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

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

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT)

{thang.vu, tim.schlippe, tanja.schultz}@kit.edu, franziska.kraus@student.kit.edu

Abstract

This paper presents our latest efforts toward LVCSR systems for five Eastern European languages such as Bulgarian, Croatian, Czech, Polish, and Russian using our Rapid Language Adaptation Toolkit (RLAT) [1]. We investigated the possibility of crawling large quantities of text material from the Internet, which is very cheap but also requires text post-processing steps due to the varying text quality. The goal of this study is to determine the best strategy for language model optimization on the given domain in a short time period with minimal human effort. Our results show that we can build an initial ASR system for these five languages in only twenty days using RLAT. On the multilingual GlobalPhone speech corpus [2], we achieved a word error rate (WER) of 16.9% for Bulgarian, 32.8% for Croatian, 23.5% for Czech, 20.4% for Polish, and 36.2% for Russian.

Index Terms: automatic speech recognition, rapid language adaptation, RLAT, Eastern European languages

1. Introduction

The performance of speech and language processing technologies has improved dramatically over the past decade with an increasing number of systems being deployed in a large variety of applications. However, most efforts are focused on a small number of languages with economic potential and a large speaker population with significant information technology needs. With more than 6,900 languages in the world, the biggest challenge today is to rapidly port speech processing systems to new languages with low human effort and at reasonable cost. Our Rapid Language Adaptation Toolkit (RLAT) [1] aims to significantly reduce the amount of time and effort involved in building speech processing systems for new languages. RLAT provides innovative methods and tools that enable users to develop speech processing models, collect appropriate speech and text data to build these models as well as evaluate the results allowing for iterative improvement. In this paper, we describe our latest improvement of these tools and their application to five Eastern European languages, namely Bulgarian [5], Croatian [6], Czech [7] [8], Polish [9], and Russian [10]. Despite a large speaker population (about 300 Million speakers in total), only a small number of research groups studied speech processing systems in these languages so far. All five languages belong to the Slavic branch of the Indo-European language family and have several language characteristics in common. These characteristics provide many challenges for speech and language processing, such as a rich morphology resulting in large vocabulary growth and high out-of-vocabulary (OOV) rates.

In this paper, we apply our RLAT to these five languages to build initial speech recognition systems in a very short time frame, with minimal human effort and at low cost. Furthermore, we investigate how to make best use of massive amounts of text data from the Internet, and evaluate the impact of amount and quality of the retrieved material.

2. Slavic Languages and Data Resources

2.1. Peculiarities of Slavic Languages

The five languages Bulgarian, Croatian, Czech, Polish, and Russian investigated in this paper all belong to the Slavic branch of the Indo-European language family, which totally contains about 20 languages and dialects. Bulgarian and Croatian are South-Slavic languages, Czech and Polish are West-Slavic and Russian belongs to the East-Slavic branch. Russian has by far the largest speaker population (more than 165M), Polish the second largest (about 56M), while Czech and Bulgarian (both about 12M) as well as Croatian (7M) have significant smaller number of speakers. Slavic languages are well known for their rich morphology, caused by a high inflection rate of nouns using various cases and genders. With respect to the sound system, Slavic languages make use of a large number of palatal and palatalized consonants, which often are grouped with related non-palatalized consonants or form pairs of complex consonantal clusters. By contrast, the vowel inventory is very small for all languages. Polish has five basic vowels plus two nasal vowels, the other four languages only use the five basic vowels. Due to the rich morphology, word order is less important than in English and can thus be used as a mean of accentuation. Grammatical similarities exist between Czech, Polish, and Russian which use seven cases and three tenses. Bulgarian and Croatian use seven tenses instead. While Croatian applies seven cases, Bulgarian has no cases. A further peculiarity is the use of grammatical aspect which denotes the temporal flow of a described event or state. Grammatical aspect is also used in English present tense, e.g. "I swim" versus "I am swimming". Bulgarian is the only Slavic language that employs articles. This peculiarity is a result from the Balkan Sprachbund, a linguistic area that includes several Balkan and South-Slavic languages. Due to the shared origin of the Slavic languages, elementary verbal communication across languages is possible.

