Summary report for project

Machine Translation for English Retrieval of Information in Any Language (MATERIAL)

For year 2017

Submitted to Raytheon BBN Technologies

By Brno University of Technology

Lead author: Dr. Martin Karafiát

Automatic Speech Recognition

MATERIAL target data significantly differs from BUILD set (telephone conversations). Mainly Topical Broadcast (TB) and News Broadcast (NB) data because speakers talk with different speaking style, use different channel and have different conversational topic.

Improvements of Acoustic and Language models are necessary for good performance. First, we experimented with telephone channel simulation instead of simple down-sampling but no improvement was observed, therefore we stick with down-sampling.

Next, we incorporated Machine Translation (MT) and Wikipedia data into Swahili LM. It required various text normalizations:

• Expansion of digits

• Detection of acronyms (all-caps up to 2-4 letter)

• Conversion to lowercase, removal of punctuation

• Filtering by set of graphemes from 'build_train',

• New LM contained 88k of unigrams, 1,180k bigrams and 433k trigrams which was significant boost from original one trained on BUILD set transcriptions only (25k unigrams, 131k bigrams and 16k trigrams). Interpolation weights was estimated on ANALYSIS1 set (‘build_train’ - 0.48, 'build_MT' - 0.28, ‘wiki' - 0.24).

The following tables show results on Swahili DNN system. Over 10% gain on Out-Of-Vocabulary (OOV) and Word Error Rate (WER) was reached by adding new data.

LM	CS - %OOV	TB - %OOV	NB - %OOV	Overall
BUILD	9.5	17.9	21.0	16.8
BUILD+MT+Wiki	8.1	7.2	5.6	7.0

LM	CS - %WER	TB - %WER	NB - %WER	Overall
BUILD	45.5	75.0	69.6	67.2
BUILD+MT+Wiki	45.3	67.1	48.4	57.9

News Broadcast (NB) and mainly Topical Broadcast (TB) contains a lot of background noise.

Therefore, we experimented with data enhancement based on denoising auto-encoder trained on telephone data. We show the results in the following table.

Enhancement	CS - %WER	TB - %WER	NB - %WER	Overall
None	45.3	67.1	48.4	57.9
Auto-encoder	46.0	61.4	46.0	54.4

A small degradation was observed on Conversational Speech (CS) but on Topical Broadcast we obtained a 5.7% gain.