Indian Journal of Science and Technology, *Vol 10(1), DOI: 10.17485/ijst/2017/v10i1/109954, January 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Language Models Creation for the Tatar Speech Recognition System

Aidar Failovich Khusainov*

Kazan Federal University, Kazan, Kremlevskaya st., 18, Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Levo-Bulachnaya st., 36a; Aidar.f.k.21@yahoo.com

Abstract

Objectives: The article presents the experiments on the creation of different language models for the Tatar language. N-gram statistical models are used with five different smoothing techniques. **Methods**: These models can be used in various applications: machine translation systems, spell checking, etc. The study intended to use the patterns in the system of Tatar speech automatic recognition. Taking into account the specifics of the Tatar language, consisting in a rich morphology, speech recognition systems may use not only words but also the building blocks of words as basic modeling units: syllables, morphemes, etc **Finding**: The following essential elements were chosen for a complete analysis of Tatar language models development: word, morpheme, morph (statistically selected component of a nutshell), the stem and affix chain, syllable and letter. Thus, some models constructed for all combinations of 2-, 3-, 4-grams, smoothing techniques and essential elements of the language. Besides, an experiment showing the possibility of a language model development based on word classes conducted. **Conclusion**: According to performed experiment results the conclusions are made about the quality of the Tatar language grammar description, the degree of coverage lexicon, and required vocabulary volume for each type of constructed models.

Keywords: Automatic Speech Recognition, Class-Based Models, Language Model, N-Grams, Tatar Language

1. Introduction

The task of language model creation arises during a variety of tasks solutions, from spell checking to machine translation systems. In all cases, a language model is designed to describe the existing language patterns and be able to evaluate the probability of certain word sequence pronouncing.

A set of grammatical rules that would describe a phrase structure possible in the context of this subject area may act as a language model for a particular class of problems. For example, the rules of a language model may allow only the necessary repetition of numbers (depending on telephone number format) in the task of subscriber's telephone number recognition. Logical operators (e.g., "OR" operator) and a word group names are often used to record grammatical rules. Speech Recognition Grammar Specification (SRGS)[1] was developed by W3C international consortium for the unification of grammar recording for speech recognition systems.

However, it is impossible to describe all possible phrases for more general recognition tasks. In such cases, statistical n-gram model is used as a language model[2]. N-gram model assumes that the probability of a spoken word may be calculated by preceding word sequence and if you can calculate the likelihood of a nutshell appearance in a phrase one may calculate the likelihood of a whole pronounced sentence.

The probabilities of each word in different contexts are determined during a language model preparation based by large text corpora. The ratio of word sequence observations is taken for the evaluation of conditional probabilities:

$$P(w_i | w_1,\ldots, w_{i-1}) = \frac{N(w_1,\ldots, w_i)}{N(w_1,\ldots, w_{i-1})},$$

*Author for correspondence

Where $N(w_i,...,w_j)$ - the number of word sequence observations $w_i,...,w_j$ in a cospus.

From a theoretical point of view, the more information we know about already uttered words, the more accurate is the assessment of the current word likelihood. However, in practice, one has to limit an analyzed context using 1 or 2 previous words to assess the probability that is 2- or 3-grams, respectively. Limitation is caused by the computational complexity of created models: the number of 3-gram model parameters from 100 000 words may be up to (10^{20})[3]. Another significant problem is the lack of text data for full model training: many word sequences are not represented in a corpus or presented there in a small number of times, insufficient for an accurate assess of probabilities. The availability of at least one subsequence of words in a pronounced phrase which is not met during the stage of training (with a zero probability) will lead to the nullification of an entire phrase probability. To overcome this situation, the methods of probability smoothing are developed, which methods are called on behalf of their creators[4].

There are the varieties of the described above statistical n-gram model. For example, the models based on classes of words, allow increasing the dimension of n-grams by existing text corpora; trigger models simulate the relationship of word pairs in a longer context[5]. Another statistical model is the copula-based density estimation of mismatch between training and testing that can improve the accuracy of classification up to 7% for Automatic Speech Recognition (ASR) task[6].

It is worth noting separately the type of n-gram model which is based on the elements smaller than a word (particle-based model). In this case, the words are presented in the form of morphemes, and the analysis of statistical regularities takes place between them, rather than whole words. This feature is valuable in the cases of language recognition with a rich morphology, for example, for in flexional and agglutinative languages, for instance, for Finnish, Turkish, Estonian, Hungarian, and Russian.

