*УДК 81-23*
*DOI 10.47388/2072-3490/lunn2024-68-4-24-40*

# АВТОМАТИЗАЦИЯ ПОИСКА КОЛЛОКАЦИЙ В ТЕКСТЕ: МЕТОД СТАТИСТИЧЕСКОЙ ОБРАБОТКИ *VS* МОДЕЛЬ ГЕНЕРАТИВНОГО ПРЕДОБУЧЕННОГО ТРАНСФОРМЕРА

**Д. И. Галюченко**

Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации, Москва, Россия

Компьютерная лингвистика — область исследования, неразрывно связанная с автоматической обработкой текстов на естественном языке. В последние годы она стала особенно актуальной благодаря развитию новых технологий, таких как модели генеративных предобученных трансформеров (GPT). Эти модели способны обрабатывать долгосрочные зависимости в тексте, что делает их перспективными для поиска коллокаций — семантически связанных словосочетаний. Цель исследования заключается в сравнении эффективности двух методов поиска коллокаций: статистической обработки естественного языка (Statistical NLP) и GPT-4 Turbo. Для этого была разработана программа, использующая меру статистической зависимости PMI, и проведен сравнительный анализ с результатами GPT-модели. Материалом исследования послужила статья 5 Европейской конвенции по правам человека.

Оценка особенностей и возможностей применения методов автоматизации поиска коллокаций в тексте в виде статистической обработки текстов естественного языка и модели генеративного предобученного трансформера (GPT) в контексте автоматизации лингвистических исследований позволяет лучше понять разнообразие подходов и сгенерировать наиболее подходящий метод автоматизации поиска коллокаций в тексте для своих собственных исследований. В ходе исследования описываются отличия в подходе к анализу и пониманию текстов на естественном языке в случае выбора метода, использующего модель GPT, и метода статистической обработки текстов на естественном языке и проводится сопоставительный анализ полученных результатов. Оба подхода имеют свои преимущества и ограничения, и выбор между ними зависит от конкретных задач и ресурсов. В некоторых случаях комбинирование этих методов может привести к лучшим результатам в обработке текстов на естественном языке.

Ключевые слова: NLP; GPT; статистические методы; коллокации; лингвистика; компьютерная лингвистика.

**Цитирование:** Галюченко Д. И. Автоматизация поиска коллокаций в тексте: метод статистической обработки *vs* модель генеративного предобученного трансформера // Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова. 2024. Вып. 4 (68). С. 24–40. DOI 10.47388/2072-3490/lunn2024-68-4-24-40.

**Automatic Collocation Retrieval in Texts:
Statistical Processing Method vs. Generative Pre-Trained Transformer Model**

**Danil I. Galyuchenko**

Moscow State Institute of International Relations
of the Ministry of Foreign Affairs of the Russian Federation, Moscow, Russia

Computational linguistics is a field of research that is closely related to automatic processing of natural language texts. It has become highly relevant in recent years due to the development of new technologies such as generative pre-trained transformer (GPT) models. These models are able to deal with long-term dependencies in text, which makes them promising for searching collocations — semantically related word combinations. The aim of the study is to compare the performance of two collocation retrieval methods: Statistical Natural Language Processing (Statistical NLP) and GPT-4 Turbo. For this purpose, a program using the PMI statistical dependency measure was developed, and a comparative analysis with the results of the GPT model was carried out. The research material was Article 5 of the European Convention on Human Rights. In terms of automating language research, evaluation of features and possibilities in applying methods of automated collocation search in texts through statistical processing of natural language texts and the generative pre-trained transformer (GPT) model for collocation search in texts allows for a better understanding of the diverse approaches and helps select the optimal method automated collocation search in texts for specific research projects. The study describes the differences in the approach to analysis and understanding of natural language texts in the case of the GPT model-based method and the statistical natural language text processing method, and compares the results obtained. Both methods have their advantages and disadvantages, and the choice between them depends on specific tasks and available resources. In some cases, combining these methods can lead to better results in natural language text processing.

**Key words:** NLP; GPT; statistical methods; collocations; linguistics; computational linguistics.
**Citation:** Galyuchenko, Danil I. (2024) Automatic Collocation Retrieval in Texts: Statistical Processing Method vs. Generative Pre-Trained Transformer Model. *LUNN Bulletin*, 4 (68), 24–40. DOI 10.47388/2072-3490/lunn2024-68-4-24-40.

## 1. Введение

Компьютерное языкознание и его прикладные отрасли являются одними из самых популярных областей исследования в связи с появлением новых технологий, которые представляют интерес с точки зрения автоматизации выполнения разного рода прикладных задач или генерации текстов на естественном языке.

Представляется актуальным оценить особенности и возможности применения методов автоматизации поиска коллокаций в тексте в виде статистической обработки текстов естественного языка (*Statistical Natural Language Processing*, *Statistical NLP*), на примере метода на основе правил и модели гене-

ративного предобученного трансформера (*Generative Pre-trained Transformer*, *GPT*) при поиске коллокаций в тексте с целью генерирования наиболее подходящего метода автоматизации эффективного поиска коллокаций в тексте.

В качестве гипотезы исследования выступает предположение о том, что новые *GPT*-модели способны превзойти традиционные методы статистической обработки естественного языка в связи с преимуществами, которые они получают от разработанного механизма внимания для обработки долгосрочных зависимостей в тексте.

