Моделирование и анализ информационных систем. Т. 25, № 4 (2018), с. 435–458
Modeling and Analysis of Information Systems. Vol. 25, No 4 (2018), pp. 435–458

Тезаурусы

DOI: 10.18255/1818-1015-2018-4-435-458

УДК 004.912

Русскоязычные тезаурусы: автоматизированное построение и применение в задачах обработки текстов на естественном языке

Лагутина Н. С., Лагутина К. В., Адрианов А. С., Парамонов И. В.

получена 1 августа 2018

Аннотация. В работе выполнен обзор существующих электронных русскоязычных тезаурусов и методов их автоматического построения и применения. Авторы провели анализ основных характеристик тезаурусов, находящихся в открытом доступе, для научных исследований, оценили динамику их развития и эффективность в решении задач по обработке естественного языка. Были исследованы статистические и лингвистические методы построения тезаурусов, которые позволяют автоматизировать разработку и уменьшить затраты на труд экспертов-лингвистов. В частности, рассматривались алгоритмы выделения ключевых терминов из текстов и семантических тезаурусных связей всех типов, а также качество применения получившихся в результате их работы тезаурусов. Для наглядной иллюстрации особенностей различных методов построения тезаурусных связей был разработан комбинированный метод, генерирующий специализированный тезаурус полностью автоматически на основе корпуса текстов предметной области и нескольких существующих лингвистических ресурсов. С использованием предложенного метода были проведены эксперименты с русскоязычными корпусами текстов из двух предметных областей: статьи о мигрантах и твиты. Для анализа полученных тезаурусов использовалась комплексная оценка, разработанная авторами в предыдущем исследовании, которая позволяет определить различные аспекты тезауруса и качество методов его генерации. Проведённый анализ выявил основные достоинства и недостатки различных подходов к построению тезаурусов и выделению семантических связей различных типов, а также позволил определить потенциальные направления будущих исследований.

Ключевые слова: тезаурус, семантические отношения, автоматическое построение тезауруса, автоматическое выделение связей, выделение ключевых слов

Для цитирования: Лагутина Н. С., Лагутина К. В., Адрианов А. С., Парамонов И. В., "Русскоязычные тезаурусы: автоматизированное построение и применение в задачах обработки текстов на естественном языке", *Моделирование и анализ информационных систем*, **25**:4 (2018), 435–458.

Об авторах:

Лагутина Надежда Станиславовна, orcid.org/0000-0002-6137-8643, канд. физ.-мат. наук, доцент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: lagutinans@rambler.ru

Лагутина Ксения Владимировна, orcid.org/0000-0002-1742-3240, студентка,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: lagutinakv@mail.ru

Адрианов Алексей Сергеевич, orcid.org/0000-0002-3073-0982, студент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: alex.a4.25@yandex.ru

Парамонов Илья Вячеславович, orcid.org/0000-0003-3984-8423, канд. физ.-мат. наук, доцент,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: Ilya.Paramonov@fruct.org

Введение

Тезаурус — это словарь терминов естественного языка, включающий в себя систему связей этих терминов [1]. Тезаурус может использоваться и как информационно-поисковый ресурс и как источник или эталон терминологии. Связи между словами являются материалом для построения лексико-семантических сетей для извлечения знаний, для определения семантической близости слов. Формализованность тезауруса позволяет легко автоматизировать его применение. Многие исследователи подчеркивают важность построения электронных тезаурусов и перспективу их использования в автоматических системах обработки текстов [2, 3].

Решая задачи в области автоматической обработки текстов на естественном языке, авторы обнаружили тот факт, что тезаурус является удобной моделью предметной области. Авторы успешно использовали автоматически сгенерированный тезаурус для построения альтернативной системы навигации для электронного туристического ресурса Open Karelia [4], а также для анализа тональности журнальных статей [5]. Следует отметить, что эти исследования проводились с использованием англоязычных текстов и применением соответствующих методов, алгоритмов и лингвистических ресурсов.

Попытка напрямую использовать разработанные подходы к анализу текстов на русском языке не дала столь же успешного результата. Самый поверхностный анализ ситуации показал низкое качество получаемых тезаурусов, в частности малое количество выделяемых связей между словами. Кроме того, применение в использованных алгоритмах внешнего электронного тезауруса РуТез как альтернативы англоязычного тезауруса WordNet существенно не повлияло на качество решения. Это побудило авторов провести анализ методов построения русскоязычных тезаурусов, а также более внимательно рассмотреть существующие электронные русскоязычные тезаурусы.

В 2015 году в работах [6, 7] был проведен краткий анализ русскоязычных тезаурусов. Авторы соответствующих обзоров отмечают недостаточный объём этих ресурсов, в частности по сравнению с WordNet, трудность интеграции с существующими алгоритмами и системами автоматической обработки текстов, сложность или невозможность использовать их, в том числе в коммерческих целях. Однако цифровые ресурсы могут достаточно быстро меняться, поэтому было бы интересно определить динамику и тенденции развития русскоязычных тезаурусов.

Существующие тезаурусы не могут полностью охватить весь спектр вопросов автоматической обработки текстов, особенно в предметных областях, поэтому задача построения новых тезаурусов и как самостоятельных ресурсов, и как вспомогательных инструментов является актуальной. К сожалению, авторам данной работы не удалось найти обзоры методов автоматизированного построения русскоязычных лексических ресурсов. Таким образом, одной из целей данной статьи стало описание

методов, применяемых при автоматизации построения тезаурусов, и определение направлений развития этих методов.

Представленная работа состоит из трех частей. В первой части проанализированы характеристики электронных тезаурусов русского языка, определена динамика развития этих ресурсов и оценены возможности их применения к задачам автоматической обработки текстов. Во второй части авторы рассматривают методы выделения терминов и связей между ними, применяемые современными исследователями при построении лексических ресурсов, чтобы определить основные тенденции и возможные направления развития в этой сфере, в первую очередь для построения тезаурусов. Третья часть описывает собственный опыт построения русскоязычных тезаурусов предметных областей. В заключении подведены итоги работы.

1. Обзор существующих электронных русскоязычных тезаурусов

Для английского языка существует бесплатный лингвистический ресурс WordNet (https://wordnet.princeton.edu), который можно назвать эталонным тезаурусом. Этот проект разрабатывается с начала 80-х годов XX века и активно используется в исследованиях. Для русского языка существует несколько проектов подобного рода.

1.1. Свойства и характеристики тезаурусов

Самым большим по объему является тезаурус RuTez. Этот проект разрабатывается Лабораторией информационных исследований во главе с Н. В. Лукашевич [8]. В его основу положены принципы построения WordNet, но модель описания сущностей отличается. Единицей тезауруса является понятие, снабженное набором терминов, значения которых соответствуют данному понятию. В качестве терминов могут выступать слова и словосочетания, количество которых может быть достаточно велико, например 20 и более. Слова и словосочетания, относящиеся к одному понятию, называются онтологическими синонимами.

Понятия связываются системой отношений. В RuTezе используются четыре типа связей: два вида иерархических отношений — «выше—ниже» и «часть—целое»; два вида ассоциативных отношений — симметричная ассоциация, связывающая понятия, очень близкие по смыслу, но не соединенные в одно понятие; и несимметричная ассоциация, связывающая два понятия, которые не могут существовать друг без друга, но не находятся в других отношениях.

На текущий момент тезаурус RuTez содержит 55 тысяч понятий, 158 тысяч слов и выражений, 210 тысяч отношений между этими понятиями. Как видно из таблицы 1, этот объем информации сопоставим с объемом WordNet по всем параметрам, кроме количества сущностей. Однако модель понятия RuTeza и модель синсета WordNet отличаются по своему смыслу, поэтому количественное сравнение в этом случае не вполне корректно.

Для некоммерческого использования по запросу можно получить тезаурус в формате XML. Кроме того, существует бесплатная версия RuTez-lite, содержащая 115

Таблица 1: Характеристики русскоязычных тезаурусов в сравнении с англоязычным тезаурусом Wordnet

Table 1: Characteristics of Russian-language thesauri in comparison with the English-speaking thesaurus Wordnet

Характеристика	РуТез	YARN	Wordnet.ru	RussNet	Wordnet
Основная единица	понятие	синсет	синсет	синсет	синсет
Количество единиц	55 тыс.	70 тыс.	20 тыс.	5,5 тыс.	117,7 тыс.
Количество слов, словосочетаний	158 тыс.	143,5 тыс.	100 тыс.	15 тыс.	155 тыс.
Виды отношений	иерархич., ассоц.	иерархич., антонимия	иерархич.	иерархич., антонимия, ассоц.	иерархич.
Число пар отношений	210 тыс.	134 тыс.	—	—	207 тыс.
Формат для использования	XML	XML	текстовые файлы	—	API
Последнее обновление	постоянно обновляется	постоянно обновляется	28.08.2008	14.06.2005	постоянно обновляется
Головная организация	Лаборатория информац. исследований	УрФУ, СПбГУ	авторский проект	СПбГУ	Принстонский университет

тысяч слов и выражений, которую можно изменять и копировать также в некоммерческих целях [9].