2.2. Speech and Text Data

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages, including Arabic, Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, English, French, German, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, Turkish, and Vietnamese [2]. The corpus contains more than 400 hours speech spoken by more than 1,900 adult native speakers. GlobalPhone is available from ELRA, the European Language Resources As-

sociation. In each language about 100 native speakers read about 100 sentences each. The read texts were selected from national newspapers from the Internet. The read articles cover national and international political news as well as economic news from 1995 to 2009. The speech data is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone. Most transcriptions are internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects. Pronunciation dictionaries for all languages cover the words in the transcripts and were manually cross-checked after a rule-based creation process. For this work, we selected five Eastern European languages from the GlobalPhone corpus, namely Bulgarian, Croatian, Czech, Polish, and Russian. Bulgarian was collected in 2003, the others in 1995 and 1999. Table 1 summarizes information about the speech data and the distribution which was used for the experiments.

Table 1: *GlobalPhone speech: Number of speakers (length of audio data in minutes) for five Eastern European languages*

Languages	Training set	Dev set	Eval set
Bulgarian	63 (1,027)	7 (149)	7 (143)
Croatian	72 (725)	10 (123)	10 (105)
Czech	82 (1,010)	10 (142)	10 (161)
Polish	79 (1,162)	10 (171)	10 (140)
Russian	95 (1,187)	10 (149)	10 (143)

To build a large text corpus for these languages, we used RLAT (see below) to collect text data from the Internet as listed in Table 2. We applied a link depth of 20, i.e. we captured the content of the given webpage, then followed all links of this page to crawl the content of the successor pages. The process was continued with the respective successors of these pages until the specified link depth is reached.

Table 2: *Text corpus size for five Eastern European languages*

Languages	Website	#Words	
		#Tokens	#Types
Bulgarian	dariknews.bg	302M	560K
Croatian	www.hrt.hr	124M	248K
Czech	www.lidovky.cz	790M	1250K
Polish	wiadomosci.wp.pl	347M	815K
Russian	www.rian.ru	565M	1000K

3. Experiments and Results

3.1. Rapid Language Adaptation Toolkit (RLAT)

RLAT is a web-based interface which aims to reduce the human effort involved in building speech processing systems for new languages. Innovative tools enable novice and expert users to develop speech processing models, such as acoustic models, pronunciation dictionaries, and language models, to collect appropriate speech and text data for building these models, and to evaluate the results. RLAT was recently extended by a "snapshot" function which gives informative feedback about the quality of text data crawled from the web. The user can specify a time interval when new language models are automatically built based on the harvested data. The quality of the language models can be evaluated based on criteria, such as perplexity, OOV rate, n-gram coverage, vocabulary size, and WER given a test corpus and a speech recognizer. Furthermore, we used also language identification between target language and English to make the crawling process more efficient.

3.2. Baseline Speech Recognizers

To rapidly build baseline recognizers for the five languages, we applied the rapid bootstrapping function in RLAT which is based on a multilingual acoustic model inventory. This inventory was trained earlier from seven GlobalPhone languages [3]. To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds. The closest match is derived from an IPA-based phone mapping. In this work, we did a phone mapping for each language and trained with RLAT five different acoustic models. They used the standard front-end by applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 43 dimensions containing 13 Melscale Frequency Ceptral Coefficients (MFCC), their first and second derivatives, zero crossing rate, power and delta power. A Linear Discriminant Analysis transformation reduces the feature vector size to 32 dimensions. The acoustic model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. The initial language models of the baseline systems were trained from all utterances of the training data and show very high OOV rates on the development set (Bulgarian: 11.2% , Croatian: 12.1%, Czech: 13.9%, Polish: 16.9%, and Russian: 22.3%). The performance of these baseline systems was measured in terms of WERs on the development set after k-means clustering and 6 iterations of Viterbi training based on the intial state alignment produced by RLAT. The results were 63% WER for Bulgarian, 60% for Croatian, 49% for Czech, 72% for Polish, and 61% for Russian.

