Вопросы по курсу “Компьютерная лингвистика и информационные технологии (2020)”

САМЫЙ ГЛАВНЫЙ ПРИЗНАК ДЛЯ СЛОВА: ФЛЕКТИВНЫЙ КЛАСС И ОКОНЧАНИЕ ОНИ ДАЮТ ПОЛНУЮ ГРАММАТИЧЕСКУЮ ИНФОРМАЦИЮ О СЛОВЕ ВНЕ КОНТЕКСТА

1. Предмет компьютерная лингвистика. Основные направления деятельности.

Компьютерная лингвистика - научная дисциплина, которая изучает естественный язык(текст) с целью формализации его структуры.

Мы пытаемся сформулировать правила работы с текстом.

Понятия — базовые строительные блоки языка и речи, на основе которых формируются единицы более высоких уровней.

Понятия в текстах выражаются **словами**, но чаще **словосочетаниями**. *Важно само понятие, а не слова и словосочетания.*

Смысл понятия проявляется полностью только через всю систему его отношений со всеми другими понятиями языка.

Предложение - вторая по значимости единица смысла. Из них формируются **сверхфразовые единства** (последовательности связного текста).

Сверхфразовое единство – фрагмент текста, объединенный неким смыслом. Предложение (Вася пил пиво).

Понятия – слова и словосочетания, которые являются ярлыками наших мыслительных образов.

Направления деятельности – все то, что связано с процессами анализа и синтеза: решение проблем формализации текстовой информации.

Лексема — основная форма слова, в которой заложен весь смысл всех слов, связанных этим смыслом, называемая *словоизменительная парадигма*

Словоизменительная парадигма — все слова (словоформы), которые изменяются по числам, родам, лицу и числу (для глаголов: переходность, валентность, возвратность).

2. Методы компьютерной лингвистики. Их краткая характеристика.

- **Статистические методы анализа** разноязычных репрезентативных корпусов текстов, позволяютщие выявить лексический состав национальных языков и соотнести его с параметром покрытия этой лексикой текстов на этом языке.
- **Принцип лингвистической аналогии** позволяет выявлять и реализовать трансформационные закономерности словоизменения и словообразования, многократно сократить объемы словарей и грамматических таблиц, а также успешно решить задачи, которые не поддающиеся решению алгоритмическими методами.

Методы компьютерной лингвистики – это те методы анализа текстовой информации, которые используются в рамках лингвистики.

• ***статистические методы*** - создание различного рода частотных словарей и анализ лексического состава и их статистических характеристик.
(изучение лексического состава: мы их изучаем по различным категориям, прежде всего по их морфологической сущности: к каким классам слов они относятся, какие формы встречаются – анализ на низшем уровне (анализ системы понятий); второй уровень смысловой значимости – предложения: ***синтаксический анализ*** целью которого является построение разных моделей и их статистические характеристики)
методы морфологического анализа изучает морфологические характеристики слов (работает со словами)
семантико-синтаксического анализ изучает синтаксические характеристики (работает с предложениями)
концептуальный анализ изучает смысловую структуру тестов в виде системы понятий и их отношений (работает с текстом)

• *лингвистические методы* (ниче не сказал)

3. Декларативные и программные средства. Их оптимальное соотношение.

Декларативные средства - та система словарей, грамматических таблиц и разных текстовых ресурсов, которая обеспечивают нам решение различных задач компьютерной лингвистики.

Программные средства – средства работы с конкретными элементами этих декларативных средств.
Оптимальное соотношение: (нихуя полезного не сказал, долго пиздел просто так…)
Выкладки:
Создаем различные словарные ресурсы и стараемся реализовать работу с ними путем простейших и быстродействующих процедур.
Требуют огромных затрат, но если использовать при их создании средства по автоматизации, то можно значительно их сократить.

4. Задачи лингвистического обеспечения систем обработки текстовой информации

Основные задачи – это все, что заключается в формализации текстового представления с целью их использования в различных технологических процессах, которые уже не требуют никакой формализации. То есть они работают с текстами как с числовой информацией.

Задачи, которые стоят перед системой обработки тестов: *система кластеризации, классификации, сематического поиска* и *машинного перевода* и т.д.

Для задачи кластеризации не требуется глубокого понимания смысловой структуры текста, достаточно выделить основные единицы смысла и их пересекать.

Задача машинного перевода требует колоссальных усилий по формализации теста и для обеспечения понимания системой смысловой структуры текста и смысла каждого понятия. Это обеспечивается системой двуязычных словарей. Причем работа с двуязычными словарями предполагает не только поиск, она еще предполагает некоторые интеллектуальные процедуры выбора этих понятий по следующим признакам:

1) наиболее длинные представления понятия по количеству слов является наиболее точным;

2) лучше выбирать эти длинные понятия не из общих словарей, а из тематических или пользовательских словарей.

5. Базовые процедуры систем обработки текстовой информации.

Сказал, что можно сказать кратко: **графематика, морфология, семантико-синтаксический анализ, концептуальный анализ.**

Мы работаем с очищенным текстовым представлением, из которого потом можно выделить его смысловые составляющие.

Есть задача выделения слов и выделения предложений.

- Процедура ***морфологического анализа*** - обеспечить обработку слов.
- Процедура ***семантико-синтаксического анализа*** - обеспечивает обработку предложений и строит их синтаксические структуры.
- Процедура ***концептуального анализа*** – выбирает систему понятий из текстов и устанавливает смысловые связи между этими понятиями.

6. Языковая единица (языковый знак). Мыслительный и акустический образ. Означаемое и означающее. Представление мыслительного образа в тексте.

Согласно утверждению основоположника традиционной лингвистики Ф. Соссюра - языковая единица (языковый знак) связывает не вещь и ее название, а понятие и акустический образ. Акустический образ является не материальным звучанием, а психическим отпечатком звучания, представлением, получаемым о нем посредством органов чувств. Таким образом, языковая единица (языковый знак) представляет собой ассоциацию двух психических сущностей: понятия и акустического образа. Первая из них получила название означаемое, вторая – означающее. Связь, соединяющая означающее с означаемым, произвольна (никак не мотивирована)

Есть человек, который содержит в своем сознании неизвестно в какой форме огромное количество мыслительных образов.

Есть средство коммуникации между людьми, в виде акустической коммуникации и в виде цифровой коммуникации, представленной в текстовом образе.

Акустическая информация – ярлыки, которые вытаскивают мыслительные образы.
Акустический образ – система взаимосвязанных понятий. Там есть и мыслительные образы, и связи.
Означаемое – то, что сидит в голове.
Означающее – то, что обозначают в написанном виде.
Представление мыслительного образа представлено в письменном тексте в виде слова или словосочетания. Причем в разных контекстных окружениях он может иметь разные текстовые представления (слова изменяются по числам, родам, падежам и тд, словосочетания тоже изменяются по своим законам).

7. Трансформация представления наименований понятий в текстах.

Понятия могут изменяться.

