Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines

Vadim Fomin [1]
Daria Bakshandaeva [1]
Julia Rodina [1]
Andrey Kutuzov [2]

[1]National Research University Higher School of Economics [2]University of Oslo

May 29, 2019

Contents

1 Introduction
2 Previous work
3 Our datasets
4 Our corpora
5 Word embeddings
6 Baseline results
- Local methods
- Global methods
- Results

7 Conclusion

Introduction

TL:DR

- Word meaning ≈ word contexts [Firth, 1957]
- Changes in contexts ≈ changes in meaning
  - a.k.a. semantic shifts.

Introduction

TL:DR

- Word meaning ≈ word contexts [Firth, 1957]
- Changes in contexts ≈ changes in meaning
  - a.k.a. semantic shifts.
- Cultural changes influence the contexts
- Studies in automatic tracing of semantic shifts for Russian require publicly available datasets and strong baselines.
- We provide those.

Introduction

Contributions

- Dataset of short-term semantic shifts in Russian adjectives, based on news texts
- Re-packing a dataset of long-term semantic shifts during the Soviet period
- Experimenting with well-established baseline algorithms for semantic shift detection, testing them on the datasets

Contents

1. Introduction
2. Previous work
3. Our datasets
4. Our corpora
5. Word embeddings
6. Baseline results
   - Local methods
   - Global methods
   - Results
7. Conclusion

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]
  - Training models separately for each time bin:
    - Comparing distances between a given word and all others (second-rank similarity) [Yin et al., 2018]
    - Aligning embedding spaces [Hamilton et al., 2016]

Previous work

- Hand-picking examples
  [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]
  - Training models separately for each time bin:
    - Comparing distances between a given word and all others (second-rank similarity) [Yin et al., 2018]
    - Aligning embedding spaces [Hamilton et al., 2016]
  - Training models jointly across time bins
    [Bamler and Mandt, 2017, Yao et al., 2018]

Contents

1. Introduction
2. Previous work
3. Our datasets
4. Our corpora
5. Word embeddings
6. Baseline results
   - Local methods
   - Global methods
   - Results
7. Conclusion

Our datasets

'Micro' dataset

- 2000 — 2014: 15 years of Russian news texts
- 20 adjectives for each year pair (2000-2001, 2001-2002, etc...)
- selected randomly, biased towards the words chosen by the *Global Anchors* method (more details further)
- 14 year pairs $\times$ 20 words = 280 entries
- Manual annotation by 3 annotators
- 3 class labels:

Label	Meaning
0	no semantic shift
1	somewhat shifted
2	significantly shifted

Our datasets

Socio-cultural semantic shifts in adjectives in 2014, as compared to 2013 (excerpts from the 'Micro' dataset)

Class	Adjective	English translation
2	крымский	*'Crimean'*
2	приёмный	*'1) adopted; 2) something receiving'*
2	луганский	*'of Luhansk'*
1	правый	*'1) right; 2) right-wing'*
1	кипрский	*'Cyprian, Cypriot'*
0	серый	*'gray'*
0	балетный	*'of ballet'*

Our datasets

Examples of words labeled as 2 in the 'Micro' dataset

Years	Adjective	English translation
2000—2001	македонский	'*Macedonian*'
2006—2007	несогласный	'*Dissident*'
2008—2009	свиной	'*Swine*'
2009—2010	греческий	'*Greek*'
2011—2012	координационный	'*Coordinating*'

Our datasets

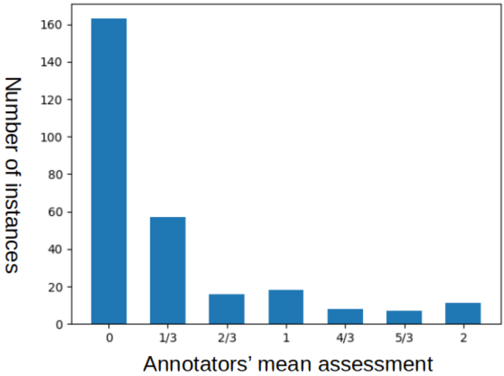


Figure: Mean values of annotators' scores, 'Micro' dataset

Our datasets

'Macro' dataset

- Originally from [Kutuzov and Kuzmenko, 2018]
- We publish it in a machine-readable form.
- Changes from Pre-Soviet through Soviet times
-

	Nouns	Adjectives
Target	38	5
Filler	152	20

- 2 class labels (no shift / shift)

Our datasets

word	label	word	label
отделение	1	тюрьма	0
секция	1	влияние	0
богадельня	1	весна	0
особа	1	уверенность	0
уклон	1	красавица	0
молодец	1	жених	0
передовой	1	заказ	0

Table: Example entries from the 'Macro' dataset

Contents

1. Introduction
2. Previous work
3. Our datasets
4. Our corpora
5. Word embeddings
6. Baseline results
   - Local methods
   - Global methods
   - Results
7. Conclusion

Our corpora

The corpora we used correspond to the datasets.

