Вкладка 1

#### 

[1. Естественно-языковой интерфейс: определение, характеристики, примеры. 3](#_kue3nbxl0vzv)

[2. Задачи создания естественно-языкового интерфейса. 4](#_ga3t0uth3naj)

[3. Отличия естественно-языкового интерфейса от других видов интерфейса. 4](#_qm4vhv2mll0z)

[4. Компьютерная лингвистика: определение, задачи, направления. 5](#_o1xzbhrpu03)

[5. Прикладная лингвистика: определение, сходства и различия с компьютерной лингвистикой. 7](#_7b42hh2fvsok)

[6. Причины возникновения компьютерной лингвистики. 7](#_b1pctxl6c6pa)

[7. Отрасли знаний и научные направления, связанные с компьютерной лингвистикой и проблемой разработки естественно-языкового интерфейса. 8](#_t81a5ih1vmo8)

[8. Понятия «общение», «диалог», «язык», их взаимосвязь друг с другом и с проблемой создания естественно-языкового интерфейса. 9](#_2vaefd9nifde)

[10. Правила и способы общения. Примеры реализации диалога человека с компьютерной системой. 10](#_4p2q93mds8ez)

[11. Диалоговая система: определение, компоненты. Проблемы, возникающие в процессе реализации естественно-языкового диалога человека и компьютерной системы. 11](#_l43nfyre4b06)

[12. Моделирование диалога. 13](#_snmrm84mn8nq)

[**13. Понятие «язык»: определение, проблемы формализации. 14**](#_d45gbxwbs7bg)

[14. Слово, словоформа, лексема. Примеры. 15](#_lfusbqb5qrpi)

[15. Функции языка. Виды языков. 16](#_zi4qmdeh159i)

[16. Уровни обеспечения систем машинного перевода текстов 16](#_ft0yhrc53bst)

[17. Основные понятия семиотики. 17](#_t5a76zwjyelx)

[18. Знаковая система: определение, свойства, примеры. 18](#_77t6uqyurboz)

[19. Омонимия: определение, виды, примеры, достоинства и недостатки. 20](#_vhcpjalzjk9g)

[20. Синонимия: определение, виды, примеры, достоинства и недостатки. 20](#_cl399nemd1pc)

[21. Автоматизация обработки текста: этап семантического анализа. 21](#_y33e07bu2cjs)

[22. Типовая структура базы знаний для решения задач автоматической обработки текста естественного языка. 22](#_5aak04mnktmn)

[24. Уровни изучения текста, связь с разделами лингвистики. 24](#_u4sp9ej1lrg4)

[25. Задача информационного поиска. 24](#_z58druf3bxvv)

[26. Задача информационного поиска. Документальные системы информационного поиска. 25](#_5u574cwviw9o)

[27. Задача информационного поиска. Фактографические системы информационного поиска. 26](#_kosput31ln8b)

[28. Документальные системы информационного поиска. Поисковая система Google. 26](#_f5ilqjjb6jvu)

[29. Проблемы машинного перевода. 28](#_ihtae023bfrj)

[30. Системы непрямого машинного перевода. Их лингвистическое обеспечение. 29](#_3adc59ffe3za)

[31. Автоматизация обработки текста: этап синтаксического анализа. 30](#_t1m6bm8toqw)

[32. Задача автоматического реферирования документов. Метод построения классического реферата. 31](#_ktr26ko2ku2o)

[33. Задача информационного поиска. Структурно-функциональная схема системы информационного поиска. 32](#_qeu1y1gxhaiz)

[34. Задача информационного поиска. Метапоисковые средства поиска, особенности построения и функционирования. 35](#_fh4gt3q6qtde)

[35. Задача информационного поиска. Метапоисковые утилиты, особенности функционирования. 36](#_y7jp5thpte84)

[37. Документальные системы информационного поиска. Поисковая система Yandex. 38](#_bacw5fk2t97b)

[38. Задача автоматического реферирования документов, подходы к решению. 39](#_txchgphpgsir)

[39. Задача информационного поиска. Поиск по запросу булевого типа. 40](#_clq91ssmmz2o)

[**41. Задача информационного поиска. Векторная модель индексирования и поиска документов. 41**](#_njh55s56lbk5)

[42. Способы описания синтаксической структуры предложения. Достоинства и недостатки. 42](#_w3nenytcjn9q)

[43. Системы составляющих. 44](#_g2g7rq6dxwe9)

[44. Деревья подчинения. 45](#_mdfwzmkdf6qr)

[46. Модель СМЫСЛ – ТЕКСТ: определение, компоненты, задачи. 47](#_mhvq8wm6ypdw)

[47. Задача автоматического реферирования документов. Sentence extraction - особенности подхода. 48](#_5r9w0nglckhm)

[48. Задача автоматического реферирования документов. Summary generation - особенности подхода. 49](#_s6lx2ut94vw)

[49. Автоматизация обработки текста: этапы преформатирования текста, распознавание границ слов и предложений. 50](#_m0y84qeyqyov)

[50. Задача информационного поиска. Статистические закономерности в текстах естественного языка. Законы Дж. Зипфа. 51](#_k3zzx5c2irz)

[51. Задача информационного поиска. Хранение индексных структур: прямой и инвертированный индексы. 53](#_9wxcrlkem4gr)

[52. Задача информационного поиска. Индексирование текстовых документов. 54](#_u8s62egf5zfr)

[53. Задача информационного поиска. Способы оценки эффективности систем информационного поиска. 55](#_xi9bzjpyi3eg)

[54. Речевой интерфейс: определение, задачи. 55](#_w24v62mx0kav)

[**55. Поисковая оптимизация. Назначение и методы реализации. 56**](#_k3wm7wy8gdkj)

[**56. Обобщенная структура системы автоматического синтеза речи. 57**](#_p923u1tvd7e1)

[57. Постановка задачи автоматического синтеза речи. 60](#_v6kp6b5kgui0)

[58. Поисковая оптимизация. Факторы, влияющие на ранжирование веб-ресурсов. 61](#_gyyecs63n04g)

[**59. Обобщенная структура системы автоматического распознавания речи. 62**](#_ktct6qq4rbl1)

[60. Постановка задачи автоматического распознавания речи. 63](#_9v456t8ay0e3)

[62. Задача автоматического распознавания языка текста. Метод частотных слов, достоинства и недостатки. 65](#_lrxkwbwo4hci)

[63. Задача автоматического распознавания языка текста. Нейросетевой метод, достоинства и недостатки. 66](#_bimu0vbhilos)

[64. Структура диалоговой системы. бил 11 67](#_s4d8dgapivjy)

[65. Направления использования речевых систем. 68](#_bc8gtau6c32s)

[**66. Характеристики речевого сигнала. 68**](#_4iacujpdxib1)

[**67. Фонемы в речи. Фонетический состав русского языка. 70**](#_h993n259bv4o)

[**68. Аллофон: определение, виды, отличия от фонемы. 71**](#_qc1idatl2jah)

[**69. Использование различных разделов лингвистики при реализации речевой системы. 72**](#_7gq8wg3bxvtn)

[**70. Фонетика. Просодика. 73**](#_usqdrytwsquy)

[71. Артикуляторные особенности акустических свойств фонем. 74](#_gxq73vn9uq4l)

[72. Место и способ образования фонем. Классификация фонем. 75](#_avdsfby654sb)

[73. Коартикуляция, редукция и др. явления в речи. 77](#_lrvon02vw0r1)

[74. Просодические характеристики речи. 77](#_rtrfo9sfhcvr)

[75. Физические параметры речевого сигнала, с помощью которых описывается просодия речи. 78](#_uixr2lsp5lji)

[76. Функции интонации в речи. 78](#_1vhm1qs9ie1e)

[77. Ударение: понятие, назначение, типы. 79](#_3o45dfsu2dm)

[78. Синтагма: понятие, типы, примеры. 80](#_hz7po4njj9aa)

[79. Акцентная группа: понятие, типы, примеры. 81](#_ugj5apkxo2nt)

[80. Основные интонационные типы конечных акцентных групп. 81](#_44iu1h7i52na)

[81. Состав акцентной группы. 82](#_h1gbvwsiebpt)

[82. Задача автоматического распознавания языка текста. 83](#_kl0b49uecuhg)

[83. Задача автоматического распознавания языка текста. Метод n-грамм, достоинства и недостатки. 83](#_uzrysivmr25r)

[84. Задача автоматического распознавания языка текста. Алфавитный метод, достоинства и недостатки. 84](#_c4ux9i56wc6s)

[85. Задача автоматического распознавания языка текста. Метод коротких слов, достоинства и недостатки. 85](#_3vgl5u4avc5c)

[86. Полиинформативность речевого сигнала. Виды информации, заключенной в речевом сигнале. 86](#_or2gmhe6q7gj)

[87. Методы синтеза речевого сигнала. 87](#_lvf61kid96q8)

[88. Структура синтезатора речи по тексту. 88](#_hm1ae4bcz0up)

[89. Лингвистический процессор. 89](#_z99me9x0nhi0)

[90. Просодический процессор. 90](#_8r4327chv5tw)

[91. Фонетический процессор. 91](#_wgvc4v8z4m3v)

[92. Акустический процессор. 91](#_yew8aihjdy1s)

[93. Задача стемминга текстов естественного языка. 92](#_lknpkpkf211m)

[94. Структура речевого сигнала. Спектрально-временные характеристики речевого сигнала. 93](#_1p00x18veyra)

[95. Синтез речевого сигнала. Методы синтеза: параметрические и компиляционные. 95](#_lwptiorct5d8)

[Параметрические методы: 95](#_y857oopm1fhi)

[Компиляционные методы: 96](#_z9n00cf003sz)

[Сравнение методов: 96](#_stfpn7u2ysf)

[96. Анализ речевого сигнала. Методов анализа: цифровой фильтрации, быстрых алгоритмов преобразования Фурье, Кепстральных коэффициентов. 96](#_egw7hh45immr)

#### 

#### **1. Естественно-языковой интерфейс: определение, характеристики, примеры.**

Естественно-языковой интерфейс (Natural Language Interface, NLI) - это система взаимодействия между человеком и компьютером, которая позволяет пользователю вводить команды или делать запросы при помощи естественного языка, вместо использования формализованных команд или языка программирования.

ЕЯ интерфейс предполагает трансляцию естественно-языковых конструкций на внутримашинный уровень представления знаний. Для этого необходимо решать задачи морфологического (распознавание и проверку правильности написания слов по словарям), синтаксического (разложение входных сообщений на отдельные компоненты (определение структуры) с проверкой соответствия грамматическим правилам внутреннего представления знаний и выявления недостающих частей) и семантического (установление смысловой правильности синтаксических конструкций) анализа и синтеза высказываний на естественном языке (решает обратную задачу преобразования внутреннего представления информации в естественно-языковое).

Характеристики естественно-языковых интерфейсов:

1. Распознавание естественного языка: Система способна анализировать и понимать естественный язык, включая синтаксис, семантику и контекст.

2. Генерация ответов: Способность системы генерировать понятные и соответствующие запросу ответы на естественном языке.

3. Обработка запросов: Система обрабатывает запросы пользователей и выполняет соответствующие операции, используя алгоритмы обработки естественного языка и другие технологии ИИ.

4. Контекстуальное понимание: Возможность учесть контекст и предыдущие запросы для более точного понимания запросов пользователя.

Примеры естественно-языковых интерфейсов:

1. Поисковики: гугл, яндекс, и тд
2. AI-чаты: gemini, gpt yandexgpt
3. Голосовые помощники: Siri от Apple, Google Assistant, Amazon Alexa, Алиса.
4. Чат-боты: Боты в мессенджерах или на веб-сайтах, такие как чат-боты в Facebook Messenger или на сайтах служб поддержки.
5. Системы умного дома: Интерфейсы для умных устройств в доме, которые могут реагировать на команды пользователя на естественном языке, например, управление освещением или устройствами в доме.

#### **2. Задачи создания естественно-языкового интерфейса.**

Создание естественно-языкового интерфейса - это сложная задача, которая включает в себя различные аспекты и требует комбинации компетенций в области лингвистики, искусственного интеллекта, обработки естественного языка и пользовательского опыта. Ключевые задачи при создании естественно-языкового интерфейса:

1. Анализ текстов на естественном языке: как устроены текст и предложения, из которых он состоит.

Для этого необходимо решать задачи морфологического (распознавание и проверку правильности написания слов по словарям), синтаксического (разложение входных сообщений на отдельные компоненты (определение структуры) с проверкой соответствия грамматическим правилам внутреннего представления знаний и выявления недостающих частей) и семантического (установление смысловой правильности синтаксических конструкций) анализа.

1. Понимание текстов (контекст, смысл слов, выделение ключевых понятий)
2. Синтез высказываний на естественном языке (решает обратную задачу преобразования внутреннего представления информации в естественно-языковое), сложность заключается в выборе подходящей структуры ответа.
3. Распознавание речи: обеспечение точного распознавания речи, чтобы система могла правильно интерпретировать аудиосигналы и преобразовывать их в текст.
4. Интеграция с базой знаний: обеспечение доступа к обширной базе знаний, чтобы система могла отвечать на вопросы и предоставлять информацию в соответствии с запросами пользователя.

#### **3. Отличия естественно-языкового интерфейса от других видов интерфейса.**

Под пользовательским интерфейсом понимается система программных решений, реализующих поиск, получение, просмотр и обработку информации из внешнего хранилища, как правило, структурированного источника данных.

При этом естественно-языковой интерфейс – это разновидность пользовательского интерфейса, который принимает на вход и обрабатывает запросы на естественном языке. Таким образом, реализация ЕЯ-интерфейса предполагает удобный для неподготовленного пользователя диалоговый режим коммуникации и должна способствовать достижению целей пользователя, которые определяются его информационными потребностями.

Важными особенностями процесса общения, направленного на удовлетворение информационных потребностей пользователя, которые необходимо учитывать при проектировании и реализации ЕЯ-интерфейсов независимо от специфики решаемых пользователями задач, являются:

1. **Изменяемость.** Информационная потребность пользователя не может быть заранее четко определена в спецификациях на разработку системы общения. Напротив, она неизбежно изменяется в ходе разработки и эксплуатации системы.

2. **Несовпадение взглядов на мир.** Представления, имеющиеся у пользователя и системы о языке общения и проблемной области, относительно которой ведется общение, могут не совпадать. Соответственно процесс общения должен предусматривать разъяснение смысла неизвестных терминов, обнаружение и устранение несовпадающих представлений, а также предупреждение ошибочных толкований, т. е. установление общих точек зрения на обсуждаемые в процессе общения сущности.

3. **Связность общения.** Процесс общения не может быть ограничен обменом изолированными парами «вопрос – ответ», т. к. в большинстве реальных случаев информационная потребность пользователя не может быть выражена в виде одного вопроса (предложения). Часто требуется определить ситуацию, в которой возникла ИПП, т. е. предпослать запросу на решение некоторой задачи контекст, в котором эту задачу необходимо решать. Кроме того, процесс удовлетворения ИПП – решение некоторой задачи в большинстве реальных приложений требует взаимодействия, основанного на смешанной инициативе участников.

4. **«Неправильность» высказываний пользователя.** Для выражения ИПП пользователь может применить как «правильные» предложения, т. е. такие, которые будут однозначно поняты и верно обработаны системой, так и «неправильные». Неправильности возникают, во-первых, из-за того, что пользователь обычно не в состоянии учесть все ограничения системы общения в части ее возможностей и знаний, во-вторых, в связи с использованием умолчаний, характерных для естественного общения и допускающих неоднозначное толкование высказываний, и, в-третьих, из-за отклонения предложений от грамматической нормы.

#### **4. Компьютерная лингвистика: определение, задачи, направления.**

Вследствие внедрения новых информационных технологий во все сферы человеческого общения прикладная лингвистика развивается по направлению автоматизации основных задач, оптимизации коммуникации, т. е. развивается такое направление прикладной лингвистики, которое назвали компьютерной лингвистикой.

**Компью́терная лингви́стика** — научное направление в области математического и компьютерного моделирования интеллектуальных процессов у человека и животных при создании систем искусственного интеллекта, которое ставит своей целью использование математических моделей для описания естественных языков. Свое начало она берет в 50–60-х гг. XX в., а ее появление связано с попыткой решить задачу автоматического перевода текстов с одного языка на другой (Джорджтаунский проект).

Центральными научными проблемами компьютерной лингвистики являются:

– проблема моделирования процесса понимания смысла текстов (перехода от текста к формализованному представлению его смысла)

– проблема синтеза речи (перехода от формализованного представления смысла к текстам на естественном языке).

**Перед компьютерной лингвистикой стоят задачи** лингвистического обеспечения процессов сбора, накопления, обработки и поиска информации. Наиболее важными из них являются:

1. Построение лингвистических процессоров, обеспечивающих общение пользователей с автоматизированными интеллектуальными информационными системами на естественном языке, или на языке, близком к естественному
2. Информационный поиск, в том числе автоматическое индексирование документов и запросов пользователя. Основной задачей здесь является поиск необходимой пользователю информации согласно его запросу (релевантной его запросу). С этой точки зрения именно смысловая обработка документов и запросов пользователя становится особенно актуальной.
3. Автоматическое реферирование текстов. Подразумевает автоматическое выделение на основе анализа текста наиболее важной с определенной точки зрения информации из документа и представление ее пользователю в том или ином виде. Это может быть часть оригинального текста, набор взятых из 9 него предложений, выделенных по заданному критерию, информация, соответствующая основным типам знаний, – реферат в виде списка объектов (ключевых слов) или фактов, в виде иерархии объектов и т. д.
4. Автоматизация инженерии знаний. Речь идет о том, чтобы на основе автоматического анализа текста выделить из него знания, по крайней мере соответствующие трем основным их типам: объектам (классам объектов), фактам и правилам, отображающим закономерности внешнего мира (предметной области).
5. Машинный перевод (МП) текстовых документов с одного ЕЯ на другой (другие). Целью данной задачи является получение адекватного представления входного текста в терминах некоторого другого (других) ЕЯ. Система МП может использоваться как средство собственно перевода документов пользователя, так и как составная часть других информационных систем, например многоязычных систем информационного поиска.
6. Автоматическая классификация текстовых документов. Включает в себя: автоматическую категоризацию, т. е. распределение документов по заранее созданным категориям; автоматическую кластеризацию, т. е. распределение документов по автоматически генерируемым группам или иерархиям групп; генерацию таксономий, т. е. создание иерархий концептов или тематических категорий.
7. Визуализация набора текстовых документов. Целью данной задачи является представление в графической форме, удобной для понимания пользователя, информации о некотором наборе документов и (или) информации, содержащейся в них. Как правило, получаемые представления имеют вид деревьев, графов или диаграмм. Например, это может быть диаграмма объектов входного набора документов с указанием частоты их встречаемости в документах.
8. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ;

#### **5. Прикладная лингвистика: определение, сходства и различия с компьютерной лингвистикой.**

**Прикладная лингвистика** представляет собой научную область, которая занимается применением знаний лингвистики к практическим задачам. Эти задачи включают в себя изучение иностранных языков, разработку языковых технологий, анализ коммуникаций, преподавание языков, перевод и многое другое. Прикладная лингвистика охватывает широкий спектр областей, в которых применяются языковые знания для решения конкретных практических проблем.

Обычно выделяют следующие направления:

* лексикография — теория и практика составления словарей;
* лингводидактика — наука о разработках методик обучения иностранному языку;
* терминоведение — наука об упорядочении и стандартизации научно-технической терминологии;
* переводоведение — теория перевода
* квантитативная лингвистика (изучение частотных свойств языковых единиц и их связь с другими свойствами, такими как фонетическая сложность, морфологическая сложность, многозначность, возраст).

Основные направления прикладной лингвистики, связанные с практическими приложениями:

* Компьютерная лингвистика (computational linguistics)
* Лингвистическая экспертиза (например, в судебной практике)
* Политическая лингвистика (анализ политического дискурса)

С функциональной точки зрения, прикладная лингвистика может быть определена как академическая дисциплина, в которой целенаправленно изучаются и разрабатываются способы оптимизации различных сфер функционирования языковой системы.

Компьютерная лингвистика — это область знаний, которая занимается компьютерным моделированием владения естественным языком и решением прикладных задач автоматической обработки текстов и речи.

Сходства прикладной лингвистики с компьютерной лингвистикой заключаются в их общей связи с обработкой языка. Различия между этими двумя областями включают в себя тот факт, что компьютерная лингвистика более технически ориентирована и сосредоточена на разработке алгоритмов и программного обеспечения для анализа текста и речи с использованием технологий и компьютерных ресурсов. В то время как прикладная лингвистика более ориентирована на конкретные языковые приложения и проблемы, такие как перевод, обучение языку, исследование языковых изменений, общения и многое другое. Основной упор делается на практическое применение языковых знаний для решения реальных проблем и задач в различных областях.

#### **6. Причины возникновения компьютерной лингвистики.**

Компьютерная лингвистика берет начало в 50–60-х гг. XX в., а ее появление связано с попыткой решить задачу автоматического перевода текстов с одного языка на другой (Джорджтаунский проект).

Возникновение данного направления обусловлено и другими факторами:

1. Потребность в обработке больших объемов текстовой информации: появление интернета и развитие цифровых технологий привели к значительному увеличению объема текстовой информации. Эффективная обработка и анализ этой информации стали крайне важными.
2. Рост интереса к искусственному интеллекту (ИИ): благодаря развитию компьютерных технологий и появлению мощных компьютеров стало возможным использование ИИ для решения сложных задач, связанных с обработкой естественного языка. С появлением первых концепций искусственного интеллекта (1950-е годы) началась разработка систем, способных взаимодействовать с людьми через язык (диалоговые системы).
3. Развитие лингвистики, в том числе появление формальных методов описания языка (например, генеративной грамматики Н. Хомского) позволило разрабатывать алгоритмы, которые могли бы анализировать языковые структуры.
4. Необходимость поиска информации в больших объемах данных: увеличение числа текстов создало спрос на системы, способные находить релевантные документы или информацию по запросу.

#### **7. Отрасли знаний и научные направления, связанные с компьютерной лингвистикой и проблемой разработки естественно-языкового интерфейса.**

Компьютерная лингвистика образовалась на стыке лингвистики, математики, информатики и искусственного интеллекта. Так как разработка естественно-языковых интерфейсов носит междисциплинарный характер, необходимо перечислить другие научные и информационные направления, знания из которых могут быть полезными при разработке таких интерфейсов:

* психолингвистика - дисциплина, которая находится на стыке психологии и лингвистики. Изучает взаимоотношение языка, мышления и сознания. Возникла в 1953 году. Психолингвистика занимается:

– описанием речевых сообщений на основе изучения механизмов порождения и восприятия речи;

– изучением функций речевой деятельности в обществе;

– исследованием связи между речевыми сообщениями и характеристиками участников коммуникации (превращение намерений говорящего в сообщения, интерпретация их слушающим);

– анализом речевого развития в связи с развитием личности.

* когнитивная психология - раздел психологии, изучающий познавательные способности человека, особенности организации его памяти, процессов восприятия, запоминания, хранения и обработки информации. Очень многие результаты когнитивной психологии берутся за основу разработок в области искусственного интеллекта. Существует также современное научно-техническое направление под названием cognitive science (когнитивная наука), которое очень активно развивается за рубежом;
* эргономика - научно-прикладная дисциплина, занимающаяся изучением и созданием эффективных систем, управляемых человеком; Эргономика изучает действия человека в процессе работы, скорость освоения им новой техники, затраты его энергии, производительность и интенсивность при конкретных видах деятельности.
* физика, если речь идет о распознавании голоса(акустика)

Очевидна также непосредственная связь компьютерной лингвистики со следующими

научными направлениями:

* лингвистика - наука о языке (естественном);
* филология;
* языкознание, труды по языкознанию.

#### **8. Понятия «общение», «диалог», «язык», их взаимосвязь друг с другом и с проблемой создания естественно-языкового интерфейса.**

Язык – это знаковая система, предназначенная для порождения, передачи и хранения информации. Язык имеет свои правила, структуру и семантику.

Общение – это процесс обмена информацией, включающий передачу и восприятие сообщений. Это может включать в себя различные формы выражения, такие как вербальные и невербальные сигналы. Диалог представляет собой разновидность общения, это процесс достижения его участниками определенных согласованных целей путем обмена связанными высказываниями, выраженными в языке, о некоторой проблемной области. Применительно к диалогу между пользователем и компьютером под общением понимают процесс обмена взаимосвязанными высказываниями, выраженными в языке, направленный на достижение целей пользователя, т. е. на удовлетворение информационных потребностей пользователя (ИПП).

В общем случае процесс общения не может быть сведен к обмену изолированными парами высказываний «вопрос – ответ». Высказывания участников общения образуют связный текст – дискурс, имеющий, как правило, достаточно сложную структуру. Связность дискурса обеспечивается как лингвистическими (родовидовыми, анафорическими, модальными, стилистическими согласованиями, согласованиями пресуппозиций и т. п.), так и экстралингвистическими (ситуативными) средствами, т. е. с помощью временных, причинноследственных и других связей, существующих в проблемной области.

Естественно-языковой интерфейс– это система предоставляющая возможность взаимодействия с ней пользователя посредством естественного языка. Это создает возможность общения с системой более естественным образом, подобным общению между людьми. В ЕЯИ система должна способствовать пониманию и генерации естественного языка, что представляет собой технологическую сложность в области обработки естественного языка.

**9. Задача машинного перевода текстов.**

**Машинный перевод** — процесс перевода текстов (письменных, а в идеале и устных) с одного естественного языка на другой с помощью специальной компьютерной программы. Так же называется направление научных исследований, связанных с построением подобных систем. Его целью является получение адекватного представления входного текста в терминах некоторого другого (других) ЕЯ. Система МП может использоваться как средство собственно перевода документов пользователя, так и как составная часть других информационных систем, например многоязычных систем информационного поиска.

В машинном переводе существует несколько подходов и технологий:

– Прямой машинный перевод (текст на исходном языке не подвергается структурному анализу за пределами морфологии. Такой перевод использует большое количество словарей и является пословным, если не считать небольшой грамматической корректировки, например, касательно порядка слов и морфологии)

– Машинный перевод на правилах

Машинный перевод Интерлингва (перевод осуществляется через промежуточную (семантическую) модель текста исходного языка, не зависящую от конкретного языка модель, из которой может быть сгенерирован перевод на любой язык.)

Машинный перевод Трансфер (основан на идее Interlingua с использованием сопоставительного анализа двух языков. Сначала текст на исходном языке переводится в абстрактную или промежуточную модель исходного языка, которая затем преобразуется в модель целевого языка, чтобы потом окончательно оформиться в текст на целевом языке)

– Машинный перевод на корпусах текстов (корпусный подход в машинном переводе использует совокупность (корпус) параллельных двуязычных текстов. )

Статистический машинный перевод

Машинный перевод на примерах

Примеры взаимодействия человека и машины при выполнении машинного перевода:

* С постредактированием: исходный текст перерабатывается машиной, а человек-редактор исправляет результат.
* С предредактированием: человек приспосабливает текст к обработке машиной (устраняет возможные неоднозначные прочтения, упрощает и размечает текст), после чего начинается программная обработка.
* С интерредактированием: человек вмешивается в работу системы перевода, разрешая трудные случаи.
* Смешанные системы (например, одновременно с пред- и постредактированием).

