


ВВЕДЕНИЕ В ЯЗЫКОЗНАНИЕ

Приложения вычислительной ЛИНГВИСТИКИ

Глава 6. Познание языка





Эта лекция посвящена компьютерному языку. Полезность компьютеров во всех возможных областях, связанных с языком, не подвергается сомнению. Мы уже затронули некоторые из них в нашем исследовании различных подполей вычислительной лингвистики, таких как интерфейсы естественного языка для различных видов компьютерных программ. В данной лекции мы обсудим некоторые из наиболее распространенных областей применения, начиная от использования компьютеров и заканчивая проверкой грамматики лингвиста на соответствие фактическому языку.

Грамматические правила, используемые компьютерами, могут не совпадать с грамматическими правилами, которые лингвисты создают для человеческого языка. Грамматика, созданная лингвистами, является моделью лингвистической компетентности. Компьютеры и люди – разные. Они достигают аналогичные цели по-разному. Так же, как эффективный летательный аппарат не является копией какой-либо птицы, хотя на стадии разработки имеются птичьи летательные аппараты. Эффективные грамматические правила для компьютеров не всегда напоминают грамматику человеческого языка.

Несмотря на это, компьютеры могут быть смоделированы для описания грамматики лингвиста. Точная грамматика, которая является настоящей моделью компетенции говорящего, должна быть способна генерировать все предложения. Неспособность генерировать грамматическое предложение указывает на ошибку в грамматике, равно как и генерация строки, которую ораторы считают неграмматической. Хотя в реальной речи мы можем производить (технически) неграмматические фрагменты-фразы, высказывания языка и т. д. Благодаря нашим знаниям о грамматических правилах, в случае, если мы замечаем, что кто-то неправильно использует грамматику, мы говорим об этом. Но поскольку объем грамматики очень большой, для проверки их лингвистического описания необходимы огромные объемы тестирования, и именно здесь важную роль играют компьютеры.

Компьютерные модели грамматики лингвистов относятся к 1960-м годам, когда программы тестирования генеративной грамматики английского языка были разработаны синтаксистами в Калифорнийском Университете в Лос-Анджелесе.

Такие модели все еще разрабатываются для тестирования новых теорий грамматики. Специалисты вычислительной лингвистики разрабатывают компьютерные программы для генерации предложений и имитации обработки этих предложений людьми с использованием правил, включенных в различные современные лингвистические теории. Вычислительные модели показывают, что можно использовать зафиксированную грамматику в производстве и понимании языка, что, по меньшей мере, частично является успешным описанием человеческой компетенции языка. Но говорить, что такого рода грамматика, произведенная вычислительной машиной, является истинной моделью обработки человеческого языка еще очень рано.

Джонатан Свифт предсказал один из видов применения компьютеров для языка: статистический анализ. Частота повторяемости букв, звуков, морфем, слов, категорий слов, фраз, которые могут быть быстро и точно вычислены для любого корпуса тела языковых данных, будь то текстовые или устные. Анализ частоты повторяемости миллиона слов письменного американского английского языка показывает десять наиболее часто встречающихся слов: «the», «of», «and», «to», «a», «in», «that», «is», «was» и «he». Эти слова составляли около 25 процентов слов в лингвистическом корпусе американского английского языка. Лидирующим словом стало «the», которое составило 7 процентов лингвистического корпуса. Аналогичный анализ устного американского английского показал несколько иные результаты. «Победителями» стали «I», «and», «the», «to», «that», «you», «it», «of», «a» и «know». Они составили почти 30 процентов лингвистического корпуса устного американского английского. Это лишь одно из различий между устным и письменным языком, найденных методом анализа корпусов. Все английские предлоги, за исключением «to», чаще встречаются в письменном, чем в устном английском языке. Также, не удивительно, ненормативная лексика и табуированные слова гораздо чаще используются в разговорной речи, чем в письменном языке.

Анализ частоты повторяемости слов проводился по существующим текстам и раскрывали стили авторов. Например, анализируя различные книги Библии, можно понять, кто написал определенные отрывки. В примечательном исследовании Федералистические бумаги, авторство спорного документа было приписано Джеймсу Мэдисону, а не



Александру Гамильтону. Это было достигнуто путем сравнения статистических анализов рассматриваемой статьи с результатами известных работ двух авторов.

