Уровень 1

|  |  |
| --- | --- |
| Слов | 67 |
| Символов (без пробелов) | 539 |
| Символов (с пробелами) | 605 |
| Символов в латинской графике | 95 |
| Чисел | 3 |
| Средняя длина слов | 8,04 |

Упражнение 1

Упражнение 2

Средняя длина слов данного текста соответствует средней длине слов в русском языке.

Средняя длина слов данного текста в русском языке составляет 7,31, а в латинском 5,35.

Средняя длина слов во всем тексте 8,04.

Упражнение 3

а) Ø норвежский

ö немецкий

ß греческий

ρ греческий

ω греческий

ё русский

 š чешский

б) th английский

sch немецкий

šč чешский

в) et французский

the английский

der немецкий

och немецкий

için турецкий

|  |  |  |
| --- | --- | --- |
| Критерий | Язык 1: Русский | Язык 2: Английский |
| Типичные артикли | - | The, a(an) |
| Указательные местоимения | Тот, этот, такой, таков, столько | This, these, that, those |
| Местоимения 3-го лица | Он, она, оно, они | He, she, it, they |
| Отдельные формы вспомогательных глаголов | быть | Be, have, do, will, shall |
| Основные предлоги и союзы | А, или, но, и, да, что, как, чтобы, с, под | And, then, but, on, of, or, if, after |
| Другие частотные слова | Человек, год, иметь, дать, лучший, молодой | Say, go, long, think, see |

Упражнение 4

Упражнение 5

|  |  |
| --- | --- |
| Буква | Встречаемость, % |
| о | 67, 8.66 |
| а | 59, 7.62 |
| е | 54, 6.98 |
| в | 40, 5.17 |
| н | 36, 4.65 |
| и | 36, 4.65 |
| л | 34, 4.39 |
| д | 29, 3.75 |
| с | 29, 3.75 |
| к | 26, 3.36 |
| у | 21, 2.71 |
| м | 21, 2.71 |
| р | 19, 2.45 |
| т | 18, 2.33 |
| п | 15, 1.94 |
| ь | 15, 1.94 |
| ы | 14, 1.81 |
| х | 13, 1.68 |
| г | 13, 1.68 |
| ш | 12, 1.55 |
| й | 10, 1.29 |
| б | 10, 1.29 |
| ж | 9, 1.16 |
| э | 9, 1.16 |
| я | 8, 1.03 |
| ч | 8, 1.03 |
| ё | 3, 0.39 |
| ц | 1, 0.13 |
| щ | 1, 0.13 |

Купание медвежат

Наш знакомый охотник шёл берегом лесной реки и вдруг услышал громкий треск сучьев. Он испугался и влез на дерево.

Из чащи вышли на берег большая бурая медведица и с ней два весёлых медвежонка. Медведица схватила одного медвежонка зубами за шиворот и давай окунать в речку.

Медвежонок визжал и барахтался, но мать не выпускала его, пока хорошенько не выполоскала в воде.

Другой медвежонок испугался холодной ванны и пустился удирать в лес.

Мать догнала его, надавала шлепков, а потом — в воду, как первого.

Очутившись снова на земле, оба медвежонка остались очень довольны купанием: день был знойный, и им было очень жарко в густых лохматых шубках. Вода хорошо освежила их. После купания медведи опять скрылись в лесу, а охотник слез с дерева и пошёл домой.

Упражнение 6

Все прилагательные оканчиваются на -а

Все существительные оканчиваются на -о

Грамматического рода в эсперанто нет.

Окончание существительных в стихотворениях может быть опущено и заменено апострофом

В эсперанто есть два падежа: Винительный и Именительный

Глаголы в настоящем, прошедшем и будущем не изменяются.

Упражнение 1

Лишнее в приведенном ниже списке с точки зрения компьютерной семантики: липа.

Лишнее в приведенном ниже списке с точки зрения компьютерной грамматики: ромашка.

