Evaluating three corpus-based semantic similarity systems for Russian

Arefyev N. V. (narefjev@cs.msu.su) Lomonosov Moscow State University & Digital Society Laboratory, Moscow, Russia

Panchenko A. I. (panchenko@lt.informatik.tu-darmstadt.de), TU Darmstadt, Darmstadt, Germany

Lukanin A. V. (artyom.lukanin@gmail.com), LLC “SoftPlus”, Chelyabinsk, Russia

Lesota O. O. (cheesemaid@gmail.com), Lomonosov Moscow State University, Moscow, Russia

Romanov P. V. (romanov4400@gmail.com), 1C Company, Moscow, Russia

Systems

Manually created lexico-syntactic patterns: **PatternSim**

Panchenko, A., Morozova, O., Naets, H. (2012). A Semantic Similarity Measure Based on Lexico-Syntactic Patterns

Sabirova, K., Lukanin, A. (2014). Automatic Extraction of Hypernyms and Hyponnyms from Russian Texts

Word cooccurence counts: **GNG**

Bullinaria, J., Levy, J. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study

Neural network: **Word2Vec**

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space

Corpora

Name	Description	Tokens	Documents	Size, Gb
wiki	Russian Wikipedia	238,052,379	1,159,723	3
web	Russian Web Pages	567,914,057	890,551	7
librusec	Lib.rus.ec book collection	12,902,854,351	233,876	149
Google Ngrams	Russian Google N-Grams	67,137,666,353* *raw corpus	591,310* *raw corpus	---

PatternSim: **wiki** (lemmatized)**, web**

GNG: **Google Ngrams**

Word2Vec: **wiki** (lemmatized/non-lemmatized), **librusec**

PatternSim — patterns

Patterns — extract semantically similar words from corpus

1. такие/таких/таким NP, как NP[, NP] и/или NP — such NP as NP[, NP] and/or NP

В России распространены такие {[овощи]=HYPER} как {[морковь]=HYPO}, {[помидоры]=HYPO}, {[капуста]=HYPO} и {[лук]=HYPO}

Such {[vegetables]=HYPER} as {[carrot]=HYPO}, {[tomato]=HYPO}, {[cabbage]=HYPO}, and {[onion]=HYPO} are popular in Russia

=> vegetable, carrot, tomato, cabbage, onion

2. NP, такие/таких/таким как NP[, NP] и/или NP — NP such as NP, NP[, NP] and/or NP

…{систем [верований]=HYPER}, таких как {[шаманизм]=HYPO},{[политеизм]=HYPO},{[пантеизм]=HYPO}, {[анимизм]=HYPO}

3. NP: NP[, NP] и/или NP — NP: NP, [, NP] and/or NP

…мир, передаваемый человеку через {его [ощущения]=HYPER}: {[зрение]=HYPO}, {[слух]=HYPO}, {[обоняние]=HYPO}, {[осязание]=HYPO} и другие

PatternSim — patterns

4. NP[, NP][, а также/также как [и]/и/или] другие/другим/других/о других NP

…использованием зараженных вирусом {[шприцев]=HYPO}, {[игл]=HYPO} и {других {медицинских и парамедицинских [инструментов]=HYPER}}

5. NP [и по сей день/в <A> время] это/есть/является/был [самый] <A> NP

{Рентгеноструктурный [анализ]=HYPO} и по сей день является самым {распространенным [методом]=HYPER} определения структуры вещества…

6. NP, включая/в том числе [и]/включительно, NP[,NP] и/или NP

…сосудистую систему {[позвоночных]=HYPER} животных, в том числе {[человека]=HYPO} и некоторых беспозвоночных

7. NP, [а] особенно/в особенности/особо, NP[, NP] и/или NP

L-Аргинин входит в {состав пептидов и [белков]=HYPER}, особенно {высоко содержание [аргинина]=HYPO}…

8. NP, как например/в частности/например/кпримеру, NP[, NP] и/или NP

Это является одной из основных проблем при проектировании устройств {цифровой [электроники]=HYPER}, в частности, {цифровых [фотоаппаратов]=HYPO}…

9. Виды/типы/формы/разновидности/сорта NP, как NP[, NP] и/или NP

Такие виды {[оружия]=HYPER} как {[шпага]=HYPO} и {[рапира]=HYPO} тоже причисляют к мечам, что не

