Spelling Correction for Russian: A Comparative Study of Datasets and Methods

Alla Rozovskaya
Department of Computer Science
Queens College, CUNY
arozovskaya@qc.cuny.edu

Abstract

We develop a minimally-supervised model for spelling correction and evaluate its performance on three datasets annotated for spelling errors in Russian. The first corpus is a dataset of Russian social media data that was recently used in a shared task on Russian spelling correction. The other two corpora contain texts produced by learners of Russian as a foreign language. Evaluating on three diverse datasets allows for a cross-corpus comparison. We compare the performance of the minimally-supervised model to two baseline models that do not use context for candidate re-ranking, as well as to a character-level statistical machine translation system with context-based re-ranking. We show that the minimally-supervised model outperforms all of the other models. We also present an analysis of the spelling errors and discuss the difficulty of the task compared to the spelling correction problem in English.

1 Introduction

The spelling correction task has been a fundamental Natural Language Processing (NLP) problem ever since the origins of the field and has enjoyed a lot of attention in the NLP research. It is not surprising, since correcting spelling mistakes is of practical relevance for various higher-level NLP tasks and downstream applications dealing with noisy data, such as named entity recognition, dependency parsing, information retrieval, topic modeling, machine translation, essay scoring, speech recognition, automatic text correction (van Strien et al., 2020). Running a spellchecker is now a common pre-processing step performed in essay scoring (Flor, 2012a), grammatical error correction (Chollampatt and Ng, 2018; Rozovskaya and Roth, 2016; Grundkiewicz and Junczys-Dowmunt, 2018) and numerous other applications. Nevertheless, even for English, performance of spellchecking tools is not as good as one would expect, especially in noisy domains (Flor et al., 2019). Kantor et al. (2019) evaluate three publicly available spellcheckers on English learner data and find that the highest recall achieved is that of 69%, and the best precision is 57%, which indicates that the task is far from being solved. Further, their simple in-house implementation outperforms by a large margin all of the common publicly available spellcheckers.

One reason for the slow progress on the task might be the lack of common benchmark datasets. As a result, proposed methods are being evaluated either on isolated spelling errors extracted from a corpus without context[1] or on artificially created datasets. Recently, Flor et al. (2019) released a dataset annotated for spelling errors in English learner essays and provided an evaluation of a minimal supervision system that combines features based on the misspelled word itself and the context in which it appears. They report strong performance on that corpus, as well as competitive results on a dataset from the medical domain.

We address the problem of correcting non-word spelling mistakes in Russian, a language with rich morphology. Our goal is two-fold: first, to implement various established spelling methods with known results for English and determine how they perform on Russian. Our second goal is to perform cross-corpus comparison, by evaluating on several Russian datasets that contain diverse data (texts by native Russian speakers in the social media domain, as well as Russian learner texts).

We implement four models. First, we use two baselines that do not take context into account: Aspell spellchecker and the model proposed in Kantor et al. (2019) that was shown to outperform Aspell and several other grammar check-

[1] *Context* in this work refers to the sentence (or the n-gram window) in which the misspelled word occurs.

Proceedings of Recent Advances in Natural Language Processing, pages 1206–1216
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_136

ers for English. We then implement two models that also take into account contextual information when proposing a correction: a statistical machine translation (SMT) approach and a minimally-supervised method. The minimally-supervised model follows the approach of Flor et al. (2019). This model is compared against a character-level SMT spellchecker that takes context into account by incorporating a word-based language model (LM) as well as other context features at the re-ranking level (Chollampatt and Ng, 2018). We evaluate these four methods on three Russian datasets that contain annotated spelling mistakes. We perform a detailed error analysis and identify the challenges pertaining to the Russian language. We show that even though spelling correction of non-word errors is considered to be an easy task, performance on a morphologically-rich language is challenging and leaves a wide gap for future research.

This paper makes the following contributions: (1) we implement and evaluate four established approaches to spelling correction and evaluate these on three Russian datasets; (2) we show that the minimally-supervised approach outperforms the other methods and is the most robust; (3) we perform error analysis identifying challenges of spelling correction for Russian.

Section 2 reviews related work on the spelling correction of Russian and on the established methods well-studied for English. Section 3 describes the three datasets of spelling errors used in this work. Section 4 presents the models. In Section 5, we present the results, and in Section 6 we perform error analysis of the results. Section 7 concludes.

