УДК 004.82:004.912

Проблемы извлечения терминологического ядра предметной области из электронных энциклопедических словарей

Кононенко И.С. (Институт систем информатики СО РАН),
Ахмадеева И.Р. (Институт систем информатики СО РАН),
Сидорова Е.А. (Институт систем информатики СО РАН),
Шестаков В.К. (Институт систем информатики СО РАН)

Статья посвящена проблемам автоматического построения терминологической системы предметной области. Предложен метод извлечения терминов предметной области на базе электронных энциклопедических источников данных. Особенностью предлагаемого подхода является тщательный анализ структуры термина, распознавание ошибок на базе их лингвистической классификации, автоматическая генерация лексико-синтаксических шаблонов, представляющих многокомпонентные термины, и использование набора эвристических методов обработки «особых» терминов. Использование энциклопедических словарей позволяет автоматически сформировать эталонный список наименований понятий и применять его для оценки качества формируемых словарей.

Ключевые слова: *извлечение терминов, многословный термин, терминологическое ядро предметной области, омонимия, предметный словарь.*

1. Введение

В данной работе рассматриваются проблемы автоматического формирования первоначального концептуального состава онтологии научной предметной области (ПО), который определяется совокупностью терминов – слов и словосочетаний, являющихся наименованиями понятий моделируемой ПО. Источником соответствующих знаний могут служить электронные тексты различных жанров: учебники и предметные указатели к ним, научные статьи и монографии, научно-популярная литература, терминологические и энциклопедические словари.

Анализ литературы [4, 5, 8] показывает, что при извлечении терминологии из большого массива текстов используются подходы, объединяющие лингвистические и статистические

методы. Для формирования списков терминов-кандидатов, удовлетворяющих заданным лингвистическим условиям, используется метод шаблонных конструкций, описывающих классы языковых выражений. В зависимости от типа учитываемой в конструкции языковой информации, применяемые в различных работах шаблоны подразделяются на грамматические [12], лексико-грамматические [5, 8] и лексико-синтаксические [1, 9]. Извлечение терминов-кандидатов сопровождается накоплением статистики встречаемости и подсчетом весов для фильтрации и сортировки полученных списков. В результате процедуры в список терминов-кандидатов попадают не только сложившиеся в данной области обозначения основных специальных понятий, но и многочисленные общенаучные, периферийные и авторские термины, которые, как показано в работе [2], характеризуются большой степенью варьирования языковой формы. В этой ситуации необходим этап экспертной оценки, на котором ранжированные списки предъявляются эксперту для отбора истинных терминов.

Задача формирования терминологического ядра онтологии научной ПО требует строгого отбора сложившихся в данной области обозначений специальных понятий. Многие исследования, ориентированные на извлечение и формализацию предметных знаний в виде онтологии, опираются не на корпус разножанровых текстов, а на более доступные структурированные источники, такие как специализированные глоссарии, предметные указатели, терминологические словари и энциклопедии. Эти источники преимущественно используются в задаче извлечения знаний о взаимосвязях понятий, например, родовидовых отношений, отношений эквивалентности и т.п. [6, 7, 10-11, 14-16]. Однако не менее перспективно использование словаря как эксплицитного источника понятийного ядра ПО и стандарта его терминологического представления. Это снимает проблему отбора основного ядра терминов экспертом (соответствующая работа проделана составителями словарей), что не исключает необходимости накопления статистической информации, учитывая наличие не вошедших в словарь периферийных терминов, общенаучных терминов и терминов смежных дисциплин.

Задачей данного исследования является разработка методов извлечения терминологического ядра предметной области из структурированных интернет-источников, содержащих энциклопедические данные научных областей знаний. Используемый метод предполагает наличие в таких источниках эталонного массива надежных терминов, который может быть автоматически извлечен и использован для оценки качества формируемых терминов. Для достижения поставленных целей 1) проведен эксперимент по автоматическому извлечению терминов с помощью базовой технологии, основанной на методе грамматических шаблонов, 2) проанализирован полученный массив терминов-кандидатов и выявлены

проблемные ситуации, 3) предложены дополнительные эвристические методы для решения выявленных проблем и 4) проведен эксперимент по извлечению терминов с помощью расширенной методики и осуществлена сравнительная оценка эффективности двух методик.

2. Электронный словарь как источник терминов

В качестве материала для данного исследования были выбраны два электронных энциклопедических словаря – словарь терминов теории графов (объемом 151 словарных статей) и толковый словарь по искусственному интеллекту (объемом 564 словарных статьи) – в предположении, что данные источники содержат достаточно представительное ядро надежных терминов и их взаимосвязей, характерных для соответствующих ПО. Исследуемые источники являются энциклопедическими словарями, содержащими термины и их дефиниции, что позволяет использовать их не только для извлечения терминов, но и в дальнейшем ставить задачу извлечения отношений между понятиями на основе терминов и их толкований. Общим свойством двух словарей является тесная связь с общенаучной терминологией и терминологией смежных дисциплин, поскольку исследования в этих областях носят междисциплинарный характер.

Контент в обоих словарях имеет HTML-разметку, что облегчает выделение границ текстовых сегментов, а систематическое использование в словарях гиперссылок может служить для выделения соответствующих вхождений терминов в текст дефиниций.

Рис.1 демонстрирует словарные статьи из словаря по ИИ и словаря по теории графов, представленные в виде текста с сохраненной разметкой.

В структуре словарной статьи выделяются две части – левая и правая. В левой части статьи представлено заголовочное слово (заглавный термин или множество заглавных терминов), выделенное заголовочными тегами (словарь по ИИ) или жирным шрифтом (словарь теории графов). Совокупность заглавных терминов представляет словник словаря. Подавляющее большинство заглавных терминов представлено номинативными конструкциями. Правая часть статьи – это зона дефиниций, которая содержит дефиницию-толкование.

Рис. 1. Примеры представления терминов в энциклопедических словарях.

Термины словника, используемые в толковании, выделены гиперссылочными тегами (словарь по ИИ) или курсивом (словарь теории графов). В первом примере приведен многозначный заглавный термин, которому соответствуют две альтернативных дефиниции. В правой части вместо толкования могут быть представлены синонимичные заглавные термины, маркированные тегами-ссылками на соответствующие словарные статьи и лексическими маркерами (“то же, что”, “см.”, “синоним для”). Зоны грамматических и стилистических помет отсутствуют, что характерно для словарей энциклопедического типа.

Оба словаря используют прием ввода вложенных дефиниций в правую часть словарных статей: толкование термина *простая цепь* в статье *цепь* (словарь теории графов), толкование термина *унификатор* в статье *унификация* (словарь по ИИ).

