МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет Механико-математический
Кафедра Программирования

Направление подготовки Математика и компьютерные науки

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Череватенко Юлия Александровна
(Фамилия, Имя, Отчество автора)

Тема работы: Методы построения семантико-синтаксических моделей для извлечения информации из текста

«К защите допущена»

Заведующий кафедрой,
д.ф.-м.н., профессор

Марчук А.Г /_____
(фамилия , И., О.) (подпись, МП)

«…»……20…г.

Научный руководитель

к.ф.-м.н., доцент ММФ

Сидорова Е.А. /_____
(фамилия , И., О.) (подпись, МП)

«…»……20…г.

Дата защиты: «…» ……20…г.

Новосибирск, 2023

Содержание

Введение

В настоящее время существующие методы обработки текстов не всегда позволяют достичь высокой точности и полноты извлечения информации, особенно при работе с текстами определенных предметных областей. Уровень семантической обработки естественного языка все еще недостаточен по ряду причин, в том числе из-за неоднозначности различных языковых единиц – от слов до наборов предложений [1]. Появление в последнем десятилетии различных инструментов семантической обработки языка подтверждает востребованность алгоритмов для такого типа обработки.

Анализ текста позволяет извлечь полезную информацию, которая может быть использована для создания различных баз данных, систем знаний и классификаторов. Однако, русскоязычные тексты имеют сложную структуру и разнообразие, что затрудняет автоматический анализ. Существующие методы компьютерного анализа текстов не всегда показывают высокие результаты на уровне семантического анализа предложений сложной структуры. К тому же, проблема многозначности в языке является одной из основных проблем в области обработки естественного языка.

Разработка новых методов и алгоритмов, основанных на семантико-синтаксических моделях, может значительно улучшить качество обработки текстов и повысить эффективность работы систем, использующих

естественный язык. Для автоматического извлечения информации используются методы, основанные на статистических и лингвистических критериях [2]. Статистические критерии основаны на подсчете частоты употребления слов и словосочетаний в тексте или коллекции текстов, а также на вычислении статистических параметров на основе этих частот. Лингвистические критерии учитывают типичную синтаксическую структуру информации и свойства конкретного программного обеспечения, в рамках которого используются слова и словосочетания.

В связи с этим актуальность данной темы обусловлена тем, что использование семантико-синтаксических моделей может значительно повысить качество извлечения информации из текстов.

Целью работы является разработка методов построения семантико-синтаксических моделей для извлечения информации из текстов.

Для достижения этой цели были поставлены задачи:

- проанализировать существующие методы извлечения информации и подходы к семантико-синтаксическому анализу текстов;
- разработать представление унарных моделей и билексем и формат представления семантико-синтаксических моделей в текстовом виде;
- разработать алгоритм поиска информации на основе моделей;

- реализовать систему поиска и проверки моделей;
- провести экспериментальные исследования с помощью собранного корпуса текстов ограниченной тематики.

Объектом исследования является семантико-синтаксическая языковая модель, предметом исследования – методы извлечения информации с помощью семантико-синтаксических моделей.

В работе были использованы следующие методы исследования: семантическая классификация лексики для определения структуры информации предметной области, методы семантико-синтаксического анализа для определения синтаксической и семантической структуры высказываний, корпусный анализ для оценки покрытия текста моделями, методы компьютерного моделирования для создание программного обеспечения, методика экспериментального исследования для оценки качества извлечения информации на основе семантико-синтаксических моделей.

Материалом для исследования послужил корпус биографических текстов, включающий 200 текстов объемом в 2959 предложений [3].

Новизна данного исследования заключается в развитии формальных средств представления семантико-синтаксических моделей, предназначенных для извлечения информации определенной предметной области.

Теоретическая значимость работы определяется вкладом в развитие семантико-синтаксического анализа и возможностью использовать результаты в лингвистических исследованиях.

Практическая значимость данной работы заключается в том, что созданная программа может быть использована в системах локального семантико-синтаксического анализа, а также в других приложениях, включающих в себя обработку естественного языка.

Глава 1. Обзор

Обзор посвящен методам семантико-синтаксического анализа, а также приведены примеры семантико-синтаксических анализаторов для русскоязычных текстов.

Анализ текста предполагает извлечение семантически связанной, полезной для пользователя информации. Однако, существующие методы компьютерного анализа текста не полностью обеспечивают потребности конечных пользователей, связанные с обработкой русскоязычных текстов.

Существует четыре основных уровня анализа текста на естественном языке, которые выполняются последовательно: морфологический, синтаксический, семантический и прагматический. Морфологический анализ направлен на определение морфологических свойств слова и его основных словоформ. Синтаксический анализ направлен на определение связей между словами и их частями в предложении, в то время как семантический анализ направлен на выявление значения слов и именованных групп. Прагматический анализ текста имеет целью раскрыть взаимодействие между автором и читателем, а также определить, какая информация в тексте будет полезной для читателя, основываясь на его типологии.

В рассматриваемых статьях [4, 5] представлен анализ вопроса автоматизации процесса извлечения ключевых слов и фраз из текста.

Существует множество методов, позволяющих решать данную проблему, среди которых выделяются лингвистические, статистические, спектральные и гибридные подходы. Лингвистические методы основываются на семантических данных о слове и используют онтологии. Статистические методы, в свою очередь, используют численные данные о встречаемости слова в тексте. Результаты ряда исследований свидетельствуют о том, что каждый из перечисленных методов обладает своими недостатками. Лингвистический подход к извлечению ключевых слов вводит значительные трудности на ранних этапах, связанные с необходимостью разработки онтологий, а область эффективного применения статистических моделей ограничена языками с бедной морфологией; как правило, имеются проблемы для естественных языков с богатой морфологией, в частности, для русского языка. Поэтому большинство авторов статей приходят к выводу, что наиболее точным является использование гибридных подходов, которые объединяют эти методы в одном алгоритме.

