DOI: 10.5281/zenodo.15321591
Link: https://zenodo.org/records/15321591

АЛГОРИТМЫ СИНТАКСИЧЕСКОГО И СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Арипова Гулчехра Ишанкуловна
Преподаватель кафедры «Современные информационные технологи»
Узбекский государственный университет мировых языков
gulchehraaripova2020@gmail.com

Аннотация. *В данной статье рассматриваются современные алгоритмы, применяемые для синтаксического и семантического анализа текстов в рамках задач автоматической обработки естественного языка. Особое внимание уделено сравнению традиционных формальных методов синтаксического анализа (таких как нисходящий и восходящий парсинг) с нейросетевыми моделями нового поколения. Также исследуются подходы к семантическому интерпретированию текста, включая метод распределённых представлений слов и трансформерные архитектуры. Проанализированы сильные и слабые стороны различных алгоритмических решений, а также их применимость в задачах машинного перевода, извлечения информации и диалоговых систем. Статья поднимает вопросы эффективности, интерпретируемости и вычислительной сложности алгоритмов, что актуально в контексте стремительно растущих объёмов текстовой информации.*

Ключевые слова. *синтаксический анализ, семантический анализ, компьютерная лингвистика, обработка естественного языка, нейросетевые модели, трансформеры, парсинг, представление смысла.*

Введение

Развитие цифровых технологий и стремительный рост объёмов текстовой информации в интернете, научных базах и пользовательских данных обуславливают необходимость эффективной автоматической обработки естественного языка. Одним из ключевых направлений в области компьютерной лингвистики и искусственного интеллекта является анализ структуры и смысла текстов — синтаксический и семантический анализ, обеспечивающие интерпретацию языковых данных машинами.

Синтаксический анализ ориентирован на определение грамматической структуры предложения, включая выявление связей между словами и построение дерева зависимостей. Семантический же анализ направлен на извлечение смысла, стоящего за синтаксической формой, и включает в себя задачи интерпретации лексических значений, контекстов и намерений. Сочетание этих двух видов анализа лежит в основе таких прикладных задач, как автоматический перевод, извлечение фактов, ответы на вопросы, генерация текстов и ведение интеллектуальных диалогов.

Современные алгоритмы синтаксического и семантического анализа претерпели значительную трансформацию: от формальных грамматик и алгоритмов парсинга к глубоким нейросетевым моделям, таким как трансформеры и языковые модели последнего поколения. Эти алгоритмы позволяют моделировать не только грамматические зависимости, но и контекстуальные связи, важные для адекватной интерпретации текста.

Настоящая статья посвящена систематическому обзору и сравнительному анализу существующих подходов к синтаксическому и семантическому анализу текста. Особое внимание уделяется алгоритмам, находящим применение в системах обработки естественного языка, их архитектуре, эффективности и перспективам дальнейшего развития.

СИНТАКСИЧЕСКИЙ АНАЛИЗ ТЕКСТА: АЛГОРИТМЫ И МОДЕЛИ

Синтаксический анализ — одно из ключевых направлений в области автоматической обработки естественного языка, связанное с установлением грамматической структуры предложений. В широком смысле под синтаксическим анализом понимается процедура сопоставления последовательности слов (лексем) с грамматическими правилами определённого языка, что позволяет установить корректность построения фраз и выстроить формализованное представление их внутренней структуры. Единицей синтаксического анализа, как правило, выступает отдельное предложение.

С одной стороны, синтаксический анализ решает задачу распознавания: определяет, соответствует ли данное предложение нормам грамматики конкретного языка. Например, в английском языке артикль располагается перед определяемым словом, тогда как в русском языке прилагательные должны согласовываться с существительными в роде, числе и падеже. Эти функции находят применение в системах автоматической проверки грамматики в текстовых редакторах и языковых помощниках.

С другой стороны, и это принципиально важно для задач интерпретации, синтаксический анализ служит промежуточным этапом построения смысловой структуры текста. Он формирует основу, на которой в дальнейшем осуществляется семантический анализ, связывающий лексические единицы с их значениями в предметной области. При этом синтаксическая структура не содержит прямых интерпретаций значений, но обеспечивает необходимую опору для таких процедур, как разрешение синтаксических неоднозначностей, анализ актантной структуры и идентификация отношений между составляющими текста [1,2].

