SpellRuEval: the First Competition on Automatic Spelling Correction for Russian

Alexey Sorokin[1,2,3], Alexey Bajtin[4],
Irina Galinskaya[4], Tatiana Shavrina[1,3]

[1]Moscow State University, [2]Moscow Institute of Science and Technology,
[3]General Internet Corpus of Russian, [4]Yandex

Dialogue
22nd International Conference on Computational Linguistics
Moscow, RSUH, 4th June, 2016

Applications of spellchecking

- Spellchecking has broad applications:
  - Search queries correction, information extraction.
  - Text editors.
  - Preprocessing in NLP tasks (parsing, fact extraction, etc.)

Applications of spellchecking

- Spellchecking has broad applications:
  - Search queries correction, information extraction.
  - Text editors.
  - Preprocessing in NLP tasks (parsing, fact extraction, etc.)
- For social media (blogs, social networks, information sites) we also need text normalization (correction of informal expressions to their formal variants).
- Especially important — when creating annotated corpora of social media.

Applications of spellchecking

- Spellchecking has broad applications:
  - Search queries correction, information extraction.
  - Text editors.
  - Preprocessing in NLP tasks (parsing, fact extraction, etc.)
- For social media (blogs, social networks, information sites) we also need text normalization (correction of informal expressions to their formal variants).
- Especially important — when creating annotated corpora of social media.
- We need a benchmark algorithm for spellchecking.

Applications of spellchecking

- Spellchecking has broad applications:
  - Search queries correction, information extraction.
  - Text editors.
  - Preprocessing in NLP tasks (parsing, fact extraction, etc.)
- For social media (blogs, social networks, information sites) we also need text normalization (correction of informal expressions to their formal variants).
- Especially important — when creating annotated corpora of social media.
- We need a benchmark algorithm for spellchecking.
- And a benchmark dataset...

Spellchecking contests

- HOO Shared Task (Dale et al., 2010, 2011; assisting authors in writing)
- Microsoft Spelling Alteration Workshop (Wang, 2011; search query correction)
- CoNLL 2013 Shared Task (Ng, 2013; grammar error correction)

Spellchecking contests

- HOO Shared Task (Dale et al., 2010, 2011; assisting authors in writing)
- Microsoft Spelling Alteration Workshop (Wang, 2011; search query correction)
- CoNLL 2013 Shared Task (Ng, 2013; grammar error correction)

Other languages:

- QALB-2014 (Arabic; Mohit, 2014; automatic text correction)
- QALB-2015 (Arabic; Rozovskaya, 2015)

Spellchecking contests

- HOO Shared Task (Dale et al., 2010, 2011; assisting authors in writing)
- Microsoft Spelling Alteration Workshop (Wang, 2011; search query correction)
- CoNLL 2013 Shared Task (Ng, 2013; grammar error correction)

Other languages:

- QALB-2014 (Arabic; Mohit, 2014; automatic text correction)
- QALB-2015 (Arabic; Rozovskaya, 2015)

Text normalization:

- Twitter lexical normalization (ACL-IJCNLP Shared Task; Baldwin et al., 2015)

Goals of SpellRuEval Shared Task

- Few works address Russian spellchecking.
- But spellchecking is a *language-dependent* task.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Goals of SpellRuEval Shared Task

- Few works address Russian spellchecking.
- But spellchecking is a *language-dependent* task.
- There is *NO* baseline algorithm.
- There is *NO* reference corpus.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Goals of SpellRuEval Shared Task

- Few works address Russian spellchecking.
- But spellchecking is a *language-dependent* task.
- There is *NO* baseline algorithm.
- There is *NO* reference corpus.

Our goals

- Create a reference typos corpus for future research.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Goals of SpellRuEval Shared Task

- Few works address Russian spellchecking.
- But spellchecking is a *language-dependent* task.
- There is *NO* baseline algorithm.
- There is *NO* reference corpus.