При извлечении информации из текста иногда необходимо учитывать семантические связи между словами. Решение данной задачи может быть осуществлено посредством нескольких подходов: использование паттернов (шаблонов), статистические методы, основанные на подсчете встречаемости, применении нейросетей и т. п., и различные эвристические методы. Надлежащее составление паттернов может обеспечить высокую точность

извлечения, но необходимо проделать значительную работу по составлению паттернов для обеспечения полноты поиска, не нарушая при этом согласованности между ними. В противоположность, статистические методы требуют немного вводных размеченных данных и показывают хорошие результаты при увеличении обучающей выборки. Однако их недостатком является низкая начальная точность на выборках небольшого и среднего размера, что требует применения дополнительных методов для ее улучшения.

Достаточно давно изучается возможность использования семантико-синтаксических моделей для обработки текста. Например, в работе Бледнова А. М. [6] проводится исследование и разработка технологий семантико-синтаксического анализа. Автором были созданы статистические и векторные модели текста, а также информационная технология семантико-синтаксического анализа, которая может использоваться для автоматизации многих процессов, связанных с созданием и использованием информационных ресурсов. В данной работе автор проводит экспериментальное исследование с использованием программного комплекса «ТЕКСТАН», в результате которого точность семантического анализа показала 90%, а вероятность правильного выделения ключевых слов составляла 95%.

Одной из самых известных семантико-синтаксических моделей, использующихся для обработки естественного языка, является модель

«Смысл - Текст», разработанная И.А. Мельчуком [7]. Основная задача модели заключается в преобразовании заданного смысла в текстовое представление. Другими словами, модель "смысл - текст" с помощью грамматических правил и лексических данных преобразует смысловое представление на естественном языке в соответствующие ему слова и фразы.

Для работы модели "смысл - текст" используется синтаксический анализатор текста, который разбивает входной текст на составные части и определяет связи между словами и их грамматические категории. Синтаксический анализ в данной модели разделяется на поверхностно-синтаксический и глубинно-синтаксический уровни.

Однако, несмотря на достижения и преимущества модели "смысл - текст", она все еще имеет некоторые ограничения, связанные с сложностью семантического анализа естественного языка. Решение этой проблемы требует дальнейшего развития технологий обработки естественного языка и создания новых методов и алгоритмов.

На основе модели "смысл - текст" построена система ЭТАП (Электронный Текстовый Анализатор Поиска) [8], которая используется для автоматизированного анализа текстов на русском языке. ЭТАП работает на основе методов машинного обучения и использует синтаксический анализатор для обработки текста. Он использует грамматику русского языка

для распознавания синтаксических отношений между словами в предложении, после чего анализирует семантические отношения между словами, чтобы определить, какая информация находится в тексте. Одним из преимуществ системы ЭТАП является высокая точность в извлечении информации из текста. Это объясняется тем, что система использует широкий словарь синтаксических, семантических и лексических правил русского языка.

Но у данного анализатора также есть и недостатки, такие как ориентированность на русский язык, некорректная интерпретация некоторых слов и фраз, ошибки при обработке текстов со сложными грамматическими конструкциями или двусмысленными выражениями. Многие минусы ЭТАПа связаны со слабостями модели "смысл - текст" [9], например, проблематичность атомизации семантики и отсутствие ясной технологии составления толково-комбинаторных словарей.

В статье [10] представлена разработка синтаксического анализатора на основе системы ЭТАП. В отличие от ЭТАПа, данный анализатор применяет восходящий метод для построения деревьев зависимостей, больше ориентируется на грамматику и использует модифицированный способ создания синтаксических связей между словами. По результатам экспериментального исследования, их анализатор показывает лучшие результаты по сравнению с ЭТАПом (процент правильных структур без меток

зависимостей у ЭТАПа составлял 47,40%, у анализатора 54.16%, с метками зависимостей 36,88% у ЭТАПа, 41,92% у анализатора). Однако исследуемый анализатор иногда совершал ошибки, которые не возникали при работе с ЭТАПом.

Также, существует система АОТ (Автоматическая обработка текста) [11] - русскоязычная система синтаксического анализа и обработки текста. Она позволяет извлечь семантику и логическую структуру из текста, распознать ключевые слова, даты, адреса и другие сущности. Система АОТ состоит из нескольких компонентов, таких как орфографический корректор, морфологический анализатор, синтаксический анализатор и словарь русского языка. Система АОТ демонстрирует высокую точность в синтаксическом анализе и извлечении информации из текстов на русском языке путем использования словарей и правил анализа синтаксических и семантических отношений русского языка.

Слабые стороны данной системы заключаются в использовании грамматики непосредственно составляющих, которая подходит для языков с преимущественно строгим порядком слов, но недостаточно хорошо себя показывает в языках с произвольным порядком слов в предложении. Помимо этого, анализ строится на общей лексике, в связи с чем могут возникнуть проблемы с анализом текстов, ограниченных определенной предметной областью.

Также рассмотрим две прикладные системы обработки естественного языка GATE и RCO.

Система GATE (General Architecture for Text Engineering) [12] используется для обработки текста. Разработана в Университете Шеффилда, Великобритания. Она представляет платформу для создания и использования приложений на основе обработки естественного языка и анализа данных. GATE поддерживает обработку текста на различных языках и имеет множество инструментов для выполнения разных задач. Среди функций GATE - морфологический и синтаксический анализ, выделение сущностей, обнаружение связей и многое другое.

Система GATE имеет несколько преимуществ, таких как платформенную независимость, поддержку разных форматов данных (XML, RDF, OWL и другие) и графический интерфейс, который ускоряет создание и тестирование решений. Важным отличием является то, что у GATE существует много инструментальных расширений, представляющих отдельные программы, которые используются для различных задач.

Однако, у системы GATE есть и свои недостатки, например, ориентированность на англоязычные тексты. Помимо этого, ее сложность может привести к трудностям при работе с ней для новичков.

