УДК 81’33:004.822

# Генерация лексико-синтаксических паттернов онтологического проектирования на основе вопросов оценки компетенции

*Овчинникова К. А. (Новосибирский национальный исследовательский государственный университет),*

*Сидорова Е. А. (Институт систем информатики СО РАН)*

Работа посвящена исследованию проблем автоматизации создания онтологий научных предметных областей с применением методов автоматического анализа текстов на естественном языке. Целью работы является разработка методов автоматической генерации лексико-синтаксических шаблонов для извлечения информации и пополнения онтологий на основе анализа содержательных паттернов онтологического проектирования для научных областей знаний, разрабатываемых в рамках концепции Semantic Web. Паттерны онтологического проектирования представляют собой структурированной описание понятий верхнего уровня в терминах классов, атрибутов и отношений, а также включают вопросы оценки компетенции на естественном языке, служащие для понимания и корректной интерпретации свойств и связей понятия пользователями. В статье предложен подход к генерации лексико-синтаксических паттернов на основе вопросов оценки компетенции. Процесс генерации лексико-синтаксических паттернов включает генерацию предметного словаря, выделение сущностей онтологии и формирование структуры паттернов на основе свойств Data Property и Object Property, и генерацию семантических, грамматических и позиционных ограничений. Вопросы оценки компетенции используются для выявления грамматических и позиционных ограничений, необходимых для поиска онтологических отношений в текстах. Для эксперимента использовалась онтология «Поддержка принятия решений в слабоформализованных областях» и корпус научных текстов той же предметной области. В ходе эксперимента получены следующие результаты: степень неоднозначности сгенерированных шаблонов - 1.5, F1-мера оценки качество поиска атрибутов и отношений объектов - F1-мера составила 0,77 для атрибутов и 0,55 для отношений соответственно. Сравнение результатов, полученных для шаблонов без грамматических ограничений, и результатов, полученных для шаблонов с

грамматическими ограничениями, показало, что добавление ограничений существенно улучшает качество извлечение объектов онтологии.

***Ключевые слова:*** *лексико-синтаксический паттерн, генерация паттернов, вопросы оценки компетентности, пополнение онтологии, онтология научной деятельности, паттерны онтологического проектирования, извлечение информации.*

# 1. Введение

Под извлечением информации понимается процесс автоматического извлечения полезного материала из текстов некоторой предметной области, его обработка и использование. Интерес к этой процедуре повышается благодаря большому объему неструктурированной информации в Интернете.

Извлечение информации из текстов узкоспециализированных научных областей представляет наибольший интерес из-за проблемы недостаточного количества размеченных данных. Это усложняет использование методов машинного обучения и глубокого обучения и является причиной использования других методов. Альтернативным подходом является использование методов на основе знаний и их дальнейшая интеграция с методами машинного обучения.

Семантическая паутина (англ. Semantic Web) [6] — часть глобальной концепции развития сети Интернет, целью которой является реализация возможности машинной обработки информации, доступной во Всемирной паутине. Основной акцент концепции делается на работе с метаданными, однозначно характеризующими свойства и содержание ресурсов. Онтология и язык ее описания является одним из способов стандартизации представления знаний и информации, поддерживающем машинную обработку. Развитие этих инструментов может быть объяснено востребованностью онтологий [8, 16] как способа стандартизации знаний о предметных областях, хранения, навигации и поиска хорошо структурированных данных. Унификация средств представления онтологий и создание банка готовых решений [15], включающих стандартные образцы сущностей онтологий, ставит перед исследователями новые задачи, а именно необходимость создать механизмы использования образцов готовых решений для проектирования и разработки пользовательских онтологий, а также инструменты для ее автоматизированного пополнения.

В задачах пополнения онтологий используются разные методы: методы машинного обучения (правила кластеризации или ассоциации) [9, 10], итерационные методы с использованием графов [18] и на основе шаблонов (паттернов) [5, 12, 14].

Для упрощения процесса разработки и дополнения онтологий в некоторых исследованиях [7, 19] уже более десяти лет используется подход, основанный на использовании паттернов онтологического проектирования (ОП). Они представляют собой задокументированные описания проверенных решений общих проблем онтологического моделирования. Одним из видов паттернов ОП являются паттерны содержания, описывающие фрагменты онтологии предметной области. Авторы методологии XD (eXtreme Design methodology) [6] предлагают добавлять в паттерн содержания не только описание одного онтологического класса, его атрибутов и отношений, но и вопросы оценки компетенции (ВОК). Вопросы оценки компетенции (ВОК) - это выраженные на естественном языке вопросы к структурным элементам класса, служащие для понимания и корректной интерпретации свойств и связей понятия пользователями. Еще одним типом паттернов ОП являются лексико-синтаксические шаблоны (ЛСП). Эти паттерны представляют собой структурные образцы конструкций языка, отражающие их лексические и поверхностные синтаксические свойства. ЛСП определяют отображение языковых единиц текста в онтологические структуры. Исследователи могут использовать ЛСП при решении задачи автоматического построения онтологий на основе корпуса текстов на естественном языке.

