Word Sense Induction for Russian: Deep Study and Comparison with Dictionaries

Konstantin Lopukhin, **Anastasiya Lopukhina**, Boris Iomdin

Dialogue 2017, Moscow.

What are word senses?

Single sense view: core representation + variations

Separate sense view: a list of senses

→ lexeme — a word in one of its lexical meanings (Moscow Semantic School)

Аудитория in two dictionaries

1. a large room in some institution: ***поточная аудитория, учебная аудитория***
2. people in this room: ***молодежная аудитория, завоевать аудиторию***
3. a large group of people for which some information is prepared: ***читательская аудитория, расширение аудитории***
4. A great number of consumers with some particular characteristics: ***целевая аудитория***

1. a large room in some institution: ***читать спецкурс в аудитории***
2. people in this room: ***ответить на вопросы аудитории***

   / a group of readers or spectators: ***завоевать симпатии широкой читательской аудитории***

WSI method

Input: raw text without annotations.

Output:

- Discovers senses and describes them.
- Groups contexts into senses.

Practical applications of WSI

- Part of NLP pipeline
- A tool for lexicographers

Works for:

- new words, new senses
- domain-specific senses
- resource-constrained languages

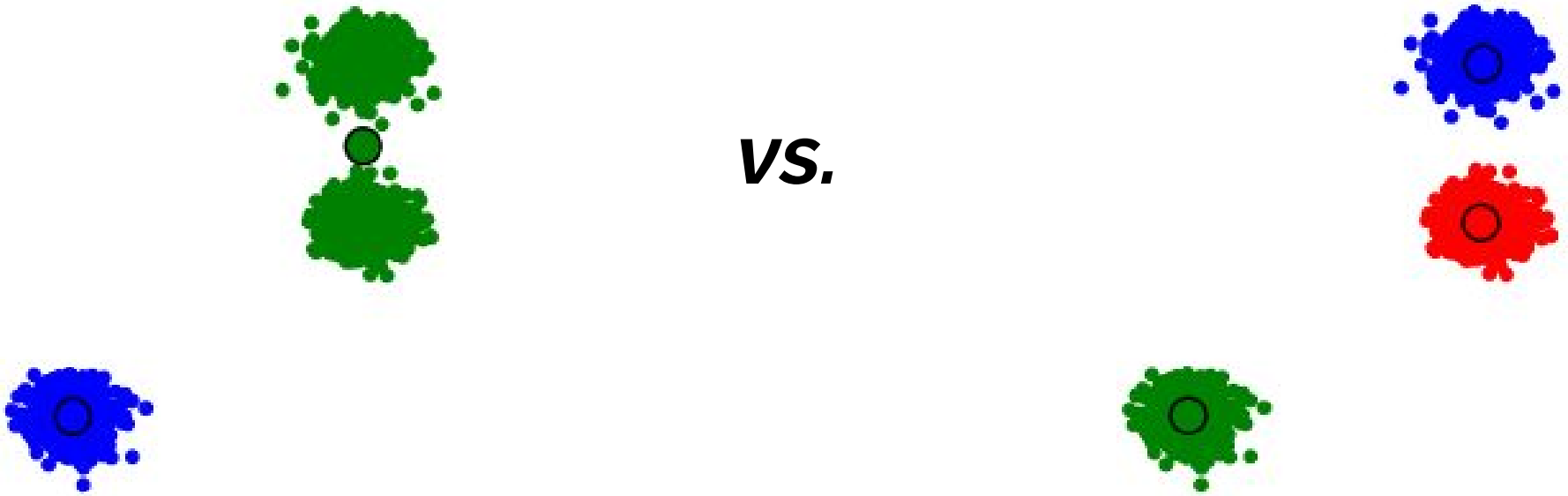
WSI evaluation

Qualitative: compare WSI sense descriptions with dictionaries and intuition.

Quantitative: compare WSI method clustering with “gold standard” clustering **for contexts**.

Quantitative WSI evaluation

How similar are two clusterings?



V-measure

Harmonic mean of **homogeneity** and **completeness**.

Homogeneity: one cluster has contexts of only one sense.
Completeness: all contexts of one sense lie in one cluster.

V-measure favors large number of clusters.

Adjusted Rand Index (ARI)

Rand Index:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a = # pairs that are in one cluster both in *X* and in *Y*

b = # pairs that are in different cluster both in *X* and in *Y*

Adjusted Rand Index is adjusted for randomness:

ARI of a random clustering is 0.

Evaluated WSI methods

- LDA
- Word2Vec neighbours
- Context clustering
- AdaGram

Latent Dirichlet allocation (LDA)

Topic modelling: discovering topics in documents.

