Бручес Елена Павловна

«Методы и алгоритмы распознавания и связывания сущностей для построения систем автоматического извлечения информации из научных текстов»

Диссертация на соискание степени кандидата технических наук

по специальности 05.13.17 – Теоретические основы информатики

Работа посвящена исследованию и разработке методов для решения нескольких задач автоматического извлечения информации из научных текстов: обнаружения сущностей, отношений между ними и связывания найденных сущностей с внешней базой знаний. Предложен метод извлечения научных терминов, основанный на частичном обучении и слабо зависящий от области знаний. Адаптированный метод извлечения отношений базируется на технике "обучения без примеров" (zero-shot learning) и позволяет решать задачу в условиях ограниченного количества размеченных данных. Перенос обучения моделей с английского языка на русский показал свою эффективность и может быть потенциально применим для широкого круга малоресурсных языков. Предложенный алгоритм автоматического связывания научных терминов с сущностями в базе знаний учитывает сущности не только верхнего уровня, но и вложенны. Проведены эксперименты, позволяющие оценить качество работы реализованных методов и алгоритмов. Для экспериментов был создан размеченный корпус научных текстов на русском языке, который находится в открытом доступе.