Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

AUTOMATIC DETECTION OF STRESS IN RUSSIAN BY MISSPELLINGS IN CORPUS

Alexeyevsky D. A. (dalexeyevsky@hse.ru),
Lipunova A. E. (bennet.ray@yandex.ru)

Higher School of Economics, Moscow, Russia

While writing words people sometimes make mistakes. A hypothesis was put forward that mistakes in vowels occur more frequently in unstressed positions than in stressed positions. Here we employ it to automatically detect stressed positions in previously unknown words by searching for words that have only one vowel that is never misspelled within selected corpus. We investigate and report possibilities and limitations of this approach for stress detection.

The approach to detect word stress presented here is usable for adding word stress in neologisms and proper nouns into existing words stress dictionaries. Another possible application of the proposed approach is to detect stress changes in usus and watch the spread of changed stress diachronically and within different societies.

Here we present an implementation of the program to detect stress in corpus with mistakes. We tested the program on corpus of Russian Twitter texts and report reasonable precision (67%) in cases where detection was possible (approximately 1% since not all possible vowel misspellings were presented in most word-forms and due to imposed restrictions on encountered word-forms). While this already gives hope to make a useful tool for annotating neologism, we note that this approach for detecting was previously uninvestigated and a lot improvements are yet possible.

Key words: stress, stress dictionary, parsing, misspellings, misprints

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ УДАРНЫХ ПОЗИЦИЙ В СЛОВАХ РУССКОГО ЯЗЫКА ПО КОРПУСУ ОШИБОК

Алексеевский Д. А. (dalexeyevsky@hse.ru),
Липунова А. Е. (bennet.ray@yandex.ru)

Высшая школа экономики, Москва, Россиия

В работе рассматривается гипотеза о том, что в безударных позициях слова носитель чаще совершает ошибки, поскольку русский язык имеет фонематическое письмо, где каждому письменному символу может соответствовать несколько фонем. Мы предлагаем использовать эту гипотезу для того, чтобы порождать словарь потенциальных ударений в слове, основываясь на случях в корпусе, когда опечатки были совершены во всех позициях, кроме одной. Мы рассматриваем и представляем возможности и ограничения, связанные с таким подходом автоматической расстановки ударений.

Такой подход к применению ударения ценен для определения ударений в неологизмах и именах собственных и в последующем добавлении их в соответствующие словари, для определния изменения ударения в узусе, для изучения распределения расстановки ударения диахронически внутри определенных сообществ людей.

Также мы представляем реализацию программы для расстановки ударений для корпуса, содержащего ошибки. Мы протестировали программу на русскоязычном корпусе Твиттера. Программа демонстрирует низкую полноту, предсказывая ударения менее, чем для 1% слов. Однако при этом демонстрирует высокую точность: 67% для словарных и несловарных слов. Таким образом, такие результаты дают надежду создать на основе нашей программы полезное приложение для разметки неологизмов. Стоит добавить, что такой подход к автоматической расстановке ударений нигде ранее не рассматривался и таким образом он открывает широкое поле для дальнейших исследований.

Ключевые слова: ударение, словарь ударений, опечатки, ошибки, орфографические ошибки

1. Introduction

Mistakes are widely distributed in written speech. Meanwhile, some types of mistakes may indicate phonetic features of words, including the stress position. As the basis of our work we admit a hypothesis that mistakes in vowels occur more frequently in unstressed positions than in stressed positions [5]. Here we employ this hypothesis to detect stressed positions by searching for words with only one unchanging vowel within selected corpus. We examine and report about possibilities and limitations of this approach for stress detection.

Within this paper we consider only mistakes in vowels and do not take into account mistakes resulting from misprints or intentional distortion of spelling rules (e.g. “превед”).

One important property of Russian language is unpredictability of stress positions in words. Detection of stressed position in words is an important part of language knowledge, notable application being speech synthesis, rhyming dictionaries, poetic text analyzers, language learning materials, automatic poetry generators [12]. Stress detection is typically based on dictionaries with simple rules for unknown words.

The approach to detect word stress presented here is usable for adding word stress in neologisms and proper nouns into existing words stress dictionaries. Another possible application of the proposed approach is to detect stress changes in usus and watch the spread of changed stress diachronically and within different societies.