Russian Context Optimizer Fact Extractor (RCO FE) [13] - это инструмент компьютерного анализа русскоязычной текстовой информации. Он использует контекстные шаблоны для извлечения фактов и сущностей из текстов. RCO FE основана на RCO Pattern Extractor [14]. Система предоставляет лингвистический анализ текста, включая грамматический и смысловой анализ языка, и имеет интерфейс для чтения результатов анализа и использования их другими программами.

Подход к семантической интерпретации [15], используемый в RCO FE, обеспечивает в среднем около 95% точности и 60% полноты при извлечении из текста описаний событий и фактов в соответствии с заданными семантическими шаблонами. В основном данный компонент используется для анализа текстов экономической и общественно-политической тематики.

Самым главным недостатком RCO является ее коммерческое распространение. Помимо этого, для создания шаблонов и обучения системы необходимо иметь соответствующие знания, система не подходит для новичков.

Проанализировав все перечисленные выше системы и технологии семантико-синтаксического анализа, можно сделать вывод о востребованности альтернативного семантико-синтаксического анализатора, который будет отвечать следующим требованиям:

1. Свободный доступ;
2. Возможность создания и редактирования моделей (шаблонов);
3. Возможность использования собственного словаря, размеченного для конкретной предметной области;
4. Локальный поиск информации по моделям.

Глава 2. Семантико-синтаксическая модель

В этой главе дается определение семантико-синтаксическим моделям, рассматривается их формальное описание и формат представления в текстовом виде.

2.1. Формальное представление

Семантико-синтаксическая модель (ССМ) или модель управления - это структура, определяющая взаимосвязи между главной (предикатной) лексемой и другими лексемами. Модель управления задается перечнем валентностей лексемы, синтаксической компонентой, в которой определяется необходимый набор морфологических характеристик для каждой валентности, и семантической компонентой, где определяются семантические характеристики, которыми должны обладать словоформы, соответствующие определенной валентности, и указываются семантические отношения, соответствующие синтаксическим отношениям, определенным в модели управления.

В модели управления валентность определяется актантом - элементом, указывающим на заполняемую в предложении некоторую валентность определенной лексемы [16]. Актанты содержат морфологические и семантические признаки, которые определяют соответствующую смысловую группу и форму слова, которую следует использовать.

Формально ССМ языка характеризуется набором:

$< S, M, C, L >$,

где

$S = \{s_i\}$ – множество семантических атрибутов,

$M = \{M_i\}$ – множество морфологических атрибутов и каждый морфологический атрибут M_i является множеством морфологических значений $\{m_{ij}\}$,

$C = < P, \{A_i\} >$ – множество моделей, где P – множество предикатов, а $\{A_i\}$ - множество актантов.

$L = < l_i, g_i, b_i >$ – множество лексем, где l_i – точные лексемы, g_i – обобщенные лексемы, а b_i – билексемы.

Семантико-синтаксическая модель характеризуется парой $C = < P, \{A_i\} >$, где P – это предикат, в качестве которого может фигурировать множество лексем из множества L. С предикатом связаны актанты. Актант A_i – это множество пар $< S_v, \{M_v\} >$, где $S_v = s_i \& \ldots \& s_j$ – конъюнкция семантических атрибутов из множества S общего списка $s_i, \ldots, s_j \in S$, $\{M_v\}$ – множество конъюнкций морфологических атрибутов

$M_v = m_{ij}\&\ldots\&m_{kl}$ из множества M общего списка, т.е. для каждой конъюнкции семантических атрибутов может быть представлено несколько конъюнкций морфологических атрибутов. Каждый морфологический атрибут M_i является множеством морфологических значений $\{m_{ij}\}$.

Унарные модели также принадлежат множеству C и состоят из одного актанта. Множество предикатов P пустое. Унарные модели характеризуются набором пар $< S_v, \{M_v\} >$, где $S_v = s_i\&\ldots\&s_j$ – конъюнкция семантических атрибутов, $\{M_v\}$ – множество конъюнкций морфологических атрибутов $M_v = m_{ij}\&\ldots\&m_{kl}$.

Каждая точная лексема l_i – это набор $< N, S_l, M_l, L_l >$, где N – нормальная форма слова, S_l – конъюнкция семантических ограничений $S_l = s_i\&\ldots\&s_j$, $s_i,\ldots, s_j \in S$, M_l – конъюнкция морфологических ограничений $M_l = m_{ij}\&\ldots\&m_{kl}$, $m_{ij},\ldots, m_{kl} \in M$, Lex_l – конъюнкция лексических признаков, $Lex_l = m_{ij}\&\ldots\&m_{kl}$, $m_{ij},\ldots, m_{kl} \in M$.

Каждая обобщенная лексема g_i – это набор $< S_l, M_l, L_l >$, где S_l – конъюнкция семантических ограничений $S_l = s_i\&\ldots\&s_j$, $s_i,\ldots, s_j \in S$, M_l – конъюнкция морфологических ограничений $M_l = m_{ij}\&\ldots\&m_{kl}$,

$m_{ij},\ldots,m_{kl} \in M$, Lex_l – конъюнкция лексических признаков, $Lex_l = m_{ij}\&\ldots\&m_{kl}$, $m_{ij},\ldots,m_{kl} \in M$.

Каждая билексема b_i – это набор пар $< l_l, c_l >$, где l_l – лексема из множества лексем L, c_l – ССМ из множества C, в множестве предикатов которой есть l_l.

2.2. Формат представления семантико-синтаксических моделей

Был разработан текстовый формат представления ССМ, включающий описание семантических и морфологических атрибутов, используемых моделями, а также лексем, моделей и билексем.