Topic ⇒ word sense.

Document ⇒ word context.

Topic modelling doesn't work great with short documents.

Word2Vec neighbours

Approach: take top 100 most similar words, cluster with k-means, merge close clusters.

Assume:

- word vector will capture properties of all senses
- each sense has several monosemous neighbours

Context clustering

Vector representation of context that captures its meaning: PMI-weighted average of Word2Vec word vectors.

Cluster vectors of a large number of contexts of one word: spherical k-means, merge close clusters.

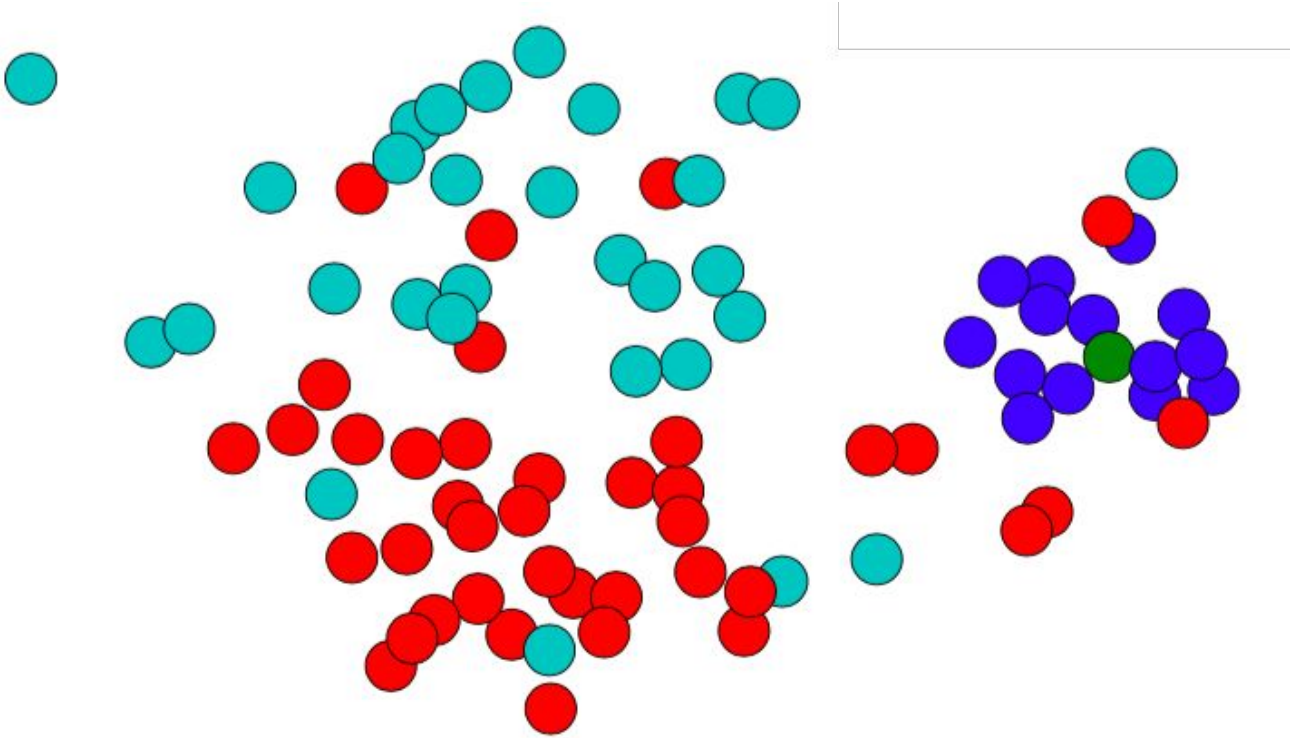
Context representation

Bag of words, large window: ± 10 words.

Some words are more important (depending on the target).

Word2Vec captures word usage from the whole corpus.

Context representation



AdaGram (Bartunov et al. 2015)

Non-parametric extension of Word2Vec skip-gram.

A vector for each sense of a word.

Different words have different number of senses.

Like in Word2Vec, sense vectors are built for all words in the corpus at once.

AdaGram (Bartunov et al. 2015)

Word2Vec skip-gram:

$$p(v|w,\theta) = \frac{\exp(in_w^T out_v)}{\sum_{v'=1}^{V} \exp(in_w^T out_{v'})}$$

AdaGram — a vector per sense (only for central word):

$$p(v|z=k,w,\theta) = \frac{\exp(in_{wk}^T out_v)}{\sum_{v'=1}^{V} \exp(in_{wk}^T out_{v'})}$$

AdaGram vs. Context clustering

Very similar context representation.

