



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Естественно-языковое управление программным обеспечением на основе теории К-представлений на примере управления файловой системой

Александр Александрович Разоренов

Научный руководитель: профессор, д.т.н. В.А. Фомичев

НИУ «Высшая школа экономики» (ВШЭ), Школа бизнес-информатики,
Каф. инноваций и бизнеса в сфере информационных технологий.

105187, Москва, Кирпичная улица, 33

e-mail: arazorenov@hse.ru

В течение последних 7 лет в мире наблюдается повышенный интерес исследователей к автоматической семантической обработке естественно-языковых предписаний пользователя. Основные приложения: естественно-языковые интерфейсы (ЕЯ-интерфейсы) для взаимодействия с роботами, виртуальными персонажами и для управления программным обеспечением (Artzi, Zettlemoyer 2013; Bastianelli et al. 2013; Babes-Vroman et al. 2012; Chong, Pucella, 2003; Misra et al. 2015; She et al. 2014 и др.)

Ещё в 1980-х было проведено исследование, сравнивающее производительность пользователей при работе с обычным и ЕЯ-интерфейсами с системой Lotus 1-2-3. Оно показало более высокую производительность пользователей при использовании ЕЯ-интерфейса по сравнению со стандартным GUI.

Проведённое позднее аналогичное исследование (Kaufmann, Bernstein 2007) подтвердило этот результат.

- Siri
- *Собеседник HD*
- *Anboto*
- *Braina Project*
- *NLUI Server*

При построении семантически-ориентированных ЕЯ-интерфейсов одной из центральных является задача преобразования вводимого пользователем текста на естественном языке (русском, английском и др.) в семантическое представление (СП) для дальнейшей обработки в рамках прикладной программной системы, для которой создается ЕЯ-интерфейс.

Анализ литературы показывает, что для формирования СП предписаний главным образом используются:

1. Логика предикатов первого порядка
2. Представление смысла средствами λ -исчислений (Lambda-Calculus Meaning Representation)
3. Абстрактное представление смысла (Abstract Meaning Representation)
4. Теория К-представлений (Theory of K-representations)

Текущее состояние предметной области

Критерий	Логика первого порядка	Abstract Meaning Representation	Lambda-Calculus Meaning Representation	Теория К-представлений
Поддержка слов «некоторый», «все», «несколько» и т.п.	-	-	-	+
Функциональные выражения	+	-	+	+
Отношения	+	+	+	+
Маркировка частей семантического представления переменными	-	+	-	+
Логические связи	+	в расширениях	+	+
Кванторы $\forall$ и $\exists$	+	неявный $\exists$	+	+
Сложные понятия	-	ограниченно	-	+
Упорядоченные наборы	-	-	-	+
СП утверждений	+	+	+	+
СП вопросов	-	$\pm$	-	+
СП предписаний	-	-	-	+
Адаптированность к различным языкам	+	-	+	+
Различие между единичными и множественными объектами	$\pm$	-	$\pm$	+

В рамках теории К-представлений определяется понятие концептуального базиса и на его основе с помощью специальных правил построения строится два языка:

- Язык типов сущностей предметной области
- Язык семантических представлений (К-представлений)

Язык типов сущностей предметной области предназначен для контроля смысловой корректности строящихся К-представлений. Например,

- правильное К-представление:

файл*(Наименование, "a.txt")

- неправильное К-представление:

~~диск*(Размер, папка*(Наименование, "Документы"))~~

Для построения семантических представлений ТКП определяет 10 частичных операций: $Op[1], \dots, Op[10]$

Полученный таким образом язык семантических представлений (К-представлений) позволяет отражать семантическую структуру произвольно сложных предложений и связных текстов (дискурсов)

Например, команде пользователя

Move file "a.txt" to folder "Desktop" if size of file "a.txt" less than "2000"

может соответствовать K-представление

$\text{If1}(\text{Less1}(\text{Size1}(\text{certain file1} * (\text{Name}, "a.txt")), "2000"),$
 $\text{movement1} * (\text{Destination}, \text{certain folder1} * (\text{Name}, "Desktop"))$
 $(\text{Source}, \text{certain file1} * (\text{Name}, "a.txt"))))$

