Семантический анализатор русскоязычного текста для вопросно-ответной системы

Мочалова Анастасия (*stark345@gmail.com*)
Мочалов Владимир (*sensorlife@mail.ru*)

Актуальность проблемы

Задачи

- Разработка математической модели онто-семантического анализатора (ОСА) русскоязычного текста;
- Разработка и программная реализация:
  - алгоритма работы ОСА;
  - алгоритма работы ВОС, основанной на результатах работы ОСА;
- Разработка архитектуры ВОС, основанной на сравнении онто-семантических графов.

Семантическая зависимость

- R^* – множество названий сем. зависимостей;

  R^* = {«Действие», «Место», «Время», «Кто», «Цель», «За_кем», «От_чего», «С_кем», «Перед_чем», …}

  $|R^*| = 126$

- α, β – неделимые смысловые единицы

Будем говорить, что две неделимые смысловые единицы α и β связывает семантическая зависимость с именем R: R *из* R^* (обозначим $R(\alpha, \beta)$), если верно одно из двух высказываний:

- «β является R для α»,
- от α к β можно задать вопрос R

- ПРИЗНАК(*вечер, теплый*)
  («теплый» является Признаком для «вечера»)
- ВМЕСТО_КОГО(*прочитать, вместо Иван*)
  («прочитать» вместо кого? – вместо Ивана)

Математическая модель ОСА

Входные данные :

$T = \{t_1, \ ..., \ t_n\}$ – анализир. текстк, $\forall t_i$ – синтаксема

$S = \{s_1, \ ..., \ s_k\}$ – множество БОСП

$s_i = \left(P(T_i^0), R(T_i^1, T_i^2), pos_{del}, sp_i \right)$ – БОСП

$$\left[\left(T_i^1 \cup T_i^2 \right) \subset T \right] \wedge \left[\left(T_i^1 \cap T_i^2 \right) \neq \varnothing \right]$$

$$P\left(T_i^0 \right) = \left\{ p\left(t_i^1 \right), ..., p\left(t_i^r \right) \right\},$$

где $p\left(t_i^j \right)$ – мн-во морф. и онт. свойств t_i^j

Выходные данные :

$$R = \left\{ R_1\left(T_1^{R_1}, T_2^{R_1} \right), R_2\left(T_1^{R_2}, T_2^{R_2} \right), ..., R_m\left(T_1^{R_m}, T_2^{R_m} \right) \right\}$$

Математическая модель ОСА

Очередь с приоритетом $Q = \{ q_1, ..., q_z \}$

$\forall q_i = \left(x_i^1, x_i^2, x_i^3 \right) : \qquad x_i^1 = T_i^1, \qquad x_i^2 = sp_i, \qquad x_i^3 = pos_{del}$

$h(q_i)$ – приоритет элемента q_i

$h(q_i) = x_i^2 \times L - x_i^3$

$h_{\max}$ – максимальное значение приоритетов элементов из Q

Операции над Q **:**

- $isEmpty(Q)$ – *проверяет Q на наличие в ней элементов*
- $Remove(Q)$ – *удаляет и возвращает эл – т из Q с наивысшим h*
- $Insert(Q, q_i)$ – *добавляет новый элемент в Q*

Математическая модель ОСА

Операции над Q

$$isEmpty(Q) = \begin{cases} 0, & \text{если } |Q| = 0 \\ 1, & \text{иначе} \end{cases}$$

$$Remove(Q) = \left(x_{del}^1, x_{del}^2, x_{del}^3 \right) : \quad h(q_{del}) = h\left(x_{del}^1, x_{del}^2, x_{del}^3 \right) = h_{\max}$$

$$Insert\left(Q, \left(x_i^1, x_i^2, x_i^3 \right)\right) = \begin{cases} Q, \text{ если } \exists \left(y_i^1, y_i^2, y_i^3 \right) \in Q : \begin{cases} \left(y_i^1 = x_i^1 \right) \wedge \\ \left(y_i^2 = x_i^2 \right) \wedge \\ \left(y_i^3 < x_i^3 \right) \end{cases} \\ Q, \text{ иначе} \end{cases}$$

Математическая модель OCA

Для $\forall s_i \in S$:

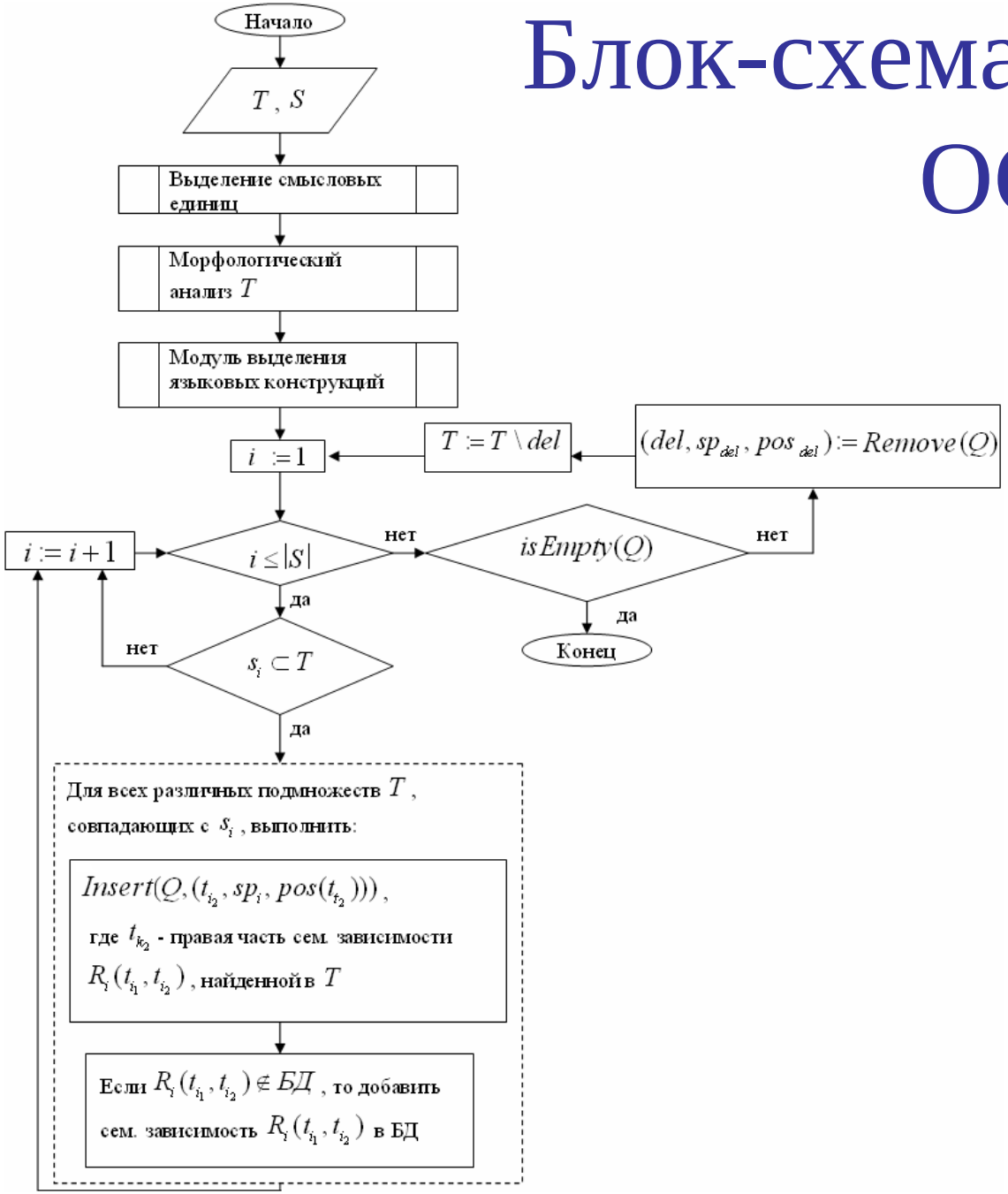
$$In(T, s_i) = \begin{cases} 1, \text{ если } \left(p(t_d) \in p\left(t_i^1\right)\right) \wedge \ \ldots \ \wedge \left(p\left(t_{d+r-1}\right) \in p\left(t_i^r\right)\right) \\ 0, \text{ иначе} \end{cases}$$

$$\left[\left(In(T, s_i) = 1 \right) \wedge \left(R_j\left(T_1^{R_j}, T_2^{R_j}\right) \notin R \right) \right] \Rightarrow R \cup R_j\left(T_1^{R_j}, T_2^{R_j}\right)$$

$isEmpty(Q) = 1 \Rightarrow$ Работа OCA завершается

$$isEmpty(Q) = 0 \Rightarrow \begin{cases} Remove(Q) = \left(x_{del}^1, x_{del}^2, x_{del}^3 \right) \\ T \setminus x_{del}^1 \end{cases}$$

Блок-схема алгоритма ОСА




РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2015613430

Программа семантического анализа текста, основанная на базовых семантических шаблонах с удалением

Правообладатель: Мочалова Анастасия Викторовна (RU)

Автор: Мочалова Анастасия Викторовна (RU)

Заявка № 2015610583
Дата поступления 28 января 2015 г.
Дата государственной регистрации в Реестре программ для ЭВМ 16 марта 2015 г.

