Чуксин Б.Г., Труфанов А.И. Жанровая специфика в семантических связях сетевых моделей текстов // Международный журнал информационных технологий и энергоэффективности.– 2026. –Т. 11 № 6(68) Ч. 1 с. 117–121

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/



УДК 004.912

ЖАНРОВАЯ СПЕЦИФИКА В СЕМАНТИЧЕСКИХ СВЯЗЯХ СЕТЕВЫХ МОДЕЛЕЙ ТЕКСТОВ

[1] **Чуксин Б.Г., Труфанов А.И.**
ФГБОУ ВО "ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ", Иркутск, Россия (664074, Иркутская область, город Иркутск, ул. Лермонтова, д. 83), e-mail: [1] bagisolo@mail.ru

В работе описывается методология построения графов совместной встречаемости на материале русскоязычного корпуса Taiga и представлены результаты пилотного эксперимента на подвыборке из 19 текстов трёх жанров — новостей, художественной прозы и нон-фикшн. По каждому тексту строился взвешенный граф с окном w=5; на полученных графах считались пять сетевых метрик. Проверка критерием Краскела-Уоллиса показала: три метрики из пяти — средняя длина кратчайшего пути (p=0.019), плотность (p=0.011) и модульность (p=0.012) — различаются между жанрами на статистически значимом уровне. Это даёт основания рассматривать графовый подход как рабочий инструмент жанровой классификации для русскоязычных текстов.

Ключевые слова: Графовые модели текста, граф совместной встречаемости, сетевые метрики, модульность, жанровая классификация, Taiga Corpus, NetworkX.

GENRE SPECIFICITY IN SEMANTIC CONNECTIONS OF NETWORK-BASED TEXT MODELS

[1]Chuksin B.G., Trufanov A.I.
IRKUTSK NATIONAL RESEARCH TECHNICAL UNIVERSITY, Irkutsk, Russia (664074, Irkutsk city Lermontova st., 83), e-mail: [1] bagisolo@mail.ru

This paper presents a methodology for constructing co-occurrence graphs from Russian-language texts of the Taiga corpus and reports the results of a pilot experiment on a sub-sample of 19 texts across three genres — news, fiction, and non-fiction. A weighted co-occurrence graph (window w=5) was built for each text, and five network metrics were computed. The Kruskal-Wallis test revealed statistically significant genre differences in three metrics: average shortest path length (p=0.019), graph density (p=0.011), and modularity (p=0.012). The findings support the viability of the graph-based approach for genre classification of Russian texts.

Keywords: graph-based text models, co-occurrence graph, network metrics, modularity, genre classification, Taiga Corpus, NetworkX.

Введение

Семантические связи в тексте — штука парадоксальная: любой читатель их чувствует, но формально описать непросто. Векторные подходы вроде TF-IDF, Word2Vec или BERT решают задачу в лоб — превращают текст в точку пространства. Удобно, но связи между конкретными словами при этом исчезают. Графовые модели устроены иначе: ребро между двумя узлами и есть тот самый объект изучения, а значит, к нему применим весь аппарат теории сетей [1].

В данной работе ставится задача разработать и проверить методологию построения co-occurrence графов на материале корпуса Taiga — одного из крупнейших русскоязычных

корпусов — и выяснить, меняется ли структура таких графов от жанра к жанру. Вопрос не абстрактный: графовые нейронные сети всё активнее применяются в задачах обработки текста [8], и то, каким именно окажется входной граф, напрямую влияет на качество модели. Пока выбор типа графового представления нередко делается на глаз — настоящая работа пытается дать для этого выбора хоть какое-то эмпирическое основание. На русскоязычном материале с полноценной статистической проверкой такой анализ проводится впервые.

Обзор литературы

Графовые модели текста как исследовательское направление сформировались в начале 2000-х годов. Одной из первых значимых работ, продемонстрировавших практическую применимость таких моделей, стала система TextRank [1], которая использовала граф совместной встречаемости для автоматического реферирования. Параллельно Ferrer i Cancho и Solé показали, что лексические сети обладают свойствами малого мира — низким диаметром и высоким коэффициентом кластеризации [2], что открыло новый угол зрения на организацию языка как системы.

Амансио и соавторы систематически исследовали связь между сложностью языка и топологическими свойствами текстовых графов [3], а Quispe и соавторы показали, что метрики co-occurrence сетей обладают высокой дискриминирующей способностью при классификации текстов [4]. Для семантических графов на основе дистрибутивного сходства принципиально важна работа Biemann и Riedl [5], где обосновывается идея связывания слов по близости их контекстуальных векторов.

