Sub-word Language Modeling for Russian LVCSR

Sergey Zablotskiy(✉) and Wolfgang Minker

Institute of Communications Engineering, University of Ulm,
Albert-Einstein-Allee 43, 89081 Ulm, Germany
{sergey.zablotskiy,wolfgang.minker}@uni-ulm.de

Abstract. Russian is a highly inflected language with rich morphology. It is characterized by the low lexical coverage, high out-of-vocabulary (OOV) rate and perplexity. Therefore, the large vocabulary continuous speech recognition (LVCSR) of Russian and languages with similar morphology still remains to be a challenging task. Augmenting the full-word language model by fragments is a well-known approach targeting this challenge which also allows us to recognize missing words in the lexicon (open vocabulary recognition). In this paper we suggest a novel "double-sided" approach for marking word fragments, which reduces the WER by up to 3.7 % absolute (20.8 % relative) compared to the full-word baseline and by up to 1.1 % absolute (7.2 % relative) compared to the corresponding sub-word baseline, tested on evaluation set. Moreover, the type of word decomposition (syllables or morpheme-like units), their smallest size and optimal number of non-fragmented words were also investigated for Russian LVCSR.

Keywords: Russian speech recognition · Double-sided marking · Syllables · Morphemes · Morphs

1 Introduction

Similarly to the other Slavic languages Russian is a highly inflected language with a complex mechanism of word formation. 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 and almost no auxiliary words. The entire grammatical information is embedded into a word itself by the use of various grammatical affixes. This leads to the abundance of word forms.

The loose word order of Russian language also sophisticates the process of LVCSR contributing to the data sparsity and up to several times higher perplexities comparing to English [8].

There are different approaches in literature addressing the same problem for Russian and other highly inflected or agglutinative languages. Thus, in [17] two advanced language modeling techniques for Russian were investigated: factor

A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 413–421, 2015.
DOI: 10.1007/978-3-319-23132-7_51

language models (LMs) and recurrent neural networks (RNNs). The factor LM is the N-gram model extended by word features. Trained on an entire text corpus, the RNNs outperformed significantly the factor LMs. However, an aggregation of the factor LMs, RNNs and trigrams showed further improvement compared to the standalone RNNs. A syntactico-statistical method was suggested in [9], which estimates more accurate N-gram models for synthetic languages like Russian.

A very popular approach to reduce the lexicon size and increase an average frequency of its elements [15] is the employment of sub-word units. In [8,15] morphemes are used as the smallest linguistic components with semantic meaning. Syllables [15,18] are often chosen from the speech production point of view. Various statistically derived units and elements augmented by their pronunciation (graphones) are successfully exploited as well [2,6,15]. The use of graphones allows to capture diverse sub-lexical pronunciation on the LM level rather than the lexical level [15]. Several attempts are made to automate the word decompounding [12] or optimize the lexicon (e.g. by using the discriminative learning method exploiting the perceptron algorithm [1]).

In this work we focused on the word decomposition approach, since it has already proven its efficiency for multiple languages with rich morphology. It is able to recognize OOV-words as a combination of sub-words (so called open vocabulary recognition [2,7]). Moreover, this approach is highly portable and, therefore, does not cost much efforts and code changes if moving from one ASR system to the other. In this paper we suggested a novel approach for sub-word marking, which does not sophisticate the process of backward full-word synthesis, but significantly improves the word recognition accuracy.

2 Methodology

2.1 Sub-lexical Units

Syllables. The algorithm of the full-word division into syllables is quite straightforward and absolutely deterministic according to the principle of rising sonority. Each syllable consists of one vowel and null to several consonants.

Morphs. For decomposing of full-words into morphemes or morpheme-like units there exist dictionary-based and unsupervised approaches [6]. The disadvantage of the former one is the necessity to have the decomposition mapping for all the words. Even large existing dictionaries for Russian can not guarantee the availability of every single word.