Подход представленный в работе [14] описывает использование лексико-синтаксических паттернов, соответствующих онтологическим паттернам проектирования, для пополнения онтологии. В отличие от паттернов Херст [11], они носят более общий характер, что позволяет охватывать большее количество вхождений сущностей в текстах. Такие паттерны показывают высокую полноту и низкую точность, тогда как паттерны Херст, наоборот, демонстрируют высокую точность извлечения информации, но низкую полноту. Разработка таких шаблонов является достаточно кропотливой работой, поэтому существует необходимость автоматизации данного процесса.

В связи с частым отсутствием размеченных данных одним из активно развиваемых подходов является генерация закономерностей на основе небольшого количества информации, представленной в справочных материалах, толковых словарях или непосредственно в онтологиях. В данной работе для автоматического формирования лексико-синтаксических моделей предлагается использовать вопросы оценки компетентности и научный словарь. Экспериментальное исследование проводится на материале, представленном на портале «Поддержка принятия решений в слабоформализованных областях» (https://uniserv.iis.nsk.su/rdms/), и корпусе научных текстов той же предметной области.

## 2. Подход к автоматизации пополнения онтологии

В ходе работы [19] были разработаны паттерны содержания для некоторых понятий, характерных для большинства научных предметных областей: *Объект исследования, Предмет исследования, Метод, Задача, Раздел науки, Научный результат, Деятельность, Проект, Персона, Организация, Публикация, Информационный ресурс и др*. Также был определен набор вопросов оценки компетенции для каждого из этих паттернов. С помощью них были выявлены ограничения для паттернов и описаны требования к ним, которые были представлены в виде аксиом и ограничений. Для каждого паттерна, представляющего концепт научной предметной области (НПО), мы определили набор ключевых атрибутов, однозначно идентифицирующих конкретный экземпляр класса.

Для реализации автоматического пополнения онтологии для каждого паттерна содержания строится набор лексико-синтаксических шаблонов, который описывает различные способы представления информации в научных текстах на основе извлеченной информации.

Рассматривая ЛСП как инструмент пополнения онтологии, мы определили две ключевые задачи для достижения поставленной цели. Первой задачей является извлечение имен объектов (в том числе не входящих в словарь) и значений их атрибутов. Во-вторых, это создание объектов на основе структуры классов онтологий. В соответствии с этими задачами были выделена два типа ЛСП: терминологический и информационный. В работе [4] предложена архитектура системы для пополнения онтологии на основе лексико-синтаксических паттернов, реализующая алгоритм пополнения. Предлагается использовать следующие технологии: система извлечения предметной лексики из текстов и построения словарей KLAN [3], система анализа текстов на основе шаблонов PatTerm [17], система анализа текстов FATON [1]. Взаимодействие с онтологией обеспечивает специально разработанный модуль, использующий средства поддержки онтологически-ориентированного программирования из библиотеки owlready2 [13].

ЛСП автоматически строятся на основе словарей общенаучной и предметной лексики и актуальной версии онтологии научной предметной области. На Рис. 1 представлена схема взаимосвязей компонентов системы, участвующих в генерации ЛСП.