AdaGram:

- more principled induction method
- more computationally efficient.

Context clustering:

- easier to implement and tweak.

Quantitative evaluation: setup

8 polysemous nouns, 10 polysemous verbs.

100 - 500 annotated contexts from RuTenTen11 and RNC.

Senses from the Active Dictionary of Russian.

Measure V-measure and ARI.

Quantitative evaluation: results

	Nouns	Verbs	average
LDA	0.12	0.02	0.07
Context clustering	**0.34**	**0.14**	**0.24**
AdaGram	0.25	**0.13**	0.18

Adjusted Rand Index for the word sense induction task

Qualitative evaluation: setup

15 nouns:

- 7 polysemous with 3 to 9 AD senses
- 8 monosemous, 5 have new/slang senses (e.g. *баян*)

Divide induced senses into groups:

- quality senses (good)
- duplicates (bad)
- hard to interpret: unclear or mixed (very bad)

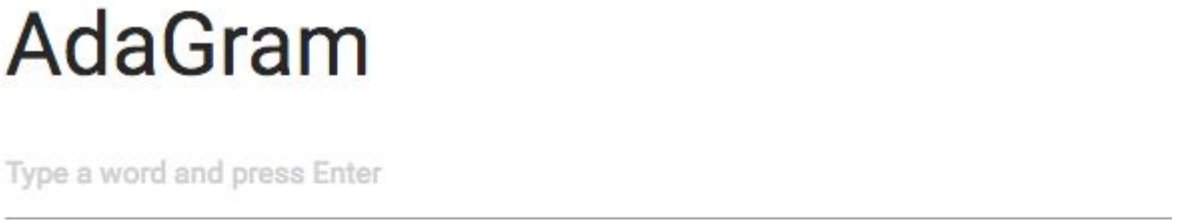
Qualitative evaluation: results

	Quality senses	Duplicate senses	Hard to interpret
Word2vec neighbours	2.4	1.1	**0.5**
Context clustering	2.8	**1.0**	0.9
LDA	1.8	2.1	1.3
AdaGram	**3.6**	3.7	2.5

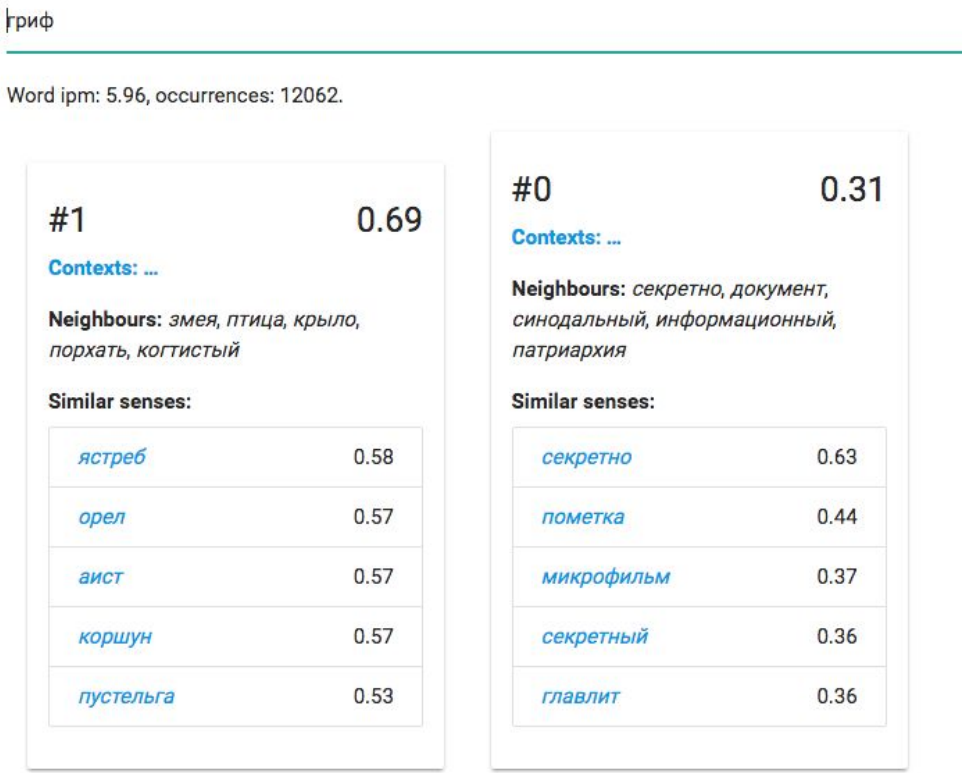
Average number of quality, duplicate and unclear senses.

