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

Moscow, June 1–4, 2016

SPELLRUEVAL: THE FIRST COMPETITION ON AUTOMATIC SPELLING CORRECTION FOR RUSSIAN

Sorokin A. A. (alexey.sorokin@list.ru)[1,3,4],
Baytin A. V. (baytin@yandex-team.ru)[2],
Galinskaya I. E. (galinskaya@yandex-team.ru)[2],
Rykunova E. D. (alenarykunova@gmail.com)[3],
Shavrina T. O. (rybolos@gmail.com)[1,4]

[1]Lomonosov Moscow State University, Moscow, Russia

[2]Yandex, Moscow, Russia

[3]Moscow Institute of Physics and Technology, Dolgoprudny, Russia

[4]General Internet Corpora of Russian, Moscow, Russia

This paper reports on the first competition on automatic spelling correction for Russian language—SpellRuEval—held within the framework of "Dialogue Evaluation". The competition aims to bring together groups of Russian academic researchers and IT-companies in order to gain and exchange the experience in automatic spelling correction, especially concentrating on social media texts. The data for the competition was taken from Russian segment of Live Journal.

7 teams took part in the competition, the best results were achieved by the model using edit distance and phonetic similarity for candidate search and n-gram language model for their reranking. We discuss in details the algorithms used by the teams, as well as the methodology of evaluation for automatic spelling correction.

Key words: spelling correction, automatic spelling correction, language of social media, automatic methods for processing Russian

SPELLRUEVAL: THE FIRST COMPETITION ON AUTOMATIC SPELLING CORRECTION FOR RUSSIAN

Сорокин А. А. (alexey.sorokin@list.ru)[1,3,4],
Байтин А. В. (baytin@yandex-team.ru)[2],
Галинская И. Е. (galinskaya@yandex-team.ru)[2],
Рыкунова Е. Д. (alenarykunova@gmail.com)[3],
Шаврина Т. О. (rybolos@gmail.com)[1,4]

[1]МГУ им. М. В. Ломоносова, Москва, Россия;

[2]Яндекс, Москва, Россия;

[3]МФТИ, Долгопрудный, Россия;

[4]ГИКРЯ, Москва, Россия

В этой статье обсуждается первое соревнование по автоматическому исправлению опечаток на материале русского языка, SpellRuEval, прошедшее в рамках проекта “Dialogue Evaluation”. Целью соревнования является сравнение разнообразных методов и подходов, применяемых для и исправления опечаток, а также обмен опытом между научными коллективами и IT-компаниями, имеющими свои успешные разработки в этой области. Соревнование проводилось на материале блогов Живого Журнала.

В данной статье подробно разбираются результаты, полученные от 7 коллективов, участвовавших в соревновании, сравниваются подходы, применённые участниками соревнования. Наилучшие результаты были достигнуты моделью, использовавшей редакционное расстояние для поиска кандидатов и комбинацию взвешенного редакционного расстояния и n-граммной языковой модели для отбора наилучшего исправления. Также в статье подробно обсуждается методика оценки качества автоматического исправления опечаток.

Ключевые слова: исправление орфографии, автоматическое исправление орфографии, язык социальных медиа, исправление опечаток

1. Introduction

SpellRuEval is the first competition aimed to make a framework for evaluation of automatic spelling correction systems for Russian and cooperation and experience exchange of scientific groups. Today, when huge amounts of data are collected from Russian internet resources (e.g. Yandex Blogs, RuTenTen Corpora, Russian Araneum Corpora and GICR), automatic processing of this data is an unavoidable problem [Manning 2011]—misspells widely hinder morphological, syntactic and semantic parsing of the texts. By the estimation of [Baytin, 2008], 15% of all the queries in Yandex have

at least 1 error, and by the data of [Shavrina, Sorokin, 2015] nearly 8% of the out-of-vocabulary words (further "OOV") are typos. Moreover, for some data sources the percentage of typos may reach 40% (Private communication, GICR).

Hence, there emerges a bulk of actual challenges for NLP-researches: which error detection model for Russian internet text is the best—dictionary look-up or rule-based? Which models are the best for isolated error-correction and which are better for context errors? How to raise the quality of real-word error detection and correction? Is there any dependency between dictionary size and recall of spelling detection? Which algorithms of machine learning give the best results for spelling correction on social media texts? All these problems we have faced during the preparation of the competition procedure and the analysis of the results.