совсем верно

10. NP - вид/тип/форма/разновидность/сорт NP

{[Хобби]=HYPO} — вид {человеческой [деятельности]}, некое занятие, увлечение…

PatternSim — algorithm

Input: Terms C, Corpus D

Output: Similarity matrix S [C × C]

1. $K_C \leftarrow$ from D extract sentences matching patterns

2. $e_{ij} \leftarrow$ number of sentences in K_C from which both c_i and c_j were extracted

3. $s_{ij} \leftarrow f(e_{ij})$

$$s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$$

p_{ij} — the number of distinct patterns which extracted the pair (c_i, c_j)

b_{*j} / b_{i*} — the number of distinct words related to c_i / c_j

$P(c_i)$ / $P(c_j)$ — probability of c_i / c_j estimated on the corpus

4. Normalize S

GNG — algorithm

Highly dimensional (~1M) sparse vectors

PPMI (positive pointwise mutual information):

$$a_i = \max(0, \log(\frac{P(w, cw_i)}{P(w)P(cw_i)}))$$

Sim(w_i, w_j) = cos(vector(w_i), vector(w_j))

Bullinaria, J., Levy, J. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study

GNG — data

Google Ngrams — Ngrams for n=1..5

- collected from Google Books Corpus — for Russian: 67B tokens / 590'000 volumes
- Librusec: **x5**, Russian Wikipedia: **x280**

Only ngrams which occured in at least **40 volumes**!

<N-gram> <Year> <Match_cnt> <Volume_cnt>

...

каша оказалась_VERB	1964	1	1
каша_NOUN между_ADP	1901	1	1
кашицы _ADJ_	1839	1	1
каше продукты	1872	3	3
каша перловая_VERB	2000	1	1

Context of w — words to the right extracted from 5-grams starting with w

(right context window of size 5)

GNG — results

аборигенкалькулятор	вдв
корнями вросший в	тарификатор расчетчик	генерал и по воздушно полковника десантные в полковник

Hi Oleg,

Yes, I would believe that the frequency thresholding on 3-grams, 4-grams, and 5-grams would result in many fewer co-occurrences happening of two words than actually occur in books.

Unfortunately we ***cannot release the full list of Ngrams****.*

Good luck,

Yuri

Word2Vec — algorithm

Skip-gram + negative sampling

the probability of word *w* to appear in context *c*

$$P(D = 1|w, c; \theta) = (1 + e^{-V_c W_w})^{-1}$$

optimizing parameters (word vectors *W* and context vectors *V)* to maximize probability of real data and minimize probability of random data

$$\theta^* = arg\ max \prod_{(c,w)\in corp} P(D = 1|w, c; \theta) \prod_{(c,w)\in rand} (1 - P(D = 1|w, c; \theta))$$

Word2Vec — corpus

Lib.rus.ec — part in FB2 format

- Select books in Russian (lang=ru)
- Convert to plaintext
- Preprocessing
  - convert ё to e
  - insert spaces around punctuation marks
  - remove digits and special characters

Word2Vec — similarity

$$Sim(w_i, w_j) = cos(vector(w_i), vector(w_j))$$

For out-of-vocabulary words:

- split by dash

sim(актриса, актер-статист) = sim(актриса, [актер, статист]) = sim(актриса, актер) = 0.75

sim(actress, dummy-actor) = sim(actress, [dummy, actor]) = sim(actress, actor) = 0.75