Помните пример «хромая рыжая собака», потом идет «собака» (родовое понятие), потом идет «животное» (тоже родовое понятие), потом идет «оно» (местоимение – понятие представляем в виде его анафоры).

Мы берем текст, делим его на предложения, в предложении находим вручную подлежащее, глагольную группу, дополнения и обстоятельства, ставим связи. И в виде элементарных смысловых триад помещаем в конечный результат, при этом каждое понятие приводим к его доминантной форме.

8. Парадигматические отношения между понятиями

Парадигматические отношения между понятиями – внеконтестные отношения между понятиями – отношения типа род – вид, часть – целое, ассоциация.
(как бы вертикальная связь)

9. Синтагматические отношения между понятиями

Синтагматические отношения между понятиями - те отношения, которые получаются или устанавливаются в рамках осмысленных предложений (Вася пил - есть связь, пил пиво – есть связь).

Они выстраиваются между именными понятиями с помощью глагольных связок или с помощью предлогов, а также бывают чисто синтаксические связи, когда нет ничего, и анафарические связи.
(как бы горизонтальная связь)

Синтагматические отношения – отношения между словами в словосочетаниях или между членами предложения – горизонтальны. Это отношения сочетаемости между последовательно расположенными языковыми единицами одного уровня, которые проявляются в том, что употребление одной единицы разрешает, требует или запрещает употребление связанной с ней другой единицы одного с ней уровня.

Синтагматическая связь объединяет слова в высказывания.

Учет СО позволяет оценить точность поиска информации.

Предикативная синтагма – это связь между подлежащим и сказуемым.

Непредикативные синтагмы осуществляют в предложении функцию пояснения, дополнения и уточнения главных членов предложения.

10. Иерархия единиц смысла языка и речи. Их краткая характеристика.

Единицы смысла в языке и речи - Основными единицами смысла в естественных языках являются понятия, предложения и сверхфразовые единства (связный текст). ***Понятия*** являются минимальными (базовыми) единицами смысла. Из наименований понятий составляются ***предложения***, которые имеют предикативную структуру, т.е. в них указываются признаки объектов и (или) отношения между ними. ***Сверхфразовые единства*** представляют собой последовательности предложений, объединенные общим смыслом. В человеческом сознании понятия представляют классы объектов, а в целом система понятий каждого языка является системой категоризации действительности. При этом важно подчеркнуть, что разные языки могут иметь разные системы категоризации действительности, т. е. могут отличаться друг от друга составом используемых в них понятий.

В естественных языках и в речи основными единицами смысла являются морфемы(корни слов, суффиксы и префиксы), слова, словосочетания, фразы и различного рода сверхфразовые единства. Эти единицы в совокупности представляют собой иерархическую систему, в которой *смысловое содержание единиц более высокого уровня не сводимо или не полностью сводимо к смысловому содержанию составляющих их единиц более низкого уровня* (смысл единиц более высокого уровня не может быть “вычислен” на основе информации о смысле единиц более низкого уровня и о информации о связях между этими единицами). ***Минимальной единицей, обозначающей понятие, является слово, но большинство понятий обозначается устойчивыми словосочетаниями и фразами.***

11. Определение наименование понятия как социально значимого мыслительного образа.

Понятие (концепт) - это социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания. Под устойчивыми фразеологическими словосочетаниями он понимает не только идиоматические выражения, но и любые повторяющиеся отрезки связных текстов длиной от двух до десяти-пятнадцати слов (более длинные устойчивые словосочетания встречаются редко). В развитых языках мира (русском, английском, немецком, французском и др.) количество различных наименований понятий достигает нескольких сотен миллионов. Большинство из них обозначаются словосочетаниями, смысл которых не сводим к смыслу составляющих их слов. Слова, входящие в состав словосочетаний, обозначают лишь некоторые признаки понятий, позволяющие отличать их друг от друга, но не исчерпывающих их содержания. Содержание понятий в полном объеме интерпретируется только в “душе” человека - в его внутреннем мире, где “все связано со всем”.

12. Наименование понятия как единица смысла. Ее характеристика

Это прежде всего слово или словосочетание, которое обозначает реальный объект объективной реальности. Несет четкий смысловой образ.
Сущность и наименование понятие – это примерно одно и то же, то, за чем стоит некий мыслительный образ.

13. Предложение - высказывание как единица смысла. Ее характеристика

Высказывание — **это** грамматически правильное повествовательное **предложение** определенного языка, которое выражает некоторый смысл и является либо истинным, либо ложным, но не тем и другим сразу.

14. Сверхфразовое единство как единица смысла. Ее характеристика

В текстах это в какой-то степени абзац.

Некий мыслительный образ, оформленный или представленный в виде нескольких контактно расположенных предложений.

Сверхфразовые единства представляют собой последовательности предложений, объединенные общим смыслом. В человеческом сознании понятия представляют классы объектов, а в целом система понятий каждого языка является системой категоризации действительности. При этом важно подчеркнуть, что разные языки могут иметь разные системы категоризации действительности, т. е. могут отличаться друг от друга составом используемых в них понятий.

15. Смысловое представление текста.

Смысловое представление текста – представление текста в виде семантической сети.
То есть есть система понятий и их система отношений.

16. Анафорические связи в текстах.

Те связи между наименованием понятия и его антецедентом. Например, связь местоимение – слово.

17. Система флективных классов как базовая классификация слов русского языка.

Разработана для классификации ***значимых*** и ***незначимы****х* слов: тех слов, которых в языке большее количество.
Белоногов разделил все слова на 2 категории:

- Слова, которые подчиняются стандартным законам словоизменения и словообразования;

- которые чем-то отличаются от этой системы (старые короткие слова, которые появились в процессе эволюции языка), все знаки препинания, предлоги, союзы, местоимения.

Разделил все слова по типам их словоизменения:

Существительные: по роду, по одушевленности.

Прилагательные: по роду и числу.

Глаголы: инфинитив, личной формы (изменяется только по лицу и числу), прошедшего времени (изменяется по роду и числу), краткое причастие.

4.1 Идея

Рас квалифицировать слова по их типу и видоизменению.

4.2 Состав

?140 разных классов, отличающихся окончаниями и разными грамматическими классами.

4.3 Исходные данные

Слова из словаря.

4.4 Выходные данные

Классифицированные слова по флективным классам.