2. Related works

Russian language script is phonematic. Phonematic scripts give rise for orthographical mistakes in cases where one phoneme may be represented in various graphics [5].

Unfortunately, not many computational linguistic works were published within topic of accentology for Russian language. Several programs were created to automatically detect stress position in words and to use knowledge about word stress for solving other tasks. Perhaps the best described among publicly available resources for automatic stress detection is stress detector of morpher.ru [9]. The program employs odict.ru [8] morphological dictionary. Odict is a crowdsourcing platform for updating the dictionary with new words. The dictionary is based on Zaliznyak dictionary [10] both in sense of structure and as a lexical basis.

Most of speech synthesis programs for Russian language solve the task of assigning stress position using specialized dictionaries or databases. Such dictionary would assign known words to word stress paradigms. Among such programs are “Acapela”, “Vocalizer” и “VitalVoice”. Typically such programs determine stressed syllable for unknown words using heuristics. Such heuristics may vary from very simple e. g. put stress on the middle syllable of the word, to rather advanced. For example, “VitalVoice” [6] attempts to detect widely distributed prefixes like “электро-” and exclude or assign stress positions by analyzing morphemes that attract stress.

Several programs exist to detect stress position in unknown words using machine learning. For example, this has been done for the Romanian language in [2]. No such programs are published for the Russian language.

One direction of research in word stress detection is updating the existing word stress dictionaries. Traditionally word stress dictionaries were only updated manually [1]. The upside of this approach is high quality of the resulting dictionaries. It comes with traditional downside of being very expensive. Somewhat less expected downside is that lexicographers tend to prefer conservative positions of stress in cases when usus change. Published human-generated stress dictionaries also seem to lack many of new words.

An approach was suggested in [3] to detect stress by analysing rhymes and metrics in poetic corpora. To our knowledge the approach has not been put to use yet.

3. Materials and methods

Our work relies on detecting mistakes in corpus. In order for stress detection to function properly a large corpus containing uncorrected texts within a limited timeframe is required. One such corpus is Russian corpus of Twitter [7]. Corpus contains 17,639,674 tweets spanning three months January through March 2014. The corpus only contains tweets longer than 40 characters. The corpus has 160,020,610 tokens.

In the following description of our approach we use several specialized terms to simplify the explanations. These are:

Consonant mask of a token is the token with each vowel replaced with wildcard symbol. Consonant mask allows to group similar words while distinguishing words that differ only in number of vowels, e.g. “шкал” and “шкала” have different consonant masks while “ходило”, “ходила”, “ходулю”, and “худели” have the same consonant mask “х*д*л*”. Since our work deals with spelling mistakes in vowels, all words with such mistakes have the same consonant mask, e.g. “госпожа” and “госпажа”. However, in some cases (mistakes affecting consonants, adding or removing vowels) various spellings of one word-form might have different sequences of consonants. Since we are interested in detecting stress through mistakes in vowels we ignore mistakes that modify the consonant mask.

Given two words having the same consonant mask it is our next task to tell if one of the words is misspelled version of the other e.g. “жарка” might be a misspelling of “жарко”, while “жирке” might not. Examining each modified vowel in a pair of words we define possible substitution as such modification of vowel that can be produced by informant due to ambiguity in mapping phoneme to graphics.

We created a set of rules defining the possible substitutions based on works [13], [4] that describe various realizations of vowel sounds during qualitative reduction in different positions within word. We then modified the set of rules by testing if different substitutions indeed produce misspelled words. More changes to rules were produced while testing the algorithm of stress detection by estimating the resulting clusters of potentially misspelled words:

Table 1. Possible vowel substitution depending on the vowel itself and its position in the word. In the left column all russian vowels are presented. In the first raw various possible positions in the word given. At the intersection of those two values possible substitutions of the given vowel in the desired position located

	Word start	After hard consonants	After ж, ш, ц	After soft consonants	After ч & щ	At the end of word
а	а, о	а, о	а, е, и, э, ы	-	а, е, и, э	-
е	е, и, я	-	е, и, э, ы	е, и	е, и	-
и	е, и, э	-	е, и, э, ы	е, и	е, и	-
о	а, о	а, о	а, о	-	а, е, и, о, ы	а, о
у	-	-	-	-	-	-

	Word start	After hard consonants	After ж, ш, ц	After soft consonants	After ч & щ	At the end of word
ы	-	э, ы	-	-	-	-
э	е, и, э, ы	-	е, и, э, ы	-	-	-
ю	-	-	-	-	-	-
я	е, и, я	-	-	е, и, я	-	-

Let us consider as an example unstressed vowel "a" in word "часы". According to possible substitution rules it is possible to expect the following misspellings of the word: "часы", "чесы"*, "чисы"* и "чэсы"*. Similarly the rules predict the following misspelling of word "жалеть": "жалеть", "желеть"*, "жилеть"*, "жэлеть"* и "жылеть"*.