1.1. A brief history of automatic spelling correction

Automatic spelling correction is one of the oldest problems of computational linguistics. The first theoretical works appeared already in the 60-s [Damerau, 1964]. The initial approach used edit (Levenshtein) distance [Levenshtein, 1965] to search for potential corrections of mistyped words. With the appearance of modern spell-checkers [McIlroy. 1982] in the early 80-s, the problem of spelling correction became a highly practical one. The most important papers appeared on the dawn of modern NLP era include [Kernighan et al., 1988], [Mays et al., 1991] and [Kukich, 1992], which is in excellent review of early approaches in automatic spelling correction. Further work in spelling correction was developed in two main directions: the works of the first category mainly addressed the problem of effective candidate search, which is a non-trivial problem for the languages with well-developed morphology [Oflazer, 1996], [Schulz, Mihov, 2002]. This branch also includes the research on learning adequate distance measure between the typo and the correction [Ristad, Yanilos, 1998], [Kernighan et al., 1990], [Brill, Moore, 2000], [Toutanova et al., 2002]. Other researchers mainly addressed the problem of using context when selecting the correct candidate for spelling correction. The most important works here include [Golding, Schabes, 1996], [Golding, Roth, 1999], [Hirst, Budanitsky, 2005], [Cucerzan, Brill, 2004].

The problem of automatic spelling correction includes several important subtasks. The first is to detect whether a word has correct spelling and provide a list of candidates. As observed by many researchers, most of the time the correction can be obtained from the mistyped word by single letter deletion, insertion or substitution or by permutation of two adjacent characters [Kukich, 1992]. However, in many cases this procedure yields multiple candidate words and additional features should be taken into account to select the most proper one. This is especially a problem for agglutinative languages or languages with a high number of inflected forms since a single edit operation on a word often creates another form of the same word and morphology and syntax should be used to disambiguate between them. The so-called real-word errors (when a mistyped word is again in the dictionary) constitute the most difficult problem. Several researchers addressed it [Liu, Curran, 2006], [Carlson, Fette, 2007], [Pedler, Mitton, 2010], however, all the algorithms were tested on pre-defined confusion sets,

such as ‘adopt/adapt’ and ‘piece/peace’, which makes rather problematic the application of their methods to real-word errors outside these sets.

Evaluation of spellchecking techniques presents another difficult challenge. Indeed, spelling correction is applied in different areas, mainly for Internet search and information retrieval [Ahmad, Kondrak, 2005], [Cucerzan, 2004], [Zhang, 2006], [Whitelaw, 2009] and in text editors, but also in second language acquisition [Flor, 2012] and grammar error correction [Rozovskaya, 2013]. The area obviously affects the character of typical spelling errors. Moreover, the effect of different features for spelling correction also highly depends from the application. Morphology and especially syntax give little advantage in case of search query correction, in this case the quality of the dictionary and gazetteer, as well as size of query collection used to train language and error models, is more important. In case of grammar error correction the situation is roughly the opposite. Most of spelling correction systems were tested on rather artificial or restricted datasets: the authors either asked the annotators to reprint the text without using ‘backspace’ and ‘delete’ keys [Whitelaw, 2009] or used Wikipedia [Schaback, 2007] or TOEFL essays collection [Flor, 2012]. Often the authors just randomly replaced a word by a potential misspelling, using some error model ([Carlson, Fette, 2007] etc.) Therefore it is not obvious that results obtained in one subarea could be successfully used in the other one.

2. Related Work

Most of spellchecking approaches were tested on English language, which is certainly not the most difficult for this task. First, a large collection of corpora is available for English and additional data could be easily collected from the Web. Second, English is very simple from the morphological point of view, therefore most of the problems concerning morphology or dictionary lookup even does not arise there. There are very few works for other languages with complex and diverse morphology, such as Turkish or Arabic ([Oflazer, 1996], [Mohit et al., 2014], [Rozovskaya et al., 2015]). The studies for Russian language include only [Baytin, 2008], [Panina et al., 2013] and [Sorokin and Shavrina, 2015], but all these works also address spelling correction problem in a rather restricted way.