18. Грамматические классы слов русского языка, их характеристика.
Грамматический класс слова (Мнемоническое обозначение)

№	*Классы и группы слов*	*Мнемоническое обозначение*
1	Существительное 1-102,145,146,147,150	N
2	Прилагательное 103-115,151	A
3	Сравнительная степень прилагательного 152	C
4	Глагол личной формы 116-124	V
5	Глагол прошедшего времени 125	L
6	Краткое причастие 126-130	K
7	Модальный глагол 143	M
8	Инфинитив 144	I
9	Субстантивированное прилагательное 103-110	A->N
10	Местоимение личное (в им. падеже) -	i
11	Местоимение личное (в косв. падеже) -	y
12	Числительное количественное	0 (Нуль)
13	Числительное порядковое 134-135	O (Лат.символ)
14	Союз сочинительный 153	&
15	Союз подчинительный 153	b
16	Междометие 154	J
17	Наречие 152	Y
18	Предлог 155-164	F
19	Причастие 103-110	W

Грамматический класс слова (Мнемоническое обозначение)
Слова, первым символом которых является заглавная буква, а все последующие - прописные B
Слова, состоящие только из заглавных букв A
Слова, состоящие только из прописных букв b
Последовательности цифровых символов 0
Слова состоящие из одной заглавной буквы I
Все знаки препинания приводятся в исходной форме . , - ! ?( ) % № « » @ # $ ^ & *

19. Грамматическая обобщенная синтагма как инструмент представления формы слова

Форма слова – две категории: флективный класс и грамматическое окончание.
Обобщенная синтагма - инструмент

20. Состав грамматических признаков, присущих отдельным формам слов.

Например, глагольность есть только у глаголов, одушевленность, омонимия, возвратность(нет у существительного), супплетивная форма(только у коротких слов, пример: ребенок - дети), субстантированное прилагательное (например: больной), признак типа слова(имя, местоименность, числительность), типа основы(каноническая, вариантная, например: банка - банок (вариативная основа)), нормализующее окончание, род, число и тд.

21. Состав грамматических признаков, присущих словоизменительной парадигме слова.

Словоизменительные категории - это категории,члены которых могут быть представлены формами одного и того же слова.
Словоформы могут отличаться друг от друга только тем элементом значений, который представлен морфологической категорией : белый - белая

- категория падежа и числа у существительных
- категория падежа числительных
- категория падежа, рода, числа и степени сравнения прилагательного
- категория числа, лица, времени, наклонения и рода глагола
- категория степени сравнения наречия

22. Базовый (усеченный) набор грамматический признак слов. Автоматизация их назначения.

длина окончания, флективный класс, тип словоформы, регистр слова.

Цифры грамматических признаков (ГП) слова имеют следующие цифровые обозначения:
· Род: 1 - мужской, 2 - женский , 3 – средний;
· Число: 1 - единственное, 2 – множественное;
· Падеж: 1 - именительный, 2 - родительный, 3 - дательный, 4 - винительный, 5 - творительный, 6 – предложный;
· Лицо: 1 - 1-ое лицо, 2 - 2-ое лицо , 3 - 3-ье лицо.

23. Полный набор грамматический признак русских слов

sE_FrmSen представляет собой последовательность символов лексико-грамматических классов, входящих в состав анализируемого предложения (N - существительное, A - прилагательное, F - предлог и т.д.).

sE_SklSen представляет собой последовательность символов, обозначающих последовательность членов предложения (S - подлежащее, P - сказуемое и т.д.).

Se_AtrSen указывает на наличие признаков глагольности у слов с соответствующими номерами.

sE_BndSen обозначает границы словосочетаний, где “(“ - обозначает начало словосочетания, “)” - его конец, “|” - обозначает, что граница начала и конца совпадает.

sE_Lnk1Sen и sE_Lnk2Sen отражают синтаксические связи между словами в предложении: для каждого слова («слуги») указывается его управляющее слово («хозяин»).

sE_SimplSent обозначает границы простого предложения. Символами B и E обозначены начало и конец сложного предложения, символами b и e - начало и конец простого предложения.

sE_OdnMember указывает на наличие в предложении однородных членов, которые обозначаются одинаковыми символами.

sE_BigCharSen обозначает написание слов строчными или заглавными буквами. С помощью этого признака можно установить, является ли слово аббревиатурой, именем собственным и т. д.

Далее приводятся результаты назначения словам однозначной грамматической информации в соответствии с их контекстным окружением.

sE_GiGndRes указывает на информацию о роде слова (1 - мужской, 2 - женский , 3 - средний).

sE_GiGndRes указывает на информацию о числе слова (1 - единственное, 2 - множественное).

sE_GiGndRes указывает на информацию о падеже слова (1 - именительный, 2 - родительный, 3 - дательный, 4 - винительный, 5 - творительный, 6 - предложный).

sE_GiGndRes указывает на информацию о лице данного слова (1 - 1-ое лицо, 2 - 2-ое лицо , 3 - 3-ье лицо).

24. Частотные словари как инструмент исследования статистических характеристик лексического состава текстов

Частотный словарь – словарь, который мы создаем по сверхбольшим текстам

Словоформы – какие-то формы слов какой-то парадигмы (парадигмы - то же, что и нормализованные формы слов)

Система словарей включает в себя разные формы словарей:

• Словарь супплетивных форм (словарь S) – те слова, которые встречаются чаще всего, имеют весь набор грамматической информации и позволяют, не обращаясь ни в какие другие словари, сразу да(словарь K)вать на вход

• Словарь с ограниченным набором грамматических признаков (то есть с длинной окончаний и с флективным классом) -

#TW – тип слова (имя, местоименность, числительность)
#TD – тип словаря (здесь К)

Почему взяли ограниченный набор грамматической информации? Потому что остальной набор грамматической информации можно легко вычислить трансформацией из усеченного набора в полный набор с помощью таблицы Т: если знаем флективный класс и набор грамматической информации

• Словарь концов (чтобы покрыть остаток, который не покрывается) - (словарь Е)
• Словарь семантических форм слов - (словарь С)

25. Машинная грамматика. Назначение и задачи

Нужна для получения некой формализованной информации о словах

1. Автоматизация составления и лингвистической обработки машинных словарей;
2. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ;
3. Автоматическое индексирование документов и информационных запросов;
4. Автоматическая классификация и реферирование документов;
5. Лингвистическое обеспечение процессов поиска информации в одноязычных и многоязычных базах данных;
6. Машинный перевод текстов с одних естественных языков на другие;
7. Построение лингвистических процессоров, обеспечивающих общение пользователей с автоматизированными интеллектуальными информационными системами (в частности, с экспертными системами) на естественном языке, или на языке, близком к естественному;
8. Извлечение фактографической информации из неформализованных текстов.

МГ - машинная грамматика
ГХ - грамматическая характеристика

Базовая процедура обработки текстовой информации — ***машинная грамматика*** **(МГ).**

Машинная грамматика - комплекс формальных правил, процедур и декларативных средств, обеспечивающих автоматическое преобразование текстовых представлений слов в формальное описание модели в виде совокупности грамматических и семантических характеристик.

Машинная грамматика нужна для автоматического установления морфологической структуры слов и определения полного набора их ***грамматических характеристик (ГХ)***. Для **машинной грамматики** предъявляется основное требование: правильность назначения грамматических характеристик для всей совокупности слов русского языка + высокое быстродействие.

Машинные грамматики построены на подходах, основанных на использовании *словарей слов или грамматических характеристик.* Этот подход требует больших начальных затрат на создание словаря большого объема. Подход порождает проблему: как характеризовать слова, что не вошли в словарь.