#### **10. Правила и способы общения. Примеры реализации диалога человека с компьютерной системой**.

Диалог между человеком и компьютерной системой строится на основе определенных правил и подходов, которые обеспечивают ясность, удобство и эффективность взаимодействия. Шаг диалога трактуется как пара «действие – реакция», где высказывание активного (т .е. владеющего инициативой) участника соответствует действию, а пассивного – реакции. Действие и реакция, образующие шаг диалога, могут в общем случае не соответствовать соседним (во временной последовательности) высказываниям участников. Соответствие нарушается при перехватах инициативы, который возникает в тех случаях, когда пассивный участник вместо цели, предложенной активным участником, выбирает иные цели, в частности подцели, предусматривающие преодоление локальных неудач. Например, вместо ответа на вопрос второй участник может задать встречный вопрос (т. е. совершить действие и тем самым взять на себя активную роль) и лишь после получения ответа на него ответить на первоначально заданный вопрос.

1. Понятность и доступность: система должна использовать простой, понятный язык, избегая сложных технических терминов, формулировки запросов и ответов должны быть лаконичными и однозначными.

- Пользователь: "Какая погода в Минске?"

- Система: "Сейчас в Минске -8°C, идет снег."

1. Контекстуальность: система должна учитывать предыдущую историю взаимодействий, чтобы предоставлять релевантные ответы.

- Пользователь: "А Гомель?" – запоминает что речь о погоде

- Система: "в Гомеле -5°C, солнечно"

1. Точность и релевантность: ответы системы должны быть корректными и соответствовать запросу пользователя. Если запрос неполный или неоднозначный, система должна уточнять, а не делать предположения.

- Пользователь: "Какой прогноз погоды для Минска на неделю?"

– выдает температуру и состояние, а не уровень давления, закаты-рассветы. необходимый минимум

1. Обратная связь в случае ошибок: объяснять причину и предлагать решение.

прогноз на год в городе пупкине,

либо– не могу узнать на год либо я не знаю такого города вы уверенны что он существует?-- перехват инициативы

1. Эмоциональная нейтральность: общение должно быть вежливым и нейтральным, с возможностью адаптации стиля к предпочтениям пользователя.
2. Гибкость: Компьютерная система должна быть гибкой и уметь обрабатывать разные типы запросов и команд пользователя. Она должна быть способна адаптироваться к различным стилям общения и формулировкам вопросов.

Пользователь: Я голоден. Где можно хорошо поужинать?

Система: Рядом с вами есть несколько хороших ресторанов, таких как "Итальянская кухня" и "Стейкхаус". Что вам больше нравится - паста или стейк?

#### **11. Диалоговая система: определение, компоненты. Проблемы, возникающие в процессе реализации естественно-языкового диалога человека и компьютерной системы.**

**Диалоговая система** представляет собой программу или искусственный интеллект, разработанный для взаимодействия с пользователем на естественном языке. Эти системы могут выполнять различные задачи, такие как ответы на вопросы, предоставление информации, выполнение команд и даже проведение некоторых форм развлечений.



Компоненты

**Модуль обработки естественного языка (NLP)**: **Распознавание речи (ASR)**: преобразование устной речи в текст (в случае голосовых систем) и (или) **Анализ ввода (NLU)**: извлечение смысла из текста пользователя.

**Менеджер диалога**: Определяет, что должна сделать система в ответ на запрос пользователя. Управляет ходом диалога, отслеживает контекст, хранит историю взаимодействия.

**Модуль обработки высказываний** (Task Manager): Выполняет задачи, связанные с обработкой данных и запросов. Обращается к базе данных или другим внешним источникам для получения необходимой информации. Обрабатывает результаты и подготавливает их для генерации ответа.

**Генерация текста или речи (NLG)**: Формирует текст ответа системы на основе данных, подготовленных менеджером диалога. В случае голосовых систем используется синтез речи (TTS — text-to-speech).

**База знаний и внешние источники**: Хранение данных, используемых системой, например, базы данных для поиска информации.

**Интерфейс взаимодействия с пользователем**: Текстовый интерфейс (чат). Голосовой интерфейс (микрофон и динамики).

**Проблемы**:

* **Понимание контекста:** Для успешного взаимодействия система должна понимать контекст разговора, что может быть сложно при многослойных и запутанных диалогах.
* **Распознавание речи:** Ошибки в преобразовании устной речи в текст из-за акцентов, фонового шума или качества записи.
* **Распознавание нечеткого ввода (многозначность слов или нечеткие формулировки):** Пользователи часто выражают свои мысли нечетко, и система должна правильно распознать их запросы даже при наличии омовений или неполных фраз.
* **Способность генерации естественного ответа:** Создание ответов, которые звучат естественно, является сложной задачей, особенно при обработке сленга, коллоквиализмов и многозначности.
* **Управление ошибками**: Диалоговые системы должны быть способными обрабатывать ошибки ввода и предоставлять пользователю понятные сообщения об ошибках.
* **Локальные различия в языке**: Различия в диалектах, акцентах, интонациях и стилях общения могут создавать трудности в обработке разнообразных языковых особенностей.

#### **12. Моделирование диалога.**

Моделирование диалога — это процесс проектирования и реализации взаимодействия между пользователем и диалоговой системой. Оно включает разработку логики общения, сценариев диалога и алгоритмов обработки входных и выходных данных. Основная цель — создание последовательного, естественного и эффективного общения. Основные подходы к моделированию диалога

1. **Модели на основе сценариев или правил:**Логика диалога задается вручную в виде последовательности правил и фиксированных сценариев.

**Плюсы**: Простота реализации. Прозрачность логики.

**Минусы**: Ограниченность. Трудно поддерживать естественные и сложные диалоги. Невозможность обработки непредусмотренных сценариев.

1. **Статистические (обучение на данных):** Диалог моделируется с использованием данных о взаимодействиях (история реальных диалогов).

**Методы**: Наивные байесовские классификаторы для анализа намерений. Методы машинного обучения для извлечения сущностей и управления диалогом.

**Плюсы**: Могут адаптироваться к различным сценариям. Частично автоматизируют процесс создания логики.

**Минусы**: Ограниченная гибкость при обучении на узкоспециализированных данных.

1. **Нейросетевые (на основе глубокого обучения):** Используются модели нейронных сетей для генерации и управления диалогом.
   * **Методы**:
     + Sequence-to-sequence модели (Seq2Seq).
     + Трансформеры (например, GPT, BERT).
   * **Плюсы**:
     + Высокая степень адаптации к различным сценариям.
     + Возможность создания свободных, "живых" диалогов.
   * **Минусы**:
     + Требуют значительных вычислительных ресурсов.
     + Сложность интерпретации и контроля.

Ключевые аспекты моделирования

1. **Учет контекста:** Система должна помнить, что было сказано ранее, для понимания текущего запроса.
2. **Управление состоянием диалога:** Отслеживание хода диалога и переход между его этапами.
3. **Реализация стратегии диалога:** Выбор стратегии: уточнение запроса, повтор, управление отклонениями.
4. **Генерация естественного ответа:** Ответы должны быть осмысленными, логически связанными и грамматически правильными.

#### **13. Понятие «язык»: определение, проблемы формализации.**

Язык - это система символов и правил, используемая для передачи и понимания информации между людьми. Он служит средством общения, позволяя выражать мысли, идеи, чувства и передавать знания. Язык включает в себя звуковые, визуальные и письменные элементы, а также грамматические и синтаксические структуры.

Коммуникативная функция: Эта функция связана с обменом информацией и коммуникацией между людьми для передачи мыслей, понимания и общения.

Экспрессивная (эмоциональная) функция: Язык используется для выражения эмоций, чувств, намерений и эмоционального состояния.

Информационная (когнитивная) функция: Язык используется для передачи информации, фактов, объяснений и обучения.

Регулятивная функция: Язык используется для управления поведением других людей или саморегуляции.

Мета-языковая функция: Это использование языка для обсуждения самого языка, его структуры, значения слов и т.д.

* Виды языков:

Естественные языки: Такие языки, как английский, испанский, китайский и другие, которые развивались естественным образом у людей для общения.

Искусственные языки: Языки, созданные с целью общения, например, эсперанто или логические языки программирования.

Специальные языки: Языки, созданные для определенных целей или сфер, такие как математический язык, музыкальные нотации, язык химических формул и т.д.

Знаковые языки: Такие языки, как язык жестов или брайля, используемые людьми с ограниченными возможностями в общении.

Мета-языки: Языки, используемые для описания и анализа других языков или языковых структур.

**Проблемы формализации:**

**Многозначность**: Слова и выражения могут иметь несколько значений в различных контекстах, что усложняет формализацию и точное определение.

**Изменчивость**: Язык постоянно эволюционирует, появляются новые слова, меняются значения, что затрудняет создание статичных формализированных моделей.

**Контекстуальность:** Значение выражения может зависеть от контекста, в котором оно используется, и контекст может быть неоднозначным.

**Эмоциональная окраска:** Язык часто несет в себе эмоциональную окраску, которую сложно формализовать и учесть в формальных моделях.

**Идиоматичность**: Некоторые языковые конструкции, идиомы и фразеологизмы не всегда могут быть легко поняты на основе общих грамматических правил.

**Индивидуальные особенности:** Люди могут использовать язык индивидуально, с уникальными стилями, жаргоном или личными коннотациями.

#### **14. Слово, словоформа, лексема. Примеры.**

**Слово** — (однозначное аксиоматическое обозначение в лексике) — одна из основных, структурных единиц языка, которая служит для наименования предметов, их качеств и характеристик их взаимодействий, а также именования мнимых и отвлечённых понятий, создаваемых человеческим воображением. В поисках структуры Слова современная наука сформировала самостоятельную отрасль, называемою Морфологией.

Вся совокупность Слов делится на два типа:

знаменательные — обозначающие определённые понятия,

служебные — служащие для связи слов между собой.

По грамматическому значению Слова классифицируются как части речи:

знаменательные слова — существительное, прилагательное, глагол, наречие;

подклассы — числительные, местоимения и междометия;

служебные слова — союз, предлог, частица, артикль и др.

**Словоформа** (также форма слова) — слово в узком смысле, то есть обладающая признаками слова цепочка фонем, формально отличающаяся от другой.

**Лексема** — это слово как совокупность всех его конкретных грамматических форм (например, лексему лист образуют формы лист, листа, листу, листом, а они в свою очередь являются словоформами). Словарная форма лексемы – лемма(т.е. начальная форма).

В ряде концепций в лексему включаются разные смысловые варианты слова, зависящие от контекста, в котором оно употребляется (например, соль в смысле названия вещества и в значении того, что придает остроту или интерес какому-либо высказыванию, мысли).

#### 

#### 

#### **15. Функции языка. Виды языков.**

Функции языка описывают различные способы, которыми язык используется для общения и передачи информации. В зависимости от контекста и целей коммуникации выделяют несколько основных функций языка:

* Коммуникативная функция: Эта функция связана с обменом информацией и коммуникацией между людьми для передачи мыслей, понимания и общения.
* Экспрессивная (эмоциональная) функция: Язык используется для выражения эмоций, чувств, намерений и эмоционального состояния.
* Информационная (когнитивная) функция: Язык используется для передачи информации, фактов, объяснений и обучения.
* Регулятивная функция: Язык используется для управления поведением других людей или саморегуляции.
* Мета-языковая функция: Это использование языка для обсуждения самого языка, его структуры, значения слов и т.д.

Виды языков:

* Естественные языки: Такие языки, как английский, испанский, китайский и другие, которые развивались естественным образом у людей для общения.
* Искусственные языки: Языки, созданные с целью общения, например, эсперанто или логические языки программирования.
* Специальные языки: Языки, созданные для определенных целей или сфер, такие как математический язык, музыкальные нотации, язык химических формул и т.д.
* Знаковые языки: Такие языки, как язык жестов или брайля, используемые людьми с ограниченными возможностями в общении.
* Мета-языки: Языки, используемые для описания и анализа других языков или языковых структур.

Все эти функции и виды языков играют важную роль в общении, передаче информации и понимании мира вокруг нас.

#### **16. Уровни обеспечения систем машинного перевода текстов**

Системы машинного перевода (МП) текстов обеспечиваются на нескольких уровнях, каждый из которых связан с этапами обработки текста, выбором метода перевода и особенностями работы системы. Эти уровни включают:

**Лексический уровень:** На этом уровне осуществляется анализ отдельных слов, их значений и перевод на целевой язык на основе лексических ресурсов, таких как словари. Этот этап отвечает за точное определение эквивалентов слов и устойчивых выражений.

**Морфологический уровень:** На морфологическом уровне происходит разбор и генерация словоформ с учетом их грамматических характеристик, таких как падеж, род, число. Например, преобразование базовой формы слова в нужную форму для предложения.

**Синтаксический уровень:** Здесь осуществляется разбор синтаксической структуры предложений. Построение синтаксических деревьев помогает установить связи между словами и группами слов, а также сохранить правильный порядок слов при переводе.

**Семантический уровень:** На этом уровне система учитывает значение текста, анализирует связи между словами и смысловые отношения, чтобы предотвратить ошибки, связанные с многозначностью слов.

**Прагматический уровень:** Учет целей и намерений автора текста: учет стиля, формальности, тональности текста.

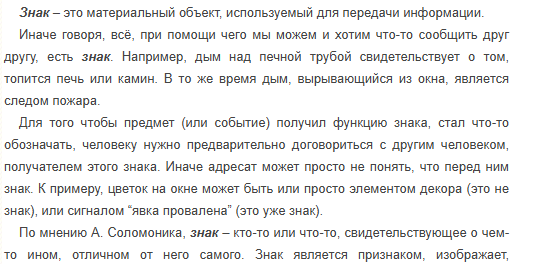
**Дискурсивный уровень:** Отвечает за обеспечение связи между предложениями и общую согласованность текста, сохраняя контекст, такие как анафорические ссылки (например, местоимения "он", "она").

**Культурный и стилистический уровень:** Учитываются идиомы, культурные особенности и стилистика текста, чтобы перевод был адекватным в конкретной культурной или профессиональной среде

#### **17. Основные понятия семиотики.**

**Семиотика** — это наука о знаках и символах, а также об их использовании и восприятии. Объект семиотики – знаки и системы знаков. Предмет – закономерности, тенденции, особенности возникновения и функционирования знаков и знаковых систем.

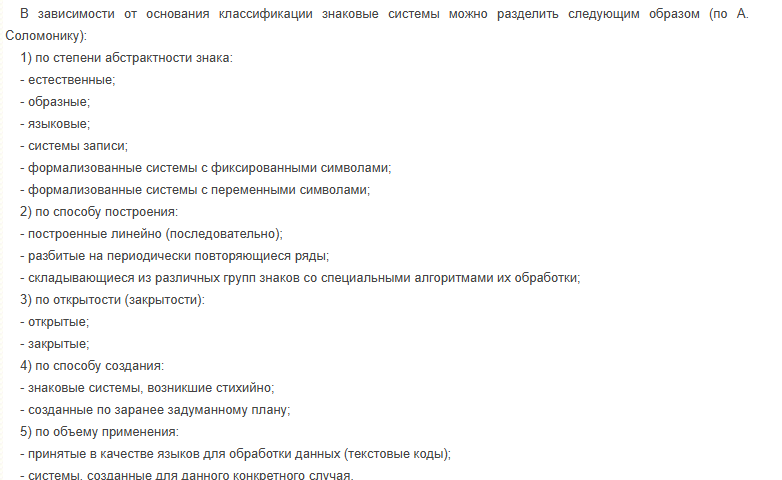
**Основные понятия семиотики** включают в себя:

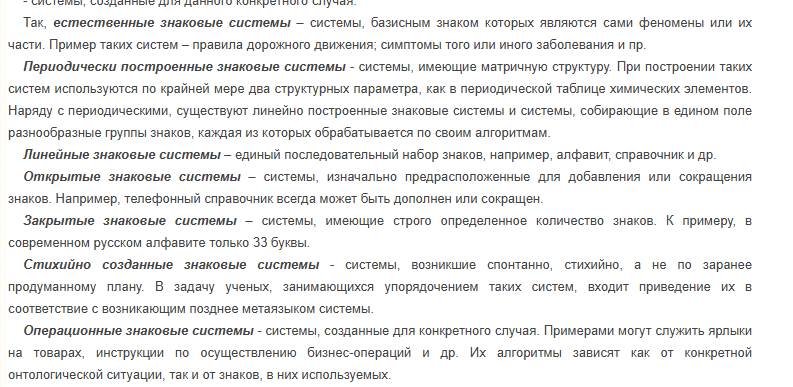
* **Знак** (символ): Элемент, который представляет собой нечто другое. Знак состоит из двух частей: сигнала (звучащего, видимого, осязаемого) и смысла (интерпретации, которую мы придаем знаку).
* **Семиозис**: процесс интерпретации знака, или процесс порождения значения.
* Ситуация использования знака – знаковая ситуация
* **Денотат и коннотат:** Денотат — прямой, буквальный смысл знака. Коннотат — дополнительные ассоциации и эмоциональные значения, связанные с знаком.
* **Знаковая система**: Совокупность знаков, объединенных между собой различными видами связей.
* **Семиотический код**: Система правил и норм, регулирующих использование знаков в конкретной ситуации или сообществе.
* Алгоритм: система правил, определяющая содержание и последовательность действий для обработки знаков(групп знаков), кодирующих родственные предметы или явления
* **Синтаксис, семантика, прагматика**: Синтаксис — правила комбинирования знаков. Семантика — их значение и интерпретация. Прагматика — контекст и практическое использование знаков в конкретной ситуации.
* **Икон, символ**: Икон — знак, который представляет объект на основе физического сходства с ним. Символ — знак, устанавливающий отношение с объектом на основе социокультурных соглашений.

#### **18. Знаковая система: определение, свойства, примеры.**

**Знаковая система –** совокупность знаков, образующих единство на основе внутренних отношений между ними и используемых в определенных сферах деятельности.

В качестве примера знаковой системы обычно приводят язык.





Свойства знаковой системы

1. Знаковость – каждый элемент системы является знаком и имеет связь с каким-либо объектом или понятием.

2. Двусторонность – знак состоит из формы (внешнего вида) и содержания (значения). Например, слово "дерево" имеет форму в виде букв, а содержание связано с реальным объектом — деревом.

3. Конвенциональность – значение знаков зависит от соглашений между участниками коммуникации. Это означает, что значение может изменяться в зависимости от культуры, языка или контекста.

4. Системность — знаки в системе взаимосвязаны и упорядочены, что позволяет передавать более сложные сообщения.

5. Многозначность — один знак может иметь несколько значений, в зависимости от контекста.

6. Универсальность — знаковые системы могут использоваться в различных областях жизни, от языка до символов и знаков в математике, науке, технике.

Примеры знаковых систем:

1. Язык — это основная знаковая система, которая используется для общения между людьми. В языке знаки — это слова, а их значения зависят от правил грамматики и синтаксиса.

2. Математика — знаки, такие как цифры и математические символы (например, +, −, =), составляют знаковую систему, позволяющую передавать количественные и логические отношения.

3. Жестовый язык — система знаков, используемых глухими людьми для общения. Это невербальная знаковая система, в которой жесты заменяют слова.

5. \*\*Трафиковые знаки\*\* — система знаков, используемая в дорожном движении. Например, знак "стоп" или "пешеходный переход" передают конкретную информацию водителям и пешеходам.

6. \*\*Двумерные графики и схемы\*\* — используются для визуального представления информации, например, диаграммы, схемы или карты. Каждая линия, цвет или форма может быть знаком, который передаёт информацию.

#### **19. Омонимия: определение, виды, примеры, достоинства и недостатки.**

Википедия:

Омо́нимы — одинаковые по написанию и звучанию, но разные по значению слова и другие единицы язык.

Виды:

* Омонимы полные (абсолютные) — омонимы, у которых совпадает вся система форм. Например, «наряд (одежда)» — «наряд (распоряжение)», «горн (кузнечный)» — «горн (духовой инструмент)», «лук — (метательное оружие)» — «лук (растение семейства Амариллисовые)», «ключ (инструмент или программный код для различных операций)» — «ключ (музыкальный знак)» — «ключ (часть сложного иероглифа в иероглифике)» — «ключ (синоним к слову „родник“)»
* Омонимы частичные — омонимы, у которых совпадают не все формы. Например, «ласка (животное)» и ласка «(проявление нежности)» расходятся в форме родительного падежа множественного числа (ласок — ласк).
* Омонимы грамматические, или омоформы — слова, совпадающие лишь в отдельных формах (той же части речи или разных частей речи). Например, числительное «три» и глагол «тереть» совпадают лишь в двух формах («к трём» — «мы трём», «три яблока» — «три сильнее!».

Достоинства(какой-то бред, интернет их не знает)

* Богатство выразительности: Омонимия позволяет языку быть более гибким и разнообразным, давая возможность использовать одно слово для выражения различных идей или концепций.
* Ирония и игра слов: Омонимия часто используется для создания игры слов, шуток и ироничных намеков, что способствует креативности и обогащает коммуникацию.
* Уменьшение размера словаря

Недостатки:

Неоднозначность: Омонимия может создавать неоднозначность в тексте или разговоре, так как одно и то же слово может иметь несколько значений. Это может затруднить точное и ясное выражение мыслей.

#### 

#### **20. Синонимия: определение, виды, примеры, достоинства и недостатки.**

Синонимия - это явление в лингвистике, когда два или более слова имеют сходное или почти одинаковое значение. Синонимы могут быть существительными, прилагательными, глаголами и другими частями речи. Они помогают обогатить язык, делая выражение более разнообразным и точным. Однако, иногда между синонимами есть тонкие различия в использовании или контексте, в котором они употребляются.

Виды синонимии:

* Полная синонимия: Слова, имеющие абсолютно одинаковые значения. Пример: "животное" и "зверь".
* Частичная синонимия: Слова, которые схожи в значении, но имеют и некоторые различия. Например, "начальник" и "руководитель" имеют схожее значение, но употребляются в разных контекстах.
* Синонимия с различиями в стилистическом оттенке: Слова с похожими значениями, но употребляемые в разных стилях речи или с разным эмоциональным окрасом. Например, "прекрасный" и "великолепный".

Примеры:

* Синонимы с полной синонимией:
  + "печь" и "духовка"
  + "радостный" и "веселый"
  + "красивый" и "прекрасный"
* Синонимы с частичной синонимией:
  + "здоровье" и "благополучие"
  + "мысль" и "идея"
  + "страна" и "государство"
* Быстрый и скорый: Оба описывают высокую скорость движения.
* Смех и хохот: Оба относятся к выражению радости или веселья.

Достоинства синонимии:

* Расширение словарного запаса: Позволяет избегать повторений и делает речь более разнообразной.
* Точность в выражении: Позволяет выбрать наиболее подходящее слово для передачи смысла.

Недостатки синонимии:

* Смысловые оттенки: Иногда синонимы могут не совсем точно совпадать по значению, что может привести к недопониманию.
* Контекст зависимости: Использование синонимов требует учета контекста, чтобы подобрать самое подходящее слово.

Синонимия обогащает язык, позволяя выбирать различные слова для передачи одного и того же значения. Однако, точный выбор между синонимами может быть ключевым для точного и четкого выражения мысли.

#### **21. Автоматизация обработки текста: этап семантического анализа.**

Семантический анализ — это этап обработки текста, направленный на выявление смысла слов, предложений и текста в целом.

Цели семантического анализа

1. Установление связей между словами, фразами и предложениями, чтобы определить, о чём говорится.
2. Выявление контекста: Учет многозначности слов и фраз, чтобы правильно интерпретировать их значения в зависимости от контекста.

Одним из подходов к семантическому анализу текстов естественного языка (ТЕЯ) является подход, базирующийся на синтаксисе. В данном случае на основе синтаксического дерева предложения и смысла слов строится соответствующая предложению предикатная формула.

Задачей в таком случае является извлечение из синтаксических деревьев в виде отношений наиболее значимых с точки зрения знаний основных типов, синтаксических структур: 1) Simple Noun Phrase (простая именная группа); 2) Verb Phrase (глагольная группа); 3) Noun Phrase Additional (именная группа, распространенная различными оборотами); 4) Complex Sentence (сложноподчиненное предложение). По ним строится лингвистический индекс (LI), который формально может быть представлен в виде,  где W – множество слов текста; POS, SYN и REL – отображения слов в множества их соответственно лексико-грамматических и синтаксических классов (меток), а также меток семантико-синтаксических отношений, предопределяющих знания основных типов. Такая модель обладает одним очень важным свойством – она допускает естественное включение в себя новых компонент в соответствии с разрабатываемым приложением базового ЛП.

Для представления результатов семантического анализа чаще всего используют предикаты первого порядка, семантические сети, фреймы и онтологии.

Следует отметить, что синтаксические структуры не всегда хорошо подходят для целей семантического анализа. Так, например, синтаксические деревья часто содержат много связей, которые не играют существенной роли для семантического анализа ТЕЯ. Решением подобных проблем является использование семантических грамматик. Правда, такие грамматики ориентированы на семантический анализ определенных предметных областей и правила составляются таким образом, чтобы необходимые семантические единицы описывались внутри одного правила. Эти правила предназначены сугубо для определенного (необходимого) уровня семантического анализа.

Ключевой проблемой на этапе семантического анализа текста ЕЯ является проблема многозначности слов. Одним из классов алгоритмов разрешения данной проблемы являются обучающиеся алгоритмы, основанные, например, на анализе контекста многозначных слов, либо так называемые самонастраивающиеся алгоритмы.

#### **22. Типовая структура базы знаний для решения задач автоматической обработки текста естественного языка.**

Структура базы знаний для решения задач обработки текста естественного языка может быть разнообразной и зависит от конкретной задачи или приложения. Однако, в общем случае, база знаний для NLP может включать следующие основные элементы:

Основные компоненты базы знаний

1. Лексическая база (словарь): Содержит слова и выражения языка с их свойствами.

- Леммы и формы слов (например, глагол "идти" и его формы: "иду", "шёл").

- Части речи (существительное, глагол, прилагательное и т. д.).

- Синтаксическая информация (спряжение, склонение).

- Фонетическая информация (ударение, транскрипция).

2. Семантическая база (семантические сети): Описывает значения слов, их синонимы, антонимы и гиперонимы (например, "собака" — это "животное"). Семантические отношения

3. Онтология: Иерархическое представление знаний о мире в терминах сущностей, их свойств и отношений.

4. Синтаксические правила и шаблоны: Описывают синтаксическую структуру предложений. Используются для парсинга текста и извлечения грамматических зависимостей.

- Примеры:

- Правила для согласования подлежащего и сказуемого.

- Шаблоны для выделения ключевых словосочетаний, таких как "прилагательное + существительное".

5. Прагматическая информация: Сведения о том, как контекст влияет на смысл текста. Примеры: Интерпретация местоимений ("он" = "Иван")

6. Корпусы текстов. могут быть размеченными (аннотированными) и неразмеченными.

7. Модули для анализа тональности и эмоций: Содержат информацию о полярности слов (позитивной, негативной или нейтральной) и их эмоциональной окраске.

8. Метаданные и индексы:

- Метаданные текстов: Информация о документах, времени создания, авторстве и другие свойства, которые могут быть полезными при анализе текста.

- Индексы для быстрого поиска: Структуры данных, ускоряющие поиск информации в больших объемах текста.

**23. Классификация систем машинного перевода текстов.**

В настоящее время различают три вида систем машинного перевода:

* Системы на основе грамматических правил (Rule-Based Machine Translation, RBMT);
* Статистические системы (Statistical Machine Translation, SMT);
* Гибридные системы;

Системы на основе грамматических правил производят анализ текста, который используется в процессе перевода. Перевод производится на основе встроенных словарей для данной языковой пары, а также грамматик, охватывающих семантические, морфологические, синтаксические закономерности обоих языков. На основе всех этих данных исходный текст последовательно, предложение за предложением, преобразуется в текст на требуемом языке. Основной принцип работы таких систем — связь структур исходного и конечного текстов.

Преимуществами систем на основе грамматических правил являются грамматическая и синтаксическая точность, стабильность результата, возможность настройки на специфическую предметную область. К недостаткам систем на основе грамматических правил относят необходимость создания, поддержки и обновления лингвистических баз данных, трудоемкость создания такой системы, а так же ее высокая стоимость.

