Automated Analysis of Verbal Fluency Task in Healthy Speakers: Preliminary Results

Galina Ryazanskaya   Mariya Khudyakova   Olga Buivolova

galka1999@gmail.com

National Research University Higher School of Economics, Moscow, Russia
Center for Language and Brain

Introduction: the Task and the Assessment

Categorical verbal fluency test - naming as many items from a semantic category as possible in one minute.

Widely used in:

- neurology
- psychiatry
- clinical linguistics

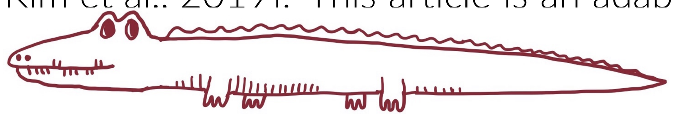
Methods of scoring:

- traditional: unique words produced
- w2v pairwise similarity of adjacent words
- clusters produced

Usually, the number of clusters is assessed manually, but it is

- time-consuming
- inconsistent (high inter-rater agreement is hard to achieve)
- no instruction exists for Russian [Drozdova et al., 2015]

Several methods of automated cluster detection were proposed, most notably [Kim et al., 2019]. This article is an adaptation of some of the methods used in [Kim et al., 2019] to the Russian language.



Data

46 healthy speakers of Russian (age, gender and education years are all linearly independent)

- ages from 18 to 75 (M = 41.17, SD = 19.53)
- 10 to 19 years of education (M = 14.94, SD = 2.54)
- 24 females

46 one category fluency task for the "animal" category transcripts, lemmatized, and the manual clustering by a trained psychiatrist for the transcripts.

Clustering Methods

- threshold cutoff:
  - at the median (c_cut_median)
  - at the mean (c_cut_mean)
  - at the 25th percentile (c_cut_p25)
  - at the average cosine similarity of each participant (c_cut_mean_local)
- sharp change (c_sharp_) at difference factors of 0.5, 0.8, 0.95, 1.05, 1.005, and 1.00001.

Model Selection

- from models (https://rusvectores.org/ru/models/) scoring the highest on semantic similarity tasks selected 4 models
- assessed number of out-of-vocabulary words (oov, e.g. "трубкозуб")
- assessed the range of cosine similarity as a measure of how well-represented animal lexicon is

Selected the model with the lowest oov and highest range of cosine similarity: tayga_upos_skipgram_300_2_2019 with a 5 bn words training set [Kutuzov and Kuzmenko, 2017].

Examples of Clustering by Different Methods

	elephant	hare	wolf	deer	kangaroo	giraffe	gopher	hamster	rabbit	penguin	ostrich	rhinoceros	crocodile	brown bear	polar bear	panda	grizzly			kolobok	boa
correct	слон	│заяц	волк	олень │	кенгуру	жираф │	суслик	хомячок	кролик │	пингвин	страус │	носорог	крокодил │	бурый медведь	белый медведь	панда	гризли │	уж	еж │	колобок │	удав
median	слон	заяц	волк	олень	кенгуру	жираф │	суслик	хомячок	кролик │	пингвин	страус	носорог	крокодил │	бурый медведь	белый медведь │	панда │	гризли │	уж │	еж │	колобок │	удав
local mean	слон │	заяц	волк	олень	кенгуру	жираф │	суслик	хомячок	кролик	пингвин	страус	носорог	крокодил │	бурый медведь	белый медведь │	панда │	гризли │	уж │	еж │	колобок │	удав
sharp 0.8	слон │	заяц	волк │	олень │	кенгуру │	жираф │	суслик │	хомячок │	кролик │	пингвин │	страус │	носорог │	крокодил │	бурый медведь	белый медведь │	панда │	гризли │	уж	еж	колобок │	удав
sharp 1.00001	слон	заяц	волк	олень │	кенгуру	жираф │	суслик	хомячок	кролик │	пингвин	страус	носорог	крокодил │	бурый медведь	белый медведь │	панда │	гризли │	уж	еж	колобок │	удав

Number of clusters and Age

In theory [Kim et al., 2019] older people produce fewer clusters And we also have the same result: p ≈ 0.01 (p<0.05), r ≈ -0.37.

	r	p
c_cut_median	-0.32566	0.0272065
c_cut_mean	-0.352979	0.0161217
c_cut_p25	-0.307397	0.0376997
c_cut_mean_local	-0.400813	0.00577282
c_sharp_1.05	-0.302898	0.0407406
c_sharp_1.00001	-0.239174	0.10941
c_sharp_0.95	-0.324345	0.0278705
c_sharp_0.8	-0.306111	0.0385493
c_sharp_0.5	-0.380397	0.00911195

All, but sharp change at 0.00001, do correlate negatively with age, the strongest being cutoff at the local mean (r ≈ -0.4, p ≈ 0.006), sharp change at 0.5 (r ≈ -0.38, p ≈ 0.009), alertcutoff at the mean (r ≈ -0.35, p ≈ 0.01).