26. Графематика. Назначение и задачи.
Нужна, чтобы разделить текст на слова и предложения

Графематический анализ – первичный этап в процессе автоматической обработки текстов на естественном языке. Основной задачей графематического анализа является выделение структурных единиц из входного текста, а именно предложений абзацев, слов, знаков препинания и т. д.

Входной текст может иметь как линейную структуру, содержащую единый фрагмент текста, так и нелинейную. В этом случае текст содержит различные структурные единицы: основной текст, заголовки, примечания, комментарии и т. д. Выходными данными этапа графематического анализа является графематическая таблица.

27. Краткое представление формы понятия

Представление в виде обобщенной синтагмы
8М: 8 – флективный класс, М- окончание

28. Краткое представление смысла наименования понятия

Смысл понятия представляется в виде его нормализованного представления. Оно может быть нормой словоизменения и словообразования.

Словоизменение – когда приводим все грамматические окончания к одному унифицированной

Словообразование – берем все суффиксы, которые есть у словообразовательной парадигмы и приводим к унифицированному представлению (Унифицированное представление словообразовательной основы для глагольных форм – инфинитив, а для именных форм – существительное)

Пример: Коллектив завода и заводской коллектив (Без словообразовательного анализа не решим, что это одно и то же, потому что завод в одном грамматическом классе –сущ, а заводской в другом – прилаг.

29. Морфологический анализ. Назначение и задачи.

Нужен для назначения грамматических характеристик слова и определения структуры в виде основы и грамматического окончания.

Морфологический анализ слов применяется с целью их членения на морфемы или сочетания морфем и получения грамматической информации, необходимой на последующих этапах обработки текстов.

***Задача морфологического анализа** – определение нормальной формы, от которой была образована данная словоформа, а также получение набора её морфологических характеристик.*

Это делается для того, чтобы в последующих этапах анализа использовать только нормальную форму слова, а не все его словоформы и использовать морфологические характеристики для проверки согласованности слов.

Морфологические характеристики – это набор пар «ключ, значение». В роли ключа выступает, например, род, число, падеж, склонение, время и другие признаки слов, используемые в русском языке. Значением является какое-либо конкретное значение, которое может принимать данный признак (ключ).

30. Морфологический анализ первого поколения. Функциональные требования

словарь K и S и таблица преобразований T
Проверка результатов обработки слов русского языка процедурой морфологического анализа разрабатываемого ЛП выполняется с целью установления вероятности

правильного назначение системы грамматических, синтаксических характеристик в текущей версии МА. Такую проверку возможно выполнить путем экспертной оценки обработанных корпусов текстов опосредствованными методами, на основе количественных и качественных оценок поэтапной обработки слов, разделенных на различные категории. Основанием для такого деления могут быть статистические данные о покрытии текстов словоформами, включенным в состав декларативных средств (анализ частотных словарей) и путем установления категорий словоформ, имеющих совпадающие наборы формальных характеристик словоформ, но входящих в различные по степени точности назначения набора характеристик словоформ словари.

Для установления вероятности правильного назначение системы грамматических, синтаксических характеристик в текущей версии МА необходимо разделить все словарные статьи системы словарей (S, K, E, C) на категории словоформ, имеющих совпадающие наборы формальных характеристик словоформ, но входящих в различные по степени точности назначения набора характеристик словоформ словари. Для решения этой задачи было предложено выполнить категоризации всех словарных статей комплекса по двум основаниям:

1. По базовым характеристикам словоформ - флективному и грамматическому классам.

2. По вхождению словоформ в состав конкретного словаря (S, K, E, C).

Результаты такой категоризации словоформ корпуса текстов приведены н в табл. 2.

Количественные характеристики вероятности правильного назначения формальных характеристик словоформ комплекса словарей по их флективным классам и вхождению в состав словарей

Флективный класс	Словарь S		Словарь K		Словарь E		Итого
	Кол.сло - воформ (в тыс.)	Вероятность назначения ГХ	Кол.сло - воформ (в тыс.)	Вероятность назначения ФХ	Кол.сло - воформ (в тыс.)	Вероятность назначения ГХ	Кол.сло - воформ (в тыс.)
001	2.7	98%	3.4	85%	8.4	75%	19.3
002	1.2	98%	3.4	85%	5.6	75%	20.0

31. Состав и назначение словарей МА первого поколения.
3 словаря:
1) коротких слов S
2) конечных буквосочетаний K
3) таблица преобразование из усеченных форм в полную форму T

32. Схема обработки слов системой словарей МА первого поколения.
из одного словаря в другой (посмотреть)

Сразу в словаре К с усеченной формой, потом Е, потом Т, выход
Смотрим в словаре исключения, если нашли выходим, если не нашли берем словарь конечных буквосочетаний К,

33. Нормализация слов. Назначение

Нормализация (лемматизация) - это лингвистическая процедура, реализующая процесс трансформации исходного слова в его нормализованную (каноническую) форму. Обычно под нормализованной формой слова понимается та его форма, которая традиционно указывается в словарях. Например, для ***существительного*** это форма именительного падежа единственного или (в случае pluralia tantum) *множественного числа*, для ***глагола*** – форма инфинитива, для ***прилагательного*** – форма именительного падежа единственного числа мужского рода.

В рамках используемой концепции необходимо различать два уровня нормализации: на уровне словоизменения и на уровне словообразования. При нормализации слов на словоизменительном уровне каноническая форма слова должна представлять всю его словоизменительную парадигму. При нормализации слов на словообразовательном уровне каноническая форма слова должна представлять всю его словообразовательную парадигму.

34. Нормализация словосочетаний. Назначение.

Процесс трансформации исходного словосочетания в нормализованную форму, где все слова приводятся к канонической форме и располагаются в определённом порядке.

Нормализация словосочетаний - это процесс приведения словосочетания к нормальной форме со строго определенным порядком слов и их грамматической формой.

Вот что-то про порядок понятий:

1. Все слова исходного наименования понятия приводятся к нижнему регистру.

2. Опорное слово должно находиться на первом слева месте и быть нормализовано на уровне словоизменения.

3. Все зависимые слова должны быть нормализованы на уровне словообразования и расположены в лексикографическом порядке.

Из состава формализованного представления должны быть исключены все знаки препинания: запятые, тире, кавычки, двоеточия, восклицательный и вопросительный знаки и др. В процессе реализации второго варианта унификации наименования понятия вначале определяется его синтаксическая структура и определяется опорное слово. Далее это слово нормализуется на уровне словоизменения и помещается на первое место слева в структуру генерируемого формализованного представления. Все зависимые слова нормализуются на уровне словообразования. Нормализация слов на уровне словообразования выполнялась в соответствии с методами и алгоритмами, изложенными в работе [10]. После этого эти зависимые слова сортируются в лексикографическом порядке и в том же порядке помещаются слова в структуру формализованного представления наименования понятия после опорного слова.

35. Унифицированное представление наименований понятий. Назначение.