В компьютерной лингвистике разработано два основных подхода к синтаксическому моделированию: грамматика составляющих и грамматика зависимостей. Первый подход, получивший развитие в трудах Н. Хомского [3], представляет предложение в виде иерархии вложенных фразовых единиц, формирующих дерево составляющих. Эта модель предполагает, что каждое предложение можно разложить на фразы, которые в свою очередь могут включать в себя подфразы и отдельные слова. Основным свойством дерева составляющих является проективность — отсутствие пересечений между подструктурами.

Альтернативой является грамматика зависимостей, заложенная в работах Л. Теньера [4] и И. Мельчука [5], где акцент делается не на группах, а на отношениях между словами — вершиной и зависимыми элементами. Такая

структура ближе к логической интерпретации высказывания и часто используется в современных NLP-системах, ориентированных на семантический разбор.

В зависимости от используемой модели синтаксический анализ может быть реализован с применением различных алгоритмов парсинга: нисходящих (top-down), восходящих (bottom-up), или комбинированных стратегий. Выбор конкретного алгоритма зависит от целей анализа, требований к скорости и точности, а также сложности обрабатываемого языка.

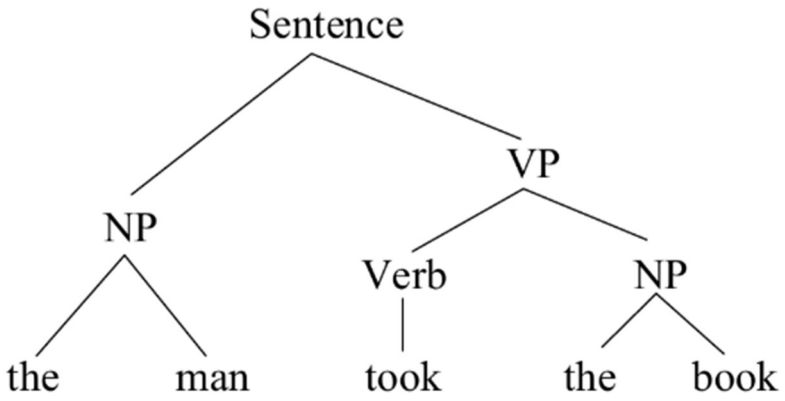

Рис 1. Первое дерево составляющих, из работы Н. Хомского [10]

Идея о том, что в структуре предложения определённые группы слов образуют единые синтаксические составляющие, основывается на наблюдении, что такие группы функционируют как грамматически связанные блоки. Эти группы, или фразовые единицы, обладают целостностью: они могут перемещаться внутри предложения, заменяться или подвергаться трансформации, сохраняя при этом внутреннюю структурную связанность. Попытка изменить только часть такой группы, как правило, нарушает грамматическую или смысловую целостность высказывания.

Грамматика составляющих, лежащая в основе формализма Хомского, представляет собой контекстно-свободную грамматику (КС-грамматику). Такая грамматическая модель позволяет строить синтаксические деревья, отражающие вложенность и иерархическую организацию составляющих. Хотя КС-грамматики имеют ограничения, они обеспечивают достаточный уровень описательной мощности для большинства конструкций естественного языка и активно используются в практике обработки текстов.

Считается доказанным, что естественные языки, включая английский, не поддаются описанию в терминах регулярных грамматик, поскольку требуют более сложных правил согласования и вложенности. В ряде исследований показано, что существуют языковые явления, выходящие за пределы КС-грамматик. Например, в швейцарском немецком наблюдаются конструкции, требующие применения контекстно-зависимых грамматик для их корректного синтаксического анализа. Тем не менее, несмотря на эти теоретические ограничения, КС-грамматики остаются одним из наиболее распространённых

инструментов формализации грамматических правил в задачах автоматического синтаксического анализа, поскольку обеспечивают приемлемый баланс между вычислительной сложностью и описательной точностью.