Врио руководителя Федеральной службы по интеллектуальной собственности Л.Л. Кирий

Базовые онто-семантические правила

$P(T_i^0)$	$R(T_i^1, T_i^2)$	pos_{del}	sp_i
Г:-:- ПРЕДЛ:в:- С:-:пр,но,врем	**ВРЕМЯ|0|1,2|**	**1,2**	**5**
последовательность слов, или неделимых смысловых единиц, с указанием названия и морфологических признаков	название сем. отношения, которое должно быть сформировано в случае обнаружения в тексте искомой последовательности	позиции слов или неделимых смысловых единиц, заносимых в очередь на удаление	приоритет сем. группы

$$s_i = \left(P(T_i^0), R(T_i^1, T_i^2), pos_{del}, sp_i \right)$$

[0] [1] [2] [0] [1] [2]

Уехал в июле → ВРЕМЯ (уехать, в июль)

БОСП на языке Drools

```
rule "957Г:-:-       ПРЕДЛ:в:-       С:-:пр,но,врем"
salience 100
when
       $w0 : Fact( partOfSpeech == "Г")
       $w1 : Fact( prev == $w0, partOfSpeech == "ПРЕДЛ", wordName == "в")
       $w2 : Fact( prev == $w1, partOfSpeech == "С", hsAttrs contains "пр",
hsAttrs contains "но", hsAttrs contains "врем")
then
       SemanticRelation sem = new SemanticRelation("МЕСТО");
       sem.setLeftAutoPosInText($w0);
       Concept conceptRight = new Concept($w1, $w2);
       sem.setRightAutoPosInText(conceptRight);
       String strIndexConcRight = conceptRight.getIndexString();
       if(hsAllIndexedConcepts.contains(strIndexConcRight) == false)
       {
              hsAllIndexedConcepts.add(strIndexConcRight);
              insert(conceptRight);
       }
       boolean changed = myQueue.addOrUpdateCheckToDelete(conceptRight, 5);
       if(changed)
              update(myQueue);
       String indexSem = sem.getIndexString();
       if(hsAllIndexedSemanticRelations.contains(indexSem) == false)
       {
              hsAllIndexedSemanticRelations.add(indexSem);
              insert(sem);
       }
end
```

Пример работы сем. анализатора (1)

[0] [1] [2] [3] [4] [5] [6] [7] [8]

Вчера отличники школы яхтенного спорта выехали в лагерь.

- ЧЕГО(отличники, школы); ***sp*** = 3;
  - *insert*(***Q***,(школы, 3, 2))
- ПРИЗНАК(спорта, яхтенного); ***sp*** = 1;
  - *insert*(***Q***,(яхтенного, 1, 3))
- МЕСТО(выехали, в лагерь); ***sp*** = 2
  - *insert*(***Q***,(в лагерь, 2, 7))

- ***Q*** = {(школы, 3, 2); (яхтенного, 1, 3); (в лагерь, 2, 7)}
  - *remove*(***Q***,(яхтенного, 1, 3))

Пример работы сем. анализатора (2)

[0] [1] [2] [4] [5] [6] [7] [8]

Вчера отличники школы спорта выехали в лагерь.

- ЧЕГО(школы, спорта); ***sp*** = 3;
  - *insert*(***Q***,(спорта, 3, 4))

Q = {(школы, 3, 2); (в лагерь, 2, 7); (спорта, 3, 4)}

  - *remove*(Q,(в лагерь, 2, 7))

Пример работы сем. анализатора (3)

[0] [1] [2] [4] [5] [8]

Вчера отличники школы спорта выехали.

- Новых сем. связей не обнаружено
- isEmpty(Q) = false

Q = {(школы, 3, 2); (спорта, 3, 4)}

- *remove*(Q,(спорта, 3, 4))

Пример работы сем. анализатора (4)

[0] [1] [2] [5] [8]

Вчера отличники школы выехали.

- Новых сем. связей не обнаружено
- isEmpty(Q) = false

Q = {(школы, 3, 2)}

- *remove*(Q,(школы, 3, 2))

Пример работы сем. анализатора (5)

[0] [1] [5] [8]

Вчера отличники выехали.

- ДЕЙСТВИЕ(выехали, отличники); ***sp*** = 17;
  - *insert*(***Q***,(отличники, 17, 1))

Q = {(отличники, 17, 1)}

- *remove*(Q,(отличники, 17, 1))

Пример работы сем. анализатора (6)

[0] [5] [8]

Вчера выехали.

- ВРЕМЯ(выехали, вчера); ***sp*** = 8;
  - *insert*(***Q***,(вчера, 8, 0))

Q = {(вчера, 8, 0)}

- *remove*(Q,(вчера, 8, 0))

Пример работы сем. анализатора (7)

[5] [8]

выехали.

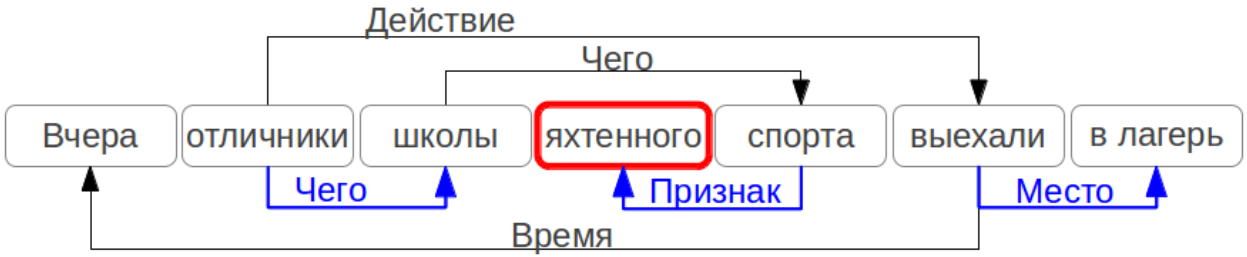
- Новых сем. связей не обнаружено
- isEmpty(Q) = true

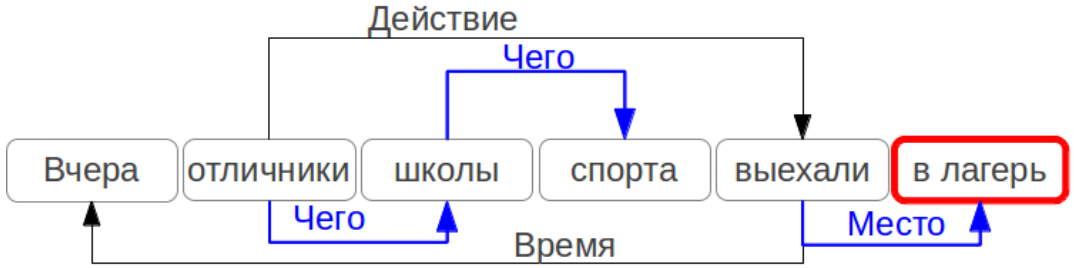
=> Конец

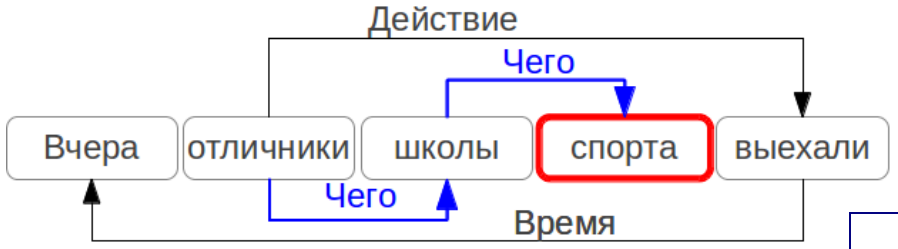
Результат работы семантического анализатора:

