On the Possibility of Embedding the Mechanism of Linguistic Anticipation into Speech Recognition Systems

Daniel Kurushin[1], Natalia Nesterova[2], Olga Soboleva[3]

[1]Perm National Research Polytechnic University, Perm, Russia, kurushin.daniel@ya.ru

[2]Perm National Research Polytechnic University, Perm, Russia, nest-nat@ya.ru

[3]Perm National Research Polytechnic University, Perm, Russia, olga.v.soboleva@gmail.com

Contribution to the State of the Art

https://doi.org/10.7251/JIT2102099K UDC: 519.76:[004.02:81'322

Abstract: The paper deals with the problems of modeling speech recognition systems. The authors proposed to use the mechanism of linguistic anticipation in the speech recognition systems. It is known that anticipation is a kind of phenomenon of anticipatory reflection, which can provide an opportunity for the subject to "look into the future." Anticipation is believed to be an effective method of improving reading technique in children as it enables to increase the speed of reading [1]. The similarity of the learning processes of the human brain and artificial neural-like algorithms allows to suggest that the inclusion of anticipation mechanisms into the operation of the speech recognition algorithm can improve the quality of the system. The paper presents the experiment carried out with the purpose to study the probability of increasing the quality of modern speech recognition systems provided that linguistic anticipation is embedded into such a system. The obtained results are discussed and possible directions for further work on this topic are considered.

Keywords: natural language processing, speech recognition systems, language models, anticipation.

Introduction

The term "anticipation" (lat. *anticipatio*) is known to be introduced into scientific discourse by V. Wundt [2] in 1880. There are other decignations for this mental mechanism, among which the following are the most well-known in Russian science: "ustanovka" (readiness to foresee and to expect future events) (D.N. Uznadze), "probabilistic forseeing" (I.M. Feigenber) and "anticipatory reflection" (P.K.Anokhin). However, today it is the anticipation that can be considered an umbrella term for "forward vision" (the literal translation of the word "anticipation" from Latin).

B.F. Lomov and E.N. Surkov (1988) define anticipation as "the ability to act and make certain decisions with a certain time-spatial anticipation in relation to expected future events" [3]. We agree that this definition of the psychological mechanism of foreseeing is rather exact, but it describes anticipation on the whole while we are interested in the so-called linguistic (language) anticipation, which is a special manifestation of the universal mechanism i.e. the effect of the subsequent language form on the form preceding it in the sequence. Linguistic anticipation allows you to predict the content of the text by the name, surname of the author, epigraph, etc., as well as to restore the missing elements of the text and guess the author's train of thought. Anticipation is considered an effective way of teaching reading techniques, because with systematic training children learn to guess a word by initial letters, a phrase by initial words, the content of the text by initial phrases, which significantly accelerates the pace of reading.

In our study, we made an attempt to consider the possibility of embedding the mechanism of linguistic anticipation into speech recognition systems. It is known that the creation of a reliable speech recognition system resistant to noise with a low error rate is one of the urgent tasks of modern robotics [4]. For the solution of this task it is necessary to involve data from applied linguistics, neurolinguistics and psycholinguistics.

Over the past decades a large number of speech recognition systems have been designed. On the one hand, there are commercial projects, such as Microsoft Speech-To-Text [5], Google Speech API [6] and Yandex SpeechKit [7]. On the other hand, there are also open source systems that allow you to control the recognition parameters yourself and integrate yourself into a third-party software. Among open-source recognizers there are such systems as Mozilla DeepSpeech [8], Kaldi [9], CMU Sphinx [10]. The main difference between open-source systems and commercial ones is their autonomy (independence from Internet connection), anonymity (achieved through source code control), as well as flexibility of settings.

However the systems mentioned above have certain problems with speech recognition often inferior to commercial ones in quality. This is quite understandable since large human and financial resources are spent on the development and maintenance of commercial systems. They have large client bases which allow to make algorithm configuration more effective. Companies such as Yandex can afford to distribute micro-tasks for training their systems to the user community through services such as "Yandex.Toloka" [11].