‘Micro’ dataset

- Newspaper subcorpus of RNC + lenta.ru
  - News texts produced in 2000,
  - News texts produced in 2001,
  - ...,
  - News texts produced in 2014,

‘Macro’ dataset

- Main body of RNC:
  - Texts produced before 1917 (75 millions tokens),
  - Texts produced in 1918—1990 (96 millions tokens),
  - Texts produced after 1991 (85 millions tokens)

Our corpora

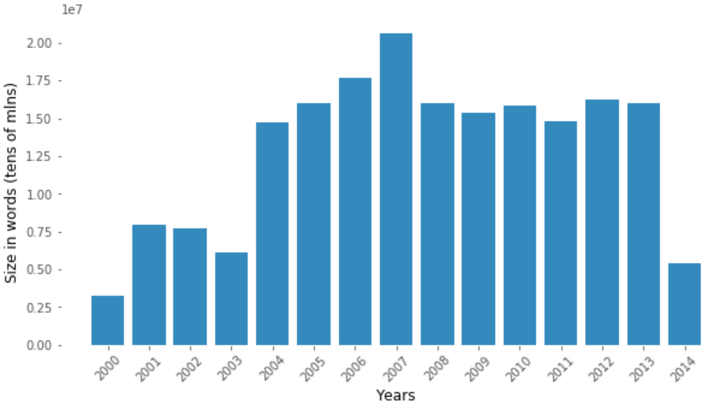


Figure: 'Micro' corpora sizes per year

Contents

1. Introduction
2. Previous work
3. Our datasets
4. Our corpora
5. Word embeddings
6. Baseline results
   - Local methods
   - Global methods
   - Results
7. Conclusion

Word embeddings

Distributional models for baselines evaluation

- 'Static' models:
  - Model trained on time bin tb_0,
  - Model trained on time bin tb_1,
  - ...
  - Model trained on time bin tb_n
- 'Incremental' models
  - Model trained on time bin tb_0,
  - Model trained on time bin tb_1, initialized with tb_0 weights,
  - ...
  - Model trained on time bin tb_n, initialized with tb_{n-1} weights.

CBOW [Mikolov et al., 2013], context window = 5, vector size 300

Contents

1 Introduction

2 Previous work

3 Our datasets

4 Our corpora

5 Word embeddings

6 Baseline results
- Local methods
- Global methods
- Results

7 Conclusion

Baseline results

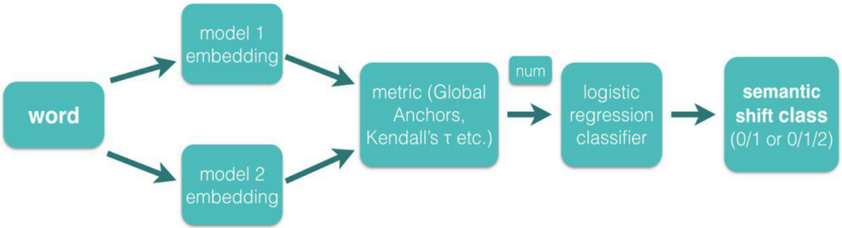


Figure: Experimental workflow

Baseline results

Local methods for semantic shift detection

Comparing words' nearest neigbors:

- Kendall's τ [Kendall, 1948]
- Jaccard distance [Jaccard, 1901]

Global methods for semantic shift detection

Comparing overall structure of semantic spaces:

- Procrustes alignment [Hamilton et al., 2016]
- Global Anchors [Yin et al., 2018]

Baseline results

Kendall's τ

[Kendall, 1948]

$$\frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j)\, sgn(y_i - y_j) \quad (1)$$

Nearest neighbors for 'луганский' ($x = 2013, y = 2014$):

x_1: иркутский → y_1: донецкий

…   …

x_7: донецкий y_{17}: иркутский

Baseline results

Jaccard distance

[Jaccard, 1901]

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (2)$$

Nearest neighbors for ‘вежливый’:
X = приветливый, общительный, уравновешенный, отзывчивый, добродушный
Y = камуфляж, неравнодушный, порядочный, здравомыслящий, незнакомый

Can you guess the years for X and Y?

Baseline results

Global Anchors

[Yin et al., 2018]

Semantic shift of word *w* from year *x* to year *y*:

$$similarities_x = (x_1, ..., x_n)$$

$$similarities_y = (y_1, ..., y_n)$$

- x_i and y_i are cosine similarities between word *w* and i^th word in vocabulary in the years *x* and *y*
- We compare global positions of *w* in the semantic space.
- Semantic similarity between different time periods = $cos(similarities_x, similarities_y)$

Baseline results

Orthogonal Procrustes Analysis

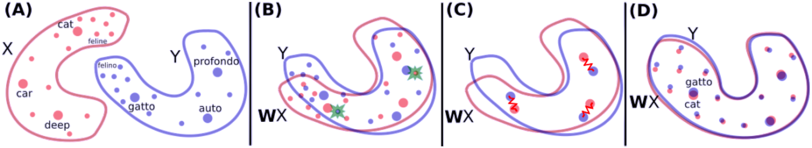
[Hamilton et al., 2016]

Given embedding matrices A and B, find an orthogonal matrix R that maps A to B.

$$B^T A = M$$

$$M = U \Sigma V^T$$

$$R = UV^T$$



Baseline results

Macro F1 scores, ‘Macro’ dataset

Models	Glob.Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.675	**0.767**	0.504	0.646	0.722
Incremental	0.598	0.681	0.475	0.576	0.617
Random choice					
≈ 0.5					

- Global methods work better
- Local methods are still applicable
- Procrustes analysis is clearly the best
- Incremental models are worse than static.