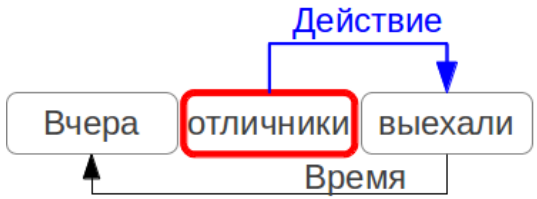
- ЧЕГО(отличники, школы)
- ПРИЗНАК(спорта, яхтенного)
- МЕСТО(выехали, в лагерь)
- ВРЕМЯ(выехали, вчера)
- ЧЕГО(школы, спорта)
- ДЕЙСТВИЕ(выехали, отличники)

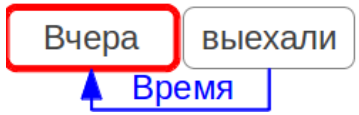
R







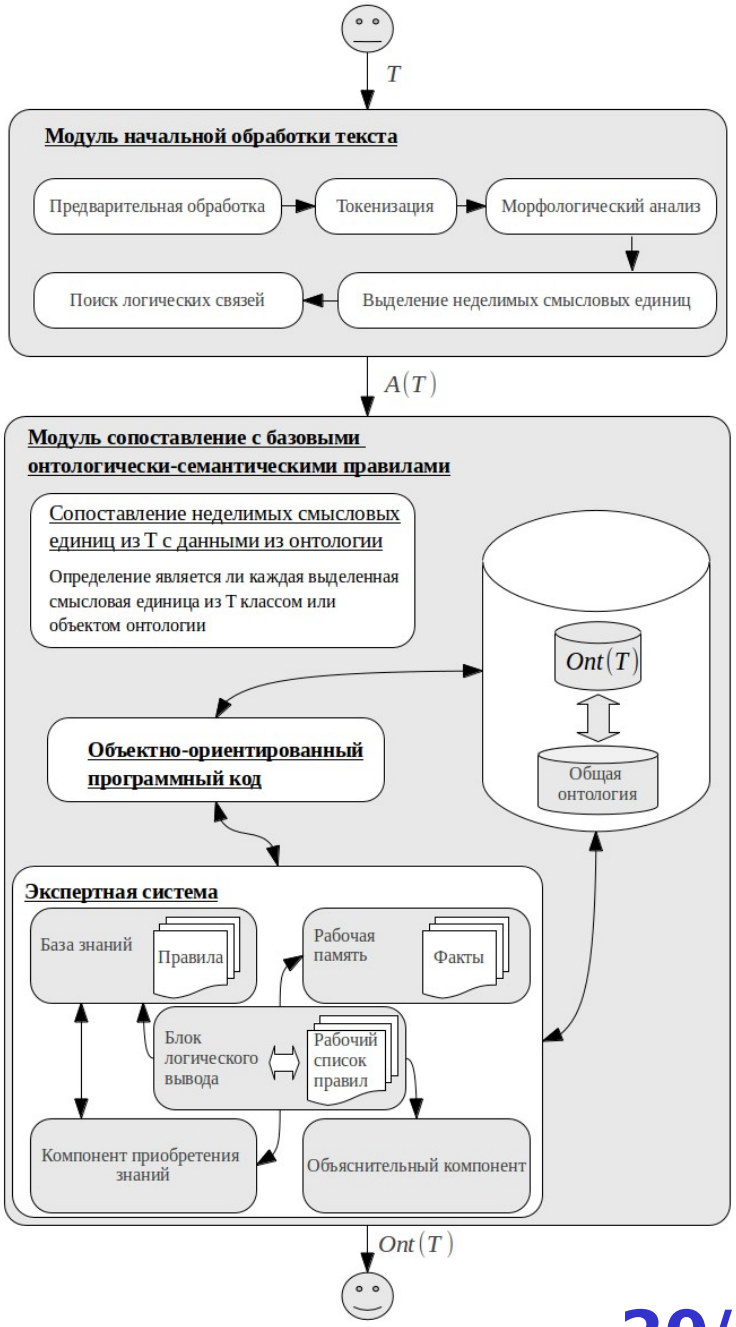




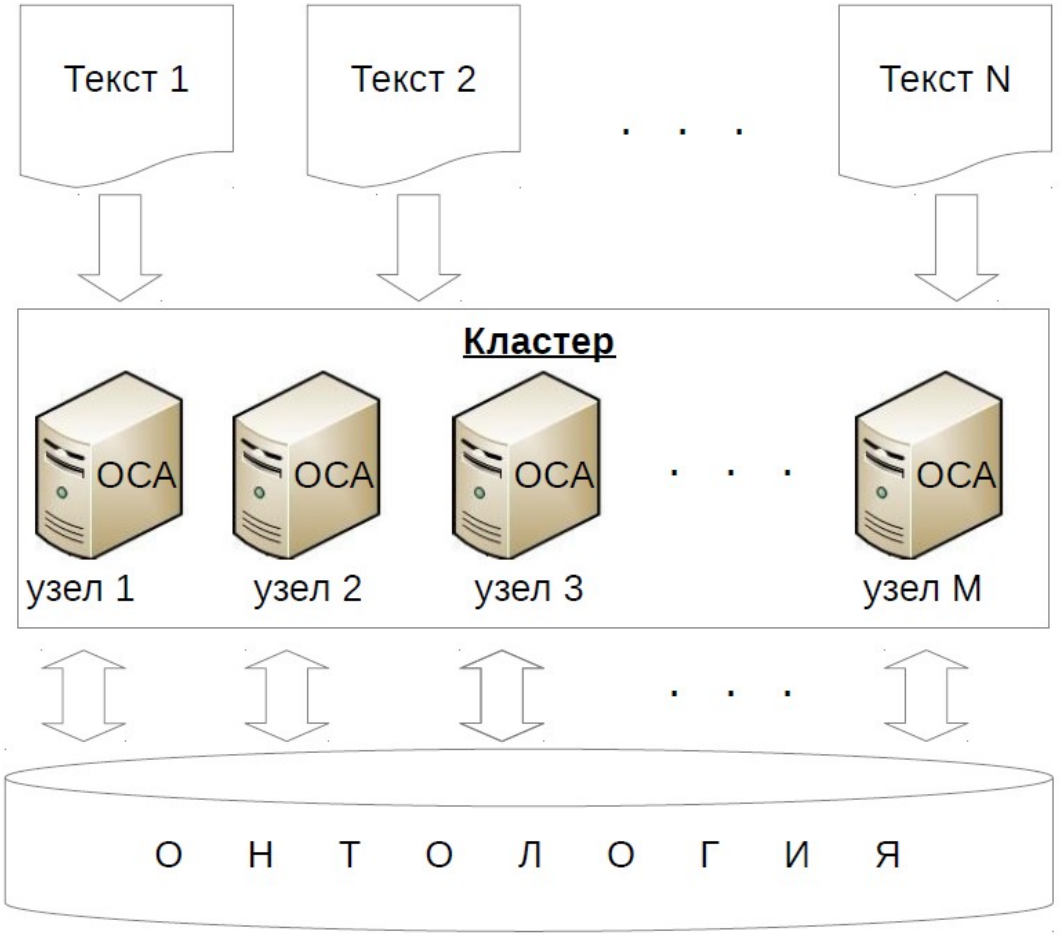
Пример работы ОСА

- *ЧЕГО(отличники, школы)*
- *ПРИЗНАК(спорта, яхтенного)*
- *МЕСТО(выехали, в лагерь)*
- *ЧЕГО(школы, спорта)*
- *ДЕЙСТВИЕ(отличники, выехали)*
- *ВРЕМЯ(выехали, вчера)*