Therefore, the data-driven tool *Morfessor* [5] for unsupervised decomposition into pseudo-morphemes (later called "morphs" as in [5]) was used. For its training an unannotated raw text is required only. Words appeared more than five times in the text corpus were used for training as it was recommended in [6]. Nevertheless, the resulted model was used for decomposing rare words as well.

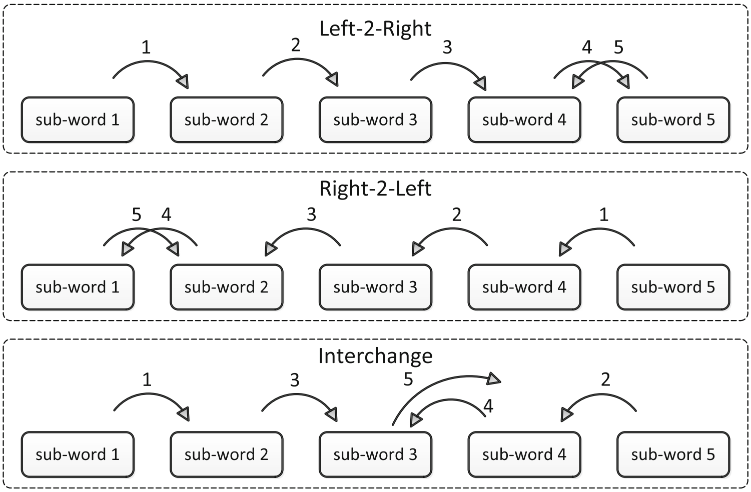

Fig. 1. Joining algorithms.

Sub-word Joining. For improving of the final WER it is recommended in [15] to adapt the word decomposition so as to avoid the very short units. Hence, here we investigated the optimal size of the smallest units for different sub-words types (syllables and morphs) as well as the algorithm of small elements' concatenation. Thus, three different joining algorithms for small sub-words were suggested: left-to-right (L2R), right-to-left (R2L) and interchange (INT). These algorithms are presented in Fig. 1. The L2R algorithm starts from the most left sub-word: if its length in letters is shorter than a threshold "min size" - join it with the neighbor to the right. The most right sub-word (if shorter than "min size") should be joint with its left neighbor. This procedure stops if all sub-words gain the required minimum size or the word boundary is reached. The R2L case is exactly inversed. The direction change for the INT happens only after reaching "min size" of the current unit. Otherwise, all transition numbers are subtracted by one. For the L2R algorithm - the largest sub-word tends to be located at the end of the word; for R2L - at the beginning, for INT - closer to the middle.

Sub-word Marking and Synthesis. Appending the "+" marker to non-terminal sub-words for easy recovery [15] ("single-sided" marking) does not take into account relative sub-word positions. Our suggestion is to use the "double-sided" marking for non-boundary sub-words. This makes the sub-word LM more context specific and more sensitive to the position of sub-units in a word.

For example, the word "прекрасное" will be decomposed as "пре+ +кра+ +сно+ +е" in the "double-sided" marking scheme and as "пре+ кра+ сно+ е" in the "single-sided" one. The "double-sided" scheme allows us to distinguish between three positional types of sub-words on the LM level. During recombination the pluses in front of sub-words are just ignored.

2.2 Text Data Collection and Normalization

To train a LM an abundance of textual resources is necessary, especially for such inflected languages like Russian. Usually the largest available digital text sources are e-books, newspaper archives and Internet articles. Most texts, especially those from newspapers, comprise lots of abbreviations and numbers. For less inflected languages they do not pose any serious challenge and can easily be substituted by full-words performing some minor grammatical adaptation. For Russian this substitution turns into a multi-step procedure involving morphological and syntactical knowledge.

Omitting such sentences causes undesired statistics falsification and such LMs poorly represent almost all the numbers and abbreviations in diverse contexts.

