Three-part diachronic semantic change dataset for Russian

Andrey Kutuzov, Lidia Pivovarova

University of Oslo, University of Helsinki

https://github.com/akutuzov/rushifteval_public

Contents

1. RuShiftEval dataset construction
   - Historical periods to compare
   - Target word list creation
2. Annotation setup
   - DURel framework
3. RuShiftEval shared task
4. Diachronic trajectory types revealed in RuShiftEval
5. Summing up

RuShiftEval dataset construction

- RuShiftEval is a manually annotated dataset of graded diachronic semantic changes for Russian nouns.

RuShiftEval dataset construction

- RuShiftEval is a manually annotated dataset of graded diachronic semantic changes for Russian nouns.
- ...but wait, we have already seen a dataset for Russian, no?

RuShiftEval dataset construction

- RuShiftEval is a manually annotated dataset of graded diachronic semantic changes for Russian nouns.
- ...but wait, we have already seen a dataset for Russian, no?

Right. This was RuSemShift

- Two sub-sets (comparisons) each covering a specific pair of time periods:
  1. *RuSemShift*$_1$: **pre-Soviet VS Soviet** times (71 words)
  2. *RuSemShift*$_2$: **Soviet VS post-Soviet** times (69 words).

[Rodina and Kutuzov, 2020]

RuShiftEval dataset construction

- RuShiftEval is a manually annotated dataset of graded diachronic semantic changes for Russian nouns.
- ...but wait, we have already seen a dataset for Russian, no?

Right. This was RuSemShift

- Two sub-sets (comparisons) each covering a specific pair of time periods:
  1. *RuSemShift$_1$*: **pre-Soviet VS Soviet** times (71 words)
  2. *RuSemShift$_2$*: **Soviet VS post-Soviet** times (69 words).

  [Rodina and Kutuzov, 2020]

What is novel in RuShiftEval?

1. Adds the third sub-set (comparison):
   - **pre-Soviet VS post-Soviet** times
2. a new single set of target nouns over all three comparisons

The annotation effort for this shared task was supported by the Russian Science Foundation grant 20-18-00206. This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

Historical periods to compare

Time periods:

1. 1700 : 1916: the period of Russian Empire before the 1917 revolution (**pre-Soviet**).
2. 1918 : 1990: the period of the Soviet Union (**Soviet**).
3. 1992 : 2016: the period after the fall of the Soviet Union (**post-Soviet**).

Historical periods to compare

Time periods:

1. 1700 : 1916: the period of Russian Empire before the 1917 revolution (**pre-Soviet**).
2. 1918 : 1990: the period of the Soviet Union (**Soviet**).
3. 1992 : 2016: the period after the fall of the Soviet Union (**post-Soviet**).

Period pairs (sub-sets):

- RuShiftEval-1 (pre-Soviet VS Soviet)
- RuShiftEval-2 (Soviet VS post-Soviet)
- RuShiftEval-3 (pre-Soviet VS post-Soviet)

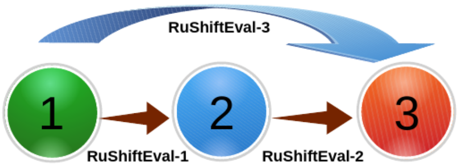
Historical periods to compare

Time periods:

1. 1700 : 1916: the period of Russian Empire before the 1917 revolution (**pre-Soviet**).
2. 1918 : 1990: the period of the Soviet Union (**Soviet**).
3. 1992 : 2016: the period after the fall of the Soviet Union (**post-Soviet**).

Period pairs (sub-sets):

- RuShiftEval-1 (pre-Soviet VS Soviet)
- RuShiftEval-2 (Soviet VS post-Soviet)
- RuShiftEval-3 (pre-Soviet VS post-Soviet)



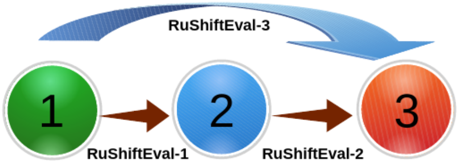
Historical periods to compare

Time periods:

1. 1700 : 1916: the period of Russian Empire before the 1917 revolution (**pre-Soviet**).
2. 1918 : 1990: the period of the Soviet Union (**Soviet**).
3. 1992 : 2016: the period after the fall of the Soviet Union (**post-Soviet**).

Period pairs (sub-sets):

- ► RuShiftEval-1 (pre-Soviet VS Soviet)
- ► RuShiftEval-2 (Soviet VS post-Soviet)
- ► RuShiftEval-3 (pre-Soviet VS post-Soviet)



Sentences for the annotation were sampled from the *Russian National Corpus* (RNC).

Target word list creation

The workflow was similar to [Kutuzov and Kuzmenko, 2018],[Rodina and Kutuzov, 2020], [Schlechtweg et al., 2020], etc.

Target word list creation

The workflow was similar to [Kutuzov and Kuzmenko, 2018],[Rodina and Kutuzov, 2020], [Schlechtweg et al., 2020], etc.

How we chose target words?