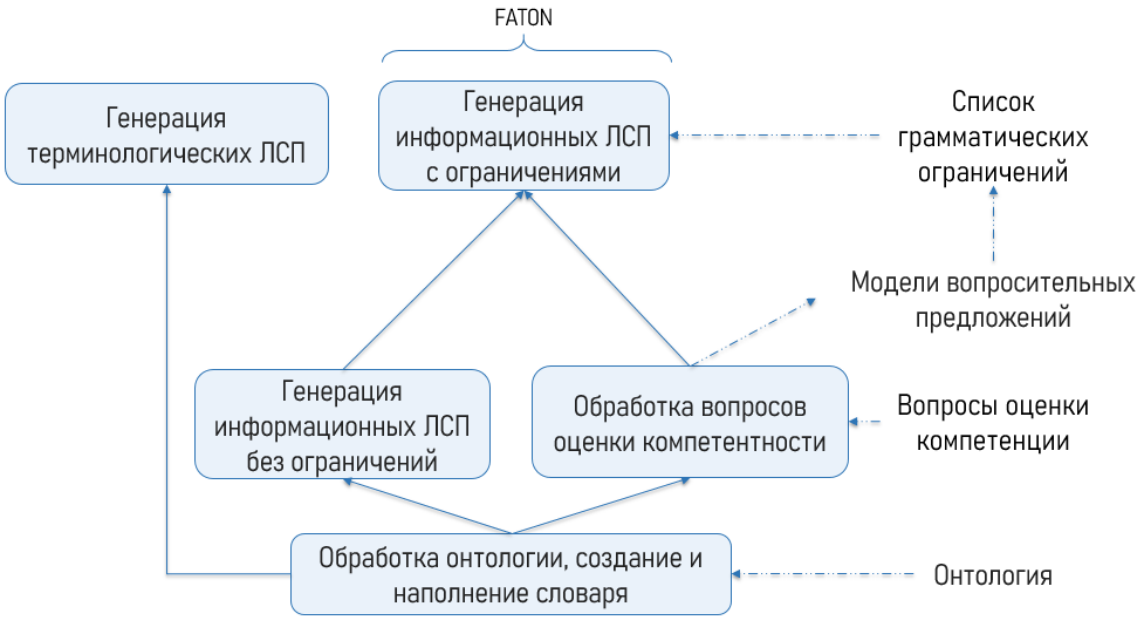


*Рис. 1. Схема взаимосвязей компонентов системы, участвующих в генерации ЛСП.*

На первом этапе генерации происходит обработка онтологии, создание и наполнение словаря. После чего происходит извлечение имен классов объектов, атрибутов и отношений из онтологии.

Данный этап подробно рассмотрен в работе [3]. На нем формируются Т-ЛСП с использованием индикаторных термов, полученных на основе онтологии, содержащие переменные с заданными свойствами. Означивание таких переменных конкретными фрагментами текста позволяет получить новые значения. В результате создаются многословные термины для словаря и шаблоны для извлечения терминов.

На следующем этапе происходит обработка вопросов оценки компетентности: удаляются лишние символы, строятся модели вопросительных предложений и извлекаются граммемы. Затем генерируются И-ЛСП, для которых требуется информация о ключевых атрибутах классов онтологий, особенно для объектов, создающих шаблоны (например, атрибут Название для Метода исследования), и отношениях с другими классами.

## 2.1. Словарь научной лексики

Для автоматического пополнения онтологии с помощью ЛСП важно обеспечить извлечение из текста специфических терминов данной научной предметной области. Для этого предлагается использовать словарь научной лексики, который включает не только общенаучную лексику, но и предметную.

Для создания общенаучного словаря было необходимо собрать тематически-нейтральный корпус научных текстов. С этой целью тексты распределялись по 5 научным коллекциям, соответствующим гуманитарным, естественным, техническим, общественным и точным наукам. Всего для проведения исследования было собрано 100 русскоязычных научных публикаций, относящихся к списку ВАК (Высшая аттестационная комиссия) или библиографической базе научных публикаций RSCI (Russian Science Citation Index) или базе научной периодики Scopus. Общий объем корпуса составляет 370,8 тыс. терминов.

Для проведения исследования был создан корпус, включающий 100 русскоязычных научных публикаций. Каждая из них относится к списку ВАК (Высшая аттестационная комиссия), библиографической базе научных публикаций RSCI (Russian Science Citation Index) или базе научной периодики Scopus. Общий объем корпуса составляет 370,8 тыс. терминов.

Следующий этап заключается в создании и обработке общенаучного словаря, который автоматически строится на основе онтологии и корпуса текстов. Для разметки научных терминов на основе анализа семантических значений, представленных в корпусе, были выделены 8 универсальных классов: *Восприятие, Ментальные, Существование, Сущность, Создание, Деятельность, Результат* и *Применение*. Такое разделение основано на семантических значениях, выраженных в предложениях. Например, универсальный класс *Деятельность* включает в себя глаголы несовершенного вида, которые определяют разворачивающееся действие, совершаемое с помощью, например, *Метода исследования*, а класс *Результат* – действие с акцентом на результат, получаемый, например, с помощью применения *Метода исследования*. Универсальные классы могут использоваться в паттернах для нахождения синонимов или в случае стандартных (универсальных) способов выражения отношений.