Статистические системы при своей работе используют статистический анализ. В систему загружается двуязычный корпус текстов (содержащий большое количество текста на исходном языке и его «ручной» перевод на требуемый язык), после чего система анализирует статистику межъязыковых соответствий, синтаксических конструкций и т. д. Система является самообучаемой — при выборе варианта перевода она опирается на полученную ранее статистику. Чем больший словарь внутри языковой пары и чем точнее он составлен, тем лучше результат статистического машинного перевода. С каждым новым переведенным текстом улучшается качество последующих переводов.

Статистические системы отличаются быстротой настройки и легкостью добавления новых направлений перевода. Среди недостатков наиболее значительными являются наличие многочисленных грамматических ошибок и нестабильность перевода.

Гибридные системы сочетают в себе подходы, описанные ранее. Ожидается, что гибридные системы машинного перевода позволят объединить все преимущества, которыми обладают статистические системы и системы, основанные на правилах.

#### 

#### **24. Уровни изучения текста, связь с разделами лингвистики.**

1. **Фонетический уровень – фонетика**  
Фонетический уровень анализа текста фокусируется на звуковых аспектах языка, таких как произношение звуков, их длительность и интонация.

2. **Морфологический уровень – морфология**

Морфологический анализ текста изучает структуру слов, включая морфемы, приставки, суффиксы и другие морфологические элементы.

3. **Лексический уровень – лексикология**

Исследует значение слов и их использование в языке. Распознавание ключевых слов и тематических доменов текста.

4. **Синтаксический уровень –синтаксис**  
Синтаксический уровень анализа фокусируется на структуре предложений, порядке слов и синтаксических связях между словами.

4. **Семантический уровень –семантика**Семантический анализ текста изучает значения слов, фраз и предложений, а также способы передачи смысла в тексте.

5. **Стилистический уровень изучения текста – стилистика**

Стилистический уровень анализа текста исследует его языковую форму с точки зрения экспрессивных, эстетических и функциональных характеристик. Этот уровень изучает, как выбор языковых средств влияет на восприятие текста, его тональность, эмоциональную окраску и соответствие жанровым или функциональным нормам.

#### **25. Задача информационного поиска.**

Информационный поиск — это процесс поиска в большой коллекции

(хранящейся, как правило, в памяти компьютеров) некоего неструктурированного

материала (обычно — документа), удовлетворяющего информационные

потребности.

Основная задача любой поисковой системы – дать пользователю ответ на

его запрос или другими словами предоставить список документов релевантных

запросу пользователя.

Подзадачи:

1. Индексация: Преобразование документов (текстов, изображений, видео и т. д.) в структурированный формат, который обеспечивает быстрый и эффективный поиск.
2. Поисковые запросы: Разработка методов для интерпретации и анализа запросов пользователей, в том числе учет синонимов, многозначности и других лингвистических особенностей.
3. Модели поиска: Разработка моделей, определяющих, какие документы наилучшим образом соответствуют запросу пользователя. К этим моделям относятся, например, векторные пространства, вероятностные модели и машинное обучение.
4. Ранжирование: Оценка и упорядочение документов в соответствии с их релевантностью для запроса пользователя. Ранжирование может зависеть от различных факторов, включая структуру запроса, а также релевантность и авторитетность документов.
5. Обработка естественного языка: Использование методов обработки естественного языка для понимания семантики запросов и документов, что улучшает точность и релевантность результатов поиска.

#### **26. Задача информационного поиска. Документальные системы информационного поиска.**

Информационный поиск — это процесс поиска в большой коллекции

(хранящейся, как правило, в памяти компьютеров) некоего неструктурированного

материала (обычно — документа), удовлетворяющего информационные

потребности.

Классическая задача ИП, с которой началось развитие этой области, – это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов. Но список задач ИП постоянно расширяется и теперь включает: вопросы моделирования, классификацию и фильтрацию документов, проектирование архитектур поисковых систем и пользовательских интерфейсов, извлечение информации, построение языков запросов и др.

Документальные системы информационного поиска (Document Information Retrieval Systems) представляют собой специализированные информационные системы, разработанные для хранения, организации и извлечения документов различных типов, таких как текстовые, мультимедийные, аудио и видеозаписи и другие. Они обеспечивают пользователям возможность эффективного поиска и доступа к информации с помощью разнообразных инструментов и методов.

Документальные системы информационного поиска включают в себя компоненты, такие как индексация документов, создание поисковых запросов, анализ результатов поиска и представление информации пользователям. Они также могут включать функции автоматической категоризации и рубрикации документов, фильтрацию результатов поиска и другие инструменты для повышения эффективности и точности поиска.

Эти системы широко используются в различных областях, включая библиотеки, научные исследования, управление документами в организациях, интернет-поиск и другие. Они играют важную роль в обеспечении доступа к огромному объему информации и повышении производительности пользователей при поиске и обработке документов.

Некоторые известные документальные системы информационного поиска:

* Microsoft SharePoint: Популярная платформа для управления документами, которая включает инструменты для хранения, поиска и обмена документами в организации.
* Apache Solr: Инструмент для поиска и анализа текстовых и мультимедийных данных, основанный на Apache Lucene.
* Elasticsearch: Распределенная система для поиска и анализа данных, которая предоставляет возможности поиска, аналитики и визуализации для различных типов документов.
* IBM Watson Discovery: Инструмент для извлечения и анализа информации из текстовых данных, включая структурированные и неструктурированные документы.

#### **27. Задача информационного поиска. Фактографические системы информационного поиска.**

Информационный поиск — это процесс поиска в большой коллекции

(хранящейся, как правило, в памяти компьютеров) некоего неструктурированного

материала (обычно — документа), удовлетворяющего информационные

потребности.

**Фактографические системы информационного поиска:**

Фактографические системы информационного поиска — это системы, ориентированные на извлечение фактов из структурированных и полуструктурированных источников данных. Они направлены на предоставление конкретных, проверенных фактов, а не просто релевантных документов.

Такие системы могут использовать методы обработки естественного языка, машинного обучения и семантического анализа для выделения ключевых фактов из текстов и баз данных. Это особенно полезно в сценариях, где требуется точная и конкретная информация, например, при ответах на вопросы или выполнении задач вроде составления фактографий.

Применение фактографических систем информационного поиска охватывает множество областей, включая медицину, научные исследования, техническую поддержку и другие сферы, где необходим доступ к конкретным фактам для принятия решений.

#### **28. Документальные системы информационного поиска. Поисковая система Google.**

Документальные системы служат для работы с документами, в которых информация хранится в виде текстовых документов (статьи, книги, рефераты, тексты законов) и графических объектов, снабженная тем или иным формализованным аппаратом поиска. Цель системы, как правило, – выдать в ответ на запрос пользователя список документов или объектов, в какой-то мере удовлетворяющих сформулированным в запросе условиям.

Автоматизация процесса информационного поиска потребовала формализации представления основного смыслового содержания документов в виде *поискового предписания* (ПП) и *поисковых образов документов* (ПОД). Поисковый образ документа (ПОД) — описание документа, в виде перечня

ключевых слов, которые могут дополняться их весами, связями и указателями

роли. По этому описанию внутри системы составляются структуры данных,

служащие для поиска документов и выдачи их из хранилищ. Такое же описание

строится для пользовательского запроса.

В процессе информационного поиска определяется степень соответствия содержания документов и запроса пользователя путем сопоставления ПОД и ПП. Решение о выдаче или не выдаче документа в ответ на запрос принимается на основе некоторого набора правил, по которому данная ИПС определяет степень смысловой близости между ПОД и ПП. Такой набор правил получил название *критерия смыслового соответствия* (КСС). Критерий может быть задан явно или неявно и базируется на понятии формальной релевантности ПОД и ПП.

*Релевантность* – соответствие содержания документа информационному запросу в том виде, в котором он сформулирован.

*Фактическая релевантность*, понимаемая как смысловое соответствие содержания выданного документа информационному запросу, может быть установлена человеком в процессе осмысления содержания документа и запроса.

Элементами ИПС являются: *Массивы документов,* выступающие в качестве объекта поиска

*Информационно – поисковый язык* (ИПЯ). Это искусственный язык предназначенный для описания содержания и формы документов и запросов для осуществления поиска.

*Правила индексирования*, следуя которым осуществляется описание средствами ИПЯ документов и запросов. В результате индексирования документа получается поисковый образ документа (ПОД), а, в результате индексирования запроса – поисковое предписание (ПП).

*Правила поиска* документов, соответствующих запросов, которые задаются в виде критерия соответствия.

*Технические средства*, с помощью которых реализуется ИПС, т.е. осуществляется

хранение и поиск информации.

Обслуживающий персонал – индексаторы и технические работники, обеспечивающие обработку и ввод в систему документов.

В состав ИПС входят четыре подсистемы: ввода и регистрации, обработки, хранения, поиска.

Документальные системы информационного поиска (Document Information Retrieval Systems) представляют собой специализированные информационные системы, разработанные для хранения, организации и извлечения документов различных типов, таких как текстовые, мультимедийные, аудио и видеозаписи и другие. Они обеспечивают пользователям возможность эффективного поиска и доступа к информации с помощью разнообразных инструментов и методов.

Документальные системы информационного поиска включают в себя компоненты, такие как индексация документов, создание поисковых запросов, анализ результатов поиска и представление информации пользователям. Они также могут включать функции автоматической категоризации и рубрикации документов, фильтрацию результатов поиска и другие инструменты для повышения эффективности и точности поиска.

**Google**- одна из самых полных и популярных зарубежных ИПС. Отличительной особенностью ИПС Google является технология определения степени релевантности документа путем анализа ссылок других источников на данный ресурс. Чем больше ссылок на какую-либо страницу имеется на других страницах, тем выше ее рейтинг в ИПС Google. Google использует алгоритм расчёта авторитетности PageRank. PageRank является одним из вспомогательных факторов при ранжировании сайтов в результатах поиска. PageRank не единственный, но очень важный способ определения положения сайта в результатах поиска Google. Google использует показатель PageRank найденных по запросу страниц, чтобы определить порядок выдачи этих страниц посетителю в результатах поиска.В 2010 году компания запустила голосовой поиск в России. Чтобы осуществить поиск, необходимо нажать в телефоне кнопку рядом со строкой поиска и произнести свой запрос, телефон отправит ваш голос на сервер, и браузер выдаст строку с распознанным вашим запросом и результатами поиска по нему.

#### **29. Проблемы машинного перевода.**

Машинный перевод имеет ряд проблем и ограничений, которые могут влиять на качество и точность перевода. Некоторые из основных проблем включают:

1. Сложности в точном переводе контекста и смысла:

* Полисемия и омонимия: Слова с несколькими значениями могут иметь разные переводы в зависимости от контекста.
* Идиомы и фразеологизмы: Значения, основанные на культурных и языковых особенностях, могут быть трудно перевести.
* Неоднозначность: Некоторые предложения могут иметь несколько интерпретаций, что затрудняет точный перевод.

2. Отсутствие контекстуального понимания:

* Неучтенный контекст: Машины часто не учитывают широкий контекст предложения, что может привести к неправильному переводу, особенно при переводе длинных текстов.

3. Недостаток данных для обучения:

* Недостаточные данные: Ограниченный доступ к параллельным текстовым данным на различных языках может снизить качество обучения моделей машинного перевода.
* Дисбаланс в данных: Неравномерное распределение данных между разными языками может привести к смещенным или менее точным результатам.

4. Проблемы с грамматикой и структурой языка:

* Различия в грамматике: Различные языки имеют разные грамматические структуры, что затрудняет точный перевод.
* Сложности с морфологией: Некоторые языки имеют сложную морфологию (например, изменение форм слов), что увеличивает сложность перевода.

**5. Специализированные термины и жаргон**: В специализированных областях (медицина, техника и т.д.) машины могут ошибаться из-за отсутствия соответствующего словарного запаса.

6. Адаптация к стилю и тону:

* Передача стиля: Перевод может терять стилистические особенности оригинала, что делает его менее точным или эмоционально выразительным.

Улучшение машинного перевода активно исследуется и продвигается с помощью методов глубокого обучения, адаптивных моделей и большего доступа к данным. Однако эти проблемы все еще остаются вызовом для достижения более точного и качественного перевода.

#### **30. Системы непрямого машинного перевода. Их лингвистическое обеспечение.**

**Система непрямого машинного перевода**, также известная как система перевода с использованием промежуточного языка или языковой структуры, отличается от прямого машинного перевода тем, что она включает этап посредник языка или языковой структуры в процессе перевода. В этой системе исходный текст сначала переводится на промежуточный язык или представляется в виде языковой структуры, после чего этот промежуточный материал переводится на целевой язык.

Машинный перевод Интерлингва (перевод осуществляется через промежуточную (семантическую) модель текста исходного языка, не зависящую от конкретного языка модель, из которой может быть сгенерирован перевод на любой язык.)

Машинный перевод Трансфер (основан на идее Interlingua с использованием сопоставительного анализа двух языков. Сначала текст на исходном языке переводится в абстрактную или промежуточную модель исходного языка, которая затем преобразуется в модель целевого языка, чтобы потом окончательно оформиться в текст на целевом языке)

Такой подход часто используется при работе с несвязанными языками, для которых сложно разработать прямую модель перевода из-за значительных различий в лексике, грамматике или структуре предложений. У этой системы есть свои преимущества и недостатки: она может обрабатывать большой спектр языков и помогает избежать проблем, связанных с прямым переводом между несвязанными языками, однако может потребовать больше времени и ресурсов из-за необходимости дополнительного этапа перевода на промежуточный язык.

**Лингвистическое обеспечение автоматизированной системы** это совокупность средств и правил для формализации естественного языка, используемых при общении пользователей и эксплуатационного персонала автоматизированной системы с комплексом средств автоматизации.

Для успешного функционирования систем непрямого перевода требуется богатое лингвистическое обеспечение, включающее следующие компоненты

1. Лексические ресурсы Билингвальные словари для соответствия слов и фраз между языками. Многоязычные онтологии для отображения универсальных понятий (например, "дом" как "место для жилья"). Корпусы текстов: Размеченные тексты на обоих языках для обучения и проверки системы.

2. Грамматические базы Морфологические правила.Синтаксические правила.

Семантические связи

3. Межъязыковая база (интерлингва): Универсальный язык или представление, например:

Упрощённый вид: "субъект — действие — объект" (SVO).

Интерлингва с логической формализацией (например, Universal Networking Language).

4. Правила трансфера:

Определяют, как преобразовать структуру и значения исходного языка в целевой язык.

Пример: Преобразование русского порядка слов (свободный) в английский (фиксированный, SVO).

5. Прагматические и дискурсивные модели:

Учитывают контекст, намерение и дискурс.

Пример: Перевод местоимений с учетом предшествующих предложений ("он" → "he" или "it").

6. Статистические и нейронные модели (в современных системах):

Используются как дополнение к традиционным подходам для учета сложных закономерностей.

Пример: Использование нейронных сетей для уточнения перевода при помощи анализа больших корпусов текстов.

#### **31. Автоматизация обработки текста: этап синтаксического анализа.**

**Автоматизация обработки текста** означает использование компьютерных систем и алгоритмов для выполнения различных задач, связанных с обработкой и анализом текстовой информации. Этот процесс включает в себя применение технологий и методов машинного обучения, обработки естественного языка (Natural Language Processing, NLP), и других инструментов, чтобы сделать обработку текста более эффективной и автоматизированной.

Задачей синтаксического анализа текста ЕЯ является распознавание в каждом его предложении синтаксических отношений и представление их, как правило, в виде функционального или синтаксического дерева, в котором словам предложения указывается их грамматическая функция и определяется тип синтаксической связи между ними.

Что касается алгоритмов синтаксического анализа, то большинство из них реализуется в соответствии с некоторой формальной грамматикой. Наиболее исследованными теоретически и получившими широкое применение в практических приложениях являются грамматики Хомского, в частности контекстно-свободные (КС) грамматики [7], трансформационные и GB-грамматики. Последние описывают два уровня синтаксической структуры (глубинный и поверхностный), связанных посредством трансформаций.

Грамматики Хомского по мощности оказались избыточными для описания синтаксиса ЕЯ и в то же время недостаточными для учета семантических аспектов языка. КС-грамматики представляются в виде продукций (правил), ставящих в соответствие нетерминальным символам в своих левых частях (до знака «=») набор терминальных и нетерминальных символов в правых частях.

Были разработаны вероятностные КС-грамматики, в которых каждому правилу ставится в соответствие некоторый параметр, отражающий вероятность того, что нетерминальный символ в левой части правила может быть расширен данным правилом. Такой механизм обеспечивает одно из возможных решений проблемы многозначности синтаксических деревьев. Были разработаны HPSG (Head-Driven Phrase Structure Grammar) [61], SFG (Systemic Functional Grammar) [62], LFG (Lexical Functional Grammar) [63], которые позволяют учитывать семантические признаки и даже всю лингвистически релевантную информацию на этапе синтаксического анализа текста.

#### **32. Задача автоматического реферирования документов. Метод построения классического реферата.**

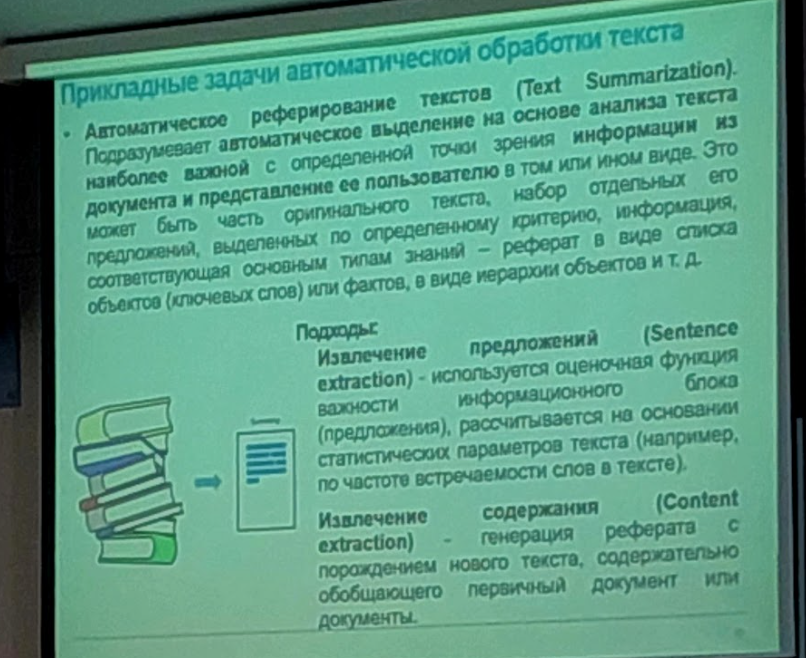
виды реферата

Классический реферат – это набор наиболее информативных предложений

текста, возможно трансформированных (удаление вводных конструкций, замена

анафоричных местоимений и т.д. с целью улучшения связности реферата и уменьшения

его объема).



**Задача автоматического реферирования документов** заключается в создании краткого обзора или резюме содержания документа, сохраняя ключевую информацию и основные идеи.

Классический метод построения реферата включает несколько этапов:

1. Извлечение ключевых предложений или фраз:

Идентификация важных фрагментов: Алгоритмы выделения ключевых предложений определяют наиболее информативные фрагменты текста.

1. Определение важности предложений:

Оценка значимости: Предложения оцениваются по различным критериям, таким как частота встречаемости ключевых слов, структурные особенности или их связь с заголовками, чтобы определить их важность в контексте текста.

1. Выбор ключевых предложений:

Формирование реферата: На основе оценки значимости предложений выбираются наиболее информативные и важные предложения для включения в реферат.

1. Сводка и генерация реферата:

Создание краткого обзора: Выбранные предложения или фразы собираются в сводку или реферат, сохраняя основную суть и ключевые аспекты исходного текста.

1. Проверка качества:

Оценка точности и качества: Результаты реферирования могут быть оценены с помощью метрик, таких как ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy) и других для оценки сходства реферата с исходным текстом.

Применение методов машинного обучения:

Современные методы автоматического реферирования также используют методы машинного обучения и нейронные сети для более точного выбора и формирования реферата, что позволяет улучшить качество и точность резюмирования текста.

Этот процесс требует анализа текста, оценки его значимости и способности выделения ключевой информации, чтобы создать сжатое, но информативное представление оригинального документа.

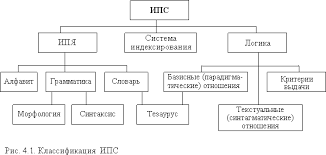
#### 

#### **33. Задача информационного поиска. Структурно-функциональная схема системы информационного поиска.**

Информационный поиск (ИП) — это процесс поиска релевантной информации (документов, данных, текстов) в большой коллекции данных на основе запросов пользователя. Целью системы информационного поиска является предоставление пользователю информации, наиболее соответствующей его запросу, с учетом содержания, контекста и релевантности.

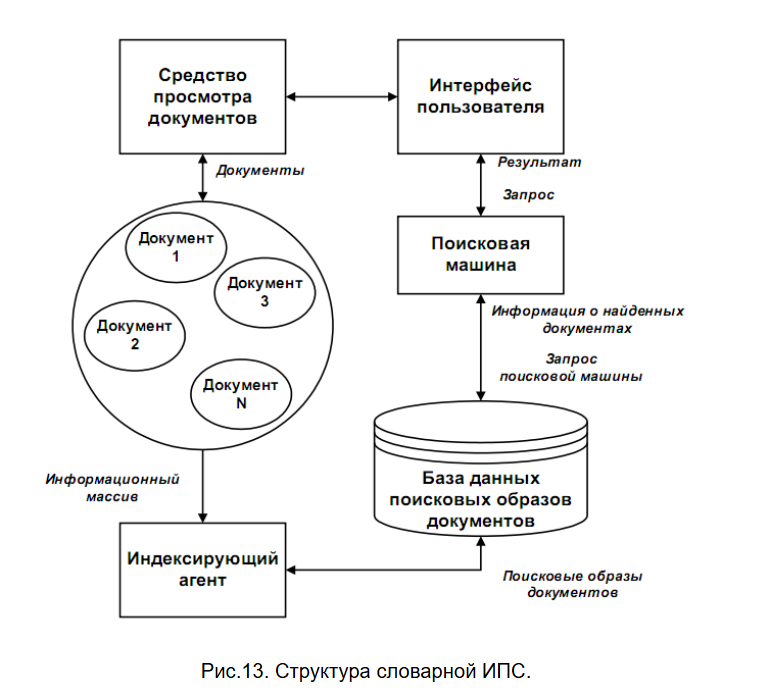
В настоящее время под ИПЯ понимается формализованная семантическая система, обеспечивающая передачу (запись) информации в объеме, необходимом для цели поиска. В качестве стандартизируемых единиц используют ключевые слова или ключевые слова в некоторых стандартизируемых выражениях (стандартизированные предложения запросов).

тезаурус Словарь, целью к-рого является наиболее полное отражение лексики того или иного языка. Словарь, отражающий систематизированную совокупность понятий какой-л. отрасли знаний.



Парадигматические отношения – это отношения между единицами одного уровня, которые объединяются в классы на основе какого- то тождества (т.е. наличия какого-то сходства) и различия. Члены одной парадигмы противопоставлены друг другу, находятся в отношении оппозиции. например, парадигма падежных форм существительного, парадигма временных форм глагола и т.д.

Синтагматические отношения. Синтагматика – это сочетаемость единиц одного уровня друг с другом и последовательность единиц одного уровня в речи. Сочетание двух слов образует словосочетание. Словосочетания могут входить в состав предложения, сочетаясь друг с другом по определенным правилам. Синтагматические отношения сравниваются часто с логическими отношениями конъюнкции (отношения «и – и»).



База документов может пред­ставлять собой простую совокупность файлов, распределенную по каталогам жесткого диска. Однако такой тип представления базы документов характеризуется двумя недос­татками: неэффективным использованием дискового пространства, низкой скоростью доступа при большом количестве файлов.

Поэтому для хранения документов применяют средства сжатия и быстрого поиска информации. В этом случае подсистема хранения представляет собой совокупность стан­дартных или специализированных средств архивации, СУБД и т.п., обеспечивающих воз­можность доступа к данным по предъявляемому идентификатору.

Далее документы поступают на вход подсистемы обработки, задачей которой являет­ся формирование для каждого документа ПОД, в который заносится информация, необхо­димая для последующего поиска документа.

При поступлении на вход системы запроса пользователя он преобразуется в ПП и пе­редается в подсистему поиска, задачей которой является отыскание в индексе ПОД, удов­летворяющих ПП с точки зрения КСС. Идентификаторы релевантных документов пода­ются с выхода подсистемы поиска на вход подсистемы хранения, которая осуществляет выдачу пользователю самых релевантных документов.

#### \*\*1. Источник данных\*\*

- Коллекция документов, текстов, данных (например, веб-страницы, базы данных).

- Данные могут быть структурированными (таблицы), полуструктурированными (XML, JSON) или неструктурированными (тексты, изображения).

#### \*\*2. Индексация данных\*\*

- \*\*Цель\*\*: Создание структурированных индексов для быстрого доступа к документам.

- \*\*Процессы\*\*:

- \*\*Предварительная обработка текста\*\*:

- Удаление стоп-слов, нормализация текста (лемматизация, стемминг).

- \*\*Разбиение текста\*\*:

- На токены (слова, фразы).

- \*\*Создание индекса\*\*:

- Инвертированный индекс (сопоставление терминов с документами).

- \*\*Хранение метаданных\*\*:

- Сведения о длине документов, частотах терминов.

#### \*\*3. Интерфейс пользователя\*\*

- Компонент, через который пользователь вводит запрос и получает результаты.

- \*\*Функции\*\*:

- Обработка запросов естественного языка.

- Подсказки для составления запроса.

- Визуализация результатов.

#### \*\*4. Обработка запросов\*\*

- \*\*Цель\*\*: Преобразовать пользовательский запрос в формат, понятный системе.

- \*\*Процессы\*\*:

- Нормализация текста запроса (лемматизация, стемминг).

- Расширение запроса (учёт синонимов, гипонимов).

- Построение поискового выражения (логические операторы, вес терминов).

#### \*\*5. Модуль поиска\*\*

- \*\*Цель\*\*: Сравнение запроса с индексом и получение релевантных документов.

- \*\*Методы поиска\*\*:

- \*\*Точное совпадение\*\*:

- Выдача документов, содержащих запрос в явной форме.

- \*\*Поиск по релевантности\*\*:

- Выдача документов, релевантных запросу по смыслу (с учётом частотности, веса слов).

- \*\*Контекстный поиск\*\*:

- Учет дополнительной информации (пользовательские предпочтения, геолокация).

#### \*\*6. Модуль ранжирования\*\*

- \*\*Цель\*\*: Упорядочить результаты по степени релевантности.

- \*\*Методы ранжирования\*\*:

- \*\*TF-IDF (Term Frequency-Inverse Document Frequency)\*\*:

- Оценка важности термина в документе.

- \*\*Модели на основе машинного обучения\*\*:

- Учет множества факторов (времени публикации, ссылок, поведенческих данных).

- \*\*Нейронные сети\*\*:

- Учет смысла текста (например, Word2Vec, BERT).

#### \*\*7. Выдача результатов\*\*

- Формирование списка найденных документов и их визуализация.

- \*\*Элементы результата\*\*:

- Заголовок, краткое описание (сниппет), ссылка на полный текст.

- Дополнительные сведения: релевантность, категория, дата.

#### \*\*8. Модули обратной связи и обучения\*\*

- \*\*Цель\*\*: Улучшение качества поиска на основе анализа поведения пользователей.

- \*\*Функции\*\*:

- Анализ кликов по результатам.

