Введение

В современном мире объемы различных видов информации, которую приходится анализировать каждому человеку, неуклонно растут, в том числе это касается и оцифрованной текстовой информации. В связи с этим возникает необходимость разработки эффективных методов анализа текстовой информации, переведенной из «бумажной» в электронную форму. К подобным задачам относится, например, поиск в электронной коллекции документов по запросу, сформулированному на естественном языке [52], компьютерный перевод текста с одного языка на другой [155], классификация оцифрованных текстов по различным признакам (стилю написания, жанру и пр.) [85] и т. п. При этом накопленные сегодня объемы текстовой информации, переведенной в электронную форму, таковы, что их эффективное использование оказывается невозможным без автоматизированных и автоматических методов решения задач информационного поиска.

С технической точки зрения задача анализа оцифрованного текста является задачей анализа некоторых наборов последовательностей символов, что с формальной точки зрения не должно иметь каких-либо принципиальных трудностей. Однако на практике оказывается, что добиться сколько-нибудь значимых результатов удается только при условии хотя бы частичного выявления смысловых отношений[1] между словами анализируемых текстовых данных, то есть фактически понимания их семантики [105]. Необходимость выявления семантических отношений многократно увеличивает сложность задач текстового анализа, так как на современном уровне человеческого понимания механизмов восприятия и анализа текстовой информации не удается создать их полноценного алгоритмического описания. Например, задача установления смысла многозначных слов в зависимости от контекста их употребления, сформулированная еще при первых попытках создания систем машинного перевода и до сих пор не получившая окончательного решения, по праву считается одной из наиболее сложных задач для искусственного интеллекта [127]. В результате, несмотря на высокое быстродействие существующих вычислительных систем, на сегодняшний день не создано программных инструментов, обеспечивающих полноценное решение задачи семантического анализа текстов.

Один из подходов, позволяющий частично учесть смысловые связи между словами и потенциально повысить качество обработки текстовой информации, основан на использовании тезауру-

[1] Далее под *семантическим отношением* мы понимаем связь слова с другими словами, входящими вместе с ним в одну семантическую систему, то есть составляющие тематические объединения слов в группы.

сов — разновидностей словарей, отражающих семантические (синонимические, антонимические, родовидовые и пр.) отношения между словами [123]. Проблемы построения и использования тезаурусов, в том числе и электронных, исследовались в трудах российских и зарубежных ученых, таких как И.В. Азарова, Л.Г. Бабенко, П.И. Браславский, Б.В. Добров, Ю.Н. Караулов, Н.В. Лукашевич, В.В. Морковкин, А.С. Нариньяни, C. Fellbaum, G.A. Miller, P. Vossen и других авторов.

Сегодня существуют электронные тезаурусы (ЭТ) для английского языка (например, Princeton Wordnet [103]), для семи европейских языков (EuroWordNet [156]) и др. В тоже время открытых русскоязычных ЭТ, полностью удовлетворяющих требованиям пользователей [18, 39], несмотря на многочисленные попытки их создания [2, 18, 23, 75], сегодня не существует.

ЭТ традиционно создается вручную группой экспертов [50], которые формируют словник, выделяют концепции, включаемые в ресурс, устанавливают семантические отношения между терминами ЭТ. Аналогичные работы приходится выполнять при создании не только ЭТ, но также толковых словарей и словарей синонимов. Принимая во внимание объемы выполняемых при этом работ и количество привлекаемых специалистов (см., например, [51, 69]), становится понятной необходимость разработки подходов, обеспечивающих хотя бы частичную автоматизацию процесса создания ЭТ.

Наиболее очевидный подход, реализующийся переводом существующих ЭТ на другие языки [23], на практике оказывается неработоспособным, поскольку положенное в его основу априорное предположение о том, что при переводе сохраняются все отношения между словами в языке оригинала, оказывается неверным. Более того, в каждом естественном языке существует безэквивалентная лексика [158]. В этой связи попытки создания ЭТ русского языка путем автоматического перевода ЭТ английского языка Princeton Wordnet не увенчались успехом, так как качество созданных продуктов оказалось весьма низким [23, 39].

В то же время некоторые типы семантических отношений удается достаточно эффективно выделять автоматически [132]. Однако качество получаемых при этом результатов оказывается столь низким, что формирование базы семантических отношений без участия экспертов оказывается невозможным. Это подтверждается в том числе итогами соревнований по определению семантической близости пар слов русского языка (конференция по компьютерной лингвистике «Диалог 2015» [133]). Анализ показал, что в среднем только 70–75% автоматически полученных результатов упорядочивания пар слов по убыванию семантической близости коррелировали с аналогичными резуль-

татами, полученными экспертами. Таким образом, разработка новых подходов к решению задачи установления семантических отношений, в том числе подходов, обеспечивающих уменьшение объемов работ, выполняемых экспертами, является актуальной.