Целью данного исследования является сравнительный анализ эффективности двух различных методов автоматизации поиска коллокаций в тексте на основе метода статистической обработки естественного языка. Исследование направлено на определение преимуществ и ограничений каждого метода в контексте выявления семантически связанных словосочетаний в тексте на примере статьи 5 Европейской конвенции по правам человека.

Обработка естественного языка (*Natural Language Processing*, *NLP*) — это ряд основанных на теоретических исследованиях компьютерных технологий автоматического анализа и представления человеческого языка (Cambria, White 2014).

Автоматический анализ текста предполагает глубокое понимание машинами естественного языка. До настоящего времени поиск, накопление и обработка информации преимущественно основывались на алгоритмах, опирающихся на текстовое представление.

Такие алгоритмы эффективно применяются для решения задач, связанных с извлечением текстов, разбиением их на части, проверкой орфографии и подсчетом количества слов. Однако при истолковании предложений и извлечении содержательной информации их возможности оказываются весьма ограниченными. Обработка естественного языка, по сути, требует его высокоуровневого понимания (Dyer 1995).

Статистический метод обработки текстов на естественном языке является основным направлением исследований в области *NLP* с конца 1990-х годов. Он опирается на языковые модели, построенные на основе таких популярных алгоритмов машинного обучения, как максимальное сходство, максимальная вероятность появления, условно-случайные распределения и системы опорных векторов. Использование большого тренировочного корпуса аннотированных текстов в процессе машинного обучения позволяет системе не только узнать валентность (в синтаксисе — способность слова образовывать синтаксические связи с другими элементами [Лингвистический энциклопедический словарь 1990]) ключевых слов (как в методе выделения ключевых слов), но и учесть

валентность других произвольных ключевых слов (например, лексическую близость), пунктуацию и частоту встречаемости слов.

С развитием вычислительных мощностей и доступом к большим объемам данных стали возможными более сложные статистические методы. Были разработаны алгоритмы для обработки более широкого спектра задач, включая морфологический анализ, выявление синтаксических связей и машинный перевод.

По мере совершенствования технологий машинного обучения появляются более сложные статистические модели, способные работать с семантикой и смыслом текстов. Появились первые нейросетевые модели, такие как рекуррентные нейронные сети (*RNN*) и свёрточные нейронные сети (*CNN*). Рекуррентные нейронные сети, в частности нейронные сети с длинной кратковременной памятью и рекуррентные нейронные сети с управлением, быстро стали лидерами среди подходов к моделированию последовательностей и решению задач трансдукции, таких как моделирование языка и машинный перевод (Wu, Schuster, Chen, Le, Norouzi, Macherey et al. 2016). С тех пор было предпринято множество попыток расширить границы применения рекуррентных языковых моделей и архитектур кодеров-декодеров (Jozefowicz, Vinyals, Schuster, Shazeer, Wu 2016).

В последнее десятилетие обработка *NLP* приобрела еще большую популярность благодаря успехам в области глубокого обучения и появлению трансформеров. Модели, такие как *BERT* (*Bidirectional Encoder Representations from Transformers*), *GPT* (*Generative Pre-trained Transformer*) и их последующие версии, обусловили большой скачок в обработке естественного языка, позволяя обрабатывать тексты с высокой степенью семантической связи.

Архитектуру трансформеров и механизм внимания придумали исследователи из *Google Research* и описали в статье *Attention Is All You Need*. Эта статья была опубликована в 2017 году и является ключевой для понимания трансформеров в области обработки естественного языка (*NLP*) (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, et al. 2017).

Механизм внимания является ключевой особенностью трансформера, которая позволила ему достичь впечатляющих результатов в различных областях обработки естественного языка и машинного обучения.

Этот механизм позволяет модели сосредотачиваться на разных частях входных данных с разной степенью важности, каждое слово обрабатывается независимо, и внимание между всеми парами слов рассчитывается одновременно, что играет большую роль в обработке долгосрочных зависимостей.

Обилие больших немаркированных текстовых корпусов не решает проблему обучения выполнению специфических задач, так как для этого требуются особые маркированные данные, которых достаточно мало, что затрудняет

адекватную работу моделей. Ученые из *OpenAI* утверждают, что значительный прогресс в решении этих задач может быть достигнут путем генеративного предварительного обучения языковой модели на разнообразных корпусах немаркированного текста с последующей дискриминативной настройкой на каждую конкретную задачу. В отличие от предыдущих подходов, были использованы преобразования входных данных с учетом задач в процессе точной настройки для достижения эффективного переноса модели при минимальных изменениях в ее архитектуре. В 2018 году была представлена первая модель *GPT* (Radford, Narasimhan, Salimans, Sutskever 2018).

Подобные высокопроизводительные системы сочетают в себе огромный набор специализированных тренировочных данных и контролируемое человеком обучение и показывают впечатляющие результаты в рамках конкретных задач, таких как ответы на вопросы, машинный перевод, понимание прочитанного и обобщение. Однако они обладают низкой гибкостью и чувствительны к малейшим изменениям в структуре распределения данных и спецификации задач.