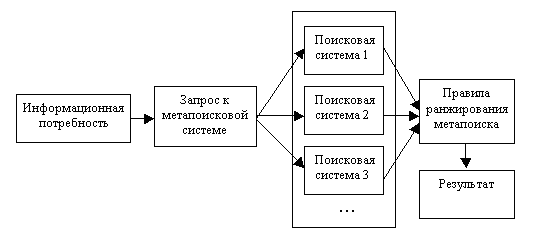
- Учёт времени, проведённого на странице.

- Использование рейтингов документов.

#### **34. Задача информационного поиска. Метапоисковые средства поиска, особенности построения и функционирования.**

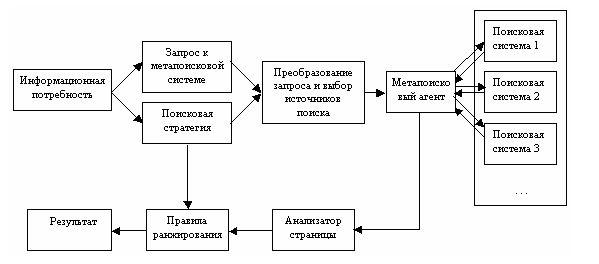
Классическая задача информационного поиска - это поиск документов, удовлетворяющих запросу в рамках некоторой статической (на момент выполнения поиска) коллекции документов.

При работе метапоисковой системы из полученного от поисковых систем множества документов необходимо выделить наиболее релевантные, то есть соответствующие запросу пользователя. Простейшие метапоисковые системы реализуют стандартный подход, представленный на схеме ниже.



В таких системах анализ полученных описаний документов не производится, что может поставить нерелевантные документы, идущие первыми в одной поисковой системе, выше релевантных в другой, чем существенно понизить качество самого поиска.

При разработке следующего поколения метапоисковых систем были учтены недостатки, присущие стандартным метапоисковым системам. Были созданы системы с возможностью выбора тех поисковых машин, в которых, по мнению пользователя, он с большей вероятностью может найти то, что ему нужно. Схема метапоисковой системы следующего поколения представлена ниже.



#### **35. Задача информационного поиска. Метапоисковые утилиты, особенности функционирования.**

Задача информационного поиска заключается в обеспечении пользователей необходимой и релевантной информацией из обширных данных, которые могут включать в себя текстовые документы, изображения, видео, веб-страницы и другие формы контента.

Метапоисковые утилиты представляют собой инструменты, объединяющие результаты поиска из различных поисковых систем или баз данных и предоставляющие пользователю совокупные результаты. Вот некоторые особенности функционирования метапоисковых утилит:

* **Параллельный поиск:**
  + Одной из основных особенностей метапоисковых утилит является параллельный поиск. Они отправляют запросы к нескольким поисковым системам одновременно, ускоряя процесс поиска и обеспечивая более полные результаты.
* **Интеграция результатов:**
  + Метапоисковые утилиты агрегируют результаты из различных источников и представляют их в удобном для пользователя формате, обычно в виде списка или таблицы.
* **Устранение дубликатов:**
  + Для повышения качества результатов, метапоисковые утилиты могут автоматически обнаруживать и удалять дубликаты, предотвращая повторную выдачу одних и тех же результатов.
* **Фильтрация и сортировка:**
  + Пользователям предоставляются инструменты для фильтрации и сортировки результатов поиска. Это может включать в себя фильтрацию по типу контента, дате публикации, рейтингу и т.д.
* **Поддержка различных источников:**
  + В дополнение к поисковым системам, метапоисковые утилиты могут интегрировать результаты из других источников, таких как социальные сети, онлайн-магазины, новостные сайты и т.д.
* **Оптимизация запросов:**
  + Метапоисковые утилиты могут использовать различные техники для оптимизации запросов пользователя, включая автоматическое исправление опечаток, расширение синонимов и дополнительные поисковые подсказки.
* **Метрики качества:**
  + Некоторые утилиты могут использовать метрики для оценки качества результатов отдельных поисковых систем и ранжирования их в соответствии с релевантностью.

**36. Задача информационного поиска. Метапоисковые машины, особенности функционирования.**

Задача информационного поиска заключается в обеспечении пользователей необходимой и релевантной информацией из обширных данных, которые могут включать в себя текстовые документы, изображения, видео, веб-страницы и другие формы контента.

Метапоисковые машины - это компьютерные системы, которые обрабатывают запросы пользователей и осуществляют поиск информации в нескольких различных поисковых системах одновременно. Они объединяют результаты от разных поисковых систем и представляют пользователю скомпилированный список результатов. Вот некоторые особенности функционирования метапоисковых машин:

1. **Объединение результатов**: Основная особенность метапоисковых машин - это объединение результатов от разных поисковых систем в единый список. Они отправляют запросы к разным поисковым системам, получают результаты и комбинируют их в единый ответ, который предоставляется пользователю.

2. **Ранжирование результатов**: Метапоисковые машины могут использовать различные алгоритмы для ранжирования результатов от разных поисковых систем. Они могут учитывать различные факторы, такие как релевантность, авторитетность и популярность страниц, чтобы определить порядок результатов в общем списке.

3. **Устранение дубликатов**: Важным аспектом работы метапоисковых машин является устранение дубликатов. Поскольку один запрос может быть отправлен нескольким поисковым системам, результаты могут содержать одни и те же страницы. Метапоисковые машины могут использовать различные методы для определения и удаления дубликатов, чтобы предоставить пользователю уникальные результаты.

4. **Обработка запросов**: Метапоисковые машины могут обрабатывать различные типы запросов, включая текстовые запросы, изображения, видео и другие форматы. Они могут использовать различные алгоритмы и методы для анализа и обработки запросов, чтобы предоставить наиболее релевантные результаты.

5. **Настройка поиска**: Некоторые метапоисковые машины позволяют пользователям настраивать параметры поиска, такие как выбор поисковых систем, фильтров и других параметров. Это позволяет пользователям получать более индивидуализированные результаты, соответствующие их конкретным потребностям.

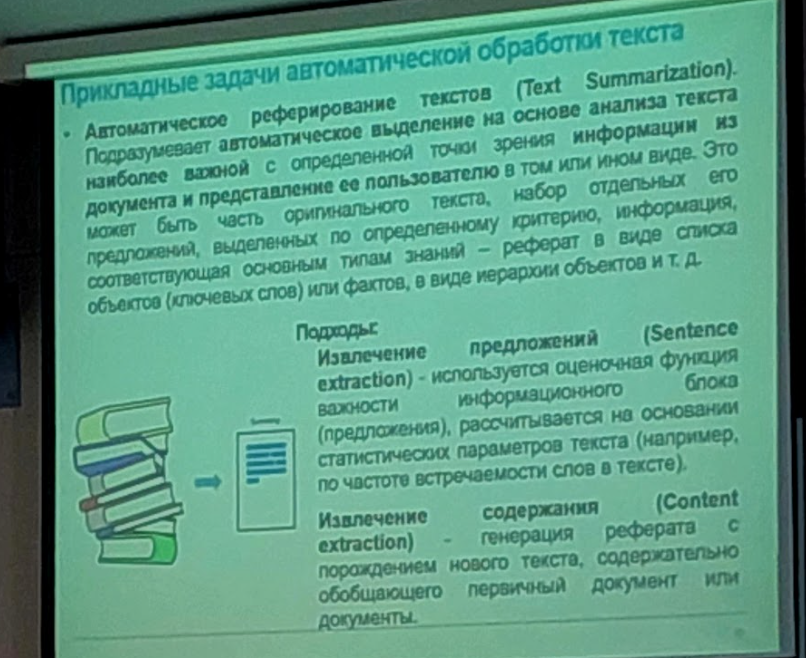
#### **37. Документальные системы информационного поиска. Поисковая система Yandex.**

Документальные системы информационного поиска предназначены для эффективного поиска, извлечения и предоставления информации в ответ на запросы пользователя. Поисковая система Yandex является одной из наиболее популярных в России и странах СНГ. Вот некоторые особенности и характеристики поисковой системы Yandex:

* **Алгоритмы Ранжирования:**
  + Yandex использует собственные алгоритмы ранжирования для определения порядка отображения результатов поиска. Алгоритмы учитывают различные факторы, такие как релевантность страницы, авторитетность ресурса, структура сайта и другие параметры.
* **Персонализированный Поиск:**
  + Система Yandex предоставляет персонализированный поиск, учитывая предыдущие запросы и предпочтения пользователя. Это позволяет системе предоставлять более релевантные результаты в соответствии с интересами каждого конкретного пользователя.
* **Расширенные Возможности Поиска:**
  + Yandex предоставляет дополнительные возможности поиска, такие как поиск изображений, видео, новостей, картинок и многих других типов контента. Это обеспечивает более широкий спектр поисковых запросов.
* **Локальные Сервисы:**
  + Yandex предлагает различные локальные сервисы, такие как Карты, Транспорт, Маркет, Погода и др. Это позволяет пользователям получать разнообразную информацию в одном месте.
* **Обработка Естественного Языка:**
  + Система Yandex использует технологии обработки естественного языка для более точного понимания запросов пользователя и предоставления релевантных результатов.
* **Автоматический Перевод:**
  + Yandex также предоставляет сервис автоматического перевода текстов с и на различные языки. Это может быть полезно для пользователей, работающих с иноязычным контентом.
* **Работа с Внешними Ресурсами:**
  + Поисковая система Yandex производит индексацию веб-страниц и обеспечивает доступ к большому объему информации в интернете. Ранжирование результатов осуществляется на основе алгоритмов поиска.
* **Рекламные Сервисы:**
  + Кроме обычного поиска, Yandex предоставляет рекламные сервисы, такие как Яндекс.Директ, который позволяет рекламодателям размещать рекламу в результатах поиска.

Поисковая система Yandex постоянно развивается, внедряя новые технологии и улучшая свои сервисы для обеспечения более эффективного и удобного поиска информации для пользователей

#### **38. Задача автоматического реферирования документов, подходы к решению.**



В общем случае задача реферирования такова: по одному или нескольким документам на входе нужно составить краткий и беглый реферат, содержащую самую важную информацию из входных документов. В этом определении "краткий" означает, что рефераты должны быть меньше входных документов, "беглый" означает, что они должны быть грамматически корректными и связными. "Важность" информации определяется из контекста и предметной области.

Классический метод построения реферата включает несколько этапов:

Извлечение ключевых предложений или фраз: Идентификация важных фрагментов: Алгоритмы выделения ключевых предложений определяют наиболее информативные фрагменты текста.

Определение важности предложений: Оценка значимости: Предложения оцениваются по различным критериям, таким как частота встречаемости ключевых слов, структурные особенности или их связь с заголовками, чтобы определить их важность в контексте текста.

Выбор ключевых предложений: Формирование реферата: На основе оценки значимости предложений выбираются наиболее информативные и важные предложения для включения в реферат.

Сводка и генерация реферата: Создание краткого обзора: Выбранные предложения или фразы собираются в сводку или реферат, сохраняя основную суть и ключевые аспекты исходного текста.

#### **39. Задача информационного поиска. Поиск по запросу булевого типа.**

**Центральная задача ИП** – помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

**Логическая модель поиска** – это модель поиска информации, в которой вы можете обрабатывать любой запрос в форме логического выражения AND, OR и NOT. Документ рассматривается как серия слов. Если этот термин существует в документе, соответствующая переменная принимает значение «True» (логический блок и документ считается релевантным) или «False» (логический нуль и документ не имеет значения).

Модель обработки логического запроса информации имеет несколько недостатков:

* Многие электронные документы могут выводиться для конкретного запроса, и пользователь должен добавить дополнительные условия для запроса, чтобы снизить результирующую выборку.
* Поиск осуществляется путем проб и ошибок. Поэтому запрос отображается в сложной логической формуле, которая требует от пользователя не только знания предметной области, но и знания правил для создания формул.
* Полученная выборка не может быть оценена по релевантности, так как имеются только два значения релевантности («релевантные» и «нерелевантные»).

Преимущества булевой модели включают предсказуемость результатов информационного поиска.

**40. Задача информационного поиска. Поиск по естественно-языковому запросу.**

Центральная задача ИП – помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

Классическая задача ИП, с которой началось развитие этой области, – это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов. Но список задач ИП постоянно расширяется и теперь включает: вопросы моделирования, классификацию и фильтрацию документов, проектирование архитектур поисковых систем и пользовательских интерфейсов, извлечение информации, построение языков запросов и др.

Поиск по естественно-языковому запросу (Natural Language Processing, NLP) - это метод поиска информации, который позволяет пользователям задавать запросы на естественном языке, подобно тому, как они общаются в повседневной жизни. Это позволяет более естественным образом взаимодействовать с системами поиска без необходимости использовать ключевые слова или специальные запросы.

Технологии обработки естественного языка позволяют компьютерам понимать, интерпретировать и отвечать на человеческий язык. Поэтому, когда пользователь вводит запрос на своем естественном языке, система NLP обрабатывает и анализирует этот запрос, понимая его смысл и выявляя общие темы и ключевые точки. Затем система переводит этот запрос во внутренний формат для поиска, чтобы предоставить наиболее релевантные результаты.

Такие технологии широко применяются в различных областях, включая интернет-поиск, виртуальных помощников, аналитики данных и другие. Они позволяют пользователям получать более точные и релевантные результаты, основанные на их естественном выражении запросов.

#### 

#### 

#### **41. Задача информационного поиска. Векторная модель индексирования и поиска документов.**

Метода к лабораторной 1

Информационный поиск — это процесс поиска в большой коллекции (хранящейся, как правило, в памяти компьютеров) некоего неструктурированного материала (обычно — документа), удовлетворяющего информационные потребности.

Поисковый образ документа (ПОД) — описание документа, в виде перечня ключевых слов, которые могут дополняться их весами, связями и указателями роли. По этому описанию внутри системы составляются структуры данных, служащие для поиска документов и выдачи их из хранилищ. Такое же описание строится для пользовательского запроса.

Поисковый образ запроса (ПОЗ) — описание пользовательского запроса, в виде удобном для поисковой системы. Структура поисковых образов для разных поисковых систем может быть различной, однако поисковый образ запроса и поисковый образ документа должны иметь одинаковую структуру в пределах одной поисковой системы.

Представление документов и поиск информации в массиве разделим на две

модели. Следуя этой логике, векторной будем называть модель описания

информационного массива, а линейной - модель поиска информации в массиве.

В векторной модели информационного потока можно выделить несколько

основных понятий: словарь, документ, поток и процедуры поиска и коррекции

запросов.

Под словарем понимают упорядоченное множество терминов, мощность

которого обозначают как D.

Документ - это двоичный вектор размерности D. Если термин входит в

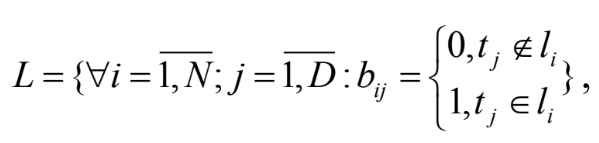
документ, то в соответствующем разряде этого двоичного вектора проставляется 1, в

противном же случае - 0. Обычно все операции в линейной модели индексирования и поиска документов выполняются над поисковыми образами документов, но при этом

их, как правило, называют просто документами.

Информационный поток или массив L представляют в виде матрицы

размерности N×D, где в качестве строк выступают поисковые образы N документов.



где tj – термин;

li – документ;

bij – значение в j-й позиции i-го вектора.

При таком рассмотрении можно определить процедуру обращения к

информационной системе формулой:



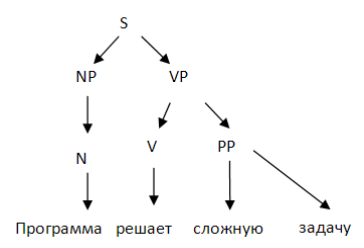
где q – вектор запроса;

r – вектор отклика системы на запрос.

#### **42. Способы описания синтаксической структуры предложения. Достоинства и недостатки.**

В настоящее время можно отметить два основных способа представления синтаксической структуры, вариациями которых можно считать все другие способы. Если использовать термин "грамматика" для обозначения системы связей, то два наиболее популярных способа называются "грамматикой непосредственно составляющих" (constituent grammars) и "грамматикой зависимостей" (dependency grammars).

Грамматика непосредственно составляющих (У нас была как система составляющих) отмечает в предложении наиболее связанные между собой по смыслу слова.



Грамматика непосредственно составляющих построена на факте пространственного следования групп, составляющих предложение, – именных, глагольных, обстоятельственных и т.п. Здесь возможны разные виды связей. Предложение в целом рассматривается как совокупность групп, сводящихся к двум главным, образующим предложение, – группе имени существительного (подлежащего) и группе глагола (сказуемого).

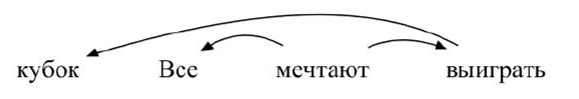
Достоинства грамматики непосредственно составляющих (ГНС):

* Простота понимания и легкость создания.
* Подходит для обучения и определенных задач.

Недостатки ГНС:

* Недостаточная точность в отражении семантики.
* Ограниченная гибкость, особенно для сложных языков.
* Проблемы с амбигвитетом и необработка зависимостей.

Грамматика зависимостей (У нас была как деревья подчинения) базируется на представлении об управлении, и в ее рамках наиболее популярен способ обозначения связей посредством стрелок.



Грамматикой зависимостей в узком смысле называется теория синтаксической структуры предложения, в которой все [связи](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D0%B2%D1%8F%D0%B7%D1%8C) в предложении рассматриваются как [подчинительные](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%B4%D1%87%D0%B8%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5_(%D1%81%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%81)), вершиной предложения признаётся [сказуемое](https://ru.wikipedia.org/wiki/%D0%A1%D0%BA%D0%B0%D0%B7%D1%83%D0%B5%D0%BC%D0%BE%D0%B5) или его знаменательная часть, а [предлоги](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B5%D0%B4%D0%BB%D0%BE%D0%B3) описываются как [управляющие](https://ru.wikipedia.org/wiki/%D0%A3%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)) связанными с ними формами [существительных](https://ru.wikipedia.org/wiki/%D0%98%D0%BC%D1%8F_%D1%81%D1%83%D1%89%D0%B5%D1%81%D1%82%D0%B2%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5).

Достоинства деревьев подчинения:

* Отражают зависимостные отношения между словами.
* Более точны в описании семантики предложения.

Недостатки деревьев подчинения:

* Больше сложности в построении и понимании.
* Могут быть менее интуитивными для новичков.
* Труднее обрабатывать некоторые явления, такие как свобода порядка слов.

Грамматика составляющих используется преимущественно в описаниях языков с фиксированным порядком слов, грамматика зависимостей – в описаниях языков со свободным порядком слов.

#### 

#### 

#### **43. Системы составляющих.**

Система составляющих (или древовидная система составляющих) - это метод анализа синтаксической структуры предложения, основанный на разделении предложения на составляющие или компоненты, которые образуют его структуру. Она представляет предложение в виде древовидной структуры, где каждая вершина дерева представляет конструкцию (например, слово, группу слов или фразу), а ребра определяют синтаксические отношения между этими конструкциями.

Основные принципы системы составляющих:

* Разделение предложения на составляющие:
  + Предложение разбивается на более мелкие части, составляющие его структуру.
  + Эти составляющие могут быть словами, группами слов, фразами или частями речи.
  + Например, в предложении "Собака гонится за кошкой" составляющие могут быть: "Собака" (подлежащее), "гонится" (сказуемое), "за кошкой" (дополнение).

Иерархическая структура:

* + Составляющие имеют иерархическую организацию, где более крупные составляющие могут содержать более мелкие.
  + Например, фраза "очень большой зеленый дом" имеет составляющие: "очень большой" (определение), "зеленый" (определение), "дом" (существительное).
* Представление в виде древовидной структуры:
  + Каждая составляющая представляется как узел дерева.
  + Синтаксические отношения между составляющими выражаются в виде ребер дерева.

Преимущества системы составляющих:

* Наглядность: Предоставляет наглядное представление синтаксической структуры предложения в виде древовидной структуры.
* Удобство анализа: Легко интерпретируется и используется лингвистами для анализа синтаксической структуры.
* Используется в компьютерной лингвистике: Этот метод используется для разработки синтаксических анализаторов для обработки естественного языка.

Недостатки системы составляющих:

* Не всегда точно отражает смысл: Может не учитывать семантические связи и некоторые аспекты смысла.
* Сложность в построении: Требует определенных знаний и опыта для построения структуры составляющих.

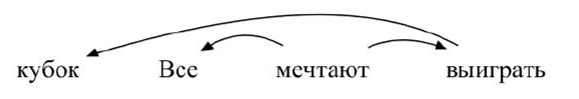
Системы составляющих используются в области компьютерной лингвистики для анализа и обработки языка, но они не всегда являются самым точным методом описания синтаксической структуры предложений из-за своих ограничений в описании семантики и более сложных языковых конструкций.

#### **44. Деревья подчинения.**

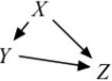
В методе деревьев подчинения предложение рассматривается как конечное множество (элемент множества — словоупотребление). Всякое дерево, для которого данное предложение служит множеством узлов, называется деревом подчинения для данного предложения.Любое слово может быть объявлено корнем дерева, а остальные слова ему непосредственно или опосредованно подчинены.

**Синтаксическая зависимость- антисимметрична**, т. е. если словоформа *X* — вершина словоформы *У,* то отсюда следует, что словоформа *У* не является вершиной словоформы *X.*

**Свойство проективность**. Дерево проективно, если:

* а) ни одна из стрелок не пересекает другую стрелку и
* б) никакая стрелка не накрывает вершину (корень)
* 

**Синтаксическая зависимость -антитранзитивна**. Это означает, что никакая словоформа *X* не может быть вершиной одновременно двух словоформ *У* и *X,* таких, что *У* — вершина *X,* следовательно структура

недопустима.

Достоинства деревьев подчинения:

* Отражают зависимостные отношения между словами.
* Более точны в описании семантики предложения.

Недостатки деревьев подчинения:

* Больше сложности в построении и понимании.
* Могут быть менее интуитивными для новичков.
* Труднее обрабатывать некоторые явления, такие как свобода порядка слов.
* **Структура дерева:**
  + Предложение представляется в виде дерева, где каждое слово является узлом, а связи между словами - рёбрами.
  + Рёбра обозначают синтаксические зависимости между словами, где одно слово является зависимым от другого.
* Типы связей:
  + Рёбра представляют типы зависимостей, такие как подлежащее, сказуемое, дополнение и т.д.
  + Каждое слово (узел) может иметь несколько зависимостей с другими словами.

**45. Автоматизация обработки текста: этап лексико-грамматического анализа.**

лексико-грамматический анализ (ЛГА) (его задачей является определение лексико-грамматического класса каждого слова входного текста с учетом

контекста и заранее заданного списка этих классов для ЕЯ; например, для английского языка их количество может достигать около 200 (это, например, JJ –

прилагательное; VB – глагол; MD – модальный глагол; NN – существительное

единственного числа; RB – наречие; ATI – определенный артикль; CC – союз

[34]), для белорусского и русского – более 1000);

задачей лексико-грамматического анализа текста является определение лексико-грамматической категории каждого его слова с учетом контекста. Множество всех лексико-грамматических категорий ЕЯ обычно задается заранее разработанным классификатором его лексико-грамматических свойств. Традиционно лексико-грамматическая классификация ЕЯ основана на разделении слов на части речи [46, 47]. При этом каждой из категорий дается некоторое условное обозначение (код, класс) в соответствии с выбранной системой кодирования. В результате ЛГА входному тексту ставится в соответствие в идеале единственная последовательность лексико грамматических кодов, по одному для каждого его слова, обычно называемая кодовой цепочкой.

Как правило, задача ЛГА решается в два этапа в предположении, что экспертами заранее построен так называемый эталонный словарь языка, в котором каждому его слову приписаны возможные для него вне контекста ЛГК в соответствии с заданным классификатором лексико грамматических свойств языка. На первом этапе поиском по эталонному словарю каждому слову входного текста ставятся в соответствие все возможные для него ЛГК, на втором – с учетом контекста снимается омонимия, т. е. для каждого слова определяется его единственно возможный ЛГК. Поскольку требуемый контекст не выходит за рамки одного предложения, то и задача ЛГА текста

обычно решается последовательно, начиная с его первого предложения и заканчивая последним.

Для решения задачи лексико-грамматического анализа текста также существует два основных подхода: - аналитический; - лингвистический. В основу первого, например, может быть положена однородная цепь Маркова первого порядка с конечным числом состояний [48], которая является математической моделью грамматики, порождающей ограниченный естественный язык [45]. Точность ЛГА в этом случае составляет порядка 95 % при размере тренировочного корпуса в 1 млн словоупотреблений [49]. Что касается скорости ЛГА, то задача может быть решена за время, близкое к линейному, что, безусловно, является большим достоинством рассматриваемого подхода, в рамках которого реализованы такие известные системы ЛГА, как PARTS Tagger 34 [50], Xerox Tagger [51] и TnT Tagger [52]. Однако реализация алгоритма и его настройка требуют достаточно трудоемкой работы с тренировочным корпусом. Лингвистический подход к ЛГА текста опирается на знания, выявленные экспертами в процессе качественного анализа естественного языка и предельно точно отражающие его истинные механизмы. Языковая компетенция в этом случае формализуется с помощью лингвистических правил, которые эксплицитно описывают конкретные языковые ситуации с указанием того, как необходимо проводить ЛГА текста в данной ситуации

#### 

#### **46. Модель СМЫСЛ – ТЕКСТ: определение, компоненты, задачи**.

**Модель СМЫСЛ – ТЕКСТ** - модель речевой деятельности человека, сопоставляющая любому смыслу все выражающие его тексты, а любым текстам — содержащиеся в нем смыслы. При этом смысл определяется как то общее, что имеется у всех различных высказываний, признаваемых и используемых носителями языка в качестве равнозначных (перифраз).

**Компоненты:**

Семантика:

1. Представление значения слов, фраз, предложений и текста в контексте их истинного смысла и значимости.
2. Анализ семантических связей между словами и фразами.

Структура текста:

1. Анализ структуры текста на уровне предложений, абзацев, разделов и документов.
2. Понимание связей между частями текста для выявления их смысла.

**Задачи:**

Семантический анализ:

1. Понимание значений слов и фраз в контексте их использования в тексте.
2. Анализ семантических отношений между словами и выявление их значения в предложениях.

Понимание текста:

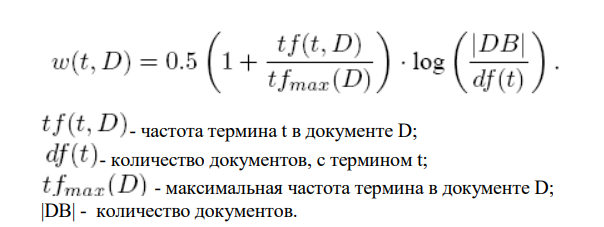
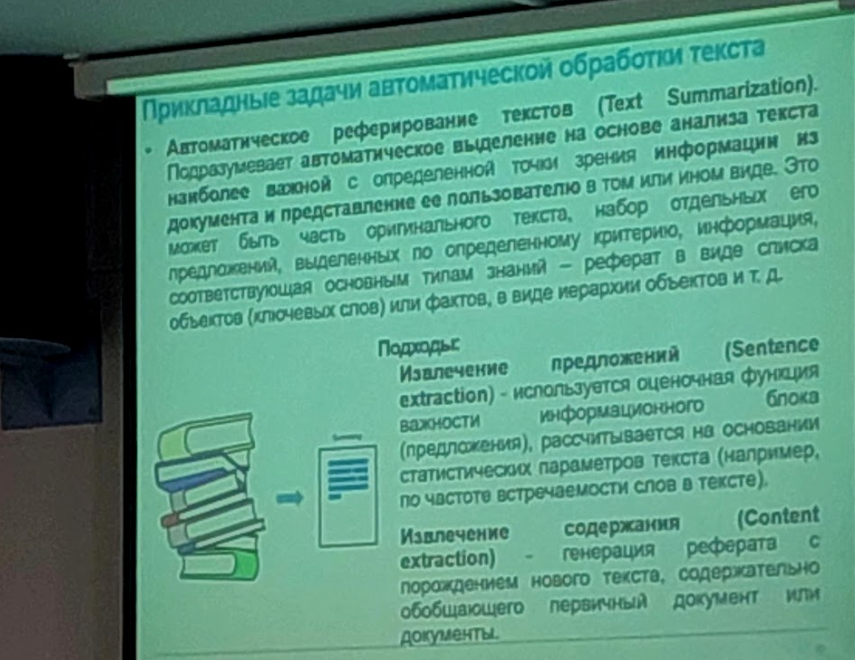
1. Связывание смысла текста с его структурой и содержанием.
2. Понимание общего контекста и целей текста.