Наименования понятий, состоящие из отдельных слов, представляют собой, как правило, существительные (N), и значительно реже глаголы (I) или наречия (Y). Наименования понятий, выражающиеся словосочетаниями, представляют собой сочетания двух или нескольких самостоятельных слов, связанных друг с другом по смыслу или грамматически. Эти словосочетания могут включать в свой состав существительные (N), прилагательные (A), предлоги (F), глаголы в форме инфинитива (I), глаголы личной формы (V), глаголы прошедшего времени (L), краткие причастия (K), и деепричастия (D).

36. Морфологический анализ второго поколения. Функциональные требования

Общие Требования:

• Высокое быстродействие, реализуемое на основе оптимальной архитектуры комплекса словарей и грам.табл. и рацион. схемы обработки анализируемых словоформ (то есть 90% потока слов идут по очень быстрой схеме обработки)

• Высокое качество обработки (в точных словарях сидит вся точная информация, которую мы многократно проверяем разными способами)

• Возможность определения полного набора грамматич. и синтаксич. характеристик анализируемого слова (определили полный набор характеристик, знаем как их можно автоматически назначить, базовые признаки – флективный класс и окончание)

• Высокая степень автоматизации при создании и ведения комплексноа словарей и грамматич. таблиц

Частные требования:

• Базирование на системе флективных классов, чтобы обеспечить возможность автоматизации

• Разработанные программные средства для исследования лексического состава и частотных характеристик

• Установленный лексический состав словарей, их структура и определена архитектура программной реализации

• Набор грамматических и семантических характеристик

Основная идея МА в том, что многоуровневая система анализа слов должна обеспечить возможность разделения потоков анализируемых словоформ по различным схемам их обработки.

1) Основной поток слов (85-92%) должен быть обработан по самой короткой и быстродействующей схеме с помощью словаря S. Он должен содержать 40-50 тыс. словоформ с полным набором грамматической информации

2) Меньший поток слов (7-12%) должен быть обработан по менее быстродействующей схеме с помощью словаря K. Он должен достигать 100-120 тыс. словоформ с усеченным набором грамматической информации. Обогащаем набор ГИ полной ГИ по таблице T, нормализуем по таблице N. Назначается набор семантических признаков с помощью таблицы C.

3) И совсем небольшой поток словоформ (0.5-3%) должен быть обработан по последней наиболее тяжелой схеме. Обратный поиск в словаре Е на наибольшее совпадение конечного буквосочетания словоформы с одним из элементов словаря (100-350 тыс)

37. Реализация требования быстродействия МА второго поколения

Эту задачу удалось достигнуть путем трансформации словарей основ слов большого объема (свыше 400 тыс. основ слов) в две небольшие грамматические таблицы – таблицу конечных буквосочетаний слов (словарь КБС), предназначенную для обработки слов по методу аналогии и таблицу, включающую все “служебные” и короткие слова (словарь СКС), которые не включены в первую таблицу [6,11] .

Использование этих таблиц позволило существенно упростить алгоритм морфологического анализа и обеспечить высокую скорость работы. Обработка слов этим алгоритмом выполняется следующим образом: вначале производится поиск анализируемого слова в таблице “служебных” и коротких слов и, если оно там находится, ему назначается соответствующий набор грамматических признаков с помощью таблицы соответствия

флективного класса слова (ФК), его окончания (ОК) и набора грамматической информации (ГИ). Если это слово не было обнаружено в словаре СКС, то производится поиск в таблице КБ. После этого слову также назначается набор грамматической информации по словарю СФКГИ. Ниже приведен алгоритм морфологического анализа слов русского языка.

38. Реализация требования высокого качества обработки лексического состава МА второго поколения

Основные требования:

1) повышение быстродействия функционирования анализатора не менее чем на 50%;

2) повышение качества декларативных средств анализатора (увеличение совокупности грамматических и семантических характеристик слов до 30 элементов) и значительное увеличение их покрывающей способности;

3) сокращение трудозатрат на разработку декларативных средств и программного обеспечения анализатора не менее чем на 50%.

Пути реализации требований:

1. Повышение быстродействия анализатора можно обеспечить включением в его состав дополнительной быстродействующей процедуры, обрабатывающей основной поток (не менее 80%) текстовых словоформ. Такая процедура должна использовать словарь, содержащий полный набор грамматических и семантических характеристик наиболее часто употребляемых форм слов русского языка.

2. Повышение качества декларативных средств можно обеспечить путем расширения спектра грамматических (формообразующих и словообразующих) и семантических характеристик. Автоматизированное назначение этих характеристик как словоформам, так и нормальным формам слов должно обеспечиваться контролируемыми технологическими процессами создания и ведения словарей. Формат словарных статей должен быть реализован в виде списковой структуры (название признака – значение признака).

3. Повышение покрывающей способности словарей можно обеспечить путем включения в состав словарного комплекса словарей словоизменительных парадигм слов, представлять которые будут их нормальные формы.

4. Сокращение трудозатрат на разработку программных средств и повышения их быстродействия можно решить путем разработки относительно простых алгоритмов с небольшой вычислительной сложностью.

39. Реализация требования максимального покрытия лексического состава текстов МА второго поколения

40. Состав и назначение словарей МА второго поколения.

Словарь S — Прямой словарь, в котором собрана самая частая лексика с назначенным ***полным*** набором грамматических и семантических характеристик (небольшой словарь, зато полная информация). Обрабатывает высокочастотный поток текстовых словоформ, включает в себя все служебные слова и наиболее встречающиеся формы слов.

Если слово попало в словарь S, оно получило полный набор характеристик, большие никуда обращаться не надо.

Словарь K — в нем нет полного набора грамматических и семантических характеристик. Он дает ***усеченный*** набор, но при всем этом он обеспечивает полный цикл обработки слов. Предназначен для обработки остального потока текстовых словоформ, которые не попали в словарь S.

Словарь E — обратный словарь, обеспечивает назначение ***усеченного*** набора грамматических характеристик словоформ. Работает только со словоформами, относящимися к регулярной трансформационной системе словоизменения, **не *покрытой словарями S и K.***

Словарь C — обеспечивает назначение семантических признаков всем членам словоизменительных парадигм слов. Для каждой нормальной формы назначен набор семантических характеристик: ***одушевленность*** и **дополнительный грамматический класс #DK**.

41. Схема обработки слов системой словарей МА второго поколения.
Словарь S, вроде скинет схему

Словарь S —Прямой словарь, в котором собрана самая частая лексика с назначенным ***полным*** набором грамматических и семантических характеристик (небольшой словарь, зато полная информация). Обрабатывает высокочастотный поток текстовых словоформ, включает в себя все служебные слова и наиболее встречающиеся формы слов.

Если слово попало в словарь S, оно получило полный набор характеристик, большие никуда обращаться не надо.

Словарь K — в нем нет полного набора грамматических и семантических характеристик. Он дает ***усеченный*** набор, но при всем этом он обеспечивает полный цикл обработки слов. Предназначен для обработки остального потока текстовых словоформ, которые не попали в словарь S.