Предметный словарь создается как расширение словаря общенаучной лексики и включает в себя слова и словосочетания (термины), относящиеся к конкретной предметной области. Система предметно-ориентированных классов основана на структуре онтологии НПО, отражая иерархию ее объектов и отношений. Имена классов терминов генерируются на основе названий онтологических элементов в соответствии с шаблоном: <название_класса.название_отношения> или <название_класса. название_аттрибута>. Например, отношение *Метод исследования.используется_в*.

В словарной статье хранится вся информация, необходимая для извлечения термина из текста и поддержки следующих этапов анализа текста. Каждый термин словаря, найденный в

тексте, снабжен морфологической и семантической информацией, которые впоследствии используются при создании и применении ЛСП.

Научный словарь представляет собой интеграцию двух словарей: универсального и тематического. Поэтому он включает в себя две самостоятельные иерархии лексико-семантических классов: универсальную иерархию из словаря общенаучной лексики и предметно-ориентированную иерархию, основанную на онтологии (Рис. 2).

| ***Универсальный класс*** | ***Список лексем*** | ***Предметно-ориентированный класс*** |
|---|---|---|
| **ментальные** | объяснять, определять, трактовать, расценивать, рассматривать, … | Метод.представлен_на |
| **создание** | предложить, ввести, разработать, описать, создать | Метод.создан_в<br>Метод.имеет_автора |
| **применение** | применять, применяться, использовать, использоваться | Метод.используется_в<br>Метод.применяется_к |

*Рис. 2. Фрагмент лексико-семантических классов созданного словаря.*

Все словарные термины отмечены признаками из предметной и/или универсальной иерархии. Лексико-семантические признаки словаря используются при описании ЛСП как способа обозначения терминов предметной области науки с определенной семантикой.

## 2.2. Вопросы оценки компетентности

Анализ вопросов оценки компетентности (ВОК) позволяет выявить начальные синтаксические свойства языковых выражений, описывающих связи между понятиями предметной области, которые могут быть впоследствии уточнены на основе корпуса текстов [18]. Таким образом, вопросы оценки компетентности могут быть использованы при генерации лексико-синтаксических паттернов для уточнения синтаксических и позиционных ограничений.

Можно выделить 5 видов вопросов оценки компетентности:

- вопросы, не содержащие вопросительных слов (*«Применяется ли метод к объекту исследования?»*);
- вопросы, не содержащие значимых вопросительных слов (*«**Какой** объект исследования исследуется в деятельности?»*);
- вопросы, содержащие вопросительные слова, напрямую не обладающие семантикой, относящей его к атрибуту класса онтологии (*«**Когда** была дата начала проекта?»*);

- вопросы, содержащие вопросительные слова, напрямую не обладающие семантикой, относящей его к классу онтологии (*«**Кто** использует метод?»*);
- вопросы, содержащие вопросительные слова, не несущие семантической нагрузки, которую можно связать с некоторым классом онтологии (*«**Как** называется задача?»*).

Был составлен отдельный список вопросительных слов третьей и четвертой группы и добавлено соотношение с атрибутами или классами онтологии:

*где: Географическое место.Название,*

*когда: Информационный ресурс.Дата, Публикация.Дата*

*кто: Персона.Фамилия, Организация.Название*.

Для каждого отношения в онтологии были разработаны 1-3 вопроса оценки компетентности. Например, для отношения Метод.имеет автора были предложены следующие вопросы:

*Кто придумал метод?*

*Кем предложен метод?*

*Кто является автором метода?*

В данном случае все вопросы являются вопросами третьего типа, т.к. для обозначения неизвестного субъекта в русском языке обычно используется местоимение *кто* (и производные от него).

Для отношения *Задача.решается* был предложен только один вопрос:

*«Какая задача решается в разделе науки?»*

Наличие только одного варианта можно объяснить тем, что данное отношение не предполагает вариативности названия актантов в предложении, а данный предикат полно отражает отношение между ними.

Для каждого вопроса оценки компетентности строится модель вопросительного предложения.

Модель вопросительного предложения состоит из объектов, для которых известны соотнесенность с названием онтологического класса, атрибута или отношения и необходимые грамматические категории.

В нашем подходе предложено две модели. Формально их можно представить в следующем виде:

$$M1=\langle O1, Rel, O2\rangle \qquad (3.1)$$

– для связи Rel между объектами $O_1$ и $O_2$.

$$M2=\langle O,D\rangle \quad (3.2)$$

– для добавления атрибута D или создания объекта O.