Конкорданция идет еще дальше анализа частоты повторяемости слов, так как она указывает местоположение в тексте каждого слова и его окружающего контекста. Конкорданция абзаца предшествующего предыдущему, не только показывает, что слово «words» было использовано пять раз, но также будет указывать, в какой строке абзаца оно использовался, и укажет его контекст. Если бы вы выбрали «window» из трех слов с обеих сторон для контекста, алфавитный указатель слов для слова «words» выглядел бы следующим образом:

Of one million	words	of written American
Most frequently occurring	words	the, of, and
These "little"	words	accounted for about
Percent of the	words	in the corpus
Profane and taboo	words	

Конкорданция, может иметь ограниченную полезность из-за ее «сырой» природы. Способ уточнения конкорданции – это анализ коллокаций. Коллокация – это возникновение двух или нескольких слов в коротком промежутке друг от друга в лингвистическом корпусе. Здесь важно найти доказательства того, что присутствие одного слова в тексте влияет на появление других слов. Такой анализ должен быть статистическим и включать большие образцы, чтобы показать значительные результаты. В вышеназванной конкорданции слова «word» недостаточно данных, так как мы анализировали только один абзац. Если бы мы выполнили конкорданцию по всей этой книге, появились бы такие шаблоны как «words» и «written», «words» и «taboo», «words» и «of» и, скорее всего, такие шаблоны чаще использовались бы с друг другом, чем «words» и «million».

Конкорданция звуков с помощью компьютера может выявить модели в поэзии, которые было бы почти невозможно обнаружить человеку. Использование компьютеров позволяет литературоведам легче изучать поэтические и прозаические особенности, такие как ассонанс, аллитерация, метр и ритм.

Сегодня компьютеры могут выполнять утомительную механическую работу, которую когда-то нужно было тщательно обрабатывать бумагой и карандашом. Ни первые словари Сэмюэля Джонсона, ни обновленный Оксфордский словарь английского языка (OED), ни стандартные словари не подходят для специалистов вычислительной лингвистики, которым нужен огромный объем информации об отдельных словах и морфемах для понимания компьютера, генерации естественного языка, машинного перевода и т. д. Поэтому поле вычислительной лексикографии касается не только создания стандартных словарей, но и создания электронных словарей, специально предназначенных для специалистов вычислительной лингвистики.

Специалисты вычислительной лингвистики нуждаются в нижеследующей информации:

- фонематическая транскрипция;
- фонетические варианты (диалектные, социальные);
- разделение на слоги;
- синтаксические категории;
- семантические свойства (абстрактные, человеческие, одушевленные и т. д.);
- число (например, «people» множественное, person единственное);
- пол (например, «ship» является женским);
- «c-selection» (для «murder» требуется прямое дополнение);
- «s-selection» («murder» требует человека и дополнение);
- стилистический уровень («ain't» является неформальным, «rad» – сленг, «fuck» – табу и т.д.);
- синонимы, антонимы, возможные омофоны и т. д.

«Wordnet» – это онлайн-словарь с десятками тысяч вводных данных, необходимых для специалистов вычислительной лингвистики с акцентом на семантические отношения.

Другие подобные проекты создаются для европейских языков (проект «EuroWordNet»), балканских языков (проект «BalkanNet») и различных других языковых сообществ.



Когда-то миллион слов письменного английского было огромным числом. Сегодня число 361 000 слов выросло на 361 млрд. слов английского языка. Слова составлены из примерно 5 миллионов книг, что составляет 4% всех книг, изданных на английском языке. Анализ этого корпуса, исследование, названное «культуромикой», создает увлекательные идеи в области лексикографии и грамматических изменений, а также во многих социальных науках. Установив общепринятый критерий, что любая последовательность букв, которая произошла с частотой, превышающей один миллиард с 1500-х годов, почти полмиллиона «неизвестных» слов, таких как «slenthem» – «тип музыкального инструмента», теперь считаются словами английского лексикона. На самом деле, по оценкам, 52% английской лексики – большинство слов, используемых в английских книгах, состоит из лексической «темной материи», недокументированной стандартными словарями.

