МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Кафедра 319 «Системы интеллектуального мониторинга»

Семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных (Linked Open Data)

Автор:
студент группы М3О-435Б-18
Урубков В.С.
Научный руководитель:
д.т.н. профессор Фомичев В.А.

Москва, 2022

Цели и задачи

Необходимо реализовать семантически-ориентированный естественно-языковой интерфейс для взаимодействия с системой взаимосвязанных открытых данных.

Для этого необходимо:

- выбрать подход к описанию семантического представления текста на естественном языке;
- разработать алгоритмы для реализации преобразования «ЕЯ-запрос → SPARQL-запрос».

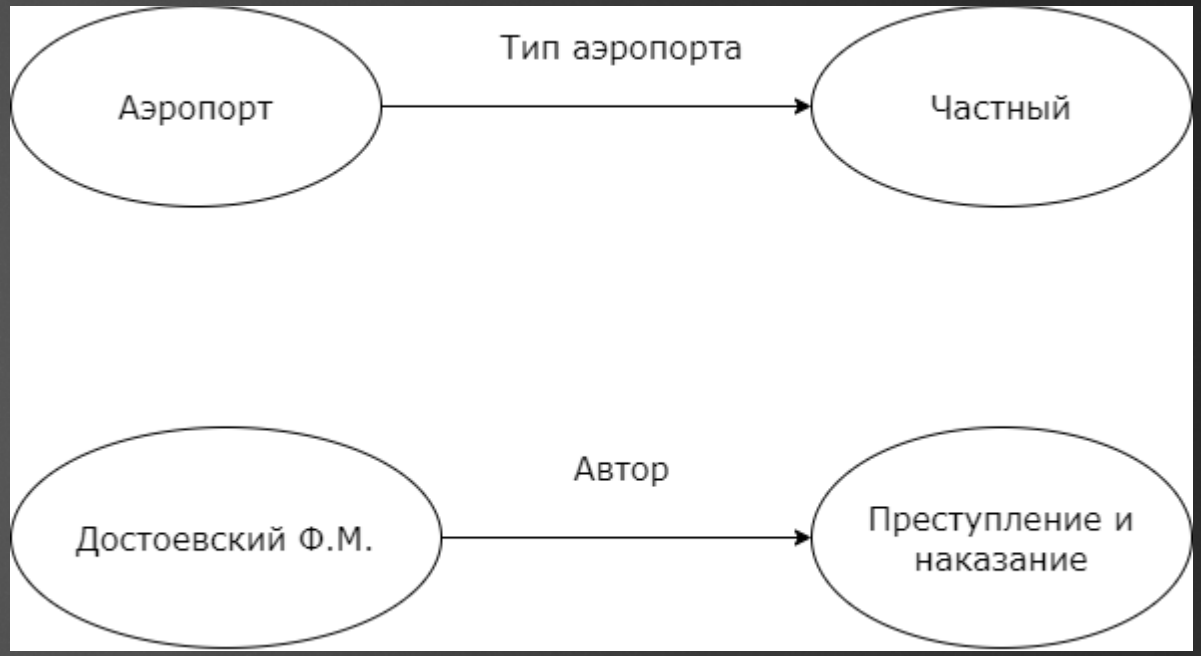
Linked Open Data (LOD)

Система LOD – огромный размеченный ориентированный граф, состоящий из элементарных графов, представляющих тройки вида (субъект, предикат, объект).

RDF

Resource Description Framework – язык для создания распределенных баз знаний (онтологий).