Поясним некоторые обозначения:

O, $O_1$ и $O_2$ представляют собой наборы (Name, Grammatic), где Name – название онтологического класса, Grammatic – множество граммем,

D представляет собой набор (DataProperty, DataProperty_type, Grammatic), где DataProperty– название атрибута класса, DataProperty_type – тип атрибута, Grammatic – множество граммем,

Rel представляет собой набор (PWord, ObjectProperty, Grammatic), где PWord – конкретный предикат, ObjectProperty – онтологическое отношение, а Grammatic – множество граммем.

Анализ вопросов оценки компетентности позволяет выделять некоторые грамматические ограничения на паттерны.

### 2.2.1. Грамматические ограничения

Анализ вопросов оценки компетентности проводится по трем направлениям: анализ предиката и анализ его первого и второго актантов.

Была составлена таблица, в которой были описаны грамматические ограничения для вопросов оценки компетентности. Под грамматическими ограничениями мы понимаем морфологические свойства аргументов и синтаксические связи между ними.

Анализ вопросов оценки компетентности проводится для каждого вопроса с точки зрения граммем каждого релевантного слова. К релевантным словам относятся те, которые имеют непосредственное отношение к объектам онтологии: названия классов и атрибутов и предикаты. Каждому предикату сопоставляются его грамматические категории и название отношения в онтологии, а каждому актанту – грамматические категории и название класса или атрибута онтологии (Таблица 1).

Все морфологические категории слов рассматриваются с точки зрения релевантности и нерелевантности. Во время анализа выделяются нерелевантные грамматические категории. Например, было решено не учитывать ***лицо*** глагола. В вопросе *«Кто предложил метод исследования?»* глагол стоит в 3 лице, как и в любом другом предложении, в котором подлежащее выражено именем собственным (или именами собственными). Компонент *Персона* определяет объект одноименного класса, в котором подразумевается наличие имени собственного, поэтому, если подлежащее в предложении, из которого будет извлекаться

информация, будет выражено чем-то другим, то предложение не будет обрабатываться конструкцией <*Персона*, *Метод.имеет автора*, *Метод исследования*>, поэтому в данном случае указание на лицо глагола можно опустить.

Рассмотрим некоторые релевантные граммемы. Граммема ***времени*** релевантна для тех случаев, когда однозначно определено время их действия. Так, например, для атрибута *Дата* (создания) класса *Метод* в паттерн будет добавляться прошедшее время глагола. Граммема ***падежа*** у существительных и местоимений считается релевантной для всех вопросов.

Таким образом, был составлен список релевантных граммем для каждой части речи:

- Глагол: аспектуальность, переходность, число, время;
- Причастие: аспектуальность, залог, число, время
- Существительное: падеж, число, род
- Местоимение: падеж

## 2.3. Генерация лексико-синтаксических паттернов

Процесс генерации И-ЛСП можно разделить на несколько этапов:

- предобработка вопросов оценки компетентности;
- морфологический анализ вопросов оценки компетентности;
- сопоставление с элементами онтологии;
- построение моделей вопросительных предложений;
- генерация разных видов И-ЛСП на основе моделей.

На первом этапе происходит приведение вопроса оценки компетентности к нужному формату для проведения дальнейших этапов. В частности, удаление знаков препинания.

«В какой деятельности участвует Организация?» =>

«В какой деятельности участвует Организация»

На этапе морфологического анализа выделяются релевантные граммемы для каждого релевантного слова. Далее происходит сопоставление этих слов с классами, атрибутами и отношениями в онтологии*:*

Деятельность Class: Персона, Grammemes: П.п.

Организация Class: Организация, Grammemes: И.п.

участвует Rel: Организация.участвует в деятельности, Grammemes: [невозврат., несов. вид, неперех., ед. ч., н.в.]

В конце обработки ВОК каждое слово будет сопровождаться информацией о нем:

- часть речи;
- грамматические значения;
- название онтологического класса или атрибута;
- тип атрибута (инициализирующий, Data Property, Object Property).

После обработки вопросов оценки компетентности строятся модели вопросительных предложений. Для каждого вопроса в модель добавляется:

- сопоставленные с релевантными словами объекты онтологии (классы и атрибуты);
- предикат в его изначальной форме;
- указывается тип атрибута (если он есть);
- список грамматических ограничений.