Our goals

- Create a reference typos corpus for future research.
- Compare existing algorithms for spelling correction.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Goals of SpellRuEval Shared Task

- Few works address Russian spellchecking.
- But spellchecking is a *language-dependent* task.
- There is *NO* baseline algorithm.
- There is *NO* reference corpus.

Our goals

- Create a reference typos corpus for future research.
- Compare existing algorithms for spelling correction.
- Determine a baseline algorithm.

Data description

- Source: LJ subcorpus of GICR.

Data description

- Source: LJ subcorpus of GICR.
- Types of errors:
  - typos (мнея $\mapsto$ меня),
  - orthographic errors (митель $\mapsto$ метель),
  - cognitive errors (компания $\mapsto$ кампания),
  - intentional incorrect writing (хоцца $\mapsto$ хочется, ваще v вообще),
  - grammatical errors (agreement etc.) (он видят $\mapsto$ он видит),
  - errors in hyphen and space positioning (както $\mapsto$ как-то),
  - mixed usage of digits and letters in numerals (2-ух $\mapsto$ двух),
  - usage of digits instead of letters (в4ера $\mapsto$ вчера)

Data description

- Source: LJ subcorpus of GICR.
- Types of errors:
  - typos (мнея $\mapsto$ меня),
  - orthographic errors (митель $\mapsto$ метель),
  - cognitive errors (компания $\mapsto$ кампания),
  - intentional incorrect writing (хоцца $\mapsto$ хочется, ваще v вообще),
  - grammatical errors (agreement etc.) (он видят $\mapsto$ он видит),
  - errors in hyphen and space positioning (както $\mapsto$ как-то),
  - mixed usage of digits and letters in numerals (2-ух $\mapsto$ двух),
  - usage of digits instead of letters (в4ера $\mapsto$ вчера)
- Other errors (not corrected):
  - foreign words including cyrillic (e.g Ukrainian or Belorussian),
  - informal abbreviations (прога $\mapsto$ программа)
  - punctuation errors (all punctuation is omitted)
  - capitalization errors
  - non-distinction of *e* and *ë* letters

Data sources

- Initial data: about 10000 sentences with typos automatically extracted from GICR.

Data sources

- Initial data: about 10000 sentences with typos automatically extracted from GICR.
- About 5000 sentences after manual inspection (removal of frequent patterns).

Data sources

- Initial data: about 10000 sentences with typos automatically extracted from GICR.
- About 5000 sentences after manual inspection (removal of frequent patterns).
- About 500 sentences with real-word errors (*не/ни*, *кампания/компания*) added.

Data sources

- Initial data: about 10000 sentences with typos automatically extracted from GICR.
- About 5000 sentences after manual inspection (removal of frequent patterns).
- About 500 sentences with real-word errors (*не/ни*, *кампания/компания*) added.
- Finally, more than 5000 sentences given to annotators.

Annotation procedure

- Annotators were asked to correct the whole sentence:

  **Ну вапрос канешна интиресный*

  Ну вопрос конечно интересный

Annotation procedure

- Annotators were asked to correct the whole sentence:

  **Ну вапрос канешна интиресный*

  Ну вопрос конечно интересный

- Annotators were allowed to express their doubt when meeting controversions.

Annotation procedure

- Annotators were asked to correct the whole sentence:

  **Ну вапрос канешна интиресный*

  Ну вопрос конечно интересный
- Annotators were allowed to express their doubt when meeting controversions.
- Each sentence processed by 3 annotators, only 2500 sentences with no conflicts between annotators were retained.

Annotation procedure

- Annotators were asked to correct the whole sentence:

  **Ну вапрос канешна интиресный*

  Ну вопрос конечно интересный
- Annotators were allowed to express their doubt when meeting controversions.
- Each sentence processed by 3 annotators, only 2500 sentences with no conflicts between annotators were retained.
- 1500 sentences without typos added as filters.
- Finally, 4000 sentences were collected, 2000 for training and 2000 for testing.