Основной структурой языка RDF являются триплеты – упорядоченные тройки вида (субъект, предикат, объект)



SPARQL

SPARQL (рекурсивный акроним SPARQL Protocol and RDF Query Language) – язык запросов к данным, представленным в формате RDF.

```
select distinct ?name
where {
    ?town rdf:type dbo:City.
    ?town dbo:country dbr:Russia.
    |
    ?town rdfs:label ?name
    filter(lang(?name) = "ru")
}
```

name
"Ульяновск"@ru
"Астрахань"@ru
"Бийск"@ru
"Челябинск"@ru

Подходы к описанию семантического представления текстов на естественном языке

Абстрактное представление смысла

Теория К-представлений В.А. Фомичева

Грамматика Монтегю

Сравнение подходов

Критерий	Абстрактное представление смысла	Грамматика Монтегю	Теория К-представлений В.А. Фомичева
Язык текста	Английский	Английский	Русский, Немецкий, Французский, Английский
Типы текстов	Повествовательные предложения	Повествовательные предложения и вопросы	Фразы-высказывания, повествовательные тексты, команды, вопросы
Допустимая структура текста	Отдельное предложение	Отдельное предложение	Связный текст наравне с отдельными предложениями

Лингвистическая база данных

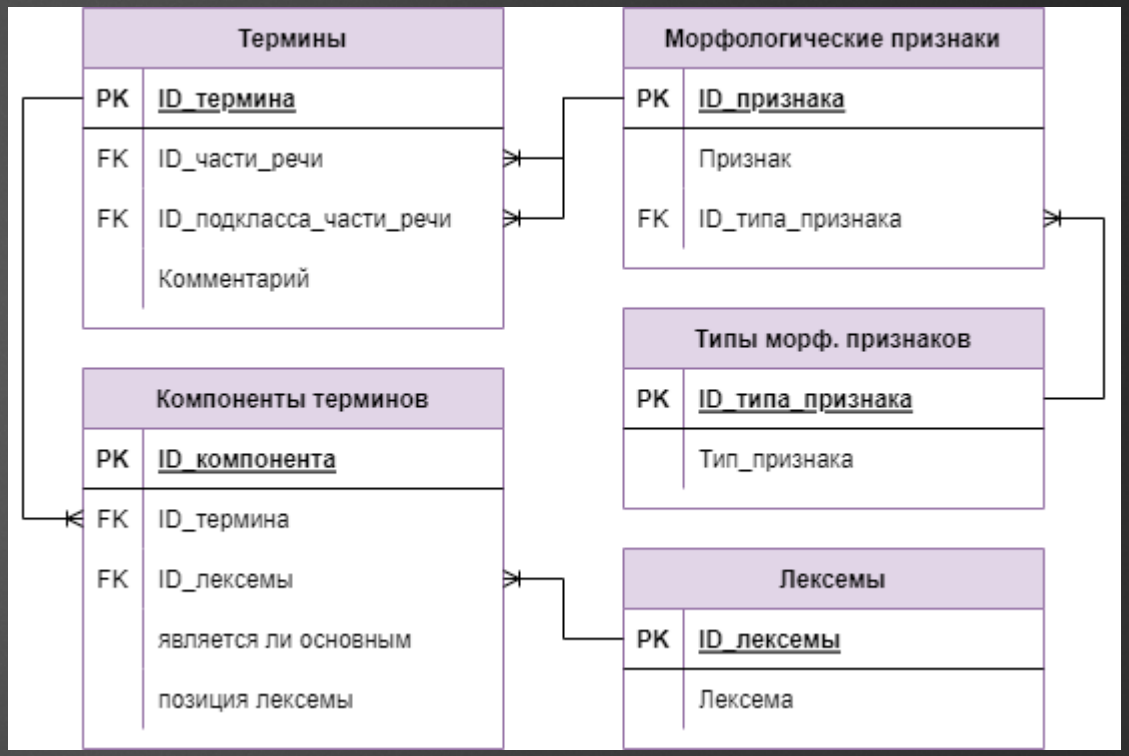
Состоит из:

- Морфологической базы данных
- Лексико-семантического словаря
- Словаря предложных фреймов

Морфологическая база данных

Содержит информацию о лексемах, терминах, а также набор возможных морфологических признаков.

Для определения морфологических признаков используется нейросетевая библиотека DeepMorphy.

