Detecting lexical stress for Russian language

Stanisław Frejlak and **Jacek Rutkowski** and **Jakub Bednarz**
s.frejlak@student.uw.edu.pl
j.rutkowski2@student.uw.edu.pl
j.bednarz2@student.uw.edu.pl
Supervisor: Spyridon Mouselinos
Uniwersytet Warszawski

Abstract

In many Slavic languages (Russian, Ukrainian, Serbo-Croatian, Bulgarian) the position of the word stress is not as predictable as in Polish - rather, it is lexically encoded. Moreover, there are many pairs of words that have distinct meanings but differ only in the position of the stress. Thus, pronunciation training is essential for learning these languages. In a self-directed study of the language, however, a reference for proper pronunciation, like a stress-annotated dictionary, may be impossible to obtain or tedious to use. Of interest are therefore tools for automatic detection and correction of mispronounced words in speech samples, as well as annotation of text fragments with stress markers. In this paper, we fine-tune a pre-trained Wav2Vec2 model to recognize stressed syllables in Russian speech.

1 Introduction

Lexical stress is a syllable-level phonological feature. Stressed syllables have tend to be spoken longer, louder and with higher pitch than the other syllables. In the Russian language the difference between stressed and non-stressed syllables is far more significant, since vowels in the non-stressed syllables are pronounced in a different way. In particular, this difference can lead to misunderstanding of an L2 Russian speaker by the natives speakers. It is highly problematic for Russian learners because in this language the lexical stress does not follow any rules unless very specific and complicated. We provide a tool to predict stress annotations for utterances consisting of one to a few sentences. Our model is able to automatically stress-annotate audio files for which a transcription without stress-annotations is given. To the best of our knowledge, no such tools exists for Slavic languages, specifically the Russian language. To address these needs and answer these questions, we present an end-to-end system for generating text transcriptions with actual and canonical stress markers. One component of the system is a speech recognition module, which we base on ‘wav2vec‘, a recent self-supervised model for generating rich audio representations. There exist versions of this backbone pretrained specifically for Russian, which we finetune on the National Corpus of Russian Language, in which are provided audio samples along with stress markers and text transcriptions.

The other part of the system is a module for detecting the actual stress in the speech sample, which we shall implement jointly with the recognition module from the audio representations.

The rest of the paper is organized as follows: Section 2 presents prior work on stress detection. Section 3 presents potential problems. Section 4 introduces the datasets and in particular our experiments and results thereof. Finally, in Section 5 we present our conclusions.

2 Related work

The attempts to automatically detect the lexical stress by exploiting some specific extracted features of the sound date back to the 60s, when Liebemann (Lieberman, 1960) developed a system of automatic lexical stress detection for bisyllabic words using a decision tree method. He examined the relevance of fundamental frequency, syllabic duration, relative amplitude and integral of the amplitude over a syllable in the context of binary automatic lexical stress recognition. He showed that the most relevant features were envelope amplitude and higher fundamental frequency.

Later on (Morton and Jassem, 1965) investigated acoustic correlates of stress by performing an experiment with unusual data which were synthetic nonsense syllables (e.g. “sisi”, “sasa”, etc.). The authors noticed that the stressed syllables tended to have longer duration and be more intense.

After some break in more important attempts, (Aull and Zue, 1985) tried to automatically identify lexical stress for isolated words using syllable duration, syllable average energy, syllable maximum pitch value and spectral change. The prediction was based on the 1-NN algorithm, each syllable-based feature vector was compared to the reference vectors. The authors used two metrics for the evaluation: syllable-based and word-based. The syllable-based accuracy was 98% and the word-based accuracy was 87%.

(Sluijter and Van Heuven, 1996) showed that among the acoustic correlates of stress was duration. They also investigated the spectral balance which was designed to measure the intensity distribution in different frequency bands. However, (Campbell and Beckman, 1997) reported that there was no significant difference in spectral balance between vowels in stressed versus unstressed syllables.

In the paper (Xie et al., 2004) was used traditional machine learning devices such as support vector machines to build classifiers for lexical stress detection. They performed the regression based on prosodic features (relating to duration, amplitude and pitch) and vowel quality features (extracted from vowel acoustic features).

(Zhao et al., 2011) considered the case of L2 English speech and tried to identify primary stress. They adopted support vector machines based on the prosodic features like loudness, vowel duration, spectral emphasis and pitch in semitone. As a dataset they used a corpus with 200 utterances spoken by 22 Taiwanese, achieving an accuracy of 88.6%. This work is quite similar to ours since it aims to detect the lexical stress in a speech of non-natives.