- Manually picked words with changed meaning from prior linguistic work and dictionaries.
- Added 2 randomly sampled 'fillers' or 'distractors' with similar frequency distributions per each target word.
- This alone does not give us relative change strength!
- For this, human annotation is needed

111 nouns total: 12 in the development set and 99 in the test set.

Contents

1. RuShiftEval dataset construction
   - Historical periods to compare
   - Target word list creation
2. Annotation setup
   - DURel framework
3. RuShiftEval shared task
4. Diachronic trajectory types revealed in RuShiftEval
5. Summing up

DURel framework

- *Diachronic Usage Relatedness* (DURel) semantic change annotation methodology [Schlechtweg et al., 2018]:
- The degree of semantic change is a **function of mean semantic relatedness across pairs of word's occurrences in different time periods**.
- The annotators are given 2 sentences from 2 time periods containing a target word
- asked to choose a relatedness score from 0 to 4:

DURel framework

- *Diachronic Usage Relatedness* (DURel) semantic change annotation methodology [Schlechtweg et al., 2018]:
- The degree of semantic change is a **function of mean semantic relatedness across pairs of word's occurrences in different time periods**.
- The annotators are given 2 sentences from 2 time periods containing a target word
- asked to choose a relatedness score from 0 to 4:

Score	Relatedness
0	Cannot decide
1	Senses unrelated
2	Senses distantly related
3	Senses closely related
4	Senses identical

[Hätty et al., 2019]

DURel framework

- Yandex.Toloka crowd-workers assigned relatedness scores for 30 randomly sampled sentence pairs for each target word and period pair (sub-set).
- Each sentence pair annotated by 3 human raters (about 100 for each sub-set).
- Native speakers of Russian, older than 30, with a university degree.

DURel framework

- Yandex.Toloka crowd-workers assigned relatedness scores for 30 randomly sampled sentence pairs for each target word and period pair (sub-set).
- Each sentence pair annotated by 3 human raters (about 100 for each sub-set).
- Native speakers of Russian, older than 30, with a university degree.
- RuShiftEval uses **COMPARE**: the mean relatedness between two time periods.
- The 1st sentence from the *earlier* period, and the 2nd sentence from the *later* period.
- Supposed to approximate the inverted degree of semantic change for a given word.

3 period pairs: 3 scores to be predicted for each word

The inter-rater agreement is on par with other semantic change annotation efforts.

Period pairs	Krippendorff α	Spearman ρ	Judgments	0-judgments
Test set (99 words)				
RuShiftEval-1	0.506	0.521	8 863	42
RuShiftEval-2	0.549	0.559	8 879	25
RuShiftEval-3	0.544	0.556	8 876	31
Development set (12 words)				
RuShiftEval-1	0.592	0.613	1 013	7
RuShiftEval-2	0.609	0.627	1 014	3
RuShiftEval-3	0.597	0.632	1 015	2

About 30 000 human judgments in total. Publicly available, including the raw scores.

Contents

1. RuShiftEval dataset construction
   - Historical periods to compare
   - Target word list creation
2. Annotation setup
   - DURel framework
3. RuShiftEval shared task
4. Diachronic trajectory types revealed in RuShiftEval
5. Summing up

RuShiftEval shared task

RuShiftEval'2021

- Shared task collocated with the Dialogue 2021 conference [Kutuzov and Pivovarova, 2021]
- First open shared task in graded semantic change detection for Russian
- Not surprisingly, used the RuShiftEval annotations to evaluate the submissions
- Participants could train on the prior RuSemShift dataset

RuShiftEval shared task

RuShiftEval'2021

- Shared task collocated with the Dialogue 2021 conference [Kutuzov and Pivovarova, 2021]
- First open shared task in graded semantic change detection for Russian
- Not surprisingly, used the RuShiftEval annotations to evaluate the submissions
- Participants could train on the prior RuSemShift dataset

Some results of the shared task

- Contextualized architectures topped the leaderboard: XLM-R, BERT and ELMo

RuShiftEval shared task

RuShiftEval'2021

- Shared task collocated with the Dialogue 2021 conference [Kutuzov and Pivovarova, 2021]
- First open shared task in graded semantic change detection for Russian
- Not surprisingly, used the RuShiftEval annotations to evaluate the submissions
- Participants could train on the prior RuSemShift dataset

Some results of the shared task

- Contextualized architectures topped the leaderboard: XLM-R, BERT and ELMo
- The first and the second best submissions relied on the multi-lingual XLM-R model,
  - But it didn't work so well at the SemEval'2020. Why?

