AUTOMATIC COLLOCATION EXTRACTION: ASSOCIATION MEASURES EVALUATION AND INTEGRATION

Zakharov V.P.
v.zakharov@spbu.ru
Saint-Petersburg State University,
Saint-Petersburg, Russia

Collocations and Collocability

- Set phrases, idioms, multiword expressions, collocations…
- "*A collocation is a word combination whose semantic and syntactic properties can't be fully predicted on the basis of information about its constituents and which therefore should be added to the dictionary (lexicon)*" [S.Evert 2004: 17].
- But there are many set phrases whose meaning is equal to the sum of the meanings of their constituents, despite the fact that such phrases function as a single unit
- Idiomatic nature vs. stability
- A probabilistic nature of collocations
- Collocations as statistically determined set phrases. In this case, not only phrasemes and idioms, but also multiword terms, named entities (real-world objects, such as persons, locations, organisations, products, etc.,) and other types of free combinations could be regarded as set phrases.

Association measures

- **P. Pecina**:
  82 measures [Pecina P. Lexical Association Measures. Collocation Extraction, Prague. 2009. Pp. 44-45, 48]
- See also dissertation of **S. Evert** [2004].
- More popular are *T-score*, *MI*, *log-likelihood*

$$MI = \log_2 \frac{f(n,c) * N}{f(n) * f(c)}$$

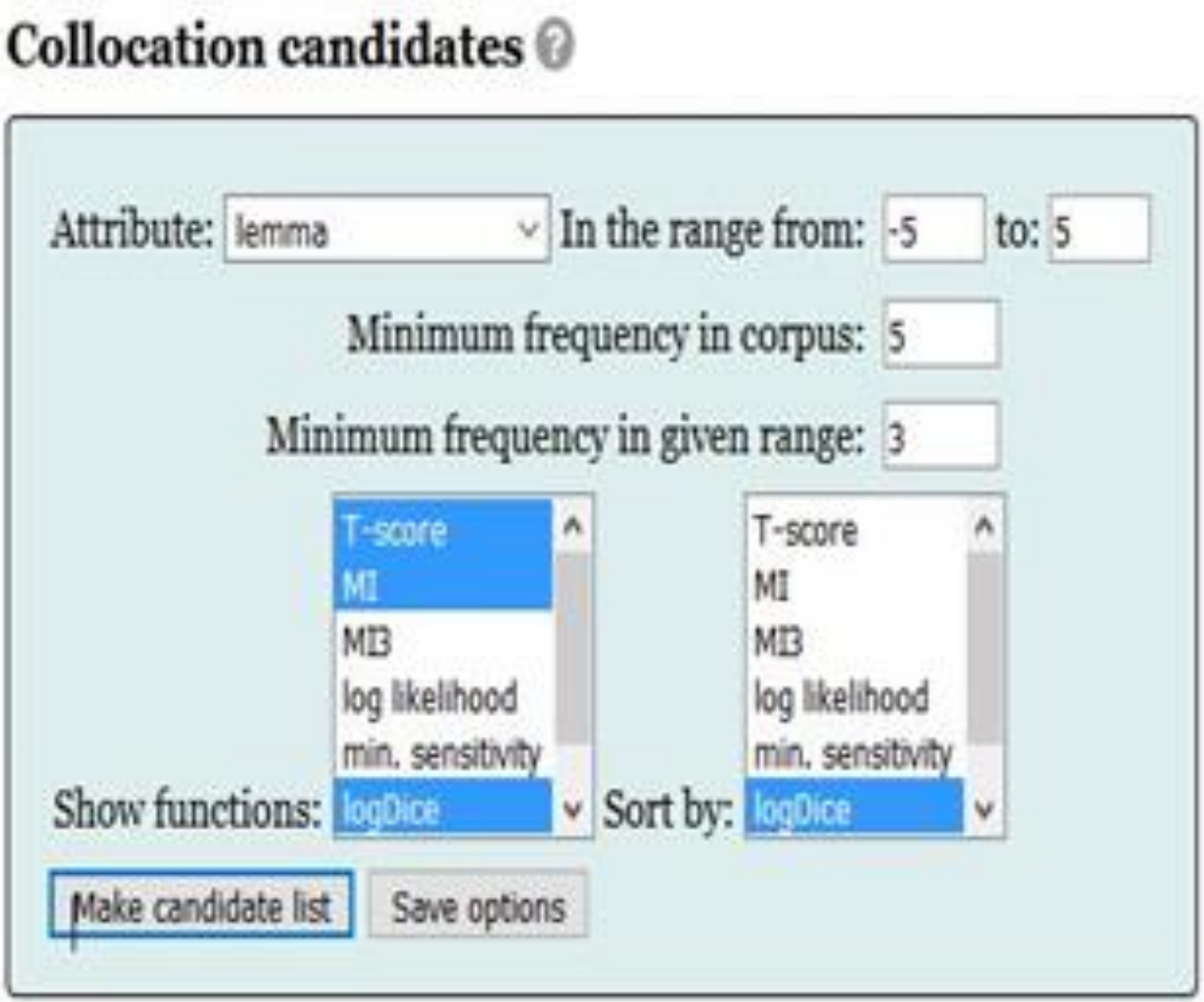
Objectives

- Evaluate association measures functionality
- Integrate collocation lists obtained by different association measures
- Suggest new ranking coefficients for collocations in combined lists
- Compare evaluation techniques

Material and tools

- Araneum corpora of Russian (http://unesco.uniba.sk)
- We used 2 corpora, Russicum Russicum Minus (120 mln. tokens), Russicum Russicum Maius (1,20 bln.)
- Access through the NoSketch Engine
- *Collocations* tool using 7 association measures: *T-score*, *MI*, *MI3*, *log-likelihood*, *minimum sensitivity*, *logDice* and *MI.log_f*

Collocations tool interface



Output of NoSketch Engine Collocation tool

Collocation candidates

Page 1 Go Next >

	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N горячий	25,719	102,240	160.014	8.809	10.086
P | N вода	38,005	672,991	193.013	6.654	9.854
P | N холодный	20,521	102,915	142.849	8.474	9.759
