Новосибирский государственный университет
Кафедра систем информатики

Разработка редактора правил для извлечения информации из текстов на естественном языке.

Студент: Ли Д. В.
гр. 0201 ФИТ НГУ
Научный руководитель:
к.ф.-м.н., с.н.с. ИСИ СО РАН
Сидорова Е. А.

Цель работы

Создание редактора правил извлечения информации из текста

ЗАДАЧИ РАБОТЫ:

- Провести исследование предметной области.
- Разработать пользовательский интерфейс редактора.
- Разработать способ обеспечения корректности составления правил на основе онтологии и признаковой системы словаря.
- Разработать форматы входных/выходных данных в виде XML-документов.
- Разработать с помощью редактора набор правил для извлечения информации о конференциях.

Актуальность

- Разработка методов автоматического извлечения информации из текстов.
- Предоставить инструмент для поддержки экспертов и лингвистов.
- Обеспечить контроль за вводимыми экспертом данными.

Модель представления знаний

- Онтология – понятия и отношения предметной области
- Предметный словарь (тезаурус) – термины описывающие понятия и отношения онтологии
- Модель извлечения фактов – описывает способы выражения фактов в тексте и их связь с элементами онтологии

Онтология

$O = \langle C, R_c, A, T, D \rangle$

- C – множество классов, описывающих понятия предметной или проблемной области;
- R_c – множество отношений, заданных на классах (понятиях);
- A – множество атрибутов, описывающих свойства понятий и отношений;
- T – множество стандартных типов значений атрибутов (*string, integer, real, date, bool*);
- D – множество доменов (множеств значений стандартного типа *string*).

Пример онтологии

class Персона (Фамилия: string; Имя: string; Отчество: string; Инициалы: string; ПолноеИмя: string; Звание: set of domen_Звания)

class Регион (Название: string; Тип: domen_Географический_Тип)

relation Включение-Регион < Регион, Регион >

class Организация (Название: string; Аббревиатура: string)

class Институт : Организация

class Филиал : Организация

…

relation Включение-Организация < Организация, Организация >

relation Сотрудник <Персона, Организация>
(Должность: domen_Должности; Дата1: data; Дата2: data)

Тезаурус

$T = \langle V, S, M, G_S \rangle$

V – конечное множество терминов

S – конечное множество семантических признаков словаря

M – конечное множество морфологических признаков словаря

G_S – конечное множество синсетов

Схема извлечения факта

< A, Res, C >

A – множество дескрипторов аргументов факта.

Res = < t, op(t), P > – результат применения схемы, где

- t – задает тип элемента (класс результирующего объекта);
- op(t) – тип операции, применяемой, если все условия СИФ выполнены (создание или редактирование объекта);
- P – множество правил для формирования результирующего объекта.

C – множество ограничений, накладываемых на характеристики аргументов факта.

Пример схемы

Scheme конференция

arg1: Term*::Номер_конференции*()

arg2: Object*::Конференция*()

Condition Position = preposition_priority, Contact = absolute

⇨ Arg2::*Конференция*(Номер: arg1.Номер)