RuShiftEval shared task

RuShiftEval'2021

- Shared task collocated with the Dialogue 2021 conference [Kutuzov and Pivovarova, 2021]
- First open shared task in graded semantic change detection for Russian
- Not surprisingly, used the RuShiftEval annotations to evaluate the submissions
- Participants could train on the prior RuSemShift dataset

Some results of the shared task

- Contextualized architectures topped the leaderboard: XLM-R, BERT and ELMo
- The first and the second best submissions relied on the multi-lingual XLM-R model,
  - But it didn't work so well at the SemEval'2020. Why?
- Using training data helps lexical semantic change detection
  - 4 top systems all train or fine-tune on *RuSemShift*

Contents

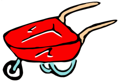
1. RuShiftEval dataset construction
   - Historical periods to compare
   - Target word list creation
2. Annotation setup
   - DURel framework
3. RuShiftEval shared task
4. Diachronic trajectory types revealed in RuShiftEval
5. Summing up

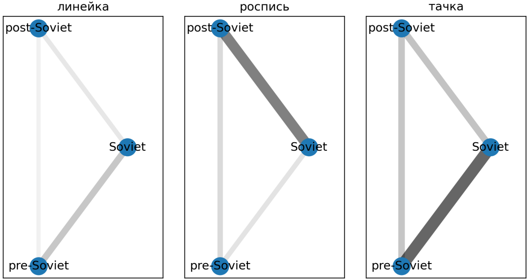
Diachronic trajectory types revealed in RuShiftEval

1. changes in every period pair, all relatedness scores are low: линейка ('carriage/ruler/series of goods')
2. change in the Soviet period VS the pre-Soviet period: роспись ('list/painting')
3. change in the post-Soviet period VS the Soviet period: тачка ('wheelbarrow/car')
4. (trivial) no changes: all three relatedness scores are high.
5. (not found) change in the Soviet period then coming back to the original meaning

Diachronic trajectory types revealed in RuShiftEval

1. changes in every period pair, all relatedness scores are low: линейка (‘carriage/ruler/series of goods’)
2. change in the Soviet period VS the pre-Soviet period: роспись (‘list/painting’)
3. change in the post-Soviet period VS the Soviet period: тачка (‘wheelbarrow/car’)
4. (trivial) no changes: all three relatedness scores are high.
5. (not found) change in the Soviet period then coming back to the original meaning





Time relatedness graphs. Nodes: time periods; edge width: relatedness scores.

Diachronic trajectory types revealed in RuShiftEval

Trajectory detection task: a toy preliminary experiment

- How good were the RuShiftEval submissions in capturing these trajectory types?
- Successful capturing is:
  - **Type 1**: percentile ranks of the scores for all 3 sub-sets are below 50
  - **Type 2**: score for the ‘Soviet:post-Soviet’ sub-set is the highest
  - **Type 3**: score for the ‘pre-Soviet:Soviet’ sub-set is the highest

Diachronic trajectory types revealed in RuShiftEval

Trajectory detection task: a toy preliminary experiment

- How good were the RuShiftEval submissions in capturing these trajectory types?
- Successful capturing is:
  - **Type 1**: percentile ranks of the scores for all 3 sub-sets are below 50
  - **Type 2**: score for the ‘Soviet:post-Soviet’ sub-set is the highest
  - **Type 3**: score for the ‘pre-Soviet:Soviet’ sub-set is the highest

Type	Example	Baseline	Top 4 systems
1	линейка (‘carriage/ruler/series of goods’)	0.5	**1.0**
2	роспись (‘list/painting’)	1.0	1.0
3	тачка (‘wheelbarrow/car’)	0.4	**0.8-1.0**

Percentages of words with correctly captured types. Baseline: diachronic CBOW and local neighbors [Hamilton et al., 2016]. Top systems: ELMo, BERT and XLM-R.

Contents

1. RuShiftEval dataset construction
   - Historical periods to compare
   - Target word list creation
2. Annotation setup
   - DURel framework
3. RuShiftEval shared task
4. Diachronic trajectory types revealed in RuShiftEval
5. Summing up

Summing up

A future sub-task?

- Performance in detecting diachronic trajectories correlates with the performance in ‘traditional’ graded semantic change...
- ...but not 100%

Summing up

A future sub-task?

- Performance in detecting diachronic trajectories correlates with the performance in ‘traditional’ graded semantic change...
- ...but not 100%
- Can be an interesting sub-task within semantic change detection...
- ...once more datasets like RuShiftEval are available...
- and ‘capturing the trajectory’ is defined more strictly.

Summing up

A future sub-task?

- Performance in detecting diachronic trajectories correlates with the performance in 'traditional' graded semantic change...
- ...but not 100%
- Can be an interesting sub-task within semantic change detection...
- ...once more datasets like RuShiftEval are available...
- and 'capturing the trajectory' is defined more strictly.

- Thanks for your attention!
- Feel free to use RuShiftEval!

`https://github.com/akutuzov/rushifteval_public`

References I

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
Cultural shift or linguistic drift? comparing two computational measures of semantic change.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019).
SURel: A gold standard for incorporating meaning shifts into term extraction.
In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.

References II

Kutuzov, A. and Kuzmenko, E. (2018).
Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes.
Quantitative Approaches to the Russian Language, page 95.

Kutuzov, A. and Pivovarova, L. (2021).
RuShiftEval: a shared task on semantic shift detection for Russian.
In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Rodina, J. and Kutuzov, A. (2020).
RuSemShift: a dataset of historical lexical semantic change in Russian.
In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

References III

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020).
SemEval-2020 task 1: Unsupervised lexical semantic change detection.
In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018).
Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.