Ю. Б. Крапивин

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА КАК НАПРАВЛЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Как показывает современная история развития научного познания мира, наиболее интересные и значимые результаты достигаются в научно-исследовательских предприятиях, объединяющих отдельные дисциплины в рамках некоторой глобальной научной проблемы. Такое объединение имело место, например, в 50-е и 60-е гг., когда возникло новое научное направление – искусственный интеллект, объединившее усилия математиков, психологов, специалистов в области робототехники, электроники, с тем чтобы научить ЭВМ в определенном смысле думать и вести себя подобно человеку (естественному интеллекту).

Обработка естественного языка (ЕЯ) – одно из наиболее динамично развивающихся направлений искусственного интеллекта в настоящее время, связанное с решением задач автоматической обработки информации, представленной на естественном языке. Центральными научными проблемами здесь являются: моделирование процесса понимания смысла текстов (переход от текста к формализованному представлению его смысла), а также синтез текста и речи (переход от формализованного представления смысла к текстам на естественном языке).

Существуя уже более полувека, оно в своем развитии опирается на результаты в области общей лингвистики, в том числе фонологии, морфологии, синтаксиса, семантики и прагматики, а также математики, информатики (computer science), психологии, социологии и др. Это, в свою очередь, означает междисциплинарный характер проводимых исследований, связанных с решением следующих прикладных задач.

1. Автоматизация инженерии знаний. Если рассматривать текст как источник знаний, то можно утверждать, что каждому типу знаний соответствуют определенные структурные единицы текста и отношения между ними, выраженные средствами естественного языка. Так, объектам соответствуют именные группы, фактам – отношения, выражаемые тройками «субъект-акция-объект», правилам – причинно-следственные отношения между фактами, выражаемые различными средствами языка. В то же время объекты, факты и правила в каждом отдельном случае становятся вполне

конкретными по своим атрибутам и отношениям, и, таким образом, речь тогда может идти в дополнение к основным о так называемых атрибутивных знаниях, таких как, например, знания о местоположении объекта, его параметрах и т.д.

2. Информационный поиск, включая автоматическое индексирование документов и запросов пользователя, реализуется с помощью систем информационного поиска. Их работа основывается на использовании процедуры индексирования, требующей лингвистической обработки и имеющей своей целью получение формального представления (поискового образа) и запроса, и документов, а также процедуры сравнения последних согласно определенному правилу – модели поиска с целью определения степени их соответствия (релевантности). В настоящее время наблюдается тенденция усиления лингвистической составляющей в системах информационного поиска. С целью повышения качественных показателей их работы необходимо осуществлять поиск во множестве документов, представленных на различных ЕЯ, а не только на языке документа-запроса (так называемая cross-language функциональность), что требует в свою очередь решения таких «лингвистически нагруженных» задач, как определение языка текстовых документов и их машинный перевод.

3. Задача автоматического определения языка текстовых документов заключается в определении языка, на котором написан документ или его часть. Решения, как правило, ориентированы на использование знаний о ЕЯ в пределах от уровня алфавита до лексико-грамматического уровня глубины ЕЯ, что вполне приемлемо по трудоемкости для их использования в промышленных системах автоматической обработки текста, а также ориентированы на возможности, предоставляемые методами машинного обучения, в том числе на базе искусственных нейронных сетей.

4. Машинный перевод (МП) текстовых документов с одного ЕЯ на другой (другие). При решении данной задачи в настоящее время применяются:

• статистические методы, в том числе и методы, основанные на примерах, суть которых состоит в том, что на основании параллельных корпусов текстов производится вычисление соответствия друг другу лексических единиц различных языков и их статистических характеристик, а также построение на их основе некоторой модели перевода, причем качество получаемого перевода напрямую зависит от объемов этих корпусов;

• лингвистические методы (методы, основанные на правилах), предполагающие построение модели перевода в соответствии с набором лингвистических правил, которые определяют нужную глубину анализа текста, а также возможную трансформацию грамматической структуры входного текста в эквивалентные ей структуры выходного текста;

• нейросетевые методы реализуются на базе различных архитектур искусственных нейронных сетей и алгоритмов их обучения, а также корпусов текстов.

5. Автоматическая классификация текстовых документов. Она включает автоматическую категоризацию, т.е. распределение документов по заранее созданным категориям; кластеризацию – распределение документов по автоматически генерируемым группам или иерархиям групп; генерацию таксономий, т.е. создание иерархий концептов или тематических категорий. Решения также используют методы машинного обучения, в том числе на базе искусственных нейронных сетей, использующих тематические, формальные, структурные признаки текста и его составляющих.

6. Автоматическое реферирование текстов. Подразумевает автоматическое выделение на основе анализа текста наиболее важной с определенной точки зрения информации из документа и представление ее пользователю в том или ином виде. Это может быть часть оригинального текста, набор отдельных его предложений, выделенных по определенному критерию, информация, соответствующая основным типам знаний – реферат в виде списка объектов (ключевых слов) или фактов, в виде иерархии объектов и т.д. Решения, как правило, реализуют подходы на основе извлечения предложений – используется оценочная функция важности информационного блока (предложения), которая рассчитывается на основании статистических параметров текста (например, по частоте встречаемости слов в тексте), или извлечения содержания – генерация реферата с порождением нового текста, содержательно обобщающего первичный документ или документы.

Для решения вышеперечисленных, а также связанных с ними задач (автоматическое выделение именованных сущностей, автоматический анализ тональности текста, автоматический синтез текста, организация вопросно-ответного и диалогового взаимодействий с пользователем на ЕЯ) требуется проведение автоматического лингвистического анализа текста, который в современных информационных системах, как правило, реализуется с помощью лингвистических процессоров, обеспечивающих следующие основные уровни такого анализа:

1) предварительное форматирование текста (преформатирование);
2) лексический анализ;
3) лексико-грамматический анализ;
4) синтаксический анализ;
5) семантический анализ.

Уровень глубины анализа текста ЕЯ, как и ЕЯ-запроса пользователя (в задачах информационного поиска, организации вопросно-ответного и диалогового взаимодействий с пользователем на ЕЯ), в каждом конкретном случае зависит от особенностей целевой задачи, в том числе и от необходимых критериев оценки ее качества решения и их значений.

Что касается приложения практических результатов рассмотренных прикладных задач, то качественное решение требуется во многих сферах бизнеса: это и контактные центры организаций, которым требуется оперировать большим потоком входящих запросов (автоматически разделять их на категории, определять темы, подбирать варианты ответов, а также информи-

ровать конечных пользователей с помощью чат-ботов), торговые и информационные интернет-площадки заинтересованы в качественном и оперативном поиске по своим каталогам, внедрении диалоговых и рекомендательных систем, а в сфере маркетинга и PR имеет место освещение деятельности компании в медиа и отслеживание, какой образ создается у аудитории.

Важной характеристикой проводимых исследований в рамках направления по обработке естественного языка особенно актуальной в современных экономических условиях является низкая стоимость материально-технической базы (в отличие от робототехники и машиностроения) при сопоставимой отдаче от результатов внедрения готовых технологических решений.