(Ferrer et al., 2015) used in their model not only prosodic features (such as pitch, energy and duration) but also spectral (tilt and Mel Frequency Cepstral Coefficient posteriorgrams). Results showed that the MFCC posteriorgrams helped improve the experimental performance. They achieved the promising error rates of 11% for L1-English and 20% for L1-Japanese systems, respectively. However, most of the words in this corpus had only two syllables, which made the whole task significantly easier than the general one. In particular, the expected value of the random classifier in such a case is 50%.

Until the recent paper of (Korzekwa et al., 2020) each attempt to solve the stress detection task based on audio data was founded on the artificially extracted features. The feature extraction makes the whole task somewhat clumsy and unwieldy. In the said paper the features are supposed to be learned from the raw audio data by means of the attention mechanism. The authors succeeded in exploiting the attention mechanism in order to automatically derive optimal syllable-level representation from frame-level and phoneme-level audio features. They were able to extract regions of the audio signal that are important for the detection of lexical stress which led them to satisfactory accuracy.

There are two popular variants of the supervised classification of the lexical stress: a) determining whether a syllable is stressed or not, b) classifying between the primary, secondary and no stress. In the Russian language, as opposed to English or German, there is only one type of stress which makes the task of the stress recognition a little bit easier.

As for the more specific task of the stress detection in the Russian language, the existing works, to the best of our knowledge, concern only the automatic annotation of the given text (Hall and Sproat, 2013), (Ponomareva et al., 2019). They use the monumental Zaliznyak's Dictionary (approximately 2m wordforms) which contains the annotated words with the sentence examples ase the training set.

As for predicting the lexical stress based on the textual data, one might easily come up with an idea of simply checking the lexical stress of the given word on some huge basis. However, as it was indicated in (Ponomareva et al., 2019), this approach is rather flawed due to the inability to predict stress of unknown words. Thus, two approaches were proposed to predict the lexical stress from the textual data in the Russian language. In (Hall and Sproat, 2013), the authors submitted the usage of Maximum Entropy Ranking introduced in (Collins and Koo, 2005) as well as classical machine learning methods like support vector machines. More modern technique was introduced in (Ponomareva et al., 2019). The proposed method consists in the usage of character level models and bidirectional RNN with LSTM nodes which led to the increase of the accuracy to 90%. Thus, the problem of predicting the lexical stress on the basis of the textual data may be considered as solved.

In our approach, we use wav2vec2 (Baevski et al., 2020) to derive automatically the features from the raw audio and then fine-tune it by applying the head neural network to predict the lexical stress.

3 Problem

Our goal is to predict the lexical stress of the word based on the raw audio data. In earlier works, a popular approach was to use audio data consisting of only one word. The data which we use consist of short utterances consisting of one or a few sentences available at the Russian National Corpus site (Grishina, 2006). They are not single words which leads to the potential problem that some bunches of words may be pronounced fast one after another and thus perceived as the single word. In our approach we do not deal with this problem which in the above-mentioned traditional approaches could indeed be considerable. In experiments we showed that for our model it does not matter whether the data are single or multiple words. As a matter of fact, in our approach it is completely insignificant whether we deal with one or more words. We extract from the raw data the syllables which are stressed, regardless of their position in the word. Therefore, our approach is much more practical, and our model could be used to produce stress-annotations for real-world data.

The other important problem is the limited amount of valuable data. The task of prediction or even simple indication of the correct lexical stress position being solved, our purpose was rather to provide a detector of lexical stress errors for Russian language students. That is why incorrectly pronounced words seem to be indispensable in order to properly train the model. Unfortunately, the amount of incorrect stress patterns is rather limited.(Korzekwa et al., 2020) proposed a method to solve this problem which consists in augmenting the training set with incorrectly stressed words generated with Neural Text to Speech (TTS).

In our approach this issue is not even mentioned since we believe that it is not necessary. Indeed, as the model learns from the raw data, it is irrelevant whether it deals with either correct or incorrect lexical stress. What it does is simply extracting information about the place of the lexical stress using the attention mechanism. We think therefore that training the model only on the correctly stressed words does not in fact lead to any overfitting.

4 Experimental Set-up

To produce predictions, we fine-tune Facebook's XLSR-Wav2Vec2 model (xls). This model was pre-trained on data from 53 different languages so it is not biased towards English language.

4.1 Training data

We train the model on short speech recordings taken from the Multimodal subcorpus of the Russian National Corpus. As the attention mechanism used in the Wav2Vec models requires quadratic time, we cap the length of the used recordings by 20 seconds. Moreover, to avoid using too much of bad quality recordings (i.e. in which recognizing the speech could be troublesome), we use only recordings in which there is only one speaker. For recordings with more than one speaker, there might be various problem, e.g. the speakers having very different speaking manner, or speaking at the same time.

Every example in the Multimodal subcorpus contains not only a recording of speech but also a stress-annotated transcription of the utterance. We use these transcriptions to produce training labels.