adagram.ll-cl.org

Libraries: training with AdaGram.jl, inference with python-adagram



Qualitative evaluation for AdaGram



AdaGram vs. Russian Dictionaries

	Apresjan, 2014	Kuznetsov, 2014	Evgenyeva, 1981-1984	Shvedova, 2007	Average
adjectives	0.44	0.72	0.68	0.66	0.62
nouns	0.50	0.70	0.72	0.74	0.69
verbs	0.35	0.61	0.68	0.71	0.61
Average	0.43	0.68	0.70	0.71	0.64

Average number of senses discovered by AdaGram (recall).

Different “senses” in AdaGram

горшок

Word ipm: 16.89, occurrences: 34188.

#0 0.33	#2 0.29	#1 0.26	#4 0.10	#3 0.03
Contexts: …	Contexts: …	Contexts: …	Contexts: …	Contexts: …
Neighbours: *цветочный, цветок, растение, грунт, клумба*	**Neighbours:** *приучать, бог, отучать, ребенок, ходить*	**Neighbours:** *глиняный, ночной, щи, звон, котел*	**Neighbours:** *переработка, деформация, межкомнатный, театрализовать, эллинский*	**Neighbours:** *адмирал, флот, крейсер, полузащитник, вмф*
Similar senses:	**Similar senses:**	**Similar senses:**	**Similar senses:**	**Similar senses:**
вазон 0.77	*памперс* 0.52	*миска* 0.75	*сосуд* 0.60	*горшков* 0.67
кадка 0.66	*садик* 0.49	*кастрюля* 0.74	*пифос* 0.58	*головко* 0.64
ваза 0.66	*ребенок* 0.49	*котел* 0.71	*амфора* 0.55	*кузнецов* 0.61
кашпо 0.65	*ребеночек* 0.49	*глиняный* 0.70	*кувшин* 0.54	*касатонов* 0.60
клумба 0.63	*доча* 0.49	*плошка* 0.69	*лепной* 0.52	*макаров* 0.58

Homonyms


топить
Word ipm: 6.38, occurrences: 12913.
#0
0.56
Contexts: ...
Neighbours: печь, печка, камин, дрова, тулуп
Similar senses:
топиться 0.79
протапливать 0.73
истапливать 0.68
печь 0.64
печка 0.63
#1
0.44
Contexts: ...
Neighbours: корабль, противник, корвет, карфагенянин, крушение
Similar senses:
утапливать 0.66
тонуть 0.56
потоплять 0.55
потоплять 0.54
утопать 0.47

Metaphorical senses

хрупкий

Word ipm: 16.29, occurrences: 32961.

#0 0.34

Contexts: ...

Neighbours: *телосложение, девушка, молодой, женщина, восемнадцатилетний*

Similar senses:

худенький	0.73
миловидный	0.72
пухленький	0.66
субтильный	0.65
хорошенький	0.65

#1 0.27

Contexts: ...

Neighbours: *плечо, стебель, цыпочки, повалиться, спина*

Similar senses:

невесомый	0.63
гибкий	0.62
худенький	0.61
тельце	0.59
обнимать	0.57

#3 0.24

Contexts: ...

Neighbours: *равновесие, стабильность, единство, мир, баланс*

Similar senses:

непрочный	0.56
прочный	0.52
неустойчивый	0.51
зыбкий	0.48
нарушать	0.47

#2 0.15

Contexts: ...

Neighbours: *материал, покрытие, сварка, обработка, оксид*

Similar senses:

прочный	0.70
непрочный	0.69
пластичный	0.63
эластичный	0.61
мягкий	0.61

Metonymic senses

аудитория

Word ipm: 35.99, occurrences: 72836.

#0 0.44

Contexts: ...

Neighbours: *целевой, выбор, пользователь, маркетинговый, спрос*

Similar senses:

целевой	0.64
рекламодатель	0.56
читательский	0.55
потребитель	0.54
охват	0.54

#2 0.33

Contexts: ...

Neighbours: *перед, пред, напыщенный, концертный, передо*

Similar senses:

публика	0.57
зал	0.56
слушатель	0.55
слушатель	0.54
зритель	0.53

#3 0.14

Contexts: ...

Neighbours: *корпус, учебный, млн, 2013, м²*

Similar senses:

лекционный	0.51
лекторий	0.49
кинозал	0.48
аудиторный	0.48
конференц-зал	0.47

#4 0.10

Contexts: ...