Упражнение 2  
An = прилагательное  
Nn = сущ в И. п.  
Ng = сущ в Р. П.  
N(A)n = сущ, образованное от прилагательного в И.п(обозначающее признак или свойство)  
N(V)n = существительное, образованное от глагола в И. п  
V = глагол  
D = наречие

Упражнение 3  
  
визит врача ↔ врачебный визит,  
N1nN2g =A(N2)nN1n  
ароматный сад ↔ аромат сада,  
A(N1)nN2n = N1nN2g  
выхожу из дома ↔ выход из дома.  
VизNg =N(V)nизNg

Упражнение 4

а) N1n из N2g ↔ A(N2)nN1n

Мед из липы-липовый мед

б) AnNnVDAaNa

Бурый мишка увидел впереди лесной шалаш.

Уровень 2

Упражнение 1

Основные периоды в развитии компьютерной лингвистики:

- 60-е годы – середина 70-х годов ХХ столетия. Разработка формальных моделей и методов, накопление начального опыта в прототипизации ЕЯ-систем. В этот период внимание уделялось формальным грамматикам, моделям языка в области формализмов, классификации грамматик…

- Середина 70-х годов – 80-е годы ХХ столетия. Разработка методов и средств обработки ЕЯ, создание первых промышленных систем общения с базами данных на ЕЯ. В этот момент появились инструменты, которые помогали лингвистам делать морфологию, а следует морфологические словари. Этот период был связан с тем, что создавались первые системы общения ЕЯ с базами данных.

- Середина 80-х годов – середина 90-х годов ХХ столетия. Разработка когнитивных моделей понимания ЕЯ и прототипов систем, использующих модели мира для понимания языка. Здесь появились когнитивные модели понимания естественного языка и некоторые прототипы систем, использующих модели мира для понимания ЕЯ.

- Середина 90-х годов ХХ столетия – начало XXI века. Переход от лингвистика предложения к лингвистики текста, разработка методов и средств обработки ЕЯ текстов. Появление первых коммерческих систем обработки ЕЯ-текстов.

Характеристики V (современного) этапа развития исследований и разработок в области компьютерной обработки ЕЯ-текстов:

- Автоматической обработке подвергаются не искусственные (модельные) тексты, а реальные документы и, в общем случае Web-контент.

- Происходит обработка не единственных текстов, а мультиязычных коллекций документов.

- Обрабатываемые документы содержат опечатки, орфографические ошибки, аграмматичности и другие реальные препятствия на пути к их правильной интерпретации.

- Целью обработки становится не просто получение внутреннего представления смысла документа, а представления результатов в форматах, удобных для эффективного зрения знаний с учётом их постоянного пополнения и последующего использования.

Hot Topics (горячие точки в исследованиях и разработках на сегодняшний день (на момент выпуска лекции)):

- IE-системы извлечения информации из текстов.

- SCC-системы семантической классификации и кластеризации.

- MIE-системы для обработки мультиязычных коллекций документов, которые могут быть получены из Интернета, новостных лент, блогов, корпоративных и персональных баз данных и других источников.

Упражнение 2

Естественные пользовательские интерфейсы

Самый естественный интерфейс для людей - человеческая речь

Самые ранние системы пытались сопоставлять сигналы человеческой речи со словами ,но так как все говорят по-разному, то система давала сбой, потому что один человек может что-то говорить с разной интонацией и в разном контексте

Методы распознавания речи основанные на статистических моделях используются, например, в банках, но в них так же много ошибок(потому что частота ошибок ошибок в произвольной речи около 20-25%)

Технология глубоких нейронных сетей увеличила уровень распознавания на 30% и уменьшение допускаемых ошибок

Bing Translator переводит страницы и видео для пользователей на родной язык, но эта система так же не идеальна(потому что он переводит текст через переводчик и создает понятную структуру текста на родном языке пользователя, следовательно информация искажается)

Сбой при анализе и распознавании звучащей речи произошел потому, что программа не понимала их акцента, соответственно программа была составлена не корректно, и нужны были изменения.