Разрабатывается и обновляется РуТез в первую очередь экспертами Лаборатории информационных исследований. Этот подход гарантирует качество тезауруса, но требует значительного времени для пополнения и изменения данных. Авторы обращают внимание на отзывы о своем проекте и учитывают существенные замечания. В целях интеграции с существующими мировыми онтологиями и тезаурусами на других языках в 2017 году РуТез был автоматически преобразован в формат WordNet. В результате создан тезаурус RuWordNet, содержащий 111,5 тысяч слов и выражений русского языка [10].

Еще один активный проект по созданию тезауруса русского языка — Yet Another RussNet (YARN) [11]. Разработчики используют модель полностью соответствующую WordNet, в основе которой лежат синсеты — группы синонимов и квазисинонимов, объединённые общим лексическим значением. Синсеты связываются между собой иерархическими отношениями и отношениям антонимии, устанавливаемым между противоположными по значению терминами. Дополнительно есть отношения между словами, а также межъязыковые связи между синсетами YARN и WordNet. В качестве начального наполнения была использована информация из Викисловаря.

Отличительной чертой данного проекта является организация дополнения тезауруса на основе краудсорсингового подхода: любой желающий после регистрации

на сайте YARN может участвовать в добавлении и редактировании данных. Авторы проекта декларируют контроль качества с помощью выделенного среди пользователей ядра редакторов, которые могут утверждать или отменять изменения, а также запрещать дальнейшее редактирование отдельных элементов тезауруса. Таким образом качество проекта напрямую зависит от квалификации выбранных редакторов.

В настоящий момент YARN содержит 143 508 слов, 69 799 синсетов, 104 906 неотредактированных пар синонимов, 29 764 неотредактированных родо-видовых отношений. Это меньше, чем в РуТез и WordNet, но если сравнить объем YARN в 2015 [7] и 2018 годах, то очевидна динамика роста. Этот факт, по-видимому, обусловлен применением автоматизации при добавлении данных из ресурсов, аналогичных Викисловарю, и краудсорсингового подхода.

Неоспоримым преимуществом проекта является то, что YARN распространяется по лицензии, разрешающей использовать материалы в исследовательских и коммерческих приложениях, копировать и изменять данные с условием ссылки на источник. Содержимое тезауруса предоставляется в XML-формате.

В открытом доступе находится еще один тезаурус, являющийся попыткой перевода WordNet — Wordnet.ru или Русский Wordnet [12]. Система построения ресурса полностью автоматическая и не предусматривает экспертной оценки. Авторы использовали эвристический алгоритм проверки соответствия синсетов английского языка синсетам русского, полученного в результате прямого перевода. Таким образом им удалось сгенерировать около 25 тысяч синсетов, 100 тыс. слов и словосочетаний.

Авторы разработали программу, визуализирующую Русский Wordnet, а текстовые файлы с данными к ней можно скачать в виде отдельного архива. К сожалению, дата модификации файлов в архиве последней версии — 28 августа 2008 года. Малый объем и отсутствие развития данного проекта показывают, во-первых, сложность построения полностью автоматического инструмента создания тезауруса, во-вторых, невозможность построения полноценных лексических ресурсов языка только на основе перевода без учета национальных особенностей.

Можно упомянуть ещё один электронный тезаурус, разработанный на кафедре математической лингвистики Санкт-Петербургского государственного университета — RussNet [13]. Модель ресурса полностью соответствует WordNet. Авторы создавали тезаурус, максимально совместимый с европейским многоязычным тезаурусом EuroWordNet, в рамках которого предложена система межъязыковых отсылок (Inter-Lingual-Index, ILI), дающих возможность переходить от терминов одного языка к сходным, но не обязательно тождественным словам другого. Следование этой системе позволяет использовать тезаурус для многоязычного поиска. Следует также отметить, что этот ресурс содержит самое большое количество разных типов связей, что обусловлено работой высококвалифицированных экспертов-лингвистов. К сожалению, закрытость данного проекта и отсутствие изменений с 2005 года сводит на нет все его преимущества. Однако в настоящий момент осуществляется интеграция RussNet и YARN [14]. Это позволит не только увеличить объем тезауруса, но и повысить его качество, так как RussNet создавался группой высококвалифицированных экспертов.

Среди открытых тезаурусов предметных областей авторам удалось обнаружить

только русскоязычную версию многоязычного тезауруса сельскохозяйственной терминологии AGROVOC [15]. Перевод AGROVOC на русский язык был выполнен специалистами Центральной научной сельскохозяйственной библиотеки в 2011 году.

Таким образом, в настоящий момент лингвистическими ресурсами, аналогичными англоязычному WordNet, для русского языка можно считать два тезауруса — РуТез и YARN. Все три ресурса сопоставимы по объему данных, как по количеству слов и словосочетаний, так и по количеству связей между ними. Преимуществом РуТеза, по мнению авторов, является качество, обеспеченное коллективом экспертов, а также разумный подход к развитию и преобразованию. Преимуществом YARN является его совместимость с международными лексическими ресурсами. Справедливости ради следует отметить, что разработчики РуТеза создали такую же совместимую версию — RuWordNet, но она пока меньше источника по объему. Также в качестве положительной стороны YARN нужно указать возможность его коммерческого использования.

1.2. Применение существующих тезаурусов к решению задач

Анализ электронных русскоязычных тезаурусов был бы неполным без обзора работ, использующих эти ресурсы для решения конкретных задач.

Существующие тезаурусы используются как источник терминологии. Авторы работы [16] используют РуТез как основу для построения общественно-политического тезауруса национального языка (татарского). В этом примере РуТез выступает в первую очередь как обычный филологический словарь тезаурусного типа. Авторы подчеркивают возможность использования такого рода лексических ресурсов в различных приложениях для автоматической обработки новостных документов, правовых актов или сообщений в социальных сетях. Поэтому полученный тезаурус преобразован и опубликован в облаке лингвистических открытых связанных данных — Linguistic Linked Open Data (LLOD) [17].

Следует отметить, что с целью передачи специфики татарской лексико-семантической системы авторы активно используют большое количество внешних источников. Такими источниками являются, во-первых, существующие двуязычные татарско-русские словари, в том числе специализированные общественно-политические, во-вторых, большое количество медиатекстов и текстов официальных документов. Текстовые документы предметной области необходимы, чтобы найти подходящие татарские термины для замены устаревших слов или добавления отсутствующих понятий.

Как источник терминов РуТез используется в работе [18]. Авторы решают задачу автоматической рубрикации текстов. Каждому понятию тезауруса ставится в соответствие рубрика текста. В исследуемых текстах ищутся слова, которые есть в РуТезе. На основе полученных данных строится модель тематического представления каждого текста и определяется рубрика. Эксперименты показали высокое качество результатов. Однако наличие узкопредметной рубрики потребовало дополнения тезауруса новыми терминами.

Структура тезауруса может быть использована как образец для построения удобного в использовании лингвистического ресурса. В полном соответствии со

структурой РуТез построен тезаурус по безопасности [19]. Этот тезаурус используется в специализированной информационно-аналитической системе, в том числе для автоматической текстовой классификации документов. Наполнение данными осуществляется из других источников: нормативно-справочная литература, специализированные текстовые коллекции, новости из средств массовой информации по тематике безопасности.

Связи между словами в тезаурусах являются ключевой информацией в методах расчета семантической близости терминов. В работе [20] сделана попытка автоматизировать оценку ответов учащихся на вопросы открытого теста. Анализ ответов на русском языке осуществлялся с использованием РуТеза и Википедии. С их помощью рассчитывалась семантическая близость слов ответа учащегося и эталонного ответа, расположение слов в тексте не учитывалось. По результатам экспериментов авторы отмечают, что качество работы системы сильно понижалось в случае появления синонимов, например, когда эталонный ответ содержал дублирующие варианты возможного ответа. Однако авторы используют только отношения «выше—ниже» и никак не учитывают синонимические отношения между словами.

Все виды связей из РуТеза используются автором работы [21] при расчете информации о схожести терминов. Это один из факторов схожести в разработанном методе построения групп семантически близких слов и выражений, описывающих тематические узлы кластера новостных сообщений. Метод был применен к задаче автоматического реферирования новостей.

Тезаурус может быть эталоном при оценке качества работы методов автоматической обработки текстов. Автор работы [22] предложил метод автоматического группирования семантически близких слов. Для оценки результатов работы использовались материалы РуТез и YARN. Качество результата было не очень высоким, особенно в случае сравнения с тезаурусом РуТез. Это объясняется тем, что метод автора базируется на графе синонимов, а синсет из YARN гораздо ближе к определению синонимов, чем единица тезауруса РуТез — понятие.

