Кластеризация лексики для решения задачи диагностики афазии[1]

Хоменко Анна Юрьевна, akhomenko@hse.ru
Исаков Данила Андреевич, issakov.daanila@mail.ru
Бальба Дарья Петровна, dpbalba@edu.hse.ru
НИУ ВШЭ – Нижний Новгород

Ранняя диагностика когнитивных нарушений является важной задачей в психиатрии, неврологии и педагогике, так как она позволяет запустить восстановительный процесс вовремя и предупредить осложнения. Клиническая лингвистика позволяет выявить эти нарушения при помощи анализа устной речи. Данная задача традиционно решается вручную, однако она нуждается в цифровых инструментах, которые помогут быстро и объективно оценить, требуется ли человеку помощь специалиста.

Для быстрой оценки расстройств мышления в различных клинических популяциях используется тест на семантическую вербальную беглость, который применяется при проведении фундаментальных исследований психики. Его суть заключается в воспроизведении максимального количества слов, принадлежащих к одной семантической категории за ограниченное время (как правило, за минуту).

В данной работе мы реализуем алгоритм автоматической кластеризации лексики для решения задачи создания компьютерной модели диагностики расстройств афатического спектра на основе результатов прохождения респондентами теста на семантическую вербальную беглость, собранных в исследовании О.В. Буйволовой и ее коллег [3], по трем категориям – «животные», «профессии» и «города» – вслед за Lundin N. B. et al [2]. До настоящего момента используемый алгоритм не применялся к русскоязычному материалу, что обусловливает новизну работы. Этот алгоритм может служить в качестве функциональной диагностической модели, пригодной для использования дефектологами и логопедами.

Исследование проводилось на речевом материале контрольной группы неврологически здоровых лиц (КГ) – 94 человека – и лиц с афазией различного генезиса (ЛСА) – 68 человек – на русском языке. При создании модели использовался следующий алгоритм: сначала слова, полученные в результате тестирования, были трансформированы в многомерные векторы при помощи дистрибутивно-семантической модели geowac_lemmas_none_fasttextskipgram_300_5_2020[2], преобразующей слова в эмбеддинги типа fastText [1]. Для выделения кластеров применялась модель, принимающая решение о смене кластера на основе косинусной близости слов в последовательности: для последовательности слов A, B, C, D переключение находится после слова B в случае, когда $S(A, B) > S(B, C)$ и $S(B, C) < S(C, D)$, где $S(A, B)$ – косинусная близость между векторами слов A и B [2].

Затем были найдены такие метрики результатов кластеризации, которые показывают значимую разницу между ответами КГ и ЛСА, представленные в Таблице 1.

Таблица 1. Метрики кластеризации, показывающие статистически значимые различия между участниками теста на вербальную беглость

Статистические метрики	**Контрольная**	**Лица с**	**Статистическая**

[1] Исследование проведено в рамках кластера проекта «Цифровые инструменты для оценки психических расстройств» стратегического проекта «Устойчивый мозг: нейрокогнитивные технологии адаптации, обучения, развития и реабилитации человека в изменяющейся среде» НИУ ВШЭ (Приоритет 2030).
[2] RusVectōrēs: семантические модели для русского языка. // URL: https://rusvectores.org/ru/models/ (дата обращения: 12.01.2024)

	группа	**афазией**	**значимость (по t-критерию Стьюдента, α=0,05)**
t-score	52	22	***
Silhouette score	0,16	0,25	***
Количество переключений	5,91	2,39	***
TTR (мера лексического разнообразия)	0,23	0,34	
Словарный запас (кол-во уникальных слов)	412	241	

Так, значение меры ассоциативности t-score оказалось выше для КГ. Это манифестирует, что в ней ассоциации между реакциями более стандартизированы. Значение silhouette-score выше для ЛСА – это свидетельствует о большей ассоциативной связанности слов внутри одного кластера. Было выявлено, что респонденты группы ЛСА обладают большим лексическим разнообразием, что также мотивировано более стандартизированными ассоциациями в ответах КГ и присутствием большого числа окказионализмов в ответах ЛСА. При этом словарный запас (количество уникальных слов) в КГ превышает словарный запас ЛСА в 1,7 раз.

Были также проанализированы наиболее частотные коллокации внутри кластеров. Наиболее частотные коллокации для категории «животные» приведены в Таблице 2.

Таблица 2. Наиболее частотные коллокации со значением t-score для контрольной группы и группы лиц с афазией по семантической категории «животные».

Контрольная группа		**Лица с афазией**	
Коллокация	**t-score**	**Коллокация**	**t-score**
кошка-собака	23.93	*кошка-собака*	17.28
лев-тигр	13.53	*волк-медведь*	8.58
жираф-слон	9.53	*заяц-лиса*	7.91
корова-овца	7.53	*волк-лиса*	7.24
заяц-лиса	7.14	*лев-тигр*	5.90

Заметим, что каждая из этих коллокаций входит в топ-10 наиболее частотных в НКРЯ, где также обладает высокими значениями t-score. Это говорит о репрезентации стандартных для носителей русского языка коллокаций внутри кластеров ответов респондентов обеих групп, причем этот вывод релевантен для всех трех семантических категорий: «животные», «профессии» и «города». Таким образом, респонденты как без нарушений, так и с афазией склонны давать ответы, отражающие стандартные для носителей русского языка ассоциации (например, *кошка-собака, заяц-лиса, лев-тигр*), однако ответы КГ оказываются статистически более стабильными.

Отметим, что данное исследование имеет ряд ограничений: небольшой объём выборки и неравномерная репрезентация типов афазии – большинство респондентов подвержены моторной афазии, что не позволяет нам на данном этапе исследования изучать особенности отдельных типов речевых нарушений.

Так, в ходе данного исследования была создана функциональная модель для определения афатических нарушений речи у носителей русского языка. При этом она моделирует некоторые когнитивные особенности носителей языка с афазией и без неё. Полученные результаты свидетельствуют о потенциале использования методов компьютерной лингвистики для прогнозирования наличия речевых нарушений и ментальных расстройств на материале русскоязычной устной речи.

Список литературы:

1. Bojanowski P. et al. Enriching word vectors with subword information //Transactions of the association for computational linguistics. – 2017. – Т. 5. – С. 135-146.
2. Lundin N. B. et al. Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis //Psychiatry research. – 2022. – Т. 309. – С. 114404.
3. Буйволова О. В. и др. Adaptation of the Aphasia Bedside Check for Russian //Российский журнал когнитивной науки. – 2020. – Т. 7. – №. 3. – С. 45-67.