Упражнение 4

На планете Маленького принца, как на любой другой планете, растут травы полезные и вредные. А значит, есть там хорошие семена хороших, полезных трав и вредные семена дурной, сорной травы. Но ведь семена невидимы. Они спят глубоко под землей, пока одно из них не вздумает проснуться. Тогда оно пускает росток; он расправляется и тянется к солнцу, сперва такой милый и безобидный. Если это будущий редис или розовый куст, пусть его растет на здоровье. Но если это какая-нибудь дурная трава, надо вырвать ее с корнем, как только ее узнаешь. И вот на планете Маленького принца есть ужасные, зловредные семена… это семена баобабов. Почва планеты вся заражена ими. А если баобаб не распознать вовремя, потом от него уже не избавишься. Он завладеет всей планетой. Он пронижет ее насквозь своими корнями. И если планета очень маленькая, а баобабов много, они разорвут ее на клочки.

— Есть такое твердое правило, — сказал мне позднее Маленький принц. — Встал поутру, умылся, привел себя в порядок — и сразу же приведи в порядок свою планету. Непременно надо каждый день выпалывать баобабы, как только их уже можно отличить от розовых кустов: молодые ростки у них почти одинаковые. Это очень скучная работа, но совсем не трудная.

|  |  |  |
| --- | --- | --- |
| Частота слов в тексте | | |
| Слово | Кол-во слов | Частота, % |
| и | 7 | 3,6 |
| на | 5 | 2,5 |
| Семена | 5 | 2,5 |
| ее | 4 | 2 |
| если | 4 | 2 |
| не | 4 | 2 |
| это | 4 | 2 |
| а | 3 | 1,5 |
| есть | 3 | 1,5 |
| как | 3 | 1,5 |
| но | 3 | 1,5 |
| он | 3 | 1,5 |
| планета | 3 | 1,5 |
| баобаб | 2 | 1 |
| в | 2 | 1 |
| вредные | 2 | 1 |
| маленького | 2 | 1 |
| надо | 2 | 1 |
| них | 2 | 1 |
| они | 2 | 1 |
| от | 2 | 1 |
| очень | 2 | 1 |
| порядок | 2 | 1 |
| принца | 2 | 1 |
| только | 2 | 1 |
| травы | 2 | 1 |
| уже | 2 | 1 |
| безобидный | 1 | 0,5 |
| будущий | 1 | 0,5 |
| ведь | 1 | 0,5 |
| вздумает | 1 | 0,5 |

|  |  |
| --- | --- |
| Частоты Ципфа | |
| Слово | Частота, % |
| Планета (планете, планеты, планетой, планета, планету) | 3,61 |
| Семя (семена) | 2,58 |
| Трава (травы, трав, трава) | 2,06 |
| Маленький (маленького, маленькая, маленький) | 2,06 |
| Баобаб (баобабов, баобабы) | 2,06 |
| Принц (принца) | 1,55 |
| Уже | 1,03 |
| Очень | 1,03 |
| Безобидный | 0,52 |
| Будущий | 0,52 |

Упражнение 5

А) Квантитативная лингвистика – это область лингвистики, которая использует математические и статистические методы для изучения языка и его структуры. Она основывается на идее, что язык можно изучать и анализировать с помощью количественных данных, таких как частота употребления слов, длина предложений, распределение звуков и т.д.

Квантитативная лингвистика стремится к объективности и точности в изучении языка, позволяя исследователям собирать и анализировать большие объемы данных. Она помогает выявить закономерности и тенденции в языке, а также понять его структуру и функционирование.

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

В) Компьютерная лексикография отличается от традиционной лексикографии тем, что она занимается созданием и анализом словарей с использованием компьютерных технологий. Традиционная лексикография включает в себя создание бумажных словарей, в то время как компьютерная лексикография работает с электронными словарями и базами данных.  
  
