ВЫСТУПЛЕНИЕ ЗАМЕСТИТЕЛЯ ДИРЕКТОРА ФИЦ "ИНФОРМАТИКА И УПРАВЛЕНИЕ" РАН ДОКТОРА ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК Г.С. ОСИПОВА

Материал поступил в редакцию 03.12.2018 г.
Принят к публикации 25.12.2018 г.

Ключевые слова: лингвистический анализ, коммуникативная грамматика, синтаксема, бинарные отношения, индуктивное машинное обучение, контекстные правила.

DOI: https://doi.org/10.31857/S0869-5873894379-380

В глобальных информационно-телекоммуникационных сетях сегодня циркулирует $4 \cdot 10^{16}$ байт (40 петабайт) неструктурированной и полуструктурированной информации, главным образом текстов. В связи с этим возникают как минимум две группы задач: анализ отдельных текстов и анализ больших массивов, состоящих из сотен миллионов текстов.

Большинство подходов, которые используются в системах, работающих с массивами текстов, основано на информационных измерениях текста и не учитывает его лингвистической природы. Вместе с тем существует целый ряд методов и моделей серьёзного лингвистического анализа текста, а извлечение из текста полезной, то есть осмысленной информации, должно основываться именно на его семантическом анализе. Тем самым мы сталкиваемся с проблемой, и её решением может стать такой подход, который, с одной стороны, не будет игнорировать лингвистической природы текста, а с другой стороны, позволит создавать в ходе анализа структуры, пригодные для использования математических методов.

В качестве основы метода, интегрирующего лингвистические и статистические подходы для анализа больших массивов текстов, нами была выбрана так называемая коммуникативная грамматика русского языка, разработанная в Институте русского языка им. В.В. Виноградова РАН. Ключевая идея данной теории состоит в том, что семантика выражается через синтаксис, и это делает её не предметно-ориентированным, а универсальным инструментом анализа. Согласно главной гипотезе, в самом тексте содержатся минимальные смысловые единицы, "атомы" смыслов, которые нужно уметь выделять. Эти единицы получили название "синтаксемы", каждая из них имеет некоторое категориальное значение.

На множестве категориальных значений синтаксем нами определено семейство бинарных отношений. Смысл всего высказывания в таком случае есть множество категориальных значений синтаксем с заданным на нём семейством бинарных отношений. В результате построен образ высказывания – алгебраическая система с множеством синтаксем в качестве основного множества, семейством бинарных отношений на множестве синтаксем и процедурами интерпретации дуг, позволяющими реализовать аргументационные рассуждения.

Кроме того, нами были применены методы индуктивного машинного обучения для снятия семантической неоднозначности – неоднозначности значений синтаксем. Снимать семантическую неоднозначность позволяют контекстные правила, полученные с помощью методов машинного обучения. Другими словами, контекст, окружающий данную синтаксему, позволяет уточнить её значение. Приведём в качестве примера одно из таких правил: "Если встречается синтаксема в падеже <родительный> с предлогами <из, изо>, имеющая категориальный класс <локатив>, а рядом с ней встречается синтаксема в падеже <именительный>, имеющая категориальный класс <личное>, то первая синтаксема имеет категориальное значение <точка начала движения>". Всего автоматически выверено 600 таких правил.

В итоге построены точное отображение множества синтаксем в множество их значений и семейство бинарных отношений на множестве значений синтаксем. Эта структура названа нами "неоднородная семантическая сеть". Она легла в основу технологии, с помощью которой можно, в частности, анализировать отдельные тексты.

Область применения предложенного метода довольно широка. Можно извлекать информацию, например, из медицинских текстов о лекарственных методах лечения, такой анализ используется для построения медицинских баз знаний. Ещё одно направление – анализ текстовой информации с целью мониторинга военно-политической напряжённости в различных регионах. Наконец, метод востребован в наукометрических

исследованиях. В целом можно ставить и решать задачи преобразования неструктурированного текста в структурированное представление – в таблицы, графики и т. д. На этом пути удалось решить задачи релевантного семантического поиска по запросу на естественном языке (www.exactus.ru), автоматического выявления научных коллективов, научных направлений и динамики публикационной активности по направлениям на основе анализа первичных научных текстов, анализа патентной информации, оценки качества научных публикаций, обнаружения семантических дефектов, проверки соответствия структуры публикации требованиям журналов, выявления авторских терминов и описания результатов и ряд других задач. Реализованы системы с указанной функциональностью (http://expert.exactus.ru, http://demo.textapp.ru). В настоящее время, исходя из предложенного подхода, исследуются методы выявления авторской картины мира.

SPEECH OF THE DEPUTY DIRECTOR OF THE FEDERAL RESEARCH CENTER INFORMATICS AND MANAGEMENT OF RAS, DOCTOR OF PHYSICAL AND MATHEMATICAL SCIENCES G.S. OSIPOV

Received: 03.12.2018
Accepted: 25.12.2018

Keywords: linguistic analysis, communicative grammar, syntaxeme, binary relations, inductive computer-assisted learning, contextual rules.