Bi-monthly report

Yang Wei

GROUPING

Center for Speech and Language Technologies

Outline

- Finished work
- Schedule
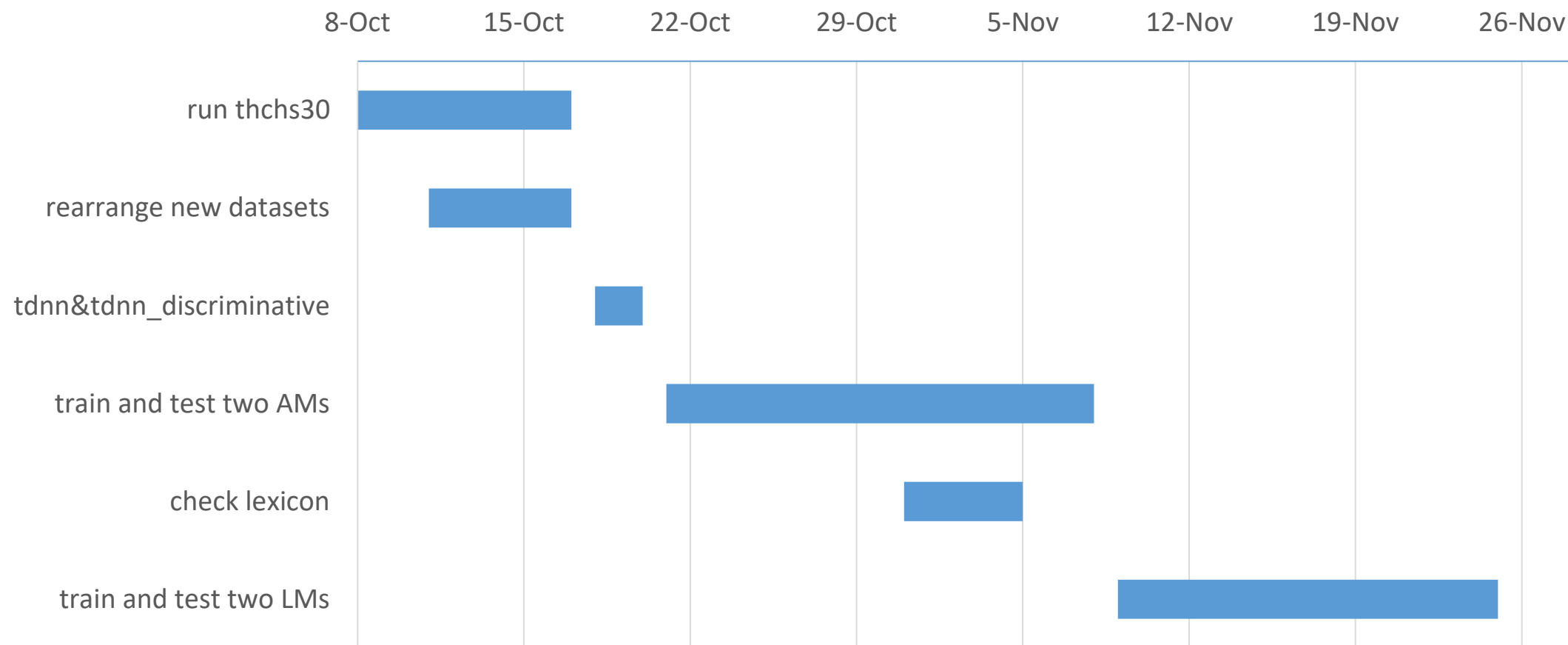- Detail
- Technology

Finished work

- For teaching
  - Run thchs30 to get familiar with the process.
  - Rearrange new datasets and decode with them.
  - Train and test with tdnn and tdnn_discriminative.
  - Train and test two AMs for comparison.
- For engineering
  - Check lexicon from Tonghuashun.
  - Train and test two LMs for comparison.

Schedule

Detail

- thchs30
  - Train/Decode
    - mono
    - tri1
    - tri2b
    - tri3b
    - tri4b
    - tri4b_dnn

Detail

- New datasets&decode
  - Rearrangement
    - wav.scp
    - text
    - spk2utt
    - utt2spk

Detail

- New datasets&decode

	exp/mono/decode_test_word/wer_19_0.0	exp/tri1/decode_test_word/wer_10_0.0	exp/tri2b/decode_test_word/wer_10_0.0	exp/tri3b/decode_test_word/wer_10_0.0	exp/tri3b/decode_test_word.si/wer_9_0.0	exp/tri4b/decode_test_word/wer_11_0.0	exp/tri4b/decode_test_word.si/wer_10_0.0	exp/tri4b_dnn/decode_test_word/wer_7_0.0
RESULTS	50.88	35.97	32.14	29.47	33.65	28.07	31.50	23.68
decode	50.70	36.09	32.22	29.61	33.46	28.10	31.51	24.54

Detail

- tdnn&tdnn_discriminative
  - Advantages
  - Adjustment

Detail

- AM training
  - Train
    - train_myhexin_20161019
    - train_100h
  - Test
    - LM
      - Hunhe_zhongzi_and_add_and_PPL_5yuan_5e14.lm.utf8.1e-5.3gram.gz
      - lm.3gram.1e-5
      - lm.3gram.1e-7
    - testset
      - test_myhexin_20161019
      - test_1000ju   test2000ju

Detail

- AM training

wer	train_myhexin_20161019			train_100h		
	test_myhexin_20161019	test_1000ju	test_2000ju	test_myhexin_20161019	test_1000ju	test_2000ju
lm_default*	16.95	54.65	65.59	28.51	19.04	33.58
lm.3gram.1e-5	10.91	62.76	69.81	17.64	27.20	40.43
lm.3gram.1e-7	8.51	58.73	64.91	13.11	19.80	33.16
lm_default=Hunhe_zhongzi_and_add_and_PPL_5yuan_5e14.lm.utf8.1e-5.3gram.gz						

Detail

- LM training
  - Dataset
    - 40G finance corpus
    - 64G universe corpus
  - Preprocess

split → clean → segment

Detail

- LM training
  - Train
    - 14 finance files
    - 22 universe files
  - Test
    - AM
      - …/tonghuashun/v0.1/am
    - testset
      - test_myhexin_20161019
      - test_1000ju   test2000ju

Detail

- LM training

wer	test_myhexin_20161019	test_1000ju	test_2000ju	
fin_3gram_1e-7	6.33	36.76	45.17	
uni_3gram_1e-7	8.05	27.78	38.08	
hybrid_3gram_1e-7	6.45	28.89	39.66	
given_finance*	6.47	37.51	45.77	
given_universe*	12.23	20.79	31.77	
given_finance=lm.3gram.1e-7				
given_universe=Hunhe_zhongzi_and_add_and_PPL_5yuan_5e14.lm.utf8.1e-7.3gram.gz				

Technology

- Linux
- Python
- Shell

Thanks a lot