Token nest is a set of tokens that are considered similar by our algorithm. The exact definition of similarity varies at different stages of processing.

4. The approach

First step of preprocessing is to tokenize corpus and build token frequency dictionary. The 160 million token Twitter corpus has 1.9 million unique tokens.

Next step is to build consonant mask for each token and group tokens with the same mask into preliminary token nests.

As described above, such nests are likely to contain completely unrelated words like "плита", "плиты", "плата" and "плато". The next task is to split the token nests into smaller ones so that only tokens that are possible misspellings of each other are in the same nest. To achieve this we build a graph for each nest such that each token in the nest is a vertex in the graph. We connect two tokens with an edge if one of the words might be a misspelled version of the other. Each pair of vowels in the two words must form a possible substitution in order to create an edge. In the resulting graph we detect token nest as a clique using Bron—Kerbosch algorithm.

The next step is to detect the stress position within the given token nest. For an every token nest the stress is detected by searching words where only one vowel in word remain the same while the others change. In these cases we might count this vowel as a stressed one with reliable precision. For cases where several vowels remain unchanged in the different spellings of the word-forms, the program makes assumptions regarding stress position based on the fact that any of unchanging vowels can be a stressed one. The cases where all vowels of the word-form change in its different spellings are also possible, however, this cases indicate the presence of homonyms or misspellings of the stressed vowel.

5. Results

We created a proof of concept program[1] which is capable of detecting stress position within the given corpus of russian texts with derivation from literary stressing norm.

[1] Available at https://bitbucket.org/Lipunova_A/autostressdetection

We train the program on russian part of Twitter corpus collected from January to March 2014 with 160.0 million word-form volume. Twitter, but not blog or mass media articles corpus is chosen as the most suitable one due to wide language variety presented. From the corpus 1.9 million unique word-forms were gathered and formed into 1.88 million token nests.

We assess performance of the program for dictionary words and unknown words.

To test performance on the dictionary words we restricted the token nests produced by selecting only the nests satisfying the following criteria:

1. At least one of tokens in the token nest has frequency above 1,000 (approx 6.25 IPM). This criterion allowed us to avoid the majority of the words with misprints and uncommon words;
2. Exactly one token in the nest is found in the dictionary. This limitation allowed us to avoid cases where two different word-forms fall under a single consonant mask and met the rules of possible substitutions, e.g. "ворона" vs "варана". If this limitation was not adopted, the program would refer both words to the same word-form and incorrectly identify the stress position;
3. The consonant mask for the nest has at least two vowels, since there are no unstressed vowels in monosyllabic words;
4. The most frequent word is found in the stress dictionary. This criterion was adopted in order to automatically check the results of the program work;
5. The program predicted some stress in the nest, since nests with no predictions (cases with only one word-form in the nest or nests with all vowels changing) do not carry any meaningful information for our research.

Since not all possible vowel misspellings were presented in most word-forms, and also due to these limitations only 795 token nests remained in the set, of those the stress position was correctly predicted for 540, yielding 67% precision.

Additionally we performed a manual assessment of the program performance on infrequent word set rich with neologisms. For the assessment we filtered the entire set of token nests to contain at least one token with IPM between 1 and 5. We selected 1000 of such nests for manual assessment. The program predicted some stress position on only 87 of 1000 tokens yielding very low recall of 0.9% due to the fact that we looked only at nests with exactly defined stress position. The list includes dissyllabic and rarely trisyllabic words in which only one vowel remained unchanged. Of the 87 tokens with predicted stress 58 were assessed by the expert to have the correct stress, yielding the same high 67% precision.