Революция культуромики также дала результаты об историческом изменении неправильных глаголов. Неправильные глаголы прошедшего времени живут бок о бок со своими обычными «-ed» соперниками (например, «dove» и «dived»), которые угрожают вытеснить их. Культуромический анализ сообщает нам, что в период между 1800 и 2000 годами шесть неправильных глаголов прошедшего времени «burn», «chide», «smell», «spell», «spill» и «thrive» («burnt», «chid», «smelt», «spelt», «spilt», «throve») стали правильными, а в то же время два правильных прошедших глагола времени «light»/«lighted» and «wake»/«waked» стали неправильными: «light»/«lit» и «wake»/«woke». «Правильность» глагола определяется через процент экземпляров, в которых используется обычная форма: например, в 1800 году слово «chided» было использовано на 10%, а в 2000 году на все 90%. Даже норма изменения может быть вычислена из корпуса; «chide» была самым быстрым глаголом; «spill» регулируется с постоянной скоростью; но «thrive» с перебоями начал меняться. Сегодня глагол, движущийся быстрее к неправильности, это – «sneak», уже 1% англоязычного населения использует «snuck» вместо «sneaked». Что интересно, регуляция неправильных «-t» – «burnt», «smelt», «spelt», и «spilt», возникла в Соединенных Штатах. Неправильные «-t» все еще цепляются за жизнь в британском английском, но едва ли им это удастся. Каждый год огромное количество людей, сравнимое с населением города Кембридж, используют «burned» вместо «burnt».

На момент написания этой книги мы затрагиваем языковые возможности, которые возникают в результате изучения обширных корпусов. На сегодня 12% всех книг (около 15 миллионов) были оцифрованы, и процесс все еще продолжается. Более того, некоторые периодические издания еще не вошли в корпус. Есть мир увлекательных исследований, с которым можно познакомиться в анализе корпусов.

В социальных науках твиттерология используется для анализа тональности текста путем подсчета частоты появления положительных эмоциональных слов, таких как «awesome» и «fantastic», а также отрицательных слов, таких как «panic» и «fear». Таким образом, можно увидеть ритм повседневной жизни внутри сообщества или даже внутри страны.

Разведсообщества, такие как ЦРУ, используют компьютеры для анализа целых 5 миллионов твитов в день, чтобы измерить настроение региона после крупного события, как, например, убийство Усамы бен Ладена. Также правительственные учреждения собирают, казалось бы, бессмысленные сленги и жаргоны, чтобы помочь выявить личность автора и политические пристрастия, а также его склонность противостоять власти или совершить акты насилия.

Лингвисты в Университете Карнеги-Меллона изучают различия диалектов разных регионов и демографических групп с помощью твиттера. Огромное население Северной Калифорнии в твиттере широко использует слово «hella», например «It's hella cold out there».

Различные регионы склонны к различным фонетическим вариантам написания. Например, жители Нью-Йорка предпочитают «suttin» и «sumthin» (для слова «something»), а калифорнийцы пишут «koo» или «soo» для слова «cool». Смайлики тоже отличаются от региона к региону и от социальной группы к социальной группе.

Инвазивность терминов от одного диалекта к другому также может быть измерена посредством твиттера. Перед твиттером все равны, поэтому использование таких американизмов, как «janitor», «parking lot» и «teenager» в Британии может также быть раскрыто твиттерологией. Обычно твиттерология ограничивается очень узкой формой общения, которая



называется тривиальной и поверхностной. Недаром лингвист Ноам Хомский предупредил нас, что «твиттер – это не средство серьезного общения». Несомненно, выводы о языке, основанные на твиттерологии, должны быть тщательно отобраны. Научная полезность твиттерологии, на наш взгляд, является спорным вопросом, на который мы сможем ответить, наверное, в течение второго десятилетия этого столетия.

Многие люди используют функции поиска в Интернете для поиска информации. Вы вводите ключевое слово и волшебным образом компьютер показывает веб-сайты, содержащие информацию по этому ключевому слову. Этот процесс является примером поиска информации. Это может быть так же тривиально, как поиск веб-сайтов, которые содержат ключевое слово, точно так же, как оно введено, но чаще используется расширенный лингвистический анализ. Веб-сайты возвращаются и даже ранжируются в зависимости от частоты появления ключевого слова, различных морфологических форм ключевого слова, синонимов ключевого слова и понятий, семантически связанных с ключевым словом. Например, ключевое слово «bird» может извлекать информацию, основанную на «bird», «birds», «tobird» (глагол инфинитив), «birdfeeders», «waterbirds», «avian», «sparrow», «feathers», «flight», «migration» и в баскетболе великий «Larry».