Сначала описываются семантическая и морфологическая таблицы.

```
[SemanticTable]
SemanticAttr_Name1
SemanticAttr_Name2
…
[MorphTable]
MorphAttr_Name1: <val1, val2 …>
MorphAttr_Name2: <val1, val2 …>
…
-----
```

После таблиц описываются модели:

```
CF CaseFrame_Name {
Aktant Aktant_Name1 :
     SemanticAttr_Name1 <val1, val2> <val3, val4>
Aktant Aktant_Name2 :
     SemanticAttr_Name1 & SemanticAttr_Name2 <val1, val2>
     SemanticAttr_Name3 <val1, val2>
}
UCF UnCaseFrame_Name : SemanticAttr_Name1 <val1, val2>
-----
```

Например, модели могут записываться следующим образом:

```
CF Родиться (birth) {
     Aktant Кто :
               Персона <Сущ, им>
     Aktant Где :
               Город <Сущ, пр, в>
               Страна <Сущ, пр, в>
}
UCF Национальность (nationality) : Национальность <Сущ> <Прил>
```

После моделей идет список лексем.

```
LEX Lex_name {
     SemLim SemanticAttr_Name1 & SemanticAttr_Name2
     MorLim <val1, val2>
     LexF <val1, val2>
     CF :
     CaseFrame_Name1
}
BLEX Blex_name {
     LEX Lex_name1 :
        SemLim SemanticAttr_Name1 & SemanticAttr_Name2
        MorLim <val1, val2>
        LexF <val1, val2>
     CF : CaseFrame_Name1
     LEX Lex_name2 :
        SemLim SemanticAttr_Name1 & SemanticAttr_Name2
        MorLim <val1, val2>
        LexF <val1, val2>
     CF : CaseFrame_Name2
}
```

Пример обобщенной лексемы:

```
LEX {
     SemLim Должность
     MorLim
     LexF <Сущ>
     CF :
          Должность (occupation)
}
```

Пример точной лексемы:

```
LEX родиться {
     SemLim
     MorLim
     LexF <Глаг>
```

```
     CF :
          Родиться (birth)
}
```

Пример билексемы:

```
BLEX направить служить
{
     LEX направить :
          SemLim
          MorLim
          LexF <Глаг>
     CF :
          Направить (professional_events)
     LEX служить :
          SemLim
          MorLim
          LexF <Глаг>
     CF :
          Служить (professional_events)
}
```

В примере билексемы описаны 2 лексемы: направить и служить. В тексте лексема “направить” будет являться предикатом модели “Направить (professional_events)”, а лексема “служить” – предикатом “Служить (professional_events)”.

Глава 3. Алгоритм анализа текста

Для поиска информации в тексте на основе семантико-синтаксических моделей был разработан алгоритм. Поиск информации заключается в анализе каждого предложения отдельно и разбивается на несколько составных частей: поиск моделей согласования, поиск билексем и поиск унарных моделей.

Анализ предложения проводится следующим образом:

1. Создается массив всех слов предложения.
2. Для каждого слова из массива:
    a. Ищется совпадающая по нормальной форме лексема из списка лексем. При совпадении:
        i. Запускается поиск моделей с данной предикатной лексемой.
        ii. Запускается поиск билексем, в состав которых входит данная лексема.
    b. Запускается поиск унарных моделей, признакам которых удовлетворяет данное слово.

Опишем алгоритм поиска моделей по лексеме. На вход дается сама лексема и массив слов в предложении.

1. Создается пустой массив результатов R.
2. Для каждой модели, связанной с данной лексемой:

a. Создается пустой массив пар слов и актантов W для результата.
b. Для каждого слова в предложении:
    i. Если слово совпадает с лексемой по нормальной форме, то это - предикатная лексема. Слово добавляется в массив пар слов и актантов без актанта. Больше в массиве предикатных лексем быть не может.
    ii. Запускается проверка принадлежности слова к одному из актантов данной модели. Если принадлежит, то в массив W добавляется это слово и актант.
c. Массив W и данная модель добавляется в массив R.

3. Возвращается R.

Проверка принадлежности слова к некоторой данной модели происходит следующим образом:

1. Для каждого актанта данной модели, для каждой семантической группы признаков данного актанта:
    a. Проверяется наличие всех семантических признаков у слова из конъюнкции семантических признаков данной семантической группы. Если несовпадение, то переход к следующей семантической группе.
    b. Проверяется совпадение всех морфологических признаков у слова хотя бы одной конъюнкции морфологических признаков

данной семантической группы. Если несовпадение, то переход к следующей семантической группе.

c. Возвращается $true$.

2. Возвращается $false$.

На вход алгоритму поиска билексем даётся сама билексема, массив слов в предложении и индекс положения одной из лексем в массиве слов. Индекс нужен, чтобы избежать повторов билексем в предложении и не засчитывать одну и ту же билексему дважды.

1. Создается массив пар моделей и массивов слов CW и счетчик лексем .
2. Для каждого слова в предложении от индекса одной из лексем и далее:
    a. Ищется совпадающая по нормальной форме лексема из списка лексем. Если совпадений нет, переход к следующему слову.
    b. Проверяется принадлежность лексемы к данной билексеме. Если нет, то переход к следующему слову.
    c. Для каждой модели, связанной с данной лексемой в билексеме:
        i. Запускается поиск моделей с данной предикатной лексемой (алгоритм см. выше).
        ii. Результат записывается в массив CW.
    d. Прибавляется 1 к счетчику лексем.
    e. Если число в счетчике лексем равно числу лексем в билексеме, то выходим из цикла.

3. Если число в счетчике лексем равно числу лексем в билексеме, то возвращаем данную билексему и массив CW.

На вход алгоритму поиска унарных моделей подается слово и унарная модель.

1. Для каждой семантической группы признаков унарной модели:
    a. Проверяется наличие всех семантических признаков у слова из конъюнкции семантических признаков данной семантической группы. Если несовпадение, то переход к следующей семантической группе.
    b. Проверяется совпадение всех морфологических признаков у слова хотя бы одной конъюнкции морфологических признаков данной семантической группы. Если несовпадение, то переход к следующей семантической группе.
    c. Возвращается унарная модель и слово.

Быстродействие алгоритма оценивается как $O(n^2)$ для поиска моделей согласования, $O(n^2 l)$ для поиска билексем и $O(n)$ для поиска унарных моделей, где n - количество слов в предложении, l - количество лексем в билексеме.