P | N теплый	17,807	136,681	132.868	7.860	9.493
P | N чистый	17,979	149,708	133.460	7.742	9.484
P | N очистка	15,994	72,028	126.148	8.629	9.458
P | N литр	14,550	63,939	120.326	8.664	9.338
P | N сточный	13,481	13,791	116.041	10.767	9.329
P | N питьевой	13,381	14,006	115.608	10.734	9.318
P | N температура	17,888	248,530	132.704	7.004	9.313
P | N минеральный	11,955	45,044	109.108	8.886	9.092
P | N стакан	11,157	42,665	105.400	8.865	8.997
P | N пить	10,524	76,555	102.168	7.937	8.846
P | N грунтовый	9,469	11,598	97.242	10.508	8.824

Table 1

Collocates	Co-occurrence count	Candidate count	MI.log_f score
Сточный (sewer)	12479	13791	100,505
Питьевой (drinkable)	11288	14006	97,878
Грунтовый (ground)	8672	11598	94,132
Кипяченый (boiled)	3635	4502	86,016
Горячий (hot)	20665	102240	84,393
Минеральный (mineral)	9409	45044	78,146
Холодный (cold)	15172	102915	77,386
Талый (melt)	1863	2701	77,295

Table 2: Correlation

Measure	Spearman coefficient
T-score	0.9999
MI	- 0.5620
MI3	0.5469
log likelihood	0.9775
min. sensitivity	0.6158
logDice	0.5326
MI.log_f	0.4932

Association measures functionality

Ranks	T-score	MI	MI3	log likeli-hood	min. sensi-tivity	log-Dice	MI.l-og_f
1-10	0	5	4	2	5	6	8
11-20	4	3	4	4	2	2	3
21-30	2	2	3	3	4	1	4
31-40	1	7	4	3	0	4	6
41-50	1	1	3	2	2	2	2
51-60	2	5	2	4	1	2	2
61-70	0	5	1	4	0	2	1
71-80	3	4	7	0	2	1	3
81-90	1	4	4	2	2	2	1
91-100	3	3	1	2	3	0	1
Всего	**17**	**39**	**33**	**26**	**21**	**22**	**31**

Evaluation: methodology

Ashmanov I., Grigoryev S., Gusev V., Kharin N., Shabanov V. (1997), Using Statistical Method for Intelligent Computer-Based Text Processing [Primenenie statisticheskih metodov dlja intellektual'noj komp'juternoj obrabotki tekstov] / The Proceedings of the Dialog’97 International Seminar on Computational Linguistics and Its Applications, pp. 33–37.

Ranks	T-score	MI	MI3	log likeli-hood	min. sensi-tivity	log-Dice	MI.l-og_f
1-10	0	5	4	2	5	6	8
1-30	6	10	11	9	11	9	15
1-50	8	18	18	14	13	15	23
1-70	10	28	21	22	14	19	26
1-100	17	39	33	26	21	22	31

Evaluation: methodology (2)

A weight of each element of the characteristic set (5, 4, 3, 2, and 1, respectively).
Each element is “weighed”: each of 5 precision values is multiplied by the its weight and divided by 15 (the sum of the weights).