МЕТОДЫ ПОСТРОЕНИЯ СИНТАКСИЧЕСКИХ ДЕРЕВЬЕВ СОСТАВЛЯЮЩИХ

Так как естественные языки в ряде случаев могут быть эффективно аппроксимированы с использованием контекстно-свободных грамматик, для синтаксического анализа текстов применяются соответствующие формальные методы и алгоритмы, разработанные в рамках теории формальных языков. Однако, в отличие от искусственных языков, таких как языки программирования, для которых синтаксический разбор, как правило, может быть осуществлён за линейное или почти линейное время, обработка естественного языка сопровождается высокой степенью неоднозначности и усложнённой грамматической структурой.

Одной из ключевых проблем при синтаксическом анализе естественного языка является неоднозначность, которая может проявляться на различных уровнях — от морфологического до структурного. На этапе морфологического анализа омонимия может затруднить выбор корректной части речи, падежа или лексического значения, что, в свою очередь, влияет на синтаксическую интерпретацию. При этом контекстно-свободные грамматики, используемые в синтаксическом анализе, часто содержат множество недетерминированных правил, допускающих альтернативные структуры разбора, что приводит к необходимости возвратов в алгоритмах и существенно снижает их вычислительную эффективность.

Особенно остро проблема встаёт при наличии синтаксической (структурной) омонимии, когда одно и то же предложение допускает несколько корректных с точки зрения грамматики вариантов анализа. В ряде случаев даже семантический контекст не позволяет однозначно выбрать предпочтительный вариант. Классический пример — предложение «Мы встречали поэта из Грузии», в котором возможны как трактовка «мы (встречали поэта) из Грузии», так и «мы встречали (поэта из Грузии)». Обе интерпретации соответствуют грамматическим и семантическим нормам, что демонстрирует сложность автоматической обработки даже при использовании формальных моделей.

Проблема неоднозначности привела к развитию целого спектра методов построения деревьев составляющих, среди которых можно выделить:

– рекурсивный нисходящий парсинг (top-down parsing) — метод, строящий дерево от корня к листьям, подвержен проблемам возвратов;

– восходящий парсинг (bottom-up parsing) — строит дерево от слов к корневому узлу, эффективен при использовании таблиц предсказаний;

– алгоритмы Эрли, CYK и Глена — универсальные алгоритмы анализа, применимые к произвольным КС-грамматикам, особенно актуальны при высокой степени неоднозначности;

– статистические и нейросетевые модели — сочетают парсинг с вероятностной оценкой вариантов разбора, что повышает точность за счёт обучения на корпусах.

Для повышения качества синтаксического анализа нередко используется дополнительная семантическая информация, позволяющая исключать грамматически возможные, но смыслово маловероятные интерпретации. Тем самым синтаксический анализ всё чаще рассматривается в контексте его интеграции с семантическими компонентами обработки текста.

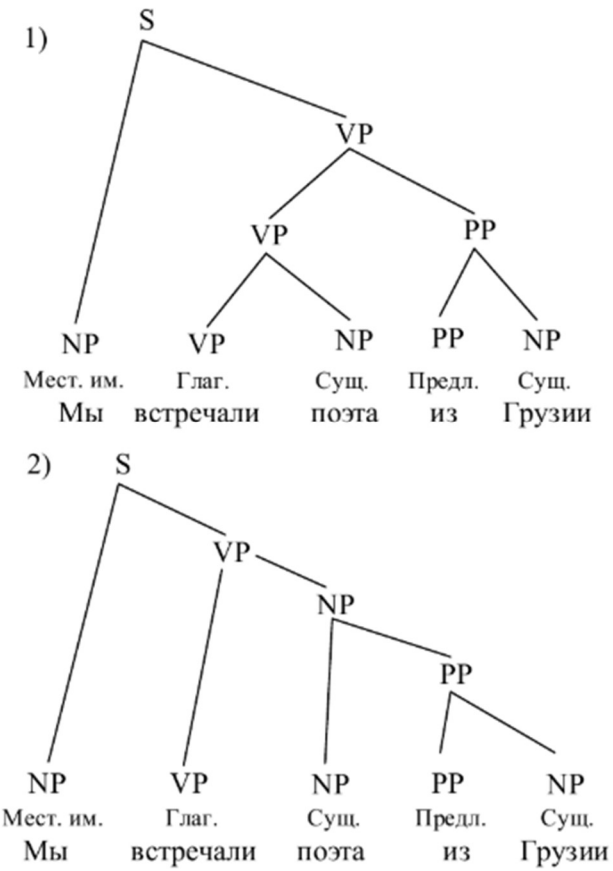

Рис 2. Пример двух вариантов синтаксического разбора предложения «Мы встречали поэта из Грузии», допустимых с точки зрения семантики

