Abstract

Code-switching refers to the alternate use of two or more languages (or dialects) during the conversation. This phenomenon has been observed in many multilingual communities across the globe. Therefore, handling code-switching by the spoken input systems is very much required for efficient human-machine interaction. However, due to the lack of domain-specific resources, the research in this domain is somewhat limited compared to the monolingual case. This thesis aims to address the acoustic and language modeling challenges in code-switching automatic speech recognition (ASR) tasks. In addition to that, a Hindi-English code-switching corpus has been created towards addressing the data scarcity issue.

The early works on code-switching ASR happen to employ the hybrid framework typically devel- oped for the monolingual case. The created Hindi-English code-switching corpus is first evaluated in the hybrid framework. The hybrid framework comprises of three sub-modules, namely, a pro- nunciation model, an acoustic model, and a language model. The end-to-end (E2E) framework has recently emerged as a viable alternative to the hybrid systems in the ASR domain. Unlike the hybrid framework, the E2E framework does not require the phonetically labeled training data, and also does not include any explicit pronunciation model. In the case of code-switching ASR, for multiple languages being involved, these attributes become more attractive. Motivated by that, in this thesis, the E2E framework has been explored for developing the code-switching ASR systems.

In the existing code-switching E2E ASR works, the target set is derived by merely combining the character sets of the languages involved. Such systems would suffer from high confusability among the cross-language targets due to the broad acoustic similarity among sound units involved in the code-switching language pairs. To avoid such a confusability, a common phone set cover- ing the underlying languages in code-switching is defined and used as the reduced target set for training the models. Interestingly, the reduced target set based E2E ASR system outperformed the combined target set one in terms of the target error rate (TER). But, a reverse trend was noted when those target sequences were converted to word sequences, i.e., for computing the word error rate (WER). This degradation in WER is because of the enhanced confusability among the homophones (the words having identical pronunciation but different spellings) within or across the languages involved. For addressing the same, a context-dependent target-to-word (T2W) transduc- tion scheme has also been proposed, which employs an explicit error model and a language model. The proposed T2W transduction scheme is noted to achieve a relative improvement of 22% over the naive transduction scheme in the context of Hindi-English code-switching E2E ASR. Further, to enhance the context information in the code-switching data, a novel textual feature referred to as the code-switching location (CSL) feature and a modified parts-of-speech (POS) tagging scheme have also been proposed. On evaluating these features by incorporated into factored language model (FLM), a significant reduction in perplexity score has been noted. With the use of these FLMs in the proposed T2W transduction scheme, a further improvement in the WER score is achieved. The proposed system outperforms the existing one and yields a TER of 18.1% along with a WER of 29.79% on the created Hindi-English code-switching corpus. Despite the proposed approaches being evaluated for the Hindi-English code-switching case, they are generic enough to be applied for any other code-switching context.

-Sreeram Ganji