Annotation procedure

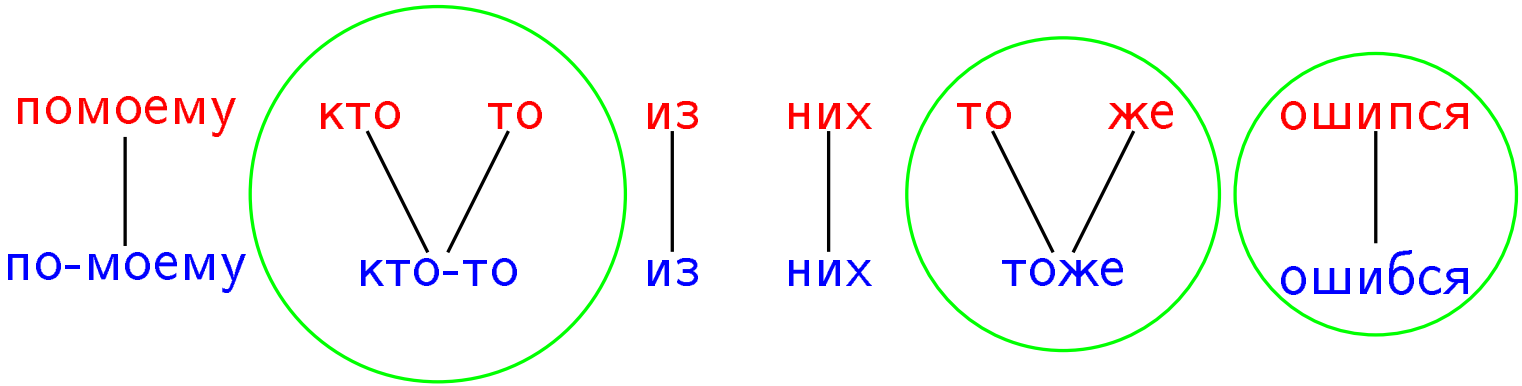
- Annotators were asked to correct the whole sentence:

  **Ну вапрос канешна интиресный*

  Ну вопрос конечно интересный
- Annotators were allowed to express their doubt when meeting controversions.
- Each sentence processed by 3 annotators, only 2500 sentences with no conflicts between annotators were retained.
- 1500 sentences without typos added as filters.
- Finally, 4000 sentences were collected, 2000 for training and 2000 for testing.
- During testing procedure, these 2000 sentences were uniformly spread among 100000 taken from the same sample.

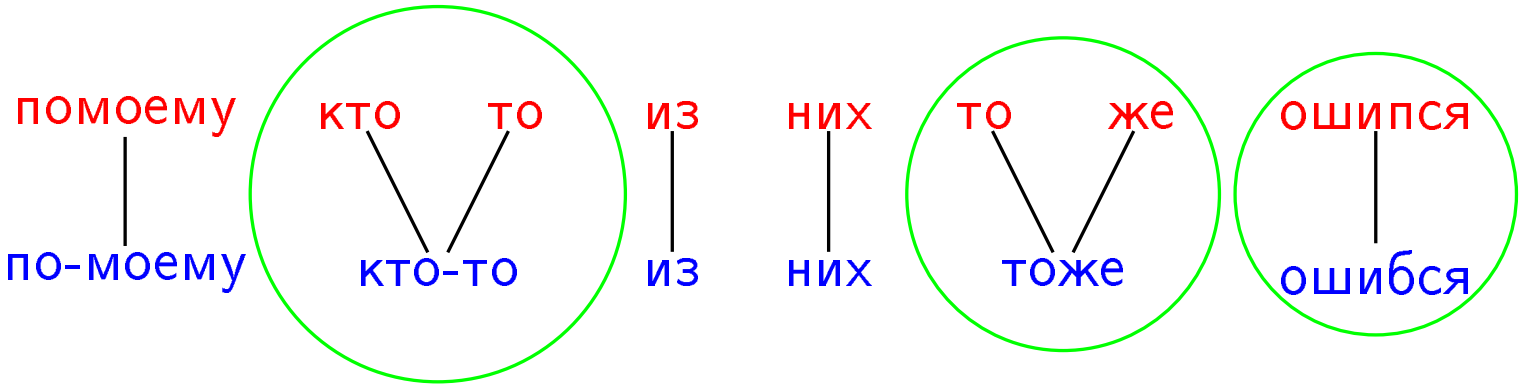
Measuring performance

- First, we aligned the source sentence with its correction:



Measuring performance

- First, we aligned the source sentence with its correction:



- We obtain a list of non-identical pairs (tokens):

кто то	кто-то
то же	тоже
ошипся	ошибся

- They are called *nontrivial tokens*.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Measuring performance

- We collected the set S_{corr} of nontrivial tokens in source-etalon alignment.

Measuring performance

- We collected the set S_{corr} of nontrivial tokens in source-etalon alignment.
- Analogous set S_{part} is obtained from source-system alignment.
- Source-system tokens were forced to have the same source part as source-etalon where possible.

Measuring performance

- We collected the set S_{corr} of nontrivial tokens in source-etalon alignment.
- Analogous set S_{part} is obtained from source-system alignment.
- Source-system tokens were forced to have the same source part as source-etalon where possible.
- We use the following measures:

$$P = \frac{|S_{corr} \cap S_{part}|}{|S_{part}|} \qquad R = \frac{|S_{corr} \cap S_{part}|}{|S_{corr}|} \qquad F1 = \frac{2PR}{P+R}$$

Measuring performance

- We collected the set S_{corr} of nontrivial tokens in source-etalon alignment.
- Analogous set S_{part} is obtained from source-system alignment.
- Source-system tokens were forced to have the same source part as source-etalon where possible.
- We use the following measures:

$$P = \frac{|S_{corr} \cap S_{part}|}{|S_{part}|} \qquad R = \frac{|S_{corr} \cap S_{part}|}{|S_{corr}|} \qquad F1 = \frac{2PR}{P+R}$$

- The number of properly corrected sentences was also measured.

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.
- Candidates are usually searched on Levenstein distance 1.
- For (say, **свена*) we have ...

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.
- Candidates are usually searched on Levenstein distance 1.
- For (say, **свена*) we have ...

свеча, вена, смена, стена, сцена, сена

- Which word to select?

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.
- Candidates are usually searched on Levenstein distance 1.
- For (say, **свена*) we have ...

свеча, вена, смена, стена, сцена, сена

- Which word to select?
- The word which better fits context:

Свеча горела на столе, ___ горела.

Лошади дали две охапки ___.

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.
- Candidates are usually searched on Levenstein distance 1.
- For (say, **свена*) we have ...

свеча, вена, смена, стена, сцена, сена

- Which word to select?
- The word which better fits context:

Свеча горела на столе, свеча горела.

Лошади дали две охапки сена.

Expected approach

- Suppose we have an out-of-vocabulary word (say, **свена*), which is a possible typo.
- Candidates are usually searched on Levenstein distance 1.
- For (say, **свена*) we have ...

свеча, вена, смена, стена, сцена, сена

- Which word to select?
- The word which better fits context:

Свеча горела на столе, свеча горела.

Лошади дали две охапки сена.

- We expect some kind of language model for candidate selection.

Difficult cases

Our dataset contained several difficult cases:

- Real-word errors:

на сайте асинхронные машины

откапал всю нужную информацию

Difficult cases

Our dataset contained several difficult cases:

- Real-word errors:

  на сайте асинхронные машины

  откапал всю нужную информацию

- Space and hyphen errors:

  ладно когда он какбы из под полы

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina

[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Difficult cases

Our dataset contained several difficult cases:

- Real-word errors:

  на сайте асинхронные машины
  откапал всю нужную информацию

- Space and hyphen errors:

  ладно когда он какбы из под полы

- Colloquial expressions distant from formal counterpart:

  а ваще, Жень, я, оказываецца, убежденный урбанист.

