Russian Sub-Word Based Speech Recognition Using Pocketsphinx Engine

Sergey Zablotskiy and Maxim Sidorov
Institute of Communications Engineering, Ulm University, Ulm, Germany

Keywords: Russian, LVCSR, Sub-words.

Abstract: Russian is a synthetic language with a large morpheme-per-word ratio and highly inflective nature. These two peculiarities increase the lexicon size for Russian automatic speech recognition (ASR) by tens of times in comparison to that for English covering the same out-of-vocabulary (OOV) rate. The employment of sub-word units is a widely spread state-of-the-art approach to reduce the abundant lexicon and lower the perplexity (PP) of the language model. The choice of sub-word units affects the accuracy of the entire speech recognition system, its performance as well as the complexity of the spoken phrase synthesis. Here, different recognition units are investigated using pocketsphinx-engine while recognizing the vocabulary of several million word forms. A designed text normalization approach is also briefly presented. This rule-based algorithm allows keeping diverse Russian abbreviations and numerals in the language model (LM) and avoiding the statistics distortion. The approach is directly applicable and useful for Russian text-to-speech translation as well.

1 INTRODUCTION

Similarly to the other Slavic languages Russian is highly inflective with a large morpheme-per-word ratio. Five basic parts of Russian speech (a noun, a verb, an adjective, a numeral and a pronoun) are inflected according to different grammatical categories: 6 cases, 3 genders, etc. There are no articles or auxiliary words at all. Each word has its own meaning. All the grammatical information is embedded into the word itself by the use of grammatical affixes. For example, “oni pryg*ajut*” (they jump), “oni pryg*ali*” (they were jumping), “on pryg*nut*” (they will jump), “oni pryg*nuli*” (they have jumped). A lot of various prefixes form different meaning nuances of one basic word. For instance, “on igral” (He was playing), “on *do*igral” (He has stopped playing), “on *vy*igral” (He has won), “on *pro*igral” (He has lost) and many others.

The above mentioned language peculiarities result in an abundance of word forms. For example, the verb “delat”’(to do) has more than 100 differently spelt word forms. The vocabulary of 2.3 million word forms used in this study corresponds to 130 thousand lemmas only.

Up-to-date non-server ASR systems handle the lexicon size of several hundred thousand words, if the real time factor (RTF) is expected to be less than 1 using a standard PC. Such an abundant Russian lexicon sophisticates the employment of the full-word N-gram approach owing to the excessively large computational time and the necessity of huge statistical data collection. A significantly larger amount of textual data is also required due to the very relaxed word order constraints in Russian. Even the subject and predicate have no predefined position in the sentence or may be omitted completely in a grammatically correct sentence. Although, some word permutations are meaningless or do not sound naturally.

There are different approaches in literature addressing the same problem for Russian and other highly inflective or agglutinative languages. The common way to reduce the lexicon is the employment of sub-word units. In (Byrne et al., 2000; Arsoy et al., 2009; Karpov et al., 2011) morphemes are used as the smallest linguistic components having a semantic meaning. The syllables (Xu et al., 1996; Shaik et al., 2011) are often chosen from the speech production point of view. Different statistically derived units and units appended by their pronunciation are exploited as well (Shaik et al., 2011). In some cases the algorithms are even able to recognize the OOV-words as the combination of sub-words (Bisani and Ney, 2005). Each type of unit has its own advantages and drawbacks. For Russian and many other languages the number of syllables is significantly smaller than number of mor-

Zablotskiy S. and Sidorov M..
Russian Sub-Word Based Speech Recognition Using Pocketsphinx Engine.
DOI: 10.5220/0005148008400844
In *Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics* (ASAAHMI-2014), pages 840-844
ISBN: 978-989-758-040-6

phemes for the same full-word dictionary. However, the major challenge and the source of SR errors is the word synthesis out of recognized syllables, since they are mostly smaller in size.

To make the choice of proper Russian sub-words easier, the comprehensive comparison of different units is presented in this paper from the SR perspective. All experiments are conducted under the same conditions for comparison to be fair.

We also briefly describe here data used for acoustic (AM) and language modeling. A special issue is the preprocessing of texts. Particularly, the handling of abbreviations, Arabic and Roman numerals, is worth noting, since their inflection mostly depends on the context and grammar and is not the trivial task for Russian.

