Применение больших языковых моделей для выявления значений слов в задаче обнаружения лексико-семантических изменений

Д. В. Кокосинский*

Рассматривается обнаружение лексико-семантических изменений на основе кластеризации словоупотреблений по матрице их семантических близостей. Большие языковые модели оценивают близость точнее компактных специализированных моделей, но на порядки дороже, поэтому в предлагаемом методе QUEST-cluster они применяются точечно, в рамках байесовского подбора порога кластеризации. Система на основе QUEST-cluster повышает среднюю корреляцию Спирмена с эталоном до 0,85 против 0,755 у системы на основе компактной модели.

Ключевые слова: лексико-семантические изменения, выявление значений слов, большие языковые модели, байесовский подбор порога.

1. Введение

В задаче обнаружения лексико-семантических изменений [1] по словоупотреблениям целевых слов из двух временных периодов требуется ранжировать слова по степени изменения их значений. Качество измеряется корреляцией Спирмена с эталонным ранжированием.

В подходе на основе выявления значений словоупотребления слова из обоих периодов кластеризуются, каждый кластер считается значением, а степенью изменений служит расстояние Йенсена–Шеннона между распределениями словоупотреблений двух периодов по кластерам — стандартная мера расхождения распределений, принятая в эталонах задачи [1]. Подход реализован в системе CC+JSD [2], где компактная специализированная модель строит матрицу семантических близостей словоупотреблений, а корреляционная кластеризация разбивает её на кластеры.

При этом большие языковые модели оценивают семантическую близость точнее компактных специализированных моделей [3], но запрос

* *Кокосинский Денис Владиславович* — аспирант каф. алгоритмических языков ф-та ВМК МГУ, e-mail: deniskokossskappa@gmail.com, ORCID: 0000-0001-5642-8725.

Kokosinskii Denis Vladislavovich — PhD student, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

к ним требует приблизительно в 10^5 раз больше вычислений, и вычисление полной матрицы близостей обходится неприемлемо дорого.

2. Метод QUEST-cluster

Пусть $U = \{u_1, \ldots, u_n\}$ — словоупотребления целевого слова. *Слабый* оценщик семантической близости (компактная специализированная модель) вычисляет матрицу близостей r_{ij} для всех пар, а запрос к *сильному* оценщику (большой языковой модели) на несколько порядков дороже. Ответ сильного оценщика b бинарный и равен 1, если значение слова в паре одно и то же, и 0 иначе. Требуется построить разбиение U по матрице r и не более чем $B \ll n^2$ ответам сильного оценщика. Качество разбиения оценивается по недоступным методу ручным аннотациям метрикой AnnotRI [2] — долей аннотированных пар, на которых решение «один кластер или разные» совпадает с решением аннотаторов, в среднем по словам.

Значительную часть ошибок CC+JSD составляют ошибки гранулярности. Корреляционная кластеризация [2] штрафует на $|r_{ij} - \theta|$ пары, у которых знак $r_{ij} - \theta$ противоречит разбиению, так что порог θ задаёт гранулярность значений. Порог θ_0 системы CC+JSD един для всех целевых слов, но может не подходить отдельному слову, и тогда его значения дробятся или укрупняются. Оракульный выбор порога каждого слова по аннотациям повышает среднее согласие с ними с 0,796 до 0,862. Метод QUEST-cluster подбирает такой порог θ_w по ответам сильного оценщика как точку шкалы слабого оценщика, в которой сильный оценщик меняет ответ.

Вероятность ответа $b = 1$ сильного оценщика для пары со слабой близостью r описывается логистической моделью отклика

$$p(r) = \sigma\big((r - \theta_w)/\beta\big), \tag{1}$$

где θ_w — искомый порог ($p(\theta_w) = 1/2$), $\beta > 0$ — ширина переходной зоны. Параметры оцениваются байесовским выводом на сетке $\Theta \times \mathcal{B}$, где Θ и $\mathcal{B}$ — конечные множества кандидатов для θ_w и β, с априорным распределением $\pi(\theta, \beta) \propto \exp\big(-(\theta - \theta_0)^2/2\sigma_0^2\big)$. По ответам $(r_1, b_1), \ldots, (r_m, b_m)$ апостериорное распределение вычисляется с точностью до нормировки:

$$S_m(\theta, \beta) = \pi(\theta, \beta) \prod_{t=1}^{m} p_{\theta,\beta}(r_t)^{b_t} \big(1 - p_{\theta,\beta}(r_t)\big)^{1-b_t}, \qquad p_{\theta,\beta}(r) = \sigma\big((r-\theta)/\beta\big).$$

Подбор порога по маргинальному распределению $\sum_\beta S_m(\theta, \beta)$ выполняется в три шага.