Пример модели вопросительного предложения:

$$M = \langle (\text{Деятельность}, [\text{П.п.}])\ (\text{участвует}, \text{Организация.участвует в деятельности}, [\text{невозврат., несов. вид, неперех., ед. ч., н.в.}])\ (\text{Организация}, [\text{И.п.}]) \rangle \quad (3.5)$$

На последнем этапе на основе созданных моделей осуществляется генерация И-ЛСП.

**Scheme**

arg1: Object::Деятельность (Падеж: ‘пр’)

arg2: Term::Организация.участвует в деятельности (Число: ‘ед’)

arg3: Object::Организация (Падеж: ‘им’)

Condition Contact (arg1, arg2) = Contact_Object,

Contact (arg2, arg3) = Contact_Object

⇨ arg3::Организация (участвует в деятельности: arg1.Название)

На схеме показан простой вариант связывания объекта класса *Метод* с объектом класса *Персона*. Для этого рассматриваются уже три аргумента: объект класса *Персона*, термин лексико-синтаксического класса *Метод.имеет автора* и объект класса *Метод исследования*. Для указания на возможность аргументов быть разделенными используется контактность (Contact) в разделе условий (Condition). Такой паттерн будет обрабатывать случаи вида: *«Метод “мозгового штурма” был разработан в 1953 г. американским консультантом Осборном»*.

Предложенные этапы позволяют формировать информационные лексико-синтаксические модели на основе вопросов оценки компетентности с учетом грамматических значений каждого из них.

## 3. Оценка качества генерации

Во время генерации из 209 вопросов оценки компетентности было сформировано 200 неповторяющихся моделей И-ЛСП. Уменьшение количества моделей по сравнению с количеством вопросов можно объяснить тем, что, несмотря на разные вопросы для отношений, ограничения на объекты могли совпадать.

Была выбрана онтология «Поддержка принятия решений в слабоформализованных предметных областях» (https://uniserv.iis.nsk.su/rdms/) [21], на основе которой был создан словарь предметной области (129 объектно-ориентированных классов и 689 терминов) и корпус, состоящий из 31 научного текста той же предметной области, для проведения эксперимента по извлечению информации с использованием сгенерированных шаблонов. Эксперименты проводились в системе FATON.

Были использованы стандартные метрики точности, полноты и $F_1$-меры для анализа результатов извлечения информации.

В Таблице 1 представлены результаты генерации паттернов с грамматическими ограничениями (они выделены серым) и без грамматических ограничений.

*Таблица 1. Результаты генерации паттернов*

| **Степень неоднозначности** | | **Точность** | | **Полнота** | | **$F_1$** | |
|---|---|---|---|---|---|---|---|
| - | 1,5 | 0,97 | 0,70 | 1,0 | 0,96 | 0,98 | 0,81 |

Из таблицы видно, что генерация паттернов с грамматическими ограничениями дала результаты ниже, чем без грамматических ограничений.

Ожидалось получение полноты, равной 1.0, поскольку вопросы оценки компетентности охватывают все отношения в онтологии. Было подсчитано, что 45 вопросов оценки компетентности не сформировали паттерны. 9 из них оказались избыточными, что привело к уменьшению количества неповторяющихся моделей по сравнению с вопросами. В оставшихся же не были обнаружены отношения, чему способствовал ряд причин.

Во-первых, невозможность сравнения некоторых онтологических отношений со словами в ВОК. К таким отношениям можно отнести *являетсяЧастьюРесурса*, *являтьсяЧастью* и а*втор-Персона*. Во-вторых, были замечены опечатки в названиях онтологических

отношений и в связях в базовой онтологии. Так, например, отношение *Организация.участвует в деятельности* связывает не два онтологических класса *Организация* и *Деятельность*, а онтологический класс *Организация* и онтологическое отношение *Организация.участвует в деятельности.*

## 4. Эксперименты на корпусе научных текстов

Для проведения экспериментов по извлечению информации с помощью сгенерированных паттернов были выбраны онтология «Поддержка принятия решений в слабо формализованных областях» и корпус, состоящий из 31 научного текста той же предметной области. Эксперименты проводились в системе FATON.

Для формализации полученных данных для каждого текста были автоматически созданы таблицы, в которых описаны созданные онтологические объекты, добавленные атрибуты и связи. Они выводятся в каждой новой строчке в той последовательности, в которой встречаются объекты.

Для созданных онтологических объектов строка содержит название онтологического класса, атрибут *Название* и конкретную лексему. Для добавленных атрибутов объектов строка включает название онтологического класса, название типа атрибута и конкретную лексему его наименования. Для созданных связей между объектами указывается класс и идентификатор первого объекта, название отношения, класс и идентификатор второго объекта.