At the end of the text corpus processing we can obtain four different types of results:

1. All vowels in the nest remain unchanging, since only one word-form variation presented in the nest, e.g. "авангард" (frequency=1,070);
2. Some of the vowels in the nest changing, e.g. "абсолютное" (frequency=289) vs "обсолютное" (frequency=1); "младенца" (frequency=774) vs "млоденца" (frequency=4). In this case the program can make an assumption about stress position. However, due to misspellings in the stressed vowel or homonymy some errors may occur, e.g. "побольше" (frequency=4,489) vs "побальше" (frequency=1).

3. Only one vowel in the nest remain unchanging, e.g. “вполне” (frequency=7,312) vs “впалне” (frequency=1); “отличноoooo” (frequency=8) vs “отличнааааа” (frequency=2) vs “атличнааааа” (frequency=2). Here it is possible to uniquely determine the stress position in the word-form. As in the previous type of results, some errors related to incorrect spelling of stressed vowel or homonymy may occur, e.g. “многих” (frequency=7,588) vs “мнагих” (frequency=1); “воры” (frequency=1,121) vs “вары” (frequency=66). Incorrect attribution of various word-forms to a single nest may affect the result precision either, e.g. “смешно” (frequency=14,765) vs “смешна” (frequency=74). Though in some cases such mistakes may not affect on the correctness of the results, e.g. “админом” (frequency=175) vs “админам” (frequency=74) vs “одминам” (frequency=1).
4. All of the vowels in the nest changing, e.g. “вором” (frequency=179) vs “варам” (frequency=3); “вошло” (frequency=795) vs “вошла” (frequency=1,903) vs “вашло” (frequency=1).

6. Discussion

Rather predictably ambiguous words provide obstacles for word stress detection. In terms of this work we ignore problems with two dictionary word-forms in token nest (e.g.: “ворон” vs “варан” vs “воран”*). Such cases were excluded from tests on the dictionary words, but were unaccounted for in manual verification and were counted as the program errors. In this work we don’t cope with homonymy cases (like “зАмок”—“замОк”), but it could be useful in further investigations. In future work on this topic could be considered a problem of misspelled word-forms or word-forms with intentional distortion of spelling rules, as it will significantly improve precision of the predictor.

In this paper we did not consider other phonetic misspelling in vowels, as, for example, excessive designation of hardness and softness after hushing sounds (e.g., “чудо” vs “чюдо”; “жизнь” vs “жызнь”; “хорошую” vs “хорошюю”), since in this positions the variability of vowel presented in token nests carries no information about the stress position in the word. In terms of this work spellings “чюдо” and “чудо” belong to different token nests, qualitative reduction cases are not examined. In addition, we did not take into account cases where the vowel sounds contain the [j] sound, e.g. “ёжик” и “йожик”; “яблоко” и “йаблоко”.

There are a few cases where the rules for possible substitutions fail. For example, manual testing revealed that substitutions of letters “o” to “a” in the end of the word occur not only in case of misspellings in the unstressed vowel, but as an expression of different word-forms of one word (cf.: “про него” vs “пра него”, “холодно” vs “холодна”). Although the last example can also be the wrong spelling of one word-form. Therefore, such cases require a detailed study and analysis of the morphological characteristics of the words. This calls to put more research into investigating interaction between morphology, orthography and stress.

On this moment Russian Twitter corpus we use includes unfiltered data like kazakh tweets (~250 tweets in whole corpus). This obviously adds noise to the results. Proper corpus filtering might give a slight improvement of the results.

We currently employ no user account identification in the program. However, preliminary classification of users into well-literate, poor-literate and additionally into adepts of different subcultures might yield noticeable improvement to the results and aid in coping with homonymy.

It is possible to merge the proposed algorithm with other means of stress detection in some kind of voting algorithm. For instance the stretched vowel is expressing the emotional status and in the most cases is the stressed one (e.g.: “хооолодно”, “приве-е-ет”) and it is possible to add such cases to the token nests with some additional preprocessing. Another promising approach to combine with the program is to add morphemic-associated stress arrangement, as suggested in [11].

One interesting point is that the algorithm presented can be modified to aid in neologism detection. The first step of such modification would be to use Levenstein distance instead of consonant mask clustering.

7. Conclusions

To sum up, we examined a method of stress detection for which no known equivalents exist. The method uses misspelling cases in unstressed vowels in large arrays of text. In this work we used Russian Twitter corpus as a test big data selection due to closeness of examples to the native live language and wide variety of possible misspellings presented.

