Speech and Language Resources for LVCSR of Russian

Sergey Zablotskiy [1], Alexander Shvets [2], Maxim Sidorov [3], Eugene Semenkin [4], Wolfgang Minker [5]

[1,5] Institute of Communications Engineering, University of Ulm, Germany
[2]Institute for System Analysis of RAS, Moscow, Russia
[3,4]Insitute for System Analysis, Siberian State Aerospace University, Krasnoyarsk, Russia
[1,5]{sergey.zablotskiy,wolfgang.minker}@uni-ulm.de
[2]alexandershvets@mail.ru, [3]sidorov.math@mail.ru, [4]eugenesemenkin@yandex.ru

Abstract

A syllable-based language model reduces the lexicon size by hundreds of times. It is especially beneficial in case of highly inflective languages like Russian due to the abundance of word forms according to various grammatical categories. However, the main arising challenge is the concatenation of recognised syllables into the originally spoken sentence or phrase, particularly in the presence of syllable recognition mistakes. Natural fluent speech does not usually incorporate clear information about the outside borders of the spoken words. In this paper a method for the syllable concatenation and error correction is suggested and tested. It is based on the designed co-evolutionary asymptotic probabilistic genetic algorithm for the determination of the most likely sentence corresponding to the recognized chain of syllables within an acceptable time frame. The advantage of this genetic algorithm modification is the minimum number of settings to be manually adjusted comparing to the standard algorithm. Data used for acoustic and language modelling are also described here. A special issue is the preprocessing of the textual data, particularly, handling of abbreviations, Arabic and Roman numerals, since their inflection mostly depends on the context and grammar.

Keywords: LVCSR, Russian, language modelling, sub-word units

1. Introduction

Russian is a highly inflective language with a large morpheme-per-word ratio. Five basic parts of Russian speech (noun, verb, adjective, numeral and pronoun) are inflective. These peculiarities of the language result in the abundance of word forms. For example, the verb "delat"'(to do) has more than 100 differently spelt word forms according to corresponding grammatical categories: "delal" (He was doing), "delala" (she was doing), "delaju" (I am doing), "delaet" (he does), etc. The extraction of all the possible word forms out of 160k lemmas yields over 3,7M words.

Up-to-date non-server ASR systems handle the lexicon of only several hundreds of thousands of words. Such an abundant Russian lexicon sophisticates the employment of the well-developed N-gram approach owing to the excessively large computational time and the necessity of the huge statistical data collection. Significantly larger amounts of textual data are also required due to relaxed word order constraints.

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-units: morphemes (Byrne et al., 2000; Arsoy et al., 2009; Karpov et al., 2011), syllables (Xu et al., 1996; Shaik et al., 2011), graphones (Shaik et al., 2011) or statistically derived units. In some cases it is even possible to recognize OOV-words as the combination of sub-units (Bisani and Ney, 2005). While morphemes or "stem plus ending" units are the most popular ones, the number of Russian syllables is significantly smaller than the amount of morphemes and, moreover, the syllable error rate is noticeably lower. However, there are no rules in Russian for the syllable concatenation. The same syllables may often be located in a different part of a word causing the ambiguousness of a standard backward synthesis.

In this paper we present an advanced approach for the concatenation of recognized syllables. The designed syllable language model has a lexicon of about 12k syllables that enables a very fast recognition step. The resulted syllable chain, however, does not incorporate any clear information about the boundaries of underlying words for fluent speech. It may contain more than 60 syllables in one sentence. An examination of all the possible syllable combinations including a limited number of recognition errors takes an unacceptable amount of time with an increase of the syllable number in a sentence. Therefore, the search for a final sentence out of recognized syllables is performed by the designed co-evolutionary asymptotic probabilistic genetic algorithm (CAPGA). Its major advantage is a very few number of parameters to be assigned and faster performance comparing to the standard GA.

We also describe here data used for acoustic and language modelling. A special issue is the preprocessing of the textual data, particularly, handling of abbreviations, Arabic and Roman numerals, since their inflection mostly depends on the context and grammar.

2. Speech Corpus

The ISABASE-2 (Bogdanov et al., 2003) corpus used in our work is one of the largest high-quality speech corpora for Russian. It was created by the Institute of System Analysis of the Russian Academy of Science with the support of the Russian Foundation of Fundamental Research in collaboration with a speech group of the Philological Faculty of Moscow State University.