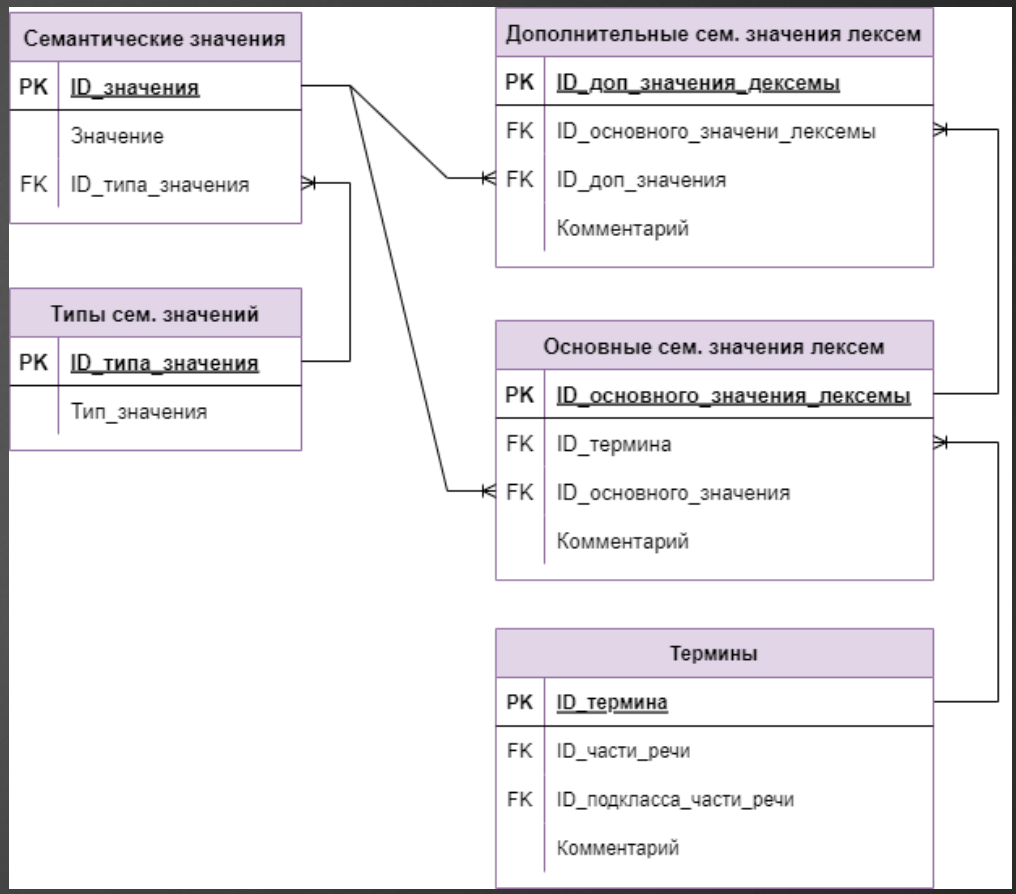


Лексико-семантический словарь

Содержит семантические значения лексем и фреймов.