Для семантического парсинга (семантико-синтаксического анализа) предписаний пользователя на естественном языке используются следующие подходы:

- *Сопоставление с образцами команд или регулярные выражения*
- *Контекстно-свободные грамматики*
- Комбинаторные категориальные грамматики (CCG)
- Семейство алгоритмов семантического парсинга, предоставляемых ТКП: SemSyn, SemSynt1, SemSynt2

В рамках проводимого диссертационного исследования были предложены:

- **Новая формализация лингвистической базы данных**
- **Новая формализация понятия «семантико-синтаксической структуры текста»**
- **Новый алгоритм семантико-синтаксического парсинга входного текста**
- **Программный комплекс для управления файловой системой**

ЛБД является упорядоченным набором (В, MorphBs, LexFrames, TermDB), где:

- **В – Концептуальный базис** – определяется в рамках теории К-представлений
- **MorphBs – Морфологический базис** – определяет множество допустимых морфологических признаков лексических единиц, отношение частичной упорядоченности на этом множестве, а также ряд отображений
- **LexFrames – Словарь лексических фреймов** – определяет возможные взаимосвязи между лексическими единицами текста и соответствующие этим связям шаблоны СП
- **TermDB – Терминологическая база данных** – определяет понятия через ограниченный набор базовых понятий предметной области

Морфологический базис определяет множество допустимых сочетаний морфологических признаков, например:

{сущ, ед.ч.}

{сущ, ед.ч., предл.пад.}

и отношение частичной упорядоченности на нём:

{сущ, ед.ч.} $\leq$: {сущ, ед.ч., предл.пад.}

Кроме того, морфологический базис определяет отображения:

- лексем в ассоциированные наборы морфологических признаков
- лексем в базовые словоформы
- лексем в элементы первичного информационного универсума

Словарь лексических фреймов состоит из упорядоченных наборов вида

- (type1,morph1,type2,morph2,prep,template,direction)

где:

- type1, type2 – типы К-представлений, соответствующих главной и подчинённой лексеме
- morph1, morph2 – морфологические признаки, соответствующие главной и подчинённой лексеме
- prep – предлог, стоящий перед подчинённой лексемой, возможно пустой
- template – шаблон К-представления
- direction – порядок следования лексем (если важен)

Словарь лексических фреймов и морфологический базис применяются при построении *графа* возможных *семантико-синтаксических связей* входного текста.

Данный граф является:

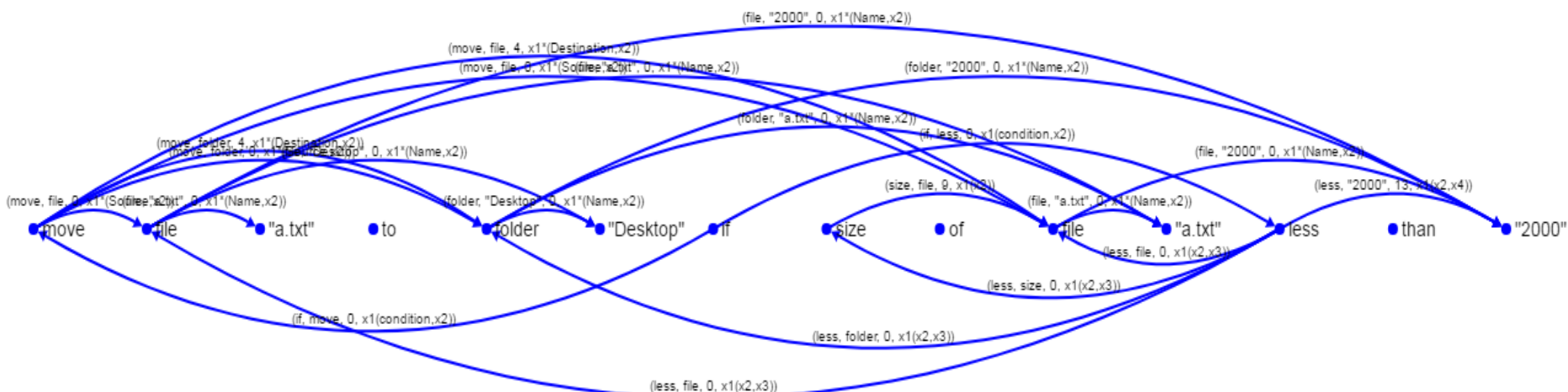
- ориентированным
- ациклическим
- размеченным