Использование в обработке естественного языка:

1. Поддержка систем машинного перевода, вопросно-ответных систем и других приложений, требующих понимания текста.

#### 

#### **47. Задача автоматического реферирования документов. Sentence extraction - особенности подхода.**



В подходе к автоматическому реферированию, основанному на извлечении предложений (Sentence Extraction), основной идеей является выбор и извлечение наиболее важных предложений из исходного текста для создания краткого реферата.

Основные особенности подхода **Sentence extraction:**

* **Полное понимание текста:**
* Извлечение информативных предложений:
  + Алгоритмы выбирают предложения, которые наиболее точно передают ключевую информацию из оригинального текста.
  + Основной акцент делается на информативности предложений и их способности передать основную суть текста.
* Оценка важности предложений:
  + Используются различные методы для определения важности предложений, такие как TF-IDF, машинное обучение (например, нейронные сети), частота встречаемости ключевых слов и другие статистические или семантические подходы.
* Критерии выбора предложений:
  + Предложения оцениваются по различным критериям, включая наличие ключевых слов, их взвешенное вхождение, структурные особенности (например, заголовки, подзаголовки), семантическую значимость и др.
* Создание краткого обзора:
  + Выбранные предложения объединяются в краткое содержание или реферат, передающий основную информацию и ключевые аспекты оригинального текста.

Преимущества подхода "Sentence extraction":

* Эффективность: Отбираются наиболее информативные части текста.
* Простота реализации: Относительно легко внедряемый подход.
* Сохранение структуры текста: Часто сохраняется структура оригинального текста.

Недостатки подхода "Sentence extraction":

* Потеря контекста: Извлеченные предложения могут потерять контекст и связность информации.
* Субъективность выбора предложений: Выбор наиболее важных предложений может зависеть от алгоритмов и критериев оценки важности.
* Не всегда генерирует качественный реферат: Могут быть упущены важные детали или контекстуальные аспекты текста.

"Sentence extraction" - это популярный метод автоматического реферирования, который находит применение в системах автоматической обработки текста для создания кратких обзоров или рефератов документов.

#### 

#### **48. Задача автоматического реферирования документов. Summary generation - особенности подхода.**

Автоматическое реферирование документов, также известное как генерация краткого содержания (Summary Generation), представляет собой задачу создания краткого и информативного обзора текста, содержащего основные идеи и ключевую информацию. Этот процесс включает в себя ряд особенностей и подходов:

* Использование методов машинного обучения:
  + Генерация резюме часто основана на методах машинного обучения, таких как нейронные сети, рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN) и трансформеры.
* Извлечение ключевой информации:
  + Алгоритмы реферирования фокусируются на выделении ключевых фраз, предложений или абзацев, содержащих существенную информацию.
* Оценка важности текстовых элементов:
  + Модели оценивают важность каждого элемента текста с целью выделения наиболее значимой информации.
* Абстракция и избегание дублирования:
  + Системы генерации резюме стремятся создать краткое содержание, избегая повторений и лишних деталей.
* Генерация сжатых и легко читаемых текстов:
  + Резюме создаются с учетом краткости и читаемости, чтобы предоставить пользователю информацию в более доступной форме.
* Использование предобученных моделей:
  + Многие современные системы реферирования используют предобученные языковые модели, такие как GPT (Generative Pre-trained Transformer) для более точного и разнообразного создания резюме.
* Обработка различных типов данных:
  + Системы реферирования могут работать с различными типами данных, включая текст, изображения и звук, что позволяет создавать краткое содержание для разнообразных источников.
* Оптимизация под конкретные задачи:
  + Многие системы реферирования разрабатываются и оптимизируются для конкретных видов текстов (научные статьи, новостные статьи, техническая документация) с учетом особенностей каждого жанра.

Генерация резюме – это сложная задача, требующая сбалансированного подхода между извлечением ключевой информации и созданием читаемого и информативного текста. Она продолжает развиваться благодаря прогрессу в области искусственного интеллекта и глубокого обучения.

#### **49. Автоматизация обработки текста: этапы преформатирования текста, распознавание границ слов и предложений.**

**Предобработка текста** переводит текст на естественном языке в формат, более подходящий для машины для дальнейшей работы. Предобработка состоит из различных этапов, которые могут отличаться в зависимости от задачи и реализации. Далее приведен один из возможных наборов этапов:

* Перевод всех букв в тексте в нижний или верхний регистры;
* Удаление цифр (чисел) или замена на текстовый эквивалент (обычно используются регулярные выражения);
* Удаление пунктуации.
* Обычно реализуется как удаление из текста символов из заранее заданного набора;
* Удаление пробельных символов (whitespaces);
* Токенизация (обычно реализуется на основе регулярных выражений);
* Удаление стоп слов;
* Стемминг;
* Лемматизация;
* Векторизация.

Существует несколько способов определения **границ слов**:

* Правила на основе символов: Этот метод использует знаки препинания и пробелы для разделения текста на слова.
* Лексические анализаторы: Эти инструменты используют более сложные правила, которые могут включать синтаксический и грамматический анализ. Они более эффективны для языков с сложной морфологией, таких как русский, где одно слово может иметь множество форм.
* Статистические методы: Эти методы используют машинное обучение для определения границ слов. Они обучаются на больших текстовых корпусах для выявления вероятных границ слов, что особенно полезно в языках, где границы слов не всегда очевидны, как, например, в китайском.

Существует несколько способов определения **границ предложений**. Один из наиболее простых и широко распространенных — основанный на использовании пунктуации, такой как точка, вопросительный и восклицательный знаки. Однако этот подход не всегда даёт точные результаты, так как в тексте могут встречаться аббревиатура с точкой (например, «Mr.»), сокращения (например, «г. Петербург»), и другие случаи, когда точка не является признаком конца предложения. Для повышения точности определения границ предложений можно использовать машинное обучение. Алгоритмы машинного обучения, такие как рекуррентные нейронные сети или скрытые марковские модели, могут обучаться на больших корпусах текста и выявлять закономерности, которые помогут правильно разделять предложения. Это позволяет справляться с сложными случаями, когда определение границ предложений на пунктуации достаточно затруднительно.

#### **50. Задача информационного поиска. Статистические закономерности в текстах естественного языка. Законы Дж. Зипфа.**

Центральная задача ИП – помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

Классическая задача ИП, с которой началось развитие этой области, – это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов. Но список задач ИП постоянно расширяется и теперь включает: вопросы моделирования, классификацию и фильтрацию документов, проектирование архитектур поисковых систем и пользовательских интерфейсов, извлечение информации, построение языков запросов и др.

Также, перед движками ИП ставятся некоторые задачи по обработке естественных языков, что включает в себя морфологический анализ, разрешение лексической многозначности и так далее.

Методы информационного поиска

*Полнотекстовый поиск* – автоматизированный документальный поиск, при котором в качестве поискового образа документа используется его полный текст или существенные части текста. Пример системы полнотекстового поиска – любой интернет-поисковик, например Яндекс или Google.

*Семантический поиск* – процесс поиска документов по их содержанию. Для реализации механизма семантического поиска необходимо осуществить перевод содержания документов и запросов с естественного языка на информационно-поисковый язык и составление поисковых образов документа и запроса. При семантическом поиске находится множество документов без указания адресов

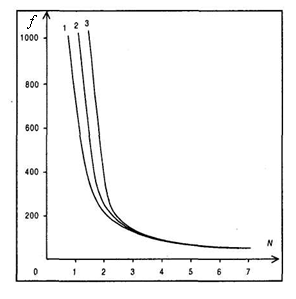
*Документальный поиск* – процесс поиска в хранилище информационно-поисковой системы первичных документов или в базе данных вторичных документов, соответствующих запросу пользователя. Существуют два вида документального поиска: библиотечный, направленный на нахождение первичных документов, и библиографический, направленный на нахождение сведений о документах, представленных в виде библиографических записей.

Первый закон Зипфа «Ранг ‑ частота»

*Частотой встречаемости слова* называется величина, равная числу вхождений слова в текст. Вероятность обнаружения некоторого слова в тексте *p* равна отношению частоты его вхождения к общему числу слов в тексте. Если все слова одного разговорного языка или просто достаточно длинного текста упорядочить по убыванию частоты их вхождения и пронумеровать, то значение частоты вхождения каждого слова окажется обратно пропорциональным его порядковому номеру. Порядковый номер слова в таком списке является рангом слова и обозначается *R*. Если несколько разных слов имеют одинаковые частоты, то учитывается только одно из них. Выявленная закономерность описывается первым законом Зипфа: *Произведение частоты встречаемости слова на его ранг приблизительно постоянно для любых текстов определенного языка. f R = C,* где *f* ‑ частота встречаемости слова, C — константа Зипфа.

Второй закон Зипфа «количество ‑ частота»

В первом законе не учитывался такой факт, что разные слова могут входить в текст с одинаковой частотой. Зипф установил, что частота и количество разных слов *N*, входящих в текст с данной частотой, также связаны между собой определенной зависимостью. Если построить график, отложив по оси ординат частоту вхождения слова, а по оси абсцисс — количество разных слов, характеризуемых одинаковой частотой, то получившаяся кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов в пределах одного языка. Однако и межъязыковые различия невелики. На каком бы языке текст ни был написан, форма данной кривой Зипфа останется неизменной. Могут незначительно отличаться лишь коэффициенты, отвечающие за наклон кривой. Данное свойство иллюстрируется данными рис. 2., на котором показаны кривые для французского (кривая 1), английского (кривая 2) и русского (кривая 3) языков.



#### **51. Задача информационного поиска. Хранение индексных структур: прямой и инвертированный индексы.**

Задача информационного поиска состоит в том, чтобы эффективно находить информацию в больших объемах данных, когда пользователь запрашивает определенные ресурсы или документы. Хранение данных в индексных структурах - ключевой аспект для быстрого и точного поиска информации.

1. Прямой индекс:

Прямой индекс представляет собой структуру данных, в которой каждому документу назначается уникальный идентификатор, а затем создается индекс, связывающий этот идентификатор с набором терминов или слов, встречающихся в этом документе. Пример:

Документ 1:

- Идентификатор: 1

- Слова: "кот", "собака", "дом"

Документ 2:

- Идентификатор: 2

- Слова: "собака", "птица", "лес"

Такой индекс позволяет быстро найти документы, содержащие определенные слова. Однако, он не обеспечивает быстрый поиск по словам, так как не подразумевает прямого доступа к документам, содержащим конкретное слово.

1. Инвертированный индекс:

Инвертированный индекс строится на основе прямого индекса. Он представляет собой структуру данных, в которой каждому уникальному слову (термину) в коллекции документов соответствует список идентификаторов документов, содержащих это слово. Пример:

Слово "кот":

- Документы: 1

Слово "собака":

- Документы: 1, 2

Слово "дом":

- Документы: 1

Такая структура позволяет быстро находить документы, содержащие конкретные слова. При поиске по слову происходит обращение к инвертированному индексу, что позволяет найти документы, в которых это слово встречается, и затем извлечь эти документы из хранилища.

Инвертированный индекс эффективен для быстрого поиска по словам, так как позволяет избежать просмотра всех документов в коллекции при запросе. Он широко используется в поисковых системах для эффективного поиска и ранжирования документов по запросам пользователей.

#### 

#### **52. Задача информационного поиска. Индексирование текстовых документов.**

**Задача информационного поиска:**

Задача информационного поиска заключается в нахождении и предоставлении пользователю информации, наилучшим образом соответствующей его запросу. Это включает в себя обработку, анализ и организацию больших объемов текстовых данных для эффективного поиска.

**Индексирование текстовых документов:**

Индексирование - это процесс создания индекса, который упрощает поиск и ускоряет доступ к информации. В контексте информационного поиска, индексирование текстовых документов включает в себя следующие шаги:

* **Предобработка текста:**
  + Текстовые документы проходят через этап предобработки, включающий удаление стоп-слов (часто встречающихся, но малоинформативных слов), лемматизацию (приведение слов к их базовой форме), удаление стоп-слов и знаков препинания.
* **Токенизация:**
  + Разделение текста на отдельные слова (токены). Это помогает в создании словаря терминов, который будет использоваться для построения индекса.
* **Создание обратного индекса:**
  + Обратный индекс - это структура данных, связывающая каждое слово с документами, в которых оно встречается. Для каждого термина в словаре создается список документов, содержащих этот термин.
* **Взвешивание терминов:**
  + Присвоение каждому термину веса, который отражает важность термина в конкретном документе или коллекции документов. Это может включать в себя использование весов, основанных на TF-IDF (term frequency-inverse document frequency).
* **Хранение индекса:**
  + Индекс может быть хранен в памяти, на диске или в базе данных в зависимости от объема данных и требований к производительности.
* **Обновление индекса:**
  + При поступлении новых документов или изменении существующих необходимо обновлять индекс, чтобы отражать текущее состояние коллекции текстовых данных.
* **Поиск и ранжирование:**
  + Поиск выполняется с использованием индекса. По запросу пользователя система находит соответствующие документы, а затем ранжирует их по степени релевантности запросу с использованием алгоритмов ранжирования.

Индексирование текстовых документов играет ключевую роль в эффективной работе информационных поисковых систем, обеспечивая быстрый и точный доступ к требуемой информации.

#### **53. Задача информационного поиска. Способы оценки эффективности систем информационного поиска.**

Центральная задача ИП – помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

Классическая задача ИП, с которой началось развитие этой области, – это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов. Но список задач ИП постоянно расширяется и теперь включает: вопросы моделирования, классификацию и фильтрацию документов, проектирование архитектур поисковых систем и пользовательских интерфейсов, извлечение информации, построение языков запросов и др.

Оценка эффективности систем информационного поиска может проводиться различными способами, например:

1. *Поисковые запросы и релевантность результатов*: Оценка проводится путем анализа поисковых запросов, представленных пользователями, и степени соответствия результатов их ожиданиям.
2. *Анализ использования*: Изучение того, как часто и каким образом пользователи взаимодействуют с системой поиска, например, какие запросы они выполняют, сколько времени проводят на поиске, какие результаты они выбирают и т.д.
3. *Тестирование пользователей:* Проведение тестов и опросов среди пользователей для оценки их удовлетворенности результатами поиска и простотой использования системы.
4. *Метрики качества поиска:* Использование различных метрик, таких как точность, полнота, F-мера, доля кликов и др., для количественной оценки производительности системы поиска.
5. *Сравнительные исследования:* Проведение сравнительных анализов эффективности различных систем информационного поиска по отношению к друг другу или к стандартным наборам данных.

Эти методы позволяют комплексно оценить производительность систем информационного поиска с точки зрения их релевантности, удобства использования и других аспектов, что помогает оптимизировать их работу в соответствии с потребностями пользователей.

#### 

#### **54. Речевой интерфейс: определение, задачи.**

**Речевой интерфейс** — это форма взаимодействия между человеком и компьютерной системой, основанная на использовании звуковой речи. В речевом интерфейсе пользователь взаимодействует с системой, произнося слова и фразы, вместо использования клавиш, мыши или других устройств ввода.

**Задачи речевого интерфейса**:

* **Распознавание речи** (ASR - Automatic Speech Recognition): Одной из основных задач речевого интерфейса является преобразование звуков речи пользователя в текст. Это обеспечивает компьютеру понимание того, что было сказано.
* **Обработка естественного языка** (NLP - Natural Language Processing): После распознавания речи система должна анализировать и понимать смысл сказанного, учитывая синтаксическую и семантическую структуру фразы.
* **Генерация речи** (TTS - Text-to-Speech): На этом этапе текст, сгенерированный системой или полученный от пользователя, преобразуется в звуковые волны, чтобы система могла ответить пользователю устно.
* **Интерпретация команд и выполнение задач:** Речевой интерфейс должен быть способен интерпретировать команды, задачи и запросы пользователя, а затем выполнять соответствующие действия.
* **Обеспечение навигации и контроля:** Пользователь должен иметь возможность управлять приложениями, устройствами или системами при помощи голосовых команд.
* **Адаптация к интонации и эмоциям**: Речевой интерфейс должен учитывать интонации и эмоции в голосе пользователя для более точного понимания его намерений.
* **Обучение и персонализация:** Речевые интерфейсы могут обучаться и адаптироваться к индивидуальным особенностям пользователя, улучшая качество взаимодействия.
* **Обеспечение безопасности и конфиденциальности:** Работа с речевыми данными требует внимания к вопросам безопасности и конфиденциальности информации.

#### 

#### **55. Поисковая оптимизация. Назначение и методы реализации.**

Поисковая оптимизация (SEO - search engine optimization) - это процесс улучшения видимости и ранжирования веб-сайта в естественных (органических) результатах поисковых систем. Основная цель SEO - привлечение большего количества органического трафика на веб-сайт путем улучшения его позиций в результатах поисковых запросов, которые пользователи вводят в поисковые системы.

Методы реализации SEO включают следующие элементы

1. Ключевые слова: Исследование и оптимизация ключевых слов является одним из важных аспектов SEO. Это включает в себя выбор релевантных ключевых слов или фраз, которые наиболее вероятно будут использоваться пользователями при поиске информации. Включение этих ключевых слов в контенте веб-сайта, метатегах, заголовках страниц и других элементах структуры страницы помогает поисковым системам определить тематику и релевантность сайта.

2. Оптимизация контента: Контент на веб-сайте должен быть информативным, уникальным и понятным для пользователей. Хороший контент, обогащенный ключевыми словами и плотно связанный с темой, будет более вероятно ранжироваться выше в поисковых результатах. Оптимизация контента также включает в себя использование заголовков, подзаголовков, списков, ссылок и других элементов для улучшения его структуры и читабельности.

3. Техническая оптимизация: Важным аспектом SEO является техническая оптимизация веб-сайта. Это включает в себя обеспечение быстрой загрузки страниц, удобной навигации, правильного использования метатегов, организацию URL-адресов, устранение технических ошибок и других факторов, которые могут влиять на индексацию и ранжирование веб-сайта.

4. Ссылочная стратегия: Качество и количество входящих ссылок (backlinks) также оказывают значительное влияние на ранжирование веб-сайта. Важно получать качественные ссылки с релевантных и авторитетных веб-сайтов. Это можно достичь через создание интересного контента, активное участие в социальных сетях и блогах, а также через партнерские и гостевые публикации на других сайтах.

5. Мобильная оптимизация: В наше время все больше пользователей используют мобильные устройства для поиска информации. Поэтому важно, чтобы веб-сайт был оптимизирован для мобильных устройств. Это включает в себя использование адаптивного дизайна, оптимизацию скорости загрузки на мобильных устройствах и удобную навигацию на сенсорных экранах.

\* Ранжирование — это сортировка сайтов поисковыми системами и размещение их в результатах выдачи.

#### 

#### 

#### **56. Обобщенная структура системы автоматического синтеза речи.**

Обобщенная структура системы автоматического синтеза речи включает несколько ключевых компонентов:

1. **Обработка текста**

- Токенизация текста для анализа.

- Лингвистический анализ. Анализ грамматической и семантической структуры текста.

2. **Преобразование текста в речь**

- Преобразование слов в последовательности фонем (звуков).

3. **Генерация речи**

- Определение интонации и ритма.

4. **Выбор голоса**

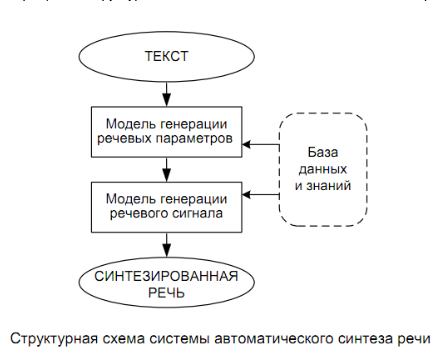
- Выбор подходящей голосовой модели, которая определяет характеристики голоса.

5. **Синтез речи**

- Генерация звукового сигнала, который звучит как естественная речь.

6. **Оценка качества**

- Анализ полученной речи с использованием метрик для определения степени соответствия натуральной речи.



Под системами автоматического синтеза речи (иначе их еще называют синтезаторами речи) понимают системы, преобразующие орфографический текст и другую информацию в звучащую речь. Общепринятое в английской литературе обозначение – TTS (Text To Speech) System – системы преобразования текста в речь.

Под моделью генерации речевых параметров понимается блок преобразования входного орфографического текста в последовательность параметров, с помощью которых можно описать речь. Это могут быть артикуляторные параметры, либо параметры, связанные с акустикой речи, либо другие параметры, набор которых определяется, исходя из того, какая информация заключена в речевом потоке и каким образом она описана.

Модель генерации речевого сигнала – это блок преобразования речевых параметров в речевой сигнал, который воспринимает пользователь системы. Данный блок сопряжен с динамиками и в некоторых реализациях синтезаторов речи представляет собой только соответствующую аппаратную часть речевого интерфейса, а в некоторых – аппаратно-программную.

Фактически система автоматического синтеза речи – это совокупность двух компонент, которые часто называют синтезатором речевых параметров и синтезатором речевого сигнала. Оба этих компонента реализуются не только как набор программ, но и используют некую базу данных и знаний, содержащую информацию об особенностях организации естественного языка и о закономерностях, которые следует учитывать при синтезе речи. Кроме того, синтезатор речевого сигнала имеет аппаратно-программную реализацию, так как для того чтобы мы услышали звук, необходимо наличие, как минимум, звуковой платы и динамиков, подключенных к компьютеру. Таким образом, на выходе мы получаем звучащую синтезированную речь.

GPT

Система автоматического синтеза речи (Text-to-Speech, TTS) преобразует текстовую информацию в аудиосигнал, имитирующий естественную человеческую речь. Такая система состоит из нескольких ключевых модулей, каждый из которых выполняет определённые задачи для достижения естественного звучания речи.

Основные этапы и структура системы TTS

1. Ввод текста

* Принимается текст, который нужно преобразовать в речь.
* Источник текста: ввод пользователем, системы диалогов, базы данных.

2. Лингвистическая обработка текста

Этот этап обеспечивает разбиение текста на осмысленные части и подготовку данных для синтеза.

3. Фонетическая обработка

На этом этапе текст преобразуется в последовательность фонем — минимальных звуковых единиц языка.

* **Фонетическая транскрипция**:
  + Преобразование текста в фонемы.
  + Пример: слово "дом" → /d/, /o/, /m/.
* **Построение просодической модели**:
  + Просодия включает темп речи, длительность звуков, паузы, тон и громкость.
  + Определение интонации, связанной с эмоциональной окраской или типом предложения.

4. Аудио-синтез

Основной этап, на котором звуковой сигнал создаётся на основе фонетической и просодической информации.

* **Методы синтеза речи**:
  1. **Конкатенативный синтез**:
     + Используются записанные фрагменты речи (дифоны, трифоны), которые соединяются в единый звуковой сигнал.
     + Преимущество: естественное звучание.
     + Недостаток: ограниченная гибкость.
  2. **Синтез на основе параметров**:
     + Речь генерируется с использованием моделей параметрического звука.
     + Преимущество: компактность системы.
     + Недостаток: менее естественное звучание.
  3. **Нейронный синтез (WaveNet, Tacotron)**:
     + Используются нейронные сети для прямого генерации звукового сигнала.
     + Преимущество: высокая естественность, плавность речи.
  4. **Формантный синтез**:
     + Создание звукового сигнала путём моделирования характеристик речевого аппарата.
     + Преимущество: высокая гибкость.
     + Недостаток: звучание менее естественное.

5. Постобработка и выдача сигнала

* Финальная обработка звукового сигнала для устранения шумов и улучшения качества.
* Конвертация аудио в требуемый формат (например, .wav, .mp3).
* Выдача результата: воспроизведение через динамик или сохранение файла.

#### **57. Постановка задачи автоматического синтеза речи.**

Постановка задачи **автоматического синтеза речи** (TTS, Text-to-Speech) включает в себя определенные шаги и критерии, целью которых является создание эффективной системы генерации естественной и выразительной речи из текстовых данных.

Вот более подробное описание **постановки задачи**:

* **Определение входных данных**: Определение формата входных данных, которые могут включать в себя текстовые фразы, предложения, абзацы или даже длинные текстовые документы.
* **Лингвистический анализ**: Разбор входного текста с целью выделения лингвистических особенностей, таких как слова, фразы, грамматическая структура и другие лингвистические аспекты.
* **Формирование текста**: Генерация промежуточного текста или лингвистической репрезентации текста, которая будет использоваться для последующего синтеза речи.
* **Выбор речевых параметров**: Определение параметров речи, таких как интонация, темп, длительность звуков, высота тона и другие просодические характеристики. Эти параметры зависят от желаемого стиля и эмоционального оттенка синтезируемой речи.
* **Фонетический анализ**: Преобразование лингвистического анализа в последовательность фонем (звуков), которые должны быть синтезированы.
* **Выбор метода синтеза речи:** Выбор между различными методами синтеза речи, такими как конкатенативный синтез (на основе записанных звуков), синтез с использованием генеративных моделей, артикуляторный синтез и др.
* **Синтез речи**: Процесс генерации аудиосигнала, соответствующего синтезируемой речи на основе выбранных параметров и модели.
* **Оценка качества**: Проведение оценки качества синтезированной речи с использованием объективных и/или субъективных мер. Это может включать в себя анализ интонации, понимаемость речи, природность звучания и другие аспекты.
* **Оптимизация и обучение:** При необходимости, процесс обучения моделей на больших объемах данных для улучшения качества синтеза и адаптации к различным стилям и языкам.

gpt

**Задача автоматического синтеза речи (Text-to-Speech, TTS)** заключается в преобразовании текста в естественно звучащую речь, сохраняющую смысл и эмоциональную окраску оригинального текста. Система синтеза должна быть способна адекватно передавать звуковую форму текста для использования в разнообразных приложениях, таких как голосовые помощники, системы озвучивания, устройства для людей с нарушением зрения и др.

Подзадачи автоматического синтеза речи

1. **Обработка текста (Text Processing)**:
   * Нормализация текста: преобразование чисел, аббревиатур, символов.
   * Лингвистический анализ: синтаксический и морфологический анализ.
   * Расстановка ударений и просодических акцентов.
2. **Фонетическая обработка (Phonetic Processing)**:
   * Преобразование текста в последовательность фонем.
   * Построение просодической модели: темп, интонация, паузы.
3. **Синтез аудио (Audio Synthesis)**:
   * Генерация звукового сигнала на основе фонемной последовательности.
   * Применение интонационных и просодических параметров.
4. **Калибровка параметров**:
   * Настройка параметров громкости, скорости, выразительности.
5. **Оценка качества (Evaluation)**:
   * Тестирование синтезированной речи на понятность, естественность и точность передачи смысла.