Такие компании, как «Google», превратили поиск информации в многомиллиардное предприятие, которое финансировало вход «Google» в несколько других областей обработки информации, включая производство огромных корпусов, которые используют культурамики.

В общем, поиск информации – это использование компьютеров для поиска и отображения данных, полученных из, возможно, очень больших баз данных. Вход в систему поиска информации состоит из слов, заявлений или вопросов, которые компьютер анализирует лингвистически, а затем использует результаты для просеивания через базу данных для соответствующей информации. В настоящее время сложные информационно-поисковые системы идентифицируют полезные шаблоны или отношения в корпусах или других репозиториях компьютеров с использованием передовых лингвистических и статистических анализов. Термины интеллектуального анализа данных, обнаружения знаний используются в настоящее время для высокоразвитых информационно-поисковых систем.

Ключевое слово «bird» может возвращать больше информации, чем может быть прочитано за десять жизней, если происходит тщательный поиск в Интернете. Для получения информации о дате выпуска данной работы, поиск составил 637 миллионов обращений, по сравнению с 200 миллионами три года назад и 122 миллиона семь лет назад. Большая часть данных повторилась, и некоторая информация перевешивала бы другую информацию. С помощью программ суммирования компьютеры могут устранить избыточность и определить наиболее характерные особенности информации. Мировые лидеры, руководители компаний и даже преподаватели университетов, все из которых, возможно, пожелают переварить большие объемы текстовых материалов, таких как отчеты, газеты и научные статьи, могут извлечь выгоду из процессов обобщения, предоставляя материал в машиночитаемой форме, что становится все более актуальным, поскольку мы уже входим во второе десятилетие XXI века.

Типичным сценарием будет использование поиска информации для доступа, скажем, к ста статьям о птицах. Статьи могут составлять в среднем 5000 слов каждая. Применяются программы суммирования, которые могут быть установлены для уменьшения статьи на определенную сумму, скажем 1/10 или 1/100. Человек читает окончательный результат. Таким образом, 500 000 слов могут быть сокращены до 5000 или 10000 слов, содержащих наиболее подходящую информацию, которая затем может быть прочитана через десять или двадцать минут.

Программы суммирования варьируются от упрощенного «печатать первое предложение каждого абзаца» до сложных программ, которые анализируют документ семантически, чтобы идентифицировать важные моменты, часто используя «векторы понятий». Вектор концепции – это список значимых ключевых слов, присутствие которых в абзаце является мерой значимости этого пункта и, следовательно, указывает, следует ли включать содержание этого пункта в обобщение.

В сводном документе содержатся концепции из максимально возможного количества ключевых абзацев с ограничениями длины.



Проверка орфографии и, возможно, в будущем, проверка произношения – это приложения вычислительной лингвистики, которые отличаются изоощренностью от бессмысленных, грубых поисков в словаре, до тех, у кого достаточно интеллекта, чтобы опознать «youг», когда это должно быть «you're», или когда вам нужно использовать слово «bаgе» вместо «beаg». Часто можно найти проверки орфографии в качестве передних концов в информационно-поисковых системах, проверяя ключевые слова, чтобы предотвратить ошибки в поиске. Большинство систем электронной почты также выполняют проверку орфографии, хотя эта функция может быть нежелательной при текстовом режиме из-за большого количества нестандартных применений. Более того, как видно из следующего стихотворения, проверки орфографии не могут заменить тщательное редактирование:

I have a spelling checker,
It came with my PC.
It plane lee marks four my revue
Miss steaks aye can knot sea.

A checker is a bless sing,
It freeze yew lodes of thyme.
It helps me right awl stiles to reed,
And aides me when aye rime.

To rite with care is quite a feet
Of witch won should bee proud,
And wee mused dew the best wee can,
Sew flaws are knot aloud.

Необходимость перевода между языками никогда не была выше, чем в современном глобальном обществе, и объем и сложность задачи делают ее в высшей степени подходящей для использования в вычислительной технике. Первое использование компьютеров для обработки естественного языка началось в 1940-х годах с попытки разработать автоматический машинный перевод. В то время перевод рассматривался так, как будто он расшифровывал код. Целью автоматического перевода является ввод речи или письменного отрывка на исходном языке и получение грамматического прохода эквивалентного значения на целевом языке (выход). В первые годы машинного перевода считалось, что эта задача может быть выполнена путем ввода в память компьютера словаря исходного языка и словаря с соответствующими морфемами и словами целевого языка. Программа перевода попыталась сопоставить морфемы входного предложения с фразами целевого языка. К сожалению, часто случался процесс, который ранние экспериментаторы с машинным переводом называли «language in, garbage out» («принимается язык, а выходит мусор»).