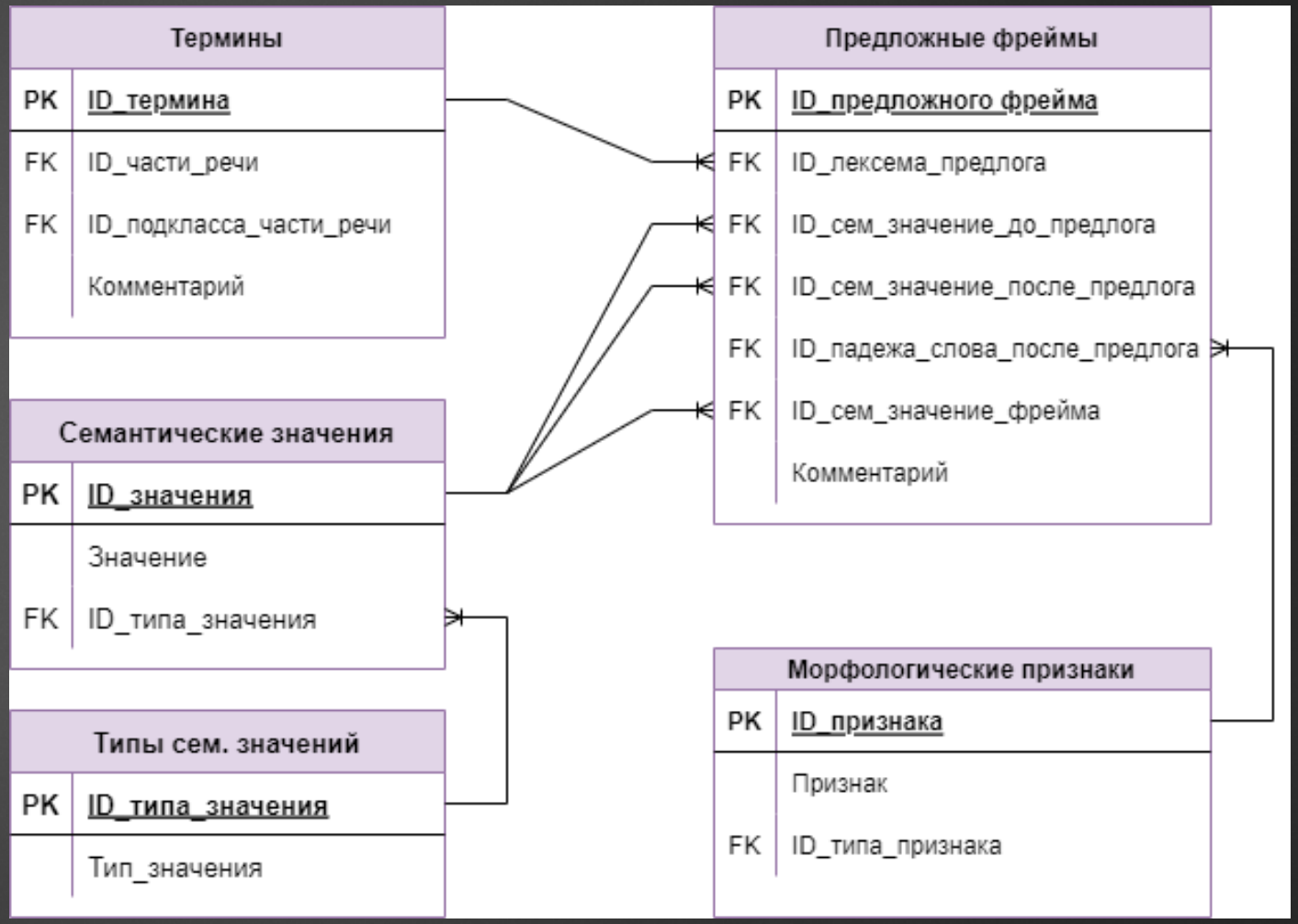
Типы семантических значений:

- Основное значение
- Дополнительное значение
- Значение фрейма



Словарь предложных фреймов

Предложный фрейм описывает смысловое отношение двух существительных, связанных предлогом, в том числе нулевым



Структура входного запроса на русском языке

Фрагмент1 Сущ1 Предлог Фрагмент2 Сущ2 Фрагмент3

- Фрагмент1 и Фрагмент2 являются либо пустой цепочкой, либо последовательностью прилагательных
- Сущ1 и Сущ2 – существительные
- Фрагмент3 является либо пустой цепочкой, либо искусственным именем, либо словосочетанием, определяющим сравнение с числом (например, «меньше 50000» или «не больше 60»)

Примеры запросов

Возможные входные запросы:

- Планета с самым большим радиусом
- Одноместные многоцелевые боевые самолёты российского производства
- Экспериментальные летательные аппараты Китая
- Широкофюзеляжные самолёты компании Airbus
- Частные аэропорты Германии
- Канадские города с населением меньше 50000