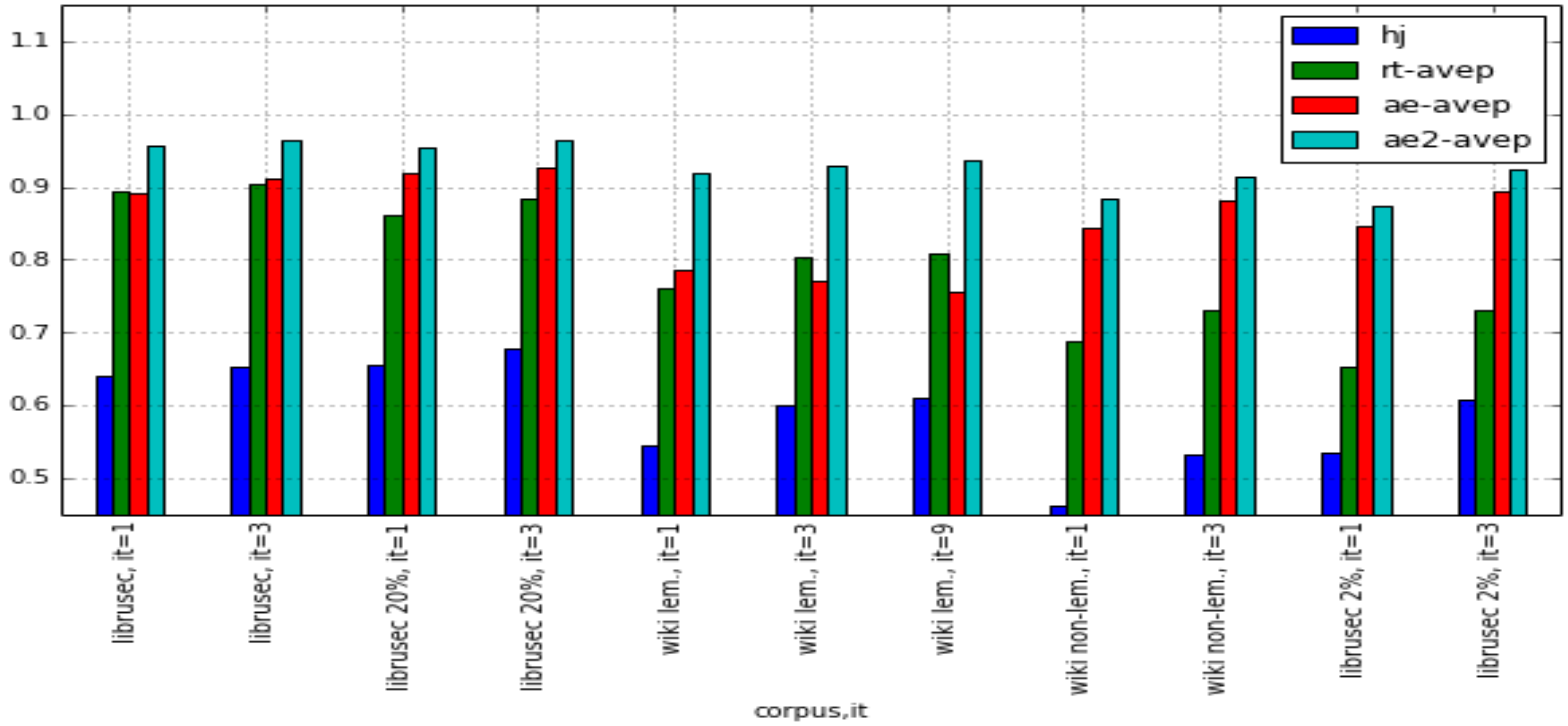
- remove prefixes

sim(автотехника, автомототехника) = sim(автотехника, [мототехника, техника]) = sim(автотехника, мототехника) = 0.64

sim(auto-vehicles, auto-motor-vehicles) = sim(auto-vehicles, [motor-vehicles, vehicles]) = sim(auto-vehicles, motor-vehicles) = 0.64