Из обзора работ видно, что тезаурусы используются в широком спектре исследований — РуТез чаще, YARN реже, — однако практически всегда совместно с другими лингвистическими ресурсами. В случаях анализа предметных областей эти тезаурусы можно применить очень ограниченно. В частности, нехватку терминов предметной области в использованных электронных ресурсах отмечают авторы работы [20].

Универсальные лексические ресурсы РуТез и YARN содержат полезную информацию, удобны для использования, кроме того, они динамично развиваются. Однако для решения огромного количества задач автоматической обработки текста нужны тезаурусы предметных областей. В силу сложности и трудоёмкости построения этих ресурсов их достаточно мало. Тем не менее, уровень развития автоматизированных методов построения тезаурусов делает эту задачу выполнимой, поэтому авторы хотели бы обратить внимание на актуальность построения открытых русскоязычных тезаурусов самых разных предметных областей.

2. Особенности автоматизированного построения русскоязычных тезаурусов

Задачу построения тезауруса можно разделить на две большие подзадачи: выделение ключевых слов и выделение связей между словами.

2.1. Выделение ключевых слов

Выделение ключевых слов — самая изученная часть процесса построения тезауруса, а также многих других задач автоматической обработки текста, в том числе на русском языке. Основную роль здесь играют статистические методы. Для английского языка способы выделения ключевых слов, включая предварительную обработку текста, хорошо исследованы и верифицированы [24]. Для построения русскоязычных тезаурусов применяются аналогичные методы, но количество исследований, тщательно оценивающих качество, значительно меньше.

При формировании РуТеза применялся автоматизированный способ получения терминологии [25]. Для каждой предметной области из рассматриваемого списка (математика, физика, химия, биология, геология) были сформированы коллекции документов (от 3 000 до 8 000 документов, объемом от 50 до 90 Мб каждый). Источниками коллекций являлись документы, доступные в интернете: материалы школьных уроков, рефераты, университетские лекции, материалы специализированных сайтов. Для выявления терминов были применены два алгоритма выделения терминоподобных слов и словосочетаний [26].

Первый алгоритм выделял существительные, прилагательные, согласованные пары и тройки прилагательных и существительных, а также заранее заданные конструкции (существительное — существительное в родительном падеже и т. п.). Второй алгоритм выделял часто повторяющиеся слова и группы в несколько слов. Полученные слова и словосочетания упорядочивались по убыванию частотности и убыванию количества содержащих их документов. Для проверки и добавления ключевых слов было проведено сопоставление с терминами Общественно-политического тезауруса, разработанного авторами ранее.

Следует отметить, что Общественно-политический тезаурус и в дальнейшем РуТез разрабатываются коллективом исследователей с конца 90-х годов. Основным способом оценки качества является экспертная оценка. Это гарантирует высокое качество, но лишний раз подтверждает сложность и трудоемкость работы. Решение задач автоматической обработки текстов требует развития методов быстрого построения узкопредметных тезаурусов и оценки их качества.

Большинство работ, описывающих автоматическое построение собственных тезаурусов или близких к ним структур, таких как лексико-семантические сети, онтологии предметных областей, при выделении терминов используют частоту встречаемости слов в используемом корпусе текстов и меру семантической близости терминов [27]. Частота встречаемости слов определяется количеством вхождений слова в корпусе текстов. Мера семантической близости предназначается для количественной оценки семантической схожести терминов. Эта характеристика может вычисляться разными способами и часто опирается на построение для каждого сло-

ва контекстного многомерного вектора и дальнейшего анализа пространства таких векторов.

Автор работы [28] решал задачу построения тезауруса, а точнее лексико-семантического поля вокруг заданного термина. В качестве исходного материала был выбран термин «двигатель» и корпус ruTenTen 2011, содержащий 14,55 миллиардов слов. В качестве статистических характеристик слов использовались степень семантической близости данного слова к ключевому и частоты данного слова в корпусе. Слова упорядочивались по степени семантической близости. Также на основе набора шаблонов и подсчёта частоты встречаемости были выделены словосочетания из двух и более слов. Однако окончательное формирование тезауруса осуществлялось экспертом.

Только на основе частоты встречаемости выбираются термины в работе [29]. В данной статье исследуется применение метода автоматической разработки тезаурусов предметных областей для построения тематических онтологий, применяемых в учебном процессе. Тезаурус строился на основе корпуса текстов предметной области. К сожалению, из работы неясно, каким образом определялись связи между терминами. Все этапы формирования тезауруса оценивал эксперт.

В статье [30] предлагается способ расчета весовых коэффициентов для определения степени значимости термина для заданной предметной области. Авторы рассматривали корпус научных статей, в рамках которого определили три числа. Первое — это ранг частоты, который позволяет уравнивать значимости самых встречаемых терминов любых текстов и одновременно распределять значимость терминов внутри одного текста. Второе — соответствие тематике текста. Тематическую группу научного текста формировали, выделяя термины из заголовка и подзаголовков, если при этом частота встречаемости терминов будет высокой в самом тексте; в этом случае вклад в значение весового коэффициента термина считается равным 1, иначе — 0. Третий коэффициент рассчитывался на основе принадлежности термина к содержательно смысловым (по терминологии авторов) блокам научной статьи. Авторы выделили индикаторы и маркеры четырех основных блоков: проблема, опыт, решение, итог. Если термин используется в предложении, содержащем формальный признак того или иного блока, то его вес корректируется на соответствующую величину. При этом если термин встретился в более чем одном блоке, его вес изменяется на сумму соответствующих величин. В качестве итоговой оценки из трех чисел рассчитывался интегральный весовой коэффициент термина. Таким образом, для анализа научных работ была предложена система метрик, тесно связанная со структурой текстовых документов, характерных для этой области.

Описанный способ был использован при разработке подхода автоматического определения тематики научного документа [31]. Эта работа показывает, что для разных предметных областей и сами методы, и отдельные детали методов могут значительно отличаться или изменяться. Хотя анализ структуры документа чаще применяется в алгоритмах извлечения знаний из текстов, тем не менее, такой подход оправдан при построении тезаурусов предметных областей, если используемый корпус текстов или его часть имеет заранее понятный формат.

В работах [32, 33] описана разработка русского тонального лексикона РуСентиЛекс. Большая часть этого лингвистического ресурса была построена вручную, но часть его терминов была выделена из текстов из социальной сети Twitter при по-

мощи алгоритма, комбинирующего бинарные классификаторы Logistic Regression, LogitBoost и Random Forest. Точность выбора терминов оценивалась на основе англоязычного корпуса отзывов, термины которого уже были размечены по тональности, и достигала 78,6 %. Отметим, что оценка качества работы алгоритма для русскоязычных текстов проводилась экспертами вручную, без какой-либо автоматизации. Авторы также доказали практическую полезность лексикона для тональной классификации текстов, организовав соревнование SentiRuEval-2016, участники которого разработали алгоритмы классификации с применением РуСентиЛекс, показавшие качество классификации в диапазоне 55–68 % для F-меры.

Отдельно хотелось бы остановиться на оценке качества выделения ключевых слов. Для этого существуют стандартные числовые характеристики: точность, полнота и F-мера [24]. Однако в статьях, посвященных работе с русскоязычными текстами, расчет этих характеристик осуществляется редко. В основном результат верифицируется экспертом, чаще всего без упоминания каких-либо числовых параметров, кроме количественных. Скорее всего, это связано с отсутствием размеченных тематических корпусов текстов на русском языке. Этот факт отмечается и в исследовании [34]. Авторы конкретных работ не выкладывают в общий доступ разработанные и используемые ими корпуса текстов. Конечно, во многих случаях это обусловлено объективными причинами, но наличие даже небольшого количества размеченных русскоязычных корпусов текстов существенно бы способствовало развитию автоматизации обработки данных в этой области.

В работе [35] качество отбора словосочетаний как терминов предметной области оценивалось с использованием существующей онтологии по естественным наукам и технологиям (OEHT). В качестве меры была выбрана средняя точность выделенных терминов, значения которой оказались в диапазоне от 59 % до 75 % в зависимости от характеристик выделяемых словосочетаний. К сожалению, этот метод можно применить только в тех немногих предметных областях, для которых существуют доступные онтологии. Кроме того, в этом случае встает вопрос о формировании корпуса текстов для исследования.

Следует отметить, что для анализа русскоязычных текстов используется достаточно узкий набор методов выделения ключевых слов. Согласно выводам исследователей, качество работы применяемых подходов и ряда стандартных инструментов, не зависящих или почти не зависящих от языка, является удовлетворительным в рамках решаемых задач. Однако малое количество и не слишком высокое качество существующих электронных русскоязычных тезаурусов, особенно по сравнению с английским языком, по мнению авторов, напрямую связано с недостаточным исследованием всего существующего спектра методов выделения ключевых слов.