Neighbours: *зашептать, губа, нагибаться, уставляться, обхватить*

Similar senses:

трапезная	0.59
кабинка	0.58
гостиная	0.58
дежурка	0.58
ризница	0.58

Novel senses

горячий

#1 0.10

Contexts: ...

Neighbours: *линия, точка, след, телефон, консультация*

Similar senses:

495	0.34
оперативный	0.33
след	0.33
пунктирный	0.33
невозврат	0.33

винт

#2 0.36

Contexts: ...

Neighbours: *одноклассник, фига, инет, зафрендить, файл*

Similar senses:

комп	0.64
винд	0.62
флешка	0.61
материнка	0.60
процессор	0.59

вышка

#2 0.27

Contexts: ...

Neighbours: *матч, чемпионат, четвертьфинал, сош, выигрывать*

Similar senses:

премьерка	0.48
премьер-лига	0.43
лч	0.41
еврокубок	0.40
плей-офф	0.40

#1 0.70

Contexts: ...

Neighbours: *блог, сайт, статья, раскрутка, видео*

Similar senses:

постить	0.66
запощать	0.64
фотка	0.60
запостить	0.60
опубликовывать	0.60

Less “senses” than in dictionaries

резать

Word ipm: 20.76, occurrences: 42017.

#2	0.51

Contexts: ...

Neighbours: *слух, национальность, геноцид, нато, прилюдно*

Similar senses:

колоть	0.53
бить	0.51
зарезать	0.51
правда-матка	0.51
травить	0.49

#1	0.25

Contexts: ...

Neighbours: *кромка, ветер, окутывать, мелькание, дуновение*

Similar senses:

кромсать	0.63
полосовать	0.61
взрезывать	0.60
режущий	0.59
полоснуть	0.58

#0	0.24

Contexts: ...

Neighbours: *обжаривать, огурец, соломка, шпинат, овощ*

Similar senses:

нарезать	0.81
порезать	0.77
разрезать	0.74
кружочек	0.74
соломка	0.72

Limitations

AdaGram does not distinguish senses which differ in argument structure (usually with causative component):

Парикмахер бреет клиента / Я брею голову у одного и того же мастера

Я погасил костер водой / Дождь погасил костер

Usages? Senses?

ВОСК

Word ipm: 7.15, occurrences: 14468.

#1 0.49

Contexts: ...

Neighbours: *озарять, свеча, скользнуть, невидящий, свет*

Similar senses:

оплывать	0.64
восковой	0.62
стеарин	0.61
смола	0.59
олово	0.58

#2 0.21

Contexts: ...

Neighbours: *эпиляция, бикини, техник, купальник, сооружение*

Similar senses:

паста	0.65
лак	0.64
наноситься	0.64
акрил	0.63
клей	0.63

#0 0.18

Contexts: ...

Neighbours: *пчелиный, парафин, мг, маточный, г*

Similar senses:

ланолин	0.77
прополис	0.73
канифоль	0.72
парафин	0.71
скипидар	0.70

#3 0.12

Contexts: ...

Neighbours: *северский, чернигов, скот, черкасский, нелегальный*

Similar senses:

пенька	0.61
ворвань	0.57
мех	0.53
мед	0.53
пушнина	0.52

курица	2	0

GO

курица: 2 ↔ 0

2 ↔ 0 similarity: 0.33
Word ipm: 24.72, occurrences: 50023.

#2 0.46

гурьба, сквозь, охать, эдан, нашептывать

цыпленок	0.75
петух	0.69
индюк	0.69
индюшка	0.69
курочка	0.68

#0 0.26

салат, отварной, жареный, свинина, похудеть

мясо	0.91
жареный	0.90
отварной	0.87
тушеный	0.87
цыпленок	0.86

Similar pairs

Word	Sense 1	Sense 2	Closeness
цыпленок	1	0	0.75
поросенок	0	1	0.59
яйцо	3	0	0.59
яйцо	3	4	0.59
птица	4	0	0.58
цыпленок	2	0	0.56
утка	1	0	0.56
курочка	1	0	0.54

яйцо

Word ipm: 58.70, occurrences: 118783.

#3	0.28

бластер, пульсирующий, фаррелл, пакс, файр

клюв	0.61
яичко	0.60
курица	0.59
цыпленок	0.58
дракон	0.51

#0	0.22

1, взбивать, мука, масло, г

сметана	0.88
желток	0.85
сахар	0.84
молоко	0.83
творог	0.83

adagram.ll-cl.org

Libraries: training with AdaGram.jl, inference with python-adagram