Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Maria Obedkova
Název práce	Data-driven Pronunciation Generation for ASR
Rok odevzdání	2019
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Mgr. Nino Peterek, Ph.D.	**Role**	Oponent
Pracoviště	ÚFAL		

Text posudku:

The author of the thesis Maria Obedkova explores automatic phonetic dictionary generation, promising field of recent speech recognition research. Phonetic dictionaries are mostly based on work of phoneticians, who analysed many pronunciation variations of each word and selected the most representative variant (and for a couple of words more than one). Another type of phonetic dictionaries (G2P) is using grapheme correspondence.

Maria Obedkova compares these traditional methods with recent data-driven algorithms based on large data and machine learning. There is mentioned successful article with fully data-driven phonetic dictionary generation, but it is using own generated acoustic elements instead of phones.

The main idea of the thesis is an automatic extension of classical phonetic dictionary by pronunciation variants through machine learning methods.

The approach of Maria Obedkova is based on mapping of MFCC speech vectors of words utterances to representative I-vectors (with the help of twin recurrent neural networks), finding clusters of these I-vectors, and guessing the phonetic transcription of the clusters by Viterbi phonetic recognition (decoding). Clusters than should represent new pronunciation variants.

Training data consist of 106.1 hours of English speech and test data of 0.6 hour. They come from VoxForge corpus, which covers many English dialects. The training part is used to prepare triphone KALDI acoustic models for finding word boundaries, serves as the source of MFCC vectors for I-vector mapping and for training of NNET3 acoustic models (for Viterbi phonetic decoding and for final ASR experiments).

All generated pronunciation dictionaries are finally used for KALDI models training, testing and finally compared with classical CMU phonetic dictionary recognition results. The comparison showed, that the best automatic method could achieve results 1,5 % worse than 9,47% WER of classical phonetic systems . Also G2P lexicon has better WER of 9.78%. The cluster method (without discriminative training based on phone edit distance) has WER of 19.99%.

Maria Obedkova guesses, that the problem is in word embedding (fix-sized I-vector reduces to much word's MFCC vectors) or in weak Viterbi phonetic decoding (that is also my guess, because phonetic recognisers are usually weak without grammar and lexicon information).

The master thesis is written in clear form, with adequate introduction to ASR and phonetic dictionary generations. There are many links to articles and books, showing good preparation for the theory and experiments. Source codes are good commented, using

accelerated GPU libraries in Tensorflow framework, showing understanding of the theory and good programming experience. The scripts have hard-coded file paths from experiment environment, but that was not trouble by source code reading.

Maria Obedkova made a lot of perfectly structured experiments with precise evaluation. The goal was to beat the classical phonetic dictionary, that was not achieved, but showed many new ASR setups and hope for future improvements, that are described in the conclusion chapter of the thesis.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 31.8.2019 **Podpis**