Эксперименты были проведены для паттернов без грамматических ограничений и паттернов, в которые в процессе генерации были добавлены грамматические ограничения. Результаты приведены в таблице ниже (Таблица 2), где серым цветом выделены значения для паттернов без ограничений.

*Таблица 2. Результаты извлечения сущности онтологии*

| / | **Точность** | | **Полнота** | | **$F_1$** | |
|---|---|---|---|---|---|---|
| Attributes | 0,34 | 0,83 | 0,97 | 0,72 | 0,50 | 0,77 |
| Relations | 0,09 | 0,38 | 1,0 | 1,0 | 0,17 | 0,55 |

Все ожидаемые объекты извлекаются из текстов, поэтому они не были внесены в таблицу. Полнота добавления атрибутов и отношений также оказалась высокой, что говорит об извлечении большей части (или всех) ожидаемых атрибутов и отношений, однако низкая

точность указывает на извлечение большого количества неправильных атрибутов и отношений.

Полученные результаты показывают, что паттерны с ограничениями значительно увеличивают точность добавления атрибутов, но при этом уменьшают полноту по сравнению с паттернами без ограничений.

Увеличение точности извлечения отношений было достигнуто сначала добавлением грамматических ограничений, а затем регулированием добавления связи с самим собой. Количество полученных объектов в онтологии будет совпадать с ожидаемым.

Анализ ошибок помог выявить причины ошибочных результатов:

- слишком строгие ограничения на падеж аргументов для паттернов, добавляющих атрибуты объектам онтологии;
- недостаточно строгие ограничения на позицию аргументов (расположение относительно друг друга) для паттернов, создающих связи между объектами онтологии;
- списки рассматриваются как одно предложение, из-за чего формируются ненужные связи;
- объектам добавляется связь с самим собой.

Увеличение полноты можно достичь с помощью добавления паттернов, в которых не будет грамматических ограничений, но будет ограничение на контактность.

Для улучшения точности можно добавить предварительный анализ, который позволил бы разбить список на фрагменты текста, чтобы избежать ненужной связи между частями списка. Также необходимо убрать возможность аргументов создавать связи с самим собой.

## 5. Заключение

Данная работа является продолжением цикла работ, посвященных методам автоматической генерации лексико-синтаксических паттернов онтологического проектирования, предназначенных для извлечения информации из текстов и пополнения онтологии. В работе предложен подход к генерации лексико-синтаксических паттернов на основе онтологических паттернов НПО. Особенностью подхода является использование вопросов оценки компетентности для извлечения грамматических свойств языковых выражений, используемых для представления понятий в текстах.

Во время проведения экспериментов были получены следующие результаты. При генерации степень неоднозначности паттернов составила 1,5. При извлечении информации F1-мера составила 1,0 для объектов, 0,77 для атрибутов и 0,55 для отношений. В целом, полученные результаты показывают, что паттерны с грамматическими ограничениями значительно увеличивают точность извлечения атрибутов и отношений, но при этом уменьшают полноту по сравнению с паттернами без ограничений. Проведенный анализ ошибок позволяет сделать предположение, что добавление новых паттернов, которые будут содержать только ограничение на взаиморасположение терминов в текстах, позволит увеличить полноту. Что касается точности, то предлагается проводить предварительный анализ, который позволит, в частности, разбивать список на фрагменты текста во избежание неправильных связей между частями списка.

Дальнейшие исследования будут связаны с а) апробацией подхода на других предметных областях, б) расширением обучающего корпуса ВОК утвердительными предложениями (например, определениями из энциклопедических источников знаний), что может дать более полную картину возможных ограничений, в) рассмотрением других типов грамматических ограничений (например, согласование в числе и роде), г) применением для генерации терминологических паттернов.

## Список литературы

1. Гаранина, Н. О. Мультиагентный подход к извлечению информации из текстов и пополнению онтологии / Н. О. Гаранина, Е. А. Сидорова // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–2015), 6 — 8 октября 2015 г., Новосибирск. – Новосибирск: Институт математики им. С. Л. Соболева СО РАН, 2015. –Т.1. — С. 50-59.
2. Кононенко И.С., Сидорова Е.А. Методика разработки лексико-семантических паттернов для извлечения терминологии научной предметной области // Системная информатика. 2022. № 20. С. 25-46.
3. Сидорова, Е. А. Комплексный подход к исследованию лексических характеристик текста / Е. А. Сидорова // Вестник СибГУТИ, №3, 2019. – С. 80-88.
4. Загорулько Ю.А., Сидорова Е.А., Загорулько Г.Б., Ахмадеева И.Р., Серый А.С. Автоматизация разработки онтологий научных предметных областей на основе паттернов онтологического проектирования // Онтология проектирования. – 2021. – Т.11, №4(42). - С.500-520. – DOI: 10.18287/2223-9537-2021-11-4-500-520.