Разметка:

- *вершины* – порядковые номера лексем входного текста
- *дуги* – шаблоны К-представлений + дополнительная информация

Например, для входного текста:

Move file "a.txt" to folder "Desktop" if size of file "a.txt" less than "2000"



- Разбиение входного текста на отдельные лексемы
- Построение графа семантико-синтаксических связей, используя лингвистическую базу данных
- Поиск и фильтрация остовных деревьев графа
- Построение К-представлений для каждого остовного дерева

Переместить файл "a.txt" в папку "Документы"

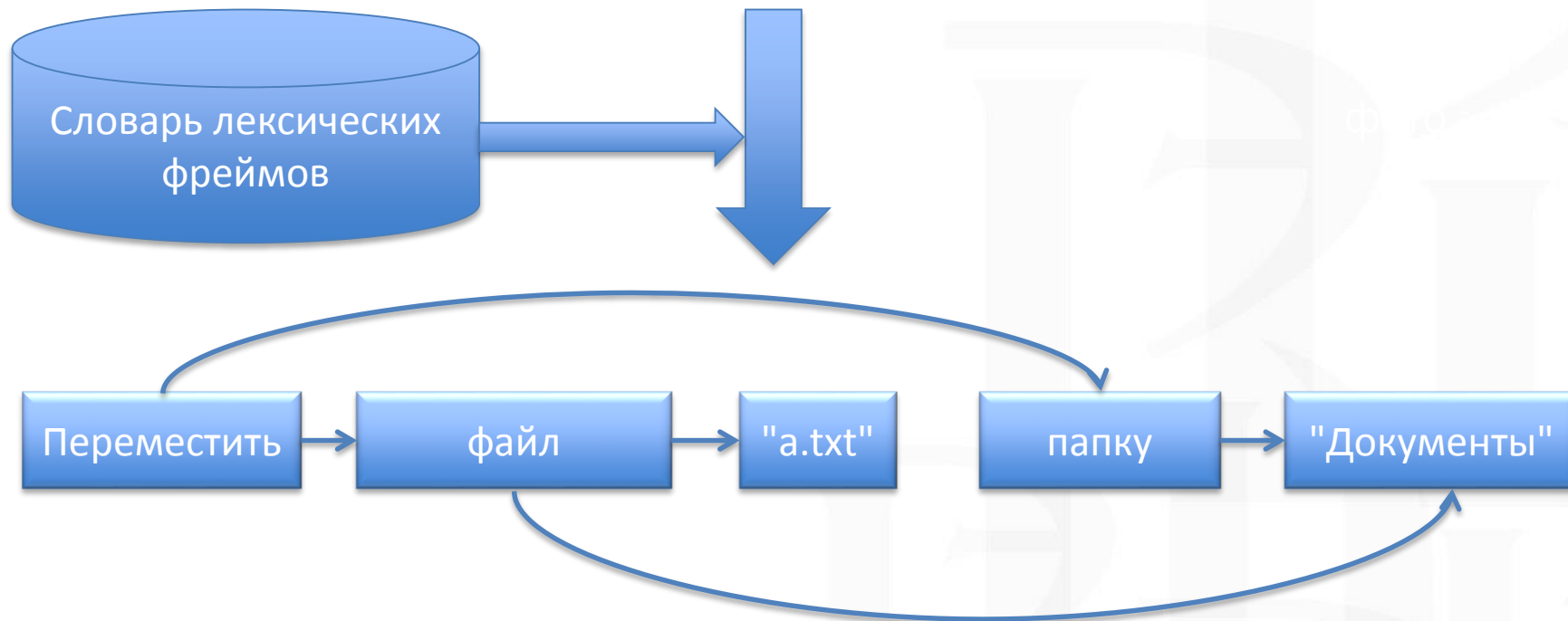


1. Переместить
2. файл
3. "a.txt"
4. в
5. папку
6. "Документы"

Этап зависит от входного языка:

- для русского языка: простое разбиение на отдельные слова за исключением текста в кавычках и имён собственных
- для английского есть ряд ситуаций, когда необходимо учитывать группы слов, так как от этого могут зависеть морфологические признаки.