4.2 Training labels

Contrary to the most popular usage of the Wav2Vec models, we fine-tune not to produce transcriptions of utterances. Instead, we only want to classify syllables as stressed or not stressed. There are various ways in which one could decide of what the model output should look like for this task.

In our approach, we use a simple pre-processing for the ground-truth transcriptions. We remove all consonants, special characters and spaces, we replace all non-stressed vowels with token #, and we equate iotized vowels with their non-iotized equivalents, i.e. we use the same token for pairs я-а, е-э, ё-о, ю-у. We use different tokens for и and though. This way, our token set has 7 elements. For example, a sentence Ма́ленькой ёлочке хо́лодно зимо́й after pre-processing looks like а##о##о###о.

Theoretically, as the goal is to only predict which syllables are stressed, one could use a set of only two tokens corresponding to a stressed and to a non-stressed syllables. However, we expect that learning distinctions between various vowels might be helpful during the training. Moreover, transcriptions containing information about vowels are much better interpretable which is particularly useful shall the model be deployed to aid producing a stress-annotation of audio data for which a transcription without a stress-annotation is given.

On the other hand, one could train the model to produce full transcriptions of utterances and not remove the consonants. However, this task seems to be considerably more difficult than mere prediction of which syllables are stressed. Such a model would require more time to train. Moreover, in this set-up during training, the model might not focus on predicting the lexical stress as this information would have a limited contribution to the loss function.

Another approach would be to create distinct tokens for non-stressed vowels, instead of equating them like in our approach. However, this seemingly simple addition would be in fact very troublesome. In Russian language, non-stressed vowels are pronounced in a reduced manner (detailed rules in (red)). The exact rules of the reduction are very complicated and include a lot of exceptions. Moreover, the exact pronunciation might depend on such factors as the speed of speech, or a dialect of the speaker. The task of mapping vowels in non-stressed syllables to sounds is almost unfeasible. On the other hand, failing to do so and simply adding distinct tokens for every letter would introduce a sub-task for the model which is virtually impossible to learn.

On the other hand, when it comes to stressed syllables, there is almost a 1-to-1 mapping between pronounced sounds and letters, if equating letters corresponding to the iotized and non-iotized vowels (detailed rules in (iot)). Therefore, our approach seems to be the most natural choice.

4.3 Training set-up

To train our model, we use a (hug) toolkit which is built on HuggingFace's tools. We use batch size of 24. Other than that, we do not change default parameters given in the toolkit. To measure our results, we use the `CER` (Character Error Rate). `WER` (Word Error Rate) is not applicable for our task as the model is not trained to split the utterance into words. We publish our code at github (our).

We performed two training runs. The first one used a set of 187 recordings where every recording was a fragment of a different program[1] to assure diversity of data. In total, it contained recording of 44 minutes length. In the second run, we lifted the requirement of using only one example from one program. In total, there were 1806 recording with total length of 6 hours and 46 minutes.

5 Results and discussion

5.1 Results

We compare the results of the model trained on low data (44 minutes) for 62 epochs (called `LD62` in the table), the results of the model trained on big data (6 hours 46 minutes) for 4 epochs (`BD4`), 9 epochs (`BD9`) and 19 epochs (`BD19`).

We evaluate our model on a subset of training data, on an evaluation data taken from the Russian National Corpus which were not seen during the training, on a bunch of high-quality recordings of Russian lectors reading poems taken from a website (sti), and on recording of a foreigner reading the same poems. The stress-annotations for the poems were made by hand by one of the authors of this paper. Similarly, the latter evaluation dataset was recorded by one of the authors.

We present the results in table 1

	`BD4`	`BD9`	`BD19`	`LD62`
Train data	0.278	0.134	0.076	0.058
Eval data	0.324	0.188	0.187	0.259
Poems nat.	0.161	0.068	0.067	0.099
Poems for.	0.156	0.125	0.117	0.171

Table 1: Comparison of CER (Character Error Rate) of various models on various datasets

5.2 Discussion

Comparing the results on train dataset and evaluation dataset (both taken from the same source, the Russian National Corpus), we can see that the model trained on low data for 62 epochs and the model trained on big data for 19 epochs overfit to the training data. The latter model in the previous stages of the run did not suffer from this issue this much as the discrepancy between the results on the two datasets is not big.

Since as far as we know, there were no experiments performed before with such an experimental set-up, we cannot compare the achieved CER with any baseline. When looking at the transcriptions produced by our models, it can be noted that they are rather aligned with the ground-truth, however, their quality is not enough to produce correct transcriptions of speech recordings. This being said,

[1] In the Russian National Corpus, one long recording is cut into short fragments and each corpus example is such a fragment of a long recording.

recordings taken from the Russian National Corpus are also not of the best quality. Sometimes they contain background noise, sometimes the speech contains sounds which are impossible to transcribe. Therefore, one should not expect to get a CER close to 0 on this dataset.