3.3. "Quick and Dirty" Text Processing

To improve the language models we started the RLAT crawling process with link depth of 20 for one website per language (see Table 2). The collected text data was roughly processed by removing HTML tags, code fragments, and empty lines. This raw text data was then used to create language models. Figures 1, 2,

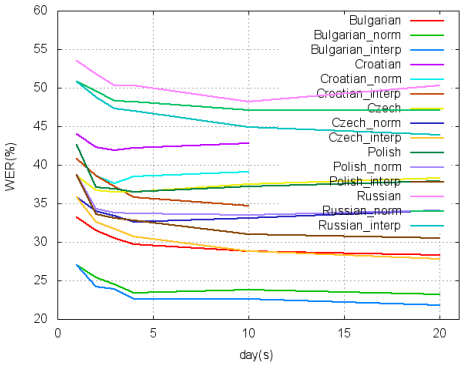

Figure 1: *WER over days of text crawling*

and 3 track the development over time of the text crawling process in terms of language model quality (perplexity, OOV rate) and its impact on the speech recognition performance (WER). The OOV rate clearly decreases over the time, while the per-

plexity increases over time. This latter effect is most likely a result of the increasing noise due to the rough text processing. For morphologically rich languages such as these five Eastern European languages, a method to select the decoding vocabulary is a big challenge. We started collecting the 100K most frequent words and defined the frequency of the last occurring word in the list as threshold. All words that occur more often than this threshold were selected. Day by day we increased the threshold by one, but only if the number of entries in the vocabulary increased. If not, we used the previous threshold. After 20 days, the decoding vocabulary for Bulgarian was 140K, for Croatian 160K, for Czech 197K, for Polish 179K, and for Russian 196K words. This method works quite well in order to control the growth of vocabulary and perplexity on one side and to decrease the OOV rate on the other side. Finally, we generated a pronounciation dictionary for all words in the selected decoding vocabulary using Sequitur G2P [11].

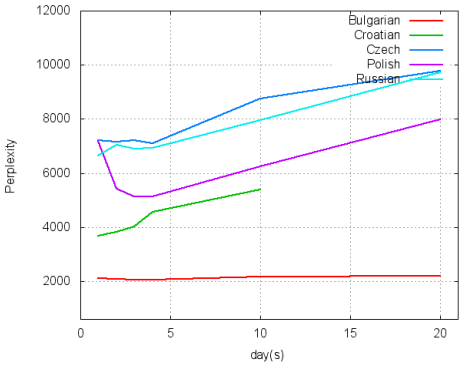

Figure 2: *Perplexity over days of text crawling*

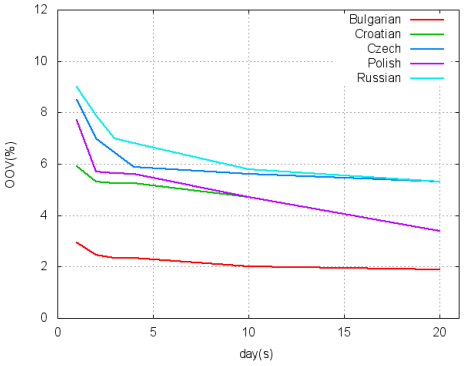

Figure 3: *OOV rate over days of text crawling*

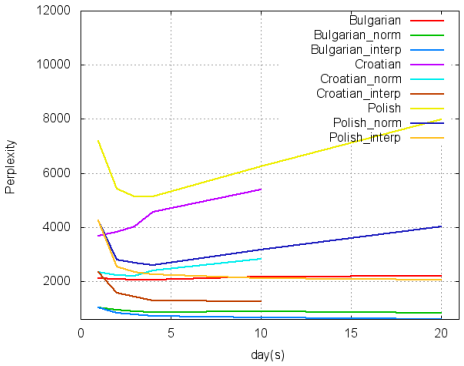

Figure 4: *Perplexity trend (non-normalized, normalized and interpolated) for 3 languages over days of text crawling*