Особенности словаря по ИИ:

(а) “*вложенный*” термин является заглавным, ему соответствует собственная словарная статья, дефиниционная часть которой представляет собой ссылку с маркером “термин объясняется в статье”;

(б) ввод иллюстративного материала в зону дефиниции с помощью лексических маркеров “*примером может служить*”, “*например*”. Так, отношение "*быть супругом*" приводится как иллюстрация термина *отношение симметричное*;

(в) использование в дефиниционной части графических инициальных сокращений, которые терминами не являются, а служат окказиональной заменой термина/компонента термина в пределах его словарной статьи*: каузация – К., диссонанс когнитивный – Д.К.*

Словарь теории графов отличают следующие особенности:

(а) нестандартные способы ввода синонимов – в левой части, путем перечисления после заглавного термина, возможно, в скобках: *Вершина,Узел**;** Вполне несвязный граф(пустой граф,нуль-граф)**;** Независимое множество вершин(известное также каквнутренне устойчивое множество);*

(б) широкое использование в толкованиях числовых записей величин, буквенных обозначений переменных (на базе латиницы), различных символьных последовательностей, характерных для обозначений в языке математики; переменные, как правило, выделяются разметочными тегами.

(в) окказионально используемое (3 случая) нестандартное представление словарной статьи, когда заглавный термин вводится в автонимном контексте с предикатом называния: *Полным графомназывается граф, в котором…*

(г) активное использование инверсной структуры терминов (почти 50 процентов многословных заглавных терминов), т.е. изменение прямого порядка слов в терминологическом словосочетании с выведением слова, несущего максимальную смысловую нагрузку, в позицию ведущего слова: *область предметная, автомат детерминированный*. Этот прием используется в различных номенклатурных списках, предметных указателях и т.п., позволяя сгруппировать родовой и видовые термины в одной зоне словаря.

Основную массу заглавных терминов составляют существительные, а также именные группы с зависимыми от главного существительного согласованными прилагательными или причастиями и существительными в родительном падеже. В обоих словарях заголовочные слова (лексемы и словосочетания) преимущественно представлены в основной (нормальной) форме, которая определяется именительным падежом лексемы или главного существительного.

3. Извлечение терминологии

Научный дискурс и жанр электронного словаря определяют специфику лексико-семантических, морфологических, синтаксических и структурных параметров анализируемых текстов. В данной главе описываются особенности процесса анализа текстового контента

электронных энциклопедических словарей и построения на их основе терминологического ядра предметной области.

Процесс извлечения терминологии включает такие этапы как а) выделение основного текстового контента и очистка от незначимых элементов, б) сегментация, направленная на выделение структурных сегментов текста, определяющих границы дальнейшего анализа, в) графематический анализ, обеспечивающий токенизацию и выделение нетекстовых элементов (формул, гиперссылок, числовых данных, обозначений и пр.) и иноязычных вкраплений, г) лексико-морфологический анализ (лемматизация, определение лексико-грамматических признаков, представление парадигмы, нормализация), д) выделение терминоподобных словосочетаний (идентификация на основе предопределенных грамматических моделей и нормализация), е) применение различных эвристик для снятия омонимии и улучшения качества извлечения терминов. Рассмотрим подробнее особенности данного процесса.

3.1. Сегментация

Процесс сегментации начинается с очистки исходного HTML документа от лишней разметки. В тексте оставляются только значимые теги: <b>, <a>, <i>, которые затем используются при анализе структуры документа.

Затем в очищенном тексте выделяются сегменты (фрагменты текста): «Определение», «Левая часть», «Правая часть». Выделение таких сегментов осуществляется на основе их декларативного описания (множество шаблонов сегментов), которое строится экспертом в зависимости от структурной организации словарной статьи электронного словаря (Рис.2).

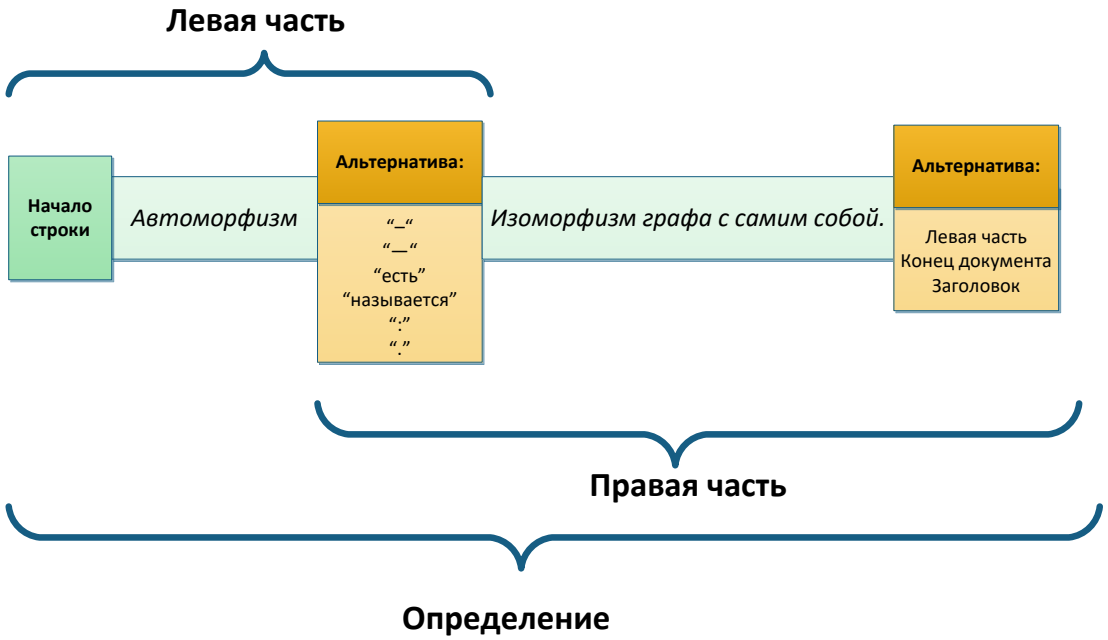

Рис. 2. Структура словарной статьи.

Для описания шаблона сегмента нужно указать тип сегмента, ограничения, которым должен удовлетворять сегмент (в зависимости от типа сегмента), а также уникальное имя, позволяющее использовать этот сегмент как часть в определении более сложных шаблонов. Пример декларативного описания сегментов, построенного для сегментации словаря терминов по теории графов, представлен на Рис. 3. Заметим, что представленные шаблоны позволяют выделять сегменты в электронных словарях схожей структуры, например, таких как глоссарии Википедии.

```
{
        "type":"or",
        "name":"Разделитель_определения",
        "segments": ["s/–", "s/—", "s/есть", "s/называется", "s/:", "s/."]
},
{
        "type":"classic",
        "name":"Левая_часть",
        "begin":"Начало_строки",
        "end":"Разделитель_определения"
},
{
        "type":"classic",
        "name":"Правая_часть",
        "begin":"Левая_часть",
        "end": ["Левая_часть", "__end__", "t/h2"]
},
{
        "type":"sequence",
        "name":"Определение",
        "segments":["Левая_часть", "Правая_часть"]
}
```

Рис. 3. Декларативное представление сегментов.

