Russian Stress Prediction using Maximum Entropy Ranking

Keith Hall **Richard Sproat**
Google, Inc
New York, NY, USA
{kbhall,rws}@google.com

Abstract

We explore a model of stress prediction in Russian using a combination of local contextual features and linguistically-motivated features associated with the word's stem and suffix. We frame this as a ranking problem, where the objective is to rank the pronunciation with the correct stress above those with incorrect stress. We train our models using a simple Maximum Entropy ranking framework allowing for efficient prediction. An empirical evaluation shows that a model combining the local contextual features and the linguistically-motivated non-local features performs best in identifying both primary and secondary stress.

1 Introduction

In many languages, one component of accurate word pronunciation prediction is predicting the placement of lexical stress. While in some languages (e.g. Spanish) the lexical stress system is relatively simple, in others (e.g. English, Russian) stress prediction is quite complicated. Much as with other work on pronunciation prediction, previous work on stress assignment has fallen into two camps, namely systems based on linguistically motivated rules (Church, 1985, for example) and more recently data-driven techniques where the models are derived directly from labeled training data (Dou et al., 2009). In this work, we present a machine-learned system for predicting Russian stress which incorporates both data-driven contextual features as well as linguistically-motivated word features.

2 Previous Work on Stress Prediction

Pronunciation prediction, of which stress prediction is a part, is important for many speech applications including automatic speech recognition, text-to-speech synthesis, and transliteration for, say, machine translation. While there is by now a sizable literature on pronunciation prediction from spelling (often termed "grapheme-to-phoneme" conversion), work that specifically focuses on stress prediction is more limited. One of the best-known early pieces of work is (Church, 1985), which uses morphological rules and stress pattern templates to predict stress in novel words. Another early piece of work is (Williams, 1987).

The work we present here is closer in spirit to data-driven approaches such as (Webster, 2004; Pearson et al., 2000) and particularly (Dou et al., 2009), whose features we use in the work described below.

3 Russian Stress Patterns

Russian stress preserves many features of Indo-European accenting patterns (Halle, 1997). In order to know the stress of a morphologically complex word consisting of a stem plus a suffix, one needs to know if the stem has an accent, and if so on what syllable; and similarly for the suffix. For words where the stem is accented,

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 879–883,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

	acc	unacc	postacc
Dat Sg	гор'оху	г'ороду	корол'ю
	gor’oxu	g’orodu	korolj’u
Dat Pl	гор'охам	город'ам	корол'ям
	gor’oxam	gorod’am	korolj’am
	’pea’	’town’	’king’

Table 1: Examples of accented, unaccented and postaccented nouns in Russian, for dative singular and plural forms.

this accent overrides any accent that may occur on the suffix. With unaccented stems, if the suffix has an accent, then stress for the whole word will be on the suffix; if there is also no stress on the suffix, then a default rule places stress on the first syllable of the word. In addition to these patterns, there are also postaccented words, where accent is placed uniformly on the first syllable of the suffix — an innovation of East and South Slavic languages (Halle, 1997). These latter cases can be handled by assigning an accent to the stem, indicating that it is associated with the syllable *after* the stem. Some examples of each of these classes, from (Halle, 1997, example 11), are given in Table 1. According to Halle (1997), considering just nouns, 91.6% are accented (on the stem), 6.6% are postaccented and 0.8% are unaccented, with about 1.0% falling into other patterns.

Stress placement in Russian is important for speech applications since over and above the phonetic effects of stress itself (prominence, duration, etc.), the position of stress strongly influences vowel quality. To take an example of the lexically unaccented noun город *gorod* ‘city’, the genitive singular г'орода *g’oroda* /gˈɔrədə/ contrasts with the nominative plural город'а *gorod’a* /gərʌdˈa/. All non-stressed /a/ are reduced to schwa — or by most accounts if before the stressed syllable to /ʌ/; see (Wade, 1992).