В связи с этим те же ученые в 2019 году продемонстрировали новую языковую модель, названную *GPT-2*, которая обучается решению самых разных задач без явного контроля при обучении на новом массиве данных, состоящем из миллионов веб-страниц, называемом *WebText* (Radford, Wu, Child, Luan, Amodei, Sutskever 2019).

В опубликованных работах был продемонстрирован существенный прирост эффективности решения многих задач и прохождения тестов *NLP* за счет предварительного обучения на большом корпусе текстов с последующей тонкой настройкой на конкретную задачу. Несмотря на то что по своей структуре модель не заточена под какую-то конкретную задачу, ее все еще необходимо дообучать на огромных массивах данных для эффективного решения отдельных задач. В отличие от машины, человеку не нужно такое большое количество примеров, чтобы чему-то научиться, ему хватит и нескольких. Разработчики продолжили двигаться в этом направлении и старались научить модель выполнять разные задачи на как можно меньшем количестве примеров. В 2020 году они представили новую авторегрессивную языковую модель *GPT-3*, которая применяется без каких-либо модификаций или точных настроек, а задания и малочисленные примеры задаются исключительно через текстовое взаимодействие с моделью (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal et al. 2020). Следовательно, работа в данном направлении продолжается.

Научная новизна исследования определяется тем, что в нем не только был выполнен обзор опубликованных работ по изучаемой теме, но и проведен самостоятельный эксперимент. В нем была самостоятельно разработана программа для поиска коллокаций в тексте с использованием меры статистической

зависимости *PMI*, на основе выбранного материала проведено наблюдение результатов работы данной программы и проведен сопоставительный анализ с результатами работы *GPT*-модели *Copilot* на основе *GPT-4 Turbo* при выполнении аналогичной работы.

**2. Характеристика материала и методов исследования**

Материалом исследования послужила статья 5 Европейской конвенции по правам человека, на основе которой был произведен поиск устойчивых N-грамм методом использования меры статистической зависимости *PMI* и *GPT*- модели *Copilot* на основе *GPT-4 Turbo*. В работе применялись следующие методы: поиск литературы по теме исследования, описательный метод, наблюдение, сравнение, измерение, проведение эксперимента, индуктивный анализ, синтез.

**3. Результаты исследования и их обсуждение**

Следует подчеркнуть, что в последнее время внимание исследователей приковано к вопросу о методах и алгоритмах, позволяющих автоматически выявлять и анализировать коллокации (Smadja 1991, 1993; Covington 1993). По определению Choueka, в *NLP* коллокация — это последовательность расположенных рядом слов, которые часто встречаются вместе (Choueka, Klein, Neuwitz 1983). Теоретически последовательности могут быть любой длины, но в реальности они содержат от двух до шести слов.

Frank Smadja выделил три типа коллокаций (Smadja 1993). Первые имеют предикативные связи, они состоят из двух слов, неоднократно употребляемых вместе в аналогичной синтаксической связи. Ко вторым относятся строгие именные группы, они включают в себя непрерывные последовательности слов, которые часто нельзя разделить на более мелкие составные части без потери смысла. Третьи являются фразовыми шаблонами, состоящими из идиоматических фраз, которые имеют особую ценность для генерации языка.

Однако статистические методы, как правило, характеризуются слабой семантической чувствительностью, т. е., за исключением однозначных ключевых слов, другие лексические элементы или совместно повторяющиеся элементы в статистической модели имеют незначительную предиктивную ценность по отдельности. В результате статистические методы классификации текста работают с приемлемой точностью только при достаточно большом объеме исходного текста. Таким образом, хотя эти методы могут классифицировать текст на уровне страницы или абзаца, они не очень хорошо работают с более мелкими единицами текста, такими как предложения или фразы (Cambria, White 2014).

Коллокация в лингвистике — устойчивое словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого. Слово, которое сохраняет свое значение, называется ключевым, или свободным, компонентом. Слово, выбор которого определяется традицией, зависит от ключевого компонента и должен храниться в памяти (в словаре), называется несвободным компонентом. Большая часть коллокаций выражает ограниченное количество стандартных смыслов, названных в модели «Смысл — Текст» лексическими функциями (Мельчук 1974; Борисова 1995, 1996).

Коллокации по синтаксически главному слову делятся на глагольные и именные, по лексическому составу коллокации делятся на несоставные, незаменяемые и неизменяемые.

Также можно встретить название N-грамма — последовательность из N элементов. С семантической точки зрения это может быть последовательность звуков, слогов, слов или букв. На практике чаще встречается N-грамма как ряд слов. Последовательность из двух элементов часто называют биграммой, последовательность из трех элементов — триграммой. Последовательность не менее четырех и выше элементов обозначается как N-грамма, N заменяется на количество последовательных элементов (Hagberg, Schult, Swart 2008).

- частотные методы (Justeson, Katz 1995);
- подсчет среднего значения и дисперсии (Smadja 1993);
- c-value (Nazarenko, Zweigenbaum, Habert, Bouaud 2001);
- методы, основанные на теореме Байеса (Pearl 1985);
- T-критерий (или критерий Стьюдента) (Church, Hanks 1989);
- критерий хи-квадрат (Church, Gale 1991);
- метод отношения правдоподобия (Damerau 1993);
- взаимная информация (MI & PMI) (Church, Gale, Hanks, Hindle 1991).
- методы, основанные на мерах ассоциативности и машинном обучении (Хохлова, Еникеева 2020);
- методы, основанные на семантической близости и синтаксической связности (Seretan 2010);
- методы, основанные на нейронных сетях и векторных представлениях слов (Shao, Gouws, Britz, Goldie, Strope, Kurzweil 2017).