Глава 4. Программные компоненты для семантико-синтаксического анализа.

В данной главе рассматривается программный компонент, созданный для анализа текста и используемые им библиотеки.

4.1. SemanticVoc

Для работы с ССМ использовалась библиотека SemanticVoc. Данная библиотека предназначена для чтения, записи, редактирования и обработки моделей.

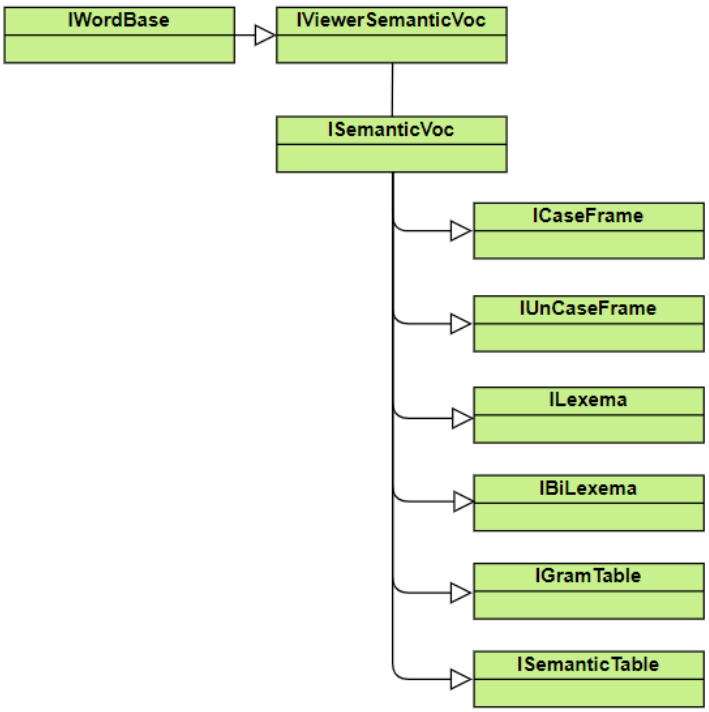

Рис. 1. Структура классов SemanticVoc

На рисунке 1 представлена структура классов SemanticVoc с основными функциями, необходимыми пользователю. Класс IViewerSemanticVoc содержит функции чтения, записи и обработки моделей. Наследуемый класс

ISemanticVoc помимо вышеперечисленных функций содержит функции редактирования моделей. Класс IWordBase определяет слова, классы ICaseFrame, IUnCaseFrame, ILexema, IBiLexema определяют обычные и унарные модели, лексемы и билексемы соответственно. Классы IGramTable, ISemanticTable определяют грамматическую и семантическую таблицы.

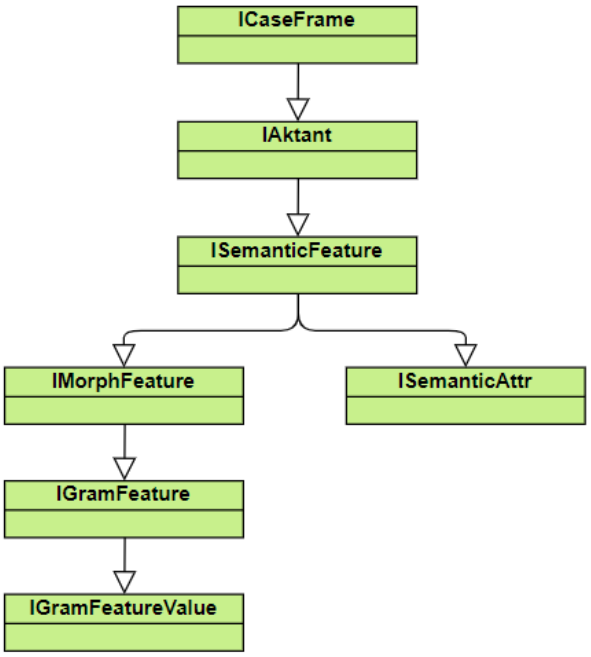

Рис. 2. Классы модели управления

На рисунке 2 представлена схема классов, составляющих семантико-синтаксическую модель. Объект класса ICaseFrame состоит из вектора актантов (IAktant), которые, в свою очередь, состоят из вектора групп семантических признаков (ISemanticFeature), каждая из которых включает в себя вектор семантических атрибутов (ISemanticAttr) и морфологических признаков (IMorphFeature). Морфологический признак состоит из вектора грамматических характеристик, каждая из которых содержит своё значение.

На рисунке 3 показаны схемы классов лексем. Объект класса билексемы состоит из вектора пар лексемы (ILexema) и вектора моделей (ICaseFrame). Сама лексема состоит из вектора семантических ограничений (ISemanticAttr), векторов морфологических ограничений и лексических признаков (IGramFeature) и вектора связанных с ней моделей (ICaseFrame).

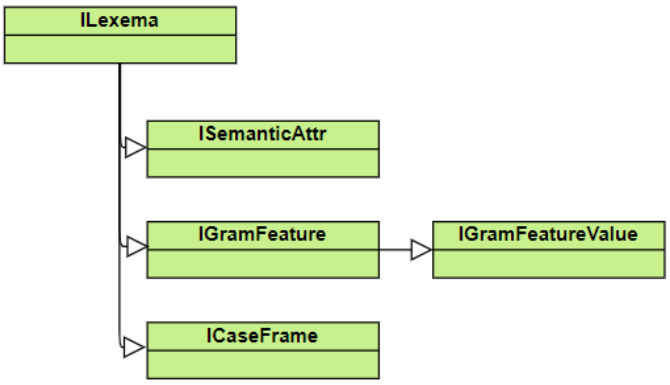

Рис. 3. Классы лексемы

Автором в библиотеку были добавлены классы билексем, унарных моделей и их реализация, а также парсер нового формата представления моделей.