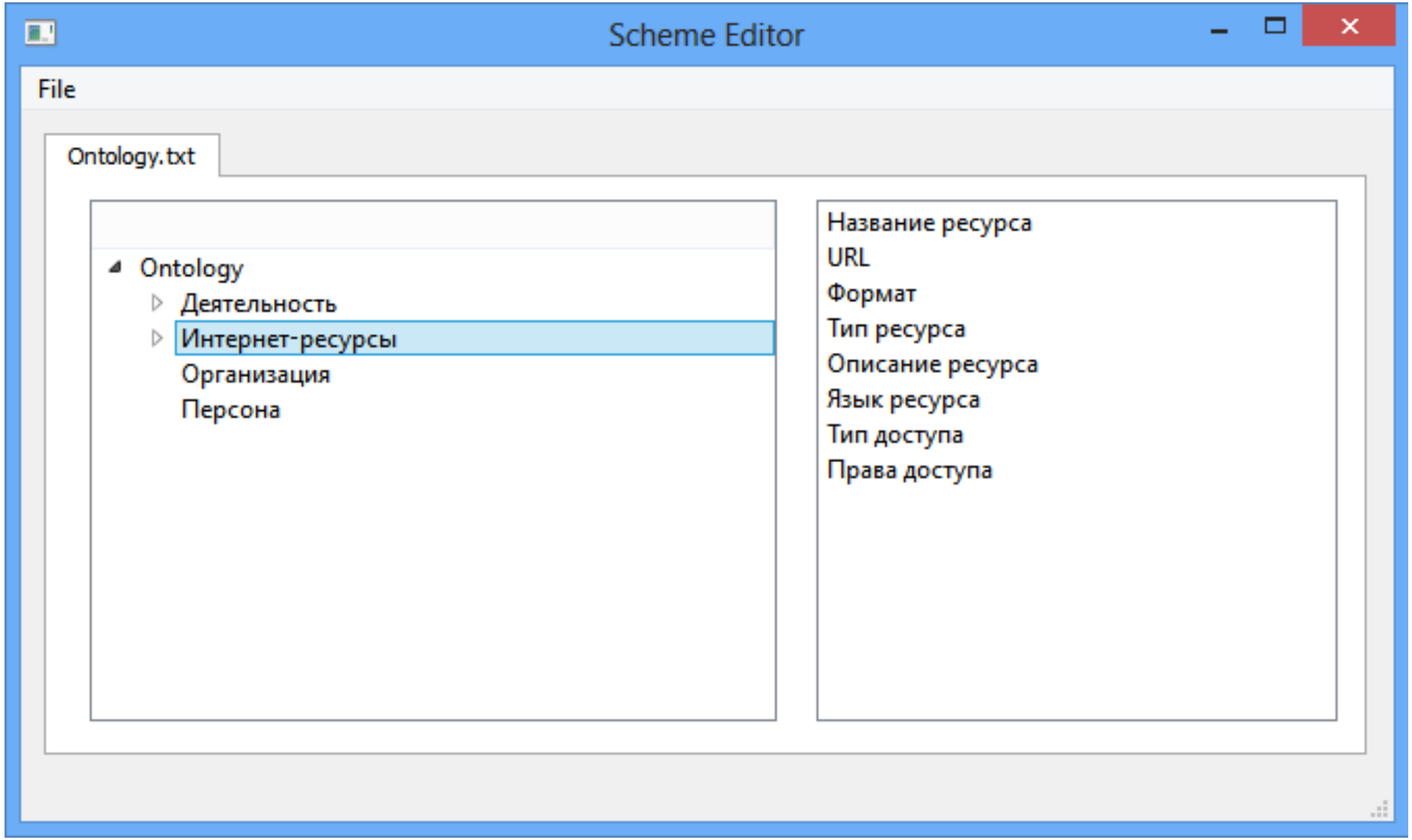
Пример XML-представления схемы

```
<?xml version="1.0" encoding="Windows-1251"?>

<FATON version="1.0" description="FATON Scheme of Facts">
    <Bank Name="MagArhive" Type="Schemes">

        <Scheme NumArg="2" CountType="0" Segment="">
            <Argument ObjectType="TERMIN" ClassName="фио" TypeCompare="EQUAL">
                <Condition Type="MORH" Operation="=" AttrName="фио-тип" AttrType="string" Data="фам" />
            </Argument>
            <Argument ObjectType="TERMIN_ALEX" ClassName="инициалы" TypeCompare="EQUAL"/>
            <ConditionStruct Res1="Arg1" Res2="Arg2" CondType="Position" Contact="CONTACT_ABSOLUTE" TextPos="POSITION_ANY" />
            <Result Type="CREATE" ObjectType="IOBJECT" ClassName="Персона" >
                <Rule AttrName="Инициалы" AttrType="string" Resurce="Arg2" FromAttrName="Value"/>
                <Rule AttrName="Фамилия"  AttrType="string" Resurce="Arg1" FromAttrName="Name" />
            </Result>
        </Scheme>
    </Bank>
</FATON>
```