2.1. First automatic spelling correction contests

In the field of automatic spell-checking for English two works can be considered as pioneer. These are Helping Our Own (HOO) Shared Tasks of 2011 and 2012 correspondingly [Dale et al., 2011] and [Dale et al., 2012]. Although the theme of the competition was set broader than just spelling correction (the main goal was to map and develop tools that can assist authors in the writing task and facilitate the processing of the typed texts), these competitions obviously exhibited the main problems of state-of-art methods and led to more specified workshops, such as Microsoft Spelling Alteration Workshop [Wang, 2011]. It was primarily concerned with correcting errors

in search queries: participant systems were evaluated on the logs of Bing search engine. The close problem of grammatical error correction was the thematics of CoNLL 2013 Shared Task [Ng et al., 2013]. However, all these competitions were held for English Language. There were no such competition for Russian and even a freely available dataset of spelling errors, such as Birkbeck corpus for English [Mitton, 1986] did not exist. The primary purpose of SpellRuEval-2016 was to fill this gap and evaluate different approaches to automatic spelling error correction for such morphologically and syntactically complex language as Russian.

2.2. First and second QALB Shared Task on Automatic Text Correction for Arabic

The first competition to succeed on automatic text normalization and spelling correction, which was carried out on not English-based materials, was the first QALB Shared Task on Automatic Text Correction for Arabic [Mohit et al., 2014]. The competition united more than 18 systems and determined a baseline of 60–70% (Precision), which is quite a progress for such languages as Arabic. By this time, there was already held the second QALB Shared Task [Rozovskaya et al., 2015] with the improvement of the baseline up to 80% of Precision. Both of the competitions were based on the Qatar Arabic Language Bank, however, they focused on slightly different goals: if the first QALB shared task was to correction of misspells, punctuation errors, extra spaces and normalization of the dialecticisms on the corpus of native speakers, the second one have added the corpora of L2-speakers in the training set, that shifted the researchers' attention to frequent mistakes made by learners of Arabic.

3. Procedure of SpellRuEval competition

3.1. Training and test data

In this section we describe the format of training and text data used in the competition. We used a Live Journal subcorpus of General Internet Corpora of Russia (GICR) [Belikov et al., 2013] to extract test sentences. We automatically selected about 10,000 sentences containing words not present in the dictionary. The sample was enriched by several hundred sentences containing real-word errors; these sentences were obtained from the same source corpus. Then we manually filtered these sentences to ensure that these sentences indeed contain typos, not rare proper names, slang or neologisms. About 5,000 remaining sentences were loaded to the annotation system. We asked the annotators to correct the typos in each sentence following a short instruction and submit the corrected sentence. If the annotator met a controversial case, supposed to be not covered by the instruction, he or she could also submit the commentary, explaining the difficulty.

The instruction contained the following items:

1. The annotator should correct:
   a) typos (*мнея → меня*),
   b) orthographic errors (*митель → метель*),
   c) cognitive errors (*компания → кампания*),
   d) intentional incorrect writing (*хоцца → хочется, ваще → вообще*),
   e) grammatical errors (agreement etc.) (*он видят → он видит*),
   f) errors in hyphen and space positioning (*както→как-то*),
   g) mixed usage of digits and letters in numerals (*2-ух →двух*),
   h) usage of digits instead of letters (*в4ера → вчера*).
2. The annotator should not correct
   a) foreign words including cyrillic (e.g Ukrainian or Belorussian),
   b) informal abbreviations (*прога → программа*)
   c) punctuation errors (all punctuation is omitted during the testing procedure—for more details, see chapter 3.2)
   d) capitalization errors (as capitalization is rather varied and informal in Russian social media texts, see also 3.2)
   e) non-distinction of “e” and “ё” letters

Most of the controversial moments in the annotation dealt with colloquial forms such as *ваще* for *вообще* and *щас* for *сейчас*. In most of the cases they can be freely replaced by corresponding formal forms without any change in meaning, except for the expressive sentences like *«Ну ты ваще»* (1) or *«Да щас!»* (2), so in the latter cases there is no typo to correct. But obviously the border between these cases is very subtle so we deliberately decided to correct such colloquial forms in all the sentences.