2.2. Выделение связей

Существует два типа алгоритмов выделения семантических связей: статистические и лингвистические. Первые вычисляют статистические характеристики терминов в зависимости от их встречаемости в текстах, затем применяют математические методы, чтобы определить, насколько близки термины. В большинстве случаев статистические методы находят ассоциативные связи. Лингвистические методы используют существующие ресурсы: словари, тезаурусы, онтологии и т. д., откуда выделя-

ются существующие связи, или применяют лингвистические правила или шаблоны. Лингвистические методы способны находить любые виды семантических связей. Обзор этих методов можно найти в работе авторов [36].

Методы автоматического выделения семантических связей активно применяются к англоязычным текстам. Анализ статей, посвященных разработке русскоязычных лингвистических ресурсов, сразу показывает слабую степень автоматизации методов решения рассматриваемой задачи.

Разработчики РуТеза уделяют большое внимание описанию разных видов связей, которые могут присутствовать в тезаурусе [37, 38]. Однако при построении тезауруса основная часть отношений между понятиями РуТеза была получена из Общественно-политического тезауруса, а дополнялись связи экспертами.

Выделение связей предоставляют экспертам и авторы некоторых других работ [28, 29, 39]. В статье [28] явно выделена актуальность автоматизации построения узкопредметных тезаурусов и то, что важной их частью являются связи между словами.

Многие авторы используют связи между терминами из доступных лингвистических ресурсов, в основном из Википедии и РуТеза. Для решения задачи автоматического определения тематики документа [31] построена собственная онтология, в которой используются синонимические, гипонимо-гиперонимические связи из Википедии. Экспертный анализ результатов показал повышение качества работы системы при использовании онтологии, однако авторы указали на сложность построения таких лексических ресурсов. Использование РуТеза обсуждалось в разделе 1.2.

Получение связей между словами статистическими методами на основе частоты совместной встречаемости применён в работе [40]. Исследователи не строили в явном виде тезаурус, но показали, что задача классификации русскоязычных текстов решается лучше, если использовать семантические связи между терминами предметной области.

Открытый дистрибутивный тезаурус русского языка [41] строился автоматически методом Skip-Gram, реализованным в инструменте word2vec (https://code.google.com/archive/p/word2vec/), который находит семантические связи между терминами при помощи статистических алгоритмов. Авторы также автоматизировали и оценку качества тезауруса, сравнивая выделенные для него связи с существующими ресурсами с размеченными отношениями между словами: корпус BLESS, корпус когнитивных ассоциаций и другие. Средняя точность выбора связей достигала 97 %.

В работе [42] предложен полностью статистический метод автоматического построения ассоциативных связей для тезаурусов для нескольких языков, в том числе и для русского. Метод основан на подсчёте совместной встречаемости слов, сингулярном разложении матрицы «термины-на-документы» и нескольких мерах семантической близости. Для тезауруса английского языка экспертная оценка качества продемонстрировала очень высокие значения метрик: точность и полнота достигали 86–97 %. Следует отметить, что в работе недостаточно подробно описана оценка качества полученного русскоязычного тезауруса, в частности, не приведены значения метрик, которые бы описывали качество тезауруса в целом.

В работах [43, 44] описано построение ассоциативного портрета предметной области, т.е. совокупности наиболее характерных предметных и лингвистических зна-

ний, свойственных такой области. Авторы использовали методы для выделения связей на основе векторного алгоритма определения семантической близости слов. Связи не делились на отдельные виды, а рассматривались как единый набор ассоциаций. Получаемые структуры успешно применялись для решения задач [45, 46].

Хотелось бы обратить внимание на то, что в большинстве работ виды связей между словами не учитываются. Трудно сказать, связано ли это с достаточным качеством решения задач обработки текстов без учета видов связей или с трудностью определения разных типов иерархических и ассоциативных связей. Авторы данной работы показали существование отличия в степени влияния синонимических, иерархических и ассоциативных связей при использовании тезауруса для анализа тональности текстов [47]. Однако исследование влияния разных типов связей между терминами тезаурусов на качество результатов в задачах компьютерной лингвистике — мало изученная проблема, решение которой может привести к улучшению и развитию методов автоматического анализа текстов на естественных языках.

Одним из методов выделения разных типов связей является применение лексико-синтаксических шаблонов. Лексико-синтаксические шаблоны представляют собой характерные выражения (словосочетания и обороты), конструкции из определенных элементов языка. Примеры наиболее распространённых в русском языке шаблонов, в которых встречаются гипонимо-гиперонимические отношения, хорошо описаны в работе [48].

Группа разработчиков сформулировала язык для записи лексико-синтаксических шаблонов — LSPL [49]. Созданные при помощи предложенного языка шаблоны типичных фраз научной речи применялись авторами для автоматического анализа научно-технических документов на русском языке. Следует отметить, что для обработки текста были использованы традиционные словари (терминологический и морфологический), а также словарь общенаучных слов и выражений. Однако в данном случае шаблоны использованы не как средство получения связей между терминами предметной области, а как средство извлечения знаний из текста на естественном язке.

В другой работе [50] обсуждается проблема автоматического построения онтологий на основе использования лексико-синтаксических шаблонов. Автор предложил собственный язык лексико-синтаксических шаблонов и разработал на его основе программный комплекс, который позволяет хранить шаблоны и корпус текстов на русском языке в базе данных, редактировать и проводить валидацию шаблонов на корпусе русскоязычных текстов, проводить семантический анализ текстов корпуса. Оценку своего метода построения онтологий автор не выполнял, но предложил способ оценки качества работы на основе решения задачи информационного поиска.

Большинство лексико-синтаксических шаблонов разрабатываются экспертами. Было бы интересно провести исследование, описывающее и анализирующее широкий спектр таких шаблонов, применимых к построению русскоязычных тезаурусов предметных областей.

Таким образом, методы автоматического выделения связей между словами русского языка применяются исследователями в основном при построении лексических конструкций в рамках непосредственного решения задач обработки текста. Важно, что практически во всех работах отмечается существенное повышение качества решения после использования найденных связей. Это подчеркивает необходимость

разработки и исследования методов автоматизации для построения качественных тезаурусов, особенно в рамках задачи выделения связей между словами.

Интересно, что оценка лингвистических ресурсов, построенных авторами работ рассмотренных в этом разделе, как правило, проводится на основе качества решения с их помощью задач автоматической обработки текста. Это ещё раз подчеркивает нехватку размеченных экспертами эталонных корпусов текстов. Поскольку создание таких корпусов — задача очень трудоемкая, она также нуждается в методах и средствах автоматизации.

3. Эксперименты по построению русскоязычного тезауруса

Обзор существующих работ по построению тезауруса показывает, что создание специализированных лингвистических ресурсов автоматизировано недостаточно и даже после автоматизации требует большой работы эксперта-филолога во многих аспектах в целях повышения качества тезауруса для дальнейшего применения. Чтобы наглядно проиллюстрировать, каким качеством обладают тезаурусы, построенные полностью автоматически, и какие методы из области обработки естественного языка наиболее полезны для генерации тезауруса, авторы предложили комбинированный метод и провели эксперименты с созданием двух тезаурусов предметной области.

Проанализировав методы построения тезаурусов, авторы выделили наиболее подходящие, по их мнению, подходы к выделению ключевых слов и связей между ними для построения тезауруса как лексико-семантической модели предметной области. Этот раздел описывает результаты экспериментов по реализации разработанного алгоритма. Особое внимание было уделено способам выделения различных типов связей как наименее исследованной задаче.

Алгоритм построения русскоязычного тезауруса основан на алгоритме построения тезауруса, описанном в предыдущем исследовании авторов [36], и состоит из следующих этапов:

1. Выделение ключевых слов из текстов как терминов тезауруса алгоритмом TextRank [51].
2. Выделение ассоциативных связей статистическими алгоритмами.
3. Выделение синонимических связей из существующих лингвистических ресурсов и методом расстояния Левенштейна [53].
4. Выделение иерархических связей линвистическими методами.
5. Фильтрация терминов без связей.

На каждом этапе выделения связей предыдущие связи могут быть перезаписаны, например, если алгоритм определил пару терминов как синонимы, то если между ними была ранее установлена ассоциативная связь, она будет удалена, а синонимическая — записана вместо неё.

После построения основных структурных элементов тезауруса из него удаляются термины, для которых не было найдено ни одной связи, так как они бесполезны для дальнейшего применения тезауруса в обработке текстов.