The rule-based algorithm for Russian text normalization *NormyRiUS* [19] is implemented as a single Perl-script available at https://gitlab.com/serjosch/normyrius. Currently it invokes the morphological tool *Mystem* [14], which is even able to estimate morphological properties of words absent in its database. Alternatively, any morphological analyzer could be exploited, e.g. *Pymorphy2*, available at https://pymorphy2.readthedocs.org/en/latest/index.html.

In this work the text data for Russian LVCSR was collected from the following sources: books of Moshkov's library (www.lib.ru), electronic scientific magazine "Наука и жизнь" (www.nkj.ru), political and non-political articles of the newspapers "АиФ" (www.aif.ru) and "Лента.ру" (www.lenta.ru).

The collected corpus of normalized texts consists of 714 M running words. It was also used for the selection of the most frequent words to be included into vocabularies.

2.3 Phonetic Transcription

Russian is a language with a close grapheme-to-phoneme relationship. There are strict pronunciation rules [8,10] with negligibly small amount of exceptions. However, the pronunciation of each word strongly depends on the position of the emphasis. For example, the word "молоко" is pronounced as /м а л а к о!/, since the first two vowels (letters) "о" are non-emphasised.

The determination of the emphasis' position is a challenge, which is normally solved by the employment of the emphasis dictionary. However, despite of its tremendous size (2.3 M), the lexicon is still missing lots of word-forms (about 100 k words out of 500 k most frequent ones are absent in the emphasis dictionary).

To our best knowledge there are no tools for the automatic Russian emphasis detection publicly available. Therefore, for the pronunciation generation of the unknown words the data-driven grapheme-to-phoneme converter *Sequitur G2P* [3] was used. It exploits data in the form "word - phonetic transcription" to train the pronunciation model and can be applied to any arbitrary language.

For a full-word Russian lexicon two different strategies were suggested and proved empirically to have very similar efficiency (±0.1 % WER for 500 k full-word vocabulary on the Development set with different parameters). Both strategies require a rule-based transcriber. In our case it was provided by the Speech and Multi-Modal Interfaces Laboratory of the St. Petersburg Institute for Informatics and Automation of Russian Academy of Science [10]. This transcriber requires information regarding stressed vowel(s) as well as proper recognition of a letter “ё”, since it is mostly omitted in written.

The difference between two approaches is in the order of two steps: either the rule-based transcriber is applied first and the Sequitur G2P is trained on the phonetic representations including emphasized phonemes or the Sequitur G2P is trained first on the words with stress markers (emphasized letter form) and the rule-based transcriber is applied to unknown words with emphases, preliminary predicted by the Sequitur G2P. The second scenario was used in this work for the sub-word lexicon generation, since it allows us to make a rule-based grapheme-phoneme alignment (transcriber property). The advantage of the rule-based alignment is the exact pronunciation borders between sub-words after splitting of full words. The same scenario was used for syllables and morphs.

3 Experimental Setup

The ISABASE-2 [4] is one of the largest high-quality continuous read speech corpora for Russian. It was created and provided for our experiments by the Institute of System Analysis of the Russian Academy of Science. A lexical material of the database consists of three non-intersecting sets:

- R-set: 70 sentences with sufficient allophone coverage for training.
- B-set: 3060 sentences, also used for training.
- T-set: 1000 sentences for testing.

The sets B and T were chosen from newspaper articles and Internet pages of different domains.

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 the 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 males and females.

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. The utterances of the T-set were split into 2 equal parts (with non-intersecting speakers) for the development (Dv) and evaluation (Ev) SR sets. All speakers were non-professional speakers living in Moscow and having mostly the Moscow pronunciation accent.

Every utterance is presented as a separate Wav-file (22050 Hz, 16 bit, downsampled to 16 kHz) along with its information file. The total duration of speech is more than 34 h including 70 min of the development and test material.

In all experiments the word is considered to be an OOV only if it is absent in the vocabulary and can not be composed from in-vocabulary fragments [15].

The acoustic modeling is performed according to [7] (re-estimation of CART, LDA with fastVTLN). The *SRILM toolkit* [16] was used to estimate the backoff 5-gram LMs with Kneser-Ney Smoothing [11] for full-word and hybrid vocabularies. SR results were obtained using the *RWTH ASR system* [13].