3.4. Text Normalization

To reduce the noise in the language models we improved the text normalization by four post-processing steps, (1) special characters were deleted, (2) digits, cardinal numbers, and dates were mapped into text form to match the dictionary, (3) punctuation was deleted, (4) all text data was converted to lowercase. Particularly the second step involved some linguistic knowledge as in Slavic languages the textual form of numbers changes with gender, numerus, and case of the referring noun. The four post-processing steps gave significant relative WER reductions of 15% for Bulgarian, 7% for Croatian, 10% for Czech, 12% for Polish, and 6% for Russian. But in spite of decreasing WER in the first few days, the WER of all languages still increases up to the 20th day. We suggest that the reason lies in the growth of perplexity, that means enlarging the text corpus provides good generalization of the language model but does not always help for a specified test set.

3.5. Language Model Interpolation

The former experiments indicate that massive text data crawling decreases the OOV rate significantly. However, the overall increase in perplexity limits the positive impact on WER performance. To smooth perplexity variations over the period of crawling and to speed up language model building for large amounts of data collected over many days, we investigated the following linear interpolation scheme. For 20 days, every day one language model was built based on the text data crawled on that day. The final language model was created based on a linear interpolation of the collection of 20 daily language models. The interpolation weights were computed using the SRI Language Model Toolkit [4], optimized on the development sets. Figures 1, 2, 3, and 4 show that not only the OOV rate, but also the perplexity of the language model improved daily. The resulting language models gave perplexities (OOV rates) of 602 (1.9%) for Bulgarian, 1,268 (4.7%) for Croatian, 2,011 (5.0%) for Czech, 2,055 (3.3%) for Polish, and 2,114 (5.3%) for Russian. The consistent WER improvements achieved by interpolated language models are depicted in Figure 5.

3.6. Text Data Diversity

The experimental results reported so far were based on text data that had been harvested starting from one particular website per language (see Table 2). This makes the crawling process fragile, especially if the starting page is poorly chosen. In our experiments we found that in case of Croatian the crawling process prematurely finished after 10 days, retrieving only a relatively small amount of text data. Also, the increasing WER for Croatian after the third day of crawling, indicated that the crawled text data was suboptimal. In the following experiment we investigate the impact of text data diversity to language model quality and WER. In our experiments text diversity was increased by picking additional websites as starting points for our RLAT-crawling process and by limiting the days of crawling to up to five days. Interpolated language models were built based on these additional data in the same way as described above. Ta-

ble 4 summarizes the location of websites, days of crawling, and the performance of the resulting language models (based on the development set). While the performance of the website-specific language models is quite low, we achieve significant improvements by interpolating them with the 20-day language model (see above). We saw the largest gain for Croatian language, for which the perplexity decreased from 1,268 to 813 and the OOV rate decreased from 5.2% to 3.6%. Finally, we evaluated the speech recognition systems of all five Eastern European languages using the different language models. Table 3 compares WERs based on the development and evaluation set using the interpolation of the 20-day language model with the model from the additional websites (+add. websites), the additional interpolation with the model from the speech transcription training data (+training utts) and the best model with use of 500K decoding vocabulary(+500K dict). Figure 5 summarizes the improvements of the speech recognition systems for all five languages.

Table 3: *WER [%] for five Eastern European Languages*

Language / LM	BG	HR	CZ	PL	RU
+ add. websites (dev)	20.4	30.5	26.5	27.2	41.0
+ training utts (dev)	20.0	28.9	25.3	24.3	40.3
+ training utts (eval)	16.9	32.8	24.8	22.3	36.6
+ 500K dict (eval)	17.6	33.5	23.5	20.4	36.2

Table 4: *Summary of LM performance based on additional data from various websites (on development set)*