Схема работы ОСА



Формирование онто-семантических графов



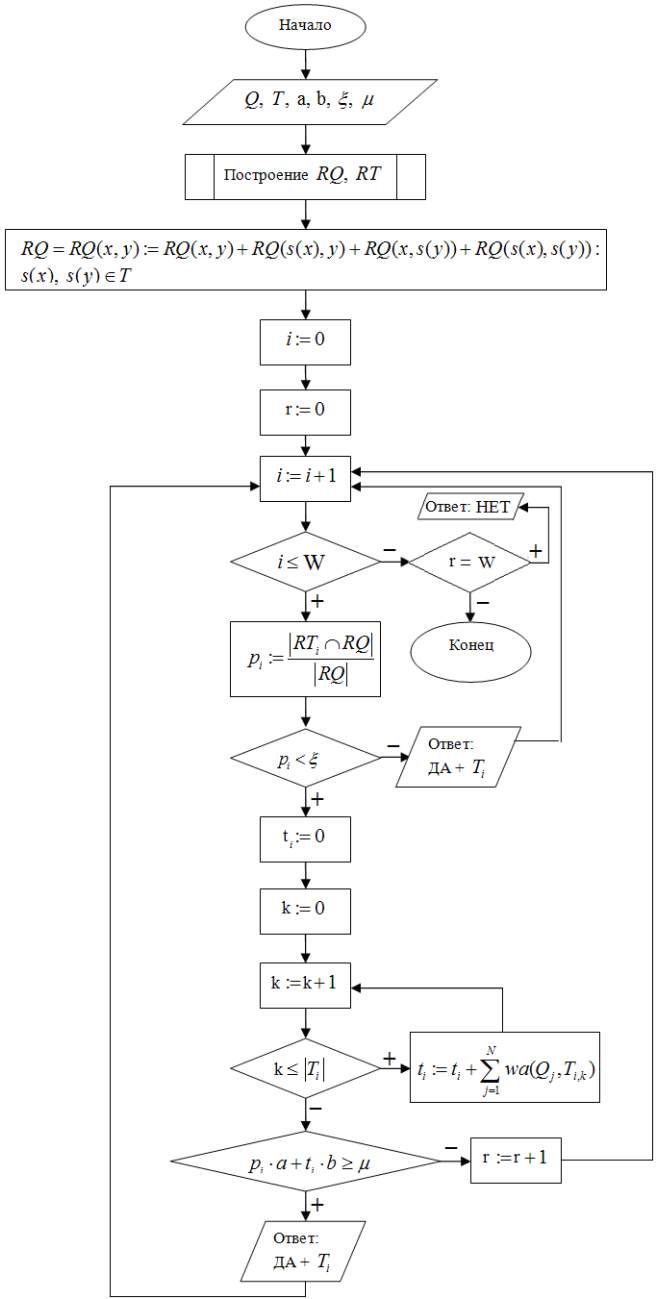
Тестовая конфигурация

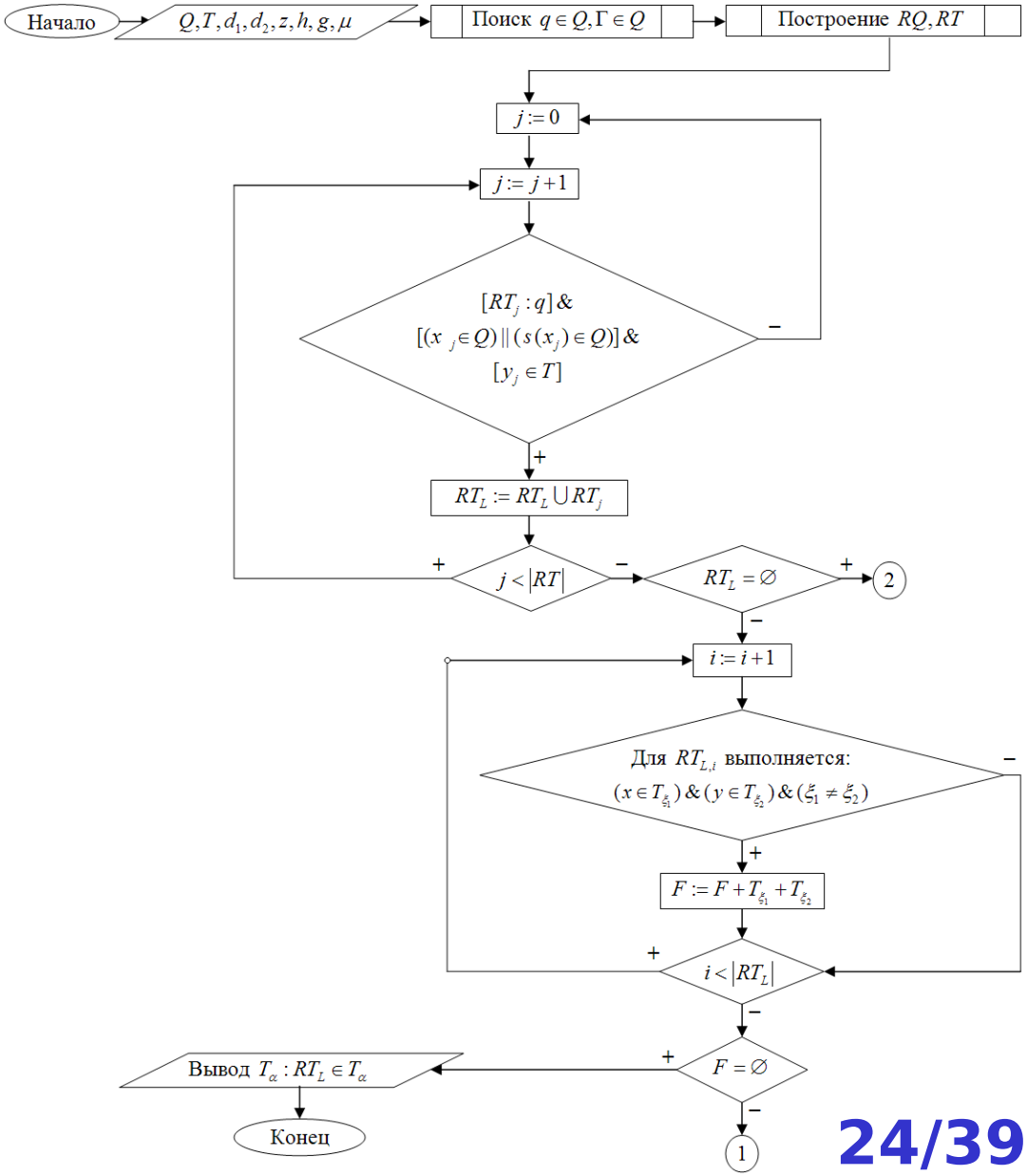
- Процессор Intel(R) Core(TM) i7 CPU 920 2.67 ГГц,
- 24 Гб оперативной памяти DDR3 1333 МГц,
- 1 Тб жесткий диск с 7200 оборотами в секунду,
- 64-х битная ОС Ubuntu 12.04.2 LTS.