2 Related Work

A *non-word misspelling* is a spelling error, such that the resulting string is not a valid word in the language. This is different from real-word (context-sensitive) errors, for example confusing "their", "there" and "they're" (Wilcox-O'Hearn et al., 2008). Context-sensitive errors also subsume grammar errors made by non-native speakers (e.g. confusing "a" and "the"), but these typically are addressed using a different set of methods (Ng et al., 2014).

Most of the spelling correction research has been focused on the English language. When dealing with a language that has rich morphology, such as Russian, specific challenges may arise. For example, the rich morphology of Russian, as we show, affects the candidate generation algorithm, where a substantially higher number of competing candidates is being generated, including those that are morphological variants of the same lemma. There is very little spelling work on other languages with complex and diverse morphology. For instance, Oflazer (1996); Mohit et al. (2014); Rozovskaya et al. (2015) address a variety of errors in Arabic, including grammar and usage errors, but they do not focus on spelling.

Previous studies on Russian spelling mainly addressed correcting spelling errors in search queries (Baytin, 2008; Panina et al., 2013), which is a special subtask of spelling correction, as the surrounding context for candidate selection is not considered or is considered in a quite restrictive way. Sorokin et al. (2016) introduced the first competition on spelling correction for Russian, which focused on correcting texts collected from Russian social media websites. Sorokin (2017) presents a follow-up study, where they show that the use of morphological information for candidate selection is beneficial for languages with well-developed morphology, such as Russian. We use the corpus released in this competition and show that it is quite different from the other two corpora used in this work.

Approaches to non-word spelling correction Broadly speaking, the approaches to correcting non-word spelling errors can be broken down into those that only consider the characteristics of the target token when ranking correction candidates, and those that also take into account contextual information. Among the former are those that compute edit distance (Levenshtein, 1966; Damerau, 1964) and phonetic similarity between the misspelling and the candidate correction (Toutanova and Moore, 2002).

One standard approach to correcting non-word spelling errors follows the noisy channel model formulation (Shannon, 1948). This approach incorporates non-contextual information, such as the edit distance and phonetic similarity between the misspelling and the candidate correction, and the candidate frequency (Kernighan et al., 1990; Church and Gale, 1991; Toutanova and Moore, 2002). Essentially, weights for different edit operations are estimated from a large training corpus of annotated spelling errors. However, this approach requires a lot of supervision: thousands of annotated errors paired with their corrections are used to estimate probabilities associated with each edit. While the noisy channel model can also incorporate contextual information, in general, adding new features from a variety of sources is not straightforward in the noisy channel formulation.

Flor et al. (2019); Flor and Futagi (2012) proposed a minimally-supervised model that combines contextual and non-contextual information. In Flor et al. (2019), they evaluate the model on two spelling corpora: an English learner corpus and a corpus from the biomedical domain, showing competitive results. Importantly, unlike the noisy channel model, their model only requires a small amount of supervision and is robust on out-of-domain data. In this work, we describe an implementation of this model for Russian.

SMT methods for Spelling Correction Character-level statistical machine translation has been widely used for spelling correction of natural data as well as OCR post-correction, which can be viewed as a subtask of spelling correction. Neural network (NN) approaches, in particular, seq2seq models have recently been used for spelling correction. We do not evaluate NN methods in this work, as we have very limited amounts of training data. For an analysis and evaluation of NN approaches for spelling correction, we refer the reader to Schnober et al. (2016) and Amrhein and Clematide (2018).

3 Datasets

We use three Russian datasets annotated for misspellings. The first one, RULEC-GEC, is a learner corpus collected at the University of Oregon and consists of essays written by learners of Russian as a foreign language and heritage speakers (Alsufieva et al., 2012; Rozovskaya and Roth, 2019). The dataset was corrected and annotated by native Russian speakers and is error-coded. It is annotated exhaustively for various grammar and usage errors, and contains a large proportion of spelling errors, especially for heritage speakers (over 42% of all errors), and over 18% of all errors in the foreign group. We only focus on mistakes that are marked as spelling errors. The corpus is partitioned into training, development, and test. Since we focus on the spelling errors, we evaluate only with respect to those mistakes and ignore other annotated errors in the data.

