Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit (RLAT)

Ngoc Thang Vu, Tim Schlippe, Franziska Kraus, Tanja Schultz

thang.vu@kit.edu

Cognitive Systems Lab

Karlsruhe Institute of Technology

1. Overview

- LVCSR systems for five Eastern European Languages such as Bulgarian, Croatian, Czech, Polish, and Russian using Rapid Language Adaptaion Toolkit (RLAT)
- Crawling and processing large quantites of text material from the Internet
- Strategy for language model optimization on the given development set in a short time period with minimal human effort

2. Slavic Languages and data resources

- Well known for their rich morphology, caused by a high reflection rate of nouns using various cases and genders
  Ex.: nowy student, nowego studenta, nowi studentci
- Text corpus size for five Eastern European languages

Languages	Website	#Tokens	#Types
Bulgarian	dariknews.bg	302M	560K
Croatian	www.hrt.hr	124M	248K
Czech	lidovky.cz	790M	1.25M
Polish	wiadomosci.wp.pl	347M	815K
Russian	www.rian.ru	565M	1M

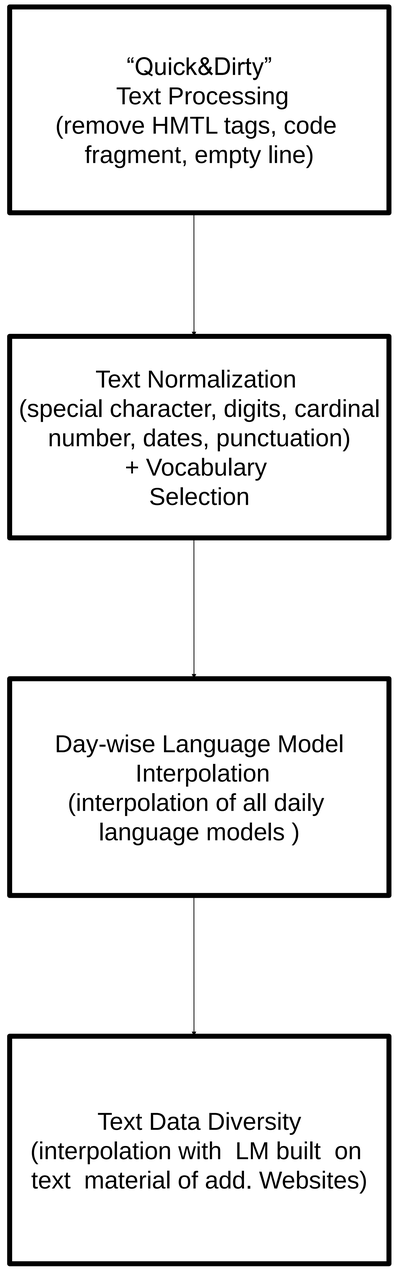
- GlobalPhone speech data: ~20h for each language, 80% for training, 10% for dev and 10% for evaluation

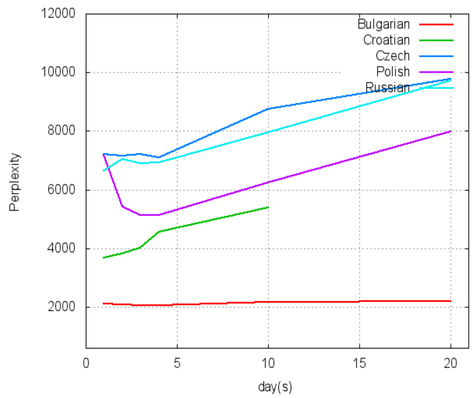
3. Baseline sytems

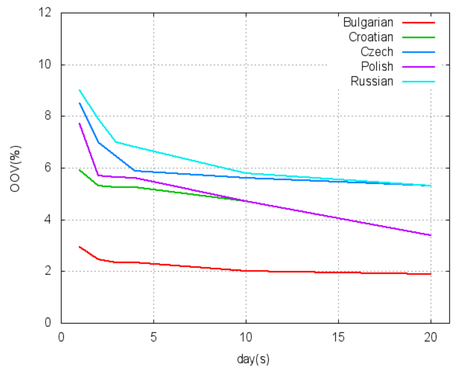
- Rapid bootstrapping based on multilingual acoustic model inventory
- Hamming window of 16ms with 10 ms overlap
- 13 MFCC, 1. and 2. derivatives, zero-crossing
- LDA -> 32 dimension
- 3 state left-to-right HMM, GMM with diagonal covariances
- LM built with utt. of training data:
- WERs of 63%(BL), 60%(HR), 49%(CZ), 72%(PL), 61%(RU)

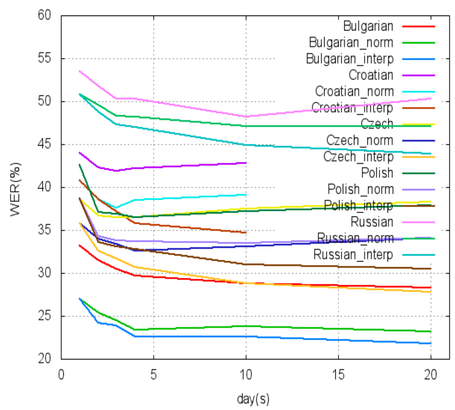
4. Experiments and Results

LM Optimization strategy:









5. Conclusion

- Crawling and processing a large amount of text material from WWW using RLAT
- Investigation of the impact of text normalization and text diversity on the quality of the language model in terms of perplexity, OOV rate and its inluence on the WER

Final Language Models:

Languages	OOV(%)	PPL	#Tokens	#Types
Bulgarian	1.2	543	405M	274K
Croatian	3.6	813	331M	362K
Czech	3.8	2,115	508M	277K
Polish	2.9	1,372	224M	243K
Russian	3.4	1,675	2931M	293K

- ASR sytems in a very short time period and with minimum human effort
- The best systems give WERs of 16.9% for BL, 32.8 % for HR, 23.5% for CZ, 20.4% for PL and 36.2% for RU on the evaluation set