Г) Современные методы анализа текста включают в себя различные подходы и технологии, такие как:  
1. Частотный анализ слов - подсчет количества вхождений каждого слова в тексте для определения наиболее часто встречающихся слов.  
2. Машинное обучение и искусственный интеллект - использование алгоритмов машинного обучения для анализа текста, классификации текстов по темам, извлечения информации и других задач.  
3. Синтаксический анализ - определение структуры предложений и связей между словами в тексте.  
4. Сентимент-анализ - определение тональности текста (положительной, отрицательной или нейтральной) с целью понимания эмоциональной окраски текста.  
5. Тематическое моделирование - выявление тематик и ключевых слов в тексте с помощью статистических методов.  
  
Эти методы позволяют проводить более глубокий и точный анализ текстов, что может быть полезно для различных целей, таких как информационный поиск, машинный перевод, анализ социальных медиа и другие области.

Упражнение 6

Ментальная карта квантитативной лингвистики:

1. Статистический анализ текста:

- Частотный анализ слов и грамматических конструкций

- Zipf's Law (Закон Ципфа) и его применение в лингвистике

- Методы анализа корпусов текстов

2. Моделирование языка:

- Марковские модели для прогнозирования последовательностей слов

- Модели n-грамм для оценки вероятности последовательностей слов

- Модели Языковых машин Тьюринга

3. Семантический анализ:

- Векторное представление слов и семантические пространства

- Методы анализа семантической близости и ассоциаций

- Word Embeddings (вложения слов) и их применение в задачах обработки естественного языка

4. Машинное обучение в лингвистике:

- Классификация текстов и анализ тональности

- Методы кластеризации текстов

- Глубокое обучение и нейронные сети в задачах обработки текстов

5. Компьютерная лингвистика:

- Обработка естественного языка (Natural Language Processing)

- Автоматическое извлечение информации из текстов

- Методы автоматического перевода и генерации текстов

Упражнение 7

Введение

Исследования в области корпусной лингвистики обычно решают сразу две задачи: создание рабочих инструментов и использование этих инструментов для изучения лингвистических феноменов. При этом рабочие инструменты чаще всего предполагают более или менее конкретные лингвистические феномены, которые планируется изучить, а интересы исследователя связаны с возможностями, предоставляемыми рабочими инструментами.

Создание рабочих инструментов для анализа включает в себя построение моно­ и многоязыковых корпусов текстов в рамках области, интересующей исследователя, а также средства разметки корпусов, поиска в них и статистической обработки результатов поиска. Эти ресурсы используются в собственно лингвистических исследованиях, предполагающих анализ языка с экспериментальной точки зрения, т.е. исследование того, какие слова, выражения, грамматические конструкции, типы развития дискурса действительно употребляются носителями языка, как часто и для каких целей.

История

В связи с тем, что компьютер представляет собой удобное средство хранения и обработки больших объемов информации, первые проекты по созданию коллекций текстов и корпусов появились сразу же, как только лингвисты получили доступ к компьютерам с достаточными возможностями для хранения текстов. Первая версия корпуса была представлена простым текстовым форматом (с небольшим количеством структурной разметки для выделения абзацев, заголовков, цитируемых фрагментов и т.п.). Позднее корпус был дополнен разметкой частей речи и морфологических признаков слов

Второй похожий британский проект Bank of English начал создаваться в 1980­е годы, в 1989 его объем достиг 20 млн. слов, в 2000 ­ 600 млн. слов. Этот корпус ориентирован на отслеживание изменений в словоупотреблении (monitor corpus) путем регулярного пополнения новыми текстами и сравнения частотных параметров, например, таких как изменение частоты слов и грамматических конструкций, появление новых слов и т.п. Этот корпус служил основой создания словаря Collins COBUILD English Dictionary [9] и ряда английских грамматик, использующих корпусный данные.