Структура семантического представления

A (B1, R1, C1) (B2, R2, C2) … (Bn, Rn, Cn)

- **A** – обозначение понятия на русском языке (самолёт, автомобиль, компания и т.д.)
- **B1**, **B2**, …, **Bn** – имена смысловых параметров семантического представления на русском языке
- **R1**, **R2**, …, **Rn** – имена бинарных отношений на русском языке
- **C1**, **C2**, …, **Cn** – обозначения значения параметра или второго атрибута отношения на русском языке

Неоднозначность именования в онтологиях

Предикаты, описывающие одно и то же отношение между объектами, могут иметь разные имена даже в рамках одной онтологии.

Город	Предикат
Оттава	population
Москва	populationTotal
Ульяновск	p
Северодвинск	pop2010census

Недостаточная связанность данных

В онтологии YAGO, использующей систему типов Schema:

- У любых объектов Автомобилей отсутствуют содержательные предикаты, хотя информация для них есть
- Объект Город связывается со страной, в которой он располагается, с помощью строки с комментарием типа «Это столица России»

rdfs:comment	"сталіца Pacei"@be-tarask
	"Столица на Русия"@bg
	"capital de Rússia"@ca
	"hlavní město Ruské federace"@cs
	"Hovedstad i Rusland"@da

Принципы преобразования параметров запросов к LOD

Для обеспечения перевода запроса к LOD на естественном языке в запрос на языке SPARQL необходимо заранее связывать параметры запроса (отношения и некоторые значения) с аналогичными параметрами онтологии.

Отношение	Предикаты
Колич-Жителей	population
	populationTotal
	p
	pop2010census
Россия	dbr:Russia

Компонент разрешения имен

Необходим для связывания параметров К-представления с параметрами онтологии

Преобразование семантического представления в SPARQL-запрос

SPARQL-запрос условно можно поделить на части следующих типов:

- Заголовок
- Тройки равенства
- Тройки сравнения

Заголовок SPARQL-запроса

В заголовке определяется тип искомой сущности на основе понятия (**A**), указанного в K-представлении входного запроса

Самолёт

```
SELECT DISTINCT ?var1
WHERE {
VALUES ?var2 {dbo:Aircraft} .
?var1 rdf:type ?var2 .
```

Тройки равенства

(Страна, =, Россия)

```
VALUES ?p3 {dbo:country} .
VALUES ?var3 {dbr:Russia} .
?var1 ?p3 ?var3 .
```

(Экипаж, =, 4)

```
VALUES ?p5 {dbo:Crew}
.?var1 ?p5 4.
```

(Радиус, =, #макс#)

```
VALUES ?p4 {dbo:radius,
dbo:meanRadius} .
?var1 ?p4 ?var4 .
…
} ORDER BY DESC (?var4) LIMIT 1
```

Тройка сравнения

(Колич-жителей, <, 50000)

```
VALUES ?p4 {dbp:population
dbp:populationTotal dbp:p
dbp:pop2010census} .
?var1 ?p4 ?var4 .
FILTER (?var4 < 50000) .
```