Словарь E — обратный словарь, обеспечивает назначение ***усеченного*** набора грамматических характеристик словоформ. Работает только со словоформами, относящимися к регулярной трансформационной системе словоизменения, **не *покрытой словарями S и K.***

Словарь C — обеспечивает назначение семантических признаков всем членам словоизменительных парадигм слов. Для каждой нормальной формы назначен набор семантических характеристик: ***одушевленность*** и **дополнительный грамматический класс #DK**.

В начале производится поиск анализируемой формы слова в словаре СКС и, если она там находится, ей назначается базовый набор грамматических характеристик.

Если эта словоформа не была обнаружено в словаре СКС, то производится инверсия ее буквенного состава и выполняется поиск на наибольшее совпадение ее конечного буквосочетания с буквосочетанием одного из элементов таблицы КБС.

После нахождения подходящего буквосочетания анализируемой словоформе назначаются грамматические характеристиками. Недостающая грамматическая информации устанавливается по таблице ФКГИ.

42. Характеристика и назначений словаря S МА второго поколения

Словарь S обеспечивает ***назначение полного набора грамматических и семантических характеристик наиболее частотным словоформам.***

Словарь S предназначен для обработки основного высокочастотного потока текстовых словоформ. Этот словарь включает в свой состав все служебные слова и наиболее часто встречающиеся формы слов. Словарь обеспечивает назначение набора грамматических и семантических признаков, включающего более 30 возможных характеристик.

43. Характеристика и назначений словаря K МА второго поколения

Словарь K обеспечивает ***назначение усеченного набора грамматических характеристик словоформ с аномальной трансформационной системой словоизменения.***

Словарь K предназначен для обработки остального потока текстовых словоформ. Этот словарь включает в свой состав формы слов, относящиеся к аномальной трансформационной системе словоизменения и словообразования, а также часто встречающиеся формы слов русского языка. Словарь обеспечивает назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик.

44. Характеристика и назначений словаря E MA второго поколения

Словарь E обеспечивает назначение усеченного набора грамматических характеристик словоформам с регулярной трансформационной системой словоизменения.

Словарь E предназначен для обработки словоформ, относящиеся к регулярной трансформационной системе словоизменения и словообразования слов русского языка, не покрытых лексикой словарей S и K. Этот словарь включает в свой состав конечные буквосочетания форм слов. Словарь обеспечивает назначение усеченного набора грамматических признаков, состоящего из пяти возможных характеристик.

45. Характеристика и назначений таблицы T MA второго поколения

Таблица T обеспечивает преобразование усеченного набора грамматических характеристик в полный набор грамматических характеристик.

Таблица T предназначена для преобразования усеченного набора грамматических признаков словоформ в его полный состав. Таблица обеспечивает назначение полого набора формообразующих характеристик, включающего более восьми возможных характеристик. Формат таблицы – FMA_t08.

Фрагмент таблицы T

#TO=+#FK=011	#GI=*1110*1140#OS=IA#SU=t#GK=N#TD=T
#TO=+#FK=014	#GI=*1110*1140#OS=LA#SU=t#GK=N#TD=T
#TO=+#FK=015	#GI=*1110*1140*1220#OS=MA#SU=t#GK=N#TD=T
#TO=+#FK=016	#GI=*1110*1140#OS=NA#SU=t#GK=N#TD=T
#TO=+#FK=017	#GI=*1110*1140*1220#OS=OA#SU=t#GK=N#TD=T

46. Характеристика и назначений таблицы N MA второго поколения

Таблица N обеспечивает преобразование текстовой словоформы в ее нормальную форму.

Таблица N предназначена для преобразования текстовой формы слова в его нормальную форму и содержит нормализующие окончания слов. Таблица включает следующие грамматические признаки: номер флективного класса словоформы и нормализующее окончание, соответствующее номеру флективного класса.

Фрагмент таблицы N

#FK=107	#NO=ой
#FK=110	#NO=ой
#FK=111	#NO=ий
#FK=112	#NO=+
#FK=113	#NO=й

47. Существительные -сущности и существительные – не сущности

Сущности – наименования понятий, которые имеют под собой мыслительный образ, не сущности – не являются отображением мыслительного образа

48. Состав грамматических и семантических признаков существительного

В качестве базовых признаков, на основе которых возможно идентифицировать каждую текстовую словоформу и автоматически построить ее грамматическую парадигму

были выбраны номер флективного класса словоформы и ее грамматическое окончание. При этом мы исходили из того, что флективный класс словоформы однозначно соотносится с типом ее словоизменения, кроме того этот класс также соотнесен с грамматическим классом словоформы, а также он содержит информацию о роде и «одушевленности» для грамматического класса «существительные».

При этом сочетание флективного класса слова и ее грамматического окончания позволяет однозначно установить расширенный состав грамматических характеристик конкретной словоформы в ее контекстном окружении.

Род	Семантические показатели	Грамматические показатели
Мужской род	Лицо мужского пола	Конечный твердый согласный, нулевое окончание стол. Слово путь
Женский род	Лицо женского пола	Окончание -а,-я, (жена, семья – 1 склонение), мягкий конечный согласный и окончание -и в Р.п. (жизнь – жизни – 3 склонение)
Средний род	Неживые предметы, явления (неодушевленные существительные)	Окончания -о, -е, -я (село, поле, дитя) и существительные на –мя (бремя, племя, время)
Общего рода	Лица обоих полов	Окончания -а, -я (1 склонение), обладают сильной эмоциональной оценкой (плакса, задира, зубрила)
Двуродовые	Лица обоих полов	Мягкий конечный согласный и нулевое окончание (бездарь, подлец, заморыш, молодец)
Не имеют рода	___	Слова, которые употребляются только во множественном числе (сани, ножницы, брюки)
Изменили род	___	Был: м.р.: вуаль, дуэль, клавиш. ж.р.: табель, погона, рояль, санатория, рельс, тополь Стал: м.р.: табель, рояль, санаторий, рельс, тополь, погон ж.р.: вуаль, дуэль, клавиша
Не имеют устойчивой родовой характеристики в русском языке около 10 слов:		проток – протока, манжет – манжета, ставень – ставня и др.

49. Состав грамматических и семантических признаков прилагательного

грамматические

- общее грамматическое значение - признак предмета
- постоянные морфологические признаки - качественное, относительное, притяжательное

- непостоянные морфологические признаки - степень сравнения(у качественных),краткая или полная форма(у качественных), число, род(в единственном числе), падеж(полные прилагательные)
- типичная синтаксическая роль- определение (полные) или именная часть составного именного сказуемого (краткие и полные)

семантической основой прилагательного является обозначение качества, признака, принадлежности предмета как относительно постоянное свойство

50. Состав грамматических и семантических признаков наречий
Грамматическая неизменяемость отличает наречие среди всех частей речи русского языка У него нет падежных форм, форм единственного и множественного числа, как у имен существительных или прилагательных, личных форм, как у глаголов
.
Наречие не склоняется и не спрягается.Только наречия с суффиксом -о, -е, которые образованы от качественных прилагательных, могут иметь степени сравнения: сравнительную и превосходную.