Элементарными маркерами сегментов могут быть конкретные последовательности символов или регулярное выражение, а также слова и фразы подключаемого предметного словаря. Поддерживается несколько типов шаблонов и ограничений, которым должен удовлетворять искомый сегмент.

a) Сегмент можно задать с помощью ограничений на его начало и конец, которыми могут быть другие сегменты (так задаются сегменты «Левая часть» и «Правая часть»).
b) Сегмент может быть задан последовательностью других сегментов, например, «Определение» является последовательностью сегментов «Левая часть» и «Правая часть».

c) Сегмент можно определить как альтернативу других сегментов: «Разделителем определения» может быть один из нескольких вариантов.

Таким образом, на вход сегментатору подаются шаблоны сегментов, представленные в JSON формате, и очищенный текст, а на выходе для каждого шаблона формируется множество найденных в тексте сегментов, удовлетворяющих условиям шаблона.

В дальнейшем анализе участвует только текст, соответствующий сегментам типа «Определение». При первом проходе в левой части определений выделяются заглавные термины (с помощью технологии Клан, рассматриваемой ниже). Заметим, что если в тексте присутствуют теги, то в первую очередь в качестве заглавного термина алгоритм пытается взять фрагмент, выделенный тегами. Так, на Рис. 4 внутри сегмента «Левая часть» красным отмечены вхождения сегментов «Выделение_тегами». Если выделенных тегами фрагментов несколько, то первый из них рассматривается как главный, остальные – синонимы. Так, в рассматриваемом примере термин «Двудольный граф» (выделен красным) считается главным термином, а термины «биграф» и «четный граф» его синонимами (выделены зеленым).

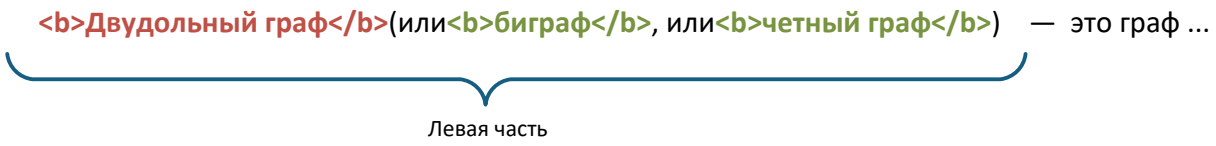

Рис. 4. Пример разбора левой части определения.

При втором проходе анализируются правые части определений и отмечаются все вхождения дескрипторов (которые мы нашли на предыдущем шаге) в определения, а также указываются их морфологические характеристики, также с помощью технологии Клан.

3.2. Технология Клан

Лексико-морфологический и поверхностно-синтаксический анализ и поиск терминов осуществляются системой Клан [12]. Данная система позволяет создавать предметно-ориентированные словари и использовать их при автоматическом анализе текстов.

Морфологический анализ осуществляется на базе модуля Диалинг (www.aot.ru), который содержит универсальный словарь русского языка и обеспечивает поиск слова в словаре, определение его грамматических признаков и нормальной формы. Также поддерживается функция предсказания [13], которая по незнакомому слову формирует гипотезы (как правило, около 3 вариантов) о его части речи, нормальной форме и других признаках. Так, например, для слова *когнитивный* были сформированы следующие предсказания:

когнитивная *Сущ, жр, но /a (пустая парадигма)

когнитивный *Сущ, жр, но /a (пустая парадигма)

когнитивный *Прил /S (ен, на, ная, нее, ней, но, ного, ное, ной, ном, ному …)

когнитивна *Сущ, жр, од, имя /b (, а, ам, ами, ах, е, ой, ою, у, ы,)

когнитивные *Сущ, жр, но /a (пустая парадигма)

когнитивных *Сущ, жр, но /a (пустая парадигма)

Особенностью морфологического представления в рамках системы Клан, является формирование морфологического класса термина и обеспечение его уникальности в рамках своего класса. Это означает, что одинаковые по написанию термины (в нормальной форме) не могут принадлежать одному морфологическому классу, а возможная лексико-семантическая неоднозначность поддерживается функционалом формирования альтернативных групп семантических характеристик.

Морфологический класс определяется частью речи, набором лексических признаков слова (например, *одушевленность* или *род* у существительных) и типом парадигмы, определяющим изменяемые морфологические признаки, которыми дополнительно снабжаются разные формы слова. Набор морфологических классов может быть изменен пользователем, в зависимости от решаемых задач, однако потребность в этом возникает достаточно редко. Исключением являются случаи, когда используются дополнительные специализированные словари терминов, например, словарь имен или географических названий, или необходимость включать в словарь слова другого языка.

Таким образом, однословный термин словаря Клан обладает уникальным идентификатором <*норма, морф_класс*>, что позволяет однозначно распознавать термин по набору его признаков и делает возможным применение методов автоматического пополнения словаря на основе корпусов текстов.

Другой важной особенностью системы является поддержка многословных терминов. Многословный термин в системе Клан – это словосочетание, сформированное по одному из правил, реализующему поверхностно-синтаксический анализ (для русского языка). Большинство многословных терминов включают от 2 до 4 слов и формируются с помощью правил вида П+С (*планарный граф*), П+П+С (*новая информационная технология*), С+Срд (*сеть петри*), П+С+Срд (*локальная степень вершины*), С+Прд+Срд (*обработка естественного языка*), С+Срд+Срд (*компонента связности графа*). Имеются также термины более сложной структуры, например, с зависимыми предложными группами С+Предл+С (*путь в орграфе, рассуждение по умолчанию*), С+Предл+С+С (*поиск в пространстве состояний*), С+Предл+П+С (*автомат с переменной структурой*) и т.д. В системе реализован

собственный компонент сборки словосочетаний русского языка, который по заданному набору слов и их грамматическим характеристикам проверяет согласование в соответствии с одной из синтаксических моделей и синтезирует нормальную форму многословного термина. Многословный термин словаря Клан однозначно идентифицируется набором *<норма, правило, <лексический_состав>>*. У такого термина можно, при необходимости, взять синтаксическую вершину – однословный термин – и грамматические признаки, которые формируются на основе грамматических признаков вершины.

Семантический компонент словаря обеспечивается иерархией семантических классов, набором семантических атрибутов, фиксацией значения (дескриптора) и механизмом формирования групп признаков у термина. Такой механизм позволяет задавать как альтернативные, так и синкретически-выраженные значения терминов. В рамках данной работы с помощью такого рода признаков фиксируются формально-текстовые признаки, отражающие, в какой части определения встретился термин, есть ли у него омонимы и т.п.