Difficult cases

Our dataset contained several difficult cases:

- Real-word errors:

  на сайте асинхронные машины
  откапал всю нужную информацию

- Space and hyphen errors:

  ладно когда он какбы из под полы

- Colloquial expressions distant from formal counterpart:

  а ваще, Жень, я, оказываецца, убежденный урбанист.

- Proper names, often colloquial

  Ну и конечно Мак покатился с ним, чтобы оценить ситуацию со стороны.

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.
  - Rank the candidates by sum of edit and language model scores.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.
  - Rank the candidates by sum of edit and language model scores.
- Three top-ranked systems:
  - Find candidates on 1 edit distance for each word.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.
  - Rank the candidates by sum of edit and language model scores.
- Three top-ranked systems:
  - Find candidates on 1 edit distance for each word.
  - Calculate language model score for each candidate sentence.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.
  - Rank the candidates by sum of edit and language model scores.
- Three top-ranked systems:
  - Find candidates on 1 edit distance for each word.
  - Calculate language model score for each candidate sentence.
  - Use them as features together with other scores (length, morphology, capitalization etc.)
- One of the systems used distributional semantics to define error positions, but it was not among the top ones.

Proposed methods

- Baseline method:
  - Find candidates on 1 edit distance for each word.
  - Calculate edit probability by multiplying probabilities of primitive edits(each nontrivial edit has fixed probability 0.1).
  - Calculate language model probability for each candidate sentence.
  - Rank the candidates by sum of edit and language model scores.
- Three top-ranked systems:
  - Find candidates on 1 edit distance for each word.
  - Calculate language model score for each candidate sentence.
  - Use them as features together with other scores (length, morphology, capitalization etc.)
- One of the systems used distributional semantics to define error positions, but it was not among the top ones.
- The winner system also uses:
  - Phonetic similarity,
  - Weighted edit model.

Results

Place	Team	Prec.	Recall	F-Measure	Corr. sent.
1	GICR, MSU	81,98	69,25	75,07	70,32
2	Orfogrammatika	67,54	62,31	64,82	61,35
3	MIPT	71,99	52,31	60,59	58,42
4	ISP RAS	60,77	50,75	55,31	55,93
	BASELINE	55,91	46,41	50,72	48,06
5	HSE	74,87	27,99	40,75	50,45
6	InfoCubes	23,50	30,00	26,36	24,95
7	NLP@CLOUD	17,50	9,65	12,44	33,96

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.
  - Combination of edit distance and trigram language model for selecting top candidates.

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.
  - Combination of edit distance and trigram language model for selecting top candidates.
  - Reranking of candidates using various features (capitalization, spaces, hyphens, etc.)

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.
  - Combination of edit distance and trigram language model for selecting top candidates.
  - Reranking of candidates using various features (capitalization, spaces, hyphens, etc.)
- Reranking model quality is far more important than dictionary size.

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.
  - Combination of edit distance and trigram language model for selecting top candidates.
  - Reranking of candidates using various features (capitalization, spaces, hyphens, etc.)
- Reranking model quality is far more important than dictionary size.
- Using morphology and semantics gives no or little gain.

Conclusions

- Optimal system is rather simple, it includes:
  - Searching for candidates by edit distance and phonetic encoding.
  - Combination of edit distance and trigram language model for selecting top candidates.
  - Reranking of candidates using various features (capitalization, spaces, hyphens, etc.)
- Reranking model quality is far more important than dictionary size.
- Using morphology and semantics gives no or little gain.

All materials available at:
https://drive.google.com/drive/u/0/folders/0B8XxHuDfyogmMnM5RFdKdWdzZmM

Спасибо за внимание!

Thank you for attention!

Alexey Sorokin, Alexey Bajtin, Irina Galinskaya, Tatiana Shavrina
[1]Moscow State University, [2]Moscow Institute of Science and Technology, [3]General Internet Corpus of Russian, [4]Yandex