- громко — говори громч е , более (менее) громко; громче всех;
- твёрдо — стало твёрж е , более (менее) твёрдо; наименее твёрдо.

По функции наречия делятся на знаменательные и местоименные.
Знаменательные местоимения обозначают признак действия, предмета и другого признака:
- взволнованно дышать;
- рубашка навыпуск;
- слегка поблёкший.

Местоименные наречия только указывают на признак, конкретно его не обозначая:
- куда положить;
- как-то невдомёк;
- везде знают.

По лексическому значению знаменательные наречия делятся на две большие группы:
- определительные (обозначают признак);
- обстоятельственные (обозначают условия действия, состояния).

Определительные наречия имеют смысловые разряды:
- наречия образа действия (шагать быстро);
- наречия меры и степени (слишком отчаянный);
- сравнительно-уподобительные наречия (плутать по-заячьи);
- наречия со значением совместности (играть вдвоём).

Обстоятельственные наречия делятся на семантические разряды:
- наречия времени (сперва, вчера);
- наречия места (сюда, слева, направо);
- наречия причины (нарочно, сгоряча);
- наречия цели (напрокат, поневоле)

51. Состав грамматических и семантических признаков деепричастий

Грамматические признаки деепричастий
Деепричастия совмещают в себе грамматические признаки глаголов и наречий.

Признаки глаголов:

Вид – совершенный (сделав, сложив) или несовершенный (делая, складывая);
Переходность (смотря фильм, запоминая дорогу) и непереходность (гуляя по улице, прыгая с дерева);
Возвратность (купаясь, одеваясь) и невозвратность (купая, надев).
Признаки наречий:

Неизменяемость (не склоняются и не спрягаются);
В словосочетаниях обычно, как и наречия, примыкают к личным формам глаголов, реже – к инфинитивам или причастиям (он говорит, смеясь; думать, работая)

52. Состав грамматических и семантических признаков местоименных существительных

53. Состав грамматических и семантических признаков местоименных прилагательных

Грамматические

Местоимения-прилагательные имеют непостоянные признаки рода, числа и падежа, в которых согласуются с существительным, к которому они относятся, склоняются местоимения-прилагательные по адъективному и смешанному склонению, в предложении бывают определением или (редко) именной частью сказуемого.

Семантические

Все **местоименные прилагательные** — изменяемые, они обладают словоизменительными категориями рода, числа и падежа. **Местоименные прилагательные** лишены противопоставления по полноте — краткости.

54. Семантико-синтаксический анализ. Назначение и задачи.

- В процессе семантико-синтаксического анализа (ССА) выполняется построение семантико-синтаксической модели текста.
- ССА - это лингвистическая процедура, обеспечивающая определения в тексте различных синтаксических конструкций и установления синтаксических связей между ними.
- Исходными данными для ССА являются результаты графематического и морфологического анализа текстов.
- Результатами работы ССА является формальная синтаксическая модель текста.

Семантико-синтаксический анализ текстов выполняет следующие задачи по созданию семантико-синтаксической модели текста:

- Реализует модель дерева зависимостей
- Реализует модель членов предложения
- Реализует модель непосредственно составляющих
- Устанавливает границы именных и глагольных словосочетаний в предложении
- Определяет главные и зависимые слова в словосочетании и устанавливается тип связи межу ними
- Определяет синтаксическую роль слов в предложении

Анализ структуры предложения

55. Синтаксические модели. Назначение и задачи.
Модель членов предложения, модель дерева зависимостей, выделение предикатно-актантной структуры

(немного, но не совсем то)

Семантическое представление представляется неупорядоченным графом («сетью»), синтаксические представления являются графическим деревом («деревом зависимостей»), морфологическое и фонологическое представления линейны

56. Построения синтаксической модели членов предложения.

Сеть зависимостей или, точнее, семантико-синтаксическая сеть является более общим типом синтаксической модели, а **дерево зависимостей** - ее частным случаем. В отличие от дерева зависимостей, семантико-синтаксическая сеть может использоваться для описания структуры единиц языка и речи любого уровня, начиная от слов и кончая сверхфразовыми единствами.

Синтаксическая структура текстов обычно описывается в терминах классов слов и их отношений. При этом в качестве классов слов могут выступать части речи (существительное, прилагательное, глагол, наречие и др.), сопровождаемые грамматической информацией, характеризующей конкретные формы слов (например, род, число, падеж, лицо и др.). В качестве отношений - отношения непосредственной доминации с той или иной степенью их дифференциации.

При построении процедур синтаксического анализа текстов на основе метода аналогии будем исходить из следующей гипотезы: одинаковым последовательностям символов классов слов соответствуют одинаковые синтаксические структуры. Предполагается, что гипотеза верна с высокой вероятностью для любых синтаксических моделей. Эта гипотеза полезна при решении как глобальных, так и частных задач синтаксического анали

57. Построения синтаксической модели дерева зависимостей.

При описании синтаксической структуры текстов удобно опереться на какую-либо ее формализованную модель, например, на модель дерева зависимостей. Согласно этой модели каждое предложение представляется в виде дерева, в узлах которого находятся слова. Слова соединяются друг с другом стрелками, выражающими отношения непосредственной доминации и направленными от подчиняющего (определяемого) слова к подчиненному (определяющему). Степень дифференциации этих отношений может быть разная. Причем, чем больше степень дифференциации, тем сложнее процесс описания текстов.

58. Предикатно-актантная структура предложения.

Основной чертой предложений является их предикативность – то есть то их свойство, что в них утверждается наличие у объектов определенных признаков и их отношений. Свойством предикативности обладают и высказывания, формулируемые на формализованных языках. Это позволяет сделать вывод, что в основе и предложений на естественном языке, и формализованных логических высказываний лежит предикатно-актантная структура, компонентами которой являются понятия-предикаты (отношения) и понятия-актанты, выступающие в роли описываемых объектов.

В естественных и в формализованных языках предикатно-актантные структуры являются теми смысловыми инвариантами, которые позволяют осуществлять автоматический перевод текстов с естественных языков на формализованные и с формализованных на естественные. Они также позволяют осуществлять автоматический перевод текстов с одних естественных языков на другие.

59. Концептуальный анализ текстов. Назначение и задачи

Построение семантической сети, объекты сети — понятия, приведенные к доминантному унифицированному виду (унифицированный вид — нормальная форма).

- Выделение наименований понятий (сущностей) выполняется на этапе концептуального анализа текстов.
- Концептуальный анализ текстов — это лингвистическая процедура, обеспечивающая выявления их понятийного (концептуального) состава, формализаци наименований и понятий и установления смысловых связей между ними.
- Исходными данными для концептуального анализа являются результаты семантико-синтаксического анализа текстов.
- Результатами работы концептуального анализа является система понятий анализируемого текста и их смысловые связи.

60. Концептуальный анализ на основе синтаксической структуры.

Представлен в виде символов обобщенных грамматических классов ***(на лекции от 25.11 сообщил, что нам это не рассказывал, удалил вопрос на лекции.)***

LV, LA таблицвыскидывал

61. Концептуальный анализ на основе обобщенных синтагм.