The second corpus, henceforth RU-Lang8 (Trinh and Rozovskaya, 2021), is a dataset collected from the online language learning platform Lang-8 (Mizumoto et al., 2011) and annotated by native speakers. The dataset contains texts by learners of a variety of foreign languages. The annotation is publicly available for research. RU-Lang8 contains 54,000 tokens split up into development and test partitions. We only use the test partition in this work for evaluation, as the models are developed and tuned on the RULEC-GEC data. RU-Lang8 differs from RULEC-GEC: the latter consists of essays written on a University setting in a controlled environment, while the Lang-8 data was collected online; the majority of texts are short paragraphs or questions posed by language learners. RU-Lang8 is thus more informal and contains data by learners of multiple first language backgrounds (unlike RULEC-GEC, whose authors are from the United States).

The third corpus, RUSpellRU, is a dataset released as part of the competition on automatic spelling correction for the Russian language, which focused on social media texts. The dataset is a collection of essays from Russian blogs and social media. This is another unique dataset, very distinct: it contains a lot of colloquialisms, slang expressions and social media spelling conventions (Sorokin et al., 2016). Since the corpus contains social media texts, the misspellings include, in addition to typos, a lot of slang and colloquial forms common in social media spelling, such as the use of digits inside the words or unconventional spellings, e.g. using phonetic spelling instead of standard one.

Statistics on the datasets, including the total number of tokens as well as the spelling error rates (percentage of tokens containing a spelling error), are shown in Table 1. We observe that the RUSpellRU dataset is the most noisy one, and its error rate is more than five times higher than in the RULEC-GEC corpus, where the percentage of tokens containing a spelling mistake is the smallest among the three. On the other hand, the RUSpellRU dataset is produced by native Russian speakers, while the other two are produced by learners of Russian and thus also contain other, grammar and usage-related errors.

Table 2 analyzes the spelling errors with respect to the type of edit – replacement, split, or merge. A *merge* is a misspelling where a space is incor-

Dataset	Token counts	Spelling errors	Error rate
RULEC-GEC (train)	83,410	1,023	1.23
RULEC-GEC (dev)	41,163	497	1.21
RULEC-GEC (test)	81,693	1,055	1.30
RU-Lang8	31,603	692	2.19
RUSpellRU	28,112	1,963	6.98

Table 1: Corpora statistics.

Corpus	Edit type		
	Repl. (%)	Merge (%)	Split (%)
RULEC-GEC	92.4	1.6	6.0
RU-Lang8	95.7	1.5	2.9
RUSpellRU	80.2	11.7	8.1

Table 2: Distribution of annotated misspellings by type (merges, splits, replacements) in the three datasets. A *merge* is a misspelling where a space is incorrectly omitted, while a *split* is a misspelling that results from an extra space being added.

rectly omitted, while a *split* is a misspelling that results from an extra space being added. The differences between the datasets are quite significant. The RU-Lang8 corpus contains the highest proportion of replacement errors (95.7%), while the social media corpus RUSpellRU contains the least proportion of replacement errors - 80.2%, while merge errors are about 5 times more common in RUSpellRU than in the other two corpora.

Finally, in Table 3 we analyze the replacement errors with respect to the edit distance between the source word and the correction. In the RULEC-GEC and RUSpellRU datasets, over 80% of replacement edits are within edit distance 1, where each type of change, including transposition errors, has a cost of 1. This analysis is consistent with findings in English corpora of misspellings (Flor et al., 2019). The RU-Lang8 corpus, however, contains a higher proportion of errors with edit distance greater than 1. Only 68.4% of errors are within edit distance of 1.

4 The Models

In this section, we describe the minimally-supervised model, the character-level SMT speller, and the two baselines that do not use context.

Corpus	Edit dist.	Perc. (%)
RULEC-GEC	1	84.0
	2	11.3
	3	2.6
	> 3	2.1
RU-Lang8	1	68.4
	2	19.5
	3	7.3
	> 3	4.8
RUSpellRU	1	83.6
	2	10.9
	3	3.6
	> 3	1.9

Table 3: Distribution of annotated misspellings (replacement errors) by edit distance to correct form, in the RULEC-GEC and RU-Lang8 datasets.

Dataset	Gold errors	Recall (%)
RULEC-GEC	1055	65.7
RU-Lang8	692	79.9
RUSpellRU	1963	71.3

Table 4: Error detection performance on the three test sets.