Problem Formulation

One of the most popular ways to improve and configure speech recognition systems with source code is to modernize the dictionary [12] and the language model of the system, because these "out of the box" entities are actually not suitable for use and they in its turn play a very important role in the recognition process.

In the course of early studies [13] it was revealed that the CMU Sphinx system achieved the best performance indicators when using medium-volume language models. But due to the fact that the medium-volume model is inherently built for a specific subject area, it can be concluded that for the correct operation of the system it is necessary to use several language models that are connected as needed.

Based on the definition of linguistic anticipation given above we understand the *anticipation* as a predictive process of loading language models into the speech recognition system relative to the intended stage of the conversation. We call a *language model* a set of coefficients formed during the training of a classifier, forming a probability distribution on a set of dictionary sequences [14].

In the course of the study three algorithms for simulation of anticipation were developed:

1) based on the classification of texts,
2) based on the transition probability matrix,
3) based on the dialog scheme.

The algorithm based on the classification of the texts is based on the idea that in the phrase uttered by a person in most cases there will be hints (perhaps difficult to notice) on the subsequent topic of the dialogue. The idea was put forward to use a classifier [13] which is given the last recognized phrase at the input, and at the output the classifier shows which of the language models the current phrase belongs to. It also shows the probabilities of correlation between the phrase and the subsequent language model.

The basis of the algorithm of the transition probability matrix was Markov chains [15]. The mathematical model of a Markov chain is based on the probability matrices of transitions of the system into possible states at each of the steps. In our case we take the existing language models as states, i.e. we have six states. It is supposed to build a transition matrix, where each cell of the matrix means the probability of transition to the next step from model to model.

The algorithm based on the dialog scheme [16] is the least labor-intensive from the point of view of implementation, but also the most unpredictable from the point of view of the result. To ensure the effectiveness of this method, it is necessary to create a reference scheme of the dialog, which is based on the language models loaded into the recognition system one at a time.

Experiment and ITS Results

The aim of the study was an experimental investigation meant to measure the performance of various speech recognition systems. Standard and predictive algorithms for using language models based on industrial volumes of data were chosen for the experiment. The initial data for building model were taken from the project "Open speech to text", published in May 2019. The collected data include 10 thousand hours of annotated oral speech collected from various Internet sources, such as books, calls, YouTube videos, etc. (a total size of over 600 gigabytes).

The test data for the experiment was a thousand pre-dictated audio files. Metrics such as WER and RTF were used to measure the performance of systems [17]. The Word Error Rate (BER), or error measure, was implemented to evaluate the accuracy of the recognition system and is described by Equation 1.

$$WER=\frac{I+D+S}{N}\quad , \qquad (1)$$

where *I* is the number of inserted characters, *D* is the number of deleted characters, *S* is the number of replaced characters, *N* is the number of characters in the recognized word. Since in some cases the numerator value may be greater than the denominator value, the *WER* value can be greater than 1.

The *RTF* metric in turn is used to estimate the time spent on speech fragment recognition. *RTF* is described by equation 2 and is the ratio of the time spent by the system on recognition. The *TPP* length of the audio file is given in seconds *LEN*.

$$RTF=\frac{TPP}{LEN} \qquad (2)$$

In the experiment the performance indicators of three autonomous speech recognition systems were compared. They are:

1) CMU Sphinx,

2) Kaldi,

3) Mozilla Deep speech.

The record of the experiment is presented in Table 1.

During the experiment it was revealed that the methods of constructing and using language models based on anticipation show a positive result only in the case of the CMU Sphinx system. The other systems worsen the indicators. Therefore one can make the following conclusion:

1) The CMU Sphinx system is based on fairly old speech recognition algorithms and therefore reducing the training sample for a given context has a positive effect on its operation.

2) More modern systems, such as Kaldi and Mozilla, are able to work with large training samples. A decrease in the sample size generally has a negative effect on classifying systems [18-20].

3) When teaching on large samples, which happens in modern systems, some patterns of dialogue development are taken into account "automatically".