Each of the 5,000 sentences was given to three annotators. Most of the annotators were the competition participants or students of linguistic and computer science departments. The annotation logs were automatically processed to select the sentences where all the three annotators gave the same answer and then manually filtered to avoid prevalence of several frequent typo patterns. Finally, about 2,400 mistyped sentences remained. The sample was extended by 1,600 correctly typed sentences obtained from the same corpora. The final sample of 4,000 sentences was randomly subdivided by two equal portions, each containing 2,000 sentences. The first half was given to the competitors as the development set. Such small size of the development set gave the participants no possibility to learn language model, however, they could use this sample to tune the parameters of their algorithm: e.g. the weighted Levenshtein distance used for candidate search or the weights of different features (error model, language model, morphology model etc.) in the final decision procedure. We also provided an evaluation script using which the participants could measure the performance of their systems on the development set. Since we have not provided any dictionary or corpora resources, the competitors were allowed to use arbitrary dictionary to search for candidates and arbitrary corpus, say, to fit the language model.

Since 2,000 sentences can be manually corrected in one or two days, they were randomly inserted into the sample of 100,000 sentences taken from the same corpus. The participants had no information about this fact and were asked to send the

answers for the whole test sample. However, the correctness was evaluated only on the 2,000 sentences from the test set.

3.2. Evaluation and metrics

The proper selection of evaluation metric for the competition was not a trivial task. A common metric for Web search spelling correction is the fraction of correctly restored queries, its direct analogue is the percentage of correct sentences. However, it is uninformative for our task: this metric cannot show the difference in performance between a system with high recall which corrects all the typos but also a lot of correctly typed words, and a system which has high precision and corrects no sentences at all. This problem could be partially remedied by calculating the number of properly and improperly corrected sentences with typos, as well as the number of "false alarms" (improperly corrected sentences without typos), but this metric is also inadequate when sentences could contain several typos. For example, consider a sentence with two typos, the described evaluation algorithm cannot distinguish a sentence with only one typo corrected from a sentence with two typos corrected and properly and one correct word changed incorrectly.

Therefore we evaluate performance in terms of individual corrections, not the whole sentences. That raises the problem of sentence alignment: in the case of space or hyphen orthographic error one word in the source sentence may correspond to multiple words in the correction, as well as many words in the source to a single one in the corrected sentence. We aligned the sentences using the following procedure:

1) First, the sentence was converted to lowercase and split by the space symbols.
2) The isolated punctuation marks were removed.
3) Since most of the punctuation symbols are not separated from the previous words, all non-alphabetic characters were deleted on both edges of each word.
4) Then the source sentence and its correction were aligned using the following algorithm:
    1. Longest common subsequence was extracted using standard dynamic programming algorithm. Words on the same position in the subsequence were aligned to each other.
    2. Each of the nonidentical groups between alignment points constructed on the previous step was aligned separately. We constructed a Levenshtein alignment between source and correction sides of the groups using standard edit distance with permutations, separating the words in groups by spaces. If an alignment point was located between the words both on the source and correction sides, then this point was added to the alignment.

Below we explain this algorithm on the sentence *«помоему, кто то из них то же ошипся»* (3) and its correction *по-моему, кто-то из них тоже ошибся*. After the removal of punctuation marks and first step of the alignment algorithm we obtain the following alignment groups:

(4)

помоему кто то		*по-моему кто-то*
из		*из*
них		*них*
то же	*ошипся*	*тоже ошибся*

When processing the pair (*помоему кто то, по-моему кто-то*), we observe that an optimal alignment matches the groups «*помоему*» and «*по-моему*» to each other. Since both these subgroups end on word edges, we obtain additional alignment pairs

(5)

помоему	*по-моему*
кто то	*кто-то*

and the remaining part

из	*из*
них	*них*
то же	*тоже*
ошипся	*ошибся*.

After constructing such alignment for all the pairs of source and correct sentences, we extracted from each sentences all the nonidentical pairs (*помоему/по-моему, кто то/кто-то* and т*о же/тоже* in the example above) and use these tokens for performance evaluation. We executed the same procedure on the pairs of source and candidate sentences, where the candidate sentences are obtained from the correction sentences. We obtain two sets S_{corr} and S_{part} containing pairs of the form ((sentence number, source token), correction token) for source-correct and source-participant alignments. Then we calculated the number TP of true positives which is $||S_{corr} \cap S_{part}||$—the number of typo tokens properly corrected by the system. To obtain the precision score we divided this quantity by $|S_{part}|$ total number of corrections made by the system. The recall score was calculated as TP / $|S_{part}|$—the fraction of typo tokens, which were corrected properly. Note that false negatives in this case are both typos for which a wrong correction was selected and the typos left without correction.