3.1. Выделение терминов

В первую очередь для тезауруса выбираются термины, которые описывают основные темы предметной области, т.е. ключевые слова. Алгоритмы выделения ключевых слов были исследованы в предыдущей работе авторов [4]. По результатам исследования обучаемые алгоритмы работают более эффективно, чем алгоритмы без учителя, но составление материала для обучения требует огромных затрат со стороны эксперта-лингвиста. Алгоритмы типа «обучение без учителя» — PageRank и Topical TextRank — демонстрируют достаточно близкие характеристики качества работы, причём Topical PageRank является вариацией TextRank, которая более точно выбирает ключевые слова для конкретных текстов. Тем не менее, итоговый набор ключевых слов для всех текстов, который и берётся для основы тезауруса, оказывается одинаков для обоих алгоритмов.

Исходя из особенностей методов выделения терминов и требования максимальной автоматизации построения тезауруса, для создания русскоязычного тезауруса был выбран алгоритм TextRank, обладающий хорошими характеристиками качества и не требующий дополнительного обучения. Эксперименты с построением тезауруса проводились на двух базах текстов: статьи о российских мигрантах, выбранные экспертами-филологами ЯрГУ, и корпус русскоязычных твитов (http://linis-crowd.org/).

Корпус статей о мигрантах содержит 103 текста, из них 20 положительных и 83 отрицательных. В среднем в каждом тексте содержится 682 слова или 4 785 символов.

Корпус твитов содержит 4 320 текстов, из них 2 160 положительных и 2 160 отрицательных. В среднем в каждом тексте содержится 157 слов или 1 044 символа.

3.2. Выделение ассоциативных связей

Для выделения ассоциаций были выбраны два алгоритма, предназначенных для поиска семантически близких терминов: латентно-семантический анализ (ЛСА) [52] и word2vec (https://code.google.com/archive/p/word2vec/).

В методе латентно-семантического анализа строится матрица «термины-на-документы», где строкам соответствуют термины тезауруса, выбранные на предыдущем этапе, а столбцам — тексты предметной области, на которых генерируется тезаурус. Элементами матрицы служат частоты встречаемости терминов в конкретных текстах. После того как матрица построена, она раскладывается по сингулярным значениям, что позволяет найти матрицу меньшей размерности, которая является достаточно точным приближением исходной матрицы. Строки итоговой матрицы интерпретируются как векторы, характеризующие взаимосвязь соответствующего термина с текстами. Эти векторы сравниваются между собой по одной из стандартных мер близости векторов, например, по косинусной мере. В конце для пар терминов, у которых векторы обладают лучшими характеристиками близости, выделяются ассоциативные связи.

Инструмент word2vec строит векторные представления слов и по косинусной мере ищет семантически близкие термины так же, как и предыдущий. Этот алгоритм требует предварительного обучения, но для него существуют уже обученные модели, в том числе и для русского языка, которыми можно свободно пользоваться. Век-

торы для слов строятся при помощи стандартных алгоритмов CBOW (Continuous Bag of Words) и Skip-Gram.

Оба алгоритма выделяют достаточно большое число качественных ассоциативных связей статистическими методами, не зависящими от языковых особенностей, поэтому они были выбраны для автоматического построения русскоязычного тезауруса.

3.3. Выделение синонимов

Выбор синонимов осуществляется тремя алгоритмами, использующими соответственно расстояние Левенштейна [53]; выделение синонимических связей из словаря синонимов Synmaster (http://usyn.ru/blog.php?id_blog=11), в котором содержится около 1 200 000 слов и от одного до 20 синонимов к каждому слову; тезаурус РуТез.

Метод, основанный на расстоянии Левенштейна, ищет однокоренные слова и формы одного и того же слова, которые в тезаурусах считаются синонимами. Расстояние Левенштейна — это число изменений (удалений, добавлений или замен букв), которое требуется для преобразования одного слова в другое. В зависимости от этого расстояния и длины слов вычисляется мера близости, которая у схожих терминов будет больше, чем у других пар.

Алгоритмы, ищущие синонимы среди лингвистических ресурсов, работают по одному и тому же принципу: из словаря или тезауруса РуТез в автоматический тезаурус добавляются только те синонимические связи, которые связывают уже существующие в тезаурусе термины, новые термины в него не записываются.

3.4. Выделение иерархических связей

Иерархические, или гипонимо-гиперонимические, связи выбираются для автоматического тезауруса лингвистическими методами морфо-синтаксических правил [54] или выделения связей из РуТеза. Выделение гипонимов и гиперонимов из тезауруса РуТез происходит по тому же алгоритму, что и для синонимов.

Метод морфо-синтаксических правил считает, что термины связаны иерархическими отношениями, если один из них включает в себя второй как строку-суффикс или часть многословного термина. Первый в обоих случаях считается гипонимом, второй — гиперонимом, так как он является более общим. Например, «таможенный союз» — это гипоним к слову «союз».

3.5. Комплексная оценка тезауруса

Метрики для оценки качества тезауруса были взяты из предыдущей работы авторов [36], в которой предлагалась комплексная оценка тезауруса, построенного полностью автоматически гибридными методами. В частности, были выбраны графовые характеристики: число терминов, семантических связей различных типов и количество компонент связности графа тезауруса и размер наибольшей компоненты.

Выбор данной комплексной оценки для исследования качества тезауруса обусловлен тем, что она была разработана авторами специально для автоматически

Таблица 2: Характеристики автоматически построенных тезаурусов для корпуса статей о мигрантах

Table 2: Characteristics of automatically generated thesauri for the corpus of texts about migrants

Методы выделения			Количество выделенных				Количество	
гиперонимов	синонимов	ассоциаций	терминов	гиперонимов	синонимов	ассоциаций	компон. связности	вершин в наиб. комп.
Extraction methods for			Quantity of extracted				Quantity of	
hypernyms	synonyms	associations	terms	hypernyms	synonyms	associations	connected comp.	vertices in max. comp.
РуТез	РуТез	ЛСА	2 321	239	484	29 345	5	2 313
РуТез	Synmaster	ЛСА	2 413	239	59 844	27 335	1	2 413
Морф	Лев	word2vec	728	20	50	5 076	1	728
Гибрид	Гибрид	Гибрид	2 413	248	60 328	32 360	1	2 413

ЛСА — латентно-семантический анализ.
РуТез — метод, основанный на использовании РуТез.
Synmaster — метод, основанный на использовании Synmaster.
Лев — метод, использующий расстояние Левенштейна.
Морф — морфо-синтаксические правила.
Гибрид — совместное применение методов ЛСА и word2vec для ассоциаций, РуТез, Synmaster и Лев для синонимов, Морф и Гибрид для гипонимов и гиперонимов.

сгенерированных тезаурусов. Её главным преимуществом является возможность оценить тезаурус по нескольким аспектам сразу: качеству терминов и семантических связей, а также структуре и её связности. Второе достоинство комплексной оценки — это автоматизация её вычисления: графовые характеристики считаются полностью автоматически, не зависят от эксперта и сторонних ресурсов, а для автоматического подсчёта качественных характеристик нужен только альтернативный тезаурус выбранной предметной области как эталонный, с которым и проводится сравнение.

К сожалению, стандартные характеристики точности и полноты выделения терминов и связей в случаях с предлагаемыми тезаурусами невозможно подсчитать автоматически, так как для выбранных предметных областей не существует тезаурусов в открытом доступе, с которыми можно было бы сравнить результаты. Поэтому оценка точности выделения тезаурусных терминов и связей проводилась экспертом вручную.

3.6. Результаты экспериментов

Для экспериментов с методами выделения тезаурусных связей был построен программный стенд, реализующий описанный выше метод построения русскоязычного тезауруса. Для реализации стенда и вычисления оценочных характеристик тезауруса использовались языки программирования Python и Java и библиотеки NLTK и Gensim, в которых реализованы стандартные методы обработки данных на естественном языке.

В таблице 2 представлены графовые оценки качества тезауруса для корпуса статей о мигрантах, описанного в подразделе 3.1. Очевидно, что предложенный авторами комбинированный метод показывает лучшие характеристики тезауруса:

Таблица 3: Характеристики автоматически построенных тезаурусов для корпуса твитов

Table 3: Characteristics of automatically generated thesauri for the corpus of tweets

Методы выделения			Количество выделенных				Количество	
гиперонимов	синонимов	ассоциаций	терминов	гиперонимов	синонимов	ассоциаций	компон. связности	вершин в наиб. комп.
Extraction methods for			Quantity of extracted				Quantity of	
hypernyms	synonyms	associations	terms	hypernyms	synonyms	associations	connected comp.	vertices in max. comp.
РуТез	РуТез	ЛСА	15 629	2 808	608	105 998	333	14 605
РуТез	Synmaster	ЛСА	15 629	2 808	1 487 526	96 436	46	15 527
Морф	Лев	word2vec	527	80	445	78	49	41
Гибрид	Гибрид	Гибрид	15 629	2 888	1 488 577	96 476	46	15 527

ЛСА — латентно-семантический анализ.
РуТез — метод, основанный на использовании РуТез.
Synmaster — метод, основанный на использовании Synmaster.
Лев — метод, использующий расстояние Левенштейна.
Морф — морфо-синтаксические правила.
Гибрид — совместное применение методов ЛСА и word2vec для ассоциаций, РуТез, Synmaster и Лев для синонимов, Морф и Гибрид для гипонимов и гиперонимов.