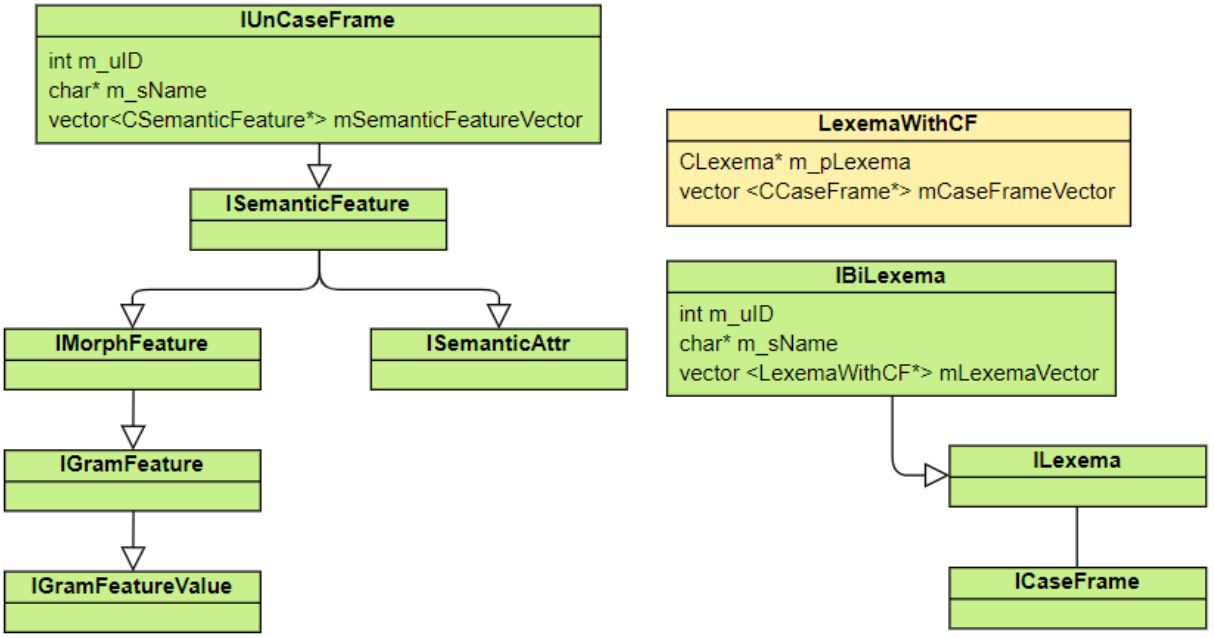

Рис. 4. Классы унарных моделей и билексем

Унарная модель состоит из вектора групп семантических признаков (ISemanticFeature). Объект класса билексемы состоит из вектора пар лексемы (ILexema) и вектора моделей (ICaseFrame).

4.2. Программа анализа текста

Программа анализа текста была создана с помощью библиотек KLAN и SemanticVoc. С помощью KLAN осуществляется работа со словарем [17] и проводится морфологический анализ текста, а с помощью SemanticVoc - работа с моделями.

Модуль анализа текста собран в отдельный класс SemAnalysis. Для представления результатов в данном классе используются следующие вспомогательные структуры:

1. AktantWord с полями IAktant - актант, и IWordBase - слово, для описания того, какое слово принадлежит какому актанту.
2. CaseFrameWords с полями ICaseFrame - модель, и вектором AktantWord для описания того, какие слова-актанты относятся к данной модели.
3. BilexemaWords с полями IBiLexema - билексема, и вектором CaseFrameWords для описания того, какие модели с какими словами принадлежат данной билексеме.
4. UnCaseFrameWord с полями IUnCaseFrame - унарная модель, и IWordBase для описания того, какое слово принадлежит данной унарной модели.

Сам класс содержит поля векторов CaseFrameWords, BilexemaWords и UnCaseFrameWord, в которых хранятся результаты анализа, а также поля с моделями (SemanticVoc) и морфологическим разбором текста (KLAN).

В данном классе реализованы алгоритмы, представленные в главе 3. Среди методов данного класса наиболее важная функция - TestSentence, принимающая на вход предложение и проводящая его анализ. Кроме того, класс содержит функции для возврата результатов анализа и несколько вспомогательных функций для TestSentence. Все результаты, полученные в ходе выполнения программы, сохраняются в полях CaseFrameWords, BilexemaWords и UnCaseFrameWord.

Входной текст разбивается на предложения. С помощью библиотеки KLAN каждое предложение распознается на отдельные слова и проводится морфологический анализ. Затем для каждого предложения вызывается функция TestSentence.

После выполнения алгоритма анализа текста результаты можно получить, обратившись к полям объекта класса SemAnalysis, где хранятся результаты анализа.

Глава 5. Экспериментальное исследование

Для оценки качества извлечения информации было проведено экспериментальное исследование тематической классификации текстов биографического корпуса. Следует отметить, что задача классификации отличается от задачи извлечения информации, поэтому результаты данного эксперимента не могут полностью отразить качество анализа текста и являются приближением к оценке качества извлечения информации.

5.1. Входные данные

Для экспериментального исследования использовался корпус биографических текстов [8]. Корпус состоял из 200 текстов и 2959 предложений в сумме. Каждое предложение относится к одному или двум тематическим классам. Всего тематических классов 13:

1. рождение (birth);
2. личные события (personal_events);
3. профессиональные события (professional_events);
4. место жительства, пребывания (residence);
5. смерть (death);
6. национальность (nationality);
7. информация о родительской семье (parenting);
8. членство (affiliation);

9. образование (education);
10. семья (family);
11. род занятий, должность (occupation);
12. прочие биографические факты (other);
13. не биографический факт (none).

По этому корпусу с помощью библиотеки KLAN был создан словарь, состоящий из 7721 термина и размеченный по 37 семантическим классам.

Для данного корпуса было создано 96 моделей на основе частотности слов в каждом тематическом классе. В таблице 1 приведено количество моделей для каждого тематического класса и количество лексем. В данном контексте тематическими моделями являются унарные модели, которые охватывают в тексте слова определенной семантики, а семантическими моделями являются модели, требующие определенной структуры в искомых фразах.