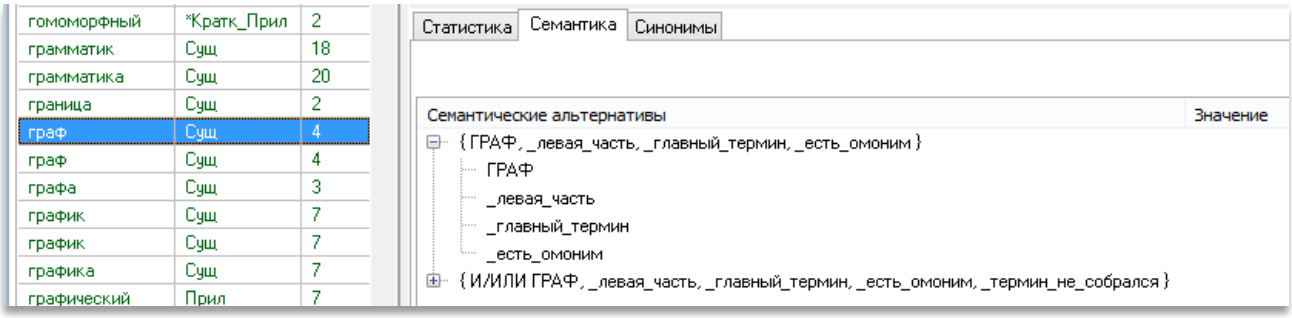

Рис. 5. Формальные характеристики термина в базовом словаре.

Для терминов также может накапливаться статистика встречаемости в тексте(текстах).

Таким образом, с помощью системы Клан реализуется терминологический анализ определений, найденных в электронном энциклопедическом ресурсе. Результатом обработки текста с помощью данной системы являются два базовых словаря лексических единиц:

– однословные кандидаты в простые термины и компоненты составных терминов и

– многословные кандидаты в составные термины.

3.3. Типизация проблем выделения терминов

При анализе списков терминов-кандидатов, полученных с помощью рассмотренных выше средств, был выявлен ряд особенностей и проблем, связанных с некорректностью и неполнотой извлечения простых и составных терминов.

3.3.1. Особенности выделения простых терминов и компонентов терминологических словосочетаний

В данном разделе приводится типизированный перечень проблем, характеризующих извлечение однословных терминов и компонентов составных терминов[1]. Корректность извлеченного из текста термина-кандидата определяется по следующим параметрам: нормальная форма (лемма), часть речи, лексические признаки, класс словоизменения, парадигма. Возможные ситуации: термину соответствует группа извлеченных омонимичных вариантов, лишь один из которых корректен; термину соответствуют только некорректные варианты; термину не сопоставлено ни одного кандидата. Учитывая это, все ошибки можно подразделить на три типа: омонимия, некорректность и неполнота.

Омонимия. Формируемый в процессе анализа текста словник содержит все леммы, имеющие омонимичные словоформы, т.е. группы лексических и лексико-грамматических омонимов из универсального словаря и группы предсказанных в сходных позициях альтернатив, различающихся по подмножеству/совокупности параметров.

(1) Наличие в универсальном словаре лексических или лексико-грамматических омонимов, лишь один из которых является термином и/или компонентом термина словаря. Частотным примером является совпадение ряда форм одноосновных существительных, различающихся по роду и/или одушевленности: *граф* (С,мр,но) – *граф* (С,мр,од) – *графа* (С,жр,но), *логика – логик*, *графика – график, метод – метода*, *клика – клик, домна – домен, мир – миро*. Более редкие примеры: *машина* (С,жр,но) – **машин* (С,од,фам).

(2) Предсказанные слова, образованные от словарных слов наращиванием стандартного префикса (*анти, мульти, полу, пере, псевдо, мета, гипер, спец, ко, су, макси-, мини-*), наследуют грамматические свойства производящих слов: **мультиграф* (С,мр,но) – *мультиграф* (С,мр,од) – **мультиграфа* (С,жр,но); *макси-код* (С,мр,но) – **макси-кода* (С, жр,но)[2].

(3) Источником омонимии в рамках одной части речи являются паронимы (*временной – временный, языковой – языковый, образный – образной*) и стилистические варианты (*знание – знанье, рассуждение – рассужденье, отношение – отношенье*), контраст которых нейтрализуется в определенных синтаксических позициях. В результате формируются неверные дубликаты терминов или словосочетаний.

(4) Предсказанная частеречная омонимия “существительное vs. прилагательное” характерна для прилагательных-компонентов составных терминов: *абдуктивный, деонтический, когнитивный, секвенциальный, кликовый*, *цикломатический*. Менее частотны частеречная

[1] Искажения, связанные с ошибками в текстах исходных словарей, не рассматриваются как ситуации некорректности.

[2] Здесь и далее звездочкой помечены некорректные гипотезы.

омонимия “прилагательное vs. причастие” (*ориентированный – ориентировать*), омонимия “существительное vs. глагол” (*фрактал – *фрактать, лок – *лочь*). Имеются и редкие варианты “существительное vs. прилагательное” (*остов – *остовый*), “существительное vs. прилагательное vs. глагол” (*решатель – *решателый – *решатеть*).

(5) Большинство иностранных фамилий, используемых в составе терминов (*универсум Эрбрана, сеть Петри, решетка Келли репертуарная*), являются незнакомыми словами и регулярно предсказываются многовариантно, например, фамилии *Эрбран* соответствуют гипотезы **эрбрана* (С,жр,но), **эрбраный* (Кр_П, кач) и **эрбрать* (Кр_Прич).

(6) Предсказание незнакомых дефисных сложных прилагательных является еще одним источником регулярной омонимии (*контекстно-связанный – *контекстный-связанный*).

Некорректность термина-кандидата определяется по подмножеству / совокупности значений приведенных выше параметров. Учитывая их взаимосвязь, можно подразделить данный тип ошибок на два подтипа.

А. Лексическая некорректность – недостатки определения словарной формы термина (неполная или искаженная форма). Так, некорректность леммы в ситуациях (7)-(8) связана с использованием графических терминоэлементов (символов, знаков, цифр, некириллических букв, вариантов графем), которые при лексико-морфологическом анализе пропускаются. Есть случаи некорректности нормальной формы в связи с тем, что общепринятая нормализация не всегда терминологична (9)-(10).

(7) Ряд терминов имеют дефисноoформленные буквенные префиксы с использованием букв латинского или греческого алфавита: *n-факторизация, k-связный*.

(8) Кириллические префиксы (*ИИ-программирование, мимд-архитектура)*, как правило, проблем не вызывают, за исключением случая нетрадиционного использования кавычек: *“лямбда”-исчисление.*

(9) В результате нормализации существительные, выступающие в качестве одиночных терминов, приводятся к виду С,им,ед. Однако стандартная нормализация по числу не всегда адекватна с терминологической точки зрения: некоторые термины-существительные (и формируемые на их основе словосочетания) представляют множественные понятия, что маркировано в толковании и отражается в выборе множественного числа существительного в качестве заглавного термина. Пример: *знания* (совокупность сведений, образующих целостное описание…).

