented in Section 5, followed by a discussion in Section 6.

3. Resources

3.1. Speech data from National Parliament

Video recordings of 32 debates that took place in the National Assembly was obtained from the SANP. The audio was extracted from the videos and converted to PCM WAVE as discussed in[4]. The average length of the resulting audio files was 3 hours. The 32 debates corresponded to around 105 hours of audio data, but approximately 27 hours was found to be unsuitable for harvesting[4].

3.2. Hansard text data

Parliamentary proceedings are transcribed in Hansard format. This format specifies how speaker turns, acoustic events other than speech, the use of different languages, etc. should be captured in the transcriptions. A total of