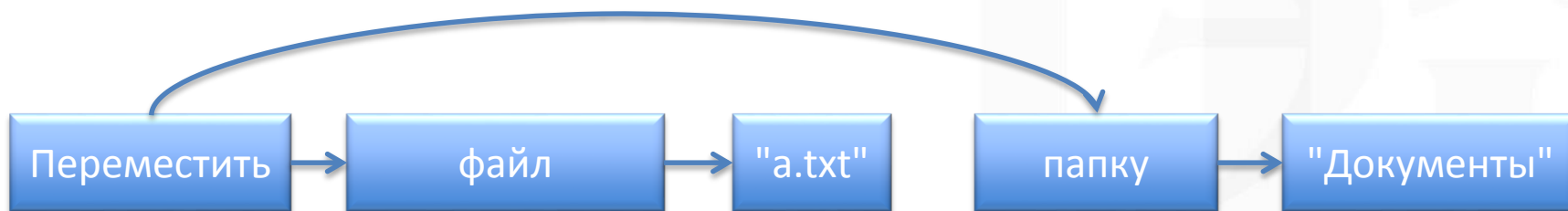
Переместить(1) файл(2) "a.txt"(3) в(4) папку(5) "Документы"(6)



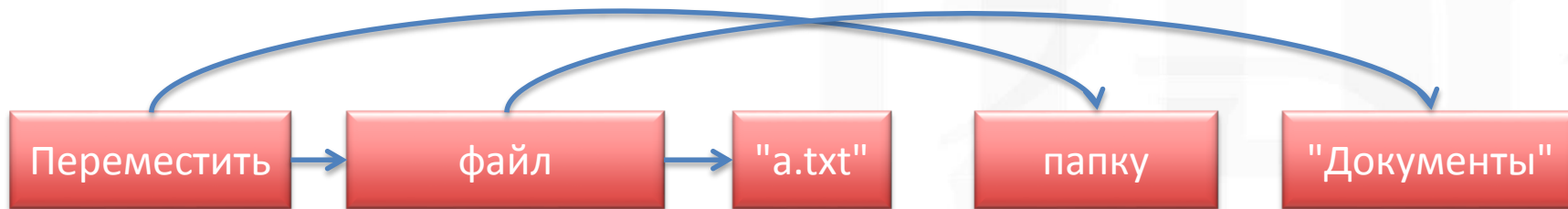
Граф ориентированный, ациклический, размеченный – каждой дуге соответствует шаблон соответствующего К-представления

Переместить(1) файл(2) "a.txt"(3) в(4) папку(5) "Документы"(6)