Понятно, что потенциально можно уменьшить долю участия экспертов при создании электронного лексического ресурса не только за счет автоматизации тех или иных этапов разработки ЭТ, но и за счет реинжиниринга процессов создания ЭТ. Например, возможно использовать краудсорсинг, т. е. привлекать для выполнения однотипных заданий большое число заинтересованных потенциальных пользователей. Анализ опыта использования краудсорсинга показывает, что данная технология успешно применяется при решении самых разных задач — например, при разметке очертаний кратеров на снимках планет [151]; получены подтверждения эффективного использовании краудсорсинга при решении задач анализа оцифрованных текстов (см., например, [89]); существуют лексические вики-ресурсы (например, Викисловарь [64]), качество которых сравнимо с источниками, созданными экспертами-специалистами [70].

Однако, несмотря на достигнутые успехи при использовании технологии краудсорсинга, на сегодняшний момент не существует единого мнения по поводу возможности его использования при создании ЭТ русского языка. В этом контексте мы провели целенаправленное исследование автоматизированных методов выявления семантических отношений для ЭТ русского языка, представленное в данной монографии, которая имеет следующую структуру.

В *первой главе* проведен анализ состояния предметной области, в том числе рассмотрены проблемы структурированного представления текстовой информации в бумажной и электронной формах и существующие подходы к их решению; лексико-семантические программные инструменты (тезаурусы Princeton WordNet и EuroWordNet, наиболее известные ЭТ русского языка — проект RussNet, тезаурусы РуТез, Russian WordNet, YARN); подходы к оцениванию качества ЭТ. Сделаны обоснованные выводы об отсутствии на данный момент открытого ЭТ русского языка, обладающего достаточными полнотой и качеством данных, о необходимости разработки количественных методов оценивания качества ЭТ и уменьшения трудозатрат экспертов при ручном наполнении ЭТ за счет автоматизации процесса создания ЭТ русского языка.

Во *второй главе* обоснованы характеристики, обеспечивающие количественное оценивание полноты ЭТ русского языка, предложены подходы, базирующиеся на использовании корпусов текстов и словников толковых словарей, обоснованы количественные характеристики для оценивания качества синонимических рядов (данные характери-

стики можно использовать применительно не только к ЭТ, но и к любым лексическим ресурсам, которые включают в себя синонимические ряды), предложен способ оценивания полноты переводных ЭТ. Для оценки адекватности введенных характеристик проведен анализ существующих русскоязычных лексических ресурсов (ЭТ, толковых словарей, словарей синонимов, Русского Викисловаря). Данный анализ проведен посредством использования полностью автоматического метода исследования ЭТ русского языка, разработанного авторами.

В *третьей главе* на примере тезауруса YARN и Русского Викисловаря проведен анализ синонимических рядов, в результате чего сделаны обоснованные выводы о существовании типовых лексикографических проблем. Для их устранения был разработан автоматизированный метод выявления эквивалентных синонимических рядов, состоящий из нескольких последовательных этапов: автоматического анализа данных и их последующей обработки с помощью краудсорсинга. Для оценивания эффективности предложенного метода выявления эквивалентных синонимических рядов были предложены характеристики, широко используемые в информационном поиске, но не применявшиеся ранее для оценивания качества синонимических рядов: точность и полнота. Также обсуждаются результаты экспериментальных исследований, подтверждающие гипотезу о возможности повышения качества синонимических рядов за счет использования краудсорсинга при решении задач анализа оцифрованных текстов. Проводится анализ результатов использования предложенного метода для извлечения синонимических рядов из существующих открытых данных, качество которых оказалось сравнимым с качеством синонимических рядов, составленных экспертами.

В *четвертой главе* описан предложенный авторами автоматизированный метод установления родовидовых отношений между существительными, основанный на выделении родового понятия из определений толковых словарей. Как было показано, такое выделение возможно сделать, применяя регулярные выражения. В основу разработанного метода положены результаты анализа определений толковых словарей, которые позволили выявить и потом описать регулярные выражения, позволяющие осуществлять автоматический поиск родовых понятий из текстов определений. Описана программная реализация метода, работающая с определениями Малого академического словаря [51]. Исходный код данной программы выложен в открытый доступ, причем этот код, как показано в главе, может быть легко адаптирован и для других толковых словарей.

В *заключении* подводятся итоги проведенного исследования.