(10) Нормализация превосходной степени качественных прилагательных, употребляемых в составе математических терминов, таких как *унификатор наибольший общий* (*наибольший*

=> *большой*), приводит к искажению смысла термина, в состав которого входит данная форма.

Б. Лексико-грамматическая некорректность диагностируется по подмножеству/совокупности значений параметров, определенных механизмом предсказания для отсутствующих в универсальном словаре слов. В (11)-(12) приведены типовые примеры ситуаций, в которых термину соответствуют только некорректные варианты.

(11) Механизм предсказаний является источником ошибок при формировании леммы и парадигмы дефисных сложных прилагательных при корректности других параметров (**аппаратный-программный* – **аппаратный-программного*), а также парадигмы существительных с изменяемой первой частью (*граф-звезда* – **граф-звезды, фрейм-прототип* – **фрейм-прототипа*).

(12) Большинство иностранных фамилий являются незнакомыми словами и предсказываются неверно. Как правило, генерируется одна или несколько ложных гипотез, например, изменяемое **крипка* (С,жр,но) от *Крипке*; неизменяемое **петри* (С,жр,но) и **петрить* (Глаг) от *Петри.* Во всех случаях отсутствует лексический признак “фам”.

Неполнота определяется в ситуации, когда термин или компонент термина в списке не представлен ни одним кандидатом.

(13) Большинство акронимов (терминологических аббревиатур, используемых в качестве эквивалентов сложных и громоздких терминов) отсутствуют в списке: из пяти терминов этого типа, представленных в словаре по ИИ, в универсальном словаре имеется только термин *СУБД.* Термин *АСУ* предсказывается неверно, остальные не извлекаются.

(14) Леммы притяжательных прилагательных на *-ово, -ево* от иностранных фамилий, регулярно употребляемые в составе терминов (*ламанов граф*), в словнике отсутствуют.

(15) Отсутствие предсказаний для некоторых фамилий, например, *Черч*.

3.3.2. Особенности выделения терминологических словосочетаний

Значительную часть общего числа терминов, включенных в словники исследуемых источников, составляют терминологические словосочетания (56,3% в словаре по теории графов и 78% в словаре о искусственному интеллекту). Их извлечение основано на синтаксических моделях, предопределенных в качестве правил формирования (извлечения и нормализации) СК в технологии Клан. Ниже приводится типизированный перечень ошибок, выявленных в результате эксперимента по извлечению многословных терминов. Все ошибки можно подразделить на три группы: неполнота, избыточность и некорректность.

Неполнота сборки термина имеет целый ряд причин, таких как наличие нелексических компонентов в составе термина, особенности репрезентации термина в тексте, неполнота стандартных моделей, отсутствие моделей для словосочетаний большой длины и инверсных словосочетаний.

А. Нелексические компоненты в составе термина.

(16) Использование символов или цифровых обозначений числа в позиции приложения *регулярный граф степени 0*.

Б. Разделители в составе термина.

(17) Использование кавычек, запятых (при осложнении постпозитивным причастным оборотом) и других разделителей: *поиск типа “сперва вглубь”*; *система, основанная на знаниях*; *область предметная, плохо структурированная*; *и/или граф*.

(18) Использование скобок (уточнение заглавного термина) *глубинная структура (предложения)*; *индукция полная (математическая)*.

(19) Разрыв термина разметочными тегами (если не рассматривать эту ситуацию как ввод релятора для различения омонимичных терминов):

- *Длинамаршрута— количество рёбер в маршруте (с повторениями)*;
- *Длинапути— число дуг пути (или сумма длин его дуг, если последние заданы)*.

В. Неполнота термина в тексте, связанная с особенностями его текстовой репрезентации, в частности, редукцией части термина (эллипсис) и использованием разрывающих термин элементов разметки.

(20) Редукция многословного термина в левой части словарной статьи:

Полным двудольнымназываетсядвудольный граф, в котором ….

(21) Редукция многословного термина в правой части словарной статьи:

- *ориентированный<a>граф*,
- *отношение называется<a>нетранзитивным, а если не выполняется ни для какой, тройки элементов, то -<a>антитранзитивным*.

Г. Неполнота стандартных моделей.

(22) Наличие примыкающих неизменяемых зависимых (наречия, частицы): *вполне несвязный граф, правила де моргана*.

(23) Наличие управляемых зависимых, отличных от генитивных: *система управления производством*.

(24) Возможность употребления страдательного причастия в адъективных позициях, не учтенная в стандартных моделях с прилагательным, кроме П+С, например, *закон исключенного третьего* по модели С+Прд+Срд.

(25) Возможность субстантивации адъективных компонентов (прилагательных и причастий), не учтенная в стандартных моделях с существительным: *дерево составляющих.*

Д. Нестандартная структура термина (длинные и инверсные термины).

(26) Именные группы большой длины при наличии вложенных именных групп в составе сложной *закон снятия двойного отрицания, компонента сильной связности графа*. В этой ситуации существуют полные покрытия стандартными моделями с пересечением и без пересечения: *закон снятия, двойное отрицание, снятие двойного отрицания*;

(27) Именные группы с инверсией позиции согласованного прилагательного. Стандартная инверсная модель С+П покрывает подавляющее большинство инвертированных терминов. Однако ряд случаев инверсии остается неучтенным. Их можно рассматривать либо как инверсный вариант исходной модели П+С+Срд (*система управления автоматизированная*) или П+П+С (*сеть семантическая интенсиональная. унификатор наибольший общий,*), либо как комбинацию исходных моделей, П+С и С+Срд+Срд (*язык представления знаний логический*), а также П+С и С+Прд+Срд (*система пятого поколения вычислительная*).

Избыточность в списке многословных терминов.

(28) Омонимия словоформ в текстовых позициях, удовлетворяющих условиям сборки по стандартным моделям; в результате наряду с верным термином формируются его “ложные” дубликаты:

- *база знаний – *баз знаний* (4 варианта сборки в позиции *<h3>база знаний</h3>* на основе омонимии словоформы *база* по роду и словоформы *знаний*, реализующей два стилистических варианта);
- *гомеоморфный граф – *гомеоморфная графа* (3 варианта сборки в позиции *гомеоморфные графы* на базе омонимии словоформы *графы* по роду и одушевленности);
- *тип данных, ациклический граф, диаметр графа, обработка естественного языка* (омонимия по одушевленности);
- *логика здравого смысла – *логик здравого смысла* (омонимия по роду);

- *машина параллельного вывода – *машин параллельного вывода, ориентированный граф – *ориентировать граф* (частеречная омонимия);
- *автомат линейно-ограниченный – *автомат линейный-ограниченный* (вариативность предсказаний леммы дефисного прилагательного).