4.1 Minimally-Supervised Spelling Correction Model

We implement the model described in Flor and Futagi (2012), Flor (2012a), that is evaluated in the original papers on the English learner corpus of TOEFL and GRE essays. It was also evaluated on the TOEFL-11 corpus as well as a corpus of biomedical English texts (Flor et al., 2019). Implementation of the model for Russian and its evaluation on the Russian data with its rich morphology is one of the contributions of the current work.

In this approach, the spelling correction task is broken down into three subtasks: (1) detection, (2) candidate generation, and (3) ranking of the candidates. We describe each step below. In our implementation, we only consider single-token spelling errors, where the original and the correction are both single tokens.

Error Detection[2] Detection of non-word spelling errors is performed using a dictionary (lexicon). Tokens that are not in the lexicon are considered to be misspelled. This is not a trivial step, as proper

[2] The detection step described here is the same for all spelling correction approaches used in this work, to make the comparisons among the algorithms fair.

Dataset	Classification of errors missed at detection step (%)					
	Proper name	Other	Context-sens.	Cap.	Grammar	Multi-token
RULEC	5.9	16.9	22.8	15.4	18.4	20.6
RU-Lang8	15.1	18.7	40.3	0.01	11.5	13.7
RUSpellRU	4.8	3.7	17.3	4.4	9.2	36.8

Table 5: Analysis of spelling mistakes missed at the detection stage. The mistakes that should have been detected are those in categories *Other* and *Proper name*. *Context-sens.* stands for context-sensitive errors, where the original token is also a valid word in the language. *Cap.* stands for capitalization errors.

names, in particular those that are foreign names, or rare words, may be missing and would be mistakenly flagged as potential misspellings. Nevertheless, recall (detecting potential misspellings) is more important than maintaining high precision in this step. Our dictionary is based on the Yandex corpus (Borisov and Galinskaya, 2014). The corpus size is over 18 million tokens, and the resulting dictionary contains 2.3 million word types. To reduce the number of false positives, for the words not in the dictionary, we also check whether the token is recognized as the last or first name by the Mystem morphological analyzer (Segalovich, 2003) (if it appears non-capitalized in non-initial sentence position) or if the stem of the word is recognized as a known stem. The recall of the detection algorithm is shown in Table 4. The lowest recall of 65.7 is achieved on the RULEC-GEC dataset, while the highest recall is obtained on RU-Lang8 (79.9%).

We further analyze the recall of the detection algorithm by classifying the spelling mistakes in the gold data that were missed (Table 5). In the RULEC-GEC dataset, 22.8% of these errors are context-sensitive spelling mistakes, i.e. spelling errors that involve confusing valid words and which are not covered in this task. 20.6% are spelling errors that are multi-token (i.e. require merging two or more tokens), while 18.4% are context-sensitive grammar mistakes (e.g. noun case) which were miscategorized by the annotator. Another 15.4% of mistakes are capitalization errors. Only 16.9% of the missed errors (category *Other*) as well as 5.9% of errors that involve spelling mistakes on proper names are in fact spelling mistakes that should have been detected at this stage. Similarly, on RU-Lang8 dataset, 40.3% of missed errors are context-sensitive errors, and the actual mistakes that were missed (categories *Other* and *Proper Name*) include 33.8% of all missed tokens. In the RUSpellRU corpus, these errors comprise 8.5%.

Dataset	Dist.	Cand. per error	Gold in cand.(%)
RULEC-GEC	1	3.1	79.4
	2	44.3	92.9
	3	313.0	95.5
RU-Lang8	1	4.6	66.0
	2	76.0	85.0
	3	412.6	91.5
RUSpellRU	1	3.7	68.1
	2	70.6	77.6
	3	361.9	78.8

Table 6: Evaluation of the candidate generation step.

If we exclude the non-relevant errors that are counted as missed, the recall of the detection stage improves to 89.2% for the RULEC-GEC corpus, 92.2% for the RU-Lang8 corpus, and 96.7% on the RUSpellRU dataset.

Candidate Generation We consider several approaches to candidate generation based on the edit distance between the source and the target strings.

Candidates are generated using the dictionary described in the previous section. Candidates include all dictionary words within edit distance that does not exceed half the length of the misspelled string; the maximum distance is set to three, as the number of candidates grows very quickly due to the rich morphology of Russian (see Table 6). For example, on average, 3 candidates are generated with edit distance of 1 for RULEC-GEC. This number increases to 313 when an edit distance of 3 is used instead. This is because, due to the morphological complexity, morphological variants of the same base word are included as different candidates (also discussed in Section 6). In English, candidates up to edit distance of 6 are included (Flor et al., 2019), but doing so would explode the search space.