Таблица 1. Количество моделей и лексем.

	Модели	Тематические модели	Семантические модели	Лексемы
Родительская семья	6	3	3	7
Личные события	10	4	6	6
Место жительства	4	1	3	10
Место работы	5	1	4	11
Национальность	1	1	0	0
Образование	10	3	7	16

Профессиональные события	28	6	22	41
Род занятий	20	3	17	46
Рождение	3	1	2	3
Семья	2	1	1	5
Смерть	4	1	3	6
Прочее	3	3	0	0
Не биографический факт	0	0	0	0
Всего	**96**	**28**	**68**	**138**

5.2. Результаты

Для оценки результатов вычислялась полнота, точность и F-мера. В данном исследовании полнота является более важной метрикой, чем точность, потому что для извлечения информации более важно найти её, где возможно.

Полнота вычислялась по формуле

$$Recall = \frac{TP}{TP + FN},$$

где TP — истинно-положительное решение, FN — ложно-отрицательное решение. В нашем случае, TP было количеством предложений с верно найденными моделями, а $TP + FN$ — всеми предложениями данного тематического класса.

Точность вычислялась по формуле

$$Precision = \frac{TP}{TP + FP},$$

где TP — истинно-положительное решение, FP — ложно-положительное решение. В нашем случае, TP было количеством предложений с верно найденными моделями, а $TP + FP$ — всеми предложениями, в которых были найдены модели данного тематического класса.

В связи со спецификой тематических моделей, они встречались в предложениях намного чаще семантических моделей, но несли в себе меньше информации. Для более точной оценки модели были проранжированы по их типу: семантические модели оценивались выше, чем тематические. Таким образом, при наличии в предложении хотя бы одной семантической модели, тематические модели больше не учитывались в статистике.

В таблице 2 приведены результаты анализа текста, его полнота и точность.

Таблица 2. Общие результаты.

	Предложения данного типа, кол.	Правильно найдено моделей, кол.	Предложения с моделями данной темы, кол.	Полнота	Точность	F-мера
Рождение	135	131	149	97,04%	87,92%	92,25%
Образование	377	298	509	79,05%	58,55%	67,27%
Родительская семья	72	56	265	77,78%	21,13%	33,23%
Смерть	111	83	86	74,77%	96,51%	84,26%
Место работы	113	84	419	74,34%	20,05%	31,58%
Национальность	14	10	45	71,43%	22,22%	33,90%

Место жительства	94	45	296	47,87%	15,20%	23,08%
Семья	48	28	53	72,92%	57,38%	64,22%
Род занятий	946	580	894	61,31%	64,88%	63,04%
Профессиональные события	490	202	642	41,22%	31,46%	35,69%
Личные события	105	11	30	10,48%	36,67%	16,30%
Прочее	319	120	418	37,62%	28,71%	32,56%
Всего	**2959**			**60,94%**	**44,68%**	**47,38%**

Все темы разделены на две группы: узкие тематики и широкие. Количество лиц и событий в группе узких тематик довольно ограничено, поэтому они использую небольшое количество моделей и семантических классов. В свою очередь, группа широких тематик обладает более разнообразным описанием событий и явлений, что требует большего количества различных моделей.

Анализ таблицы показывает значимую разницу в результатах между группами узких и широких тематик: значение полноты у узких тематик значительно выше, чем у широких. Это объясняется тем, что группа узких тематик имеет ограниченную лексику, что позволяет созданным для них моделям покрыть большую часть этой лексики, в то время как группа широких тематик обладает более богатой лексикой с различной семантикой, которую покрыть моделями намного сложнее.

На полноту влияет семантическая разметка словаря и разнообразие моделей, в то время как на точность влияет неоднозначность лексики и моделей и степень обобщенности модели.

Разберём подробнее тематические классы. Группа узких тематик имеет множество схожих черт. Тематические классы "Рождение" и "Смерть" имеют высокий показатель полноты и точности в результате наличия слов "родиться", "умереть", "похоронен" и других. Такие слова всегда присутствуют в предложениях данных тем и почти никогда не встречаются в других темах.

"Образование" имеет достаточно высокий процент полноты и относительно высокий процент точности по тем же причинам. Большинство предложений имело конструкцию "окончил что-то", "поступил туда-то", "получил такую-то степень" и т.д. или содержали названия вузов, школ, академий.

Предложения темы "Родительская семья" имели названия членов родной семьи, такие как "отец", "мать", "сестра" и другие, поэтому модели данной темы ориентировались именно на эти слова. Это является причиной относительно высокой полноты и низкой точности, так как многие слова этого семантического класса иногда встречаются в предложениях других тем. Данную проблему можно решить, если сделать модели менее обобщёнными.

Такая же проблема коснулась "Национальности" и "Места работы". Для "Национальности" была создана всего одна тематическая модель, которая отвечала за существительные и прилагательные, обозначающие национальность, так как предложения этой тематики могли выглядеть как "Он был грузином", так и "Рос в грузинской семье". Предложения "Места

работы" чаще всего говорили об участии в какой-либо организации, поэтому модели в основном состояли из слов, обозначающих принадлежность к этой организации, и названий этих организаций. Так как такие слова могли легко встретиться и в предложениях других тем, здесь точность была довольно низкой.

Тематический класс “Семья” во многом схож с “Родительской семьей”, за исключением того, что здесь использовались слова “супруг”, “сын”, “дочь” и подобные, и потому имеет схожие модели и, следовательно, те же проблемы.

“Место жительства” имеет самые низкие значения полноты и точности в данной группе. В основном это связано с тем, что, помимо глаголов “переехать”, “проживать” и подобных, словами, объединяющими все предложения данной темы, были названия городов, областей, республик и стран. Проблемой стало то, что морфологический анализатор был неспособен определить названия многих малоизвестных деревень и сёл, из-за чего пострадала полнота. Точность довольно низкая из-за того, что названия городов и сёл также используются в предложениях других тематик. Решить проблему с полнотой можно, если усовершенствовать морфологический анализатор, а с точностью, если написать модели, менее ориентированные на названия населенных пунктов и более на глаголы, означающие проживание и переезд.