Коллокации играют важную роль в нашем языковом опыте и воздействуют на то, как мы выражаем свои мысли, передаем смысл и создаем концептуальные связи. Когнитивная лингвистика помогает нам лучше понять, как коллокации связаны с нашим пониманием мира и как они отражают наши концептуальные структуры.

Безусловно, основное отличие между методами статистической обработки текстов на естественном языке заключается в подходе к обработке и анализу текста. Важно отметить, что метод статистической обработки текстов на естественном языке основан на использовании статистических методов и алгоритмов для обработки текстовых данных. Он включает в себя анализ частотности слов, коллокаций (словосочетаний), распределения терминов и других статистических показателей. Метод статистической обработки текстов на естественном языке использует вероятностные модели и статистические методы для решения задач, таких как машинный перевод, распознавание речи, классификация текстов и др.

Статистическая обработка текстов на естественном языке опирается на методы статистики и теории вероятностей для выявления семантически связанных словосочетаний. Один из популярных методов — это анализ частотности совместного появления слов в корпусе текстов. Часто встречающиеся совместные появления слов указывают на семантическую связь между ними. Для анализа частотности применяют, например, методы на основе правил.

Модель *GPT*, с другой стороны, является моделью глубокого обучения, которая использует трансформерную архитектуру для генерации и понимания текстов на естественном языке. Модель *GPT* обучается на больших объемах текстовых данных и способна моделировать сложные языковые структуры и выражать семантические отношения между словами. Таким образом, модель *GPT* может генерировать текст, отвечать на вопросы, завершать предложения и выполнять другие задачи, связанные с естественным языком.

Трансформеры, такие как *GPT*, обучаются понимать семантические зависимости между словами на более глубоком уровне. Они анализируют контекст вокруг слова и используют механизм внимания для выделения важных словосочетаний. Такие модели способны выделять коллокации, которые могли бы быть упущены при поиске их более простыми статистическими методами.

Целесообразно утверждать, что, в то время как статистические методы на основе правил опираются на анализ частотности и вероятности, трансформеры способны более глубоко анализировать семантические связи и контекст слов в тексте, что делает их более мощными инструментами для работы с коллокациями.

В целом, на наш взгляд, различие между использованием методов статистической обработки естественного языка на основе правил и модели *GPT* заключается в методологии и подходе к анализу и пониманию текстовых данных на естественном языке. Оба подхода имеют свои уникальные особенности и применимы в различных сценариях и задачах, связанных с обработкой естественного языка. Было проведено исследование об автоматическом выделении коллокации в статье 5 Европейской конвенции о правах человека (ЕКПЧ)

на английском языке общим объемом 377 слов. Выбор материала обусловлен тем, что тексты юридической тематики характеризуются высокой степенью клишированности и обилием повторяющихся последовательностей часто встречающихся вместе слов (Garner 2001). Полученные данные дают возможность описать отличия в подходе к анализу и пониманию текстов на естественном языке в модели *GPT* и методе статистической обработки текстов на естественном языке.

Критерии, которые будут учитываться при проведении сравнительного анализа, следующие: время, затраченное на выполнение задачи, релевантность полученных результатов, количество потенциально значимых коллокаций, статистическая значимость, семантическая связанность, чувствительность контекста, синтаксическая связанность.

В качестве примера работы метода статистической обработки текстов на естественном языке была написана программа на языке программирования *Python* с использованием пакета библиотек и программ *NLTK* (*Natural Language Toolkit*) для символьной и статистической обработки естественного языка с применением метода *PMI*.

Для поиска коллокаций в тексте в *NLTK* используется мера статистической зависимости *PMI*. *PMI* означает *Pointwise Mutual Information* (точечная взаимная информация). *PMI* — это мера статистической зависимости между двумя словами в тексте. *PMI* описывается как «одна из наиболее важных концепций в *NLP*», которая «опирается на предположение, что лучший способ оценить связь между двумя словами — это выяснить, насколько часто эти слова встречаются в корпусе, по сравнению с тем, насколько мы априори ожидали бы их появления случайно» (Church, Gale 1991; Jurafsky, Martin 2021) (перевод наш. — Д. Г.).

*PMI* измеряет вероятность того, что два слова будут встречаться вместе (совместная вероятность), по сравнению с вероятностью их появления независимо друг от друга (маргинальная вероятность). Если значение *PMI* положительно, это указывает на то, что два слова часто встречаются вместе и они скорее связаны, чем независимы. Если значение *PMI* отрицательно или близко к нулю, это означает, что слова встречаются вместе примерно так же часто, как независимо друг от друга. Помимо этого, в библиотеке есть возможность использования таких инструментов, как *T-Score*, *Log-Likelihood*, *Chi-Square*.