The candidate generation algorithm is evaluated

Feature name	Description
Non-contextual features	
Orthographic similarity	Inverse edit distance
Character difference	A pair comprising original and replacement character
Candidate frequency	Unigram word frequency
Contextual features	
N-gram support	N-gram counts in the 4-word window

Table 7: Description of all the features used in candidate ranking with the minimally-supervised model.

in Table 6. As the edit distance increases, the recall of the candidate generation (i.e. proportion of errors for which gold is among the generated candidates) improves, however, the number of candidates per error increases exponentially. We note, though, that, while on the RULEC-GEC and RU-Lang8 datasets, the recall increases to over 90% with the edit distance set to 3, on the RUSpellRU corpus, the highest recall achieved is 78.8%, even though 83.6% of misspellings are within edit distance of 1, as shown in Table 3. This indicates that a large number of colloquial and slang words present in the corpus are not found in the dictionary.

Ranking of Candidate Corrections The ranking step is the most challenging one and is the focus of most work on non-word spelling correction (Fivez et al., 2017). Ranking of correction candidates in the minimally-supervised model uses both the features of the misspelling-candidate pair and the contextual information. Flor (2012b) tuned feature weights manually on a set of misspellings, extracted from a corpus of TOEFL and GRE essays. In this work, similar to (Flor et al., 2019), feature weights are learned using a linear algorithm (Averaged Perceptron (Rosenblatt, 1958), implemented within Learning Based Java (Rizzolo and Roth, 2007).

We implement the following features: orthographic similarity (inverse edit distance), character-difference, candidate word frequency, and n-gram support. The features are listed in Table 7 and described below.

Orthographic similarity is computed as inverse edit distance, $1/(eDist + 1)$, where $eDist$ is the edit distance (including transpositions) between the misspelling and the correction candidate (Levenshtein, 1966; Damerau, 1964).

Character difference is a feature that encodes the specific letter change between the original and the candidate. This feature is active with replacement, deletion, and character insertion errors for candidates whose edit distance is 1. The feature is expected to reflect some common and well-known character confusions, both among native and non-native Russian writers, e.g. omitting the ь at the end of a word after character ш or incorrectly using a instead of o in an unstressed position. Note that this feature is similar to the concept of encoding phonetic similarity, which we omit in this implementation.

Candidate frequency A more frequent word is more likely to be the intended word than a rare word (Flor, 2012a). Unigram word frequency is computed for each candidate using the Yandex corpus.

N-gram support For each correction candidate, all n-grams in the window of four context words on each side are taken into account by the n-gram support feature. We use co-occurrence counts computed from a large corpus collected over the Web (235 million tokens), henceforth the Sharoff corpus.[3] The n-gram support feature is a summation over the counts of all n-grams of length 2 to 5 (excluding the unigram count of the candidate itself, since its frequency is reflected in the candidate frequency feature). For each error, the n-gram count value is normalized by the highest candidate count for that error.

For each misspelled token, with the exception of the letter difference feature, the feature scores of its candidate corrections are *normalized*, by dividing the score of the candidate feature by the highest-scoring candidate on that given feature.

4.2 The SMT Speller

We implement a character-level statistical machine translation (SMT) speller (Chollampatt and Ng, 2017). Input to the character-level SMT component is a sequence of characters that make up the unknown (misspelled) word and output is a list of correction candidates (words). In Chollampatt and Ng (2017), the misspelled words are those words that have not been observed in the source side of the parallel training data used to train the translation model. In this work, the unknown words are identified using the same detection algorithm described in Section 4.1. We do this for two reasons: first, due to lack of large amounts of parallel data

[3] The corpus was kindly shared by Serge Sharoff.

and the morphological complexity of Russian, the number of unknown words for a word-level SMT system would be too high. Second, we wish the keep the detection step fixed, which allows for a fair comparison of the re-ranking algorithms.

The character-level translation model, in line with Chollampatt and Ng (2017), is trained on pairs of misspellings and their corrections from the RULEC-GEC training corpus (774 pairs) and an additional set of 1,000 correct words selected uniformly at random from the target side of RULEC-GEC training data.[4] The language model that is part of the SMT system is a 5-gram character-level model trained on the Yandex corpus (22 million tokens). The SMT model is tuned on the misspelling-correction pairs from the RULEC-GEC development set. The character-level SMT model is tuned using MERT (minimum error-rat training) on characters, with character-level edit operation features and a 5-gram character LM.