1. Сильному оценщику предъявляются m пар с наименьшими $|r_{ij} - \theta_0|$. Если биномиальный критерий на уровне α не отвергает гипотезу $p(\theta_0) = 1/2$ по числу ответов $b = 1$, принимается $\theta_w = \theta_0$, и процедура завершается.
2. Иначе, пока центральный 90 %-интервал апостериорного распределения порога шире w и запрошено меньше K пакетов, предъявляются m неопрошенных пар с наименьшими $|r_{ij} - \theta_{\text{med}}|$, где θ_{med} — апостериорная медиана.
3. Если выборочная корреляция r и b по всем ответам отрицательна (при постоянных b — нулевая), принимается $\theta_w = \theta_0$, иначе $\theta_w = \theta_{\text{med}}$.

Результат метода — кластеризация матрицы r при пороге θ_w. Название отсылает к психофизическому методу QUEST [4].

3. Эксперименты

В системе QUEST-cluster+JSD, как и в CC+JSD [2], слабый оценщик — специализированная модель GlossReader, θ_0 — порог CC+JSD. Сильные оценщики — две конфигурации больших языковых моделей из [3]: Qwen3.5-9b в режиме рассуждения (ответ y по шкале 0–10, порог c подобран по обучающим выборкам CoMeDi) и GPT-5.3-Chat (шкала 1–4, $c = 2{,}5$). Параметры (B, m, K — по вычислительному бюджету, остальные — по гистограммам близостей GlossReader): $B = 100$ (менее 1 % пар, из них подбор порога расходует от m до Km), $m = 20$, $K = 3$, $\alpha = 0{,}115$ (отклонение числа единиц от $m/2$ не менее 4), $w = 0{,}05$, $\sigma_0 = 0{,}06$, $\mathcal{B} = \{0{,}05;\ 0{,}1;\ 0{,}15;\ 0{,}25;\ 0{,}5\}$, Θ — 57 равноотстоящих значений между порогами грубой и тонкой гранулярности из [2]. Остаток бюджета система тратит на не входящую в метод правку близостей: пары у порога θ_w, пары слабосвязных словоупотреблений и случайные межкластерные пары получают близость $\theta_w + (y - c)/h$ (h — диапазон шкалы y).

Системы сравниваются на семи бенчмарках из [2] с эталоном по расстоянию Йенсена–Шеннона. Обе конфигурации повышают согласие с аннотациями на каждом бенчмарке, а корреляцию Спирмена — на всех семи с Qwen3.5-9b и на шести с GPT-5.3-Chat. В среднем согласие растёт с 0,796 у CC+JSD до 0,832 (Qwen3.5-9b) и 0,835 (GPT-5.3-Chat), корреляция Спирмена — с 0,755 до 0,845 и 0,847. Сам метод, без правки близостей, даёт 0,814 и 0,780, то есть около половины и четверти прироста.

Список литературы

[1] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection", *Proceedings of SemEval-2020*, 2020, 1–23. DOI: 10.18653/v1/2020.semeval-1.1.

[2] D. V. Kokosinskii, N. V. Arefyev, "Automatic systems with word sense induction for lexical semantic change detection", *Numerical Methods and Programming*, **27**:3 (2026), 450–469. DOI: 10.26089/NumMet.v27r329.

[3] Д. В. Кокосинский, "Большие языковые модели в задаче оценки семантической близости словоупотреблений", *Электронные библиотеки*, **29**:4 (2026), 1133–1154. DOI: 10.26907/1562-5419-2026-29-4-1133-1154.

[4] A. B. Watson, D. G. Pelli, "QUEST: A Bayesian adaptive psychometric method", *Perception & Psychophysics*, **33**:2 (1983), 113–120. DOI: 10.3758/BF03202828.

Статья поступила 7 октября 2026 г.

Large language models for word sense induction in lexical semantic change detection

D. V. Kokosinskii

We consider lexical semantic change detection based on clustering word usages by the matrix of their semantic proximities. Large language models estimate proximity more accurately than compact specialized models but are orders of magnitude more expensive, so the proposed QUEST-cluster method uses them selectively, within Bayesian estimation of the clustering threshold. A system based on QUEST-cluster raises the average Spearman correlation with the gold ranking to 0.85, compared with 0.755 for a system based on a compact model.

Keywords: lexical semantic change, word sense induction, large language models, Bayesian threshold estimation.

References

[1] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection", *Proceedings of SemEval-2020*, 2020, 1–23. DOI: 10.18653/v1/2020.semeval-1.1.

[2] D. V. Kokosinskii, N. V. Arefyev, "Automatic systems with word sense induction for lexical semantic change detection", *Numerical Methods and Programming*, **27**:3 (2026), 450–469. DOI: 10.26089/NumMet.v27r329.

[3] D. V. Kokosinskii, "Large Language Models for Word-in-Context", *Russian Digital Libraries Journal*, **29**:4 (2026), 1133–1154 (in Russian. DOI: 10.26907/1562-5419-2026-29-4-1133-1154).

[4] A. B. Watson, D. G. Pelli, "QUEST: A Bayesian adaptive psychometric method", *Perception & Psychophysics*, **33**:2 (1983), 113–120. DOI: 10.3758/BF03202828.

Received on October 7, 2026