*NLTK* является популярным решением в области обработки естественного языка и широко используется в научном сообществе. Использование популярных библиотек способствует воспроизводимости результатов и обеспечивает стандартизацию методов анализа. Выбор также обусловлен тем, что методы на основе мер ассоциации, которые применяются при статистической обработке,

являются относительно легкими в понимании и практической реализации, в них не используются нейросети и не требуются большие вычислительные мощности.

Представляется, что выбор *PMI* позволяет избежать некоторых ограничений, таких как чувствительность к объему текста. При работе с небольшим текстом применение сложных моделей машинного обучения может быть избыточным и даже привести к переобучению. Такие методы статистического анализа, как *PMI*, могут быть более постоянными и интерпретируемыми в контексте небольших корпусов.

*PMI* вычисляется как логарифм отношения фактической вероятности совместного появления двух слов к их ожидаемой вероятности при их независимости. Формула для вычисления *PMI* между словами $X$ и $Y$ на основе частоты их встречаемости в корпусе выглядит следующим образом:

$$PMI(X,Y) = \ln(\frac{P(X,Y)}{P(X)P(Y)}),$$

где:

$P$ $(X, Y)$ — вероятность встретить слова $X$ и $Y$ вместе (совместная вероятность),

$P$ $(X)$ — вероятность встретить слово $X$ (маргинальная вероятность),

$P$ $(Y)$ — вероятность встретить слово $Y$ (маргинальная вероятность),

*ln* — натуральный логарифм.

*NLTK* предоставляет метод nltk.collocations.PMI для вычисления *PMI* между словами на основе их встречаемости в тексте. Это может быть полезным для изучения семантических связей между словами, определения коллокаций и построения ассоциативных словесных связей (Sample Usage for Collocations 2023).

В *NLTK* также предоставляется возможность выполнить поиск коллокаций в тексте на основе обычной статистики методом nltk.collocations.raw_freq.

Результаты работы подобной программы отражены на Рис. 1, где слева представлены коллокации в виде биграмм, триграмм, четыреграмм, а справа — значение *PMI* для каждой коллокации.

```
(('deprived', 'of', 'his', 'liberty'), 15.739744592794525)

(('competent', 'legal', 'authority'), 14.44643717584693)

(('lawful', 'arrest', 'or', 'detention'), 13.917742894772516)

(('detention', 'of', 'a', 'person'), 12.887995551378467)

(('liberty', 'and', 'security'), 12.44643717584693)

(('prescribed', 'by', 'law'), 12.276512174404617)

(('shall', 'be', 'entitled'), 11.416689832452878)

(('purpose', 'of', 'bringing'), 10.639082253789326)

(('arrest', 'or', 'detention'), 9.308933652096993)

(('law', 'the', 'lawful'), 8.691549673683461)

[Finished in 754ms]
```

**Рис. 1.** Результаты работы программы статистической обработки текста на естественном языке

Были получены такие коллокации, как *deprived of his liberty* 15.739744592794525 *PMI*, *Competent legal authority* 14.44643717584693 *PMI*, *lawful arrest or detention* 13.917742894772516 *PMI*, *detention of a person* 12.887995551378467 *PMI*, *liberty and security* 12.44643717584693 *PMI* и др.

Стоит отметить, что результаты работы программы могут отличаться в зависимости от используемых методов фильтрации поиска коллокаций и других настроек, добавленных в программу разработчиком.

В данном случае при поиске коллокаций были исключены вхождения, которые появляются в тексте менее двух раз, и добавлены фильтры по таким стоп-словам, как *the*, *and*, *or*, *a*, *an*, *of*, *for* и т. д., так как они представляют минимальную исследовательскую ценность. Для биграмм требованием было, чтобы ни одно из этих слов не присутствовало в словосочетании. Для триграмм были запрещены словосочетания, в которых эти слова появляются в начале или конце. Аналогичный фильтр был применен для четыреграмм. Затем все полученные N-граммы были отсортированы по значению pmi. Данным условиям удовлетворили 10 словосочетаний из 377 слов. Программа завершила свою работу за 0,754 секунды.

Кажется очевидным, что для отображения корректных и объемных результатов необходим значительный объем исходных материалов, тем не менее метод *PMI* все равно может быть применен даже на материалах с незначительным объемом данных. В данном случае, чтобы получить больше результатов, представляется целесообразным ослабить требования к фильтрации, но тогда

велик риск получить нерелевантные результаты, которые будут включать единичные вхождения без доказательств их значимости как словосочетаний.

Тем не менее полученные результаты выглядят адекватными и представляющими исследовательскую ценность. Программа выполнила анализ с минимальным количеством ошибок.

Лишь один из представленных результатов — *law the lawful* 8.691549673683461 *PMI* — не имеет лексической ценности, так как метод не учитывает вхождения любых небуквенных символов. Слова *law* и *the lawful* являются частями двух разных предложений, хоть и следуют друг за другом. Программа рассматривает текст как набор следующих друг за другом токенов-слов без привязки к синтаксису, выполняемой ими роли и связи в предложении.

Отсюда, на наш взгляд, вытекает следующий недостаток. Если бы в программе учитывалась синтаксическая функция слов в предложении, то четырехграмму *lawful arrest or detention* следовало бы поделить на две биграммы *lawful arrest* и *lawful detention*, так как *lawful* относится к двум синтаксически главным словам и является частью двух именных словосочетаний.