For each unknown word, the character-level SMT produces 100 candidates that are then rescored to select the best candidate based on the context. The rescoring is done following Chollampatt and Ng (2017) and uses word-level n-gram LM features: LM probability and the LM OOV (out-of-vocabulary) count denoting the number of words in the sentence that are not in the LM's vocabulary. The word-level n-gram LM is trained on the Sharoff corpus, using the KenLM toolkit (Heafield et al., 2013).

4.3 Further Baseline Systems

We compare to two other methods that do not make use of context information: Aspell, and a re-ranking algorithm proposed in Kantor et al. (2019) . The latter has been recently used in grammar and spelling correction research and showed good results in English.

Kantor et al. re-ranking Kantor et al. (2019) implement an approach to English spelling correction, that is quite simple but is surprisingly effective and outperforms substantially other commonly used open-source spellcheckers: Enchant, Norvig, and Jamspell. Briefly, the approach relies on a large dictionary compiled from a native corpus to identify misspelled tokens. In re-ranking, for each misspelling, the most frequent candidate correction (with a minimum count of 20) within an edit distance of 1 (transposition is treated as a distance of 1) is returned. If no such candidate exists, they check if the misspelled word can be split into two words that are in the word-count data or in the dictionary. We implement their re-ranking method (keeping the minimum count for words at 5, since Russian is a morphologically-rich language). Our list of incorrect tokens is generated using the same candidate detection step described above (Section 4.1). Only the re-ranking is different.

Dataset	**P**	**R**	F_1
RULEC-GEC	71.6	46.5	64.6
RU-Lang8	54.0	42.7	51.3
RUSpellRU	74.5	59.0	65.9

Table 8: Key results of the minimally-supervised model. Performance on the error correction, using the full set of features. The model is trained on the RULEC-train corpus. Since edit distance 1 is used, this feature is omitted.

5 Results

In all cases, the models are trained on the RULEC-train corpus and tuned on the RULEC development data. All results are reported on the test partitions of the three datasets. Key results of the minimally-supervised model on the three datasets are shown in Table 8. We observe that the performance on RU-Lang8 is significantly lower than on the other two datasets. We conjecture that this may be due to the fact that RU-Lang8 has a small proportion of errors with corrections being within an edit distance of 1 from the misspelled token (see Table 3).

The results of the minimally-supervised model and of the other models implemented in this work are shown in Table 9. The minimally-supervised model outperforms all of the other models significantly on all three datasets. The relative performance of the models on each dataset is consistent: we note that Aspell has the poorest performance. This is followed by the SMT approach and the approach by Kantor et al. (2019). The two approaches are quite close, although the SMT method has high precision on RULEC-GEC and the RUSpellRU datasets. The minimally-supervised method achieves a substantially higher recall than all the other methods. It also achieves the highest precision on RULEC-GEC and RUSpellRU, although on the RU-Lang8 corpus its pre-

[4] We also tried adding all words from the target side (without repetitions), but this did not improve the performance.

Dataset	System	P	R	F_1
RULEC-GEC	Aspell	42.2	41.2	42.0
	Kantor et al. (2019)	65.5	43.6	59.6
	SMT	70.1	34.4	58.1
	Minim.-super.	71.5	46.5	**64.6**
RU-Lang8	Aspell	33.8	7.3	19.6
	Kantor et al. (2019)	50.1	41.9	48.2
	SMT	49.9	28.5	43.4
	Minim.-super.	42.7	54.0	**51.3**
RUSpellRU	Aspell	34.2	35.8	35.0
	Kantor et al. (2019)	59.5	48.4	53.4
	SMT	66.7	20.1	44.9
	Minim.-super.	74.5	59.0	**65.9**

Table 9: Comparison of the minimally-supervised model with other systems implemented in this work.

Dataset	Edit dist.	P	R	F_1
RULEC-GEC	1	71.6	46.5	64.6
	2	67.5	51.0	63.4
	3	62.1	50.5	59.4
RU-Lang8	1	42.7	54.0	51.3
	2	46.5	51.8	50.6
	3	49.1	53.6	52.7
RUSpellRU	1	74.5	59.0	65.9
	2	67.5	59.9	63.4
	3	65.4	59.5	62.3

Table 10: Evaluation of different edit distances in candidate generation. Performance on error correction of the minimally-supervised model, using the full set of features. The model is trained on RULEC-train.

