Summary report for project

Exploiting Language Information for Situational Awareness (ELISA)

For year 2016

Submitted to Information Sciences Institute, University of South California, USA

By Brno University of Technology

Lead author: Dr. Ondřej Glembek

The BUT effort was concentrated on building ASR systems for the November 2016 Evaluations. The basic principle was to build a robust ASRs and use their output as an input to the existing MT (ISI) engines and further to the SF classifiers (USC).

Speech Evaluations - Mandarin

The BUT Mandarin ASR system is GMM-HMM-based with cross-word tied-states triphones, trained using the MPE criterion. The feature extraction was based on concatenation of two feature streams: PLP-HLDA, and Stacked Bottle-neck Neural Network (SBN). The system was trained on the HKUST Mandarin Telephone Speech (70h of speech).
The language model was built by merging the HKUST transcriptions (~1e6 characters), gazetteers collected by RPI for the summer evals (~2e5 characters), and the Gigaword Mandarin corpus (~2e9 characters), and Mandarin Wiki data (~3.5e8 characters). These corpora were shared with USC to extend the GALE Mandarin system.
Both the BUT and the USC ASR systems were part of the primary submission.

Speech Evaluations - Uyghur

ASR System

We used the native informant to perform two tasks: i) read aloud sentences (selected from the dev data of the summer text eval), and ii) transcribe randomly selected excerpts of the Uyghur DEV set. This way we compiled four sets of transcribed data (in the two sessions): R1, T1, R2, T2. We used T1 (33 sentences in approx 2.5 minutes of audio) to evaluate the systems and we used R1, R2 and T2 (~1h of data + various vocal and speed perturbations) for training the Uighur ASR in the primary submission. We have also added Uighur-transliterated Uzbek transcribed data from the Uzbek DEV set (492 utterances, 4 hours). We used PLP and MultiLingual RDT features as an input to a 2-layer acoustic modelling DNN. We used the summer eval text corpora to build the LM.
We have also build an Uzbek system, assuming the languages are mutually intelligible and thus sharing some SF-modelling compatible strings. We have submitted a system based on this ASR as one of the contrastive systems.

N-gram Based System

Based on the fact that Uighur and Uzbek languages are of the same family and (to a large extent) mutually intelligible, we have built a situation-type classification system based on phone n-grams obtained by decoding the Uzbek Development speech using unsupervised acoustic unit discovery models (trained using multi-lingual bottleneck features), extracting 3-grams and training the Naive Bayes classifiers. This system was then used to classify Uyghur evaluation data.

Our AUD model was a Bayesian phone-loop having prior over the weights of the possible phones and the parameters of the emissions probabilities. The prior over the weights was a Dirichlet Process leading to a model which, in theory, has an infinite number of phones. Learning was done using the Variational Bayes framework. More specifically, the inference estimates the posterior distribution over the emissions' parameters and also the "effective" number of phones needed to model the data. The model was initialized on forced alignments from Babel Turkish data set.

This system, however, was not used in the final evaluation due to an error causing missing scores for some SF types.