Одним из классических методов построения синтаксических деревьев в рамках контекстно-свободной грамматики является алгоритм Кокса–Янгера–Касами (CYK), представляющий собой универсальный восходящий парсер, основанный на динамическом программировании. Он применим к грамматикам, приведённым к нормальной форме Хомского (НФХ), в которой все правила имеют строго определённый формат: либо разложение нетерминала на два нетерминала, либо на один терминал.

Алгоритм CYK строит таблицу, в ячейки которой заносятся нетерминалы, способные породить соответствующие фрагменты исходного предложения. Разбор производится снизу вверх, начиная с самых коротких отрезков (обычно отдельных слов) и постепенно переходя к более длинным комбинациям. Каждая

ячейка таблицы анализируется с учётом содержимого ячеек, расположенных ниже и левее, а также с опорой на правила грамматики. Если в процессе работы в правой верхней ячейке оказывается начальный символ грамматики (обычно обозначаемый как S), то предложение считается успешно распознанным.

Однако применение алгоритма CYK имеет свои ограничения. Главный из них — необходимость приведения исходной КС-грамматики к нормальной форме Хомского, что может приводить к значительному увеличению числа правил и потере читаемости исходной лингвистической структуры. Кроме того, чтобы восстановить дерево синтаксических составляющих в терминах оригинальной грамматики, необходимо выполнять обратное преобразование, что требует дополнительных вычислительных ресурсов и увеличивает требования к памяти.

Тем не менее, несмотря на вычислительную затратность и формальные ограничения, алгоритм CYK остаётся одним из базовых методов синтаксического анализа, особенно в системах, где важна строгая формализация и детерминированность процесса синтаксического вывода. Он также применяется в образовательных целях и для тестирования корректности грамматик.

Другим важным подходом, основанным на методах динамического программирования, является алгоритм Эрли, предложенный Джеем Эрли как универсальный способ синтаксического анализа для произвольных контекстно-свободных грамматик. В отличие от алгоритма CYK, работающего снизу вверх и требующего строгого соблюдения нормальной формы Хомского, алгоритм Эрли выполняет анализ сверху вниз и более гибок в применении к грамматикам в естественном виде.

Ключевая особенность алгоритма Эрли заключается в использовании таблицы состояний, где каждое состояние представляет собой «правило с точкой» — запись грамматического правила, в которой точка указывает, насколько правило уже «выполнено» для текущей позиции в строке. Алгоритм пошагово анализирует предложение слева направо, и в зависимости от текущего состояния выполняет три возможных действия:

– предсказание — добавление новых состояний на основе правил, ожидающих нетерминал после точки;

– сканирование — сопоставление терминалов с текущим словом входной строки;

– завершение (completion) — переход к следующему состоянию, если правило полностью распознано.

Разбор считается завершённым, если в конце таблицы содержится состояние, соответствующее полному правилу для начального нетерминала грамматики (S). Алгоритм Эрли отличается высокой универсальностью и применяется в ситуациях, где требуется точный анализ с возможностью обработки неоднозначных конструкций без предварительного преобразования грамматики.

Развивая идеи универсальных парсеров, Мартин Кей предложил обобщённую модель построения синтаксического анализа, названную chart parsing. В этой схеме синтаксический анализ представляется как процесс построения графа, содержащего частично сформированные поддеревья. Основное достоинство chart parsing заключается в возможности комбинировать различные стратегии разбора — как снизу вверх (по аналогии с CYK), так и сверху вниз (по аналогии с Эрли). Путём варьирования порядка применения операций и добавления эвристических правил система может адаптироваться под конкретные задачи, повышая эффективность и точность разбора.