Here is an example for the *MI* measure that has 5 true collocates in the top ten candidates (precision is 0.5), 10 true collocates in the top thirty (precision is 0.33), 18 in the top fifty (0.36), 28 in the top seventy (70), and 39 in the top hundred (0.39).

Then, the resulting precision will be equal to:

0,5*5/15 + 0,33*4/15 + 0,36*3/15 + 0,4*2/15 + 0.39*1/15 = 0,167 + 0,088 + 0,072 + 0,053 + 0,026 = 0,406.

Normalised precision values for association measures

	t-score	MI	MI3	log likeli-hood	min. sensi-tivity	log-Dice	MI.log_f
Number of true colloca-tions	17	39	33	26	21	22	31
Norma-lised precision	0.115	0.406	0.366	0.262	0.357	0.391	0.562
Place	7	2	4	6	5	3	1

Normalised precision values for association measures for *рыба* (fish)

	t-score	MI	MI3	log likeli-hood	min. sensi-tivity	log-Dice	MI.log_f
Number of true collocations	29	32	57	50	63	62	69
Normalised precision	0,229	0,340	0,572	0,495	0,753	0,771	0,820
Place	7	6	4	5	3	2	1

Normalised precision values for association measures for *враг* (enemy)

	t-score	MI	MI3	log likeli-hood	min. sensi-tivity	log-Dice	MI.log_f
Number of true collocations	21	32	31	29	33	33	30
Normalised precision	0,266	0,505	0,362	0,373	0,506	0,613	0,532
Place	7	4	6	5	3	1	2

Combining collocation lists

	Collocates	T-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f
Сточный (sewer)		5	25	1	2	5	4	1
Питьевой (drinkable)		7	39	2	4	7	6	2
Грунтовый (ground)		13	53	4	7	13	10	3
…		…	…	…	…	…	…	…
Отвод (drainage)		60	0	35	37	64	50	29
Родниковый (spring)		--	78	70	--	--	--	30
Туалетный (cologne)		73	0	37	45	75	57	31

New ranking coefficients

- 1) the *number of association measures* that have “calculated” a given collocate (within 100 “cleaned” lines for each measure);
- 2) the *average rank* of the collocate: the sum of all ranks divided by the value “the number of association measures”;
- 3) the *normalised rank* of the collocate

the normalised rank = $log_2(1+7/n)$,

Combining collocation lists (2)

	Collocates	T-score	MI	MI3	log likelihood	min. sensitivity	log-Dice	MI.log_f	Кол-во мер	Ср. ранг	Норм. ранг
Сточный (sewer)		5	25	1	2	5	4	1	7	6.14	6.14
Питьевой (drinkable)		7	39	2	4	7	6	2	7	9.57	9.57
Грунтовый (ground)		13	53	4	7	13	10	3	7	14.71	14.71
…		…	…	…	…	…	…	…			
Отвод (drainage)		60	0	35	37	64	50	29	6	45.83	51.33
Родниковый (spring)		--	78	70	--	--	--	30	3	59.33	103.23
Туалетный (cologne)		73	0	37	45	75	57	31	6	53.00	59.36

The optimised rank

- This indicator is calculated taking into account the preference of the measures.
- It is calculated as follows: all products of non-zero ranks multiplied by the coefficient of the measure significance are summed up and are divided into the number of measures used for a given collocate.
- The measure significance coefficients are as follows: ***MI.log_f* – 0.4, *logDice* – 0.5, *min. sensitivity* – 0.6, MI – 0.7, MI3 – 0.8, *log-likelihood* – 0.9, *T-score* – 1.0**.

Of course, this is only preliminary ranking.

The optimised rank (2)

No.	Collocate	Average rank	Optimised rank
1.	Поверхностный (surface)	81.5	59.8
2.	Крещенский (baptismal)	82.0	36.0
3.	Обычный (usual)	61.0	34.1
4.	Газированный (sparkling)	63.0	27.9
5.	Качество (quality)	24.6	19.5
6.	Урез (encroachment line)	29.0	14.5
7.	Соленый (salt)	49.7	37.2
8.	Паводковый (flood)	52.5	22.5

Further work

- 1. Develop **the programming tool** that allows to make a single list of collocates with all the necessary parameters and calculate integrated ranks.
- 2. Study how the efficiency of the association measures is associated with **the width of the range** (to the left and to the right of the key word) within which collocates are selected, and estimate the degree of such efficiency.
- 3. Identify the inter-relation between **“syntagmatic” and “paradigmatic”** collocates on the one hand and **“idiomatic” and “statistical”** on the other hand within the same search results, and identify the dependence of such inter-relation on the width of the window.

Спасибо за внимание!