Websites (#days)	OOV	PPL	#Words	Vocab
Bulgarian				
24chasa.bg (2)	2.1	904	66M	153K
dnes.bg (2)	2.2	1,099	77M	169K
capital.bg (5)	1.7	808	262M	174K
Inter. LMs	1.2	543	405M	274K
Czech				
halonoviniy.cz (5)	5.2	2,699	127M	166K
respek.ihned.cz (5)	6.6	3,468	118M	173K
hn.ihned.cz (5)	5.2	2,600	127M	63K
aktualne.centrum (5)	9.5	3,792	136M	102K
Inter. LMs	3.8	2,115	508M	277K
Croatian				
index.hr (5)	4.5	1,006	71M	218K
ezadar.hz (5)	5.6	1,333	87M	187K
tportal.hr (5)	5.7	1,084	49M	143K
vecernji.hr (5)	6.3	1,884	124M	158K
Inter. LMs	3.6	813	331M	362K
Polish				
fakt.pl (5)	8.2	3,383	79M	136K
nowosci.com.pl (5)	9.0	4,824	45M	90K
wyborcza.pl (5)	3.1	1,673	100M	225K
Inter. LMs	2.9	1,372	224M	243K
Russian				
pravda.ru (3)	4.0	2,039	84M	216K
news.ru (4)	4.6	2,330	91M	222K
bbc.ru (4)	14.5	3,015	23M	34K
news.mail.ru (5)	7.2	3,098	136M	129K
Inter. LMs	3.4	1,675	334M	293K

4. Conclusions

This paper presented our latest efforts toward LVCSR systems for five Eastern European languages, i.e. Bulgarian, Croatian, Czech, Polish, and Russian using our RLAT. We described the crawling of massive amounts of text data using the RLAT-snapshot function and investigated the impact of text normalization and text diversity on the quality of the language model in terms of perplexity, OOV rate and its influence on the performance of speech recognition for the five languages. Our results indicate that initial speech recognition systems can be built with RLAT in very short time and with moderate human effort. The current best systems give word error rates of 16.9% for Bulgarian, 32.8% for Croatian, 23.5% for Czech, 20.4% for Polish, and 36.2% for Russian on the GlobalPhone evaluation set.

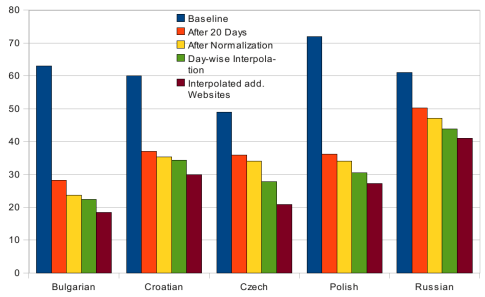

Figure 5: *Speech Recognition Improvements [WER]*

5. Acknowledgements

The authors would like to thank Edy Guevara, Dimitri Majarle, Yassine Khelifi, Wojtek Breiter, Sebastian Ochs, Zlatka Mihaylova and Chenfei Zhu for their support and discussions. This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

6. References

[1] T. Schultz and A. Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc. ICASSP Las Vegas, NV 2008.

[2] T. Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[3] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication August 2001, Volum 35, Issue 1-2, pp 31-51.

[4] A.Stolcke. SRILM - an extensible language modeling toolkit, in Proceedings of ICSLP, 2002.

[5] A. Mircheva. Bulgarian Speech Recognition and Multilingual Language Modeling, Study thesis, Uni Karlsruhe, March 2006.

[6] P. Scheytt, P. Geutner, A. Waibel, Serbo-Croatian LVCSR On The Dictation And Broadcast News Domain. International Conference on Acoustics, Speech, and Signal Processing 1998, ICASSP 1998, Seattle, USA, 01. May 1998

[7] W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka. On large vocabulary continuous speech recognition of highly inflectional language - Czech. In Eurospeech 2001.

[8] P. Icring and J. Psutka. Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary. In TSD, Železnaá Ruda, Czech Republic. 2001.

[9] J. Lööf, C. Gollan, and H. Ney. Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System. In Interspeech 2009, pages 88-91, Brighton, U.K., September 2009.

[10] S. Stüker, T. Schultz. A Grapheme based Speech Recognition System for Russian. 9th International Conference Speech and Computer 2004, SPECOM 2004, St. Petersburg, Russia, 2004.

[11] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html