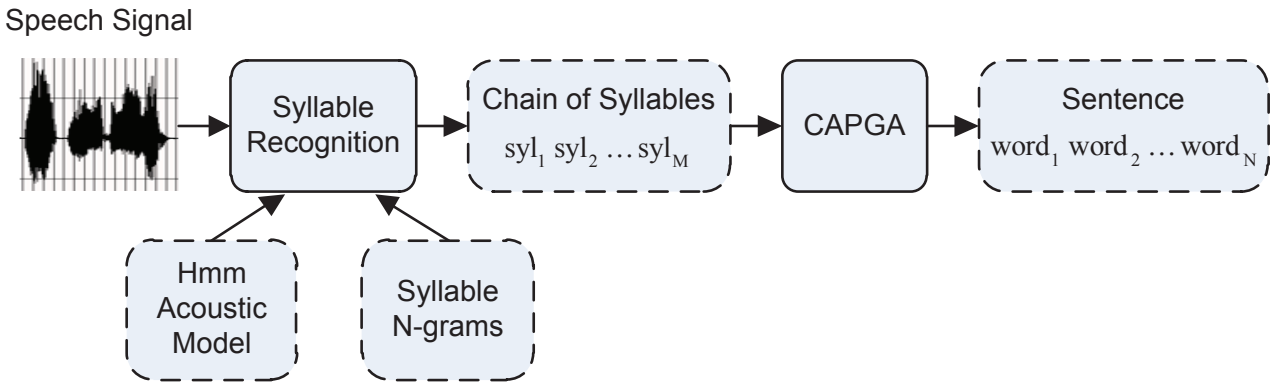

Figure 1: The syllable concatenation algorithm

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 B and T sets were chosen from newspaper articles and internet pages of different domains: politics, economics, culture, etc. Some sentences were taken without adaptation, some of them were pruned. The result sets provide sufficient allophone coverage.
Sentences from the sets R and B were spoken by 100 speakers: 50 male and 50 female. Each speaker has read 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. Thereby, 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 latter includes:

- Speaker personal information: sex, age, education, place of birth and residence, etc.;
- Experiment conditions, like date of recording and equipment used;
- 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 75 minutes of the test material.

3. Collecting and Pre-Processing of Russian Text Corpora

The largest available digital text sources are usually scanned/typed books or newspaper archives. Most texts, especially from newspapers, comprise many abbreviations and numbers. Such sentences could be simply omitted. Unfortunately, this leads to the undesired statistics falsification and the model poorly represents almost all the numbers and such abbreviations. For the less inflective languages those abbreviation and numbers 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.
The algorithm of the text pre-processing (Zablotskiy et al., 2011) is implemented by the authors as a Perl-script invoking the morphological tool "mystem" (Segalovich, 2003) which is even able to estimate the morphological properties of unknown words. The algorithm was applied for the texts of Maxim Moshkov's library (Moshkov,) and the archive of "Nezavisimaya Gazeta" newspaper (nez,). It took about 12 hours to process 1Gb of a plain text on a single Intel® Core™2 Duo machine with 6Gb of RAM.

4. The Algorithm for Syllable Concatenation

The algorithm is shown in Fig. 1. A chain of syllables is recognized using the standard HMM acoustic model and the syllable N-gram language model. The CAPGA generates different hypotheses about the probable concatenation of syllables in a phrase. Each individual is a binary string which bits mean the type of connection between syllables (see Fig. 2). The unit bit means that corresponding syllables should be joint, otherwise they belong to two neighbouring words.

Syllable Chain:	при ве тству ю дру зья
Chromosome-Solution:	1 1 1 0 1
Correct Sentence:	приветствую друзья
In English:	hello friends!

Figure 2: A chromosome example

For each hypothetical sentence (one individual of the GA) a set of closely pronounced sentences is generated. This

is done through the search for a list of likely words to each word in the hypothesis and is required for some minor recognition error correction. A score is estimated for each sentence of the found set (see Section 5.). A sentence with the highest score is attributed to the chromosome and the fitness value is equal to the score. The CAPGA iterates until the convergence or exceeding a maximum number of generations. The sentence with the highest fitness is treated as a final SR result.

5. Sentence Score Estimation

The sentence score is a numerical measure of the sentence correspondence to the available sequence of recognized syllables on the one hand and the language model on the other hand. The score estimation is shown in Fig. 3.