Применительно к русскому языку наиболее близкими к данной работе являются исследования Зарубина Кирилла Александровича и Труфанова Андрея Ивановича по сетевым метрикам русской классической литературы [6] и собственная предыдущая работа автора, где аналогичный инструментарий применялся к интернет-текстам [7]. Настоящая статья продолжает эту линию исследований, распространяя её на разнородный жанровый корпус с применением строгой статистической проверки.

Методология

Корпус и выборка. Исследование проводится на материале корпуса Taiga [9] — многожанрового собрания русскоязычных текстов. В пилот вошли тексты трёх жанров — новости, художественная проза и нон-фикшн. Социальные сети изначально тоже планировались, однако из-за особенностей формата хранения этого подкорпуса удалось извлечь лишь один текст — этого очевидно недостаточно для какого-либо сравнения, поэтому жанр из анализа исключён. В итоге выборка получилась такой: 6 новостных текстов, 6 прозаических и 7 нон-фикшн, всего 19. Каждый текст обрезался до 500 токенов.

Предобработка. Предобработка намеренно сделана минималистичной. Регулярное выражение вытаскивает кириллические последовательности, всё приводится к нижнему регистру, затем вычищаются стоп-слова — их набралось 127 единиц — пунктуация и числа. Лемматизация в пилотном эксперименте не применялась — это ограничение будет снято в полной версии исследования с использованием библиотеки Natasha [10].

Построение графа. Граф строится следующим образом: каждая уникальная словоформа становится вершиной, а ребро между двумя вершинами появляется когда соответствующие

слова оказываются в пределах одного скользящего окна шириной пять токенов. Граф неориентированный, взвешенный — вес ребра отражает, сколько раз эта пара слов встретилась рядом в тексте. Граф построен средствами библиотеки NetworkX.

Извлечение метрик. На каждом графе считались пять метрик. Средняя степень вершины и коэффициент кластеризации характеризуют локальную связность; средняя длина кратчайшего пути считалась на наибольшей связной компоненте — для несвязных графов это стандартный обходной манёвр. Плотность даёт представление о насыщенности рёбрами в целом. Модульность вычислялась алгоритмом Лувена через пакет python-louvain и отражает, насколько чётко граф распадается на тематические кластеры.

Статистический анализ. Для проверки гипотезы о жанровых различиях применялся критерий Краскела-Уоллиса — непараметрический аналог однофакторного дисперсионного анализа, не требующий нормальности распределений. Уровень значимости $\alpha = 0.05$.

Результаты

В Таблице 1 представлены усреднённые значения пяти сетевых метрик для каждого жанра пилотной выборки.

Таблица 1 - Средние значения сетевых метрик co-occurrence графов по жанрам (w=5).

Жанр	Avg. Degree	Clustering	Avg. Path	Density	Modularity
Новости (n=6)	11.270	0.646	2.987	0.081	0.601
Проза (n=6)	11.410	0.646	3.317	0.049	0.644
Нон-фикшн (n=7)	11.098	0.647	4.227	0.028	0.718

В таблице 2 представлены результаты критерия Краскела-Уоллиса.

Таблица 2 - Результаты критерия Краскела-Уоллиса. * $p < 0.05$; н.з. — различия незначимы.

Метрика	H	p	Значимость
avg_degree	1.064	0.5873	н.з.
clustering	0.012	0.9942	н.з.
avg_path	7.906	0.0192	*
density	9.064	0.0108	*
modularity	8.854	0.0120	*

Три метрики из пяти оказались жанрово чувствительными — плотность (H=9.064, p=0.011), модульность (H=8.854, p=0.012) и средняя длина пути (H=7.906, p=0.019). Степень вершин и кластеризация практически не различаются между жанрами. Последнее само по себе любопытно: при окне w=5 локальная структура графа оказывается одинаковой вне зависимости от того, новость это, роман или научпоп.

Обсуждение

Нон-фикшн выделяется сильнее всего: модульность 0.718, плотность всего 0.028, средний путь — 4.227. За этими числами стоит вполне понятная вещь: такие тексты тематически монолитны, каждый следующий абзац вводит новые термины и редко возвращается к старым. Результат — разреженная сеть с хорошо отделёнными кластерами.

Новости — полная противоположность. Плотность 0.081, пути короткие (2.987), модульность низкая (0.601). Одни и те же слова — имена, топонимы, глаголы — мигрируют через весь текст, и граф получается густым и однородным без явных тематических зон.

Проза занимает промежуточное положение по всем метрикам. Художественный текст тематически разнообразен, но нарратив формирует связность через повторяющихся персонажей и сюжетные мотивы — отсюда умеренная модульность (0.644) и плотность (0.049).