5. Aguado de Cea, A. Using Linguistic Patterns to Enhance Ontology Development / G. Aguado de Cea, A. Gomez-Perez, E. Montiel-Ponsoda, M. C. Suarez-Figueroa // In: Proc. Int. Conf. on Knowledge Engineering and Ontology Development (KEOD 2009) (Funchal - Madeira, Portugal, October 6-8, 2009), 2009. – P. 206–213.
6. Blomqvist, E. Engineering Ontologies with Patterns: The eXtreme Design Methodology / E. Blomqvist, K. Hammar, V. Presutti // Ontology Engineering with Ontology Design Patterns. Studies on the Semantic Web, 2016. – P. 23 – 50.
7. Gangemi, A. Ontology Design Patterns / A. Gangemi, V. Presutti // Handbook on Ontologies. Springer, 2009. – P. 221–243.
8. Ganino G., Lembo D., Mecella M., Scafoglieri F. Ontology population for open-source intelligence: a GATE-based solution // Software: Practice and Experience. 2018. V. 48. Is. 12.
9. Gnminger, M. Methodology for the design and evaluation of ontologies / M. Gnminger, M. Fox // Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, Canada, 1995. – 10 p.
10. Guarino, N. OntoSeek: content-based access to the Web / N. Guarino, C. Masolo, G. Vetere // IEEE Intelligent Systems, 1999. – P. 70-80.
11. Hearst, M. Automatic acquisition of hyponyms from large text corpora / M. Hearst // Conference on Computational Linguistics (COLING'92), Nantes, France, Association for Computational Linguistics. 1992. – P. 539-545.
12. Ijntema, W. A lexico-semantic pattern language for learning ontology instances from text / W. Ijntema, J. Sangers, F. Hogenboom, F. Frasincar // Journal of Web Semantics, 2012. – P. 37–50.
13. Lamy, J.-B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies / J.-B. Lamy // Artificial Intelligence In Medicine, 2017. – P. 11-28.
14. Maynard, D. Using Lexico-Syntactic Ontology Design Patterns for ontology creation and population / D. Maynard, A. Funk, W. Peters // Proceedings of the 2009 International Conference on Ontology Patterns, vol. 516, 2009. – P. 39-52.
15. Ontology Design Patterns // Ontology Design Patterns URL: http://ontologydesignpatterns.org (дата обращения: 20.10.2022). (44)
16. Petasis G., Karkaletsis V., Paliouras G., Krithara A., Zavitsanos E. Ontology Population and Enrichment: State of the Art. In: Paliouras G., Spyropoulos C.D., Tsatsaronis G. (eds). Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. LNCS, V. 6050. Springer, Berlin, Heidelberg.
17. Rosenberg, G. Handbook of Formal Language / G. Rosenberg, F. Salomaa, 1996. – 450 p.
18. Roux, C. An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions / C. Roux, D. Proux, F. Rechenmann, L. Julliard // Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000), Berlin, Germany, 2000.

19. Zagorulko Y. A. Using a System of Heterogeneous Ontology Design Patterns to Develop Ontologies of Scientific Subject Domains / Y. A. Zagorulko, O. I. Borovikova // Programming and Computer Software, 2020. – P. 273-280.
20. Zagorulko Y. A., Sidorova E. A., Akhmadeeva I. R. and Sery A. S . Approach to automatic population of ontologies of scientific subject domain using lexico-syntactic patterns // International Conference «Marchuk Scientific Readings 2021» (MSR-2021) 4-8 October 2021, Novosibirsk, Russian Federation. Journal of Physics: Conference Series, 2021, vol.2099, p. 012028. doi:10.1088/1742-6596/2099/1/012028)
21. Zagorulko Y., Zagorulko G. Application of Ontology Design Patterns for Building an Ontology of Decision Support in Weakly Formalized Domains // Proceedings of Selected Contributions to the Russian Advances in Artificial Intelligence Track at RCAI 2021, collocated with the 19th Russian Conference on Artificial Intelligence (RCAI 2021). – Vol. 3044. – P. 108–116. – CEUR Workshop Proceedings, CEUR-WS.org, 2021.