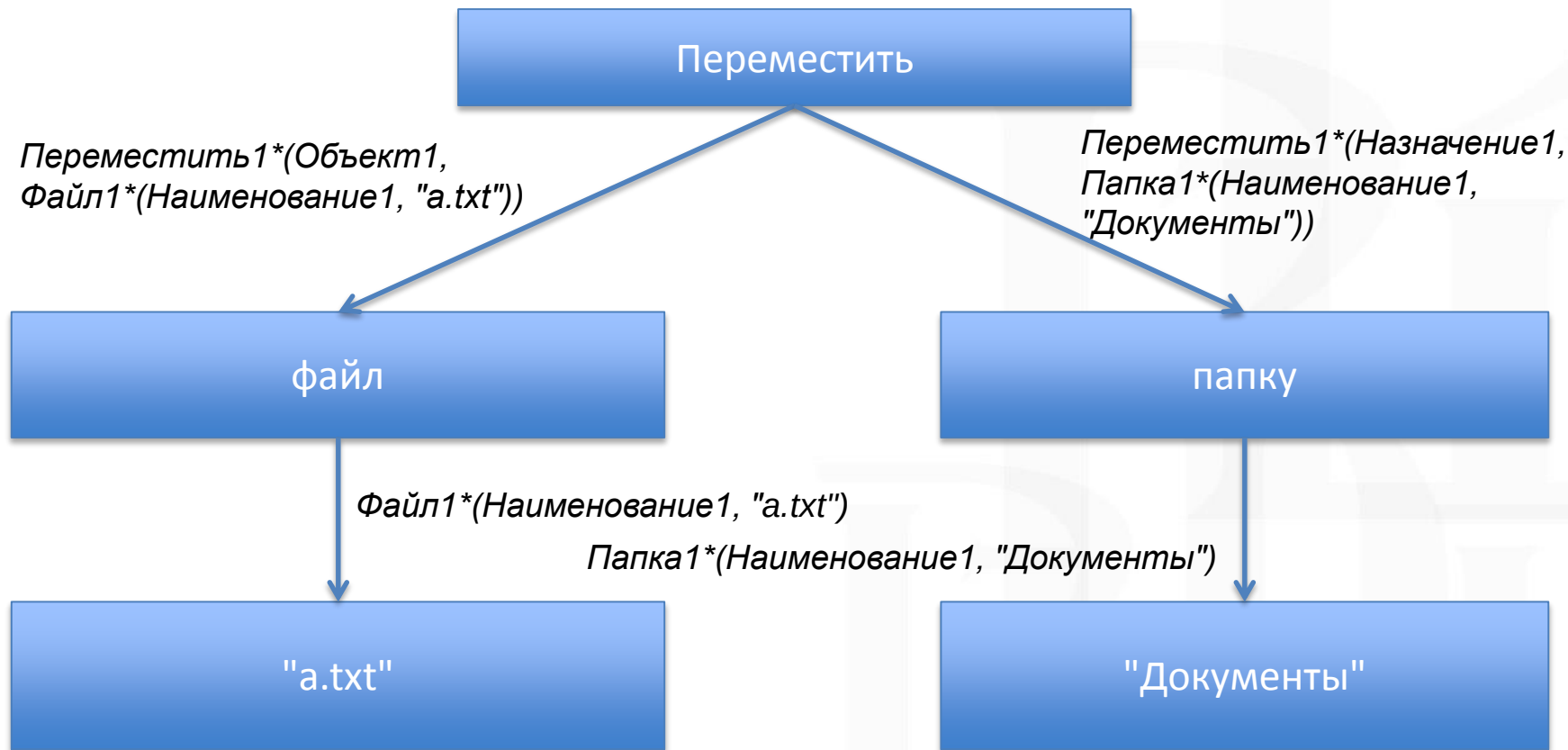
Правильное остовное дерево



Неправильное остовное дерево



Переместить(1) файл(2) "a.txt"(3) в(4) папку(5) "Документы"(6)



Итоговое представление: *Переместить1*(Объект1, Файл1*(Наименование1, "a.txt"))(Назначение1, Папка1*(Наименование1, "Документы"))*

Алгоритм может обрабатывать:

- простые предписания:
 - удали файл «a.txt»,
 - переименуй папку «Документы» в «проект»
 - и т.п.
- последовательность действий:
 - Заархивируй и отправь документ «Выводы.doc»
- предписания с условиями:
 - Если размер свободного пространства на флешке больше размера файлов в папке «Музыка», скопируй папку «Музыка» на флешку.

Преобразование ЕЯ-текстов в семантические представления

Критерий	КСГ	CCG	SemSynt	SemSynt RA
Сложность реализации	Средняя	Высокая	Высокая	Низкая
Объём требуемой БД	Высокий	Высокий	Высокая	Средний
Сложность заполнения БД	Средняя	Высокий	Средняя	Средний
Вычислительная сложность	$O(m^2n)$	«Наивная» реализация: $O(2^n)$ «Оптимизированный» алгоритм: $O(2^d n^4)$	-	$A(n)+O(n^2m)+B(n)+O(dn)$
Учёт омонимий (разных смыслов одного слова)	Нет	Да	Да	Да

Выполнение предписаний пользователей выполняется в два этапа:

- СП преобразуется в унифицированную форму, используя терминологическую базу данных ЛБД TermDB
- унифицированное СП преобразуется в команду целевой системы, используя базу знаний команд

Терминологическая база данных представляет из себя множество пар (mask, result), где и mask и result – К-представления, первое является образцом, а второе заменяющей подстановкой