Сравнение времени работы сем. анализаторов

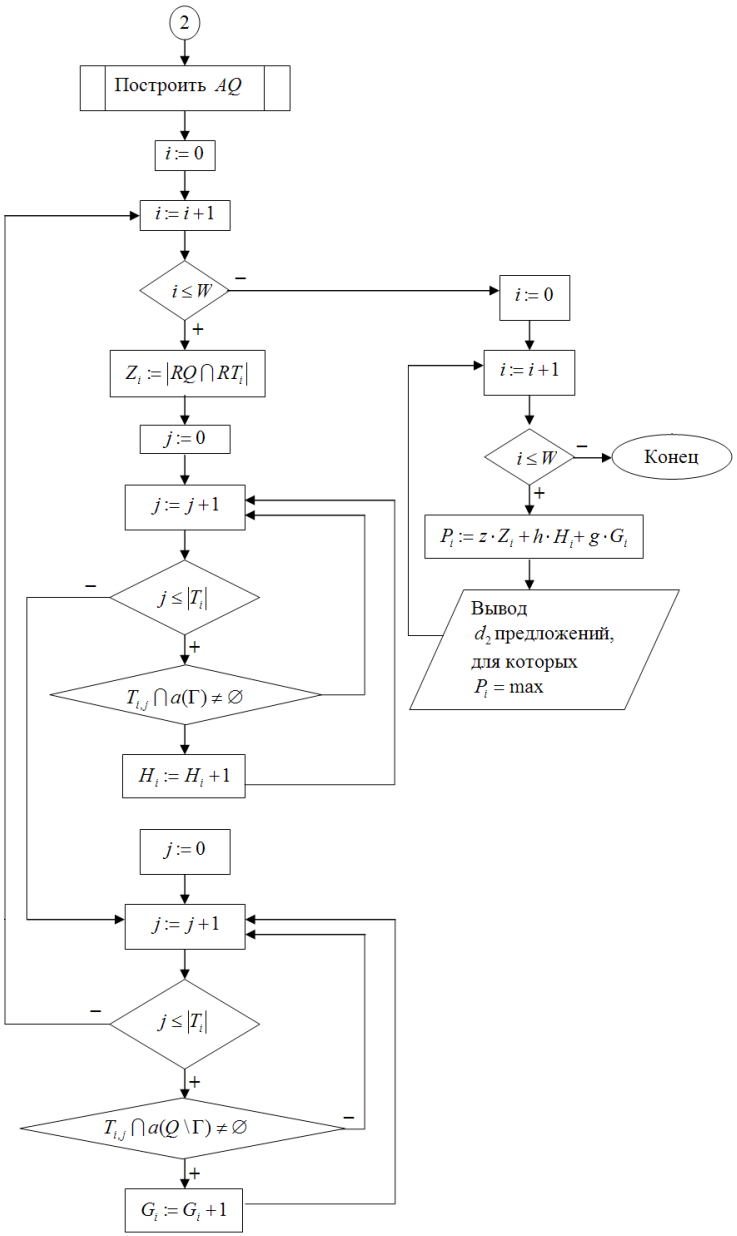
Кол-во предложений	**CA_BOC (*мс*)**	**АОТ (*мс*)**
1	32	147
10	55	1068
100	176	10284
500	740	51950
1000	1407	103188
5000	7323	515478
10000	14174	1030618
20000	31929	2061290