он содержит больше всего терминов и связей всех типов, а также обеспечивает связный тезаурус.

Наименьший вклад вносят методы, основанные на морфо-синтаксических правилах и расстоянии Левенштейна, они выделяют меньше всего связей. Наибольшее число связей выделяет метод, берущий информацию из словаря Synmaster, за его счёт тезаурус оказывается состоящим из одной компоненты связности. Тезаурус, построенный при помощи методов word2vec, расстояния Левенштейна и морфо-синтаксических правил, также связный, но он существенно меньше остальных — в нём всего 728 терминов.

Для ассоциативных связей метод ЛСА существенно превосходит word2vec: он выделяет около 29 тысяч связей, часть из которых переходит в синонимические на следующих этапах алгоритма, а word2vec выбирает около 5 тысяч ассоциаций, что приблизительно в 6 раз меньше.

Гипонимо-гиперонимических связей в текстах оказалось очень мало по сравнению с остальными типами связей, и большая их часть выделена из тезауруса общего назначения РуТез.

В экспериментах с корпусом твитов (табл. 3) повторяются те же закономерности: комбинированный метод превосходит другие методы, взятые по отдельности, лучшим методом для выделения иерархических связей оказывается метод, использующий РуТез, синонимических — метод, применяющий Synmaster, а ассоциативных — ЛСА. При этом статистический метод word2vec показывает худший результат, чем для предыдущей выборки текстов: он находит существенно меньше связей, всего 78.

Таким образом, предложенный авторами метод демонстрирует лучшие характеристики качества итогового тезауруса. Отметим, что качество результата обеспечивается в первую очередь лингвистическими методами, выбирающими связи из существующих лингвистических ресурсов. Среди статистических методов самым эффективным оказался ЛСА, он в обоих случаях выделил достаточно большое число

связей. Остальные статистические методы либо выделяют малое число синонимов (метод расстояния Левенштейна), либо демонстрируют нестабильные результаты, отличающиеся от выборки к выборке (word2vec).

Экспертная оценка точности выделения терминов и связей была проведена для автоматического тезауруса, построенного комбинированным методом для корпуса статей о мигрантах, поскольку он обладает существенно меньшими размерами, чем второй, и эксперт может оценить его за разумное время. Оценка показала, что точность выделения терминов 94,3 %, всего 137 терминов оказались некорректными; синонимов — 99,9 %, а гипонимов и гиперонимов — 98,3 %, что объясняется тем, что практически все они были выделены из надёжных лексических ресурсов.

Что касается структуры автоматически сгенерированного тезауруса, то он в первую очередь состоит из синонимов и ассоциаций. Гипонимов и гиперонимов в нём оказалось существенно меньше, и эта закономерность прослеживается для обеих предметных областей. Возможной причиной этого может быть особенность корпусов текстов, которые содержат достаточно мало терминов, находящихся в иерархических отношениях между собой.

Заключение

Анализ существующих тезаурусов позволил выделить два доступных ресурса: РуТез и YARN, являющихся общими тезаурусами русского языка. Оба тезауруса могут успешно использоваться в задачах автоматической обработки текстов. И тот и другой сопоставимы по количеству слов и связей между ними, хотя и отличаются моделями основных единиц: понятие в тезаурусе РуТез и синсет в YARN.

Однако практически полное отсутствие доступных тезаурусов предметных областей показывает актуальное направление работы для создателей тезаурусов. Также следует отметить, что открытость как самих таких лексических ресурсов, так и моделей их разработки может позволить добиться повышения качества тезаурусов, а также внести существенный вклад в решение задач автоматической обработки текстов в соответствующих областях.

Методы построения русскоязычных тезаурусов также нуждаются в развитии и анализе. Здесь стоит обратить внимание на использование методов машинного обучения, комбинации статистических и лингвистических методов.

Отдельной большой проблемой является оценка методов построения тезаурусов. С точки зрения авторов работы, можно выделить два направления в развитии этих методов. Первое направление заключается в оценивании самого тезауруса с помощью как количественных характеристик содержащихся в нем слов и связей, так и с помощью оценки его внутренней структуры. Второе направление связано с оценкой тезауруса опосредованно через качество решения задач обработки текстов с его применением. Для реализации этого подхода необходима разработка открытых размеченных корпусов текстов.

Анализ современных работ в области построения русскоязычных тезаурусов и собственный опыт авторов позволяют надеяться на успешное решение обозначенных проблем в обозримом будущем.

Список литературы / References

[1] Aitchison J., Gilchrist A. and Bawden D., *Thesaurus construction and use: a practical manual*, Psychology Press, 2000, 230 pp.

[2] Сидорова Е. А., "Подход к моделированию процесса извлечения информации из текста на основе онтологии", *Онтология проектирования*, **8**:1(27) (2018), 134–151; [Sidorova E. A., "Ontology-based approach to modeling the process of extracting information from text", *Ontology of design*, **8**:1(27) (2018), 134–151, (in Russian).]

[3] Еленевская М. Н., Овчинникова И. Г., "Хранение и описание вербальных ассоциаций: словари и тезаурусы", *Вопросы психолингвистики*, 2016, №29, 69–92; [Yelenevskaya M. N., Ovchinnikova I. G., "The storage and description of the verbal associations", *Questions of psycholinguistics*, 2016, №29, 69–92, (in Russian).]

[4] Paramonov I. et al., "Thesaurus-Based Method of Increasing Text-via-Keyphrase Graph Connectivity During Keyphrase Extraction for e-Tourism Applications", Communications in Computer and Information Science, **649**, Springer, 2016, 129–141.

[5] Shchitov I., Lagutina K., Lagutina N., Paramonov I., "Sentiment classification of long newspaper articles based on automatically generated thesaurus with various semantic relationships", *Proceedings of the 21st Conference of Open Innovations Association FRUCT, University of Helsinki, Helsinki, Finland*, 2017, 290–295.

[6] Бленда Н. А., "Обзор русскоязычных тезаурусов для решения задачи расчета семантической близости между научными публикациями", *Информационные технологии и системы*, Труды Четвертой Международной научной конференции, 2015, 70–74; [Blenda N. A., "Overview of russian-language thesauri to solve the problem of calculating the semantic similarity for scientific publications", *Information Technologies and Systems*, Proceedings of the Fourth International Scientific Conference, 2015, 70–74, (in Russian).]

[7] Поршнев С. В., "О качестве открытых электронных тезаурусов русского языка", *Сборник материалов Всероссийской молодежной школы-семинара «Актуальные проблемы информационных технологий, электроники и радиотехники – 2015» (ИТЭР –2015)*, **2** (2015), 45–48; [Porshnev S. V., "O kachestve otkrytyh ehlektronnyh tezaurusov russkogo yazyka", *Sbornik materialov Vserossijskoj molodezhnoj shkoly-seminara «Aktualnye problemy informacionnyh tekhnologij, ehlektroniki i radiotekhniki – 2015» (ITER –2015)*, **2** (2015), 45–48, (in Russian).]

[8] Loukachevitch N., Dobrov B., "RuThes linguistic ontology vs. Russian wordnets", *Proceedings of the Seventh Global WordNet Conference*, 2014, 154–162.

[9] Loukachevitch N., Dobrov B., Chetviorkin I., "RuThes-Lite, a publicly available version of Thesaurus of Russian language RuThes", *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, 2014, №13(20), 340–349.

[10] Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V., "Creating Russian WordNet by conversion", *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, 2016, №15(22), 405–415.

[11] Braslavski P., Ustalov D., Mukhin M., Kiselev Y., "YARN: Spinning-in-Progress", *Proceedings of the Eight Global Wordnet Conference*, 2016, 58–65.

[12] Сухоногов А. М., Яблонский С. А., "Автоматизация построения англо-русского WordNet", *Компьютерная лингвистика и интеллектуальные технологии*, Труды Международного семинара "Диалог", 2005, 25–31; [Suhonogov A. M., Yablonskij S. A., "Avtomatizaciya postroeniya anglo-russkogo WordNet", *Computational Linguistics and Intellectual Technologies*, Papers from the Annual conference "Dialogue", 2005, 25–31, (in Russian).]

[13] Azarowa I., "RussNet as a computer lexicon for Russian", *Proceedings of the Intelligent Information systems IIS-2008*, 2008, 341–350.