#### 

#### **58. Поисковая оптимизация. Факторы, влияющие на ранжирование веб-ресурсов.**

Поисковая оптимизация - это работа над внутренними и внешними факторами, влияющими на ранжирование сайта в поисковых системах.

Категории и примеры факторов ранжирования:

1. Технические факторы - Демонстрируют качество работы сайта:
   1. Скорость загрузки каждой страницы (оптимально 1-3 секунды)
   2. Защищенное соединение (Использование протокола https вместо http обеспечивает безопасность обмена данными.)
   3. Время ответа сервера (в идеале до 0.2 секунды)
   4. Оптимизация под мобильное устройство
   5. Отсутствие дублей контента и тд
2. Хостовые и доменные - параметры, связанные с регистрацией и историей существования сайта
   1. Возраст домена
   2. История домена
   3. Доменное имя
3. Текстовые факторы - Помогают оценить качество текстов
   1. уникальность контента
   2. семантическое ядро - При оценке контента поисковики выявляют ключевые слова, которые соответствуют запросу пользователя
   3. Скрытая семантика (LSI) - Дополнительно к ключевым словам поисковые системы оценивают контекст употребления ключевых фраз и их семантическое окружение. То есть текст должен быть связан по смыслу с употребленными ключевыми словами
4. Ссылочные факторы - ссылочные факторы охватывают использование внешних и внутренних ссылок
   1. Число исходящих ссылок
   2. Количество входящих ссылок
   3. Авторитетность ссылающихся доменов и тд
5. Поведенческие факторы - позволяют проанализировать поведение пользователей на сайте, например: время пребывания на сайте, кликабельность сниппета и тд
6. Социальные факторы - связь веб-ресурса с социальными сетями
7. Коммерческие факторы - объединяют критерии, связанные с интернет-продажей товаров и услуг, например: инфа о продавце, наличие каталога, способы оплаты и тд.

#### 

#### 

#### **59. Обобщенная структура системы автоматического распознавания речи.**

Под системами автоматического распознавания речи (САРР) понимают системы, преобразующие входную речь (речевой сигнал) в распознанное сообщение. При этом распознанное сообщение может быть представлено как в форме текста этого сообщения, так и преобразовано сразу в форму, удобную для его дальнейшей обработки с целью формирования ответной реакции системы. Изначально перед системой автоматического распознавания речи ставится задача преобразования текста в речь.

Упрощенная структурная схема системы автоматического распознавания речи:



Под моделью анализа речевого сигнала понимают блок, в задачи которого входит анализ входного сигнала, во-первых, с целью отнесения его к числу речевых, а во-вторых, для выделения в составе полученного сигнала компонент, которые являются основными для распознавания полученного сообщения.

Модель распознавания речи и принятия решения – это блок, в рамках которого осуществляется формирование распознанного сообщения на основе анализа последовательности параметров, полученных из первого блока. Например, если используется формантная модель описания речи, то на основе полученных в первом блоке частот формант строится последовательность распознанных фонем, составляющих входное сообщение. При этом осуществляется принятие решения о том, распознано ли входное сообщение правильно. При принятии решения, в частности, возможны следующие решения: сообщение распознано правильно (подтверждением этого является текст, соответствующий нормам естественного языка) либо сообщение не распознано или распознано не правильно (такое решение принимается в случае наличия в распознанном сообщении явных, трудно исправимых автоматически ошибок или вообще полной бессмыслицы).

#### **60. Постановка задачи автоматического распознавания речи.**

К основным классам задач речевого интерфейса следует отнести:

1. синтез речи – эта задача включает в себя комплекс подзадач и заключается в обеспечении возможности произнесение речи компьютером на основе произвольного орфографического текста;
2. анализ и распознавание речи – комплекс задач, включающих запись, оцифровку и анализ речи для распознавания полученного речевого сообщения компьютерной системой;
3. понимание (интерпретация) речи – это комплекс задач, связанных с анализом смысла речевых сообщений и формированием реакции (ответа) компьютерной системы. Часто указанная задача является подзадачей задачи распознавания речи;
4. распознавание голоса – комплекс задач, включающих анализ особенностей голоса говорящего с целью выявления каких-либо его индивидуальных (личностных) особенностей и качеств. Данный комплекс задач называют также верификацией и идентификацией речи;
5. компьютерное клонирование голоса и дикции – это создание близкой копии, но не биологической, а компьютерной, и не всего существа в целом (в данном случае человека), а только одной из его интеллектуальных функций: чтение произвольного орфографического текста. При этом ставится задача максимально полного сохранения персональных акустических особенностей голоса, фонетических особенностей произношения и акцента, а также просодической (интонационной) индивидуальности речи (мелодика, ритмика, динамика)

Помимо перечисленных выше задач, входящих в группу задач разработки собственно речевого интерфейса, следует также отметить, что имеется ряд вспомогательных задач, решением которых занимаются научно-технические коллективы, разрабатывающие речевые системы. Это связано с тем, что задача реализации речевого интерфейса до сих пор не решена окончательно. К таким задачам, в частности, относятся следующие:

* исследование особенностей фонетического строения речи различных естественных языков;
* исследование особенностей интонационной окраски речи различных языков;
* выявление наборов параметров для описания речи, используемых как для синтеза речи, так и для ее распознавания;
* разработка новых методов синтеза речи;
* исследование различий речи разных дикторов и, в частности, мужского и женского голоса;
* разработка новых методов распознавания речи;
* поиск оптимальных путей передачи речи по каналам связи;
* разработка специальных шумоподавляющих микрофонов;
* разработка специальной аппаратуры для исследования характеристик речи;
* разработка новых методов оцифровки и оптимального сжатия речевого сигнала;
* разработка специальных звуковых карт, ориентированных на синтез и анализ речи;
* формирование баз данных с «образцами» речи различных дикторов с целью повышения естественности звучания
* синтезированной речи;
* исследование строения речевого тракта человека и особенностей образования звуков речи;
* исследование строения органов слуха человека;
* исследование особенностей восприятия речи человеком;
* поиск путей оптимального использования речевого интерфейса в различных технических и бытовых системах и разработка соответствующих технологий и др.

**61. Поисковая оптимизация. Построение семантического ядра. Инструменты поисковой оптимизации.**   
Построение семантического ядра в контексте поисковой оптимизации относится к процессу определения ключевых тем и слов, связанных с определенной тематикой сайта или контента. Семантическое ядро имеет центральное ключевое слово, как правило высокочастотное, и все остальные ключевые слова в нём ранжируются по мере убывания частоты совместного использования с центральным запросом в общей коллекции документов.

Ключевые слова семантического ядра подбираются путём анализа услуг или товаров компании, анализа статистики запросов, статистики сайта, содержимого конкурирующих сайтов и сезонности употребления поисковых запросов. Состав семантического ядра должен максимально соответствовать представлениям целевых посетителей сайта о той информации которая на нём, по их мнению, должна присутствовать.

Для заимствования основных ключевых слов используются в том числе следующие методы:

* Сервисы «Google Analytics», «Яндекс. Вордстат» и подобные им, предоставляющие информацию о частотности употребления ключевых слов в поисковых запросах;
* анализ статистики использования основных ключевых слов, по которым осуществляется переход из результатов поиска на сайт, для которого составляется семантическое ядро (информация доступна в «Google Analytics», «Яндекс.Метрика» и им подобных);
* мониторинг видимости конкурирующих сайтов в результатах поиска по предположительно интересным запросам.

#### 

#### **62. Задача автоматического распознавания языка текста. Метод частотных слов, достоинства и недостатки.**

Распознавание языка текста — это процесс определения языка, на котором написан определенный текст. Для автоматического распознавания языка текста можно использовать различные методы и технологии.

**Метод частотных слов** (или метод частотного анализа) является одним из подходов к автоматическому распознаванию языка, и он основан на анализе частоты встречаемости слов в тексте.

Согласно методу частотных слов ПОЯ представляет собой набор слов, обладающих наибольшей частотой встречаемости в сравнении с остальными словами документов из обучающего корпуса. Вероятность принадлежности входного документа языку определяется на основании пословного сравнения с ПОЯ с учетом рассчитанных вероятностей его элементов.

Достоинства:

* **Простота и эффективность:**
  + Метод частотных слов прост в реализации и может быть эффективным для базового определения языка, особенно если тексты на разных языках имеют уникальные слова или фразы.
* **Низкие требования к вычислительным ресурсам:**
  + Для реализации этого метода не требуется больших вычислительных ресурсов. Это позволяет использовать его в условиях ограниченной вычислительной мощности.
* **Относительно быстрое выполнение:**
  + Анализ частот слов обычно происходит быстро, что делает метод подходящим для реального времени или больших объемов текстов.

Недостатки:

* **Зависимость от объема текста:**
  + Точность метода может страдать при анализе коротких текстов, так как частоты слов могут быть не репрезентативными.
* **Неэффективность в случае пересекающихся языков:**
  + В случае, если два языка сильно пересекаются по используемым словам, метод может давать ошибочные результаты.

#### 

#### **63. Задача автоматического распознавания языка текста. Нейросетевой метод, достоинства и недостатки.**

Автоматическая обработка текстов (АОТ) предполагает решение многих задач,

которые условно можно разбить на два уровня. Задачи высокого уровня представле-

ны задачами распознавания речи, реферирования текстов, распознавание языка текста, генерации документов, машинного перевода, извлечения информации, обучения языку, т. е. приложениями.

К задачам низшего уровня относят грамматический разбор, снятие смысловой много-значности, корректировку орфографии и синтаксический разбор, т. е. задачи собственно лингвистической обработки ЕЯ. К настоящему времени этот круг задач

значительно расширился и в целом охватывает всю индустрию развития и поддержки компьютерной формы существования ЕЯ.

Главная проблема при решении указанных задач состоит в необходимости об-

рабатывать неструктурированные тексты. Единый типовой алгоритм их автоматиче-

ской обработки создать не удается, поскольку конкретный вид алгоритма, в первую

очередь, определяется строем языка.

Для этого были испробованы различные методы и алгоритмы, но подробнее остановимся на нейросетевом.

**Нейросетевой метод в автоматическом распознавании языка:**

Нейросетевые методы, особенно с использованием глубокого обучения, демонстрируют выдающиеся результаты в автоматическом распознавании языка текста. Рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN), и трансформеры — это типичные архитектуры, применяемые для решения задачи распознавании языка текста.

Методы глубокого обучения могут извлекать высокоуровневые признаки из текста, учитывая контекст и долгосрочные зависимости между словами

**Достоинства нейросетевых методов в автоматическом распознавании языка:**

1. Высокая точность: Нейросетевые методы демонстрируют высокую точность в задачах распознавания языка благодаря способности изучать сложные зависимости в данных.

2. Адаптивность к различным языкам: Нейросетевые модели способны обучаться на большом объеме данных различных языков, что делает их адаптивными и эффективными для многих языковых сценариев.

**Недостатки нейросетевых методов в автоматическом распознавании языка:**

1. Требовательность к данным: Нейросетевые методы, особенно глубокие модели, часто требуют больших объемов размеченных данных для эффективного обучения.

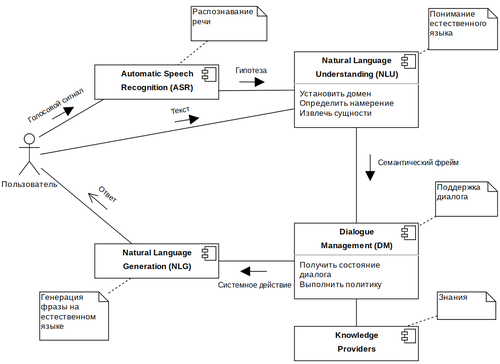
2. Вычислительная сложность: Обучение и использование глубоких моделей может быть вычислительно затратным процессом, что может быть проблемой при ограниченных ресурсах.

3. Неинтерпретируемость: Нейросетевые модели могут быть сложными и труднопонимаемыми, что затрудняет интерпретацию принятых ими решений.

В целом, несмотря на эти ограничения, нейросетевые методы остаются одними из наиболее эффективных в решении задачи автоматического распознавания языка текста.

#### **64. Структура диалоговой системы. бил 11**

По назначению можно выделить три типа диалоговых систем: общего назначения (general, чат-боты), задачеориентированные (task-oriented) и способные вести диалог на любую тему (open domain). Упрощенные версии диалоговых систем, такие как чат-боты имеют довольно простую архитектуру и состоят из набора правил и заранее подготовленных ответов.

Классический метод построения целеориентированных систем заключается в использовании цепочки модулей (конвейера)

Описание модулей:

* **ASR**. На вход поступает речь пользователя, которая затем [распознается](https://neerc.ifmo.ru/wiki/index.php?title=%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%80%D0%B5%D1%87%D0%B8) и переводится в текст. Результат работы компонента называют *гипотезой*, так как полученный текст может соответствовать исходному сообщению не полностью.
* **NLU**. Фраза в текстовом виде анализируется системой: определяется домен, намерение, именованные сущности. Для распознавания намерений может применяться обученный на [векторном представлении фраз](https://neerc.ifmo.ru/wiki/index.php?title=%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%B5%D0%B4%D1%81%D1%82%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81%D0%BB%D0%BE%D0%B2) классификатор. [Распознавание именованных сущеностей](https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0#spaCy) является отдельной задачей извлечения информации. Для ее решения используются [формальные языки](https://neerc.ifmo.ru/wiki/index.php?title=%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D1%85_%D1%8F%D0%B7%D1%8B%D0%BA%D0%BE%D0%B2), статистические модели и их комбинации. В результате работы компонента создается формальное описание фразы — семантический фрейм.
* **DM**. Состоянием диалога или контекстом является информация, которая была получена при общении с пользователем ранее. В соответствии с текущим состоянием выбирается политика поведения системы, корректируется семантический фрейм. В качестве поставщика знаний может выступать СУБД или Web API.
* **NLG**. В соответствии с выбранным действием осуществляется генерация ответа пользователю на естественном языке. Для генерации применяются [генеративные модели или шаблоны](https://neerc.ifmo.ru/wiki/index.php?title=%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D1%86%D0%B8%D1%8F_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%B0).

#### **65. Направления использования речевых систем.**

Перечислим некоторые конкретные применения речевого интерфейса, которые уже существуют и используются в различных прикладных системах. К числу

таких систем можно отнести следующие:

* системы поддержки безбумажных технологий: диктовка и формирование текстовых файлов на компьютере, системы подготовки документов, редакционно-издательские системы;
* речевые интерфейсы для пользователей-инвалидов по слуху и зрению;
* системы компьютерной телефонии (телефонные диалоговые информационно-справочные системы, включая справки по паролю, телефонные автосекретари, речевая электронная почта; речевой набор номера телефона и др.);
* системы речевого управления (информационные и навигационные системы, диспетчерские системы управления воздушным и наземным транспортом, тренажеры систем управления; интеллектуальные здания и др.);
* системы защиты доступа к базам данных, информации и объектам с использованием принципа парольных фраз ("Голосовой ключ");
* системы обнаружения голосовых сообщений (детекторы речи);
* системы защиты речевых сообщений (компрессия речи для повышения эффективности криптографической защиты речевых сообщений, повышение помехоустойчивости передачи речевых сообщений по сверхузкополосным каналам передачи данных и т.п.);
* системы-чтецы (например, система голосовых объявлений в общественном транспорте;
* системы голосового оповещения населения в чрезвычайных ситуациях);
* системы для криминалистической экспертизы на основе анализа голоса и речи;
* системы обучения языкам (в частности, иностранным), в число которых включаются также «говорящие» словари, речевые разговорники, системы обучения правильному произношению иностранных слов и т.п.;
* компьютерные системы обучения по различным предметным областям, использующие мультимодальный интерфейс;
* игровые компьютерные программы (в частности, компьютерные развивающие игры для детей).

#### 

#### 

#### **66. Характеристики речевого сигнала.**

Речевой сигнал, как и любой сложный сигнал, является результатом сложения нескольких гармонических составляющих, каждая из которых характеризуется своими значениями указанных параметров. Каждая такая составляющая называется гармоникой.

В результате разложения речевого сигнала на входящие в его состав гармонические колебания получается частотный спектр – амплитудно-частотная зависимость.

Спектр речевого сигнала выражает его частотный состав, т.е. самыми главными характеристиками речевого сигнала являются значения частот его гармонических составляющих. Именно в результате наложения этих частот и их восприятия органами слуха мы получаем возможность слышать богато украшенный звук – речевой сигнал.

С точки зрения лингвистики в речевом сигнале отражаются фонетические и просодические аспекты речи, с точки зрения акустики – акустические характеристики фонем и просодем, а информационная структура содержит дополнительную информацию о речевом сигнале, включающую его смысловое содержание, а также индивидуальные характеристики личности говорящего и передающей среды.



#### 

Спектральные характеристики всех звуковых сигналов в общем

случае подобны и описаны в большинстве книг курса общей физики.

Определяются они с помощью двух основных зависимостей,

раскрывающих их состав в полной мере: зависимость амплитуды от

времени (осциллограмма или амплитудно-временная характеристика) и

зависимость амплитуды от частоты (спектрограмма или амплитудночастотная характеристика). Особенностью же речевых сигналов

является наличие формантных областей – набора гармоник,

образующих определенный звук речи

Акустика — речевая (от греч. ἀκουστικός слуховой) раздел общей акустики, изучающий структуру речевого сигнала, процессы речеобразования и восприятия речи у человека и связанный с разработкой систем автоматического синтеза и распознавания речи.

Просо́дия (др.-греч. προσῳδία — ударение, припев; также просо́дика) — раздел фонетики, в котором рассматриваются такие особенности произношения, как высота, сила/интенсивность, длительность, придыхание, глоттализация, палатализация, тип примыкания согласного к гласному и другие признаки, являющиеся дополнительными к основной артикуляции звука[1]. Элементы просодии реализуются в речевом потоке на всех уровнях речевых сегментов (в слоге, слове, синтагме, фразе и т. д.), выполняя при этом смыслоразличительную роль. В рамках просодии изучается как субъективный уровень восприятия характеристик суперсегментных единиц (высота тона, сила/громкость, длительность), так и их физический аспект (частота, интенсивность, время)[2].

#### 

#### **67. Фонемы в речи. Фонетический состав русского языка.**

**Фонема** – это минимальная значимая звуковая единица речи.

Акустические свойства каждой фонемы обусловлены артикуляторными особенностями ее образования в речевом тракте человека, к которым относятся:

• место образования;

• способ образования.

Под **местом образования** понимается сужение артикуляторного тракта, определяющее его артикуляторную статику и резонансные свойства. Место образования гласных фонем обусловлено положением тела языка (переднее / заднее, высокое / низкое) и губ (огубленное / неогубленное). Место образования согласных фонем определяет положение смычки или щели (губное, зубное, альвеолярное, нёбное, фарингальное), а также заднее или переднее положение тела языка (разделение на мягкие и твердые согласные).

От способа образования фонем зависят энергетические и динамические особенности артикуляции фонемы. Под **способом образования** фонемы понимается то, каким образом образуется звук при прохождении воздуха через артикуляторный тракт человека во время произнесения фонемы.

По способу образования фонемы подразделяются на следующие виды:

• гласные: [у], [о], [а], [э], [ы], [и];

• плавные: [й];

• боковые (латеральные): [л] (твёрдый) и [л'] (мягкий);

• носовые: [м], [м'], [н], [н'];

• дрожащие: [р], [р'];

• звонкие: [й], [л], [л'], [м], [м'], [н], [н'], [р], [р'], [з], [з‘], [ж], [в], [в’], [б], [б’], [д], [д’], [г], [г’];

• глухие: [с], [с’], [ш], [щ’], [ф], [ф’], [х], [х’], [п], [п’], [т], [т’], [к], [к’];

• щелевые (фрикативные): [с], [с’], [ш], [щ’], [ф], [ф’], [х], [х’], [з], [з‘], [ж], [в], [в’];

• взрывные: [п], [п’], [т], [т’], [к], [к’], [б], [б’], [д], [д’], [г], [г’];

• аффрикаты: [ц], [ч’].

Гласные отличаются по месту образования положением тела языка в координатах верхний / нижний – передний / задний и положением губ в координатах огубленный / неогубленный:

• верхний, задний, огубленный – [у];

• нижний, задний, огубленный – [о];

• нижний, задний, неогубленный – [а];

• нижний, передний, неогубленный – [э];

• верхний, задний, неогубленный – [ы]; • верхний, передний, неогубленный – [и].

Согласные по месту образования отличаются положением смычки или щели (губное,

зубное, альвеолярное, заальвеолярное, нёбное), а также задним или передним положением

тела языка (мягкие и твердые):

• губные твёрдые – [м], [в], [б], [ф], [п];

• зубные твёрдые – [н], [л], [д], [р], [з], [с], [т], [ц];

• заальвеолярные твёрдые – [ж], [ш];

• нёбные твёрдые – [г], [к], [х];

• губные мягкие – [м’], [в’], [б’], [ф’], [п’];

• зубные мягкие – [н’], [л’], [д’], [р’], [з’], [с’], [т’];

• заальвеолярные мягкие – [щ’], [ч’]; • нёбные мягкие – [й], [г’], [к’], [х’].

Полный перечень 42 фонем русского языка:

[у] [о] [а] [э] [и] [ы]

[й] [л] [л'] [м] [м'] [н] [н'] [р] [р'] [п] [п'] [к] [к'] [т] [т']

[в] [в'] [з] [з'] [ж] [б] [б'] [д] [д'] [г] [г']

[ф] [ф'] [c] [c'] [ш] [щ'] [х] [х'] [ц] [ч’]

#### 

#### 

#### **68. Аллофон: определение, виды, отличия от фонемы.**

Модификации фонем, связанные с особенностями их окружения в слитном потоке речи другими фонемами, называют аллофонами. Аллофоны подразделяются на комбинаторные и позиционные.

Возникновение комбинаторных аллофонов связано с окружение данной фонемы другими фонемами. При этом, как следствие, происходит наложение артикуляции одного звука на артикуляцию другого. Это так называемый эффект коартикуляции. Примером коартикуляции является огубление согласных перед гласными [о] или [у].

Позиционные аллофоны возникают в связи с различным положением данной фонемы в слове или фразе по отношению к ударному слогу, концу или началу слова и т.п. Это явление называется процессом редукции (например, переход безударной [о] в [а], оглушение сонорных согласных [р] и [л] в конце фразы, аффрикация, т.е. удлинение шумовой фазы взрывных фонем в конце фразы и др.).

Вики:

Аллофон – реализация фонемы, её вариант, обусловленный конкретным фонетическим окружением.

Фоне́ма — минимальная смыслоразличительная единица языка. Фонема не имеет самостоятельного лексического или грамматического значения, но служит для различения и отождествления значимых единиц языка (морфем и слов))

Различаются комбинаторные и позиционные аллофоны.

Комбинато́рные аллофо́ны — реализации фонем, связанные с коартикуляцией под влиянием фонетического окружения звуков.

Примерами комбинаторных аллофонов в русском языке могут служить:

* продвинутые вперед гласные заднего ряда [а], [о], [у] после мягких согласных («пять», «тётя», «люди»);
* лабиализованные (огубленные) согласные перед гласными [о], [у];
* звонкие аффрикаты [дз], [д’ж'] на месте [ц], [ч] перед звонкими шумными («плацдарм», «ночь бы»).

Комбинаторными аллофонами также считаются назализованные гласные перед носовыми [n], [m], [ŋ] в английском языке. В некоторых языках мира комбинаторные признаки (например, назализация) могут распространяться на несколько слогов.

Позицио́нные аллофо́ны — реализации фонем, связанные с их фонетической позицией в слове или слоге. Под фонетической позицией принято понимать:

* положение звука по отношению к абсолютному началу слова (после паузы);
* положение звука по отношению к абсолютному концу слова (перед паузой);
* положение звука по отношению к ударению.

Позиционными аллофонами гласных [а], [о] в русском языке считаются гласные [ъ], [ʌ] в безударных слогах.

#### 

#### 

#### **69. Использование различных разделов лингвистики при реализации речевой системы.**

Основные разделы лингвистики и примеры использования знаний из этих разделов при реализации речевых систем. Здесь нас будут интересовать следующие разделы лингвистики: лексика; синтаксис; морфология; фонетика; семантика; прагматика.

Лексика – это активный словарь, с которым работает естественно-языковая система. Например, системы, проверяющие орфографические ошибки с учётом знаний о словообразовании, работают с 1–2 млн словоформ русского языка. Системы распознавания речи работают с меньшими словарями – это десятки, сотни, максимум тысячи слов.

Морфология – это раздел лингвистики, изучающий части речи и правила словообразования, в которых используются морфемы: приставки (префиксы), суффиксы, окончания и т.д. С использованием знаний морфологии, в частности, строятся морфологические таблицы с перечислением всех окончаний, суффиксов, парадигм, спряжений, склонений и т.д.

Фонетика – это раздел лингвистики, занимающийся изучением звукового строя естественного языка. Предметом рассмотрения фонетики являются звуки (фонемы) речи.

Синтаксис – это набор правил образования предложений, с помощью которых осуществляется построение фраз естественного языка.

Семантика занимается вопросами изучения смысла, заключенного в естественно-языковом высказывании. Это особый раздел лингвистики, который имеет тесные связи с искусственным интеллектом и проблемой формализации знаний.

На уровне прагматики рассматривается ситуация, в рамках которой происходит общение. Здесь изучается набор условий, при которых возможно или невозможно возникновение той или иной естественно-языковой фразы.

Рассмотрим примеры использования перечисленных разделов лингвистики в случае распознавания одной фразы:

«Саша кушал кашу».

Так как процесс распознавания – процедура достаточно непростая, то компьютерная система может «услышать»данную фразу по-разному, в том числе с ошибками, исправление которых возможно с привлечением знаний из различных разделов лингвистики.

Пусть система распознавания речи распознала указанную фразу следующим образом:

«Фаша кушал кашу».

Такая ситуация возможна, если, например, сообщение было передано по телефонному каналу, вследствие чего произошла потеря информации. Либо диктор, который произнес данную фразу, имеет трудности с произношением звука [c]. Итак, здесь мы имеем ошибку в первом слове. Эта ошибка легко исправима, если обратиться к лексике, т.е. к словарю. Очевидно, что при этом обнаружится, что слова "Фаша" в словаре нет, но можно найти ближайшее по звучанию – "Саша".

Другой вариант ошибки:

«Саша кушал кашо».

Здесь мы имеем ошибку на уровне морфологии. У основы "каш" окончания "о" быть не может . Поэтому снова следует выбрать ближайшее по звучанию сочетание – "кашу".

Система распознавания речи может допустить также синтаксическую ошибку:

«Саша кушали кашу».

Исправление данной ошибки возможно двумя способами: "кушал", если речь идет о мальчике, и "кушала" – о девочке. Эта неоднозначность может быть разрешена с учетом знаний о контексте.

Возможен также следующий вариант распознавания:

«Саша кушал Машу».

Здесь мы имеем уже семантическую ошибку, которая исправляется с привлечением знаний семантики и подбором слова, которое по смыслу может быть связано со словом "кушал" и является близким по звучанию к слову "Машу".

Наконец, еще один вариант ошибки распознавания:

«Саша сушил кашу».

Это прагматическая ошибка. Возможно, где-то на производстве по технологии кашу и могут сушить, но в данной ситуации это бессмыслица.

#### 

#### 

#### **70. Фонетика. Просодика.**

Фонетика (греч. phone звук) – 1) звуковой строй языка (звуки и их закономерные сочетания, слоги, ударение, интонация и т.п.), 2) раздел языкознания, в котором изучается звуковая сторона языка.