Некорректность сборки термина по стандартной модели распадается на два подтипа.

А. Некорректная нормализация.

(29) Неверная нормализация по модели П+С, где в позиции П – страдательное причастие: **пометить граф* (вместо *помеченный граф*), **расширить сеть* (вместо *расширенная сеть*), **породить подграф* (вместо *порожденный подграф*).

(30) Нетерминологичность стандартной нормализации по числу для множественных понятий – ср. (9): *кратные ребра* (несколькоребер,инцидентныходной и той же паре вершин); *гомеоморфные графы* (графы, получаемые из одного графа с помощью последовательности подразбиений ребер).

(31) Случаи рассогласования: **метод прямого волны* vs. *метод обратной волны*.

(32) Случаи неверного выбора падежной формы: **рассуждение по аналогия* vs. *рассуждение по умолчанию*.

Б. Некорректный термин определяется в ситуации, когда термин построен по стандартной модели на базе неверно (по совокупности параметров) предсказанных компонентов. В отличие от предыдущего типа, корректный вариант в списке словосочетаний отсутствует.

(33) Некорректная лемма дефисного прилагательного – **аппаратный-программное средство*.

(34) При наличии грамматически правильного предсказания компонентов-прилагательных, таких как *абдуктивный, деонтический, когнитивный*, словосочетания *вывод абдуктивный, логика деонтическая, диссонанс когнитивный* строятся по модели С+Срд на базе неверного предсказания части речи и класса словоизменения (неизменяемое существительное). Словосочетания *структура/графика/психология/модель когнитивная* строятся по модели С+Срд на базе компонентов-прилагательных, неверно предсказанных как неизменяемые существительные с леммой *когнитивная*. Заметим, что для терминов с препозицией прилагательного *когнитивная наука/психология/процесс/структура* используется верное предсказание и корректная сборка по модели А+С, что позволяет предположить ошибки перебора вариантов при наличии альтернативных предсказаний с частеречной омонимией.

(35) В отсутствие грамматически правильного предсказания компонента термина (например, притяжательных прилагательных от фамилий), словосочетания собираются по

неверным моделям С+Сим (**эйлеров цепь*), С+Срд (**гамильтон граф,*ламан граф),* либо вообще не собираются (отсутствуют *эйлеров цикл, эйлеров путь, гамильтонов цикл, гамильтонов путь*).

3.4. Эвристические методы

Рассмотренные выше проблемы выявили необходимость разработки специальных методов, позволяющих разрешать неоднозначность, формировать «недостающие» термины, улучшать качество синтеза нормальной формы терминов и т.п.

3.4.1. Сравнение с «эталоном»

Особенностью рассматриваемого жанра электронного ресурса является структура словарных статей, в которых выделяются левые части, с определяемыми терминами, и правые части – с толкованиями. Практически всегда, за исключением случаев использования предиката *называется*, в левой части термин представлен в «эталонном» виде. Под эталонном понимается нормализованная форма термина, принятая в данной предметной области, которая, как правило, представляет собой именную группу, главным термином которой является существительное в именительном падеже. Указанная жанровая особенность позволяет автоматически выделить границы терминов и построить эталонный словарь терминов, представляющих собой эталонные строковые выражения.

Анализ и корректировка базового словаря (однословных и многословных терминов, полученных системой Клан) опирается на информацию о границах эталонных терминов и сопоставление эталонного выражения с найденными в данных границах однословными и многословными терминами базового словаря.

Сравнение найденных терминов с эталонным образцом позволяет разрешить следующие проблемы путем выбора правильного варианта, совпадающего по норме с эталоном (в скобках будем указывать номера из классификатора проблем выделения терминов, представленного в п. 3.3):

a) однословная омонимия (1-4),
b) вариативность многословного термина (28), причиной которой чаще всего является вариативность компонентов (3-6).

В остальных случаях, когда правильный вариант в списке терминов отсутствует, необходимо проанализировать компонентный состав эталонного термина и сопоставленных ему вариантов.

При сравнении границ термина и его нормальной формы с эталонными можно выделить следующие виды несоответствия.

a) Неполнота определения термина в случаях, когда длина термина меньше эталонного выражения (7,8,16-19).
b) Неполнота определения термина в случаях, когда не нашлось подходящей модели сборки. Такая ситуация характеризуется покрытием эталонного выражения совокупностью терминов (22-27).
c) Формирование (синтез) неправильной нормальной формы термина (8-12,29-35) в границах эталонного термина.
    - Ошибки выбора формы зависимого компонента при стандартной нормализации (29,31,32).
    - Нетерминологичность стандартной нормализации (9,10,30).
    - Ошибки нормализации первой части дефисных терминов, как правило, прилагательных (8,11).
    - Ошибки предсказания компонента термина (12) – при отсутствии правильного варианта (35) либо при выборе неправильного варианта (34).
d) Отсутствие термина или компонента термина, определяемого в границах эталонного термина, в базовом словаре (13-15).

Предлагаемый метод решения заключается в создании корректных составных терминов на основе шаблонных описаний и опирается на разработанную типизацию ошибок формирования терминов (п.3.3.).

3.4.2. Лексико-синтаксические шаблоны

Для решения задач описания сложных составных терминов, а также для проверки различных эвристик был предложен язык лексико-синтаксических шаблонов.

В работе [8] лексико-синтаксический шаблон определяется как модель (структурный образец) языковой конструкции, в котором указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем). В [1] предусмотрена возможность описания сложного шаблона с использованием именованных подшаблонов, таких как грамматически согласованная именная группа. Т.о. шаблоны представляют собой формальные описания лингвистических свойств языковых выражений, формулируются экспертами (лингвистами или специалистами по ПО) и

используются для автоматического выделения конструкций в тексте (извлечение именованных сущностей, терминов, их связей и т.п.). Так, в недавней работе [3] шаблоны применяются в задаче извлечения и отбора терминов для предметных указателей.

Предлагаемый язык шаблонов характеризуется следующими отличительными особенностями:

1) поддержка логического комбинирования лексических элементов – элементами конструкции могут выступать однословные или многословные термины (формируемые технологией Клан), элементы разметки или произвольные символьные последовательности;
2) поддержка альтернативных вариантов при описании элементов языковой конструкции;
3) именование шаблонов, позволяющее определять все более сложные шаблоны на основе уже известных шаблонов и тем самым постепенно наращивать их мощность.

Шаблон определяет структурный образец целевой языковой конструкции, ее лексический состав и поверхностно-синтаксические свойства. Объекту, найденному по шаблону, приписываются соответствующие признаки – имя (совпадающее с именем шаблона), грамматические характеристики, вычисляемые на основе пересечения множеств характеристик составляющих элементов (если они присутствуют), и формальные (позиционные) признаки. Синтаксис языка лексико-синтаксических шаблонов, как и язык описания сегментов, основан на языке JSON, что позволяет применять стандартные парсеры и редакторы для просмотра и отладки шаблонов.