Группу широких тематик объединяет большое количество событий и явлений, которые они описывают, из-за чего конструкции предложений и их лексика становится достаточно разнообразной.

"Профессиональные события" и "Род занятий" являются достаточно близкими тематическими классами, из-за чего их модели во многом совпадают. Они различаются лишь в том, что "Род занятий" указывает на профессию или вид деятельности человека, а "Профессиональные события" - на некоторые события, связанные с его карьерой. В связи с этим, полнота "Рода занятий" выше, чем у "Профессиональных событий": для "Рода занятий" достаточно обозначить названия профессий и должностей, тогда как для "Профессиональных событий" нужно прописывать модели для совершенно разных случаев: награждений, служб, театральных постановок и т.д. "Род занятий" относится к группе широких тематик из-за большого разнообразия профессий, для которого требуется хорошо размеченный словарь.

"Личные события" отличаются ещё большим разнообразием: в личной жизни человека может происходить что угодно. Модели данной темы были разделены на небольшие подгруппы, определяющие некоторые возможные события: ранение, потеря родных, судимость и т.д. - однако их все равно было недостаточно. Повысить полноту здесь возможно только написав больше моделей, отражающих различные события человеческой жизни.

Тематический класс "Прочее" содержит предложения, которые несут в себе некоторые биографические факты, но не относятся ни к какой другой теме. Единственное, что их связывает, - это отношение к персоне, о которой рассказывается в тексте, поэтому были созданы 3 тематические модели для имени, фамилии и местоимении. Для данного класса высокие показатели полноты и точности не ожидались, так как, во-первых, не во всех предложениях есть имена, фамилии и местоимения, и, во-вторых, они характерны для предложений любой тематики.

Таким образом, были определены основные признаки, которые влияют на полноту и точность поиска моделей.

Заключение

Таким образом, в рамках работы был разработан метод построения семантико-синтаксических моделей для извлечения информации из текстов.

Были выполнены следующие задачи:

- Разработано представление унарных моделей и билексем;
- Разработан формат представления семантико-синтаксических моделей;
- Разработан алгоритм и создана программа, осуществляющая анализ текста на основе ССМ;
- Предложен подход к экспериментальной оценке качества анализа на основе тематически размеченного корпуса текстов;
- Построено 96 моделей на основе частотности слов в корпусе биографических текстов;
- Проведено экспериментальное исследование.

Целью дальнейшего исследования может рассматриваться разработка автоматического построения моделей с помощью машинного обучения на основе частотности встречаемых в тексте слов и их валентностей.

Список литературы

1. Поречный А. С. Построение семантико-синтаксической модели текстов для определения их смысловой близости // Информатика: проблемы, методы, технологии : сб. мат-лов XXI Междун. науч.-метод. конф. (Воронеж, 11–12 февраля 2021 года). Воронеж: Общество с ограниченной ответственностью "Вэлборн", 2021. С. 1488-1495.
2. Ефремова Н. Э. Методы и программные средства извлечения терминологической информации из научно-технических текстов: автореф. дис ... канд. физ.-мат. наук. М., 2013. 18 с.
3. Глазкова А.В. Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке // Труды Института системного программирования РАН. 2018. Т. 30. №6. С. 221-236.
4. Денисов М. Е., Катышев А. М., Сычев О. А., Аникин А. В.. Извлечение ключевых понятий и связей между ними из тематических текстов на русском языке // Инженерный вестник Дона. 2022. № 12(96). С. 338-345.
5. Виноградова Н. В, В. К. Иванов. Современные методы автоматизированного извлечения ключевых слов из текста // Информационные ресурсы России. 2016. № 4(152). С. 13-18.

6. Бледнов А. М. Разработка и исследование моделей и информационной технологии семантико-синтаксического анализа русскоязычного текста: автореф. дис ... канд. тех. наук. Ижевск, 2007. 20 с.
7. Мельчук И. А.. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974 (2-е изд., 1999).
8. Apresjan J., Boguslavsky I., Iomdin L., Lazourski A., Sannikov V., Sizov V., Tsinman L.. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning Text Theory // MTT 2003, First International Conference on Meaning — Text Theory (June 16–18 2003). Paris: Ecole Normale Supérieure, 2003. P. 279–288
9. Bolshakov I. A., Gelbukh A. F. The Meaning↔Text Model: Thirty Years After // J. International Forum on Information and Documentation, N 1. 2000.
10. Inshakova E. S., Sizov V. G.. An experimental rule-based parser for Russian employing the nlp resources of the Etap system // Computational Linguistics and Intellectual Technologies. Vol. 19(26), 2020. P. 387-399.
11. AOT :: Технологии. 2003. URL: http://www.aot.ru/technology.html (дата обращения: 10.05.2023)
12. Cunningham H., Maynard D., Bontcheva K., Tablan V.. GATE: A Framework and Graphical Development Environment for Robust NLP Tools

and Applications // Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002. P. 168-175.

13. RCO Fact Extractor SDK. – 2023. URL: http://www.rco.ru/?page_id=3554 (дата обращения: 10.05.2023)

14. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: сб. мат-лов XI Межд. науч. конф. Москва, 2003.

15. Ермаков А.Е., Плешко В.В. Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии, N 6. 2009.

16. Яковчук Е.И., Сидорова Е.А. Обобщенные семантико-синтаксические модели в задачах обработки текста // Тр. конф. «Наукоемкое программное обеспечение НПО-2011». Новосибирск: ИСИ СО РАН, 2011. С. 287-292.

17. Сидорова Е.А. Комплексный подход к исследованию лексических характеристик текста // Вестник СибГУТИ, №3, 2019. С. 80-88.