We calculated F1-measure as the harmonic mean between precision and recall. All the three metrics were reported by the evaluation script; however, only F1-measure was used to rank the participants. In the final version of the evaluation script we also reported the percentage of correct sentences just for comparison.

When testing the alignment procedure, we found one subtle case not captured by the alignment algorithm. Consider the source sentence «*я не сколька не ожидал его увидеть*» (6) and its correction «*я нисколько не ожидал его увидеть*» (7). Suppose the spellchecker corrected both the mistyped words but did not manage to remove the space, yielding the sentence «*Я ни сколько не ожидал его увидеть*» (8). Literal application of the procedure above gives us two nontrivial alignment groups in the source-candidate pair: «*не/ни*» and «*сколька/сколько*». Both these pairs were not observed in the reference alignment, therefore we obtain two false positives. Note that leaving the mistyped word «*сколька*» untouched yields better score since in this case only one unobserved aligned pair «*не/ни*» appears.

To improve this deficiency we made the following minor correction: we forced the alignment between source and suggestion sentences to have the same source components as in the source-correct alignment. For example, in the sentence above, the groups *«не/ни»* and *«сколька/сколько»* were joined together to obtain the pair *«не сколька/ни сколько»*, contributing one false positive instead of two.

3.3. Competition and participants

Seven research groups from 4 universities (MSU, MIPT, HSE, ISP RAS), 3 IT-companies (InfoQubes, NLP@Cloud, Orfogrammatika) and 2 cities (Moscow, Novosibirsk) successfully participated in the competition. These groups are listed in Table 1. Only best results from each group were taken into consideration and the number of attempts was not limited.

Table 1. Participants of SpellRuEval competition

Code of the group	scientific group
A	MIPT
B	GICR, MSU
C	HSE CompLing Spell
D	InfoQubes
E	ISP RAS
F	NLP@CLOUD
G	Orfogrammatika

4. Results and Discussion

All the systems presented in Table 2 used different toolkits and methods of automatic spelling correction, some of them are first time applied for Russian.

Table 2. Results of SpellRuEval competition

place	scientific group	Precision	Recall	F-measure	Accuracy
1	B	81.98	69.25	75.07	70.32
2	G	67.54	62.31	64.82	61.35
3	A	71.99	52.31	60.59	58.42
4	E	60.77	50.75	55.31	55.93
	BASELINE	55.91	46.41	50.72	48.06
5	C	74.87	27.99	40.75	50.45
6	D	23.50	30.00	26.36	24.95
7	F	17.50	9.65	12.44	33.96

We also evaluated a baseline system. Like several participant systems, it uses edit distance for search and combination of edit distance and n-gram model probability for ranking. In takes all the candidates on the distance of at most one edit from the source word and rank the obtained sentences using the sum of logarithmic edit distance score and language model score. Trigram language model was obtained using KenLM toolkit [Heafield et al., 2013] trained on the same data that was used by team B.

4.1. Methods and their efficiency

We collected the information about the methods and tools used by the competitors in Table 3 in the Appendix. Competition results show that all the teams used large dictionaries with approximately the same size, however, the difference in results is substantial. It means that the algorithms used for correction are more significant than additional data. It also proves that it is more important (and more difficult) to select a correct candidate than to find it, though several types of errors cannot be captured by basic search model based on edit distance without using additional errors lists or phonetic similarity. All the three top-ranked teams used a combination of edit distance and trigram language model for candidate ranking. It is interesting that morphological and semantic information gives no or little advantage in comparison with language model. One of the teams used Word2Vec instead of traditional n-gram model, which results in rather high precision (but not the best among all participants), though the recall was moderate in comparison with other results. It shows that Word2Vec is very successful in capturing frequent patterns, however, this method alone cannot detect all the errors. As expected, real-word errors were the most difficult to capture even by the top competitors, another source of difficult errors were misplaced hyphens and spaces. Probably, to correct such errors at least some rules of Russian orthography and grammar should be handcrafted, since such errors are too frequent and subtle to be captured by pure statistical methods. Last but not the least, it is interesting that the competition winner (and actually the three best teams) used a rather simple algorithm: find the candidates using Levenshtein distance and rerank them using language model. It offers much room for future improvement by careful integration morphological, grammar or semantic features; however, it is not an easy task, as direct incorporation of morphology gave no advantage in current competition.