Особенностью нашей задачи является необходимость автоматической генерации шаблонов, представляющих многокомпонентные термины предметной области.

Общая структура шаблона, представляющего термин, выглядит следующим образом:

```
{
   "name": "ЭТАЛОННОЕ_ИМЯ_ТЕРМИНА",
   "segments": [  // множество разнотипных элементов
            // список возможных типов элементов
            "ph/МНОГОСЛОВНЫЙ_ТЕРМИН",
            "w/ОДНОСЛОВНЫЙ_ТЕРМИН/индекс_морф_класса",
            "s/СТРОКА",
            "ИМЯ_ШАБЛОНА"
    ],
   "type": "sequence"/ "or" // последовательность или альтернативы
}
```

Данная схема демонстрирует поля json-формата для внесения данных и приводит возможные значения, используемые для описания шаблонов.

Генерация шаблонов осуществляется за счет наличия эталонного термина, формирующего имя шаблона, границ эталонного термина в левой части определения (которые обычно выделяются с помощью тегов) и покрытия данного фрагмента текста терминами словаря Клан, из которых формируется состав шаблона (тип *sequence*). Для представления альтернатив формируется шаблон с типом *or*.

Реализуются следующие варианты формирования шаблона.

A. Шаблон формируется из вложенных терминов (22-27).

```
{
    "name": "закон снятия двойного отрицания",
    "segments": [
            "w/закон/e",
            "s/ ",
            "w/снятие/k",
            "s/ ",
            "ph/двойное отрицание",
     ],
     "type": "sequence"
}
```

B. Шаблон формируется из комбинации строк и/или вложенных терминов (7, 8, 13, 15-17).

```
{
    "name": "k-дольный граф",
    "segments": [
            "s/k-",
            "ph/дольный граф"
    ],
    "type": "sequence"
}
```

Частным случаем такого шаблона является разбиение существующих терминов на компоненты (11, 33, 35).

```
{
    "name":"граф-звезда"
     "segments": [
            "s/граф-",
             "w/звезда/c"
     ],
     "type": "sequence"
}
```

C. Шаблон формируется из полного термина, но с другим эталонным именем (9, 10, 30).

```
{
    "name": "кратные ребра",
    "segments": [
            "ph/ кратное ребро"
    ],
    "type": "sequence"
}
```

D. Шаблон формирует список альтернатив (синонимов):

```
{
    "name": "автоматизированная система управления",
    "segments": [
            "ph/автоматизированная система управления",
            "s/АСУ"
    ],
    "type": "or"
}
```

или альтернативу с уточнением (18):

```
{
    "name": "глубинная структура (предложения)",
    "segments": [
            "ph/глубинная структура",
            "ph/глубинная структура предложения"
    ],
    "type": "or"
}
```

E. Шаблон формируется с использованием вложенных шаблонов.

```
{
    "name":"автоматизированная система управления предприятием"
     "segments": [
            "автоматизированная система управления",
            "s/ ",
             "w/предприятие/k"
     ],
     "type": "sequence"
}
```

В общем случае для формирования шаблона выбирается вариант A, при наличии полного покрытия терминами эталонного выражения, и вариант B – при неполном покрытии. Вариант C выбирается при отсутствии «правильных» терминов (т.е. терминов или их компонентов, совпавших с эталоном) и, как правило, комбинируется с A путем создания альтернативы термина с разбиением. Альтернативы создаются с помощью варианта D. Если термин, для которого уже создан шаблон, входит в состав другого термина, то используется шаблон E.

Термины специального вида требуют применения специализированных методов их обработки.

3.4.3. Методы обработки «особых» терминов

Метод обработки определения, использующего предикат называния. Данный метод позволит сформировать термин с помощью лексико-синтаксического шаблона на основе структурного критерия. Основная идея заключается в формировании термина как шаблона, включающего элементы, расположенные слева и справа от слова *называется (обозначается)*. При этом случай, когда в качестве компонентов шаблона выступают термины базового словаря, не вызывает трудности (вариант формирования шаблона А). Проблемными являются ситуации омонимии (здесь уже не получится использовать эталон) и наличия повторов в левой и правой частях.

Метод обработки инверсных терминов. Данный метод применяется для энциклопедических источников, в которых заглавные термины имеют инвертированный порядок слов.

В рамках метода осуществляется сборка групп прилагательных, стоящих после главного существительного, и их перестановка в начало термина перед главным существительным. Итоговый термин формируется как шаблон, включающий множество альтернатив – инверсный термин и варианты с перестановкой (если прилагательных несколько, то вариантов тоже может быть несколько).

```
{
    "name": "ЗНАНИЯ ПРАГМАТИЧЕСКИЕ",
    "segments": [
            "ph/знание прагматическое",
            "ph/прагматическое знание"
    ],
    "type": "or"
}
```

Метод обработки дефисных предсказаний. Следует отметить, что выявить наличие ошибки в предсказании дефисных терминов удается не всегда. Существуют случаи, когда нормальная форма термина предсказывается корректно (т.е. совпадает с эталонной), но парадигма неправильная (например, *граф-звезда* из случая (11)), что в дальнейшем приведет к проблемам при поиске вхождений данного термина в текстах. Поэтому для всех предсказанных терминов с дефисом осуществляется общая схема обработки по следующему принципу.

Если для термина выполняются следующие условия:

- термин – *существительное*,
- первая часть дефисного слова – *существительное*,
- у первой части нет признака *неизменяемый*

⇒ Формируется шаблон с разбивкой существующих терминов (вариант B формирования шаблонов).

В остальных случаях применяется стандартная процедура сравнения с эталоном и либо выбор совпадающего с эталоном варианта, либо формирование шаблона из частей, если совпадения с эталоном нет.

Метод синтеза сокращения. Как правило, сокращения в левых частях определений указываются в скобках. Однако, помимо сокращений, в скобках может быть указано уточнение термина (18), либо набор его синонимов. Возникает необходимость проверки, является ли конструкция в скобках акронимом или графическим сокращением (13). Такая проверка осуществляется путем сопоставления букв предполагаемого сокращения с начальными буквами слов термина. Если конструкция оказывается сокращением, то формируется шаблон с альтернативами (вариант формирования D).

Метод обработки терминов в скобках. Данный метод применяется для анализа терминов, представленных в скобках в левых частях словарных статей.

Выражение в скобках:

- множество терминов через запятую – считаем, что это синонимы;
- конструкция, включающая точки, – сокращение;
- однословный термин, содержащий заглавные буквы, – термин проверяется на сокращение методом синтеза;
- термин, вершина которого является существительным в именительном падеже, – синоним;
- другое – уточнение термина.

Если термин в скобках – уточнение или сокращение, то формируется шаблон с альтернативой (вариант формирования D). Имя шаблона будет содержать значения в скобках в случае уточнения.

