Summary report for project

Exploiting Language Information for Situational Awareness (ELISA)

For year 2015

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

By Brno University of Technology

Lead author: Dr. Ondřej Glembek

Topic Discovery

In this part, we used the provided Turkish audio data to perform the task of topic discovery. We split the data into training and test sets. Due to the limited amount of the data, we created three setups based on the amount ratio of the training/test data: 30/70, 50/50, 70/30.

Phonetic approach

In the first set of experiments, we used a phoneme recognizer to extract the phoneme probability networks---the lattices, out of which we extracted 1-best phoneme sequences. For this purpose, we built a Turkish phoneme recognizer to see, how well we could do if we did have training data for the incident language. Out of the 1-best phone-sequences, we computed the the n-gram statistics. In our case, we used the context of 2 preceding phonemes, i.e. we used tri-gram counts for the topic classification.

These tri-gram counts were then used as an input to our classifier (the backend). We experimented with:

1) TF-IDF with Naive Bayes (NB)
2) TF-IDF with Support Vector Machine (SVM)
3) Subspace Multinomial model with SVM

		Classification accuracy in %		
Train	Test	TF-IDF + NB	TF-IDF + SVM	SM + SVM
30%	70%	35.40	22.12	22.12
50%	50%	39.51	40.74	32.10

70%	30%	34.78	39.13	30.43

While listening to the provided data, we observed some interesting facts about the nature of the data set.

- Most of the files have some English speech segments in between.
- Audio files related to topic "Hydrological / Meteorological" have lot of wind noise in the background.
- Audio files related to topic "Civil Unrest" have group of people shouting (background and some times foreground).
- Audio files related to "Terrorism" have lot of gun firing, and sirens in the background.

These facts led us to the usage of “acoustic approach”, i.e. use only acoustic information to classify the topics---similar to e.g. speaker or language recognition.

Acoustic approach

We used the BUT standard Speaker Recognition *i-vector* system (the Voice Biometry Standardization initiative system) to extract a single i-vector per audio recording. This system has been proven to extract i-vectors that are suitable not only for Speaker Recognition, but also for tasks of Language Recognition, Emotion Detection, Age Estimation, LVCSR speaker adaptation, etc. The i-vector was then tak as an input for the SVM backend in a similar way as in the phonetic approach.

Train set	Test set	Accuracy [%]
30% (58)	70% (125)	27.2
50% (94)	50% (89)	29.2
70% (133)	30% (50)	36.0

The i-vector extractor has been optimized for telephone speech and its parameters have been trained on the NIST Speaker Recognition 2004-2008 telephone data, the Fisher English database, and the Swtichboard and Switchboard Cellular data.