5. Conclusions

SpellRuEval 2016 has brought together a number of IT companies and academic groups that work on Russian Web text processing and normalization, so that it became possible to compare state-of-the-art methods in the field for Russian. The results have shown that the problem of automatic spelling correction for Russian social media is far from its solution. Up to now the best results are obtained using simple combination of edit distance candidate search and trigram language model, so future improvement can be achieved by adding morphological and semantic component to this basic framework.

The competition has the following practical outcomes:

- we have measured the current baseline of automatic spelling correction for Russian: on social media the baseline method show F1-measure of 50% and sentence accuracy of 48%. State-of-the-art methods used by the competition winner achieve F1-Measure of 75% and sentence accuracy of 70%.
- Various approaches to automatic spelling correction for Russian were tested; we have compared the role of different language models (ngram vs Word2Vec), different candidate search algorithms (dictionary lookup vs dictionary-free) and relative significance of different model elements (dictionary size, edit distance, language model, morphology and semantics usage). The results show that dictionary size is not the main factor, much more important is the adequacy of ranking model. Using more fine-grained features than simple edit distance score also improves the performance slightly. However, current system gain little or no advantage from morphological or semantic information which leaves much room for future improvement.
- the manually tagged golden standard set was developed, consisting of nearly 2000 sentences with different types of mistakes (typos, grammatical, orthographic, cognitive errors etc.) and their corrected variants. The organizers hope that the training set and golden standard (available at URL http://www.webcorpora.ru/wp-content/uploads/2016/01/source_sents.txt and http://www.webcorpora.ru/wp-content/uploads/2016/01/corrected_sents.txt) will help other researchers to evaluate their algorithms;

The experience of first SpellRuEval challenge could be useful for organizers and participants in future spell checking competitions. It would be interesting to test how linguistic information such as morphology, syntax or semantics could help in this task. The methods proposed could be also helpful in similar task like automatic grammar correction or social media text normalization. We hope to present one of these tasks in future Dialogue Evaluation competitions.

Acknowledgements

We would like to thank all colleagues who participated in the annotation of the golden standard. We also thank all the teams who took part in the competition, particularly to HSE and Orfogrammatika teams for their fruitful suggestions. We also would like to thank Eugeny Indenbom and Vladimir Selegey for their deep insight and helpful advice during the organization of the competition.

References

1. *Ahmad F., Kondrak G.* (2005) Learning a spelling error model from search query logs //Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.—Association for Computational Linguistics,—pp. 955–962.

2. *Andrew Carlson and Ian Fette*. Memory-based context-sensitive spelling correction at web scale //Machine learning and applications. ICMLA 2007, Sixth international conference on.—IEEE, 2007.—pp. 166–171.
3. *Baytin A.* (2008), Search query correction in Yandex [Ispravlenie poiskovykh zaprosov v Yandekse], Russian Internet technologies [Rossijskie Internet-tekhnologii], 2008.
4. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.,* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. Proceedings of Web as Corpus Workshop (WAC-8), Lancaster.
5. *Brill E., Moore R. C.* (2000) An improved error model for noisy channel spelling correction. In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp. 286–293. Association for Computational Linguistics.
6. *Cucerzan S., Brill E.* (2004) Spelling correction as an iterative process that exploits the collective knowledge of web users. In Proceedings of EMNLP 2004, pp. 293–300.
7. *Dale R., Anisimoff I., Narroway G.* (2012) A Report on the Preposition and Determiner Error Correction Shared Task. In Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications.
8. *Dale R., Kilgarriff A.* (2011) Helping Our Own: The HOO 2011 Pilot Shared Task. In Proceedings of the 13th European Workshop on Natural Language Generation.
9. *Damerau F. J.* (1964) A technique for computer detection and correction of spelling errors. Communications of the ACM-7, pp. 171–176.
10. *Duan, H., Hsu, B.-J.* (2011) Online Spelling Correction for Query Completion. In Proceedings of the International World Wide Web Conference, WWW 2011, March 28–April 1, Hyderabad, India.
11. *Flor M.* (2012) Four types of context for automatic spelling correction //TAL.—T. 53.—№ 3.—pp. 61–99.
12. *Golding A. R., Roth D.* (1999) A winnow-based approach to context-sensitive spelling correction //Machine learning.—Vol. 34.—№ 1–3.—p. 107–130.
13. *Golding A. R., Schabes Y.* (1996) Combining trigram-based and feature-based methods for context-sensitive spelling correction //Proceedings of the 34th annual meeting on Association for Computational Linguistics.—Association for Computational Linguistics,—pp. 71–78.
14. *Heafield K., Pouzyrevsky I., Clark J., Koehn I.* (2013) Scalable Modified Kneser-Ney Language Model Estimation //ACL (2).—pp. 690–696.
15. *Hirst G., Budanitsky A.* (2005) Correcting real-word spelling errors by restoring lexical cohesion //Natural Language Engineering.—Vol. 11.—№ 01.—pp. 87–111.
16. *Kernighan M. D., Church K. W., and Gale W. A.* (1990) A spelling correction program based on a noisy channel model. In Proceedings of the 13th conference on Computational linguistics, pages 205–210. Association for Computational Linguistics.
17. *Kukich K.* (1992) Techniques for automatically correcting words in texts. ACM Computing Surveys 24, pp. 377–439.
18. *Li M., Zhang Y. , Zhu M., Zhou M.* (2006) Exploring distributional similarity based models for query spelling correction //Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting

of the Association for Computational Linguistics.—Association for Computational Linguistics, 2006.—pp. 1025–1032.
19. *Liu V., Curran J. R.* (2007) Web Text Corpus for Natural Language Processing // EACL.—2006.
20. *Manning C.* (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? Proc. of CICLing.
21. *Mays E., Damerau F. J., Mercer R. L.* (1991) Context based spelling correction // Information Processing & Management. –Vol. 27.—№ 5.—pp. 517–522.
22. *McIlroy M. D.* (1982) Development of a Spelling List. AT&T Bell Laboratories.
23. *Mitton R.* (1987) Spelling checkers, spelling correctors and the misspellings of poor spellers //Information processing & management.—1987.—Vol. 23.—№ 5.—pp. 495–505.
24. *Mohit B., Rozovskaya A., Habash N., Zaghouani W., Obeid O.* (2014) The First QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of EMNLP Workshop on Arabic Natural Language Processing, Doha, Qatar, October.
25. *Ng H. T., Wu S. M., Briscoe T., Hadiwinoto C., Susanto R. H., Bryant C.* (2014) The CoNLL-2014 Shared Task on Grammatical Error Correction. In Proceedings of CoNLL: Shared Task.
26. *Ng H. T., Wu S. M., Wu Y., Hadiwinoto Ch., Tetreault J.* (2013) The CoNLL-2013 Shared Task on Grammatical Error Correction. In Proceedings of CoNLL: Shared Task.
27. *Oflazer K.* (1996) Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction //Computational Linguistics.—Vol. 22.—№ 1.—pp. 73–89.
28. *Panina M. F., Baitin A. V., Galinskaya I. E.* (2013) Context-independent autocorrection of query spelling errors. [Avtomaticheskoe ispravlenie opechatok v poiskovykh zaprosakh bez ucheta konteksta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Bekasovo, pp. 556–568.
29. *Pedler J., Mitton R* (2010). A large list of confusion sets for spellchecking assessed against a corpus of real-word errors //Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).
30. *Popescu O., Phuoc An Vo N.* (2014) Fast and Accurate Misspelling Correction in Large Corpora. Proceedings of EMNLP 2014: Conference on Empirical Methods in Natural Language Processing. Doha, Qatar.
31. *Ristad E. S., Yianilos P. N.* (1998) Learning string-edit distance //Pattern Analysis and Machine Intelligence, IEEE Transactions on.—Vol. 20.—№ 5.—p. 522–532.
32. *Rozovskaya A., Bouamor H., Habash N., Zaghouani W., Obeid O., Mohit B.* (2015) The Second QALB Shared Task on Automatic Text Correction for Arabic. ANLP Workshop 2015, The Second Workshop on Arabic Natural Language Processing, July 30, 2015 Beijing, China.
33. *Rozovskaya A., Roth D.* (2013) Joint learning and inference for grammatical error correction //Urbana.—Vol. 51.—pp. 61801.