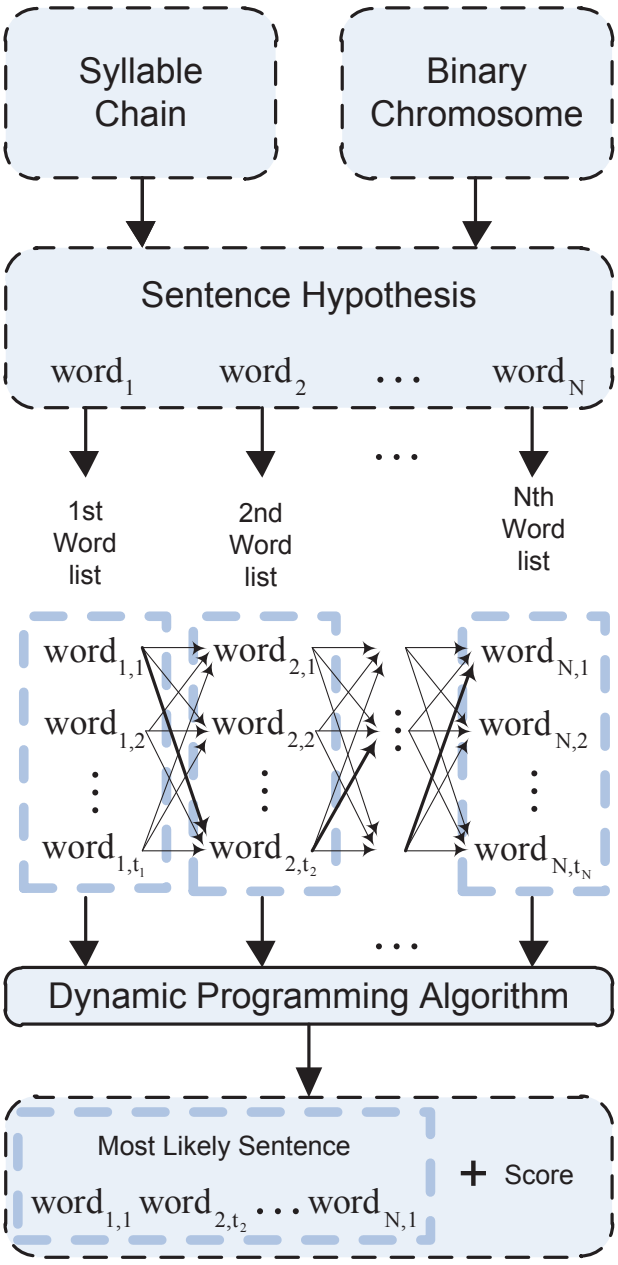

Figure 3: Sentence score estimation

The list of likely words consists of words whose likelihood coefficients are more than zero in relation to the concatenated word. The likelihood coefficient between two words is based on the Levenshtein distance (Levenshtein, 1966) with a variable operation cost between their phoneme representations where one phoneme is considered as one character. The larger distance results in the smaller likelihood coefficient.

To accelerate the performance the lexicon is stored as one single trie (Knuth, 1998). For one list the Levenshtein distance is saved in every non-terminal node to prevent unnecessary computations.

Each word in the sentence has its own list of likely words. To identify the sentence with the highest score all combinations of words (one word from one list) are checked by a dynamic programming algorithm. The score of one sentence is calculated according to Eq. 1:

$$Q = \prod_{i=1}^{N} P(w_i|w_{i-1}, w_{i-2}) \cdot \frac{\sum_{h=1}^{N} K(w_h^{con}, w_h)}{N}, \quad (1)$$

where $P(w_i|w_{i-1}, w_{i-2})$ - the conditional probability of the i-th lemma 3-gram in the hypothetical sentence, $K(w_h^{con}, w_h)$ is the likelihood of the word w_h from the h-th list with respect to the originally concatenated word w_h^{con}, N is the number of words in the sentence.

6. Co-evolutionary Asymptotic Probabilistic Genetic Algorithm

A GA is a global optimization stochastic algorithm. Its major problem is the necessity to choose its parameters. The GA performance strongly depends on this choice. The CAPGA is the combination of different techniques described below that reduce the number of manually assigned parameters.

In (Schlierkamp-Voosen and Muhlenbein, 1994) the adaptation at the expenses of competing populations is suggested. Each sub-population has its own strategy (algorithm's parameters). A redistribution of resources provides the domination of the sub-population with the best searching strategy. The modification of the co-evolutionary algorithm (Sergienko and Semenkin, 2010) uses not only the competition but also the cooperation between individual GAs. It is organized by the migration of the best individuals into all other sub-populations. This improves the performance due to the positive effect of the cooperation and weakens the algorithm's sensibility to the choice of parameters. The probabilistic modification (Semenkin and Sopov, 2005) replaces the crossover by the species generation according to the probability distribution of their elements. In this case there is no need to chose the recombination type. The asymptotic GA (Galushin and Semenkin, 2009) enables further reduction of the GA parameters, since it has an adaptive mutation which does not require any settings and performs very fast.