Фонетика занимает особое место среди лингвистических наук. Звуки речи, которые она изучает, являются способом существования языка, материальным средством его выражения. Значимые части слова (морфемы), слова в целом, грамматические формы слов, словосочетания и предложения выражаются с помощью звуков, поэтому фонетика как система тесно связана с другими системами русского языка, а фонетическая наука связана с науками, изучающими другие уровни языка. Специфика же фонетики обусловлена спецификой изучаемой ею единицы: звук, в отличие от всех других единиц языка, - одноплановая единица, имеющая только план выражения, форму, но лишенная содержания, значения. Таким образом, фонетика имеет дело с материальной стороной языка, со звуковыми средствами, лишенными самостоятельного значения.

Просодика - раздел [фонетики](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D0%BD%D0%B5%D1%82%D0%B8%D0%BA%D0%B0), в котором рассматриваются такие особенности [произношения](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B8%D0%B7%D0%BD%D0%BE%D1%88%D0%B5%D0%BD%D0%B8%D0%B5), как [высота](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D1%81%D0%BE%D1%82%D0%B0_%D0%B7%D0%B2%D1%83%D0%BA%D0%B0), [сила/интенсивность](https://ru.wikipedia.org/wiki/%D0%93%D1%80%D0%BE%D0%BC%D0%BA%D0%BE%D1%81%D1%82%D1%8C_%D0%B7%D0%B2%D1%83%D0%BA%D0%B0), [длительность](https://ru.wikipedia.org/w/index.php?title=%D0%94%D0%BB%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%8C_(%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)&action=edit&redlink=1), [придыхание](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%B4%D1%8B%D1%85%D0%B0%D0%BD%D0%B8%D0%B5), [глоттализация](https://ru.wikipedia.org/wiki/%D0%93%D0%BB%D0%BE%D1%82%D1%82%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F), [палатализация](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BB%D0%B0%D1%82%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F), [тип примыкания](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BC%D1%8B%D0%BA%D0%B0%D0%BD%D0%B8%D0%B5_(%D1%84%D0%BE%D0%BD%D0%B5%D1%82%D0%B8%D0%BA%D0%B0)) [согласного](https://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D0%B3%D0%BB%D0%B0%D1%81%D0%BD%D1%8B%D0%B5) к [гласному](https://ru.wikipedia.org/wiki/%D0%93%D0%BB%D0%B0%D1%81%D0%BD%D1%8B%D0%B5) и другие признаки, являющиеся дополнительными к основной [артикуляции](https://ru.wikipedia.org/wiki/%D0%90%D1%80%D1%82%D0%B8%D0%BA%D1%83%D0%BB%D1%8F%D1%86%D0%B8%D1%8F_(%D1%84%D0%BE%D0%BD%D0%B5%D1%82%D0%B8%D0%BA%D0%B0)) [звука](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D1%87%D0%B5%D0%B2%D0%BE%D0%B9_%D0%B7%D0%B2%D1%83%D0%BA). Просодика за­нимается интонацией, тоном, ударением.

Интонация — основной способ фонетической организации потока речи. Важнейшая ее составная часть — мелодика, т. е. повышение и понижение тона, образующее во фразе своеобразный тоновый рисунок. Кроме того, в качестве компонентов интонации выступают ритм (чередование кратких и долгих слогов), интенсив­ность (сила или слабость произносимых звуков, особенно гласных), темп (скорость протекания речи во времени), тембр (специфическая окраска, сообщающая речи те или иные эмоционально-экспрессив­ные свойства), расстановка пауз и т. п.

#### **71. Артикуляторные особенности акустических свойств фонем.**

Артикуляторные особенности представляют собой характеристики движений артикуляторных органов, таких как губы, язык, нёбо, голосовые складки и др., при произнесении фонем. Эти движения влияют на акустические свойства речи. Вот несколько артикуляторных особенностей, которые оказывают влияние на акустические характеристики фонем:

1. Место образования звука:

* Губы (билабиальные, лабиодентальные, лабиовелярные и т.д.).
* Альвеолярные (язык прикладывается к альвеолам, где производятся звуки [т], [д], [н]).
* Постальвеолярные (язык прикладывается к задней части альвеол, [ш], [ж], [ч], [дж]).
* Веларные (язык поднимается к мягкому небу, [к], [г]).
* Увулярные (язык поднимается к заднему увуле, [х], [ғ]).

2. Способ образования звука:

* Остановочные или плоцкие звуки (когда звук полностью прерывается, например, [п], [т], [к]).
* Фрикативные звуки (когда звук произносится с узким проходом воздуха, например, [ф], [с], [ш]).
* Аффрикаты (сочетание остановочного и фрикативного звуков, например, [ч], [дж]).
* Носовые звуки (звуки, производимые через нос, например, [м], [н]).
* Трель (колебания воздушного потока через артикуляторы, например, [р]).

**3.Лабиализация:**

* *Лабиализированные:* Звуки, при которых артикуляция включает движение губ. Например, [в], [у]

**4.Напряжение голосовых складок (Голосность):**

* *Глухие:* Звуки, при которых голосовые складки не колеблются. Например, [п], [т], [к].
* *Звонкие:* Звуки, при которых голосовые складки колеблются. Например, [б], [д], [г].

#### **72. Место и способ образования фонем. Классификация фонем.**

Фонемы образуются в речевых органах человека, таких как губы, язык, рёбра и гортань. Эти органы изменяют своё положение и форму, чтобы создавать различные звуки. Процесс образования фонем называется артикуляцией.

Кроме того, фонемы могут быть образованы изменением потока воздуха, проходящего через речевые органы. Например, задержка потока воздуха в альвеолярном ряду может порождать звук [т], а звук [д] производится при одновременном включении голоса.

Звуки могут также образовываться при соприкосновении органов речи, как, например, звук [м], образуемый закрытой губной позицией.

Фонема – это минимальная значимая звуковая единица речи. Любое сообщение можно написать буквами, а можно – фонемами.

Акустические свойства каждой фонемы обусловлены артикуляторными особенностями ее образования в речевом тракте человека, к которым относятся:

* место образования;
* способ образования.

Под **местом образования** понимается сужение артикуляторного тракта, определяющее его артикуляторную статику и резонансные свойства. Место образования гласных фонем обусловлено положением тела языка (переднее / заднее,

высокое / низкое) и губ (огубленное / неогубленное). Место образования согласных фонем определяет положение смычки или щели (губное, зубное, альвеолярное, нёбное, фарингальное), а также заднее или переднее положение тела языка (разделение на мягкие и твердые согласные).

От **способа образования** фонем зависят энергетические и динамические особенности артикуляции фонемы. Под способом образования фонемы понимается то, каким образом образуется звук при прохождении воздуха через артикуляторный тракт человека во время произнесения фонемы.

В соответствии с артикуляторными особенностями образования, фонемы различаются:

1. по способу образования;
2. по месту образования.

По способу образования фонемы подразделяются на следующие виды:

* гласные: [у], [о], [а], [э], [ы], [и];
* плавные: [й];
* боковые (латеральные): [л] (твёрдый) и [л'] (мягкий);
* носовые: [м], [м'], [н], [н'];
* дрожащие: [р], [р'];
* звонкие: [й], [л], [л'], [м], [м'], [н], [н'], [р], [р'], [з], [з‘], [ж], [в], [в’], [б], [б’], [д], [д’], [г], [г’];
* глухие: [с], [с’], [ш], [щ’], [ф], [ф’], [х], [х’], [п], [п’], [т], [т’], [к], [к’];
* щелевые (фрикативные): [с], [с’], [ш], [щ’], [ф], [ф’], [х], [х’], [з], [з‘], [ж], [в], [в’];
* взрывные: [п], [п’], [т], [т’], [к], [к’], [б], [б’], [д], [д’], [г], [г’];
* аффрикаты: [ц], [ч’].

Гласные отличаются по месту образования положением тела языка в координатах верхний / нижний – передний / задний и положением губ в координатах огубленный / неогубленный:

* верхний, задний, огубленный – [у];
* нижний, задний, огубленный – [о];
* нижний, задний, неогубленный – [а];
* нижний, передний, неогубленный – [э];
* верхний, задний, неогубленный – [ы]; • верхний, передний, неогубленный – [и].

Согласные по месту образования отличаются положением смычки или щели (губное,

зубное, альвеолярное, заальвеолярное, нёбное), а также задним или передним положением тела языка (мягкие и твердые):

* губные твёрдые – [м], [в], [б], [ф], [п];
* зубные твёрдые – [н], [л], [д], [р], [з], [с], [т], [ц];
* заальвеолярные твёрдые – [ж], [ш];
* нёбные твёрдые – [г], [к], [х];
* губные мягкие – [м’], [в’], [б’], [ф’], [п’];
* зубные мягкие – [н’], [л’], [д’], [р’], [з’], [с’], [т’];
* заальвеолярные мягкие – [щ’], [ч’]; • нёбные мягкие – [й], [г’], [к’], [х’].

Полный перечень 42 фонем русского языка:

[у] [о] [а] [э] [и] [ы]

[й] [л] [л'] [м] [м'] [н] [н'] [р] [р'] [п] [п'] [к] [к'] [т] [т']

[в] [в'] [з] [з'] [ж] [б] [б'] [д] [д'] [г] [г']

[ф] [ф'] [c] [c'] [ш] [щ'] [х] [х'] [ц] [ч’]

#### **73. Коартикуляция, редукция и др. явления в речи.**

Коартикуляция, редукция и другие явления в речи - это различные аспекты произношения, которые влияют на то, как произносятся звуки и слова в различных контекстах. Вот их основные характеристики:

1. Коартикуляция:

* Это явление, при котором два или более звука произносятся одновременно или последовательно, влияя друг на друга.
* Примеры: В некоторых случаях, при произношении слова, звук может изменяться из-за артикуляции следующего или предыдущего звука. Например, в слове "василиск" согласные "с" и "к" могут взаимодействовать, изменяя произношение "с".

2. Редукция:

* Это упрощение или сокращение произношения звуков или слогов в речи для экономии времени и энергии.
* Примеры: В быстрой и разговорной речи, некоторые звуки могут быть опущены или заменены более простыми звуками. Например, "немного" может произноситься как "немнО".

3. Ассимиляция:

* Это явление, при котором звук меняется под влиянием соседних звуков, становясь более похожим на них.
* Примеры: В слове "самолет" согласный звук "м" может звучать ближе к "л" из-за влияния этого звука.

4. Элизия:

* Это явление, при котором звук или звуки опускаются в произношении.
* Примеры: В разговорной речи, некоторые звуки могут опускаться, например, "вот он" может произноситься как "вотн".

5. Интонационные особенности:

* Это изменения тональности и интонации в речи, которые отражают эмоциональное состояние, ударение и намерение говорящего.
* Примеры: Повышение или понижение интонации может передавать эмоциональные нюансы или сигнализировать о типе предложения (вопрос, утверждение и т.д.).

Эти явления являются нормальными и общими в речи, особенно в разговорной форме. Они могут изменяться в зависимости от контекста, скорости речи, социокультурных факторов и степени формальности общения.

#### **74. Просодические характеристики речи.**

К просодическим характеристикам речи относят базисные (элементарные) средства, которые обычно объединяются в просодические комплексы. В число базисных средств включают просодические признаки, соотносящиеся с основными акустическими признаками звука — длительностью, силой/интенсивностью, высотой. На фразовом уровне к базисным средствам относят также значимое отсутствие звука, паузу, и особенности фонации (просодического тембра). Ряд просодических средств, таких, как придыхание, глоттализация, палатализация, тип примыкания согласного к гласному, модифицируют элементы слога. Большинство остальных просодических средств выходят за рамки локального использования и определяют характеристики языковых единиц более высокого уровня.

К комплексным просодическим средствам, которые формируются из базисных средств, относят, например, ударение и интонацию. Ударение образуется из таких компонентов, как сила/интенсивность, длительность, высота голосового тона и тембральные характеристики — по преобладанию этих фонетических компонентов в ударном слоге выделяют языки с динамическим (силовым, экспираторным), музыкальным (тоническим), количественным (квантитативным) и качественным ударением, а также языки в которых отмечается та или иная комбинация силы, длительности, высоты и тембра — языки с количественно-динамическим, качественно-количественным и другими подобными типами ударения. Интонацию формируют такие просодические элементы, как мелодика, пауза, сила/интенсивность, длительность и фонация. Ударение реализуется на уровне разных по сложности языковых единиц — на уровне слова, синтагмы и фразы (различают словесное, синтагматическое (тактовое) и фразовое ударение), интонация — только на уровне фразы.

#### **75. Физические параметры речевого сигнала, с помощью которых описывается просодия речи.**

Физические параметры речевого сигнала, используемые для описания просодии речи, включают в себя следующие аспекты:

1. *Интонация*: Изучение изменения тональности и высоты голоса в процессе речи, описываемое с помощью параметров частоты (герцы), амплитуды и длительности фраз и слов.
2. *Темп речи*: Оценка скорости речи и пауз между фразами и словами, измеряемая в словах в минуту или символах в минуту.
3. *Громкость*: Определение уровня громкости речи, который можно измерять в децибелах (дБ) или аналогичных шкалах.
4. *Ритм*: Анализ паттернов ударений и акцентов, характеризующих ритмическую структуру речи.
5. *Интонационные контуры*: Описание изменения тональности и акцентов в предложении с помощью специальных графических контуров.

.

#### **76. Функции интонации в речи.**

Интонация – единство взаимосвязанных компонентов: мелодики, интенсивности, длительности, темпа речи и тембра произнесения. Вместе с ударением интонация образует просодическую систему языка]. В отличие от сегментных фонетических единиц (фонем) и их дифференциальных признаков, не имеющих собственного плана содержания, все интонационные единицы двусторонние, то есть являются важным средством формирования высказывания и выражения его смысла.

**Функции** интонации:

1. Семантическая: выражение цели высказывания, его «коммуникативной организации», различение частей высказывания по смысловой важности, членение на тему и рему.
2. Экспрессивная: выражение эмоций говорящего, воздействие на эмоции слушающего. Реализуется интегральными средствами, что является причиной характерности последних для диалогической и художественной речи.
3. Синтаксическая: синтаксическая составляющая предложения зачастую оформляется как единая интонационная группа с общими интегральными характеристиками и одним фразовым акцентом.
4. Эвфоническая: способствование благозвучию отрезка речи, в частности его деление на соизмеримые по времени звучания фрагменты, чередование сильных и слабых ударений

Неопределенность в трактовке понятия «функция» привела к появлению разнородных по принципам и противоречивых по содержанию систем классификации функций и интонации. Различные авторы выделяют эмоциональную и интеллектуальную, вербальную и голосовую, логическую, эмфатическую и акцентирующую, эмоциональную, эмфатическую и физиологическую и другие функции [1].

Л.Р. Зиндер дал трактовку термина «языковая функция» – функцией данного языкового средства следует считать «его предназначенность для передачи соответственной языковой категории». В соответствии с этой трактовкой можно выделить следующие **функции интонации**:

1) функция членения на синтагмы;

2) функция связи между синтагмами;

3) функция различения коммуникативных типов (ситуаций);

4) функция акцентного выделения элементов синтагмы;

5) функция выражения эмоциональных значений;

6) функция передачи модальных отношений.

#### 

#### 

#### **77. Ударение: понятие, назначение, типы.**

Ударение в слове — это особое выделение (акцент) одного из слогов для придания ему особой интонации или смысловой выразительности.

**Назначение ударения:**

*Смысловое различие:* Ударение может изменять смысл слова. Например, в русском языке различие между словами "зАмок" (крепость) и "замОк" (замочек) обусловлено местом ударения.

*Ритм и интонация:* Ударение помогает создать ритм речи и ее интонацию, делая высказывание более выразительным и понятным.

Типы ударения:

По способу выделения слога из такта можно ударение подразделить на:

динамическое (экспираторное, силовое) – слог выделяется силой выдоха

квантитативное (количественное) – слог выделяется долготой

мелодическое (музыкальное, тоновое) – слог выделяется движением голосового тона.

Чистых типов ударения фактически не бывает, почти всегда один тип сопровождается другим, хотя один из них преобладает.

Динамическое ударение имеется в русском, английском, чешском и других языках. В русском языке динамическое ударение сопряжено с квантитативным.

Мелодическое ударение встречается, напр., в сербохорватском, литовском, китайском и др. яз.

Динамическое ударение бывает или сильное (в русском, английском и др. яз) или слабое (в чешском, грузинском и др.)

Функционально ударение соотносится с разными сегментными единицами языка. В зависимости от связи с той или иной единицей выделяются такие типы ударения, как[1]:

Словесное ударение — выделяет слог в составе фонетического слова.

Синтагматическое, или тактовое, ударение — выделяет слово в составе синтагмы.

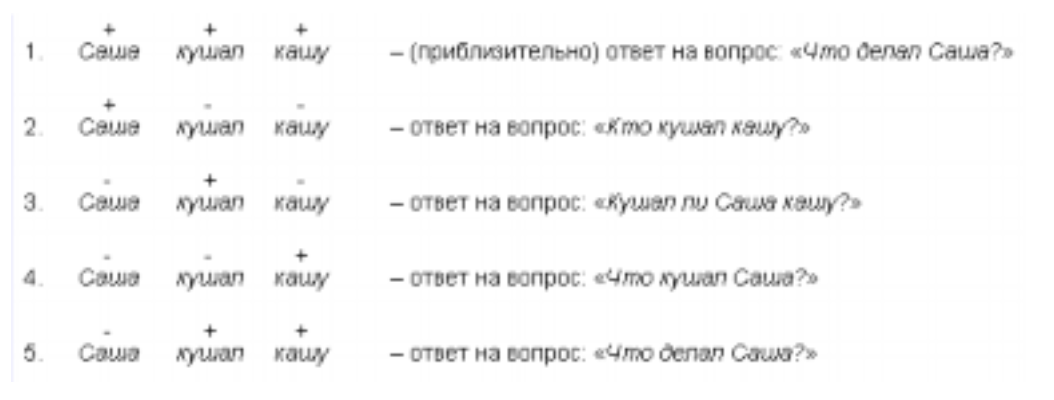
Фразовое ударение — выделяет синтагму в составе фразы.

#### **78. Синтагма: понятие, типы, примеры**.

Синтагма – это часть фразы, выделенная паузами с обеих сторон. В зависимости от позиции синтагмы во фразе можно выделить три типа синтагм:

* начальная – синтагма, которая стоит в начале фразы;
* конечная – в конце фразы;
* срединные – все остальные синтагмы, которые стоят между начальной и конечной

Паузы в речи являются ее естественной составляющей, так как при чтении текста любой человек делает дыхательные остановки. Это может быть связано с естественной потребностью сделать очередной вдох либо с необходимостью выделить интонационно какую-либо часть произносимой речи. Во втором случае имеется ввиду выделение той части фразы, которая на письме, как правило, заканчивается знаком препинания: запятой, двоеточием и т.д. Часто синтагмы никак не обозначаются в тексте, а только подразумеваются. Синтагма может состоять как из одного слова, так и из нескольких слов – двух, трех и т.д., в зависимости от фразы и от особенностей дикции говорящего.



Синтагма состоит из одной или нескольких акцентных групп (АГ). В акцентной группе каждое слово отмечено ударением. Мы выделили 2 типа ударений: сильное (основное, +) и слабое (побочное, -). При выделении акцентных групп используется правило, согласно которому в синтагме столько акцентных групп, сколько имеется сильных ударений. Если в синтагме есть слова, помеченные слабыми ударениями, то они присоединяются к словам с сильным ударением в одну акцентную группу.

Позиция синтагмы во фразе определяет интонационный тип этой синтагмы. Наиболее значимой является конечная синтагма, которая в основном определяет интонационный тип фразы. Конечная синтагма легко определяется тем, что она стоит перед каким-либо знаком препинания

В рассмотренном выше примере фразы «Саша кушал кашу» можно выделить одну синтагму (она является конечной)

#### **79. Акцентная группа: понятие, типы, примеры.**

**Акцентная группа** — это группа слов, выделенных особым образом, чтобы привлечь внимание к определенному элементу предложения и выделить его значение. Акцентные группы могут использоваться для подчеркивания информации, изменения интонации и передачи эмоциональной окраски в высказывании.

**Типы акцентных групп**:

1. Акцент на **слове**:

**Пример**: "Я не сказал, что *ты* это сделал."

1. Акцент на **фразе**:

**Пример**: "Это *он* приготовил ужин."

1. Акцент на **части предложения**:

**Пример**: "Он читал *книгу*, которую *мы* ему подарили."

1. **Эмоциональный** акцент:

**Пример**: "Я так *рад*, что ты здесь!"

1. **Логический** акцент:

**Пример**: "Он не *любит* футбол."

#### 80. Основные интонационные типы конечных акцентных групп.

Конечные акцентные группы, завершающие фразу или предложение, могут иметь разные интонационные типы, указывающие на различные нюансы смысла или эмоциональное окраску высказывания. Основные интонационные типы конечных акцентных групп включают:

1. Вопросительная интонация(Восходящий интонационный тип):

* Выражает вопрос(Этот тип интонации характеризуется подъемом интонации к концу акцентной группы).
* Интонационно поднимается в конце фразы.
  + ***Пример:* "Ты идешь в магазин?"**

2. Утвердительная интонация(Нисходящий интонационный тип):

* Подтверждает факт или выражает утверждение(Интонация опускается к концу акцентной группы).
* Интонационно опускается или плоская на конце фразы.
  + ***Пример:* "Я сегодня пошел в кино."**

3. Восклицательная интонация:

* Выражает сильные эмоции, восклицание или удивление.
* Интонационно поднимается в конце фразы.
  + ***Пример:* "Как это потрясающе!"**

4. Нейтральная интонация(Плавный тип интонации:):

* Не несет в себе особой эмоциональной окраски(В этом типе изменения интонации более плавные и менее выраженные).
* Имеет плавный переход без резких подъемов или падений в интонации.
  + ***Пример:* "Он приготовил обед."**

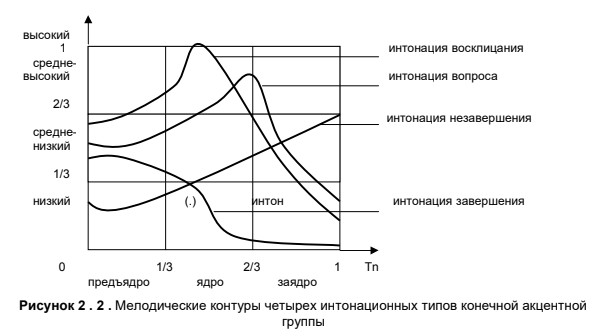
**5. Падающе-восходящий (или сочиненный) тип:**

* Описание: Интонация начинается с падения и затем переходит в подъем.
* Пример: "Он пошел в магазин и... вернулся." В начале фразы интонация опускается, чтобы выделить первую часть высказывания, а затем поднимается для выделения окончания или продолжения.

Эти интонационные типы конечных акцентных групп позволяют дополнительно выразить эмоциональную или смысловую окраску высказывания в зависимости от контекста и намерений говорящего.

#### **81. Состав акцентной группы.**

Акцентная группа (АГ) делится на три составляющие: предъядро, ядро и заядро. Это так называемые элементы акцентной группы (ЭАГ). При этом ядро считается главным элементом АГ. Правила разбивки акцентной группы на элементы достаточно просты: ядро является главной ударной гласной слова, заядро – это все звуки справа от ядра, предъядро – все звуки слева от ядра. Получается, что заядро и предъядро могут содержать различное количество звуков вплоть до их отсутствия. Например, фонетический период: «Ты поедешь сегодня? – Да!» – содержит фразу «Да!», которая является при этом и синтагмой, и акцентной группой одновременно. В этом случае один-единственный гласный звук придется делить на три части, потому что нужно выделить интонационный тип и определить ядро, заядро и предъядро.

Интонационный тип фразы определяется положением мелодического максимума относительно главного ударного гласного. Причем положению максимума в центре гласного соответствует интонация побуждения или восклицания, сдвигу его влево – интонация завершения или утверждения, а сдвигу вправо – интонация незавершенности или вопроса. Другие параметры мелодической кривой (значение максимума и ее ширина) существенны при передаче эмоциональных оттенков речи.

#### **82. Задача автоматического распознавания языка текста.**

Задача автоматического распознавания языка текста является одной из основных задач обработки естественного языка. Она заключается в определении языка, на котором написан текст, без прямого указания на этот язык.

Существует несколько подходов к решению этой задачи. Один из самых распространенных подходов - использование статистических методов, основанных на частотности букв, слов или н-грамм в тексте. Эти методы предполагают, что разные языки имеют различные частотные характеристики, которые можно использовать для их распознавания.

Другой подход - использование машинного обучения, основанного на классификации. В этом случае, модель обучается на размеченном наборе данных, содержащем тексты на разных языках, и потом применяется для классификации новых текстов на основе их языка.

Существуют и другие методы, такие как использование нейронных сетей или комбинация различных подходов для получения более точных результатов.

Задача автоматического распознавания языка текста имеет широкий спектр применений, например, в многопользовательских приложениях, мультиязычных системах машинного перевода, обработке больших объемов текстовых данных и многих других областях.

#### **83. Задача автоматического распознавания языка текста. Метод n-грамм, достоинства и недостатки.**

\* из методы к 2 лабе

В 1994 году В. Канвар и Дж. Тренкл предложили метод для определения языка и кодировки документа по его содержимому основываясь на статистиках документов, для которых язык и кодировка известны заранее. Метод основывается на подсчете частот N-грамм (подстрок длины не более N) и предположении, что примерно 300 самых частоиспользуемых N-грамм сильно зависят от языка.

В основе этого метода лежит Закон Зипфа – эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова). Например, первое по используемости слово встречается примерно в два раза чаще, чем второе, и в три раза чаще, чем третье.

Алгоритм метода заключается в нахождении частот N-грамм для всех тестовых документов, для которых известен язык, а также для каждого документа, язык которого пытаемся определить. После этого среди всех тестовых документов находим тот, для которого расстояние от его N-граммной статистики до статистики тестируемого документа минимально. После этого языком тестируемого документа считается язык найденного тестового документа.

Расстояние между статистиками подсчитывается следующим образом: все N-граммы сортируются в порядке убывания частоты их появления, затем для каждой N-граммы вычисляется разница её позиций в отсортированном списке N-грамм тестового и тестируемого документов. Расстояние между статистиками определяется как сумма разниц позиций каждой N-граммы. Если N-грамма отсутствует в профиле категории (языка), то ей назначается максимальная величина оценки несовпадения позиций N-грамм. Значение N предлагается использовать равным 5.

\* из чата

Достоинства: Метод n-грамм прост в реализации и может быть эффективным для определения языка текста, особенно для коротких текстов

Недостатки: Для более высоких значений n (например, триграммы или больше), размер словаря может значительно возрастать, требуя больше памяти и вычислительных ресурсов. Для редких n-грамм метод может столкнуться с проблемой разреженности данных, особенно при обучении на небольших корпусах текстов. Метод n-грамм не учитывает семантическую связь между словами или смысл текста, фокусируясь исключительно на последовательности символов или слов.

#### **84. Задача автоматического распознавания языка текста. Алфавитный метод, достоинства и недостатки.**

Автоматическая обработка текстов (АОТ) предполагает решение многих задач,

которые условно можно разбить на два уровня. Задачи высокого уровня представле-

ны задачами распознавания речи, реферирования текстов, распознавание языка текста, генерации документов, машинного перевода, извлечения информации, обучения языку, т. е. приложениями.

К задачам низшего уровня относят грамматический разбор, снятие смысловой многозначности, корректировку орфографии и синтаксический разбор, т. е. задачи собственно лингвистической обработки ЕЯ. К настоящему времени этот круг задач

значительно расширился и в целом охватывает всю индустрию развития и поддержки компьютерной формы существования ЕЯ.

Главная проблема при решении указанных задач состоит в необходимости об-

рабатывать неструктурированные тексты. Единый типовой алгоритм их автоматиче-

ской обработки создать не удается, поскольку конкретный вид алгоритма, в первую

очередь, определяется строем языка.