4 Evaluation

More than 10000 recognition experiments (minimum one hour long each on a state-of-the-art desktop computer) with different acoustic and language model parameters were carried out on the Dv set to achieve the results presented here.

The performance of a baseline full-word Russian LVCSR is shown in Table 1.

Table 1. Baseline WERs - full word 5-gram LMs (voc: vocabulary, Dv: development, Ev: evaluation, RTF: real time factor on Intel® Xeon® E3-1245, 3.40 Ghz machine)

Voc size	ISABASE-2 (Dv)			ISABASE-2 (Ev)		
	WER[%]	OOV[%]	RTF	WER[%]	OOV[%]	RTF
100 k	16.9	5.52	1.55	17.8	4.52	1.45
200 k	12.7	2.25	1.75	14.1	1.63	1.86
300 k	11.1	1.11	1.84	13.4	0.93	2.02

Table 2 shows the comparison between the morph baseline and its "double-sided" counterpart. Each row corresponds to the best parameters found on Dv set (including the type of joining and smallest element size). As can be seen, the suggested "double-sided" version outperforms significantly the morph baseline: 1.0 % WER absolute (7.1 % relative) for 300 k vocabulary on Ev set.

Table 2. WERs - morph based 5-gram LMs

Voc size	#full	min	join	ISABASE-2 (Dv)			ISABASE-2 (Ev)		
	words	size	type	WER[%]	OOV[%]	RTF	WER[%]	OOV[%]	RTF
Single-sided marking									
100 k	50 k	1	-	11.9	0.00	2.42	14.6	0.00	2.31
200 k	150 k	1	-	10.6	0.00	2.50	14.1	0.00	2.32
300 k	200 k	1	-	10.3	0.00	2.51	14.0	0.00	2.63
Double-sided marking									
100 k	50 k	1	-	11.2	0.00	2.40	14.4	0.00	2.42
200 k	150 k	1	-	10.1	0.00	2.65	13.2	0.00	2.60
300 k	150 k	1	-	10.0	0.00	2.98	13.0	0.00	3.00

The number of full-words means the number of the most frequent originally non-split words. After the text enrichment with resulted sub-words, the most frequent (sub-)words were selected again and, therefore, the number of full-words may deviate.

The comparison to the syllable baseline is given in Table 3. Again, the syllable "double-sided" modification outperforms not only the full-word baseline, e.g. by

Table 3. WERs - syllable based 5-gram LMs

Voc size	#full words	min size	join type	ISABASE-2 (Dv) WER[%]	OOV[%]	RTF	ISABASE-2 (Ev) WER[%]	OOV[%]	RTF
Single-sided marking									
100 k	30 k	4	R2L	11.8	0.72	1.86	15.2	0.63	1.77
200 k	150 k	4	R2L	10.7	0.40	2.01	13.5	0.25	1.88
300 k	150 k	4	INT	10.4	0.13	2.23	13.6	0.05	2.08
Double-sided marking									
100 k	70 k	3	R2L	11.6	0.27	1.94	14.1	0.25	1.87
200 k	150 k	3	R2L	10.7	0.08	2.24	13.1	0.08	2.19
300 k	150 k	4	INT	10.3	0.13	2.21	13.1	0.08	2.07

3.7 % absolute (20.8 % relative) for 100 k vocabulary, but also the syllable "single-sided" baseline, e.g. by 0.5 % absolute (3.7 % relative) for 300 k on Ev set.

For morph models the optimal size of the smallest unit is equal to one, i.e. joining of even one-letter morphs decreases the recognition accuracy of Russian LVCSR (as opposed to the results reported for German [6] and Polish [15]).

5 Conclusions and Future Work

Suggested "double-sided" marking method improves significantly the recognition accuracy without remarkable sophistication to conventional sub-word models. For 100 k and 200 k vocabularies, syllable modifications outperform morph ones on Ev set, taken into account, that the minimum syllable size is equal to 3.