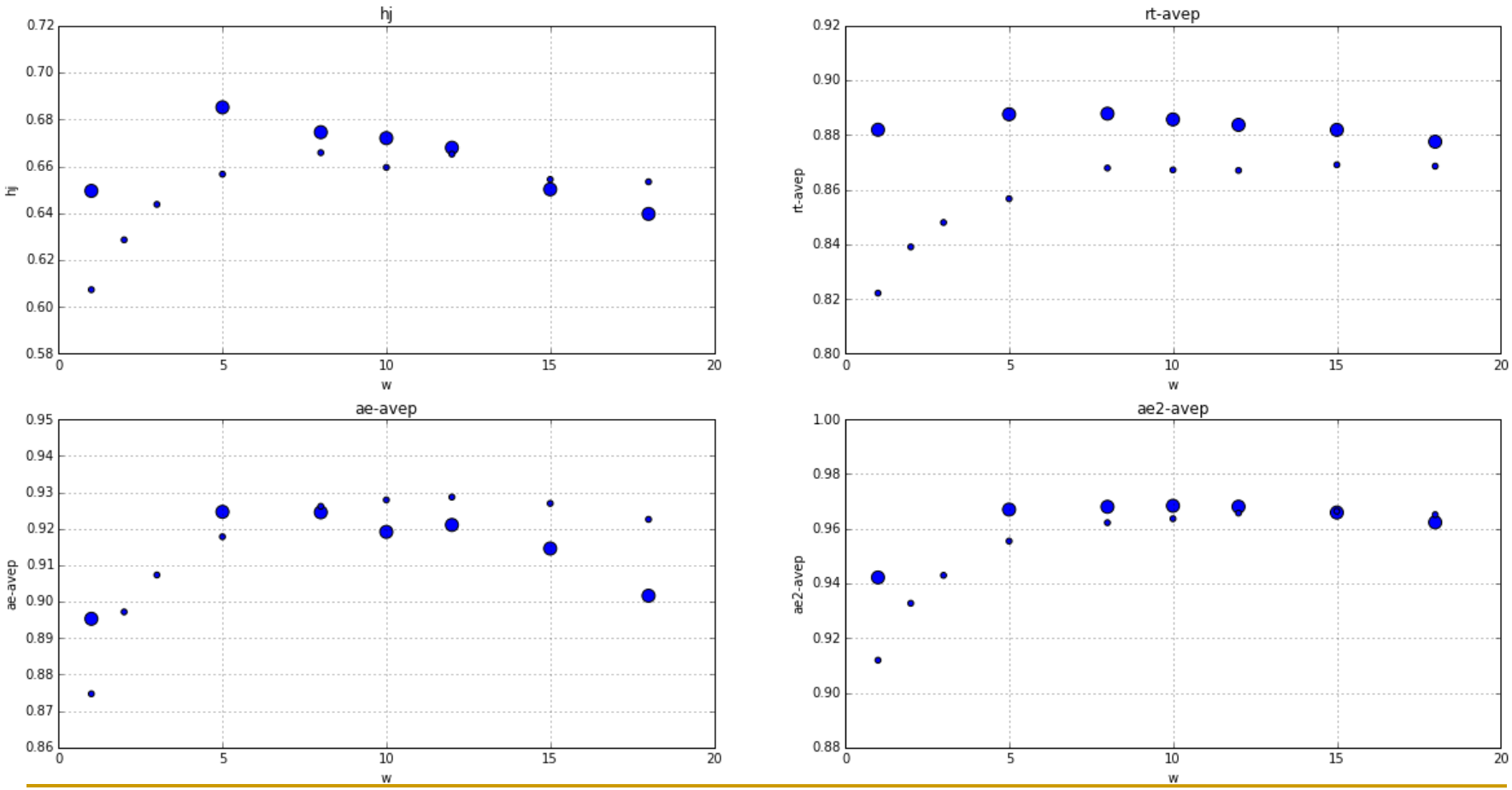
Word2Vec — metaparameters

the corpus and the number of iterations (sz500-w5)