Для этого были испробованы различные методы и алгоритмы, но подробнее остановимся на алфавитном методе.

**Алфавитный метод** — это метод распознавания языка текста, основанный на анализе статистики частоты букв в тексте. Идея заключается в том, чтобы определить, какие буквы чаще всего встречаются в тексте, и сравнить этот профиль с профилями известных языков.

**Достоинства алфавитного метода:**

1. Простота: Этот метод прост в реализации и понимании. Его можно легко применять для базового определения языка текста.

2. Низкие вычислительные затраты: В сравнении с более сложными методами, алфавитный метод обычно не требует больших вычислительных ресурсов.

3. Эффективность на коротких текстах: В случае коротких текстов, где более сложные методы могут испытывать затруднения из-за ограниченного контекста, алфавитный метод может быть относительно эффективен.

**Недостатки алфавитного метода:**

1. Чувствительность к шуму: Алфавитный метод может быть чувствителен к шуму в тексте и изменениям статистики букв, что может снизить точность.

2. Ограниченность в распознавании сложных языков: Для сложных языков, в которых статистика букв может сильно различаться в зависимости от контекста, алфавитный метод может быть менее точен.

3. Неучет грамматических особенностей: Алфавитный метод не учитывает грамматические особенности языка, что ограничивает его способность точно определить язык, основываясь только на статистике букв.

4. Неэффективность для многозначных языков: В регионах, где говорят на нескольких языках, алфавитный метод может столкнуться с трудностями в определении конкретного языка.

В целом, алфавитный метод может быть полезным для базового определения языка в простых случаях, но для более точных результатов часто применяются более сложные методы, такие как нейросетевые подходы.

#### **85. Задача автоматического распознавания языка текста. Метод коротких слов, достоинства и недостатки.**

Метод коротких слов при построении ПОЯ(поисковый образ языка) использует слова определенной длины, не превышающей заданный порог (например, слова до 5 символов, которые встречаются более 3 раз). Вероятность появления в тексте i-й лексемы рассчитывалась как отношение ее частоты к общей сумме частот всех лексем из полученного набора. Предложение входного документа разбивалось на лексемы, и лексемам, присутствующим в ПОЯ, назначались их частоты, а отсутствующим – некоторая минимальная частота. Вероятность принадлежности предложения языку рассчитывалась как произведение вероятностей его лексем

Достоинства:

* **Быстрота и простота:**
  + Метод коротких слов прост в реализации и быстр в выполнении анализа.
* **Эффективность для определенных задач:**
  + Может быть эффективным для определенных задач, таких как категоризация текста или выделение тем.

Недостатки:

* + Короткие слова, особенно частотные союзы, артикли и предлоги, могут быть общими для нескольких языков и не обеспечивать достаточной уникальности для точного распознавания языка.
  + Использование только коротких слов может ограничивать способность модели адекватно учесть особенности более длинных выражений, фраз и структур языка.
  + Некоторые стили или жанры текста могут иметь уникальные лексические особенности, которые могут не отражаться в частотных коротких словах.
  + Метод коротких слов может быть более чувствителен к размеру обучающей выборки, и при ограниченном объеме данных он может не давать надежных результатов.
  + Короткие слова могут быть подвержены шумам в данных, таким как орфографические ошибки, что может снижать точность метода.
  + Если языки имеют общие корни и лингвистические особенности, метод коротких слов может столкнуться с трудностями в точной классификации.

#### **86. Полиинформативность речевого сигнала. Виды информации, заключенной в речевом сигнале.**

Всякая техническая система связи включает такие функциональные элементы, как источники информации, кодирующие устройства, переносчики, модуляторы, системы разделения каналов, демодуляторы, декодеры и др. Естественные системы связи, в частности, устная речь, не являются в этом смысле исключением. Основной особенностью речевого сигнала, вытекающей из такого рассмотрения, является его **полиинформативность** и полимодуляционность.

**Полиинформативность** речевого сигнала заключается в многообразии типов информации, передаваемой с помощью речи (рис). Информация, заключенная в речевом сигнале, может быть разделена на два вида:

* основная, т.е. смысловое содержание речи (семантика высказывания);
* дополнительная, т.е. информация об индивидуальных особенностях говорящего, его физическом и эмоциональном состоянии, а также характеристики передающей среды.



#### **87. Методы синтеза речевого сигнала.**

Синтез речи в широком смысле представляет собой восстановление формы речевого сигнала по его параметрам. В узком смысле – это формирование речевого сигнала по печатному тексту.

Все **методы синтеза речи**, в зависимости от принципов, положенных в основу, **можно подразделить на три группы**:

– **параметрический синтез**;

– **конкатенативный, или компиляционный (компилятивный) синтез**;

– **полный синтез речи по правилам**.

Каждые из методов синтеза речевых сигналов различается сложностью алгоритма и основными принципами синтеза, используемыми в каждой отдельно взятой реализации.

**Параметрический синтез**.

Параметрический синтез речи является конечной операцией в вокодерных системах, где речевой сигнал представляется набором небольшого числа непрерывно изменяющихся параметров. Параметрический синтез целесообразно применять в тех случаях, когда набор сообщений ограничен и изменяется не слишком часто.

К **достоинствам** данного метода относится:

– возможность записи речи для любого языка и любого диктора;

– высокое качество;

– простота реализации.

Основным **недостатком** данного метода является невозможность применения для произвольных, заранее не заданных сообщений.

**Компиляционный синтез**.

Компиляционный синтез сводится к составлению сообщения из предварительно записанного словаря исходных элементов синтеза. Размер элементов синтеза составляет не меньше слова. При этом содержание синтезируемых сообщений фиксируется объемом словаря. Как правило, число единиц словаря не превышает нескольких сотен слов. Основная проблема в компиляционном синтезе – объемы памяти для хранения словаря. В этой связи используются разнообразные методы сжатия/кодирования речевого сигнала.

**Достоинствами** данного метода являются:

– простота реализации;

– возможность применения для синтеза произвольных выражений (ограниченно).

К **недостаткам** можно отнести:

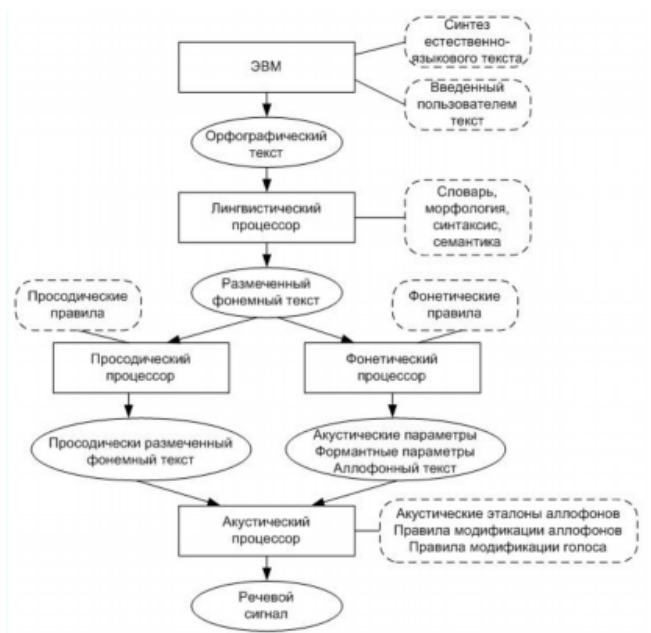
– необходимость больших затрат памяти на хранение словаря;

– необходимость использования сжатия речевого сигнала.

**Полный синтез речи по правилам**.

Метод синтеза речи по правилам базируется на запрограммированном знании акустических и лингвистических ограничений и не использует непосредственно элементов человеческой речи. Полный синтез речи по правилам обеспечивает управление всеми параметрами речевого сигнала и, таким образом, может генерировать речь по заранее неизвестному тексту. В этом случае параметры, полученные при анализе речевого сигнала, сохраняются в памяти так же, как и правила соединения звуков в слова и фразы.

#### **88. Структура синтезатора речи по тексту.**



Основными компонентами фонемного синтезатора речи по тексту являются:

* лингвистический процессор;
* фонетический процессор;
* просодический процессор;
* акустический процессор.

Как видно из данной структуры, синтезатор речи состоит из ряда процессоров (обработчиков), основная задача которых заключается в постепенной обработке входного орфографического текста. Эта обработка представляет собой разметку текста с целью его преобразования в последовательность параметров, описывающих речевой сигнал. Результирующая последовательность параметров должна быть «воспринята» акустическим процессором, задачей которого является непосредственное воспроизведение звука, т.е. собственно синтез речи.

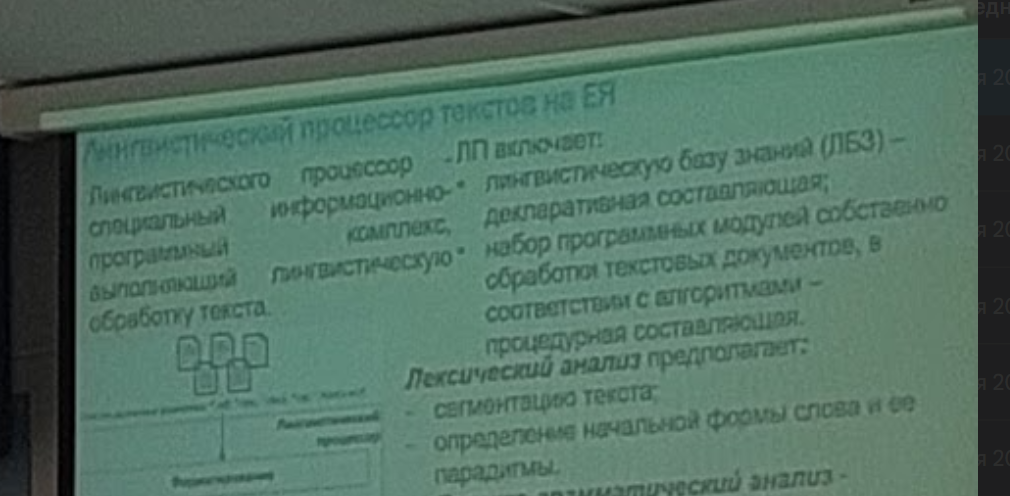
#### **89. Лингвистический процессор.**

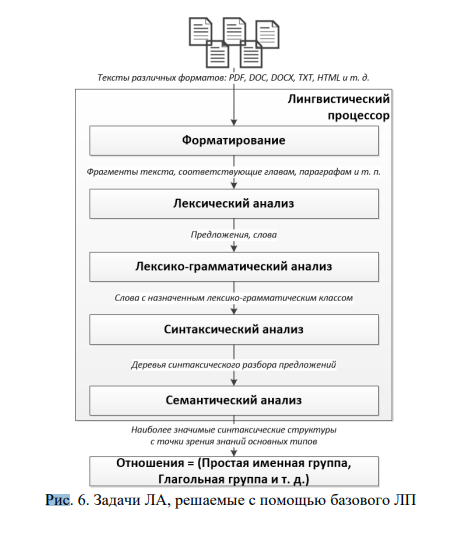
Лингвистический процессор - это компьютерная программа, которая проводит автоматический лингвистический анализ текста на естественном языке. Он состоит из нескольких компонентов, каждый из которых выполняет определенную функцию, такую как лексический, морфологический, синтаксический и семантический анализ.

Лингвистический процессор может использоваться для различных целей, таких как автоматический перевод, извлечение информации, анализ тональности, генерация текста и другие. Он может работать как в режиме онлайн-обработки, так и в режиме обработки больших объемов текста в автономном режиме.

Компоненты лингвистического процессора:

1. Лексический анализатор - отвечает за разбор текста на лексические единицы, такие как слова, числа, знаки пунктуации, идентификацию неизвестных слов и т.д.
2. Морфологический анализатор - определяет грамматическую форму слова и его часть речи.
3. Синтаксический анализатор - анализирует синтаксическую структуру предложения и определяет зависимости между словами в предложении.
4. Семантический анализатор - анализирует смысл текста и связывает его с знаниями, хранящимися в базе знаний.
5. Генератор текста - создает новый текст на основе полученных результатов анализа.
6. Интерфейс пользователя - это часть системы, которая обеспечивает взаимодействие между пользователем и лингвистическим процессором. Это может быть голосовой ассистент, приложение на мобильном устройстве или сайт с возможностью чата.
7. Каждый из компонентов лингвистического процессора выполняет свою задачу, и результаты анализа используются для решения различных задач в области обработки естественного языка.





#### **90. Просодический процессор.**

**Просодический процессор** — это часть системы синтеза речи, отвечающая за анализ и управление просодическими характеристиками речи. Процессор обрабатывает размеченный фонемный текст и добавляет просодические элементы, такие как интонация, интенсивность (громкость), длительность и частота основного тона (мелодика). Применение просодического процессора позволяет придать речи выразительность, эмоциональный оттенок и естественность.

**Основные функции просодического процессора**:

* **Деление на акцентные группы** (АГ): Процессор разделяет фонемный текст на акцентные группы (группы слов, объединенные единым акцентом).
* **Разметка акцентных групп** (ЭАГ): Каждая акцентная группа дополнительно размечается на элементы акцентных групп, такие как интонационное предъявление, ядро и заядро.
* **Установка параметров просодии**: Процессор определяет значения интенсивности (громкости), длительности фонем и частоты основного тона (мелодики) для каждого элемента акцентной группы.
* **Работа с просодической базой данных (БД) и правилами**: Процессор использует просодическую базу данных, содержащую информацию о том, как различные элементы речи связаны с просодическими характеристиками. Также могут применяться правила для уточнения просодической разметки.
* **Определение формы интонационного контура**: Процессор формирует интонационный контур, определяя, как изменяется высота тона внутри каждой акцентной группы.
* **Управление динамикой и эмоциональным окрасом**: Процессор способен управлять динамикой речи, включая громкость и интонацию, что позволяет придавать выражение и эмоциональный окрас синтезированной речи.

#### **91. Фонетический процессор.**

**Задача фонетического процессора** заключается либо в генерации последовательности акустических параметров речи,либо в выработке формантных параметров фонем, либо в подстановке позиционных и комбинаторных аллофонов.

**На выходе** фонетического процессора получается либо последовательность акустических параметров, либо формантные параметры, либо аллофонный текст. Это зависит от выбранного метода реализации синтезатора речи.

Полученная таким образом информация, т.е. просодически размеченный фонемный текст и параметры, сформированные фонетическим процессором, объединяются в одном процессоре, который называется акустическим процессором.

#### **92. Акустический процессор.**

**Акустический процессор**— электронное [устройство](https://ru.wikipedia.org/wiki/%D0%A3%D1%81%D1%82%D1%80%D0%BE%D0%B9%D1%81%D1%82%D0%B2%D0%BE) или программный комплекс, предназначенный для управления многокомпонентными звукоусилительными системами.

Прежде всего акустические процессоры (АП) характеризуются наличием определённого количества входов и выходов — это соотношение определяет применимость процессора в конкретных ситуациях. Существует множество конфигураций "входы/выходы" АП: 2/4, 2/6, 2/8, 4/8 и т.д. Разумеется, все входы АП оснащены аналого-цифровыми преобразователями, а выходы — цифро-аналоговыми.

Все функции АП организованы в блоки обработки. Важнейший из них — это [маршрутизатор](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%80%D1%88%D1%80%D1%83%D1%82%D0%B8%D0%B7%D0%B0%D1%82%D0%BE%D1%80), с помощью которого входы определённым образом назначаются пользователем на выходы, создавая таким образом определённую конфигурацию звуковой системы. Блоки обработки звуковых сигналов в АП:

* Б**лок регулировки уровней сигналов (Level)** - настроить уровень сигнала индивидуально для каждого входа и выхода.
* **Блок** [**задержки**](https://ru.wikipedia.org/w/index.php?title=%D0%97%D0%B0%D0%B4%D0%B5%D1%80%D0%B6%D0%BA%D0%B0_%D1%81%D0%B8%D0%B3%D0%BD%D0%B0%D0%BB%D0%B0&action=edit&redlink=1) позволяющий установить определённой длительности задержку индивидуально на каждый вход или выход. Это необходимо в случаях, когда компоненты звуковых систем находятся на разных расстояниях относительно слушателя и, таким образом, имеется возможность выровнять звуковой поток по времени. Соответственно, АП позволяют устанавливать задержку как в секундах, так и в метрах.
* **Блок** [**эквалайзеров**](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D0%B2%D0%B0%D0%BB%D0%B0%D0%B9%D0%B7%D0%B5%D1%80) **(EQ)**, позволяющий индивидуально настроить [частотную характеристику](https://ru.wikipedia.org/wiki/%D0%90%D0%A7%D0%A5) сигналов каждого входа и выхода. Как правило, в этих блоках применяются сложные параметрические эквалайзеры с большим количеством фильтров и их параметров. Помимо эквалайзеров как таковых, также имеются обычные обрезные фильтры.
* **Блок** [**кроссоверов**](https://ru.wikipedia.org/wiki/%D0%9A%D1%80%D0%BE%D1%81%D1%81%D0%BE%D0%B2%D0%B5%D1%80) **(X-Over)** находится только в секции выходов. Он позволяет выделить в сигнале каждого выхода определённую частотную полосу и корректно распределить сигнал в многополосных звукоусилительных системах в соответствии с параметрами её компонентов ([сабвуферов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B0%D0%B1%D0%B2%D1%83%D1%84%D0%B5%D1%80), [сателлитов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B0%D1%82%D0%B5%D0%BB%D0%BB%D0%B8%D1%82) и т.д.). Кроссоверы в АП, как правило, имеют большое количество параметров, таких как тип кроссовера, крутизна среза и т.д.
* **Блок** [**компрессора**](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%BE%D1%80_%D0%B0%D1%83%D0%B4%D0%B8%D0%BE%D1%81%D0%B8%D0%B3%D0%BD%D0%B0%D0%BB%D0%B0) **(Comp)**, позволяющий ограничить динамический диапазон сигналов каждого входа и выхода.

Подавляющее большинство акустических процессоров построено по типу древообразного меню, т.е. на его лицевой панели вы не найдёте регуляторы для каждого из параметров — все настройки осуществляются с помощью дисплея и кнопок выбора/ввода значений, плюс к этому нередко используются энкодеры. Однако для удобства некоторые функции всё же выносятся на лицевую панель, например, кнопки заглушения (Mute) каждого входа и выхода, а также кнопки быстрого доступа к наиболее часто используемым блокам обработки. Для наглядности все АП оснащаются различными индикаторами входных и выходных сигналов.

Исходя из сложной структуры акустических процессоров, очевидно, что они предназначены для работы только в профессиональных сферах, и, конечно же, пользователь должен обладать довольно обширными знаниями и опытом в области звукоусиления и обработки звука.

#### **93. Задача стемминга текстов естественного языка.**

Задача стемминга текстов естественного языка заключается в приведении слов к их основной (или базовой) форме, что позволяет учитывать различные грамматические формы слова как единый элемент. Это важная задача в обработке естественного языка, так как она помогает улучшить точность анализа текстов и извлечение смысла из них.

Простой стеммер ищет флективную форму в таблице поиска. Преимущества этого подхода заключается в его простоте, скорости, а также лёгкости обработки исключений. К недостаткам можно отнести то, что все флективные формы должны быть явно перечислены в таблице: новые или незнакомые слова не будут обрабатываться, даже если они являются правильными (например, iPads ~ iPad), а также проблемой является то, что таблица поиска может быть очень большой. Для языков с простой морфологией наподобие английского размеры таблиц небольшие, но для сильно флективных языков (например, турецкого) таблица может иметь сотни возможных флективных форм для каждого корня.

Флекти́вный строй (от лат. flectivus — гибкий) — устройство языка синтетического типа, при котором преобладает словоизменение при помощи флексий — формантов, сочетающих сразу несколько значений. Флективный строй противоположен агглютинативному, в котором каждый формант несёт только одно значение.

Таблицы поиска, используемые в стеммерах, как правило, генерируются в полуавтоматическом режиме. Например, для английского слова «run» автоматически будут сгенерированы формы «running», «runs», «runned» и «runly». Последние две формы являются допустимыми конструкциями, но они вряд ли появятся в обычном тексте на английском языке.

Алгоритмы усечения окончаний не используют справочной таблицы, которая состоит из флективных форм и отношений корня и формы. Вместо этого, как правило, хранится меньший список «правил», который используется алгоритмами, учитывая форму слова, чтобы найти его основу[3]. Некоторые примеры правил выглядят следующим образом: если слово оканчивается на 'ed', удалить 'ed'

если слово оканчивается на 'ing', удалить 'ing'

если слово оканчивается на 'ly', удалить 'ly'

Алгоритмы усечения окончаний гораздо эффективнее, чем алгоритмы полного перебора. Для разработки таких алгоритмов нужен программист, который достаточно хорошо разбирается в лингвистике, в частности морфологии, а также умеет кодировать «правила усечения». Алгоритмы усечения окончаний неэффективны для исключительных ситуаций (например, 'ran' и 'run'). Решения, полученные алгоритмами усечения окончаний, ограничиваются теми частями речи, которые имеют хорошо известные окончания и суффиксы с некоторыми исключениями. Это является серьёзным ограничением, так как не все части речи имеют хорошо сформулированный набор правил. Лемматизация пытается снять это ограничение. **Лемматиза́ция** — процесс приведения [словоформы](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%BE%D0%B2%D0%BE%D1%84%D0%BE%D1%80%D0%BC%D0%B0) к [**лемме**](https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BC%D0%BC%D0%B0_(%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)) — её нормальной (словарной) форме.

Алгоритмы усечения префикса также могут быть реализованы. Однако не во всех языках слова имеют приставки и суффиксы.

#### **94. Структура речевого сигнала. Спектрально-временные характеристики речевого сигнала.**

Речь – это сложное физическое явление. Она образуется в результате прохождения

воздушных потоков через речевой тракт человека. С акустической точки зрения звук – это колебательные движения среды. Сложность речи заключается в том, что это не один колебательный процесс. Речевой сигнал состоит из множества гармонических

составляющих, т.е. колебаний, которые периодически изменяются во времени по

синусоидальным законам (рис. 21) и описываются следующим уравнением:



Речевой сигнал, как и любой сложный сигнал, является результатом сложения нескольких гармонических составляющих, каждая из которых характеризуется своими значениями указанных параметров. Каждая такая составляющая называется гармоникой.

В результате разложения речевого сигнала на входящие в его состав гармонические

колебания получается частотный спектр – амплитудно-частотная зависимость. Например, если для сигнала частота и амплитуда первой гармоники равны F1 и А1 соответственно, а второй – F2 и А2, то получим амплитудно-частотную зависимость, представленную на рис.2.2.



Для более сложного сигнала амплитудно-частотная зависимость будет включать гораздо большее количество вертикальных отрезков с разными значениями частоты и амплитуды. В этом случае используется другое, более наглядное представление указанной зависимости в виде огибающей кривой линии.

Спектр речевого сигнала выражает его частотный состав, т.е. самыми главными

характеристиками речевого сигнала являются значения частот его гармонических

составляющих. Именно в результате наложения этих частот и их восприятия органами слуха мы получаем возможность слышать богато украшенный звук – речевой сигнал. Для разложения речевого сигнала в спектр, как правило, используется преобразование Фурье. На практике такое разложение речевого сигнала представляется в виде так называемых динамических спектрограмм (сонограмм).

#### **95. Синтез речевого сигнала. Методы синтеза: параметрические и компиляционные.**

Синтез речевого сигнала представляет собой процесс создания искусственного звукового потока, призванного имитировать человеческую речь. В настоящее время существует разнообразие методов синтеза речи, разработанных для повышения качества и натуральности сгенерированного звука.

Эти методы стремятся моделировать биологическую систему человека, основываясь на понимании процессов, происходящих в человеческом мозгу и речевом аппарате.

В настоящее время методы синтеза речевого сигнала подразделяются на два основных класса: параметрические и компиляционные (см. рисунок).

#### Параметрические методы:

* Характеристики:
  + Основаны на знаниях о процессах формирования речи у человека и параметрах, описывающих речевой сигнал.
  + Моделируют акустические процессы речеобразования, подражая речевому тракту человека.
* Два подхода к моделированию:
  + Артикуляторные процессы: строится геометрическая модель речевого тракта.
  + Акустические процессы: формируются наборы акустических параметров, таких как формантные параметры.
* Комбинированные методы:
  + Включают артикуляторно-формантный и фонемно-формантный подходы.
  + Формирование речевого сигнала основывается на двух указанных подходах.
* Трудности параметрических методов:
  + Затруднения в моделировании геометрии речевого тракта из-за ограниченных знаний об его строении.
  + Формантные модели чаще используются из-за более легкой измеримости и обобщаемости физических характеристик.

#### Компиляционные методы:

* Принцип:
  + Игнорируют физику процесса речеобразования.
  + Синтезируют речь из отрезков естественной речевой волны, компилируя их в речевой сигнал.
* Используемые отрезки:
  + Микроволны, аллофоны, дифоны и другие фрагменты речевой волны.
  + Комбинированные модели могут использовать аллофоны и дифоны.
* Достоинства и Недостатки:
  + Достоинства:
    - Естественный звук синтезированной речи.
    - Возможность "клонирования" голоса.
  + Недостатки:
    - Трудоемкость создания базы отрезков.
    - Требуются большие объемы памяти.

#### Сравнение методов:

* Параметрические методы:
  + Преимущества:
    - Простота реализации.
    - Высокое быстродействие.
    - Экономичность.
  + Недостатки:
    - Звук может звучать "роботно" из-за ограниченных знаний о физических особенностях речевого сигнала.
* Компиляционные методы:
  + Преимущества:
    - Естественный звук.
    - Возможность индивидуального "обучения" голоса.
  + Недостатки:
    - Трудоемкость создания базы отрезков.
    - Требования к объему памяти.

#### **96. Анализ речевого сигнала. Методов анализа: цифровой фильтрации, быстрых алгоритмов преобразования Фурье, Кепстральных коэффициентов.**

Анализ речевого сигнала является важной задачей в области обработки речи, и существует несколько методов, предназначенных для извлечения полезной информации из звуковых данных. Среди них выделяются методы цифровой фильтрации, быстрых алгоритмов преобразования Фурье и использование Кепстральных коэффициентов.

* Цифровая фильтрация:
  + Описание: Метод цифровой фильтрации включает применение различных фильтров к речевому сигналу для выделения определенных частотных компонент. Фильтры могут быть настроены на выделение основных характеристик речи, таких как форманты.
  + Применение: Цифровая фильтрация широко используется для улучшения качества речи, выделения особенностей речевого сигнала и подготовки данных для более сложных методов анализа.
* Быстрые алгоритмы преобразования Фурье (БПФ):
  + Описание: Преобразование Фурье используется для разложения речевого сигнала на частотные компоненты. Быстрые алгоритмы преобразования Фурье (например, алгоритм БПФ) позволяют эффективно вычислять преобразование Фурье для обработки звуковых данных.
  + Применение: Этот метод помогает выделить спектральные характеристики речевого сигнала, что полезно для распознавания речи и анализа его частотных компонент.
* Кепстральные коэффициенты:
  + Описание: Кепстральные коэффициенты представляют собой результат применения обратного преобразования к логарифму спектра мощности сигнала. Они отражают особенности формант в речи и являются важными признаками для распознавания речи.
  + Применение: Кепстральные коэффициенты широко используются в системах автоматического распознавания речи для извлечения ключевых характеристик речевого сигнала.

Эти методы анализа речевого сигнала играют важную роль в создании эффективных систем распознавания речи и обеспечивают основу для более сложных моделей машинного обучения, используемых в данной области. Они позволяют извлекать релевантную информацию из акустических данных, необходимую для понимания и интерпретации человеческой речи.