Table 1 shows that the usage of Kaldi system demonstrates the best results. The percentage of recognition errors for it was only 19%, and the indicator of the time spent on recognition is considered practically a reference for open source systems.

Conclusion

During the experiments a rather contradictory result was obtained. On the one hand, the use of anticipation for some systems gives a positive effect as expected. On the other hand, for more modern systems the effect is absent or may even be regarded as negative. In addition, the probability of erroneous recognition at 19% is still high enough to talk about the applicability of the system in real conditions. In addition, it can be noted that, for example. CMU Sphinx gives a better result using the transition probability matrix than Mozilla DeepSpeech without anticipation. This suggests that it is possible to

Table 1 *– The results of the experiment*

System	CMU Sphinx		Kaldi		Mozilla DeepSpeech	
Metric	WER	RTF	WER	RTF	WER	RTF
The original algorithm	0.79	3.2	0.19	1.1	0.48	1.6
Based on text classification	0.32	1.7	0.30	1.0	0.49	1.6
Based on the transition probability matrix	0.29	1.7	0.28	1.2	0.48	1.5
Based on the dialog scheme	0.47	1.6	0.69	1.1	0.59	1.6

WER: less is better.

RTF: Less is better.

create more advanced anticipation mechanisms and more advanced mechanisms for embedding the anticipation model into the speech recognition mechanism.

One of the solutions to this problem, in our opinion, can be a formalized representation of the subject model, presented in the form of the graph which can explicate the denotatum structure of a unit of speech. The algorithm for generating such a graph was developed by the famous Russian linguist Anatoly Ivanovich Novikov [21]. In the future, this algorithm was adapted for machine use [22], and at the moment there is reason to believe that the modeling of future speech recognition systems will be based on the mechanisms of anticipation, which, in turn, will be based, among other things, on a model of a specific subject area.

References

[1] Anticipaciya: chto eto, kak ona proyavlyaetsya i korrektiruetsya. – URL: https://glpni.ru/lekapctva/antitsipatsiya-chto-eto-kak-ona-proyavlyaetsya-i-korrektiruetsya.html (data obrashcheniya: 26.11.2021)

[2] Vundt V. Osnovaniya fiziologicheskoj psihologii. Ob elementah dushevnoj zhizni. Intensivnost' oshchushcheniya. Per. s nem. M.: URSS, 2010.

[3] Lomov B.F., Surkov E.N. Anticipaciya v strukture deyatel'nosti. M., 1980.

[4] Sal'nikov I.S. Aktual'naya problematika fundamental'nyh i prikladnyh issledovanij v oblasti komp'yuternyh tekhnologij, mobil'noj svyazi, robototekhniki i iskusstvennogo intellekta // Problemy iskusstvennogo intellekta. 2018. №3 (10).

[5] Microsoft Speech-To-Text // Microsoft Speech API. – URL: https://docs.microsoft.com/en-us/azure/cognitive-services/speech/home (data obrashcheniya: 26.11.2021).

[6] Google Speech API // Cloud STT. – URL: https://cloud.google.com/speech-to-text/ (data obrashcheniya: 26.11.2021).

[7] Yandex SpeechKit // Cloud – Raspoznavanie rechi – Tekhnologii YAndeksa. – URL: https://tech.yandex.ru/speechkit/cloud/doc/guide/concepts/asr-overview-technology-docpage/ (data obrashcheniya: 26.11.2021).

[8] A TensorFlow implementation of Baidu's DeepSpeech architecture // github.com. – URL: https://github.com/mozilla/DeepSpeech (data obrashcheniya: 26.11.2021).

[9] Kaldi // Kaldi Speech Recognition Toolkit. – URL: https://github.com/kaldi-asr/kaldi (data obrashcheniya: 26.11.2021).

[10] Alimuradov A.K., Tychkov A.YU., Zareckij A.P., Kuleshov A.P., CHurakov P.P., Kvitka YU.S. Metod povysheniya effektivnosti golosovogo upravleniya na osnove komplementarnoj mnozhestvennoj dekompozicii na empiricheskie mody // Trudy MFTI. 2017. №2 (34).