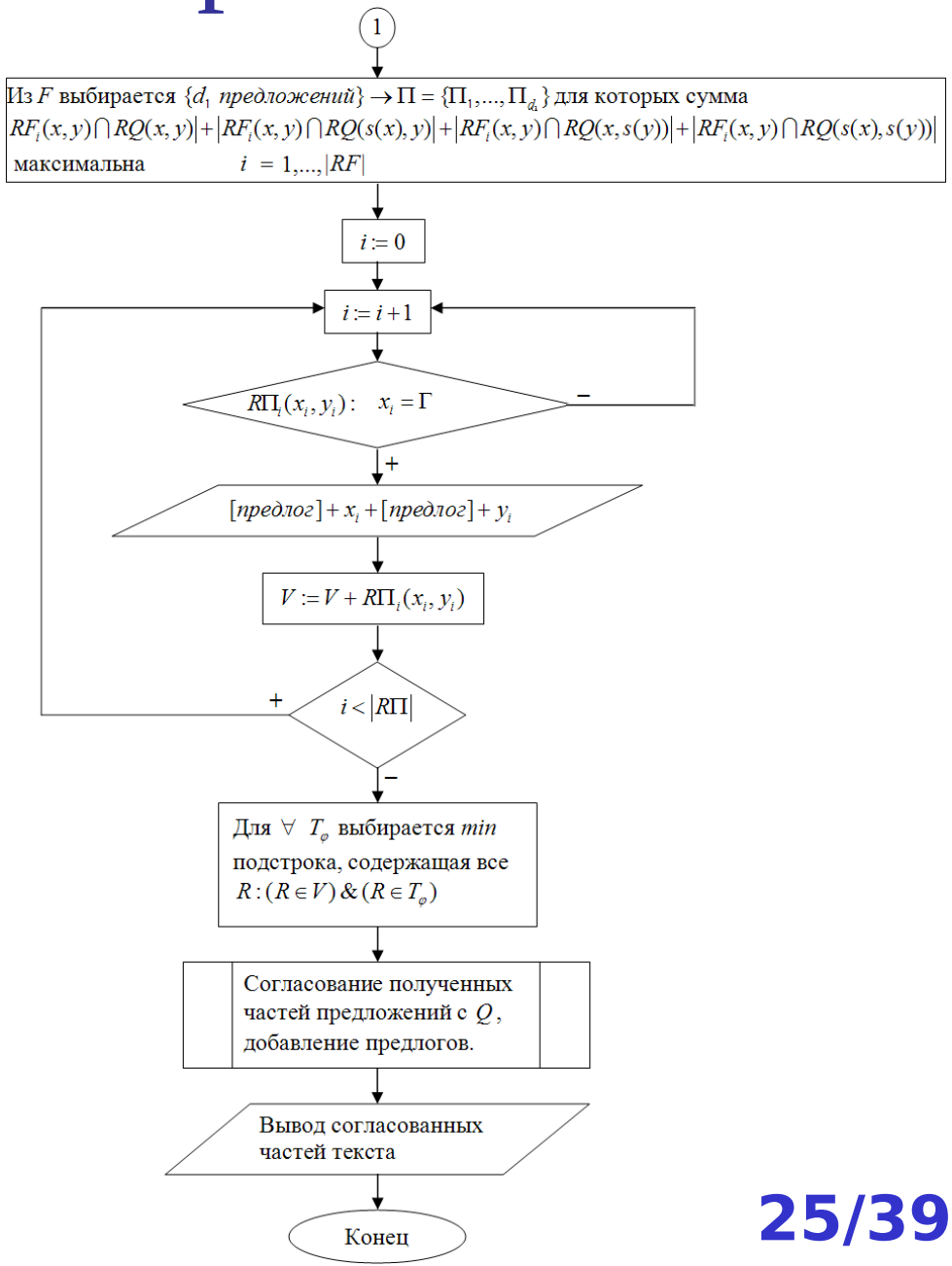
Блок-схема алгоритма ВОС





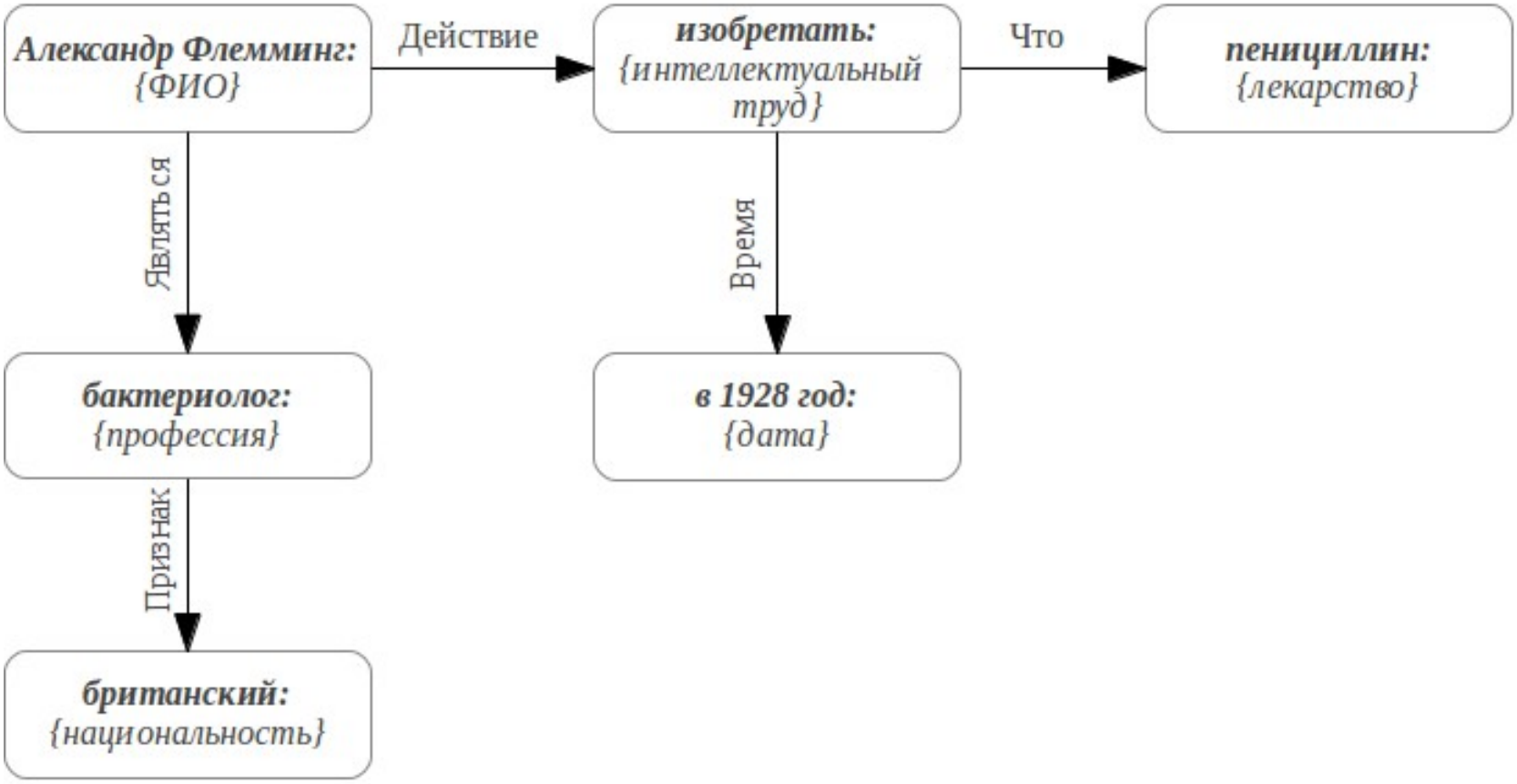
Блок-схема алгоритма ВОС



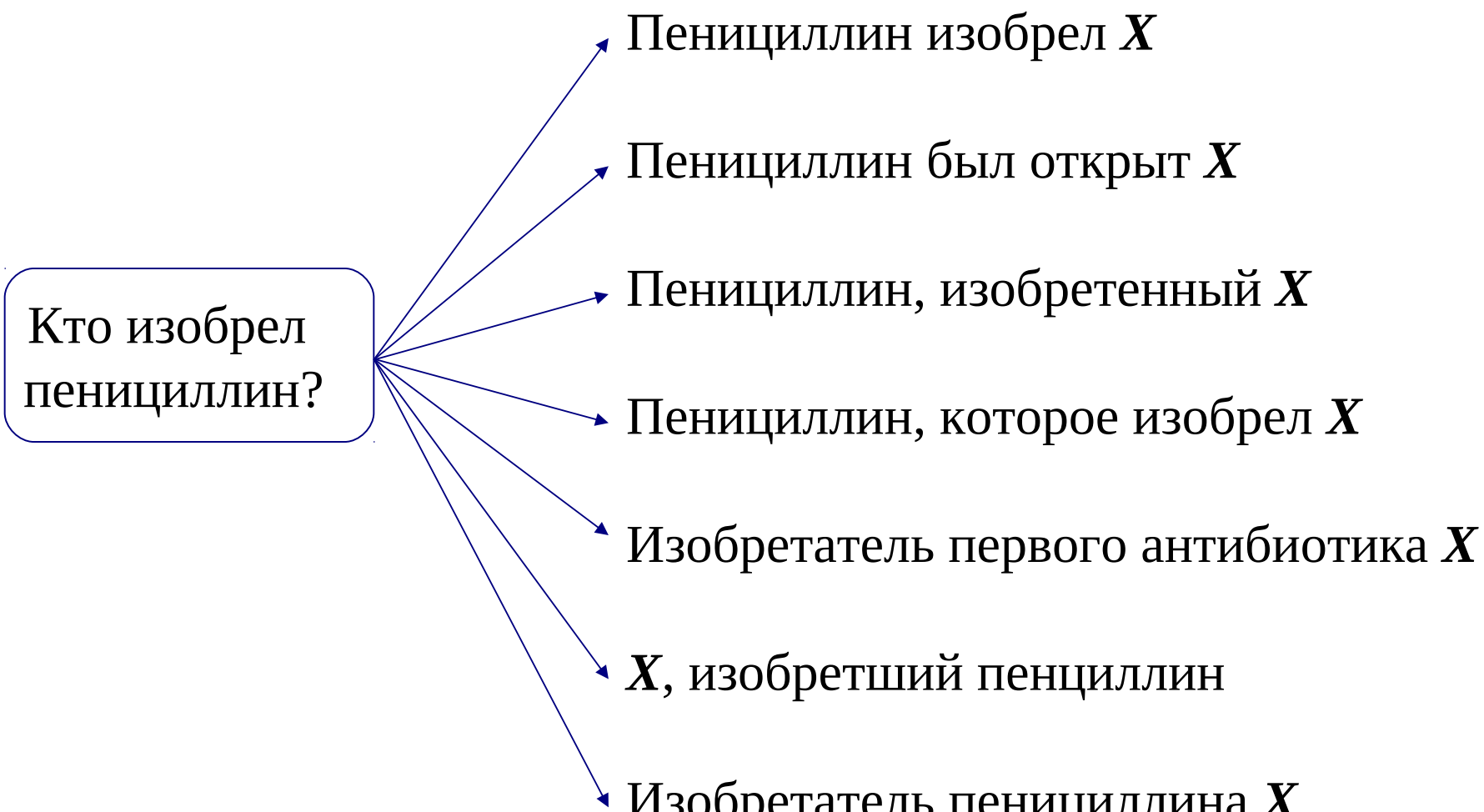


Онто-семантические графы

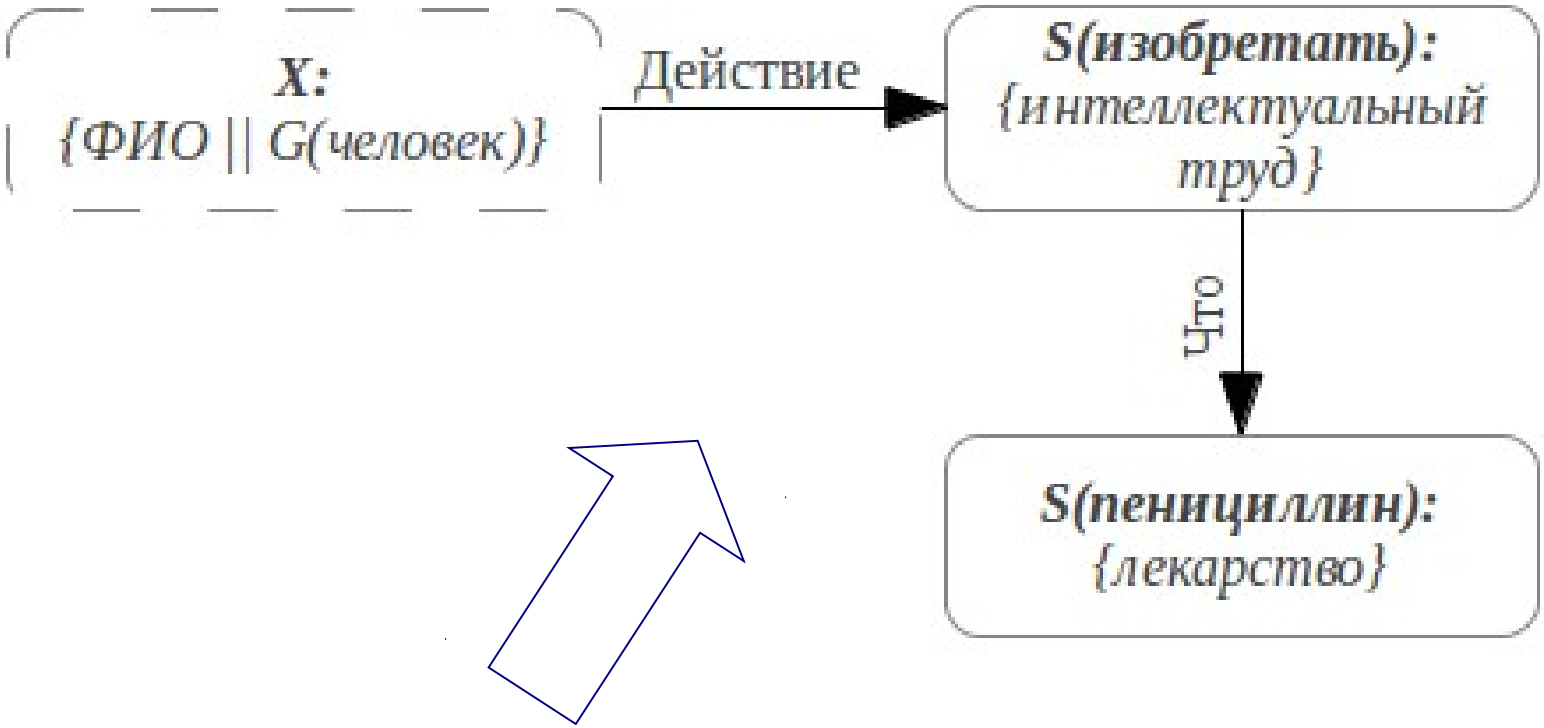
Британский бактериолог Александр Флемминг изобрел пенициллин в 1928 году



