Компьютерная лингвистика №1: зачем лингвисту компьютер

Рассказывает Светлана Тимошенко

из Лаборатории компьютерной лингвистики ИППИ РАН

(URL: https://www.youtube.com/watch?v=kxSH7UTheLE)

Ну... Что такое компьютерная лингвистика?

Нетрудно догадаться, что это компьютеры + лингвистика. Все мы знаем, что, когда появились компьютеры, мы стали знакомые вещи делать немножечко иначе. Появились сериалы – это такие очень длинные фильмы. Чаты – это такая либо мгновенная переписка, либо разговор в письменной форме. Всё то же самое, но немножечко иначе.

С лингвистикой произошло то же. Лингвисты продолжают изучать те же самые вещи, которые они изучали всегда, но делают это немножко иначе.

Что вообще изучают лингвисты? Язык и буквально всё, что в нём есть. Вот в школьных учебниках были разделы: морфология, синтаксис, лексикология, семантика. Это наши темы. И если говорить о том, что остаётся от наших исследований, то результатами изучения лексики и семантики становятся словари, а результатом работы морфологов и синтаксистов – правила. Правила обычно человек видит в учебниках.

Что же для такой работы может сделать компьютер? Оказывается, очень многое. Лингвист обобщает свои наблюдения над языком. И вот раньше, чтобы выяснить, есть слово или нет, надо было снимать с полки книжки, читать газеты, проверять собственные впечатления, спрашивать кого-то. А теперь задавать подобные вопросы лингвисту достаточно обратиться к уже готовой коллекции текстов. **Такая коллекция текстов называется корпус**. К ней, как правило, идёт **программа поиска**, т.е. отправил запрос – и тут же получил результат. Мы делаем всю ту же работу, что и раньше, только гораздо быстрее.

Но есть и содержательное приращение, потому что теперь мы не просто можем убедиться, что слово есть или его нет, мы можем увидеть, что слова есть в разной степени – нам встретилось 5 примеров или 5 тысяч примеров. Колоссальная разница! Но вместе с тем компьютер, а главным образом Интернет, поставил перед лингвистами новые задачи. Вот, например, что такое Интернет? Прежде всего поиск. А как поисковые сайты ищут? Даже можно сказать поисковые машины, потому что за той

картинкой, которую мы видим на сайте, стоит некоторый алгоритм и тоже, значит там, база данных. А ищет он, в том числе, по словам. И вот – представим себе – что пользователь написал *лук* в поле запроса. Нужно решить, **какие документы ему показывать**. Слово многозначное. Значит, прежде чем сформировать выборку, мы должны понять: в тех документах, которые у нас есть, вот это слово в том же смысле, в котором его написал пользователь, или в каком-то другом.

И вот так появляется **задача разрешения неоднозначности**, причём автоматической. Нам нужно сделать программу, которая будет смотреть на контекст и по контексту определять: вот в этом тексте в каком значении слово. Источниками наборов значений выступают старые словари. Но не только. Всё время появляются новые вещи, новые слова. Значит, в идеале наша **программа должна не только выбирать для этого текста какое-то значение из уже имеющихся, но выбирать значение для каких-то новых слов**. Автоматическим разрешением неоднозначности задачи компьютерной лингвистики, собственно, не ограничиваются. Кроме того, есть машинный перевод. **Машинный перевод** – понятно, что тут нужно вообще всё: и морфология, и семантика, и синтаксис, никак без него не обойтись.

И вот из этих кусков и строится компьютерная лингвистика.