Алгоритмы совершенствования чат - ботов в системах повышения эргономичности сайтов

Овсяник Андрей Павлович

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Анализ и понимание текстов на естественном языке занимали важную нишу в исследованиях по искусственному интеллекту с момента возникновения этой науки. В течение многих десятилетий работы в этом направлении были сосредоточены в областях, называемых "вычислительной лингвистикой" и "обработкой естественного языка". Целью вычислительной лингвистики (ВЛ, computational linguistics, CL) является "построение логико-лингвистических моделей и соответствующих им алгоритмов и программ". В отличие от ВЛ, где большое значение имеет теоретическая лингвистическая корректность и адекватность предложенных моделей, обработка естественного языка (ОЕЯ, natural language processing, NLP) сосредоточена "на моделировании всего того, что изучает лингвистика в целом".

Первоначально обе этих области, преимущественно, стремились построить точные языковые модели, основанные на формальном синтаксическом и семантическом анализе; этот подход получил название "глубинная ОЕЯ", deep NLP, DNLP (иногда применяется термин "символическая ОЕЯ" — symbolic NLP). Потребность в точных моделях была обусловлена тем, что одной из первых задач для являлся автоматический машинный перевод. Исследования в направлении глубинной ОЕЯ стимулировались развитием генеративной лингвистики, начало которой было положено с выходом в 1957 году работы Н. Хомского "Синтаксические структуры.

Со времени появления в 1990-х годах понятий "обнаружение знаний" (knowledge discovery) и "интеллектуальный анализ данных" (ИАД, data mining), исследователи обнаружили, что для многих практических задач, в первую очередь информационного поиска (в том числе поисковых машин масштаба всемирной сети Интернет и поиска для социальных сервисов), удовлетворительные результаты могут быть получены более простыми и вычислительно эффективными способами — на основе статистических данных

о тексте. Этот подход называется "поверхностная ОЕЯ", shallow NLP, SNLP (иногда — "эмпирическая NLP" — empirical NLP).

Современная область интеллектуального анализа текстов (ИАТ, text mining) является динамично развивающимся и практически востребованным направлением ОЕЯ, основанным на применении методов ИАД и машинного обучения (machine learning).

Ключевыми группами задач ИАТ можно назвать:

– классификацию (распределение текстов по заранее заданным группам) и кластеризацию (распределение текстов по группам, которые должны быть определены в процессе работы алгоритма);

– извлечение информации (information extraction, идентификация фактов и метаинформации о них в тексте) и информационный поиск;

– обнаружение возникающих трендов (emerging trend detection) — идентификация новых тем в коллекции текстов, соответствующих новым явлениям, событиям, предметам и т. д.

Важным параметром систем ИАТ является качество средств просмотра (browsing) результатов, в том числе визуализации данных и навигации, поэтому проектирование моделей уровня представления (presentation layer) можно считать четвёртой группой задач ИАТ.

Первые системы ИАТ, относящиеся к 1990-м годам, основывались на вычислительно несложных методах, разрабатываемых для каждой отдельной задачи. С развитием техники и появлением дешёвых вычислительных ресурсов, в том числе за счёт организации распределённых вычислений, в область ИАТ были привлечены более сложные методы теории вероятностей и статистики. В частности, в настоящее время зарубежные исследования по обработке текстов, преимущественно, связаны с так называемым тематическим.

Классическая задача обнаружения трендов касается обработки архивных данных за большой период времени, т. е. не подразумевает оперативную обработку текстовых документов по мере их поступления

В то же время нельзя утверждать, что ранние, более простые методы анализа данных и текстов не применимы в современных системах. Исследователями искусственного интеллекта было показано, что в ряде случаев применение интуитивных, нечётких и приближённых методов обладает существенными преимуществами в плане эффективности и прозрачности систем.