[14] Азарова И. В., Захаров В. П., Киселев Ю., Усталов Д. А., Хохлова М. В., "Интеграция тезаурусов RussNet и YARN", *Компьютерная лингвистика и вычислительные онтологии*, Труды XIX Международной объединённой научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург, 22–24 июня 2016 г., Университет ИТМО, СПб, 2016, 7–13; [Azarova I. V., Zaharov V. P., Kiselev YU.,

Ustalov D. A., Hohlova M. V., "Integraciya tezaurusov RussNet i YARN", *Kompyuternaya lingvistika i vychislitelnye ontologii*, Trudy XIX Mezhdunarodnoj objedinyonnoj nauchnoj konferencii «Internet i sovremennoe obshchestvo» (IMS-2016), Sankt-Peterburg, 22–24 iyunya 2016 g., Universitet ITMO, SPb, 2016, 7–13, (in Russian).]

[15] Сладкова О., Пирумова Л., Пирумов А., "Информационные ресурсы Интернет для специалистов сельского хозяйства", *Международный сельскохозяйственный журнал*, 2016, № 2, 44–48; [Sladkova O., Pirumova L., Pirumov A., "Informacionnye resursy Internet dlya specialistov selskogo hozyajstva", *International Agricultural Journal*, 2016, № 2, 44–48, (in Russian).]

[16] Галиева А. М., Якубова Д. Д., "Принципы представления лексики в общественно-политическом тезаурусе татарского языка", *Филологические науки. Вопросы теории и практики*, 2016, № 12-2 (66), 80–84; [Galieva A. M., Yakubova D. D., "Principles of vocabulary presentation in socio-political thesaurus of the tatar language", *Philological Sciences. Questions of theory and practice*, 2016, № 12-2 (66), 80–84, (in Russian).]

[17] Галиева А. М., Кириллович А. В., Лукашевич Н. В., Невзорова О. А., Сулейманов Д. Ш., Якубова Д. Д., "Русско-татарский общественно-политический тезаурус: публикация в облаке лингвистических открытых связанных данных", *International Journal of Open Information Technologies*, **5**:11 (2017), 64–73; [Galieva A. M., Kirillovich A. V., Lukashevich N. V., Nevzorova O. A., Sulejmanov D. SH., YAkubova D. D., "Russian-tatar socio-political thesaurus: publishing in the linguistic linked open data cloud", *International Journal of Open Information Technologies*, **5**:11 (2017), 64–73, (in Russian).]

[18] Агеев М. С., Добров Б. В., Лукашевич Н. В., "Автоматическая рубрикация текстов: методы и проблемы", *Учён. зап. Казан. гос. ун-та. Сер. Физ.-матем. науки*, **150**:4 (2008), 25–40; [Ageev M. S., Dobrov B. V., Lukashevich N. V., "Automatic Text Categorization: Methods and Problems", *Kazan. Gos. Univ. Uchen. Zap. Ser. Fiz.-Mat. Nauki*, **150**:4 (2008), 25–40, (in Russian).]

[19] Лукашевич Н. В., Добров Б. В., Павлов А. М., Штернов С. В., "Онтологические ресурсы и информационно-аналитическая система в предметной области «Безопасность»", *Онтология проектирования*, **8**:1 (27) (2018), 74–95; [Lukashevich N. V., Dobrov B. V., Pavlov A. M., SHternov S. V., "Ontological resources and information-analytical system in security domain", *Ontology of design*, **8**:1 (27) (2018), 74–95, (in Russian).]

[20] Мишунин О. Б., Савинов А. П., Фирстов Д. И., "Проблемы, возникающие в интеллектуальных обучающих системах при оценке ответов на естественном языке", *Современные проблемы науки и образования*, 2015, № 2–2, 189–199; [Mishunin O. B., Savinov A. P., Firstov D. I., "Problems of automatic free-text answer grading in intelligent tutoring systems", *Modern problems of science and education*, 2015, № 2–2, 189–199, (in Russian).]

[21] Алексеев А. А., "Тематический анализ новостного кластера как основа тематического аннотирования", *Программная инженерия*, 2014, № 3, 41–48; [Alekseev A. A., "Thematic representation of a news cluster as a basis for summarization", *Software Engineering*, 2014, № 3, 41–48, (in Russian).]

[22] Усталов Д. А., "Обнаружение понятий в графе синонимов", *Вычислительные технологии*, **22**:S1 (2017), 99–112; [Ustalov D. A., "Concept discovery from synonymy graphs", *Computational Technologies*, **22**:S1 (2017), 99–112, (in Russian).]

[23] Kolchin M., Chistyakov A., Lapaev M., Khaydarova R., "FOODpedia: Russian food products as a linked data dataset", *International Semantic Web Conference*, 2015, 87–09.

[24] Hasan K., Vincent N., "Automatic keyphrase extraction: A survey of the state of the art", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, 1262–1273.

[25] Добров Б. В., Лукашевич Н. В., "Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска", *Учен. зап. Казан. гос. ун-та. Сер. Физ.-матем. науки*, **149**:2 (2007), 49–72; [Dobrov B. V., Lukashevich N. V., "Linguistic ontology on natural sciences and technologies for information-retrieval applications", *Kazan. Gos. Univ. Uchen. Zap. Ser. Fiz.-Mat. Nauki*, **149**:2 (2007), 49–72, (in Russian).]

[26] Лукашевич Н. В., Добров Б. В., Чуйко Д. С., "Отбор словосочетаний для словаря системы автоматической обработки текстов", *Компьютерная лингвистика и интеллектуальные технологии: Тр. Международной конференции "Диалог"*, 2008, № 7(14), 339–344; [Lukashevich N. V., Dobrov B. V., Chujko D. S., "Automated analysis of multiword expressions for computational dictionaries", *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, 2008, № 7(14), 339–344, (in Russian).]

[27] Turney P. D., Pantel P., "From frequency to meaning: Vector space models of semantics", *Journal of artificial intelligence research*, **37** (2010), 141–188.

[28] Захаров В. П., "Корпусно-ориентированный подход к построению тезаурусов и онтологий", *Структурная и прикладная лингвистика*, 2015, № 11, 123–141; [Zakharov V. P., "Corpus-based approach to thesaurus and ontology construction", *Structural and applied linguistics*, 2015, № 11, 123–141, (in Russian).]

[29] Котова Е. Е., Писарев И. А., "Построение тематических онтологий с применением метода автоматизированной разработки тезаурусов", *Известия СПбГЭТУ «ЛЭТИ»*, 2016, № 3, 37–47; [Kotova E. E., Pisarev I. A., "Construction of thematic ontologies using the method of automated thesauri development", *Proceedings of Saint Petersburg Electrotechnical University*, 2016, № 3, 37–47, (in Russian).]

[30] Аюшеева Н. Н., Кушеева Т. Н., "Способ вычисления весовых коэффициентов вершин семантической сети научного текста", *Фундаментальные исследования*, 2012, № 6-3, 626–630; [Ayusheeva N.N., Kusheeva T.N., "Method of calculation of weight factors tops semantic network scientific text", *Fundamental Research*, 2012, № 6-3, 626–630, (in Russian).]

[31] Аюшеева Н. Н., Гомбожапова Т. Н., Доржаев Т. В., "Способ автоматического определения тематики научного текста", *Фундаментальные исследования*, 2016, № 8-2, 229–233; [Ayusheeva N. N., Gombozhapova T. N., Dorzhaev T. V., "An automatic scientific text topic identification method", *Fundamental Research*, 2016, № 8-2, 229–233, (in Russian).]

[32] Chetviorkin I, Loukachevitch N., "Extraction of Russian sentiment lexicon for product meta-domain", *Proceedings of COLING 2012*, 2012, 593–610.

[33] Loukachevitch N., Levchik A., "Creating a General Russian Sentiment Lexicon", *Proceedings of Language Resources and Evaluation Conference*, 2016, 1171–1176.

[34] Ванюшкин А. С., Гращенко Л. А., "Оценка алгоритмов извлечения ключевых слов: инструментарий и ресурсы", *Новые информационные технологии в автоматизированных системах*, **20** (2017), 95–102; [Vanyushkin A. S., Grashchenko L. A., "Ocenka algoritmov izvlecheniya klyuchevyh slov: instrumentarij i resursy", *Novye informacionnye tekhnologii v avtomatizirovannyh sistemah*, **20** (2017), 95–102, (in Russian).]

[35] Лукашевич Н. В., Логачев Ю. М., "Комбинирование признаков для автоматического извлечения терминов", *Вычислительные методы и программирование*, **11**:4 (2010), 108–116; [Lukashevich N. V., Logachev YU. M., "Automatic term extraction based on feature combination", *Vychislitel'nye metody i programmirovanie*, **11**:4 (2010), 108–116, (in Russian).]

[36] Лагутина Н. С., Лагутина К. В., Мамедов Э. И., Парамонов И. В., "Методические аспекты выделения семантических отношений для автоматической генерации специализированных тезаурусов и их оценки", *Моделирование и анализ информационных систем*, **23**:6 (2016), 826–840; [Lagutina N. S., Lagutina K. V., Mamedov E. I., Paramonov I. V., "Methodological aspects of semantic relationship extraction for automatic thesaurus generation", *Modeling and Analysis of Information Systems*, **23**:6 (2016), 826–840, (in Russian).]