Также могло бы быть целесообразным приводить глагольные и именные группы в начальную форму, как, например, в коллокации *deprived of his liberty*. Так как слова *deprive* и *liberty* имеют прочные предикативные связи, они часто употребляются вместе в аналогичной синтаксической связи, поэтому коллокацию можно разбить на более мелкие части без потери этих связей, например: *deprive of liberty* или *deprivation of liberty*.

Таким образом, возможно выделить следующие преимущества и недостатки метода *PMI*.

Преимущества:

1. Интерпретируемость. Метод статистической обработки текста на естественном языке основан на статистических методах и моделях, которые могут быть более интерпретируемыми. Результаты метода статистической обработки текста на естественном языке могут быть объяснены и интерпретированы с использованием статистических показателей и вероятностных моделей.

2. Более точное моделирование языковых структур. Метод статистической обработки текста на естественном языке позволяет явно моделировать языковые модели, такие как частотность слов, распределение терминов и коллокации. Это может привести к более точным результатам в определенных задачах, таких как извлечение информации и классификация текстов.

3. Требования к вычислительным ресурсам. Метод статистической обработки текста на естественном языке основан на заданных правилах и не требует значительных вычислительных ресурсов, что обеспечивает высокую скорость вычислений.

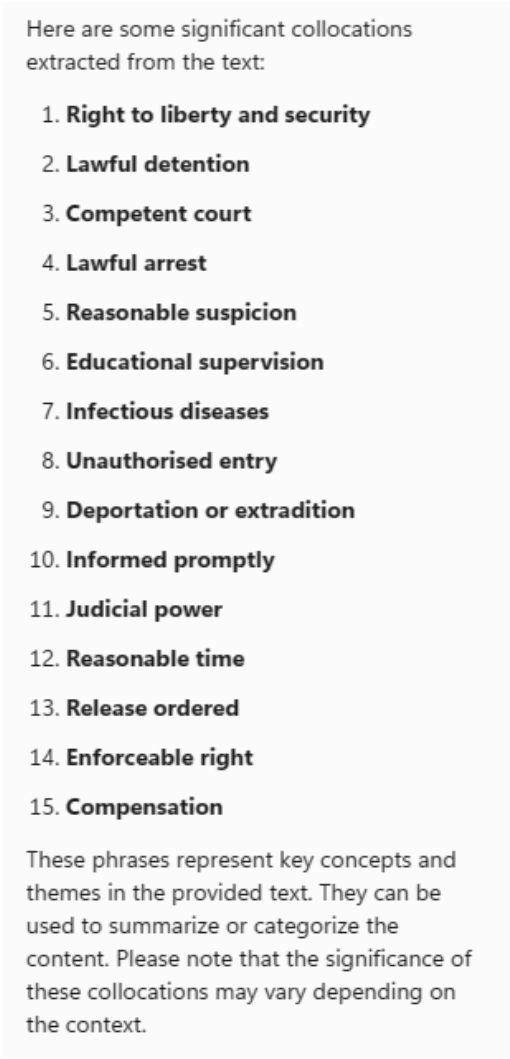
Недостатки:

1. Зависимость от качества данных. Метод статистической обработки текста на естественном языке требует больших и качественных наборов данных для своей работы. Если данные неполные или не представляют все разнообразие языка, то результаты статистической обработки могут быть ограниченными или неточными.

2. Ограничения в понимании семантики. Метод статистической обработки текстов на естественном языке может иметь ограничения в понимании семантических отношений между словами и выражении сложных смыслов. Он часто ориентирован на статистические показатели и может иметь трудности в интерпретации и понимании сложных контекстов без заранее продуманных сложных методов фильтрации.

Теперь рассмотрим, как с автоматическим поиском коллокаций справится модель генеративного предобученного трансформера, которая обучается понимать семантические зависимости между словами, анализировать контекст вокруг слова и использовать механизм внимания для выделения ключевых словосочетаний. В качестве примера работы модели *GPT* возьмем за основу популярную модель *GPT-4* в версии от компании *Microsoft*.

Ей была поставлена аналогичная задача. Запрос был следующий: «Извлеки из текста список значимых коллокаций. Предоставь ответ на исходном языке текста». На вычисления ей потребовалась 31 секунда, что почти в 41 раз дольше по сравнению с методом *PMI*. Результатами оказались следующие коллокации, которые модель *GPT* посчитала наиболее семантически связанными в тексте (см. Рис. 2).


Here are some significant collocations extracted from the text:

1. **Right to liberty and security**
2. **Lawful detention**
3. **Competent court**
4. **Lawful arrest**
5. **Reasonable suspicion**
6. **Educational supervision**
7. **Infectious diseases**
8. **Unauthorised entry**
9. **Deportation or extradition**
10. **Informed promptly**
11. **Judicial power**
12. **Reasonable time**
13. **Release ordered**
14. **Enforceable right**
15. **Compensation**

These phrases represent key concepts and themes in the provided text. They can be used to summarize or categorize the content. Please note that the significance of these collocations may vary depending on the context.