Creation of Language Models for the Tatar Language

The Tatar language belongs to the group of agglutinative languages and has a rich morphology. During the development of standard statistical language models, a problem occurs with a large number of word forms which shall be included in a dictionary. A significant number of different affixal chains, which may follow the stem of a nutshell, make it impossible to develop a dictionary of an adequate volume with a small level of Out Of Vocabulary (OOV) words. The solution of these problems, as was mentioned in par. 1, is the reduction of a base modeled unit to an element, which is smaller than a word. The following approaches were chosen as the core ones in the current study:

- Morphemes;
- Basics and affixal chains;
- Statistically selected morphs;
- Syllables;
- Letters.

2.1 Tools for Language Model Development

SRILM (Speech Technology and Research (STAR) Laboratory) was chosen as a primary tool for Tatar language model development[7]. It includes the functionality on the development of n-gram language models, interpolation algorithms of various models, the developed model quality assessment. The work with this tool consists of three stages:

- The call of ngram-count function to calculate the number of n-grams;
- The call of ngram-count function to develop a language model based on the operation results of the first paragraph with the indication the selected model smoothing function;
- The assessment of developed quality model using a test unit via n-gram function and -ppl parameter.

Besides, the software was developed to process a text unit of the Tatar language and for process automation. They include the following main modules:

- Preliminary processing of a corpus (filtration, separation into a test and a training part);
- Splitting of words of a text corpus into the modeling elements;
- Automation of all processes: creation of language models, test performance, and the drawing up of reports according to the testing results.

Let us describe the tools used during the separation of words of a text body into the elements necessary for modeling. Thus, the dividing words into separate morphemes or stems with affixal chains were carried out using Morph a morph analyzer[8]. The isolation of "morphs" - the word parts specified statistically - using Morfessor tool[9]. The splitting of Tatar words into syllables occurred by knowledge about six main types of syllables in a language (G, SG, GS, SGS, GSS, and SGSS) without taking into account the specifics of loan words separation.

2.2 Text Corpus

Initial information for language model learning is presented by the Tatar language text corpus[10]. The fragment for work obtained after the filtration procedure (the removal of repeats, the fragments of Russian and English texts, the removal of special characters, etc.) and the separation into the learning and the test parts has the following characteristics, Table 1.

Table 1. Text body features

Corpus	Training part	Test part	Total
Number of files	200 000	17 294	217 294
Number of words	64 629 794	5 180 239	69 810 033
Number of syllables	172 193 048	13 821 430	186 014 478 (2,66/word)
Number of morphemes	102 131 309	8 149 139	110 280 448 (1,58/word)
Number of morphs	86 507 729	6 950 813	93 458 542 (1,34/word)
Number of stems and aff. chains	90 253 214	7 208 004	97 461 218 (1,4/word)
Number of letters	402 356 569	32 279 979	434 636 548 (6,23/word)
Size	834 MB	67 MB	901 MB

3. Experiment Results

Taking into account the lack of publications are so far on the issue of development and comparison of different types of statistical language models for the Tatar language, the innovative scheme drawn up so as to collect a complete assessment of factor impact on the quality of the final language model. So, individual statistical models were developed and analyzed for all combinations from the following categories:

- Element type - 6 types: word, syllable, morpheme, morph, stem + affixed chain, letter;
- The dimension of n-grams: bigrams, trigrams, 4 grams (5 grams for the model on the basis of letters);
- Model smoothing algorithm - 5 types: absolute smoothing, Good-Turing, Kneser-Ney, Witten-Bell, Kneser-Ney modified algorithm.

Such indicators assessed the quality of the developed model as the probability logarithm for the test sub corpus, perplexity (model confidence level during experimental data analysis), OOV (test assembly number of elements not included in a dictionary) and a standard size (according to the number of used n-grams).

According to the results of a model development the conclusion made that the primary and modified Kneser-Ney algorithms showed best results regarding a smoothing algorithm. The data according to perplexity parameter value by morphemic model example are presented in Table 2.

Table 2. Perplexity value for the morphemic model of Tatar language at various smoothing algorithms

Smoothing	2-gram	3-gram	4-gram
Absolute	72,6082	37,2884	29,9665
Good-Turing	81,0384	42,8639	33,1613
Kneser-Ney	**72,0003**	**36,2964**	28,6693
Witten-Bell	73,193	37,2586	29,3679
Mod. Kneser-Ney	**72,0003**	**36,2964**	**27,9677**

Among 95 developed models word model demonstrated the best quality, then the models are presented based on morphemes and stem with an affixal chain, morphs, syllables and letters, Table 3.