Baseline results

Macro F1 scores, 'Micro' dataset

Models	Glob.Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.453	0.468	0.136	0.301	**0.503**
Incremental	0.462	0.459	0.194	0.326	0.442
Random choice					
≈ 0.33					

- Global methods clearly win on granular timespans
- Local methods sometimes worse than random guessing
- Combining methods is a good idea
- Still no (significant) profit from incremental models

Contents

1. Introduction
2. Previous work
3. Our datasets
4. Our corpora
5. Word embeddings
6. Baseline results
   - Local methods
   - Global methods
   - Results
7. Conclusion

Conclusion

Contributions

- Two manually annotated datasets with diachronic semantic shifts for Russian:
  - A short-term 'Micro' dataset, scale = years
  - A long-term 'Macro' dataset, scale = centuries
- Evaluating 4 algorithms, 2 of which have never been tested on Russian data

Conclusion

Contributions

- Two manually annotated datasets with diachronic semantic shifts for Russian:
  - A short-term 'Micro' dataset, scale = years
  - A long-term 'Macro' dataset, scale = centuries
- Evaluating 4 algorithms, 2 of which have never been tested on Russian data
- Providing solid ground for further studies.

Conclusion

Contributions

- Two manually annotated datasets with diachronic semantic shifts for Russian:
  - A short-term 'Micro' dataset, scale = years
  - A long-term 'Macro' dataset, scale = centuries
- Evaluating 4 algorithms, 2 of which have never been tested on Russian data
- Providing solid ground for further studies.
- Please outperform our baselines! :-)

Conclusion

Contributions

- Two manually annotated datasets with diachronic semantic shifts for Russian:
  - A short-term 'Micro' dataset, scale = years
  - A long-term 'Macro' dataset, scale = centuries
- Evaluating 4 algorithms, 2 of which have never been tested on Russian data
- Providing solid ground for further studies.
- Please outperform our baselines! :-)
- Datasets and code available:

https://github.com/wadimiusz/diachrony_for_russian

Conclusion

Future work

- Looking into word features

Conclusion

Future work

- Looking into word features
- Using models with continuous time variables
  [Rosenfeld and Erk, 2018]

Conclusion

Future work

- Looking into word features
- Using models with continuous time variables [Rosenfeld and Erk, 2018]
- Contextualized word embeddings [Peters et al., 2018]

Conclusion

Future work

- Looking into word features
- Using models with continuous time variables [Rosenfeld and Erk, 2018]
- Contextualized word embeddings [Peters et al., 2018]
- Sense-aware semantic shifts detection

References I

Bamler, R. and Mandt, S. (2017).
Dynamic word embeddings.
In *Proceedings of the International Conference on Machine Learning*, pages 380–389, Sydney, Australia.

Daniel, M. and Dobrushina, N. (2016).
Two centuries in twenty words (in Russian).
NRU HSE.

Firth, J. (1957).
A synopsis of linguistic theory, 1930-1955.
Blackwell.

References II

Hamilton, W., Leskovec, J., and Jurafsky, D. (2016).
Diachronic word embeddings reveal statistical laws of semantic change.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany.

Jaccard, P. (1901).
Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines.
Rouge.

Kendall, M. G. (1948).
Rank correlation methods.
Griffin.

References III

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014).
Temporal analysis of language through neural language models.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 61–65, Baltimore, USA.

Kutuzov, A. and Kuzmenko, E. (2018).
Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes.
In *Quantitative Approaches to the Russian Language*, pages 95–112. Routledge.

References IV

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018).
Diachronic word embeddings and semantic shifts: a survey.
In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
Advances in Neural Information Processing Systems, 26:3111–3119.

References V

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *Proc. of NAACL*.

Rosenfeld, A. and Erk, K. (2018).
Deep neural models of semantic shift.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484, New Orleans, Louisiana, USA.

Tang, X. (2018).
A state-of-the-art of semantic change computation.
Natural Language Engineering, 24(5):649–676.

References VI

Traugott, E. C. and Dasher, R. B. (2001).
Regularity in semantic change.
Cambridge University Press.

Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018).
Dynamic word embeddings for evolving semantic discovery.
In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681, Marina Del Rey, CA, USA.

References VII

Yin, Z., Sachidananda, V., and Prabhakar, B. (2018).
The global anchor method for quantifying linguistic shifts and domain adaptation.
In *Advances in Neural Information Processing Systems*, pages 9433–9444.