Предполагаемые ответы



Онто-семантический граф предполагаемого ответа

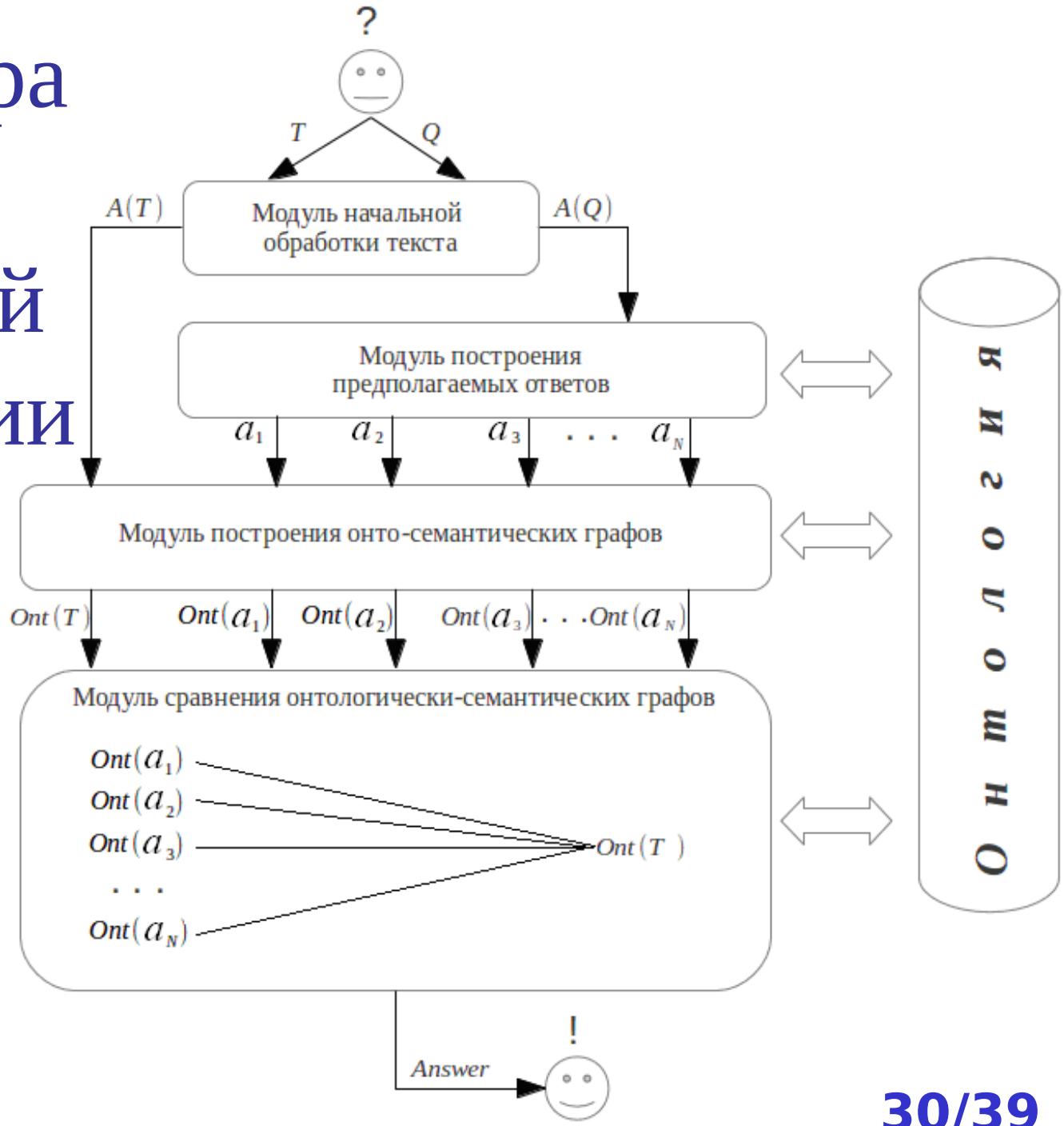


Кто изобрел пенициллин?

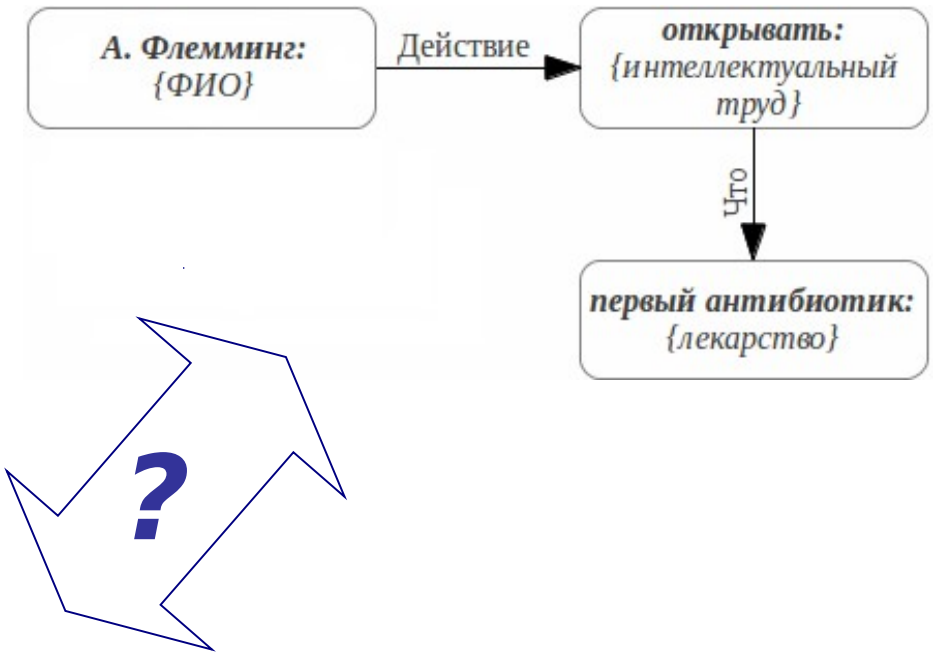
Формирование предполагаемых ответов

Предполагаемый ответ	**Пример ответа**
X был S(открыт) S(пенициллин) ***X***:{тв. падеж, ФИО || G(человек)}	В 1928 г. **Александром Флемингом** был изобретен пенициллин.
X S(открыл) S(пенициллин) ***X***:{им. падеж, ФИО || G(человек)}	**Британский бактериолог** открыл первый антибиотик в 1928 году.
X, S(открывшему) S(пенициллин), ***X***:{дт. падеж, ФИО || G(человек)}	**А. Флемину**, открывшему пенициллин, была присуждена Нобелевская премия.