34. *Schaback J., Li F.* (2007) Multi-level feature extraction for spelling correction // IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data.—pp. 79–86.
35. *Schulz K., Mihov S.* (2002) Fast string correction with Levenshtein automata // International Journal on Document Analysis and Recognition.—Vol. 5.—№ 1.—pp. 67–85.
36. *Shavrina T., Sorokin A.* (2015) Modeling Advanced Lemmatization for Russian Language Using TnT-Russian Morphological Parser. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015", RSUH, Moscow.
37. *Toutanova K., Moore R. C.* (2002) Pronunciation modeling for improved spelling correction //Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.—Association for Computational Linguistics—pp. 144–151.
38. *Vladimir I. Levenshtein,* (1965) Binary codes capable of correcting deletions, insertions, and reversals [Dvoichnye kody s ispravleniem vypadenij, vstavok i zameschchenij simvolov], Doklady Akademij Nauk SSSR.—1965.—Vol. 163.—pp. 845–848.
39. *Wang K., Pedersen J.* (2011) Review of MSR-Bing web scale speller challenge. In Proceeding of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1339–1340.
40. *Wang K., Thrasher C., Viegas E, Li X., Hsu B. H.* (2010) «An overview of Microsoft Web N-gram corpus and applications.» In Proceedings of the NAACL HLT 2010 Demonstration Session, pp. 45–48.
41. *Whitelaw C., Hutchinson B., Chung G. Y, Ellis G.* (2009) Using the web for language independent spellchecking and autocorrection. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, pp. 890–899. Association for Computational Linguistics.
42. *Corpus of Russian Student Texts* web-corpora.net/CoRST/search/index.php?interface_language=ru.
43. *Hunspell:* open source spell checking, stemming, morphological analysis and generation, 2011. http://hunspell.sourceforge.net/.
44. *Russian National Corpora* http://www.ruscorpora.ru/search-main.html.
45. *Russian OpenCorpora* http://opencorpora.org/.
46. *Spelling Alteration for Web Search Workshop.* 2011. Bellevue, USA http://spellerchallenge.com/Workshop.aspx.

Appendix 1. Analysis of methods used by SpellRuEval participants

Group code	Methods used	Dictionary size: wordforms	Dictionary size: lemmata	Error detection	Error correction	Most common mistakes of the system:	Training set (excluding SpellRuEval training set)
A	Edit distance for candidate search and language model for selection	3,700,000	230,000	Dictionary-based by edit distance, keyboard adjacency and graphic similarity	Correction is selected by trigram language model	wrong candidate ranking, syntactic errors, real-word errors	Corpus of 213,000,000 words with morphology for language model
B	Multi-level spelling correction, error model, n-grams, POS-tag n-grams	3,700,000	230,000	Each word is a potential error, the best variant is selected by the correction score	Edit distance and phonetic coding for candidate search, trigram language model and error model for ranking	semantic errors, wrong candidate ranking	50 million tokens from Live Journal
C	Word2Vec, n-grams, hybrid error model combining: 1) traditional channel model that uses single letter edits, 2) the model introduced by Brill and Moore, 3) extended version of the channel model with wider context edits	no dictionary used	no dictionary used	error detection confidence classifier, uses word2vec filtered vector scores	Autocorrector that processes words flagged as misspellings by the classifier	wrong candidate ranking, syntactic errors, undetected errors, real-word errors	Corpus of Russian Student Texts (CoRST) 2.5 million tokens + 13.8 million tokens from Blogs + 22.2 million tokens from newspapers for n-gram model
D	edit distance, automatic rule-based paradigm construction	5,000,000	260,000	Dictionary based + dictionary look-up with suitable suffixation rules	Edit distance. If there are 2 or more candidates with same distance, choice is random	wrong inflection, wrong candidate ranking	No extra-data
E	n-grams, rule-based dictionary look-up	5,095,000	390,000	Dictionary based (OpenCorpora Dictionary)	Candidate ranking with 3-grams from training set and edit distance	wrong candidate ranking, spacing errors	400 million tokens from newspapers, social networks and WikiPedia
F	chunking in dependency model, vector model, Jflex, Apache Tika	5,095,000	390,000	Dictionary based (OpenCorpora Dictionary)	Ranking the ChunkTrees with max.number of words in the sentence, then correcting wordform depending on syntactic and POS-tags.	undetected errors, orthographic errors	No extra-data used
G	n-grams, POS-tag n-grams	5,500,000	400,000	Dictionary-based	Candidate ranking with wordform 3-grams, POS-tag 2- and 3-grams, wordform frequency etc.	wrong candidate ranking, word separation errors	RNC sample of 1 million tokens with resolved homonymy