ЗАКЛЮЧЕНИЕ

Одной из современных и эффективных моделей, применяемых в задачах семантико-синтаксического анализа, является дерево зависимостей, представляющее структуру предложения в виде направленного графа, в котором слова связаны между собой отношениями зависимости. В отличие от деревьев составляющих, модель зависимостей более точно отражает синтаксические связи между словами, особенно в языках с гибким порядком слов, таких как русский, где линейное расположение элементов не всегда отражает грамматическую иерархию.

Выбор именно зависимостной модели обусловлен как её выразительностью, так и высокой вычислительной эффективностью. В частности, алгоритм Нивре (Nivre algorithm) реализует построение дерева зависимостей за линейное время O(n), что является значительным преимуществом по сравнению с классическими алгоритмами анализа на основе КС-грамматик, которые имеют вычислительную сложность $O(n^3)$ в худшем случае. Это позволяет применять зависимостной анализ в системах с жёсткими требованиями к скорости обработки и большим объёмом входных данных.

В последние годы наблюдается смещение фокуса исследований с разработки формализованных грамматик к построению моделей, основанных на машинном обучении. Такие модели обучаются на размеченных корпусах текстов, в которых каждому слову и паре слов приписаны определённые синтаксические или семантические отношения. Особенно актуальным это стало благодаря наличию больших синтаксически размеченных корпусов, таких как СинТагРус — репрезентативный корпус русского языка, содержащий обширный массив данных с указанием грамматических и синтаксических связей.

В рамках данной статьи в качестве инструмента синтаксического анализа предлагается использование MaltParser — одного из наиболее эффективных обучаемых анализаторов, способного адаптироваться под специфику конкретного языка и жанра текста. Его эффективность подтверждается широким применением в международных проектах по обработке естественного языка, а также высокой точностью при синтаксическом парсинге.

С точки зрения семантической интерпретации предложений, в данной работе в качестве основы выбрана модель ролевой семантики (semantic role labeling, SRL). Эта модель опирается на идею представления значения

предложения через идентификацию ролей участников ситуации (агент, пациент, инструмент и т. д.). Несмотря на относительную концептуальную простоту, ролевая структура оказалась чрезвычайно полезной для решения широкого спектра прикладных задач, включая извлечение информации, автоматическое аннотирование текста и построение логических моделей содержания.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Iomdin L., Petrochenkov V., Sizov V., Tsinman L. ETAP parser: state of the art // Papers from the Annual International Conference "Dialogue" (2012). — 2012. — С. 830–853.
2. Anisimovich K. V., Druzhkin K. Ju., Minlos F. R. и др. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies // Papers from the Annual International Conference "Dialogue". — 2012. — Т. 2. — С. 91–103.
3. Chomsky N. Three models for the description of language // IRE Transactions on Information Theory. — 1956. — Т. 2, № 3. — С. 113–124.
4. Tesnière L. Elements de syntaxe structurale. — Paris: Editions Klincksieck, 1959.
5. Mel’cuk I. A. Dependency syntax: theory and practice. — Albany: SUNY Press, 1988. — 428 p.
6. Jurafsky D., Martin J. H. Speech and Language Processing. — 3rd ed. — Draft, Stanford University, 2023. — 1250 p. — [Электронный ресурс]. — Режим доступа: https://web.stanford.edu/~jurafsky/slp3/.
7. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. — Cambridge: MIT Press, 1999. — 680 p.
8. Goldberg Y. Neural Network Methods for Natural Language Processing. — Morgan & Claypool, 2017. — 309 p.
9. Klein D., Manning C. D. Accurate unlexicalized parsing // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL). — 2003. — P. 423–430.
10. Vaswani A. et al. Attention is all you need // Advances in Neural Information Processing Systems. — 2017. — Vol. 30. — P. 5998–6008.
11. Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. — 2013. — Vol. 26. — P. 3111–3119.
12. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. — 2018. — 15 p. — [Электронный ресурс]. — Режим доступа: https://arxiv.org/abs/1810.04805.
13. Bojar O., Chatterjee R., Federmann C. et al. Findings of the 2017 Conference on Machine Translation (WMT17) // Proceedings of the Second Conference on Machine Translation. — 2017. — Vol. 2. — P. 169–214.