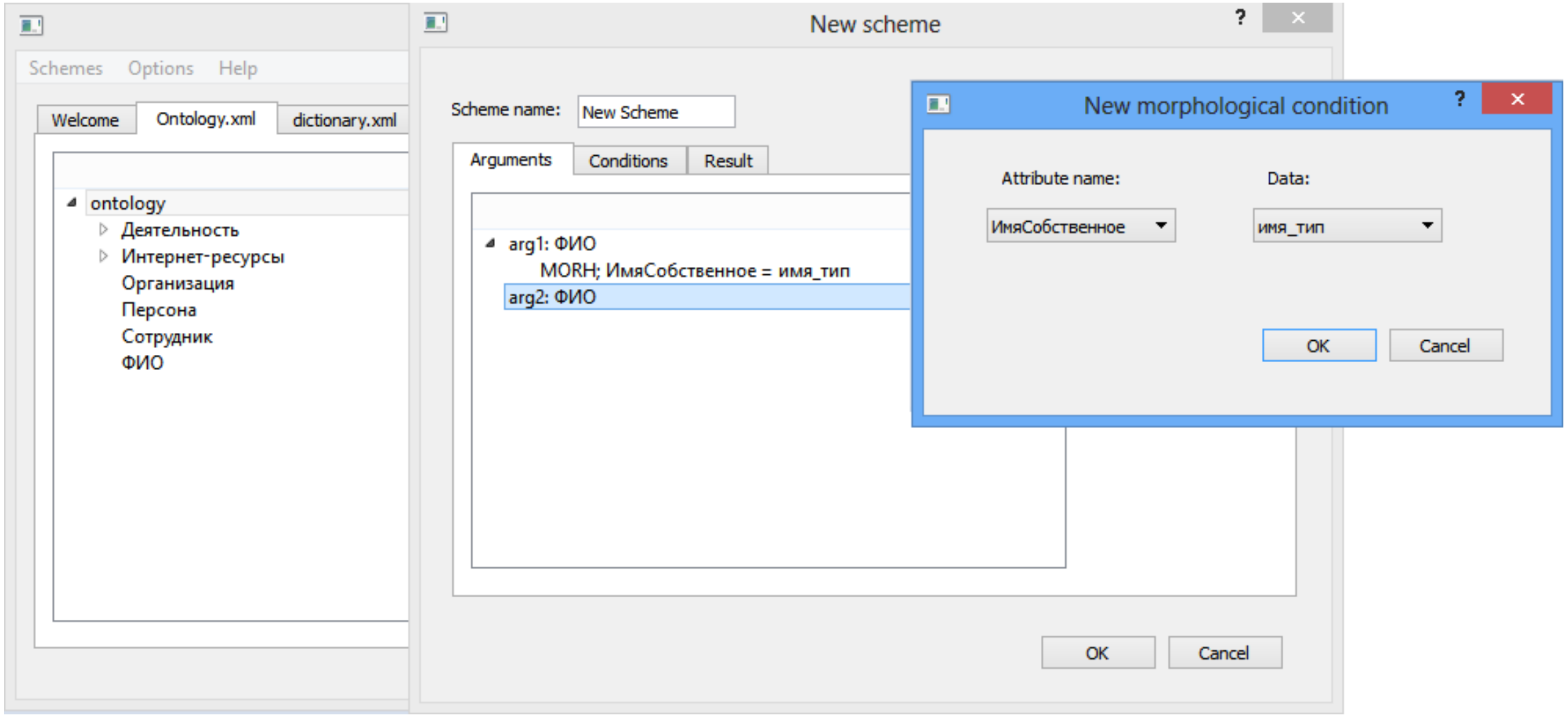
Требования к редактору

- Проверка корректности входных данных
- Возможность создания правил для различных языков
- Кроссплатформенность