2 ACOUSTIC MODELING

Since the basic idea of this study is a fair comparison of different sub-word LMs, all the phonemes (senones) were trained only once. The same set of phonemes and corresponding HMMs were used in each experiment for all sub-word units. Pronunciation dictionaries were built in a way, that the phonetic transcriptions of concatenated sub-words coincide with the corresponding full-word transcriptions. That was done by modifying a grapheme-to-phoneme converter provided by the Laboratory for Speech and Multimodal Interfaces of Russian Academy of Sciences (Kipyatkova and Karpov, 2009). The tool applies around 100 phonetic rules and requires two dictionaries: an emphasis dictionary and a ”jo”-dictionary. The former is essential, since stressed vowels are differently pronounced (e.g. letter ”o” is spellt as phoneme ”o!” if emphasised and phoneme ”a” otherwise). Stressed vowels often alter the neighbouring sounds as well. The latter dictionary is required, since in written Russian letters ”jo” are usually replaced by ”je” causing pronunciation ambiguity.

The ISABASE-2 (Bogdanov et al., 2003) corpus used in our work is one of the largest high-quality read speech corpora for Russian.

The lexical material of the speech database consists of 3 non-intersecting sets:

- R-set: 70 sentences chosen by linguists to cover all phonemes at least three times.
- B-set: 3060 sentences for training.
- T-set: 1000 sentences for testing.

The sets B and T were chosen from newspaper articles and internet pages of different domains. Some sentences were taken without adaptation while some of them were pruned in order to be mostly no longer than 10 words. The result sets provide the sufficient allophone coverage.

Sentences from the sets R and B were spoken by 100 speakers: 50 male and 50 female. Each speaker has uttered all 70 sentences from R-set and 180 sentences from B-set. For any two speakers B-subsets either coincide or do not intersect at all. Therefore, each sentence from the R-set was spoken by all 100 speakers and each sentence from the B-set was pronounced by several male and female persons.

The test set was uttered by other 10 speakers: 5 male and 5 female. Each of them read 100 unique sentences from the T-set.

All speakers were non-professional speakers living in Moscow and having mostly the Moscow pronunciation type.

Every utterance is presented as a separate Wav-file (22050 Hz, 16 bit) along with its information file. The information file includes the following:

- Speaker personal information: sex, age, education, place of birth and residence, etc.;
- Textual transcription of the utterance;
- Expected phonetic transcription;
- Data from experts (phoneticians): actual utterance transcription and estimation of the speaker’s accent type.

The total duration of speech is more than 34 hours including 70 minutes of the development and test material.

For acoustic modelling the sphinxtrain (Carnegie Mellon University, 2012) tool was utilized. Features, extracted from an acoustic signal are 13 MFCCs, their first and second derivatives. 40 Mel-filters were applied in the frequency range from 130 till 6800 Hz. 20 different AMs were trained on the whole training set and tested on the development set to find the optimum number of senones (tied states of triphones) and Gaussian mixture components. The best found parameters for Russian AM are 2000 senones and 16 Gaussians in one mixture.

The phoneme set consists of 48 sounds: 1 silence phoneme, 18 plain and 18 palatalized consonants, 6 stressed and 5 unstressed vowels (unstressed ”o” does not exist).

3 PRE-PROCESSING OF RUSSIAN TEXT CORPORA AND LANGUAGE MODELING

The largest available digital text sources are usually scanned or typed books, newspaper archives and internet articles. The most of texts, especially from newspapers, comprise a lot of abbreviations and numbers. For less inflective languages it does not pose any challenge and they can be easily substituted by full words performing some minor grammatical adaptation. For Russian this substitution turns into a multi-step procedure involving morphological and syntactical knowledge. Such sentences could be simply omitted, like it is often done for Russian SR. Unfortunately, this leads to the undesired statistics falsification and the model poorly represents almost all the numbers and abbreviations in diverse contexts.

The algorithm of the text pre-processing (Zablotskiy et al., 2011b) was implemented by the authors as a single Perl-script invoking the morphological tool "mystem" (Segalovich, 2003) which is even able to estimate precisely enough morphological properties of words absent in its dictionary. At the moment of writing the script consisted of 14 thousand lines of very compressed scripting language code. Two real examples taken from the script's log-file are presented in Fig 1.