7. Evaluation

The GAPGA was tested on the unconstrained optimization tasks (20 multi-extremal functions). In most cases, it performed better (average reliability $\bar{R}_{\text{CAPGA}} = 0,921$) than the

best individual GA ($\bar{R}_{\text{BGA}} = 0,953$) which is not known in advance. The reliability here means the ratio of restarts when the solution was found (given an assigned accuracy) to the number of all restarts. However, the CAPGA was much faster in finding a solution: on average, it needed only 2452 generations versus 3262 generations of the best GA.
This algorithm was applied to the syllable concatenation task. Average syllable error rate achieved on the speech data set described above is ranged from 10 till 20 % depending on the certain acoustic model. All the sentences (recognized and simulated) were grouped according to the level of syllable error rate to investigate the behaviour of the concatenation algorithm. The results are presented in Table 1.

Table 1: Test results of the syllable concatenation algorithm

No. of syllab.	WER (%)	Time (sec)	$\mu(N_g)$	$\delta(N_g)$	No. of sent.
Incoming syllable chain has 0% recognition errors					
9-20	7	3	2.4	0.5	860
21-50	4	8	5.3	2.9	3058
51-80	5	41	7.5	3.0	11307
Incoming syllable chain has 4-9% recognition errors					
21-50	9	11	7.1	4.6	4010
51-80	8	49	10.1	4.9	11652
Incoming syllable chain has 10-15% recognition errors					
9-20	21	6	2.6	0.7	1045
21-50	12	11	8.4	4.8	4307
51-80	9	48	12.0	5.5	11780

Here, $\mu(N_g)$ and $\delta(N_g)$ are the mean and standard deviation of passed GA generations and the last column shows the average number of analysed hypothetical sentences. Since the algorithm is stochastic and the results may be different given the same conditions every sentence was synthesised 20 times out of the same recognized chain of syllables and the results were averaged.
The algorithm is able to concatenate the syllables and to correct some minor recognition errors. By now it takes more computational time than morpheme-based approaches and much less time than the full-word recognition. However, in most cases it is more accurate and, therefore, is recommended for non time-critical domains.

8. Acknowledgements

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

9. References

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

Maximilian Bisani and Hermann Ney. 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).

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

William Byrne, Jan Hajič, Pavel Ircing, Pavel Krbec, and Josef Psutka. 2000. Morpheme based language models for speech recognition of Czech. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 1902 of *Lecture Notes in Computer Science*, pages 139–162. Springer Berlin / Heidelberg.

P.V. Galushin and E.S. Semenkin. 2009. Asymptotic probabilistic genetic algorithm. *Scientific Journal of Siberian State Aerospace University named after academician M.F. Reshetnev*, 5(26).

A. Karpov, I. Kipyatkova, and A. Ronzhin. 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), August.

Donald E. Knuth. 1998. *Art of Computer Programming, Volume 3: Sorting and Searching (2nd Edition)*. Addison-Wesley Professional, 2 edition, May.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, February.

Maxim Moshkov. Maxim mashkov's library. http://lib.ru/.

Nezavisimaya gazetta (independent gazette) newspaper. http://www.ng.ru/.

Dirk Schlierkamp-Voosen and Heinz Muhlenbein. 1994. Strategy adaptation by competing subpopulations. In *Parallel Problem Solving from Nature (PPSN III)*, pages 199–208.

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

E. Semenkin and E. Sopov. 2005. Probabilistic evolutionary algorithms of complex systems optimization. In *International Conf. Intelligent systems (AIS05) and Intelligent CAD (CAD-2005)*, volume 3. Fizmatlit.

R. Sergienko and E. Semenkin. 2010. Competitive cooperation for strategy adaptation in genetic algorithm for constrained optimization. In *IEEE World Congress on Computational Intelligence (WCCI2010)*, Barcelona, Spain.

M.A.B. Shaik, A.E.-D. Mousa, R. Schluter, and H. Ney. 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, may.

Bo Xu, Bing Ma, Shuwu Zhang, Fei Qu, and Taiyi Huang. 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, oct.

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