The stress patterns of Russian suggest that useful features for predicting stress might include (string) prefix and suffix features of the word in order to capture properties of the stem, since some stems are (un)accented, or of the suffix, since some suffixes are accented.

4 Maximum Entropy Rankers

Similarly to Dou et al. (2009), we frame the stress prediction problem as a ranking problem. For each word, we identify stressable vowels and generate a set of alternatives, each representing a different primary stress placement. Some words also have secondary stress which, if it occurs, always occurs before the primary stressed syllable. For each primary stress alternative, we generate all possible secondary stressed alternatives, including an alternative that has no secondary stress. (In the experiments reported below we actually consider two conditions: one where we ignore secondary stress in training and evaluation; and one where we include it.)

Formally, we model the problem using a Maximum Entropy ranking framework similar to that presented in Collins and Koo (2005). For each example, x_i, we generate the set of possible stress patterns $\mathcal{Y}_i$. Our goal is to rank the items in $\mathcal{Y}_i$ such that all of the valid stress patterns $\mathcal{Y}_i^*$ are above all of the invalid stress patterns. Our objective function is the likelihood, $\mathcal{L}$ of this conditional distribution:

$$\mathcal{L} = \prod_i p(\mathcal{Y}_i^* | \mathcal{Y}_i, x_i) \tag{1}$$

$$\log \mathcal{L} = \sum_i \log p(\mathcal{Y}_i^* | \mathcal{Y}_i, x_i) \tag{2}$$

$$= \sum_i \log \frac{\sum_{y' \in \mathcal{Y}_i^*} e^{\sum_k \theta_k f_k(y', x)}}{Z} \tag{3}$$

Z is defined as the sum of the conditional likelihood over all hypothesized stress predictions for example x_i:

$$Z = \sum_{y'' \in \mathcal{Y}_i} e^{\sum_k \theta_k f_k(y'', x)} \tag{4}$$

The objective function in Equation 3 can be optimized using a gradient-based optimization. In our case, we use a variety of stochastic gradient descent (SGD) which can be parallelized for efficient training.

During training, we provide all plausibly correct primary stress patterns as the *positive* set

$\mathcal{Y}_i^*$. At prediction-time, we evaluate all possible stress predictions and pick the one with the highest score under the trained model Θ:

$$\arg\max_{y' \in \mathcal{Y}_i} p(y'|\mathcal{Y}_i) = \arg\max_{y' \in \mathcal{Y}_i} \sum_k \theta_k f_k(y', x) \quad (5)$$

The primary motivation for using Maximum Entropy rather the ranking-SVM is for efficient training and inference. Under the above Maximum Entropy model, we apply a linear model to each hypothesis (i.e., we compute the dot-product) and sort according to this score. This makes inference (prediction) fast in comparison to the ranking SVM-based approach proposed in Dou et al. (2009).

All experiments presented in this paper used the Iterative Parameter Mixtures distributed SGD training optimizer (Hall et al., 2010). Under this training approach, per-iteration averaging has a regularization-like effect for sparse feature spaces. We also experimented with L1-regularization, but it offered no additional improvements.

5 Features

The features used in (Dou et al., 2009) are based on trigrams consisting of a vowel letter, the preceding consonant letter (if any) and the following consonant letter (if any). Attached to each trigram is the stress level of the trigram's vowel — 1, 2 or 0 (for no stress). For the English word *overdo* with the stress pattern 2-0-1, the basic features would be *ov:2*, *ver:0*, and *do:1*. Notating these pairs as $s_i : t_i$, where s_i is the triple, t_i is the stress pattern and i is the position in the word, the complete feature set is given in Table 2, where the stress pattern for the whole word is given in the last row as $t_1t_2...t_N$. Dou and colleagues use an SVM-based ranking approach, so they generated features for all possible stress assignments for each word, assigning the highest rank to the correct assignment. The ranker was then trained to associate feature combinations to the correct ranking of alternative stress possibilities.