Необходимо обозначить ограничения пилотного эксперимента. У пилота есть два очевидных слабых места. Первое — объём: 19 текстов дают статистику, но не дают уверенности в её устойчивости. Второе — отсутствие лемматизации: «идёт», «шёл» и «ходит» в нынешней версии — три разные вершины графа, что раздувает его и, скорее всего, занижает метрики связности. В полном эксперименте на 300 текстах оба момента будут закрыты: выборка вырастет, а Natasha обеспечит нормальную морфологическую нормализацию.

Заключение

В работе предложена и апробирована методология построения co-occurrence графов на материале русскоязычного корпуса Taiga. Пилотный эксперимент на 19 текстах трёх жанров дал конкретный результат: из пяти исследованных метрик три — плотность, модульность и средняя длина пути — жанрово специфичны, тогда как степень вершин и коэффициент кластеризации при окне w=5 оказались жанрово нейтральными. Последнее тоже результат: эти две метрики можно исключить из признакового пространства при последующей классификации.

Практический вывод прямой: плотность и модульность co-occurrence графа работают как жанровые маркеры для русскоязычных текстов. Дальнейшее исследование предполагает расширение выборки до 300 текстов, добавление лемматизации через библиотеку Natasha и сравнение co-occurrence графов с графами синтаксических зависимостей и семантического сходства. Переход от статического графа к динамическому — отслеживание того, как структура семантической сети складывается по мере последовательной обработки текста.

Список литературы

1. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts // Proceedings of EMNLP. 2004. pp. 404–411.
2. Ferrer i Cancho R., Solé R.V. The small world of human language // Proceedings of the Royal Society of London. Series B. 2001. Vol. 268. pp. 2261–2265.
3. Amancio D.R., Aluisio S.M., Oliveira O.N., Costa L.F. Complex networks analysis of language complexity // EPL (Europhysics Letters). 2012. Vol. 100, No. 5. p. 58002.

4. Quispe L.V., Tohalino J.A., Amancio D.R. Using virtual edges to improve the discriminability of co-occurrence text networks // Physica A: Statistical Mechanics and its Applications. 2021. Vol. 562. pp. 125344.
5. Biemann C., Riedl M. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity // Journal of Language Modelling. 2013. Vol. 1, No. 1. pp. 55–95.
6. Зарубин К.А., Труфанов А.И. Комплексные сети в русской классической литературе: атрибуция текста и сетевая модель // Обществознание и социальная психология. 2023. № 10 (58). С. 12–18.
7. Чуксин Б.Г., Труфанов А.И. Сетевая стилометрия русскоязычных интернет-текстов // Научные высказывания. 2024. № 12 (59). С. 24–28.
8. Yao L., Mao C., Luo Y. Graph Convolutional Networks for Text Classification // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. Vol. 33, No. 01. P. 7370–7377.
9. Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser // Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika — 2017». Saint-Petersburg: Izd-vo SPbGU, 2017.
10. Korobov Y. Natasha: NLP library for Russian [Elektronnyj resurs]. — URL: https://github.com/natasha/natasha (data obrashhenija: 15.04.2026).

References

1. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts, Proceedings of EMNLP, 2004, pp. 404–411.
2. Ferrer i Cancho R., Solé R.V. The small world of human language, Proceedings of the Royal Society of London. Series B, 2001, vol. 268, pp. 2261–2265.
3. Amancio D.R., Aluisio S.M., Oliveira O.N., Costa L.F. Complex networks analysis of language complexity, EPL (Europhysics Letters), 2012, vol. 100, no. 5, p. 58002.
4. Quispe L.V., Tohalino J.A., Amancio D.R. Using virtual edges to improve the discriminability of co-occurrence text networks, Physica A: Statistical Mechanics and its Applications, 2021, vol. 562, p. 125344.
5. Biemann C., Riedl M. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity, Journal of Language Modelling, 2013, vol. 1, no. 1, pp. 55–95.
6. Zarubin K.A., Trufanov A.I. Kompleksnye seti v russkoj klassicheskoj literature: atributsiya teksta i setevaya model [Complex networks in Russian classical literature: text attribution and network model], Obshchestvoznanie i sotsial'naya psikhologiya, 2023, no. 10 (58), pp. 12–18.
7. Chuksin B.G., Trufanov A.I. Setevaya stilometriya russkoyazychnykh internet-tekstov [Network stylometry of Russian-language internet texts], Nauchnye vyskazyvaniya, 2024, no. 12 (59), pp. 24–28.
8. Yao L., Mao C., Luo Y. Graph Convolutional Networks for Text Classification, Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, no. 01, pp. 7370–7377.
9. Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser, Proc. of International Conference «Corpus Linguistics — 2017», Saint-Petersburg, 2017.
10. Korobov Y. Natasha: NLP library for Russian. Available at: https://github.com/natasha/natasha (accessed 15.04.2026).