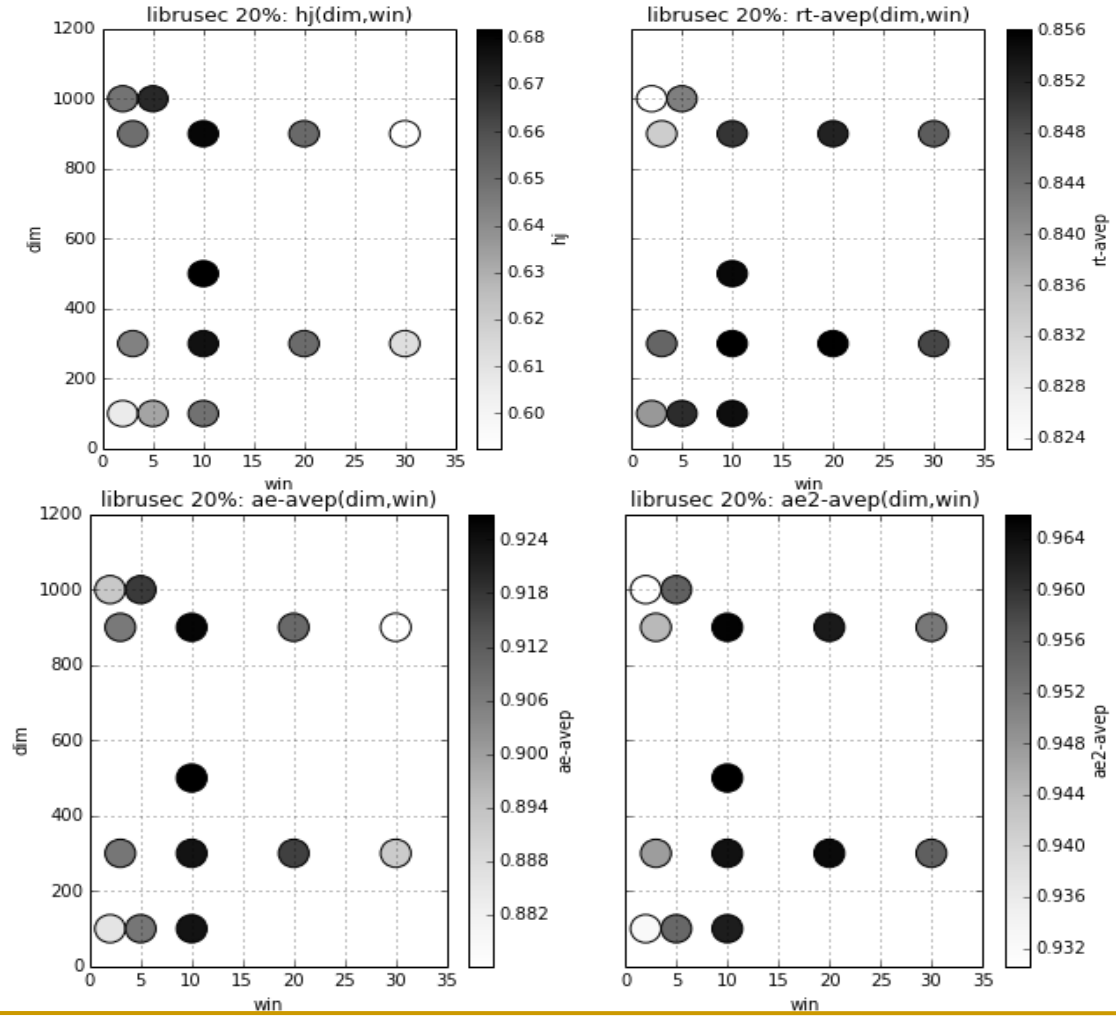
Word2Vec - metaparameters

the window size and the number of iterations (librusec 20%, sz500)



Word2Vec — metaparameters

the window size and the vector size (librusec 20%, it1)



RUSSE results

Method	Corpus	HJ max 0.763	RT max 0.959	AE max 0.956	AE2 max 0.985
patternsim	web+wiki	0.372	0.754	0.708	0.797
patternsim	wiki	0.322	0.755	0.724	0.784
patternsim	web	0.322	0.745	0.696	0.775
skipgram-dim100-win10-iter1	lib	0.621	0.847	0.912	**0.967**
skipgram-dim500-win5-iter3	lib	0.654 (6th place)	**0.903** (6th place)	0.912 (4th place)	**0.965** (5th place)
skipgram-dim500-win5-iter3	*Wiki nonlemm..*	*0.532*	*0.731*	*0.881*	*0.914*
skipgram-dim500-win5-iter3	*Wiki lemm.*	*0.601*	*0.803*	*0.771*	*0.928*
skipgram-sim500-win10-iter3	***lib***	***0.674***	***0.903***	***0.925***	***0.972***
skipgram-sim500-win10-iter3 + oov	***lib***	***0.699***	***0.918***	***0.928***	***0.975***
right-context-window	ngram	0.303	0.612	0.734	0.676

Error analysis

Error type	%	Example
1. out-of-vocabulary words	14.4 %	автомотосредство
2. words with few occurences	20.3 %	перестрахование
3. non-common form	5.3 %	переселенка
4. too abstract hypernym	23.5 %	бытность
5. unknown error	34.2 %	?

Error analysis

бытность	ожидание	горесть	эпизод	европеец	литовка
бытность	замыкание	горесть	переплет	европеец	молдаванка
бытность	лежание	горесть	крест	европеец	македонка
бытность	скапливание	горесть	происшествие	европеец	словенка
бытность	накапливание	горесть	катаклизм	европеец	исландка
бытность	нарастание	горесть	испытание	европеец	карелка
бытность	позирование			европеец	полька
бытность	возлежание				
бытность	произрастание				
бытность	сидение				
бытность	висение				
бытность	отстояние				
бытность	просиживание				
бытность	доминирование				
бытность	времяпровождение				
бытность	стояние				
бытность	расположение				
бытность	увенчание				

Thank you

Questions?