Given the discussion in Section 3, plausible additional features are all prefixes and suffixes of the word, which might be expected to better capture some of the properties of Russian stress patterns discussed above, than the much more local features from (Dou et al., 2009). In this case for all stress variants of the word we collect prefixes of length 1 through the length of the word, and similarly for suffixes, except that for the stress symbol we treat that together with the vowel it marks as a single symbol. Thus for the word *gorod'a*, all prefixes of the word would be *g*, *go*, *gor*, *goro*, *gorod*, *gorod'a*.

Substring	s_i, t_i s_i, i, t_i
Context	s_{i1}, t_i $s_{i1}s_i, t_i$ s_{i+1}, t_i s_is_{i+1}, t_i $s_{i1}s_is_{i+1}, ti$
Stress Pattern	$t_1t_2...t_N$

Table 2: Features used in (Dou et al., 2009, Table 2).

vowel	а,е,и,о,у,э,ю,я,ы
stop	б,д,г,п,т,к
nasal	м,н
fricative	ф,с,ш,щ,х,з,ж
hard/soft	ъ,ь
yo	ё
semivowel	й,в
liquid	р,л
affricate	ц,ч

Table 3: Abstract phonetic classes used for constructing "abstract" versions of a word. Note that etymologically, and in some ways phonologically, в *v* behaves like a semivowel in Russian.

In addition, we include prefixes and suffixes of an "abstract" version of the word where most consonants and vowels have been replaced by a phonetic class. The mappings for these are shown in Table 3.

Note that in Russian the vowel ё /jɔ/ is *always* stressed, but is rarely written in text: it is usually spelled as е, whose stressed pronuncation is /(j)ɛ/. Since written е is in general ambiguous between е and ё, when we compute stress variants of a word for the purpose of ranking, we include both variants that have е and ё.

6 Data

Our data were 2,004,044 fully inflected words with assigned stress expanded from Zaliznyak's *Grammatical Dictionary of the Russian Language* (Zaliznyak, 1977). These were split randomly into 1,904,044 training examples and 100,000 test examples. The 100,000 test examples obviously contain no *forms* that were found in the training data, but most of them are word forms that derive from lemmata from which some training data forms are also derived. Given the fact that Russian stress is lexically determined as outlined in Section 3, this is perfectly reasonable: in order to know how to stress a form, it is often necessary to have seen other words that share the same lemma. Nonetheless, it is also of interest to know how well the system works on words that do not share any lemmata with words in the training data. To that end, we collected a set of 248 forms that shared no lemmata with the training data. The two sets will be referred to in the next section as the "shared lemmata" and "no shared lemmata" sets.

7 Results

Table 4 gives word accuracy results for the different feature combinations, as follows: Dou et al's features (Dou et al., 2009); our affix features; our affix features plus affix features based on the abstract phonetic class versions of words; Dou et al's features plus our affix features; Dou et al's features plus our affix features plus the abstract affix features.

When we consider only primary stress (column 2 in Table 4, for the shared-lemmata test data, Dou et al's features performed the worst at 97.2% accuracy, with all feature combinations that include the affix features performing at the same level, 98.7%. For the no-shared-lemmata test data, using Dou et al's features alone achieved an accuracy of 80.6%. The affix features alone performed worse, at 79.8%, presumably because it is harder for them to generalize to unseen cases, but using the abstract affix features increased the performance to 81.0%, better than that of using Dou et al's features alone. As can be seen combining Dou et al's features with various combinations of the affix features improved the performance further.

Features	1 stress	1+2 stress
	shared lemmata	
Dou et al	0.972	0.965
Aff	0.987	0.985
Aff+Abstr Aff	0.987	0.985
Dou et al+Aff	0.987	0.986
Dou et al+Aff+Abstr Aff	0.987	0.986
	no shared lemmata	
Dou et al	0.806	0.798
Aff	0.798	0.782
Aff+Abstr	0.810	0.790
Dou et al+Aff	0.823	0.810
Dou et al+Aff+Abstr Aff	0.839	0.815

Table 4: Word accuracies for various feature combinations for both shared lemmata and no-shared lemmata conditions. The second column reports results where we consider only primary stress, the third column results where we also predict secondary stress.