Разметка корпуса

По мере роста возможностей компьютеров и увеличения количества документов, доступных в электронной форме, размеры корпусов непрерывно росли: от 500 тыс. слов в 1960­е годы до последних проектов, нацеленных на создание корпусов размеров в 1 млрд. слов [10]. Размер является важным параметром оценки корпуса, но не менее важным является состав дополнительной информации внесенной в результате обработки исходного текста.  В связи с тем, что современные средства обработки языка пока не позволяют вносить семантическую или синтаксическую информацию автоматически, а ручная разметка корпуса объемом в десятки миллионов слов нереальна, разметка большинства больших (и даже не очень больших) корпусов ограничена лишь так называемой лемматизацией (обозначением леммы) и, следовательно, возможностью для пользователя найти в тексте все словоформы произвольной лексемы.

Более детальная разметка, учитывающая морфологические или морфологосинтаксические признаки очень трудоемка, потому что лингвистически корректная разметка русских текстов даже на морфологическом уровне без внесения синтаксической информации обязательно требует ручной работы. Однако для небольшого корпуса ручная или полуавтоматическая разметка возможна не только для морфологической, но и для синтаксической или семантической информации.

В то же время, развитие технологий обработки текстов позволяет вносить все большее количество информации автоматически, в частности, идентифицировать объекты по именам или различать омонимы на основе локального контекста. Разметка текста лингвистической информацией в настоящее время чаще всего основана на языке SGML/XML, который предполагает выделение фрагментов текста.

Использование корпусов для лингвистический исследований

Любой корпус создается как средство отражения и эмпирического исследования явлений, встречающихся в языке (или подъязыке, в случае специализированного корпуса). Наличие компьютерного корпуса не меняет радикально деятельность лингвиста. Корпуса текстов использовались для создания грамматик и словарей задолго до появления компьютеров.

Создание представительных корпусов и их разметка на различных уровнях влечет за собой создание словарей и грамматик, построенных на основе корпусной методологии.

Представительный британский корпус LOB, построенный по модели Брауновского корпуса, лег в основу Quirk, et al, 1985). Создание первого большого английского корпуса Bank of English (20 млн. словоупотреблений в 1980­е годы) привело к созданию серии грамматик и словарей серии COBUILD. Создание БНК привело к появлению новой грамматики . Задача, встающая перед корпусно­ориентированной грамматикой, заключается в создании описания, которое может дает адекватный анализ для любого словоупотребления, зафиксированного в корпусе, а наибольшее внимание обращается на наиболее частотные случаи.

К созданию представительного корпуса русского языка

В настоящее время идея осуществления такого проекта возрождается (см. статью Л.А. Вербицкой, Н.Н. Казанского и В.Б. Касевича в настоящем сборнике). Предполагаемый корпус будет иметь разные составляющие. Отдельной задачей видится, например, представительный корпус XIX­го и первой половины XX­го века, по понятным причинам состоящий из письменных, прежде всего, художественных текстов; специфических проблем создания такого корпуса (и тем более трудностей, связанных с созданием корпуса древнерусских текстов) мы в настоящей статье касаться не будем.

Высказываемые ниже соображения применимы прежде всего к корпусу современного русского языка (со второй половины XX­го века). Они основаны на собственном экспериментальном проекте автора, который мог бы быть использован для дальнейшей коллективной работы в этой области. Этот проект фигурирует пока под условным названием БОКР (Большой корпус русского языка). БОКР представляет все значимые виды использования русского языка в настоящее время. Автор проекта предполагал, что относительно небольшой фрагмент этого корпуса, наилучшим образом отражающий нейтральный литературный язык второй половины XX в., мог бы быть полностью размечен морфологически – со снятой вручную омонимией (тогда как разметка остальной части корпуса будет произведена автоматически).