To investigate the problem, stress autodetection program is developed.

To test gathered results, we performed two testing sessions—firstly, auto-compare in between stress dictionary and the program results and secondly, the manual one to check the results on random selection of word-forms. Since not all possible vowel misspellings were presented in most word-forms, and also due to some limitations we imposed, the recall appeared to be not so high (1%). In the first case we get 67% of successful results of unambiguous selection proven by dictionary matches and in the second case we get the same 67% proven manually, which is a good result for using only one method to detect the stress and may be improved in further investigations while using several methods and additional features combined. Additionally, we consider the high precision of the predictor to be a very strong supporter of the hypothesis that unstressed vowels are more prone to misspelling.

This work might be extremely useful in such fast-spreading technologies as speech synthesis, rhyming dictionaries, poetic text analyzers, language learning materials and similar linguistic areas.

References

1. *Ageenko F. L., Zarva M. V.* (2001), Dictionary of Word Stresses in the Russian Language. [Russkoe slovarnoe udarenie. Slovar naricatelnyh imen.], NTs ENAS Publisher [ENAS], Moscow.

2. *Balc D., Beleiu A., Potolea R., Camelia L.* (2015), A learning-based Approach for Romanian Syllabification and Stress Assignment. Intelligent Computer Communication and Processing (ICCP), 2015 IEEE International Conference, Cluj-Napoca, pp. 37–42.
3. *Belikov V., Selegei V. P., Sharov S. A.* (2012), Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proektu Generalnogo internet-korpusa russkogo yazyka (GIKRYA)]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2012” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2012”], Bekasovo, pp. 37–49
4. *Bryzgunova, E. A.* (1963), The Practical Course of Russian Phonetics and Intonation [Prakticheskaya fonetika i intonaciya russkogo yazyka], Publishing House of MSU [MGU], Moscow, pp. 136–140
5. *Glazkov A. V.* (1999), Orthographic mistake as an object of linguistic study [Orfograficheskaya oshibka kak predmet lingvisticheskogo issledovaniya], Scientific works, Moscow Culturological Lyceum №1310. Series Philology [Uchonye zapiski Moskovskogo kulturologichskogo liceya №1310. Seriya Filologiya], vol. 3, pp. 29–30
6. *Homickevich O. G., Rybin S. V., Talanov A. O., Oparin I. V.* (2008), Automatic detection of stress position in unknown words in speech synthesis systems [Avtomaticheskoe opredelenie mesta udareniya v neznakomyh slovah v sisteme sinteza rechi], Materials of XXXVI International Philological Conference [Materialy XXXVI Mezhdunarodnoj filologicheskoj konferencii], Sankt-Petersburg.
7. *Rubcova Yu. V.* (2015), Constructing a corpus for sentiment classification training [Postroenie korpusa tekstov dlya nastrojki tonovogo klassifikatora], Software products and systems [Programmnye produkty i sistemy], vol. 1 (109), pp. 72–78.
8. *Slepov S.* Open Russian grammatic dictionary [Otkrytyj grammaticheskij slovar’ russkogo jazyka] (electronic document): http://odict.ru
9. *Slepov S.* Stresses placement programm [Programma rasstanovki udarenij] (electronic document): http://morpher.ru
10. *Zaliznyak A. A.* (1977), Russian Grammar Dictionary [Grammaticheskiy slovar russkogo yazyka], Russian language [Russkij yazyk], Moscow.
11. *Zaliznyak A. A.* (1985), From Proto-Slavic to Russian accentuation [Ot praslavjanskoj akcentuacii k russkoj], Science [Nauka], Moscow.
12. *Zelenkov Yu. G., Zobnin A. I., Maslov M. Yu., Titov V. A.* (2014), Ilya Segalovich and Development of Ideas of Computational Linguistics to Yandex [Ilya Segalovich i razvitie idej kompyuternoj lingvistiki v Yandekse], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2014” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2014”], Bekasovo, pp. 775–786.
13. *Zemskaya E. A.* (2006), Russian colloquial speech. Linguistic analysis and learning problems. [Russkaya razgovornaya rech: lingvisticheckij analiz i problemy obucheniya: uchebnoe posobie], Science [Nauka], Moscow.