Средства разработки

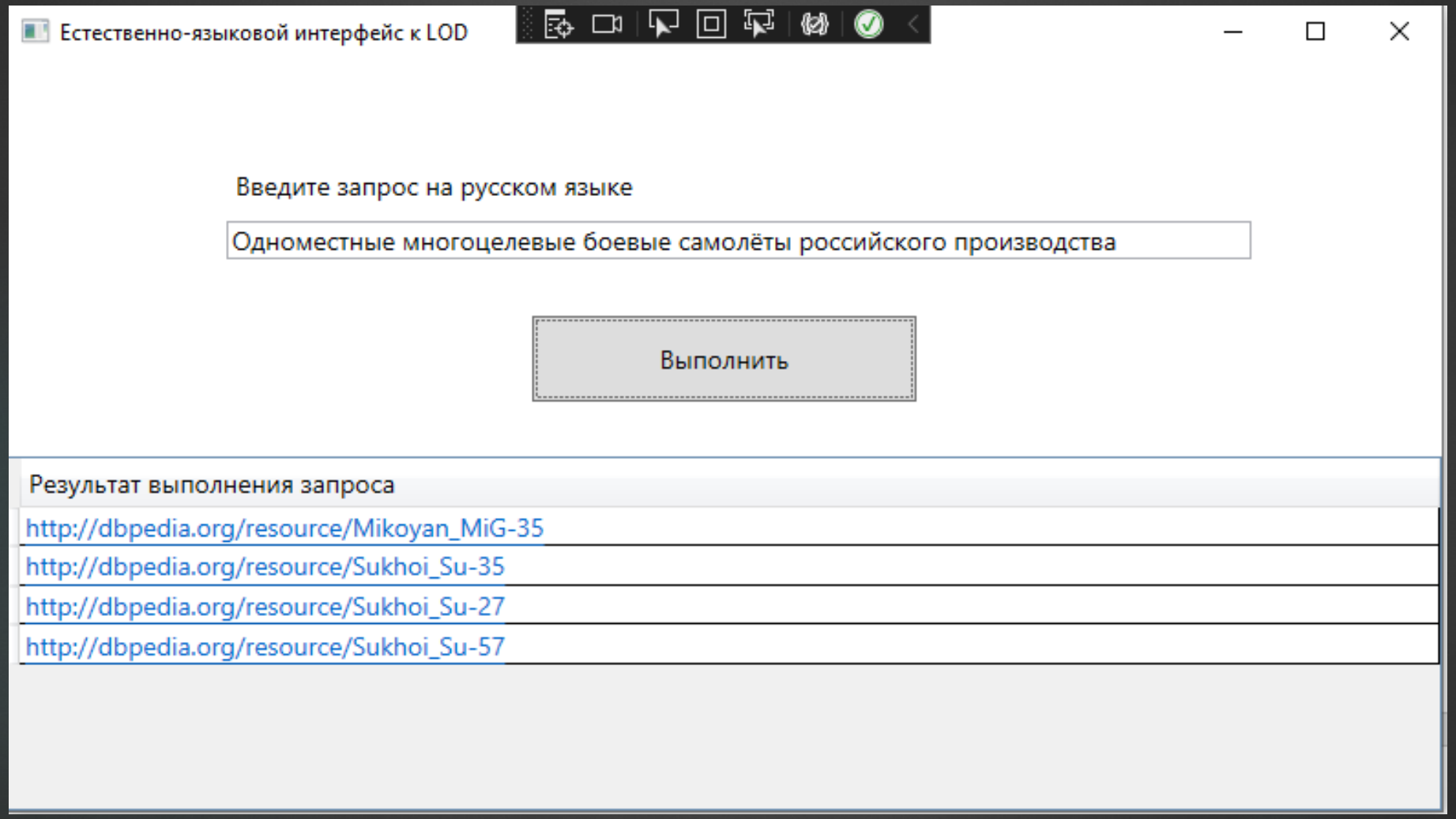
Для программной реализации приложения была выбрана платформа **.NET** (версия **.NET6**).

В качестве СУБД выбрана **PostgreSQL**.

Использовались следующие инструменты и технологии, предоставляемые платформой .NET:

- **LINQ to Entity Framwork** (взаимодействике с базой данных),
- **DeepMorphy** (определение морфологических признаков),
- **dotNetRdf** (выполнение SPARQL-запросов),
- **WPF** (оконное приложение)

Интерфейс приложения



Выводы

В ходе данной работы:

- Был разработан и реализован семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных
- Проведено сравнение подходов к формальному описания семантической структуры текстов и выбран подход теории К-представлений В.А. Фомичева
- Разработана и реализована лингвистическая база данных
- Разработаны алгоритмы для реализации преобразования «ЕЯ-запрос → SPARQL-запрос»
- Предложен принцип преобразования параметров запроса, позволяющий преодолевать проблему неоднозначности имен в онтологии