Arabic Numerals
`Old: V predelah 1,5 tys. t.`
`New: V predelah pol` *u* `tor` *a* `tysjach tonn`
(no more than 1,5 thousand tonn)

Arabic Numerals
`Old: Na 1,5 tys. t.`
`New: Na poltor` *y* `tysjach` *i* `tonn`
(for 1,5 thousand tonn)

Figure 1: Automaitic number-to-text transcription.

The algorithm was applied for the texts of Maxim Moshkov's library (Moshkov, 2012) and the archive of "Nezavisimaya Gazeta" newspaper (www.ng.ru). It takes about 12 hours to process 1Gb of plain texts on a single Intel® Core™2 Duo PC. The result texts were processed by the SRILM toolkit (A. Stolcke and Abrash, 2011) to estimate the back-off 3-gram LMs. Two smoothing techniques (Good-Turing discounting with Katz back-off smoothing (Chen and Goodman, 1998) and Kneser-Ney smoothing (Kneser and Ney, 1995)) were applied resulting in different WER (up to 2% absolute difference) on the same development set using same AMs. Therefore, due to the limitation of space, final results in Table 1are only shown for the smoothing techniques which correspond to the better performance on the development set. Mostly it is the Kneser-Ney smoothing.

4 EXPERIMENTAL EVALUATION

The recognition of sub-word units was performed by the pocketsphinx recognizer (Carnegie Mellon University, 2012). All the possible conditions for different sets were unified to make the honest comparison. The number of sub-words is determined by splitting of the available phonetic dictionary (2.3M word forms). This implies, that all the built systems are able to recognize the words from this lexicon. However, due to the different size of units, the systems may recognize different number of OOV words. In Table 1 the recognition results for development and test set are presented.

Table 1: Speech recognition results for various LMs.

LM, type	Lex. size	OOV (%)	PP	SWER (%)	WER (%)
Development set					
SylBE	36k	0	31	**21.9**	68.4
SylE	32k	0	33	22.6	67.9
MrphPA	193k	0.04	118	26.8	**55.9**
MrphA	217k	0.11	149	47.4	89.7
MrphS	397k	0	166	51	88.9
Test set					
SylBE	36k	0	31	**25.4**	74.7
SylE	32k	0	33	26.7	74.4
MrphPA	193k	0.06	118	32.9	**64.7**
MrphA	217k	0.09	149	51.9	92.1
MrphS	397k	0	166	53.6	95.6

Here, the following abbreviations are used. SylE - is the syllable LM with the words' last syllables marked by "_e". Since the number of syllables in Russian words is large on average and a 3-gram LM is used, the SylBE model is tested as well. The only difference to the previous LM is the presence of marker "_b" appended to the words' first syllables. MrphA and MrphPA are morpheme LMs which morphemes were determined according to a morphological dictionary. MrphPA is a "prefix+stem+affix" model, MrphA has only affix markers. MrphS is a pseudo-morpheme LM which lexicon consists of sub-words extracted by the Morphessor tool (Creutz and Lagus, 2005). 86% of words were split into 2 parts only. However, a lot of words were divided into the larger

number of parts (up to 7). No changes were applied to the default settings of Morphessor and corresponding morpheme division. SWER stands for the sub-word ER. The RTF for syllable LMs is 0.58, for morpheme LMs - 0.45 and for MrphS LM - 0.49.

As can be seen, the best sub-word recognition accuracy was achieved for syllable LMs, since the lexicon size is relatively small and each Russian syllable consists obligatory of one vowel which usually pronounced longer and is easier to recognize. However, after straight-forward word synthesis the "base+affix" MrphA LM turns to be the best choice for very large vocabulary Russian ASR despite of high perplexity (PP). It's also worth mentioning, that OOV-rate on the development and test sets was non-zero for morpheme LMs only. Other LMs were able to find the appropriate sub-word units to cover the unknown words. Interesting result is that the inclusion of prefixes into the morpheme LM decreases dramatically the recognition accuracy. This is explained by the small size of prefixes on average and by the fact, that lots of Russian prefixes do not include vowel sounds. The same explanation is valid for MrphS LM: the vast amount of small sub-words without vowels in the model.