Типология текстов

При создании представительных корпусов требуется максимально широкое покрытие различных типов текстов и функциональных стилей, для чего была разработана типология текстов на основе рекомендаций Синклера [24]. Англоязычная литература содержит множество предложений по классификации для создания представительных корпусов, например, [30, 31, 32, 33] и т.д. Принимая эти рекомендации необходимо иметь в виду, что среда использования языка оказывает существенное влияние на классификацию текстов. Так очевидное несоответствие между российской и англоязычной культурами выражено и на уровне жанров. Такие жанры как выделяемые в БК Foundation reports или Popular Lore с трудом допускают даже адекватный перевод на русский.

Разметка слов

Корпус предполагает проведение лемматизации и морфосинтаксической разметки текстов. Хотя многие корпуса английского языка ограничивались частеречной разметкой, в случае русского языка безусловно необходима лемматизация (иначе будет затруднен поиск многих словоформ) и приписывание граммем (это позволит исследовать предложное и глагольное управление). Поскольку выделение именных и предложных групп в русском языке автоматическими методами достаточно надежно, корпус предполагает также частичную синтаксическую разметку.  После этапа морфологического анализа словоформы с неоднозначными грамматическими характеристиками (например, род, число, падеж) составляют около 60% словоупотреблений, словоформы с неоднозначным выделением лексемы и части речи – около 30%. Однако частичный синтаксический анализ на уровне именных и предложных групп может оказать существенную помощь при разрешении некоторых видов неоднозначности, в частности: • неоднозначности падежных форм, например, новой книги, где при неоднозначности отдельно взятых словоформ у целой именной группы возможен только родительный единственного; • неоднозначности субстантивированных прилагательных: ст арший в мой ст арший брат получает только интерпретацию прилагательного (а мой – местоимения), но в ст арший группы – только существительного; • неоднозначности притяжательных и личных местоимений, например, в его книге, где невозможна интерпретация личного местоимения,  • омонимии между существительным или прилагательным и глаголом в повелительном наклонении: кривей, мой, полей, при наличии в непосредственной близости другого, однозначно определяемого, глагола, глагольная форма невозможна. Наоборот, при наличии согласования с близлежащим существительным или местоимением в именительном падеже формы были, замер, ст али, однозначно интерпретируются как глаголы. Аналогично можно разрешить и многие случаи омонимии между кратким прилагательным и наречием.

Заключение

Подавляющее большинство исследований в области корпусной лингвистики начиналось на материале английского языка. Причиной этого является не только и не столько активное развитие компьютерной техники в США а интеллектуальный климат в Британской лингвистике в 60­80 годы 20 века. В США в это время властвовал хомскианский подход, основанный на лингвистической интуиции, которая не требует наличия корпусных данных (зачастую хаотичных и зависящих от более широких контекстов высказывания), поскольку объектом изучения является возможность построения правильных языковых конструкций (well­formedness), а различие между правильными и неправильными конструкциями может быть проведено любым носителем изучаемого языка. В противоположность рационалистскому подходу, основанному на лингвистической интуиции, проводящей различие между правильными и неправильными конструкциями, эмпирический подход предполагает, что язык является ресурсом, обеспечивающим набор возможностей для коммуникации. Этот набор реализуется в дискурсе, поэтому объектом исследования в лингвистике является результат реализации этого ресурса, а именно слова и конструкции, употребленные в тексте.

Упражнение 8

Звуковые уровни представляют собой фонематическую (Sound Ideal) и фонетическую (Sound Real) транскрипции потока речи. Причем, если первая выполняется в соответствии с правилами, выработанными в рамках Санкт-Петербургской фонологической школы, учитывающей фонемную интерпретацию слов, произнесенных в полном типе, то вторая (фонетическая) транскрипция осуществляется на основе слухового анализа реального звучания отдельных звуков и/или звуков в контексте с помощью знаков Международного фонетического алфавита  
— звуки (sound real);  
— фонемы (sound ideal).