It is worth noting that the sub-words with differently located markers are counted as separate elements. As a result, the "double-sided" LMs have a priori slightly smaller variety of vocabulary entries. Nevertheless, the "double-sided" variation outperforms its "single-sided" counterpart.

It was figured out, that all OOVs should be split, even if appeared once.

A comprehensive testing of suggested methods on the other corpora is currently ongoing. Investigation of Russian graphones is also referred to future work.

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

References

1. Ablimit, M., Kawahara, T., Hamdulla, A.: Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language. Speech Commun. **60**, 78–87 (2014)

2. Bisani, M., Ney, H.: Open vocabulary speech recognition with flat hybrid models. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Lisbon, Portugal, pp. 725–728 (2005)
3. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**(5), 434–451 (2008)
4. Bogdanov, D., Bruhtiy, A., Krivnova, O., Podrabinovich, A., Strokin, G.: Tekhnologiya formirovaniya rechevykh baz dannykh. In: Organizatsionnoe upravlenie i iskusstvennyy intellekt, pp. 239–258. Editorial URSS (2003). [Technology for Creation of Speech Corpora. In: Administration and Artificial Intelligence] (in Russian)
5. Creutz, M., Lagus, K.: Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical report A81, Helsinki University of Technology (2005). http://www.cis.hut.fi/projects/morpho
6. El-Desoky Mousa, A., Shaik, M., Schluter, R., Ney, H.: Sub-lexical language models for German LVCSR. In: Spoken Language Technology Workshop (SLT), pp. 171–176 (2010)
7. Hahn, S., Rybach, D.: Building an open vocabulary ASR system using open source software. In: Interspeech, Florence, Italy (2011). http://www-i6.informatik.rwth-aachen.de/rwth-asr/manual/index.php/Open_Vocabulary_Tutorial
8. Karpov, A., Kipyatkova, I., Ronzhin, A.: Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy (2011)
9. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun. **56**, 213–228 (2014)
10. Kipyatkova, I., Karpov, A.: Creation of multiple word transcriptions for conversational Russian speech recognition. In: Proceedings of the 13th Conference "Speech and Computer" (SPECOM), St. Peterburg, Russia, pp. 71–75 (2009)
11. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: International Conference on Acoustics, Speech, and Signal Processing. ICASSP, vol. 1, pp. 181–184 (1995)
12. Pellegrini, T., Lamel, L.: Automatic word decompounding for ASR in a morphologically rich language: application to Amharic. IEEE Trans. Audio Speech Lang. Process. **17**(5), 863–873 (2009)
13. Rybach, D., Hahn, S., Lehnen, P., Nolden, D., Sundermeyer, M., Tüske, Z., Wiesler, S., Schlüter, R., Ney, H.: RASR - the RWTH Aachen University open source speech recognition toolkit. In: IEEE Automatic Speech Recognition and Understanding Workshop, Hawaii, USA (2011)
14. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA, pp. 273–280 (2003)
15. Shaik, M.A.B., El-Desoky Mousa, A., Schlüter, R., Ney, H.: Using morpheme and syllable based sub-words for Polish LVCSR. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4680–4683 (2011)
16. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: update and outlook. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Waikoloa, Hawaii (2011)
17. Vazhenina, D., Markov, K.: Evaluation of advanced language modeling techniques for Russian LVCSR. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 124–131. Springer, Heidelberg (2013)

18. Xu, B., Ma, B., Zhang, S., Qu, F., Huang, T.: Speaker-independent dictation of Chinese speech with 32K vocabulary. In: Fourth International Conference on Spoken Language Processing ICSLP, vol. 4, pp. 2320–2323 (1996)
19. Zablotskiy, S., Zablotskaya, K., Minker, W.: Automatic pre-processing of the Russian text corpora for language modeling. In: Proceedings of the XIV International Conference "Speech and Computer" (2011)