Перевод – это нечто большее, чем замена слова в слово.

Часто на переводящем языке нет эквивалентного слова, и порядок слов может отличаться, как на английском языке «redhouse», который в испанском звучит как «lacasaroja». Также трудно переводить идиомы, метафоры, сленговые слова и т. д. Переводчики справляются с этими проблемами, потому что они знают грамматику двух языков и используют общие знания предмета и мира, чтобы достичь намеченных значений. Но иногда они терпят неудачу, о чем свидетельствуют некоторые признаки «вывоза мусора», на примере вывесок для туристов в неанглоязычных странах:

«The lift is being fixed for the next day. During that time we regret that you will be unbearable» – Лифт фиксируется на следующий день. За это время мы сожалеем, что вы будете невыносимы. (Бухарестский холл гостиницы)

«The nunsharborall diseases and have no respect for religion» – Монахини питают все болезни и не уважают религию. (Швейцарский женский монастырь)

All the water has been passed by the manager» – Вся вода была передана менеджером. (Немецкий отель)



«Because of the impropriety of entertaining guest of the opposite sex in the bedroom, it is suggested that the lobby be used for this purpose» – Из-за непристойности развлекательного гостя противоположного пола в спальне предлагается использовать лобби для этой цели. (Отель в Цюрихе)

Такие «переводы» представляют трудности поиска правильных слов, но выбор слов не является единственной проблемой в автоматическом переводе. Есть проблемы в морфологии при переводе между языками.

Синтаксические проблемы одинаково сложны. Английский язык – это язык, который обладает притяжательными формами различной синтаксической сложности, такими как «that man's son's dog's foodish», или «the guy that my roommate is dating's cousin». Перевод этих предложений без потери смысла на языки, запрещающие такие структуры, требует значительной реструктуризации предложений.

Переводы отдельных слов между десятками языков с африкаанс на идиш легко доступны в Интернете, но людям необходимо с точностью переводить более длинные отрывки, такие как газетные статьи, в целях экономии времени и усилий, профессиональные переводчики очень часто используют компьютеры для их перевода, но компьютерные программы не гарантируют грамматически и семантически верный перевод.

Мы косвенно обсуждаем перевод письменных текстов. А как насчет перевода речи с одного языка на другой? С одной стороны, требуется распознавание речи – или «речь-текст». С другой стороны, требуется «текст-речь». Наиболее общий сценарий машинного перевода – речь в речи – инкапсулирует все области вычислительной лингвистики.

Этот анализ и многие подобные ему показывают полезность основанного на корпусе компьютерного подхода анализа к сложным юридическим проблемам. Для вычислительных криминалистов-лингвистов все чаще становится обычным искать информацию не в базе данных, а в других местах более изобретательными путями.

Многие преступления связаны с анонимными сообщениями, в которых важно идентифицировать человека, который посылает эту информацию. Определение личности говорящего – это использование компьютеров для оказания помощи в такой задаче, в отличие от слуховых наблюдений, которая зависит от суждения слушателей.

Полностью автоматический перевод речи на уровне профессиональных устных переводчиков, таких, которые встречаются в международных встречах, как Организация Объединенных Наций, по-прежнему остается не достижимым.

Судебная лингвистика – это подполе лингвистики, которое применяется к языку, используемому в правовой и судебной областях. Он включает авторские исследования, интерпретацию юридического языка, языковые права и его использование в зале суда, анализ заявлений (например, заметки о самоубийстве), защиту и нарушение товарных знаков (McWho?), опознать личность говорящего (кто оставил эту угрозу бомбы?), текстовую аутентификацию (например, вопросы плагиата), правдивость чтения по губам и т. д.

Вычислительная криминалистическая лингвистика – это подобласть, которая относится к компьютерным приложениям в судебной лингвистике. Мы рассмотрим три таких приложения: торговые марки, интерпретацию юридических терминов и личность говорящих.