Of particular interest are good quality audio recordings for which making a stress-annotation should be not problematic for a human annotator. A good example of such data are recordings of lectors reading poems. It should be noted that even for a significantly overfitted model BD19, CER for such data is much lower than for the training data. Analysis of transcriptions produced for these recordings suggests that the annotations could be used to aid the process of stress-annotating recordings for which a transcription without stress-annotations is given. However, a program would be needed to map the tokens from the output to the tokens of the ground-truth in the most probable way.

The last row of Table 1 is important shall the model be deployed to help in language learning process for foreigners. Lexical stress in these recordings is put on the correct syllables. However, the results show that the foreign accent is an issue. It is very interesting that for a model trained only for 4 epochs CER is very similar for poems read by a native speaker and by a foreigner, however it is not the case for models trained for more epochs. This might suggest that in the first epochs the model learns to recognize stressed syllables by general, non language specific features. On the other hand, it might be that in the next epochs it starts to use features correlated with the exact pronunciation of stressed and not stressed vowels by the native speakers. Shall the model be used for helping foreigners to learn Russian language, it might be a good idea to use recordings of foreigners' speech during training.

5.3 Mistake types

The most common type of mistake in the transcriptions produced by the big data model is omitting some of the not stressed syllables. Therefore, oftentimes the predicted transcriptions are a bit shorter than the ground-truth ones.

Less common types of mistakes are: omitting a stressed syllable, predicted a wrong vowel for a stressed syllable (e.g. confusing э with и or о with у), or erroneously recognizing a not stressed syllable as the stressed one. The latter type is especially prevalent in case of one syllable words which are not stressed. Such mistakes are difficult to avoid as in the speech there is sometimes no clear distinction between a stressed and not stressed syllable.

6 Conclusions and future work

Our paper shows a novel usage of the Wav2Vec pretrained model. With limited data and computing power, we managed to fine-tune the model to predict stress-annotations of utterances with a decent accuracy.

Other experimental set-ups might be tried in the future. One could try to use distinct tokens for various not different vowels without necessarily aiming at implementing all the vowel reduction rules. This might help to tackle the most common mistake type of omitting some of the not stressed syllables.

Our experiments showed a big performance gain when using a bigger dataset. In the future work, it would be a good idea to try to collect even more data. One could also try to collect more diverse data, e.g. include theatrical speech or recordings of foreigners' speech.

Another interesting experiment would be to train the model on data of various languages and check if it would outperform models trained on a specific language.

With more data and computational power, one could expect to achieve results which would facilitate producing high-quality transcriptions of utterances for which not stress-annotated transcription are given. Since there is much more such data publicly available than data with stress-annotations, such a model could be used to produce big datasets which might be helpful in various NLP tasks.

References

Орфоэпия. Произношение безударных гласных звуков.

Орфоэпия. Произношение гласных звуков.

Папины сказки.

Hugging sound.

Stress detection.

A Aull and Victor Zue. 1985. Lexical stress determination and its application to large vocabulary speech recognition. In *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 1549–1552. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Nick Campbell and Mary Beckman. 1997. Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications*.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda. 2015. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, 69:31–45.

Elena Grishina. 2006. Russian national corpus.

Keith Hall and Richard Sproat. 2013. Russian stress prediction using maximum entropy ranking. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 879–883.

Daniel Korzekwa, Roberto Barra-Chicote, Szymon Zaporowski, Grzegorz Beringer, Jaime Lorenzo-Trueba, Alicja Serafinowicz, Jasha Droppo, Thomas Drugman, and Bozena Kostek. 2020. Detection of lexical stress errors in non-native (l2) english with data augmentation and attention. *arXiv preprint arXiv:2012.14788*.

Philip Lieberman. 1960. Some acoustic correlates of word stress in american english. *The Journal of the Acoustical Society of America*, 32(4):451–454.

John Morton and Wiktor Jassem. 1965. Acoustic correlates of stress. *Language and speech*, 8(3):159–181.

Maria Ponomareva, Kirill Milintsevich, Ekaterina Chernyak, and Anatoly Starostin. 2019. Automated word stress detection in russian. *arXiv preprint arXiv:1907.05757*.

Agaath MC Sluijter and Vincent J Van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical society of America*, 100(4):2471–2485.

Huayang Xie, Peter Andreae, Mengjie Zhang, and Paul Warren. 2004. Detecting stress in spoken english using decision trees and support vector machines. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation-Volume 32*, pages 145–150. Citeseer.

Junhong Zhao, Hua Yuan, Jia Liu, and Shanhong Xia. 2011. Automatic lexical stress detection using acoustic features for computer assisted language learning. *Proc. APSIPA ASC*, pages 247–251.