Metrics of cutting at the local mean and sharp change at 0.5 are more strongly correlated with age than manual splits.

Nothing else is correlated (number of splits by any metric with education years or gender, age or education years with oov or average cosine similarity - but we did not expect it to)

Manual scoring

Spearman's correlation of manually calculated number of clusters with different approximations methods.

	r	p
c_cut_median	0.621444	4.04692e-06
c_cut_mean	0.623257	3.72209e-06
c_cut_p25	0.576507	2.75147e-05
c_cut_mean_local	0.680288	1.98531e-07
c_sharp_1.05	0.695671	8.03234e-08
c_sharp_1.00001	0.584366	2.0084e-05
c_sharp_0.95	0.692367	9.80138e-08
c_sharp_0.8	0.706165	4.19302e-08
c_sharp_0.5	0.731857	7.53034e-09

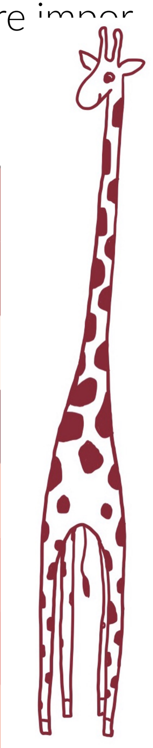
All methods of getting the number of splits are good enough at approximating manual calculation, the best being sharp change metrics with factors 0.5, 0.8, 0.95, 1.05 (r ≈ 0.73 for 0.5, r ≈ 0.7 for others) and cutoff at the local mean (r ≈ 0.7).

However, the factors of 0.5 and 0.8 produce on average too many splits (16 and 12), and the best at approximating not the ranks but the actual numbers (10 clusters on average) are probably the ones with approximately 9 clusters - sharp change metrics at 1.05 and 0.95 and cutoff at the local mean.

Quality of Cluster Boundary Positioning

Here we use accuracy, precision, recall, f1-measure and weighted f-measure to determine the quality of cluster boundaries positioning. For the task at hand precision is more important than recall.

	accuracy	precision	recall	f1-measure	f-weighted
c_cut_median	0.764904	0.735506	0.650322	0.650895	0.716729
c_cut_mean	0.773509	0.733427	0.661191	0.657087	0.717744
c_cut_p25	0.595463	0.804822	0.360821	0.455518	0.64587
c_cut_mean_local	0.7762	0.741929	0.685759	0.671965	0.729971
c_sharp_1.05	0.654006	0.771342	0.455681	0.543543	0.677481
c_sharp_1.00001	0.682829	0.751812	0.518207	0.570662	0.689635
c_sharp_0.95	0.767823	0.70666	0.656295	0.64037	0.695978
c_sharp_0.8	0.871985	0.650207	0.81263	0.684797	0.677281
c_sharp_0.5	0.969211	0.58447	0.952826	0.690515	0.633448



Cutoff at the mean is the best at catching all correct values (but poor at not-catching incorrect splits - making too many splits).

Sharp change metrics at 1.05 to 0.95 are good at only catching correct values (but bad at identifying all splits - making too little splits)

As we care more about only identifying correct splits (precision), then on identifying all correct splits (recall), we use 0.5 weighted f-measure

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

The best metrics by weighed f-measure are: cutoff at the local mean, sharp change metrics (1.05 to 0.95), and cutoff at the mean. Overall the metrics are good-ish (0.7 is not all that good but tolerable).

Results

This work successfully adapts some of the methods used in [Kim et al., 2019] to Russian language. It proves it is possible to automatically approximate both number of clusters and their positioning. For the given model the best metrics across the tasks are:

- cutoff at the local mean, that was proposed, but not realized in [Kim et al., 2019]
- sharp change measure with the factor of 0.95
- cutoff at the mean or median of the whole dataset

Future Research

The results might be improved by:

- using several raters to tag manual clustering (although high inter-rater agreement might be hard to achieve)
- tagging non-semantic associations as separate clusters, as "утка-лебедь-рак-щука" would be tagged one under current manual clustering
- solving out-of-vocabulary issues and poor representation of animal lexicon by using transfer learning (additionally training the model for the specific category)
- assessing the convexity of the curves of change in the number of clusters as one lowers the threshold
- comparing more of the models that are available for Russian language

References

[Drozdova et al., 2015] Drozdova, K., Rupchev, G., and Semenova, N. (2015). Нарушение вербальной беглости у больных шизофренией. *Социальная и клиническая психиатрия*, 25(4).

[Kim et al., 2019] Kim, N., Kim, J.-H., Wolters, M. K., MacPherson, S. E., and Park, J. C. (2019). Automatic scoring of semantic fluency. *Frontiers in psychology*, 10:1020.

[Kutuzov and Kuzmenko, 2017] Kutuzov, A. and Kuzmenko, E. (2017). *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*. Springer International Publishing, Cham.