Нюансы смысла юридического языка оспаривались на протяжении всей истории судебных систем. Например, правовое определение «фунта плоти» занимает центральное место в сюжете пьесы Шекспира «Венецианский Купец». К тому же, недавний случай зависел от законного использования слова «виза», а не как товарного знака кредитной карты, а как юридический термин, относящийся к международным поездкам. Речь шла о том, дает ли виза путешественнику безусловное право на въезд в страну выдачи виз, или если это что-то тонкое, но значительно другое. Специалист вычислительной лингвистики изучил корпус, состоящих из многомиллионных слов Банка Англии, и нашел 74 вида виз и виз, связанных с общими глаголами, такими как вопрос, отказ, подача заявки, потребность и требование, и смог успешно утверждать, что смысл визы был, по мнению обычного путешественника, видом разрешения на въезд в страну, а не на разрешение просить разрешения на въезд в страну, поскольку, если это



имело место даже с визой, путешественнику было бы отказано во въезде. Этот вывод одного британского суда продолжает иметь последствия в мире международного права, хотя вопрос до сих пор еще не полностью урегулирован.

Этот анализ и многие подобные ему показывают полезность основанного на корпусе компьютерного подхода анализа к сложным юридическим проблемам. Для вычислительных криминалистов-лингвистов все чаще становится обычным искать информацию не в базе данных, а в других местах более изобретательными путями.

Два инструментария вычислений обычно применяются для помощи в идентификации говорящих. Один показывает форму волны высказывания, которая показывает изменения амплитуды речи во времени; вторая отображает спектрограмму, которая показывает сбой частот речевого сигнала с течением времени. Оба этих графических дисплея могут быть увидены, но не услышаны, и из-за этого может быть эффективным устройством в судебном разбирательстве.

Однажды поступила угроза о бомбе. По этому сообщению был обвинен афроамериканец, родившийся и выросший в Северной Каролине. Сторона защиты привлекла вычислительного судебного лингвиста в качестве эксперта-свидетеля для проведения анализа идентификации говорящих. Анализируя многие сегменты речи, эксперт определил, что подозреваемый вряд ли мог быть тем, кто оставил эту угрозы бомбы, но также и высокую вероятность того, что говорящий не был носителем английского языка. Вот выдержка из отчета эксперта-свидетеля:

The word «goodbye» occurs in the bomb threat.

Caller: Inserts an epenthetic vowel so that the pronunciation is «good-a – bye», clearly seen in the waveform and spectrogram. No native speaker of English is likely to have this pronunciation. The caller also pronounces «bye» with a fully diphthongized /aɪ/—the way foreigners are taught.

Suspect: His «goodbye» is «goobah», without the /d/ and certainly without the epenthetic vowel. His «bye» is monophthongized and somewhat lengthened as in much speech of the south, black and white.

Эксперт взял образцы от подозреваемого и поместил осциллограммы бок о бок. Фигура слева – это «goodbye» осведомителя. Рисунок справа – это подозреваемый. Ясный звук /d/ осведомителя начинается с 0,40 с остановкой завершения (тишиной) около 80 миллисекунд. Амплитуда небольшая, но не ноль из-за шума. Эвентуальная гласная видна между 0,52 и 0,64 секунды. На 0,64 секунды начинается остановка закрытия для /b/ «bye». Справа нельзя увидеть /d/ и нет дополнительного гласного. За «gooh» следует остановка завершения /b/ в слове «bye» через 2,55 секунды.

На стороне осведомителя почти тишина остановки завершения /d/ легко проявляется в белом пространстве около 0,40 секунды, а эвентуальная гласная, вероятно, ([ə]) «schwa» хорошо видна как гласная. Расходящиеся первый и второй форманты в конце звонящего делают дифтонг слова «good-a-bye» видимым.

В спектрограмме подозреваемого нет вообще никаких признаков /d/, ни дополнительного гласного. Единственный период молчания предшествует остановке завершения /b/ в слове «bye». В конце концов, плоскость всех формантов в гласной слова «bye» показывает монофональный звук, совсем не похожий на звонившего.

Таким образом, компьютерная лингвистика является прикладной наукой, которая изучает реализацию компьютерных, языковых процессов, отношение компьютера и человека. Также компьютеры помогают ученым переводить литературу, делать языковые анализы, находить необходимые материалы из большого корпуса, решать правовые и криминальные вопросы.