[37] Лукашевич Н. В., "Квазисинонимы в лингвистических онтологиях", *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог"*, 2010, № 9(16), 307–312; [Lukashevich N. V., "Near-synonyms in linguistic ontologies", *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, 2010, № 9(16), 307–312, (in Russian).]

[38] Лукашевич Н. В., "Моделирование отношений ЧАСТЬ–ЦЕЛОЕ в лингвистическом ресурсе для информационно-поисковых приложений", *Информационные технологии*, 2007, № 12, 28–34; [Lukashevich N. V., "Modeling of PART–WHOLE Relations in Linguistic Resource for Information-Retrieval Applications", *Information Technology*, 2007, № 12, 28–34, (in Russian).]

[39] Баранюк В. В., Богорадникова А. В., Смирнова О. С., "Определение семантического содержания предметной области на основе формирования тезауруса", *International Journal of Open Information Technologies*, **4**:9 (2016), 74–79; [Baranjuk V.V., Bogoradnikova A.V., Smirnova O.S., "Defining the scope semantics by forming its thesaurus", *International Journal of Open Information Technologies*, **4**:9 (2016), 74–79, (in Russian).]

[40] Нугуманова А. Б., Бессмертный И. А., Пецина П., Байбурин Е. М., "Обогащение модели Bag-of-Words семантическими связями для повышения качества классификации текстов предметной области", *Программные продукты и системы*, 2016, № 2(114), 89–99; [Nugumanova A. B., Bessmertnyj I. A., Pecina P., Bajburin E. M., "Semantic relations in text classification based on bag-of-words model", *Software products and systems*, 2016, № 2(114), 89–99, (in Russian).]

[41] Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N., Biemann C., "Human and machine judgements for russian semantic relatedness", *Analysis of Images, Social Networks and Texts. 5th International Conference, AIST 2016*, Springer, 2016, 221–235.

[42] Rapp R., "The automatic generation of thesauri of related words for English, French, German, and Russian", *International Journal of Speech Technology*, **11**:3–4 (2008), 147–156.

[43] Галина И. В., Козеренко Е. Б., Морозова Ю. И., Сомин Н. В., Шарнин М. М., "Ассоциативные портреты предметной области—инструмент автоматизированного построения систем big data для извлечения знаний: теория, методика, визуализация, возможное применение", *Информатика и её применения*, **9**:2 (2015), 92–110; [Galina I. V., Kozerenko E. B., Morozova Yu. I., Somin N. V., Sharnin M. M., "Associative portraits of subject areas as a tool for automated construction of big data systems for knowledge extraction: theory, methods, visualization, and application", *Informatics and its Applications*, **9**:2 (2015), 92–110, (in Russian).]

[44] Kuznetsov I. P., Kozerenko E. B., Charnine M. M., "Technological peculiarity of knowledge extraction for logical-analytical systems", *Proceedings of ICAI*, **12**, 2012, 18–21.

[45] Золотарев О. В., Шарнин М. М., "Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик", *Труды Международной научной конференции CPT2014*, 2015, 92–98; [Zolotarev O. V., Charnin M. M., "Methods of extracting knowledge from natural language texts and the construction of models of business processes on the basis of allocation processes, objects, their relationships and characteristics", *Proceedings of the International Scientific Conference CPT2014*, 2015, 92–98, (in Russian).]

[46] Золотарев О. В., Шарнин М. М., Клименко С. В., "Семантический подход к анализу террористической активности в сети Интернет на основе методов тематического моделирования", *Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление*, 2016, № 3, 64–71; [Zolotarev O. V., Charnin M. M., Klimenko S. V., "A semantic approach to the analysis of terrorist activity on the internet based on the methods of topic modeling", *Bulletin of the Russian New University. Series: Complex systems: models, analysis and management*, 2016, № 3, 64–71, (in Russian).]

[47] Лагутина Н. С., Лагутина К. В., Щитов И. А., Парамонов И. В., "Анализ использования различных типов связей между терминами тезауруса, сгенерированного с помощью гибридных методов, в задачах классификации текстов", *Моделирование и анализ информационных систем*, **24**:6 (2017), 772–787; [Lagutina N. S., Lagutina K. V., Shchitov I. A., Paramonov I. V., "Analysis of Influence of Different Relations Types on the Quality of Thesaurus Application to Text Classification Problems", *Modeling and Analysis of Information Systems*, **24**:6 (2017), 772–787, (in Russian).]

[48] Sabirova K., Lukanin A., "Automatic Extraction of Hypernyms and Hyponyms from Russian Texts", *Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST'2014)*, 2014, 35–40.

[49] Большакова Е. И., Иванов К. М., Сапин А. С., Шариков Г. Ф., "Система для извлечения информации из текстов на базе лексико-синтаксических шаблонов", *Пятнадцатая национальная конференция по искусственному интеллекту с международным участием*, 2016, 14–22; [Bol'shakova E. I., Ivanov K. M., Sapin A. S., SHarikov G. F., "Sistema dlya izvlecheniya informacii iz tekstov na baze leksiko-sintaksicheskih shablonov", *Pyatnadcataya nacionalnaya konferenciya po iskusstvennomu intellektu s mezhdunarodnym uchastiem*, 2016, 14–22, (in Russian).]

[50] Рабчевский Е. А., "Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска", *Электронные библиотеки: перспективные методы и технологии, электронные коллекции*, сб. науч. тр. 11-й Всероссийской научной конференции RCDL-2009, Петрозаводск, 2009, 69–77; [Rabchevskij E. A., "Avtomaticheskoe postroenie ontologij na osnove leksiko-sintaksicheskih shablonov dlya informacionnogo poiska", *Elektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollekcii*, sb. nauch. tr. 11-j Vserossijskoj nauchnoj konferencii RCDL-2009, Petrozavodsk, 2009, 69–77, (in Russian).]

[51] Mihalcea R., Tarau P., "TextRank: Bringing order into texts", *Proceedings of Empirical Methods in Natural Language Processing – EMNLP*, ACL, Barcelona, Spain, 2004, 404–411.

[52] Wiemer-Hastings P., Wiemer-Hastings K., Graesser A., "Latent semantic analysis", *Proceedings of the 16th international joint conference on Artificial intelligence*, 2004, 1–14.

[53] Noh S., Kim S., Jung C., "A Lightweight Program Similarity Detection Model using XML and Levenshtein Distance", *FECS*, 2006, 3–9.

[54] Lefever E., Van de Kauter M., Hoste V., "Evaluation of automatic hypernym extraction from technical corpora in English and Dutch", *9th International Conference on Language Resources and Evaluation (LREC)*, 2014, 490–497.

Lagutina N. S., Lagutina K. V., Adrianov A. S., Paramonov I. V., "Russian-Language Thesauri: Automated Construction and Application For Natural Language Processing Tasks", *Modeling and Analysis of Information Systems*, **25**:4 (2018), 435–458.

DOI: 10.18255/1818-1015-2018-4-435-458

Abstract. The paper reviews the existing Russian-language thesauri in digital form and methods of their automatic construction and application. The authors analyzed the main characteristics of open access thesauri for scientific research, evaluated trends of their development, and their effectiveness in solving natural language processing tasks. The statistical and linguistic methods of thesaurus construction that allow to automate the development and reduce labor costs of expert linguists were studied. In particular, the authors considered algorithms for extracting keywords and semantic thesaurus relationships of all types, as well as the quality of thesauri generated with the use of these tools. To illustrate features of various methods for constructing thesaurus relationships, the authors developed a combined method that generates a specialized thesaurus fully automatically taking into account a text corpus in a particular domain and several existing linguistic resources. With the proposed method, experiments were conducted with two Russian-language text corpora from two subject areas: articles about migrants and tweets. The resulting thesauri were assessed by using an integrated assessment developed in the previous authors' study that allows to analyze various aspects of the thesaurus and the quality of the generation methods. The analysis revealed the main advantages and disadvantages of various approaches to the construction of thesauri and the extraction of semantic relationships of different types, as well as made it possible to determine directions for future study.

Keywords: thesaurus, semantic relationships, automatic thesaurus construction, automatic relationship extraction, keyword extraction

On the authors:
Nadezhda S. Lagutina, orcid.org/0000-0002-6137-8643, associate professor, PhD,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003, Russia, e-mail: lagutinans@rambler.ru

Ksenia V. Lagutina, orcid.org/0000-0002-1742-3240, student,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003, Russia, e-mail: lagutinakv@mail.ru

Aleksey S. Adrianov, orcid.org/0000-0002-3073-0982, student,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003, Russia, e-mail: alex.a4.25@yandex.ru

Ilya V. Paramonov, orcid.org/0000-0003-3984-8423, associate professor, PhD,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl, 150003, Russia, e-mail: Ilya.Paramonov@fruct.org