**Рис. 2.** Результаты работы GPT-4

Были получены такие коллокации, как *Right to liberty and security*, *Lawful detention*, *Competent court*, *Lawful arrest*, *Reasonable suspicion*, *Educational super-*

*vision, Infectious diseases* и др. без каких-либо статистических данных. При попытке попросить привести статистические показатели для данных коллокаций, по которым *GPT* определил их значимость, модель *GPT* привела следующие критерии: оценка частоты, оценка контекста, оценка семантической связанности, без конкретных деталей. Функция модели *GPT* заключается в генерации текстов на основе заданного контекста и понимании смысла и структуры предложений, но не в статистическом анализе текстов.

Тем не менее благодаря тому, что модель *GPT* обучается на больших объемах текстовых данных и способна моделировать сложные языковые структуры и выражать семантические отношения между словами, а также может обучаться на ограниченном числе примеров, она смогла извлечь из ограниченного объема материалов значимые коллокации, которые можно использовать для дальнейшего анализа.

Механизм внимания позволяет модели сосредотачиваться на разных частях входных данных (каждое слово обрабатывается независимо) и выделять те части текста, которые имеют наибольшую ценность для понимания.

Сразу заметен прирост в результате до 15 выделенных моделью коллокаций, что в полтора раза больше, чем результат программы, которая использует метод *PMI* для поиска коллокаций в ограниченном объеме входных данных.

Также модель корректно учитывает синтаксические связи между словами и в аналогичной ситуации правильно выделила две отдельные биграммы *lawful arrest* и *lawful detention*.

Так как модель специализируется на генерации естественного языка, поиск предиктивных связей между словами в тексте является одной из самых сильных ее сторон, что также повлияло на выбор наиболее связанных в тексте коллокаций и эффективность поиска.

Также наблюдаются некоторые различия в результатах применения этих двух методов, которые можно было бы дополнить, например добавив коллокацию *prescribed by law* из результатов статистического подхода *NLP*.

Таким образом, были выделены следующие преимущества и недостатки модели генеративного предобученного трансформера.

Преимущества:

1. Генеративная способность. Модель *GPT* способна генерировать качественный текст, который может быть использован в различных задачах, таких как генерация статей, завершение предложений, ответы на вопросы и мн. др., благодаря развитой предиктивной функции.

2. Понимание семантики. Модель *GPT* обучается на больших объемах текстовых данных и обладает способностью понимать семантические отношения между словами и выражать сложные языковые структуры, что позволяет ей

генерировать связный и информативный текст. Она может выполнять задачи, требующие понимания и интерпретации текстов на естественном языке, такие как определение тематики текста и синтаксический анализ.

3. Гибкость и адаптивность. Модель *GPT* является обучаемой моделью и может быть адаптирована к конкретным задачам. Она способна обучаться на специфических данных и адаптироваться к изменяющимся условиям. Это позволяет использовать модель *GPT* для разных задач, требующих обработки и понимания естественного языка.

Недостатки:

1. Ограниченность. Модель *GPT* может не всегда точно понимать контекст и выражать семантические отношения. Это может привести к ошибкам в генерации текста или неправильному пониманию запросов. Несмотря на значительные достижения, модель *GPT* все еще имеет ограничения в понимании сложных контекстов и специфических тематик.

2. Требования к вычислительным ресурсам. Модель *GPT* является моделью с глубоким обучением, требующей значительных вычислительных ресурсов для обучения и работы. Это может ограничить использование модели GPT на устройствах с недостаточными вычислительными мощностями или в условиях с ограниченным доступом к вычислительным ресурсам.

## 4. Заключение

Оба подхода имеют свои преимущества и недостатки, и выбор между моделью *GPT* и методом *PMI* зависит от конкретных задач, доступных ресурсов и требуемых результатов. В некоторых случаях комбинирование этих подходов может привести к лучшим результатам в обработке и пониманию текстов на естественном языке.

### Список литературы / References

*Борисова Е. Г.* Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка. М.: Филология, 1995. [Borisova, Elena G. (1995) *Slovo v tekste. Slovar' kollokacij (ustojchivyh sochetanij) russkogo jazyka* (Word in the Text. Dictionary of Collocations (Stable Combinations) of the Russian Language). Moscow: Filologiya. (In Russian)].

*Борисова Е. Г.* Коллокации. Что это такое и как их изучать? М.: Филология, 1996. [Borisova, Elena G. (1996) *Kollokacii. Chto jeto takoe i kak ih izuchat'?* (Collocations. What are they and how do we study them?). Moscow: Filologiya. (In Russian)].

*Мельчук И. А.* Опыт теории лингвистических моделей «Смысл — Текст». М.: Наука, 1974. [Mel'chuk, Igor' A. (1974) *Opyt teorii lingvisticheskih modelej «Smysl — Tekst»* (Experience of the Theory of Linguistic Models "Meaning — Text"). Moscow: Nauka. (In Russian)].