Dataset	Counts	P	R	F-s
RULEC-GEC	all feats	71.5	46.5	64.6
	no cand freq	71.2	46.1	64.2
	no char. diff	71.2	46.2	64.2
	no n-gram	67.5	43.9	61.0
RU-Lang8	all feats	42.7	54.0	51.3
	no cand freq	43.5	55.1	52.3
	no char. diff	43.2	54.7	52.0
	no n-gram	40.2	51.1	48.5
RUSpellRU	all feats	74.5	59.0	65.9
	no cand freq	74.0	58.7	65.4
	no char. diff	74.4	58.9	65.7
	no n-gram	72.1	57.1	63.8

Table 11: Feature ablation. Performance on the error correction, using an edit distance of 1. The model is trained on the RULEC-train corpus.

cision is lower than the SMT and the re-ranking approaches. Overall, among the three datasets, the performance on RU-Lang8 is the lowest.

Evaluation of different edit distance values in candidate generation Next, we evaluate performance as a function of edit distance values for candidate generation. In all cases, we use the full feature set. Results are shown in Table 10. We observe that there is no clear benefit to using an edit distance greater than 1: on the RULEC-GEC corpus, recall slightly improves, while precision drops. On the RU-Lang8 dataset, precision improves with a larger edit distance, while recall remains the same. On the RUSpellRU dataset, recall does not change, while precision drops. For this reason, we use an edit distance of 1 in candidate generation, as the number of candidates is much smaller, as discussed above. It should be noted that, while using only edit distance 1 is optimal for the scoring with the system, it is not sufficient, as there are still many misspellings in each of the datasets (as shown in Table 3) that have corrections within higher edit distances. We leave this for future work.

Feature ablation Finally, we perform feature ablation to evaluate the contribution of the various features in the minimally-supervised model. Results are shown in Table 11. The n-gram support feature is shown to be the most important: dropping this features results in performance loss of 2-3 points on each dataset. This result demonstrates the significance of the contextual information in spelling correction.

6 Error Analysis

Candidate re-ranking We perform error analysis of the candidate re-ranking component that is part of the minimally-supervised approach. From each dataset, we analyze 100 errors, on which an incorrect candidate is preferred. Results are shown in Table 12. *Morph. variant* refers to incorrect candidates that is an inflectional morphological variant of the correct suggestion. *Wrong cand.* denotes an incorrect suggestion that is not morphologically related to the correct suggestion. *Dist.* denotes corrections that have an edit distance greater than 1 to the source word. Since we currently only consider candidates within edit distance of 1, these errors cannot be corrected. *Lex. change* refers to misspellings that are also word usage errors. It is interesting to note that all three corpora have the same proportion of errors (50%) that could not be

Dataset	Mistakes by type (%)			
	Morph. variant	*Wrong cand.*	*Edit dist.*	*Lex. change*
RULEC-GEC	23	13	50	14
RU-Lang8	17	13	50	20
RU-SpellRU	20	31	49	0

Table 12: Distribution of incorrect suggestions by the candidate re-ranking algorithm. *Morph. variant* refers to an incorrect candidate that is a morphological variant of the correct suggestion. *Wrong cand.* denotes an incorrect suggestion that is not morphologically related to the correct suggestion. *Edit dist.* denotes corrections that have an edit distance greater than 1 to the source word. *Lex. change* are errors that are word usage errors, in addition to having a spelling error.

corrected due to the edit distance of the correct candidate being greater than 1. Further, the corpora have similar distributions overall regarding mistakes when selecting a morphological variant (17-23%). This shows that for languages with rich morphology, morphological variants present an issue for spelling correction, since morphological variants typically differ by one character change, and thus correction candidates typically include multiple morphological variants of the same word. A similar conclusion, although not quantified, was drawn in the RUSpellRU competition for the social media data (Sorokin et al., 2016). We confirm this finding for various corpora and quantify it. Both of the learner corpora also have grammatical errors, and some of these were mistagged as spelling mistakes (14% and 20% in the RULEC and RU-Lang8 corpora, respectively). In contrast, because the RUSpellRU corpus contains data from native speakers, it is not expected to have many grammar-related errors.