Выявление синонимов требует специальной обработки, рассмотрение этой проблемы не входит в задачи данной работы.

3.4.4. Дополнительные методы формирования терминов

Методы, указанные в данном подразделе, пока не получили достаточной экспериментальной оценки. Мы приводим их исключительно для представления более полной картины покрытия выделенных видов проблем и методов их решения.

А. Метод разрешения омонимии (вариативности) на основе статистики. Идея данного метода связана с тем, что у омонимичных вариантов термина могут быть разные парадигмы,

соответственно, анализ текста всей энциклопедии (или текстов данной ПО в целом) может дать статистический критерий – правильный вариант будет встречаться чаще.

B. Дополнительной обработки требует ситуация тегов, идущих подряд. Теги, если они присутствуют в тексте, служат границами терминов. Ошибки в расстановке тегов или внутренние теги, призванные выделить термин внутри термина, приводят к формированию «ложных» границ терминов. Однако явного критерия ошибочности в расстановке тегов пока не обнаружено.

C. Метод приоритетов. В зависимости от предметной области, пользователь предварительно может указать «приоритеты» грамматических признаков, например, исключать термины-глаголы (полностью или при неоднозначности, как в (4)), исключать одушевленные существительные, не рассматривать фамилии и т.п.

4. Результаты экспериментов

Метод автоматической оценки правильности формирования словаря терминов опирается на сравнение полученного словаря с эталонным.

В таблице 1 приведено сравнение основных качественных показателей обработки двух энциклопедических словарей базовыми методами (базовый словарь) и с добавлением эвристик (итоговый словарь), а также произведена оценка качества терминологического ядра формируемой онтологии. Терминологическое ядро – это множество имен понятий, полученных из итогового словаря путем объединения альтернатив, т.е. терминов, найденных в одной позиции в левой части словарных статей источников, с помощью лексико-синтаксических шаблонов.

Таблица 1. Результаты экспериментов.

Источник	Базовый словарь		Итоговый словарь		Терминологическое ядро онтологии		
	Полнота	Точность	Полнота	Точность	Полнота	Точность	F-мера
Теория графов	0,83	0,36	0,90	0,45	0,90	0,78	0,84
Искусственный интеллект	0,86	0,46	0,97	0,51	0,996	0,998	0,997

Как видно из приведенной таблицы, применение эвристических методов дало возможность значительно улучшить качество построения терминологический словарей. Более низкие показатели для словаря по теории графов объясняются тем, что не удалось в полной мере

решить проблему снятия омонимии для существительных, различающихся по одушевленности, например, для термина *граф* и всех образуемых от него однословных (*мультиграф, орграф, подграф, псевдограф* и т.п.) и многословных терминов.

Заключение

В работе предложена методология построения терминологического ядра предметной области на базе электронных энциклопедических источников данных. Особенностью предлагаемого похода является тщательный анализ структуры термина, распознавание ошибок на базе их лингвистической классификации, автоматическая генерация лексико-синтаксических шаблонов, представляющих многокомпонентные термины, и использование набора эвристических методов обработки «особых» терминов. Применение указанного подхода позволило значительно повысить качество создаваемых словарей по сравнению с классическими методами обработки текста (морфологическим и поверхностно-синтаксическим анализом).

Создаваемое терминологическое ядро является основой для дальнейшего построения онтологии предметной области, качество которой в значительной степени зависит от корректности выделения терминов. Следующим шагом исследования является применение созданных словарей для анализа правой части словарных статей источников и автоматическое выявление синонимов и родовидовых отношений между терминами предметной области.

Работа выполнена при поддержке Президиума СО РАН (Блок 36.1. Комплексной программы ФНИ СО РАН II.1).

Список литературы

1. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: тр. межд. конференции Диалог‘2007. М.: Изд-во РГГУ, 2007. С. 70-75.
2. Большакова Е.И., Васильева Н.Э. Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием: труды конференции. М., 2008. Т. 2. С. 174-182.
3. Большакова Е.И., Иванов К.М. Выделение терминов и их связей для предметного указателя научного текста // Сборник трудов XVI Национальной конференции по искусственному интеллекту с международным участием (КИИ-2018). Т1. М.: РКП, 2018. С. 253-261.

4. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». 2003. С. 201-210.
5. Захаров В.П., Хохлова М.В. Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. С.-Петербург: Изд-во С-Петерб. гос. ун-та, 2014. Вып.10. С. 182–200.
6. Киселёв Ю.А., Поршнев С.В., Мухин М.Ю. Метод извлечения родо-видовых отношений между существительными из определений толковых словарей // Программная инженерия. 2015. №6. С. 34–40.
7. Лезин Г.В., Клименко Е.Н., Силина Е.Ф. Онтологическая интерпретация дефиниций терминологического словаря // Прикладная лингвистика в науке и образовании: сб. трудов VII межд. конференции. СПб.: «Книжный дом», 2014. С. 50–54.
8. Митрофанова О.А., Захаров В.П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов // Компьютерная лингвистика и интеллектуальные технологии: тр. межд. конференции «Диалог–2009». М.: 2009. С. 321-328.
9. Рабчевский Е., Булатова Г., Шарафутдинов И. Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий // Труды десятой всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» –RCDL'2008. Дубна: ОИЯИ, 2008. С. 103-106.
10. Рубашкин В. Ш., Бочаров В. В., Пивоварова Л. М., Чуприн Б. Ю. Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Вып. 9 (16). М.: Изд–во РГГУ, 2010. С. 413–418.
11. Рубашкин В.Ш., Капустин В.А. Использование определений терминов в энциклопедических словарях для автоматизированного пополнения онтологий // XI Всероссийская объединенная конференция «Интернет и современное общество». СПб., 2008. С.32-39.
12. Сидорова Е. А. Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: тр. межд. конф. «Диалог–2008». М.: 2008. Вып. 7 (14). М.: Изд–во РГГУ, 2008. С. 475-481.
13. Сокирко А.В. Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. межд. конференции Диалог’2004 / Под ред. И.М.Кобозевой, А.С. Нариньяни, В.П. Селегея. М.: Наука, 2004. С. 559-564.
14. Chodorow M.S., Byrd R.J., Heidorn G.E. Extracting semantic hierarchies from a large on-line dictionary// Proc. of the 23rd Annual Conference of the Association for Computational Linguistics, Chicago, 1985. P. 299–304.

15. Dolan W., Vanderwende L., Richardson S. Automatically Deriving Structured Knowledge Bases From on-Line Dictionaries // Proc. of the Pacific Association for Com putational Linguistics. PACL Press, 1993. P. 5–14.
16. Navigli R., Velardi P. From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions // Proc. of the conference on Ontology Learning and population: Bridging the Gap between Text and Knowledge. IOS Press Amsterdam. 2008. P. 71–78.