*Хохлова М. В., Еникеева Е. В.* Методы машинного обучения применительно к задаче выделения глагольных и атрибутивных коллокаций // Компьютерная лингвистика и вычислительные онтологии. Вып. 4. (Труды XXIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2020, Санкт-Петербург, 17–20 июня 2020 г. Сборник научных статей). СПб.: Университет ИТМО, 2020. С. 54–60. [Khokhlova, Mariya V., & Enikeeva, Ekaterina V. (2020) Metody mashinnogo obucheniya primenitel'no k zadache vydeleniya glagol'nykh i atributivnykh kollokatsiy (Applying Machine Learning Methods to Verbal and Noun Phrases Extraction). In *Komp'yuternaya lingvistika i vychislitel'nye ontologii*. Vyp. 4. (Trudy XXIII Mezhdunarodnoj ob''dinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2020, Sankt-Peterburg, 17–20 iyunya 2020 g. Sbornik nauchnyh statej) (Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference "Internet and Modern Society", IMS-2020, St. Petersburg, June 17–20, 2020)). Saint Petersburg: ITMO University, 54–60. DOI 10.17586/2541-9781-2020-4-54-60. (In Russian)].

Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой; Институт языкознания АН СССР. М.: Советская энциклопедия, 1990. [*Lingvisticheskij jenciklopedicheskij slovar'* (1990) / Pod red. V. N. Yarcevoj; Institut jazykoznanija AN SSSR (Linguistic Encyclopedic Dictionary/ In Yartseva, Viktoriya N. (ed.); Institute of Linguistics, Academy of Sciences of the USSR). Moscow: Sovetskaja jenciklopedija. (In Russian)].

Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, et al. (2020) *Language Models Are Few-Shot Learners*. Retrieved from arXiv preprint arXiv:2005.14165. (2023, Jule 27).

Cambria, Erik, & White, Bebo. (2014) Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, 9 (2), 48–57.

Choueka, Yaacov, Klein, Shmuel T., & Neuwitz, E. (1983) Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. *Association of Literary and Linguistic Computing (ALLC) Journal*, 4, 34–38.

Church, Kenneth W., & Hanks, Patrick. (1989) Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 76–83.

Church, Kenneth W., Gale, William A., Hanks, Patrick, & Hindle, Don. (1991) Using Statistics in Lexical Analysis. In Zernik, Uri. (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum, 115–164.

Church, Kenneth W., & Gale, William A. (1991) Concordances for Parallel Text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*. Oxford, 40–62.

Covington, Michael A. (1993) *Natural Language Processing for Prolog Programmers*. Englewood Cliffs, NJ: Prentice-Hall.

Damerau, Fred J. (1993) Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Information Processing & Management*, 29 (4), 433–447.

Dyer, Michael. (1995) Connectionist Natural Language Processing: A Status Report. In Sun, Ron, & Bookman, Lawrence. (eds.) *Computational Architectures Integrating Neural and Symbolic Processes*. Dordrecht, The Netherlands: Kluwer Academic, 389–429.

Garner, Bryan A. (2001) *Legal Writing in Plain English: A Text with Exercises*. Chicago: University of Chicago Press.

Hagberg, Aric A., Schult, Daniel A., & Swart, Pieter J. (2008) Exploring Network Structure, Dynamics, and Function Using NetworkX. In Varoquaux, Gael, Vaught, Travis, & Millman,

Jarrod. (eds.) *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, 11–15. Retrieved from https://proceedings.scipy.org/articles/PFVC8793. (2023, Jule 27).

Jozefowicz, Rafal, Vinyals, Oriol, Schuster, Mike, Shazeer, Noam, & Wu, Yonghui. (2016) Exploring the Limits of Language Modeling. Retrieved from arXiv preprint arXiv:1602.02410. (2023, Jule 27).

Jurafsky, Daniel, & Martin, James H. (2021) Speech and Language Processing (3rd ed., draft). Chapter 6. Retrieved from https://web.stanford.edu/~jurafsky/slp3/. (2024, December 03)

Justeson, John S., & Katz, Slava M. (1995) Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, 1 (1), 9–27.

Nazarenko, Adeline, Zweigenbaum, Pierre, Habert, Benoît, & Bouaud, Jacques. (2001) Corpus-Based Extension of a Terminological Semantic Lexicon. In Bourigault, Didier, Jacquemin, Christian, & L'Homme, Marie-Claude (eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, 327–351.

Pearl, Judea. (1985) Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, Vol. 7, 329–334.

Radford, Alec, Narasimhan, Karthik, Salimans, Tim, & Sutskever, Ilya. (2018) *Improving Language Understanding by Generative Pre-Training*. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. (2023, Jule 27).

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, & Sutskever, Ilya. (2019) *Language Models are Unsupervised Multitask Learners*. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. (2023, Jule 27).

Sample Usage for Collocations. *NLTK Project*. Retrieved from https://www.nltk.org/howto/collocations.html. (2023, Jule 27).

Seretan, Violeta. (2010) *Syntax-Based Collocation Extraction* (1st ed.). Berlin, Heidelberg: Springer-Verlag.

Shao, Louis, Gouws, Stephan, Britz, Denny, Goldie, Anna, Strope, Brian, & Kurzweil, Ray. (2017) Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2210–2219.

Smadja, Frank. (1991) From N-Grams to Collocations: An Evaluation of Xtract. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA: Association for Computational Linguistics, 279–284.

Smadja, Frank. (1993) Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19 (1), 143–178.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., et al. (2017) Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*. NY: Curran Associates Inc, 5999–6010.

Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V., Norouzi, Mohammad, Macherey, Wolfgang, et al. (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Retrieved from arXiv preprint arXiv:1609.08144. (2023, Jule 27).