For primary *and* secondary stress prediction (column 3 in the table), the results are overall degraded for most conditions but otherwise very similar in terms of ranking of the features to what we find with primary stress alone. Note though that for the shared-lemmata condition the results with affix features are almost as good as for the primary-stress-only case, whereas there is a significant drop in performance for the Dou et al. features. For the no-shared-lemmata condition, Dou et al.'s features fare rather better compared to the affix features. On the other hand there is a substantial benefit to combining the features, as the results for "Dou et al+Aff" and "Dou et al+Aff+Abstr Aff" show. Note that in the no-shared-lemmata condition, there is only one word that is marked with a secondary stress, and that stress is actually correctly predicted by all methods. Much of the difference between the Dou et al. features and the affix condition can be accounted for by three cases involving the same root, which the affix condition misassigns secondary stress to.

For the shared-lemmata task however there were a substantial number of differences, as one might expect given the nature of the features. Comparing just the Dou et al. features and the all-features condition, systematic benefit for the all-features condition was found for secondary stress assignment for productive prefixes where secondary stress is typically found. For example, the prefix аэро ('aero-') as in а`эродина'мика ('aerodynamics') typically has secondary stress. This is usually missed by the Dou et al. features, but is uniformly correct for the all-features condition.

Since the no-shared-lemmata data set is small, we tested significance using two permutation tests. The first computed a distribution of scores for the test data where successive single test examples were removed. The second randomly permuted the test data 248 times, after each random permutation, removing the first ten examples, and computing the score. Pairwise t-tests between all conditions for the primary-stress-only and for the primary plus secondary stress predictions, were highly significant in all cases.

We also experimented with a postaccent feature to model the postaccented class of nouns described in Section 3. For each *prefix* of the word, we record whether the following vowel is stressed or unstressed. This feature yielded only very slight improvements, and we do not report these results here.

8 Discussion

In this paper we have presented a Maximum Entropy ranking-based approach to Russian stress prediction. The approach is similar in spirit to the SVM-based ranking approach presented in (Dou et al., 2009), but incorporates additional affix-based features, which are motivated by linguistic analyses of the problem. We have shown that these additional features generalize better than the Dou et al. features in cases where we have seen a related form of the test word, and that combing the additional features with the Dou et al. features always yields an improvement.

References

Kenneth Church. 1985. Stress assignment in letter to sound rules for speech synthesis. In *Association for Computational Linguistics*, pages 246–253.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31:25–69, March.

Qing Dou, Shane Bergsma, Sittichai Jiampojamarn, and Grzegorz Kondrak. 2009. A ranking approach to stress prediction for letter-to-phoneme conversion. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 118–126, Suntec, Singapore, August. Association for Computational Linguistics.

Keith B. Hall, Scott Gilpin, and Gideon Mann. 2010. Mapreduce/bigtable for distributed optimization. In *Neural Information Processing Systems Workshop on Leaning on Cores, Clusters, and Clouds*.

Morris Halle. 1997. On stress and accent in Indo-European. *Language*, 73(2):275–313.

Steve Pearson, Roland Kuhn, Steven Fincke, and Nick Kibre. 2000. Automatic methods for lexical stress assignment and syllabification. In *International Conference on Spoken Language Processing*, pages 423–426.

Terence Wade. 1992. *A Comprehensive Russian Grammar*. Blackwell, Oxford.

Gabriel Webster. 2004. Improving letter-to-pronunciation accuracy with automatic morphologically-based stress prediction. In *International Conference on Spoken Language Processing*, pages 2573–2576.

Briony Williams. 1987. Word stress assignment in a text-to-speech synthesis system for British English. *Computer Speech and Language*, 2:235–272.

Andrey Zaliznyak. 1977. *Grammaticheskij slovar' russkogo jazyka*. Russkiy Yazik, Moscow.