Осуществляется поиск не по буквенному представлению (как в анализе по ЭКС), а по их формам представления. Для этого берется эталонный словарь, по большому числу текстов строим шаблоны обобщенных синтагм.

В этом варианте концептуального анализа выделяемые в тексте нормализованные отрезки сверялись со словарем Обобщенных синтагм,, полученным путем обработки словаря ЭКС. Обобщенные синтагмы представляют собой сочетания символов обобщенных грамматических классов слов, входящих в состав словосочетаний эталонного словаря.

Концептуальный словарь (также идеографический или идеологический **словарь**) - это **словарь**, который группирует слова по понятию или семантическому отношению, а не упорядочивает их в алфавитном порядке.

Процесс автоматического выявления наименований понятий в текстах по данному варианту можно представить следующим образом:

Вначале выполняется членение текста на предложения, выделяются в нем слова и знаки препинания, и проводится морфологический анализ слов. На его основе выполняется пословная нормализация текста.

Производится членение текста на отрезки длиной от одного до пяти слов и для каждого такого отрезка строится обобщенная синтагма. Потом осуществляется сопоставление этих синтагм с элементами словаря обобщенных синтагм.

62. Концептуальный анализ на основе эталонного концептуального словаря.

Делим текст на отрезки (чтобы взять всевозможные фрагменты текста).

Производим нормализацию каждого слова (переводим его в нормальную форму).

Осуществляем поиск по словарю.

Пример предложения: *Если некоторому отрезку текста соответствует….* Получаем отрезки: Если, Если некоторому, Если некоторому отрезку…***и так далее до конца.*** Затем сдвигаемся на слово право и повторяем процедуру: Некоторому, некоторому отрезку, Некоторому отрезку текста…

Есть эталонный словарь, хотим найти именно те слова и словосочетания, которые в нем находятся.

Процесс выявления и формализации наименований понятий в текстах по анализу с помощью ЭКС можно представить следующим образом. Вначале выполняется членение текста на предложения, выделяются слова и знаки препинания, проводится морфологический анализ слов, и на его основе выполняется пословная нормализация текста.

После того как текст был представлен в виде совокупности нормализованных слов, производится членение текста на отрезки длиной от одного до пяти слов и осуществляется сопоставление этих отрезков с элементами словаря ЭКС. Совпавшие нормализованные отрезки текста считаются наименованиями понятий.

Далее на основе результатов семантико-синтаксического анализа текстов устанавливаются синтагматические связи между наименованиями понятий, и с помощью процедуры установления смысловых связей между понятиями производится замена родовых понятий на их видовые понятия.

63. Формализация текстового представления в виде иерархии структур единиц смысла

Есть структура слова, структура словосочетания, структура предложения, структура сверхфразового единства и структура всего текста.

Структура – набор тех метаданных, которые можно получить.

64. Формализованное представление смысловой структуры текста.

Семантическая сеть

Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы. Таким образом, семантическая сеть является одним из способов представления знаний.

В системах семантической обработки текстовой информации основной задачей является формализация представления смысловой структуры текстов – выделения в них смысловых единиц и установления связей между ними. Центральной процедурой при решении этой задачи является процедура семантико-синтаксического концептуального (понятийного) анализа текстов. Важнейшим средством автоматической смысловой обработки текстовой информации являются мощные словари наименований понятий, представленные преимущественно фразеологическими словосочетаниями. При решении задачи формализации смыслового содержания текстов необходимо методами семантико-синтаксического и концептуального анализа обработать текст, разделить его на предложения, выделить из него единицы смысла (наименования понятий) – слова и словосочетания, выражающие понятия.

65. Морфемная структура слов. Состав и характеристика состава морфем слова.

Приставка, корень, суффикс (их около 1500), грамматическое окончание (их около 80)
Здесь суффикс – это суффикс с обобщенным грамматическом окончанием или с флективным классом

Морфемная структура слова — это системно упорядоченное единство его значимых элементов (морфем и комбинаций морфем).

В составе слова различают корневые морфемы (корни), префиксы (приставки) и суффиксы (морфемы, стоящие после корня). Основную смысловую нагрузку несет корень, а префиксы и суффиксы выступают в роли модификаторов смысла.

66. Словоизменительная и словообразовательная основа.

Словоизменительная основа – без грамматическая, словообразовательная основа – без сочетания суффиксов

67. Система флективных классов русских (вопрос повторяется, см. 17 вопрос)

68. Общее представление об методе аналогии

На основании сходства предметов по каким-либо признакам делается вывод об их сходстве и по другим признакам
(не уверена, что оно)

Представление синтаксической структуры текстов в виде последовательности контактно расположенных двухсимвольных элементов обобщенных синтагм, обладающих грамматическими свойствами конкретных слов-эталонов, позволяет фиксировать грамматические и синтаксические свойства различных отрезков реальных текстов, а также дает возможность в ряде задач распознавать аналогичные по заданным свойствам отрезки текстов.

Принцип лингвистической аналогии позволяет выявлять и реализовать трансформационные закономерности словоизменения и словообразования, многократно сократить объемы словарей и грамматических таблиц, а также успешно решить задачи, которые не поддающиеся решению алгоритмическими методами.

69. Тезаурус. Назначение и задачи

Тезаурус - это полный систематизированный набор данных о какой-нибудь области знаний, который разрешает человеку или вычислительной машине в ней ориентироваться.

В отличие от толкового словаря, тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами, благодаря чему может использоваться для наполнения баз знаний систем искусственного интеллекта.

70. Онтология. Назначение и задачи

Онтология – тот же тезаурус, но с правилом вывода

Представляет собой семантическую сеть предметной области, в узлах которой назначены слова и связи между ними (род, вид, часть, целое, ассоциации и синонимы)

Каждый узел ведет свое начало от его метародового представления

71. Метаданные текста. Их характеристика и назначение.

Метаданные текста – вся информация, которую получаем на этапе их базовых характеристик

72. Технологии словарной службы. Назначение и задачи

Технологии словарной службы предназначена для автоматизированного формирования различных словарей и грамматических таблиц

73. Лингвистический процессор как инструмент формализации текстового представления информации. Назначение и задачи

Лингвистический процессор – инструмент формализации текстового представления информации в его смысловое представление. На его основе для текста создается весь набор метаданных.

74. Функции лингвистического процессора.

А. Функции «верхнего» уровня

Б. Служебные процессные функции системного словаря концептов

В. Служебные процессные функции приоритетного словаря концептов

Г. Служебные процессные функции для морфологического анализа и коррекции морфологической информации (отображение и обеспечение работы формы МА)

75. Технология кластеризации текстов. Назначение и задачи

Технология кластеризации текстов – автоматическое сопоставление текстов по их смысловому содержанию и формирование кластеров

76. Технология классификации текстов. Назначение и задачи

Технология классификации текстов – та же самая кластеризация, но заранее задан класс или рубрика в виде системы понятий, с которой мы сопоставляем системы понятий каждого класс