As we noted, one of the main issues during the statistical modeling of languages with a rich morphology is a large dictionary volume necessary to cover the vocabulary, which leads either to the reduction of system operation speed with a large vocabulary or the OOV word number increase at the dictionary volume reduction. From this perspective, the models, developed by the elements smaller than a word, showed a significant decrease

Table 3. The comparison of language models

Basic element	n-gram	Dictionary volume	Probability Log, thousand	OOV, %	Number of n-grams
Word	4	1 029 311	-12 209	1%	30,5 mln.
Morpheme	4	748 349	-12 638,7	0,5%	25,2 mln.
Morph	4	95 691	-12 772,4	0%	27,7 mln.
Syllable	4	147 957.	-14 282	0,1%	17,5 mln.
Basis+chain	4	758 752	-12 386,7	0,5%	27,5 mln.
Letter	5	51	-20 741,5	0%	3,3 mln.

Table 4. The results of model comparison with the dictionaries including 20, 50 and 200 thousand elements

Basic element	Dictionary volume	OOV	Dictionary volume	OOV	Dictionary volume	OOV
Word, 3-gram	20 thous.	17%	50 thous.	10%	200 thous.	5%
Morpheme, 3-gram	20 thous.	7%	50 thous.	5%	200 thous.	3%
Morph, 3-gram	20 thous.	3%	50 thous.	0%	200 thous.	-
Syllable, 3-gram	20 thous.	0%	50 thous.	0%	200 thous.	-
Stem+chain, 3-gram	20 thous.	5%	50 thous.	2%	200 thous.	1%

of OOV words. The dictionaries for different types of base units, consisting of 20, 50 and 200 thousand elements were developed for the experiment. The results of the developed models evaluation are presented in Table 4. The least number of items in a dictionary is necessary for a full coverage of a test sub corpus concerning the models based on syllables and statistically allocated morphs.

The Bigram model based on word classes for the vocabulary of 20 thousand elements developed in the final experiment. To select the types of words SRILM tool was also used, implementing a Brown algorithm for this purpose. The advanced model has a zero value of out-of-dictionary words, small size; however, it concedes the standard word models as a test sub corpus description. The result of 20,000-word dictionary decomposition into classes is of interest: automatically selected classes combine the words with similar meanings. For example, the names of settlements, numbers, years, names, country names, professions, etc. are allocated into separate classes.

4. Conclusions

The best quality of the Tatar language features modeling is achieved using the word n-gram models, but a significant reduction of the required dictionary size is possible using syllable and morph models with a relatively small decrease in simulation quality.

5. Summary

The necessity of language model development arises in dealing with a broad range of tasks: speech recognition, machine translation, predictive selection. In the context of agglutinative language analysis, the standard approaches based on the development of word n-gram models have serious limitations due to a rich morphology of these languages. To solve this problem during modeling, the constituent speech elements are used. In this paper, the development and the comparison of models based on words, morphemes, stems and affixal chains, morphs, syllables and letters was performed first for the Tatar language. The experiment results showed that the best modeling quality of Tatar language features is achieved by using word models, but a significant reduction in the required dictionary volume is possible by using the models of syllables and morphs at a relatively small reduction of simulation quality.

The obtained results and the models are planned to be included in the future in the Tatar continuous speech recognition system.

6. Conflict of Interest

The author confirms that the presented data do not contain any conflict of interest.

7. Acknowledgment

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

8. References

1. Speech recognition grammar specification version 1.0 [Internet]. 2014 [cited 2014 Mar 16]. Available from: https://www.w3.org/TR/speech-grammar/.
2. Manning CD, Schutze H. Foundations of statistical natural language processing. MIT – Press: Cambridge, Massachusetts; 1999. p. 704.
3. Shamna TC, Baiju KC. The emerging issues of immigrant labourers in the construction sector of Kerala. Indian Journal of Economics and Development. 2016; 4(2):1–12.
4. Srinivasan S. an exclusive cache architecture with power saving. Indian Journal of Science and Technology. 2015 Dec; 8(33):1–5.
5. Kipyatkova IS, Karpov AA. The development and research of Russian language statistical model. Trudy SPIIRAN proceedings. 2010; 12:35–49.
6. Bayestehtashk A, Shafran I, Babaeian A. Robust speech recognition using multivariate copula models. In Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing; 2016 Mar. p. 5890–4.
7. Khusainov A, Suleymanov D. Language identification system for the Tatar language. Speech and Computer, Lecture Notes in Computer Science, Springer. 2013; 8113:203–10.
8. Mathias C, Lagus K. Unsupervised discovery of morphemes. In the Proceedings of the Workshop on Morphological and Phonological Learning of Association for Computational Linguistics; 2002. p. 21–30.
9. Suleymanov D, Nevzorova OA, Khakimov B. National corpus of the Tatar language Tugan Tel: structure and features of grammatical annotation. Procedia - Social and Behavioral Sciences. 2013 Oct; 95:68–74.
10. Brown PF, Pietra VJD, Souza PVD. Class-based N-gram models of natural language. Computational Linguistics. 1992 Dec; 18(4):467–79.