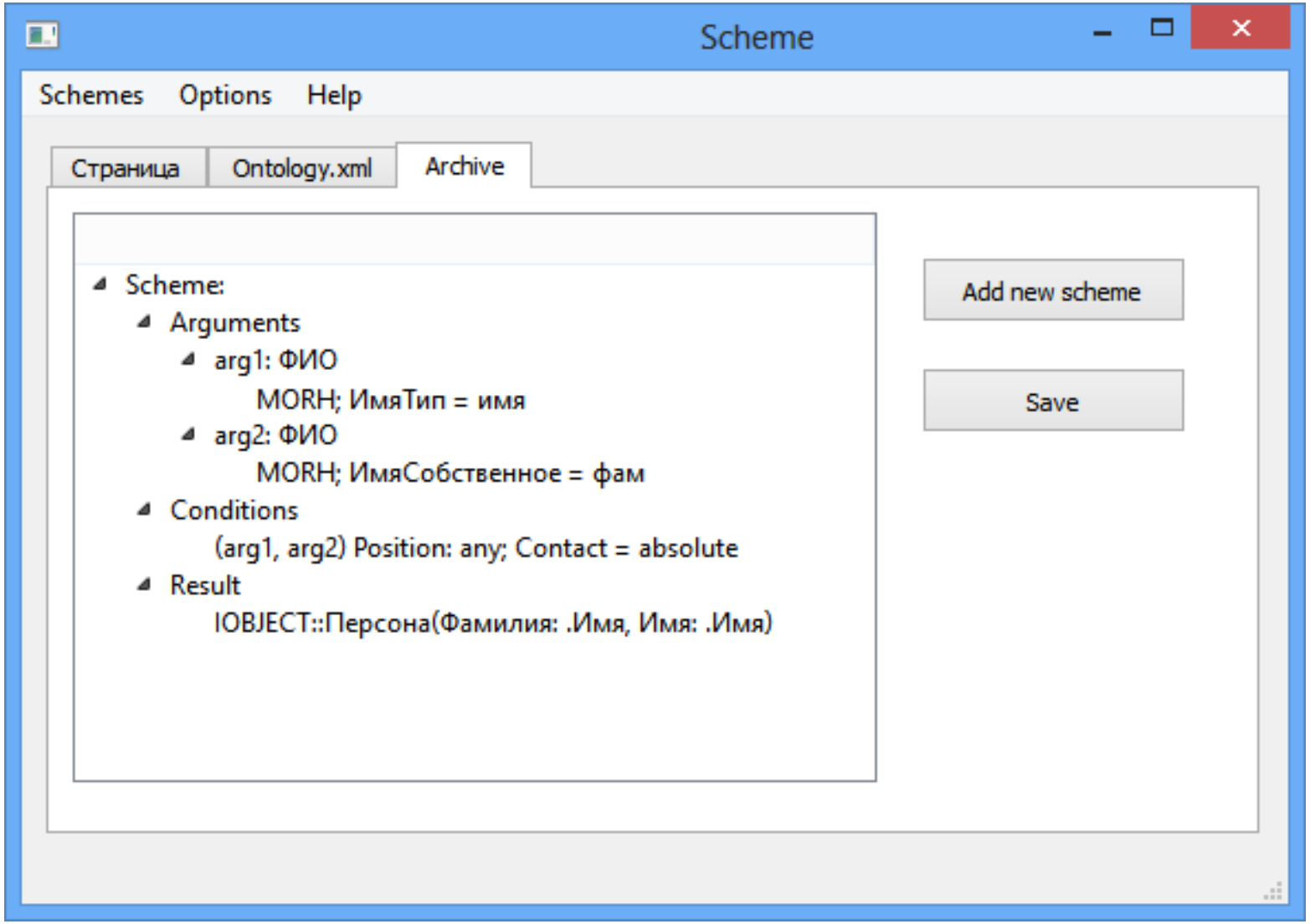
Редактор правил



Редактор правил



Редактор правил



Эксперимент

class Конференция(string:Название; string: Номер_конф)

class Место_проведение(string:Город; string:учреждение)

class Организация(string:Название)

Relation <Конференция, место проведения>

Relation <Конференция, Организатор >

Эксперимент

Scheme конференция
arg1: Term*::событие*()

⇨ Object::*Конференция*(Название: arg1.Название)

Scheme Место_проведения
arg1: Term*::город*()

⇨ Object::*Место_проведения*(Город: arg1.Город)

Scheme
arg1: Object*::Место_проведения*()
arg2: Object*::Конференция*()

Condition Position = any, Contact = object

⇨ Arg2::*Конференция*(Место: arg1.город; Организатор: arg1.Учреждение)

Результаты

- Изучена предметная область, проведен обзор существующих систем извлечения информации
- Разработан формат представления онтологий, семантических и морфологических признаков словаря в виде XML-документа
- Реализован пользовательский интерфейс редактора
- Разработан подход обеспечения корректности составленных фактов
- Проведен эксперимент

Спасибо за внимание