5 CONCLUSION AND FUTURE WORK

In this study different sub-word LMs were compared. The language and acoustic training were conducted under the same conditions, using the same textual and speech material. The "stem+affix" morpheme model outperformed significantly other LMs being an optimum trade-off between the number of sub-words and their size. Relatively low recognition accuracy is explained by the highly inflective nature of Russian and conforms with the results of other state-of-the-art research for Russian SR (Karpov et al., 2011). The word forms of one lemma in Russian sound often very similar making them hard to distinguish even for a human listener. Quite relaxed word order constraints complicates the statistical modeling of word inflection.

The graphone LMs (Shaik et al., 2011) were not investigated in this study. However, they could achieve the better recognition accuracy for Russian and should be tested in the future. As shown in (El-Desoky et al., 2009) it is better for sub-word LM to not decompose the N most frequent words. In our experiments all the decomposable words were split into parts that resulted in the low WER but this was done on purpose to compare the LMs under similar conditions.

As shown in our previous work (Zablotskiy et al., 2011a), it is possible to concatenate the sub-words even without any markers using the genetic global search algorithm capable to correct some SR errors. However, for this algorithm to work the relatively small SWERs are required. From this perspective, the syllable LM is the most suitable for Russian LVCSR.

ACKNOWLEDGEMENTS

This work is partly supported by the DAAD (German Academic Exchange Service).

REFERENCES

A. Stolcke, J. Zheng, W. W. and Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.

Arsoy, E., Can, D., Parlak, S., Sak, H., and Saraşlar, M. (2009). Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, 17(5).

Bisani, M. and Ney, H. (2005). Open vocabulary speech recognition with flat hybrid models. In *Proc. of the European Conf. on Speech Communication and Technology (Eurospeech'05)*, pages 725–728, Lisbon (Portugal).

Bogdanov, D., Bruhtiy, A., Krivnova, O., Podrabinovich, A., and Strokin, G. (2003). *Organizational Control and Artificial Intelligence*, chapter Technology of Speech Databases Development (in Russian), page 448. Editorial URSS.

Byrne, W., Hajič, J., Ircing, P., Krbec, P., and Psutka, J. (2000). Morpheme based language models for speech recognition of Czech. In Sojka, P., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 1902 of *Lecture Notes in Computer Science*, pages 139–162. Springer Berlin / Heidelberg.

Carnegie Mellon University (2012). CMUSphinx. Open source toolkit for speech recognition.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Creutz, M. and Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report A81, Helsinki University of Technology.

El-Desoky, A., Gollan, C., Rybach, D., Schlter, R., and Ney, H. (2009). Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR. In *Proc. of the 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brighton (UK).

Karpov, A., Kipyatkova, I., and Ronzhin, A. (2011). Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In *Proc. of the 12th Annual Conference of the International Speech Communication Association (Interspeech'11)*, Florence (Italy).

Kipyatkova, I. and Karpov, A. (2009). Creation of multiple word transcriptions for conversational russian speech recognition. In *Proc. of the 13th Conference "Speech and Computer" (SPECOM'2009)*, pages 71–75, St.Peterburg (Russia).

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184 vol.1.

Moshkov, M. (2012). Maxim mashkov's library.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280.

Shaik, M., Mousa, A.-D., Schluter, R., and Ney, H. (2011). Using morpheme and syllable based sub-words for polish LVCSR. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4680 –4683.

Xu, B., Ma, B., Zhang, S., Qu, F., and Huang, T. (1996). Speaker-independent dictation of Chinese speech with 32K vocabulary. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2320 –2323 vol.4.

Zablotskiy, S., Shvets, A., Semenkin, E., and Minker, W. (2011a). Recognized Russian syllables concatenation by means of co-evolutionary asymptotic algorithm. In *Proc. XIV International Conference "Speech and Computer" (SPECOM'2011)*.

Zablotskiy, S., Zablotskaya, K., and Minker, W. (2011b). Automatic pre-processing of the Russian text corpora for language modeling. In *Proc. XIV International Conference "Speech and Computer" (SPECOM'2011)*.