Архитектура ВОС, основанной на сравнении ОСГ



Сравнение ОСГ



Вычисление коэффициента схожести двух ОСГ

Когда коэффициент схожести не равен нулю?

- $Ont(\mathrm{T}^*) : Ont(\mathrm{T}^*) \subseteq Ont(\mathrm{T})$ содержит онто-семантическое отношение «*Действие*», связывающие те же аргументы, начальная форма которых совпадает (или, для неопределенных вершин – не противоречит) с аргументами отношения «*Действие*» графа $Ont(a_i)$;
- В случае наличия в вопросе вопросительного слова, в $Ont(\mathrm{T}_j)$ присутствует семантическая зависимость, определенная по вопросительному предложению как главная (для вопроса с вопросительным словом «*где*» - отношение «*Место*», для вопросительного слова «*когда*» - «*Время*» и т. д.) и аргументы этих зависимостей либо совпадают, либо (для неопределенной вершины) не противоречат.

Примеры работы ВОС

Онтология: *животные → ленивцы*

Анализируемый текст: <…> Известно, что бурые медведи живут в тайге, а медлительные ленивцы обитают в джунглях. <…>
Вопрос: Какие животные живут в джунглях?
Развернутый ответ: Медлительные ленивцы обитают в джунглях.
Краткий ответ: ленивец.

Логические связи: *панды → они*

Анализируемый текст: <…> Панды являются практически вегетарианцами и питаются в основном бамбуком. За день они съедают до 30 кг бамбука и побегов.<…>
Вопрос: Сколько съедает панда в день?
Развернутый ответ: За день они съедают до 30 кг бамбука и побегов.
Краткий ответ: 30 кг

Примеры работы ВОС

Онтология: *вечер → время*

Анализируемый текст: <…>Вчера дочь генерала читала книгу. <…>
Вопрос: Когда дочь читала книгу?
Развернутый ответ: Вчера дочь генерала читала книгу.
Краткий ответ: вчера

Синонимы: *месяц → луна*

Анализируемый текст: <…> Над елями показался месяц. <…>
Вопрос: Где показалась луна?
Развернутый ответ: Над елями показался месяц.
Краткий ответ: над ель

Результаты работы

- Разработана математическая модель ОСА русскоязычного текста;
- Разработаны и программно реализованы:
  - Алгоритм работы ОСА;
  - Алгоритм работы ВОС, основанной на результатах работы ОСА;
- Разработана архитектура ВОС, основанной на сравнении онто-семантических графов.

Апробация работы

1. Международном конгрессе по интеллектуальным системам и информационным технологиям <<AIS-IT'12>>. Дивноморское, 2012 г.
2. Международном конгрессе по интеллектуальным системам и информационным технологиям <<AIS-IT'13>>. Дивноморское, 2013 г.
3. **IIV Всероссийский конгресс молодых ученых. Санкт-Петербург, ИТМО, 2014 г.**

4. Международная Поспеловская летняя школа-семинар для студентов, магистрантов и аспирантов <<Методы и технологии гибридного и синергетического искусственного интеллекта>>. Светлогорск, 2014 г.
5. Семинар Захарова В.П. <<Корпусная лингвистика>>, СпбГУ, 2014 г.
6. IV Всероссийский конгресс молодых ученых. Санкт-Петербург, ИТМО, 2015 г.
7. VI Международная конференция <<Системный анализ и информационные технологии>>. Светлогорск, 2015 г.
8. International Conference on Knowledge Engineering and Semantic Web, 30 сентября, 2 октября, Москва, 2015 г.
9. V Всероссийский симпозиум <<Инфраструктура научных информационных ресурсов и систем>>. Санкт-Петербург, 5 - 8 октября 2015 г.
10. AINL-ISMW FRUCT, Санкт-Петербург, 9-14 ноября 2015 г.

Публикации авторов

1. **Мочалова А.В., Мочалов В.А. Интеллектуальная вопросно-ответная система // Информационные технологии. 2011. №5. С. 6-12.**
2. **Мочалова А.В. Алгоритм семантического анализа текста, основанный на базовых семантических шаблонах с удалением. // Научно-технический вестник информационных технологий, механики и оптики. – 2014. - №5.**
3. **Мочалова А.В. Создание и пополнение терминологических систем с помощью семантического анализатора // Ученые записки Петрозаводского государственного университета. - 2015.**
4. **Kuznetsov V.A. , Mochalov V.A., Mochalova A.V. Ontological-semantic text analysis and the question answering system using data from ontology // ICACT Transactions on Advanced Communications Technology (TACT) Vol. 4, Issue 4, July 2015, pp.651-658.**
5. **Mikhailova V., Mochalova A., Mochalov V., Zakharov V. "Uncovering semantic relations conveyed by Russian prepositions", Proceedings, The IEEE 18th International Conference on Advanced Communication Technology, 2016, ICACT 2016, Phoenix Park, Korea (в печати).**
6. Mochalova A. Search for answers in ontological-semantic graph. Proceedings of the AINL-ISMW FRUCT, Saint-Petersburg, Russia, 9-14 November 2015, ITMO University, FRUCT Oy, Finland. P. 174-180.
7. Мочалова А.В. Некоторые вопросы работы русскоязычной вопросно-ответной системы, использующей данные из онтологии. Труды шестой международной конференции <<Системный анализ и информационные технологии>>, Светлогорск, 2015 г.
8. Мочалов В.А., Старкова А.В. Интеллектуальный агент-менеджер // Тр. конгресса по интеллектуальным системам и информационным технологиям "AIS-IT'09". Научное изд. в 4-х томах. М.: Физматлит. 2009. Т. 3. С. 261-268.
9. Мочалова А.В. Поиск семантических и логических связей в тексте // Труды международного конгресса по интеллектуальным системам и информационным технологиям «AIS-IT'12». Научное изд. В 4-х томах. М.:Изд-во Физматлит, 2012. С. 374-379.

10. Мочалова А.В. Автоматизация создания базы фактов с помощью семантического анализатора // Труды международного конгресса по интеллектуальным системам и информационным технологиям «AIS-IT’13». Научное изд. В 4-х томах. М.:Изд-во Физматлит, 2013. С. 352-359.
11. Кузнецов В.А., Мочалов В.А., Мочалова А.В. Распределенная программная реализация упрощенного онтологически-семантического анализатора // Сборник трудов V Всероссийский симпозиум <<Инфраструктура научных информационных ресурсов и систем>> (в печати).
12. Мочалова А.В. Вопросно-ответная система, основанная на сравнении семантических графов // Сборник трудов V Всероссийский симпозиум <<Инфраструктура научных информационных ресурсов и систем>> (в печати).
13. Мочалова А.В. Лингвистические переменные в вопросно-ответных системах // Труды 1-ой Международной Поспеловской летней школы-семинара для студентов, магистрантов и аспирантов "Методы и технологии гибридного и синергетического искусственного интеллекта"
14. Мочалова А.В. Проблемы создания интеллектуальных русскоязычных вопросно-ответных систем и их интеграция с онтологиями // Сборник тезисов докладов конгресса молодых ученых, Выпуск 3. - СПб: Университет ИТМО, 2014. С. 19-20.
15. Мочалова А.В. Архитектура и программная реализация вопросно-ответной системы, использующей данные из онтологии // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. - СПб: Университет ИТМО, 2015.
16. Мочалова А.В. Функции для работы с онтологией, интегрированной с вопросно-ответной системой // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. - СПб: Университет ИТМО, 2015.
17. Мочалова А.В. Свидетельство о государственной регистрации программы для ЭВМ <<Программа семантического анализа текста, основанная на базовых семантических шаблонах с удалением>>, №. 2015613430, Россия. 28.01.2015.
18. Мочалова А.В. Свидетельство о государсвенной регистрации программы для ЭВМ <<Экспертная система для поиска семантических отношений в русскоязычном тексте с помощью базовых семантических правил с удалением>> (поданы документы)

Поддержка научными фондами

Частично работа выполнена при финансовой поддержке **РГНФ** в рамках научного проекта **№15-04-12029** "*Программная разработка электронного ресурса с онлайн-версией русскоязычной вопросно-ответной системы*".

Спасибо

за

внимание!