[11] YAndeks.Toloka. – URL: https://toloka.yandex.ru/ (data obrashcheniya: 26.11.2021).

[12] Raspoznavanie rechi s pomoshch'yu CMU Sphinx / Habr // habr.com. – URL: https://habr.com/ru/post/267539/ (data obrashcheniya: 26.11.2021).

[13] Obzor metodov klassifikacii v mashinnom obuchenii s pomoshch'yu Scikit-Learn // tproger.ru. – URL: https://tproger.ru/translations/scikit-learn-in-python/ (data obrashcheniya: 30.11.2021).

[14] Grinin Igor' Leonidovich Rabota modeli generacii teksta s pomoshch'yu nejronnyh setej kak sostavnoj sistemy: modul'nyj analiz modul' pervyj. YAzykovaya model': rabota s tekstovymi vhozhdeniyami // Innovacii i investicii. 2020. №7.

[15] CHto takoe skrytye modeli Markova / Habr // habr.com.. – URL: https://habr.com/ru/post/135281/ (data obrashcheniya: 30.10.2021).

[16] GitHub - daniel-kurushin/three_word_english // github.com URL: https://github.com/daniel-kurushin/three_word_english/blob/master/10.jpg (data obrashcheniya: 30.10.2021).

[17] Karpov A. A., Kipyatkova I. S. Metodologiya ocenivaniya raboty sistem avtomaticheskogo raspoznavaniya rechi // Priborostroenie. 2012. №11.

[18] Mingalev A. V., Belov A. V., Gabdullin I. M., Agafonova R. R, SHusharin S. N. Raspoznavanie test-ob"ektov na teplovizionnyh izobrazheniyah // KO. 2019. №3.

[19] Novikova N. M., Noaman S. A. Komp'yuternaya model' statisticheskogo raspoznavaniya izobrazhenij // Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Seriya: Ekonomika. Informatika. 2012. №13-1 (132).

[20] Mokeev V. V., Tomilov S. V. O reshenii zadachi raspoznavaniya izobrazhenij metodom glavnyh komponent i linejnym diskriminantnym analizom // KO. 2014. №4.

[21] Novikov A.I. Semantika teksta i ee formalizaciya. – M.: Nauka, 1983.

[22] Kurushin D.S., Nesterova N.M., Chovich L.I. // XIII međunarodni naučno-stručni skup Informacione tehnologije za e-obrazovanje: Zbornik radova, Banja Luka, Izdavač: Panevropski univerzitet "APEIRON", 2021. p. 81-86.

Received: September 30, 2021
Accepted: November 29, 2021

About the authors

Daniel S. Kurushin, born in Perm, Russia, holds PhD in technical sciences. He graduated from Perm National Research Polytechnic University where he works at present as an Associate Professor at the Department of Information Technology and Automated Systems. For five years he had been working at Vienna University of Technology as a researcher. His main research interests are in artificial intelligence, robotics and systems of technical vision.

Natalia M. Nesterova, Doctor in Philology, works as a professor at the Department of Foreign Languages, Linguistics and Translation of Perm National Research Polytechnic University. She received her PhD at the Institute of Linguistics of Russian Academy of Sciences (Moscow), later she received a doctorate degree at Perm State University. Her research interests are in applied linguistics, psycholingustics and speech perception mechanisms.

Olga V. Soboleva, born in Perm, Russia, holds PhD in philology. She graduated from Perm State University and at present works as an Associate Professor at the Department of Foreign Languages, Linguistics and Translation of Perm National Research Polytechnic University. She is specializing in teaching Russian as a foreign language, information technologies in linguistics and automatic natural language processing systems.

For citation

Daniel Kurushin, Natalia Nesterova, Olga Soboleva, On the Possibility of Embedding the Mechanism of Linguistic Anticipation into Speech Recognition Systems, *JITA – Journal of Information Technology and Applications Banja Luka*, PanEuropien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 11(2021) 2:99-103, (519.76:[004.02:81'322), (DOI: 10.7251/JIT2102099K), Volume 11, Number 2, Banja Luka, December 2021 (69-140), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004