7 Conclusion

In this paper, we implement four models for spelling correction for Russian and evaluate these on three diverse datasets that contain spelling mistakes. We present a comparative analysis of spelling mistakes contained in the three datasets. Evaluation results show that the minimally-supervised model outperforms two baseline models that do not use context when selecting a candidate correction, and another model that uses a character-level SMT and a language model in re-ranking. We perform feature ablation of the minimally-supervised model showing that contextual information contributes to the performance. We also carry out error analysis that reveals that one common source of errors in Russian in selecting the appropriate correction candidate is the presence of morphological variants. This study should provide insight into the spelling correction problem for languages with rich morphology.

Acknowledgments

We thank the three anonymous reviewers for their insightful comments.

References

A. Alsufieva, O. Kisselev, and S. Freels. 2012. Results 2012: Using flagship data to develop a russian learner corpus of academic writing. *Russian Language Journal*, 62:79–105.

C. Amrhein and S. Clematide. 2018. Supervised OCR error detection and correction using statistical and neural machine translation methods. *Journal for Language Technology and Computational Linguistics (JLCL)*, 33(1):49–76.

A. Baytin. 2008. Search query correction in yandex. In *Russian Internet technologies*.

A. Borisov and I. Galinskaya. 2014. Yandex school of data analysis russian-english machine translation system for WMT14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

S. Chollampatt and H. T. Ng. 2017. Connecting the dots: Towards human-level grammatical error correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI*. Association for the Advancement of Artificial Intelligence.

Kenneth W. Church and William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing*, 1:93—103.

Frederick Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):659—664.

Pieter Fivez, Simon Sŭster, and Walter Daelemans. 2017. Unsupervised context-sensitive spelling correction of english and dutch clinical free-text with word and character n-gram embeddings.

M. Flor. 2012a. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL). (Special Issue: Managing noise in the signal: error handling in natural language processing)*, 3(53):61–99.

M. Flor and Y. Futagi. 2012. On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Michael Flor. 2012b. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL)*, 53(3):61—99.

Michael Flor, Michael Fried, and Alla Rozovskaya. 2019. A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

R. Grundkiewicz and M. Junczys-Dowmunt. 2018. Near human-level performance in grammatical error correction with hybrid machine translation. In *NAACL*.

K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *ACL*.

Y. Kantor, Y. Katz, L. Choshen, E. Cohen-Karlik, N. Liberman, A. Toledo, A. Menczel, and N. Slonim. 2019. Learning to combine grammatical error corrections. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Papers presented to the 13th International Conference on Computational Linguistics (COLING 1990)*, volume 2, pages 205–210.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707—710.

T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. 2011. Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *IJCNLP*.

B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP)*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

K. Oflazer. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89.

M. Panina, A. Baitin, and I. Galinskaya. 2013. Context-independent autocorrection of query spelling errors. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013"*.

N. Rizzolo and D. Roth. 2007. Modeling Discriminative Global Inference. In *Proceedings of the First International Conference on Semantic Computing (ICSC)*, pages 597–604, Irvine, California. IEEE.

F. Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.* (Reprinted in *Neurocomputing* (MIT Press, 1988).).

A. Rozovskaya, H. Bouamor, W. Zaghouani, O. Obeid, N. Habash, and B. Mohit. 2015. The second QALB shared task on automatic text correction for arabic. In *Proceedings of the ACL Workshop on Arabic Natural Language Processing*.

A. Rozovskaya and D. Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *ACL*.

A. Rozovskaya and D. Roth. 2019. Grammar error correction in morphologically-rich languages: The case of russian. In *Transactions of ACL*.

C. Schnober, S. Eger, E.-L. Ding, and I. Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *COLING*.

I. Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Claude Shannon. 1948. A mathematical theory of communications. *Bell Systems Technical Journal*, 27:623–656.

A. Sorokin. 2017. Spelling correction for morphologically rich language: a case study of russian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*.

A. Sorokin, A. Baytin, I. Galinskaya, E. Rykunova, and T. Shavrina. 2016. SpellRuEval: the first competition on automatic spelling correction for russian. In *Proceedings of the International Conference "Dialogue 2016"*.

D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. Association for Computational Linguistics.

Kristina Toutanova and Robert Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Viet Anh Trinh and Alla Rozovskaya. 2021. New dataset and strong baselines for the grammatical error correction of russian. In *Proceedings of ACL Findings*. Association for Computational Linguistics.

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky. 2008. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In *Proceedings of CICLing-2008*, pages 605–616.