(музыка, нек файл1(Расширение, "mp3"))*

База знаний команд является множеством (mask, cmd), где mask – образец К-представления, а cmd – шаблон команды

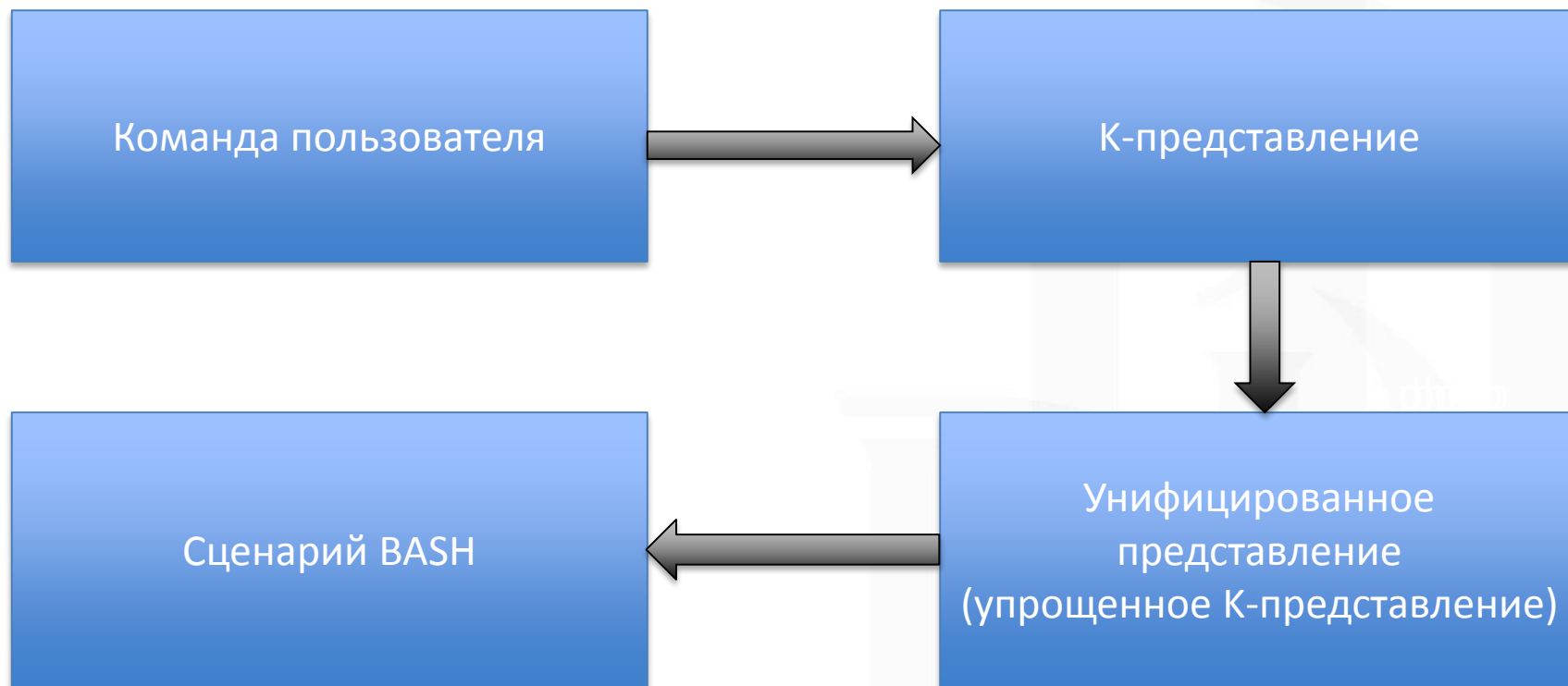
(удаление(Объект, x1), "rm -f {x1}")*

Для проверки на соответствие обрабатываемого К-представления или его части образцу было предложено отношение соответствия образцу с учётом особенностей К-представлений и исследованы его свойства, в том числе свойства позволяющие проводить оптимизацию соответствующих алгоритмов – индексацию терминологической базы данных и базы знаний команд по образцам.

Для разработки и исследования алгоритмов и методов построения ЕЯ-интерфейсов было создано приложение для управления файловой системой – Natural Language Commander (NLC): первая версия в 2014-15 гг., вторая версия в 2016 году.

