При поддержке Российского гуманитарного научного фонда, проект 15-03-00812

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА КАК ТРЕНД В ОБЛАСТИ СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Р. И. Лоскутов
Томский политехнический университет
e-mail: mxhpns@mail.ru

COMPUTER LINGUISTICS AS A TREND IN THE MODERN INFORMATION TECHNOLOGIES

R. I. Loskutov
Tomsk Polytechnic University

В настоящее время все больше ученые и крупные компании сосредотачиваются на разработке систем искусственного интеллекта: это и распознавание изображений в видеопоследовательности с использованием технологии CUDA [1], и технологии распознавания речи (например, используемые разработки компании Google), и программирование ботов (программных роботов), которые предугадывают действие игрока (используются, например, в шахматах, в 3D-играх), и конечно, искусственные нейронные сети, исследования нечеткой логики. Но в настоящем докладе внимание сосредоточено на перспективном направлении в области технологий искусственного интеллекта — это компьютерная лингвистика.

В нашем мире появляется все больше и больше информации. Часто поиск необходимой крупицы знаний может не дать нужных результатов. Так, например, поисковые сервисы интернета, не способны анализировать и выдавать ответ частного на основе общего высказывания. То есть необходима система, способная анализировать информацию, полученную из текста, и отвечать на поставленные вопросы. При этом система должна быть интерактивной.

Этим уже долгие годы занимались отчаянные исследователи, но сейчас это уже не кажется невозможным. Существует множество работ, посвященных данному вопросу: книга Мельчука «Русский язык в модели СМЫСЛ-ТЕКСТ», «Семантико-синтаксический анализ естественных языков» И. В. Смирнова и А. О. Шелманова. Также существуют уже готовые системы извлечения информации из текстов: система «ABBYY Compreno», «ЭТАП-3», «Xerox XLE».

Эти идеи открывают невероятные возможности. Скажем, Вам необходимо вычислить циклический интеграл. Вы просто обращаетесь к компьютеру с целью получения ответа с решением. Или, необходи-

мо выполнить сложную последовательность действий на компьютере — это будет сделано за Вас. Вот почему эти новые развивающиеся технологии так привлекательны для исследователей.

Систему, о которой идет речь, можно представить в виде блока, имеющего на входе текст от пользователя, а на выходе — выполняемые действия и ответы пользователю на его вопросы. Для существования данной системы необходимо использовать методы компьютерной лингвистики, которые позволяют выполнить графематический, синтаксический, семантический анализ поступающего на вход текста с целью извлечения информации из данного текста. После чего он представляется в памяти персонального компьютера в виде различных структур (деревьев, графов). Далее полученная информация анализируется в пассивном режиме (без участия пользователя). Недостаточные сведения могут быть получены в интерактивном режиме.

Это является логическим продолжением в развитии информационных технологий. Исследования, связанные с рентабельностью подобных проектов, были проведены на основе эмпирических данных, полученных из различных источников. Так, скажем, компания ABBYY до сих пор финансирует проект «compreno», а значит это имеет смысл с точки зрения основного предположительного результата действия компании — извлечения прибыли.

Альтернативой методам компьютерной лингвистики (так называемый лингвистический подход к представлению и обработке текстов) является статистический подход, заключающийся в том, что «текст представляется как упорядоченное множество последовательностей символов (слов), а обработка текстов сводится к статистической обработке встречаемости слов», но несмотря на его большую распространенность и возможность решать многие задачи обработки текстов, он принципиально не позволяет на высоком уровне качества решать многие задачи, такие как машинный перевод, извлечение фактов и т. д. Лингвистический подход, напротив, позволяет решать многие задачи обработки текстов на более высоком уровне.

Рассмотрим наиболее важные проблемы, появляющиеся на этапе разработки систем автоматизированной обработки текстов.

Первая из них — это явление различного рода омонимии, присущее всем существующим естественным языкам (русский, английский и т. д.). Такой проблемы не возникает когда дело касается формальных языков (языков программирования, например). Часто используемый пример, который подчеркивает данную проблему, следующий: «Эти типы стали есть на складе». В данном высказывании компьютеру надо будет определиться с тем, о чем идет речь: о том что определенные люди устроили пир на складе, или о том, что определенные типы

металла имеются на складе. То есть, в данном предложении словоформа «СТАЛИ» может относиться как к лексеме «СТАЛЬ» (сплав металлов), так и к лексеме «СТАТЬ» (глагол); а словоформа «ЕСТЬ» может интерпретироваться как глагол (кушать или существовать, быть) или же как частица (или военная команда). В связи с этим возникает дополнительная проблема: необходимость в анализе контекста, присущему данному отрывку.

Вторая проблема, не менее значительная, относится к тому, что согласно современным исследованиям синтаксис неразрывно связан с семантикой языка, но большинство существующих на данный момент разработок анализаторов текста строго разделяют фазу синтаксического анализа и фазу семантического анализа, что связано с основными принципами технологии программирования (крупные проекты разрабатываются так, чтобы разделить всю задачу на обособленные, не связанные между собой, что улучшает сопровождаемость разработок и ускоряет процесс кодирования за счет привлечения отдельных групп людей, каждая из которых занимается только своей, менее крупной, задачей и не вникает в подробности других задач), но вместе с этим увеличивает продолжительность анализа текста и не учитывает факт неразрывной связи синтаксиса и семантики языка.

Отметим также то, что компьютер не является человеком, а значит у него не может быть понятия красоты, нет возможности ориентироваться в нашем мире, нет понятий духовности и т. д. А значит, на вопросы типа «Тебе нравится такая-то игра?» или «Как твои дела?» ответ будет надуманным, зависящим от реализованных идей разработчиков по этому поводу.

Подводя итоги, заметим, что все перечисленные проблемы решаемы и можно утверждать, что существование полностью готовых систем подобного рода является вопросом времени и привлечет значительную прибыль для организаций, занимающихся разработками в этой области.

Литература

1. Что такое CUDA? [Электронный ресурс]. — Режим доступа: http://www.nvidia.ru/object/cuda-parallel-computing-ru.html.