Для формализованного представления естественно-языковых предписаний пользователя компьютерная система NLC использует аппарат СК-языков, а также разработанные модель лингвистической БД и алгоритм семантико-синтаксического анализа предписаний.

Natural Language Commander



Пример работы Web-версии Natural Language Commander

Command:

move files "a.txt" to folder "Desktop" if size of file "a.txt" less than "2000"



Basic representation:

if(less(size(file*(Name,"a.txt")), "2000"), move*(Destination, folder*(Name, "Desktop"))(Source, file*(Name, "a.txt")))

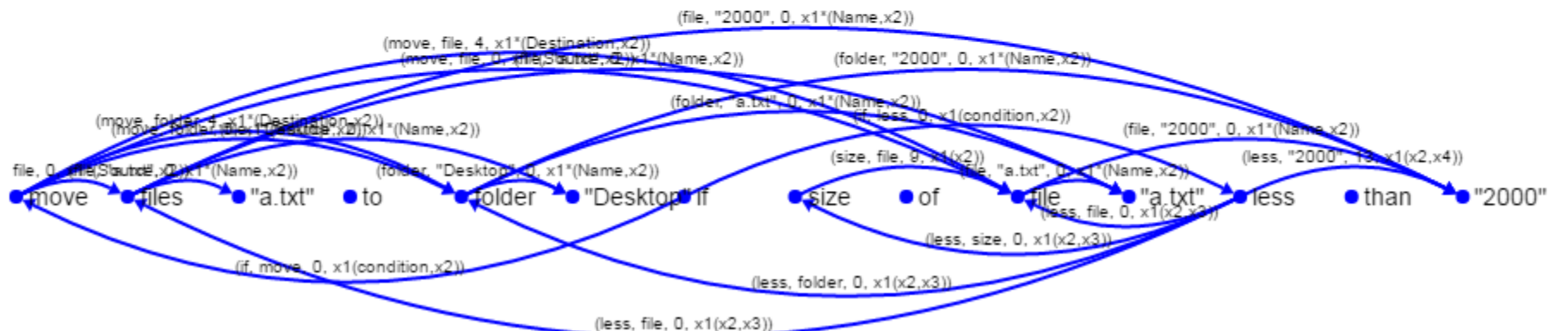
Press to start processing

Final representation:

if(less(size(file*(Name,"a.txt")), "2000"), move*(Destination, folder*(Name, "Desktop"))(Source, file*(Name, "a.txt")))

Result script:

if [\$(du -s "a.txt" | awk '{print \$1}') -lt "2000"]; then mv "a.txt" "Desktop"/; fi



Данная программа использована в двух курсах ВШЭ на английском языке:

- Semantic Information Systems, Бакалавриат, 4 курс, программа «Бизнес-информатика»
- Natural Language Processing, Магистратура, 1 курс, программа «Big Data Systems»

а также в курсе НИУ МАИ:

- Проектирование лингвистических процессоров, Магистратура, 1 курс

1. Предложенная математическая модель лингвистической базы данных является более общей по сравнению с моделями ЛБД, рассматриваемыми в теории К-представлений. В новой модели уменьшено количество используемых сущностей, в частности, за счёт унификации словарей глагольно-предложных и предложных фреймов.
2. Был изложен новый подход к формальному описанию семантико-синтаксической структуры как промежуточной структуры при формировании К-представления исходного текста.
3. Описание предложенного алгоритма является весьма компактным по сравнению с другими алгоритмами, представленными в научной литературе.

4. Ценной чертой предложенной парадигмы является возможность ее использования не только для английского языка, но и для русского и многих других естественных языков
5. По сравнению с ранее предложенными алгоритмами SemSyn, SemSynt1, SemSynt2 был расширен входной язык: поддержана обработка союза «или»; расширена поддержка обработки различных частей речи и взаимосвязей между словами, принадлежащими этим частям речи, а также обработки сочетаний «модальное слово + инфинитив».



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Спасибо за внимание!

101000, Россия, Москва, Мясницкая ул., д. 20

Тел.: (495) 621-7983, факс: (495) 628-7931

www.hse.ru