

Н. А. Власова

## К проблеме разметки текстов на русском языке для задачи извлечения фактографической информации

Аннотация. В настоящей работе рассматривается современное состояние проблемы извлечения фактографической информации из текстов на русском языке как подзадачи в направлении Information Extraction. Проанализирован опыт разметок текстов для задачи извлечения информации о событиях в рамках проектов MUC и ACE. Обосновывается и определяется понятие модели текстового упоминания о событии, описывается его структура. Текстовое упоминание о событии представляет собой синтаксически связный фрагмент текста, обозначающий событие заданного типа. Этому фрагменту ставится в соответствие фрейм события, слоты которого заполнены информацией, которая может быть получена при анализе данного фрагмента. Предполагается, что выделение в тексте синтаксически связных фрагментов, которым сопоставлен фрейм со структурированной информацией, извлеченной из данного фрагмента, будет удобным промежуточным уровнем для работы со всем многообразием способов выражения информации о событиях в русскоязычных текстах. На примере анализа текстовых упоминаний событий назначения и отставки рассматриваются сложности извлечения фактографической информации из текстов на русском языке.

*Ключевые слова и фразы:* автоматическое извлечение информации, фактографическая информация, текстовые коллекции, разметка.

### 1. Извлечение из текста информации о событии

Извлечение из текстов фактографической информации — одно из направлений общей задачи извлечения информации (Information Extraction) [1]. Одной из основополагающих особенностей Information Extraction является то, что эта задача направлена на анализ текста

---

Работа выполнена в рамках НИР «Разрешение синтактико-семантической неоднозначности в рамках задачи извлечения информации из текстов, основанное на использовании кроссмодального контекста» (№ гос. регистрации 01201354592).

© Н. А. Власова, 2014

© ИНСТИТУТ ПРОГРАММНЫХ СИСТЕМ ИМЕНИ А. К. АЙЛАМАЗЯНА РАН, 2014

© ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ, 2014

с целью выявления в нем информации определенного вида и представления ее структурированном образом, удобным для последующей обработки. Извлечение из текстов фактографической информации (fact extraction, event extraction) — это извлечение информации о событиях определенных семантических типов из больших объемов неструктурированного текста. Такими событиями могут быть: покупка и продажа акций или предприятий, увольнение и назначение, рождение и смерть и т.д. Результатом извлечения обычно является заполненный фрейм события. Например, в случае события назначения, это информация о том, что имеет место назначение, кто назначает, кого назначает и на какую должность. В текстах же информация о событиях может быть представлена по-разному. Классический «языковой образ» события — это глагольный предикат и его актанты. Например,

- (1) *Президент России Владимир Путин назначил Александру Левицкую советником главы государства.*

Извлеченная информация о событии представлена в фрейме:

Событие: *назначение*

Кто назначил Владимир Путин (Президент России)

Кого назначил Александру Левицкую

На какую должность советником (главы государства)

Однако информация об этом назначении может быть представлена в текстах совсем по-другому. Например,

- (2) *«Владимир Путин подписал Указ о назначении Александры Левицкой советником Президента Российской Федерации», — говорится в сообщении пресс-службы Кремля.*

- (3) *Первый заместитель главы аппарата правительства РФ Александра Левицкая назначена советником президента России.*

- (4) *Владимир Путин назначил нового советника президента. Им стала первый замглавы аппарата правительства Александра Левицкая.*

- (5) *Как следует из указа президента под номером № 666, подписанного 2 августа 2013 года, Александра Левицкая окончательно покидает правительство и уходит в администрацию президента: она получила пост советника главы государства.*

- (6) *Путин взял к себе в советники Александру Левицкую.*
- (7) *Александра Левицкая стала новым советником президента.*
- (8) *Однако, очевидно, что за «лояльность Кремлю» Путин наградил Левицкую должностью в Администрации президента.*
- (9) *Что же стоит за назначением Левицкой?*
- (10) *Новый советник главы государства Александра Левицкая была назначена на эту должность только вчера.*

Можно с уверенностью утверждать, что во всех приведенных выше примерах содержится информация об одном и том же событии назначения (Путин назначил Левицкую советником), и читающий текст человек, конечно, вычленил эту информацию. Однако очевидно, что представлена информация об этом событии в примерах (2-10) не так, как в примере (1). В одних случаях в предложении упомянуты не все участники события (3, 7, 9, 10). Иногда полная информация содержится не в одном предложении, а в нескольких, когда участник события при предикате, который обозначает событие, выражен местоимением, а полноценное упоминание об участнике события содержится в другом предложении (4). Иногда ситуация обозначена предикатом, который обычно не называет событие данного семантического типа (6, 8). Есть случаи, когда синтаксические группы, которые описывают участников события, не зависят непосредственно от предиката, маркирующего событие в тексте (5).

Задача извлечения фактографической информации — получить заполненный семантический фрейм события из языкового выражения независимо от того, каким образом информация о событии представлена в тексте. Такой фрейм можно было бы получить, обладая полным синтаксическим и семантическим разбором текста документа. Однако методы обработки текстов для решения задачи извлечения информации (Information Extraction) обычно ориентированы на быстродействующие алгоритмы, опирающиеся на частичный синтаксический и семантический анализ релевантных для поставленной задачи фрагментов текста. Поэтому крайне важно понять, на какие именно языковые структуры при частичном анализе текста удобнее всего опираться при автоматическом извлечении информации о событиях.

## 2. Подходы к извлечению фактографической информации, MUC и ACE

В рамках работ конференции MUC (Message Understanding Conference) был разработан поэтапный подход к извлечению информации о событиях [2, 3]. Задача по извлечению разбивается на подзадачи. На первом этапе из текста извлекается информация об именованных сущностях (Named Entity Recognition), устанавливается тождество названных в тексте по-разному объектов (coreference resolution) и строятся отношения между ними (template relation construction). Далее с использованием заранее разработанного шаблона заполняются слоты фрейма, который содержит информацию о событии (scenario template production). Необходимо отметить, что слоты фрейма должны быть заполнены на основе анализа всего разбираемого текста. Например, при заполнении слотов для фрейма «назначение» при анализе текста из примера 4 должны быть проанализированы оба предложения, установлено тождество объектов, описанных синтаксическими группами «нового советника президента» и «им», а также тождество событий, описываемых в первом и втором предложении. Эта процедура называется слияние (merging). Таким образом, фрейм события по окончании работы алгоритма извлечения информации должен выглядеть так:

Событие: *назначение*

Кто назначил	<u>Владимир Путин</u>
Кого назначил	<u>Александра Левицкая</u>
На какую должность	<u>советника</u>

Планка качества извлечения информации о событиях, заданная стандартами MUC, необычайно высока. Неудивительно, что результаты тестирования систем извлечения фактографической информации на MUC были достаточно низкими. На последней 7-ой конференции F-мера извлечения информации о событиях не превышала 51% [4]. Это связано со сложностями как в разметке, так и в разработке эффективных методов решения подзадач извлечения информации, сформулированных на MUC.

Преемником MUC в разработке стандартов для извлечения информации из текстов стал проект ACE (Automatic Content Extraction) [5, 6]. В рамках данного проекта были определены принципы извлечения информации об именованных сущностях разных типов (entities), отношений (relations) и событий (events). В отличие от MUC здесь подробнее проработаны вопросы разметки текстов для извлечения информации о событиях. Определяются следующие понятия: границы упоминания о событии (event extent) — это предложение, в котором содержится информация о событии; маркер события (event trigger) — это слово, которое обозначает событие (предикатное слово); участники события — это синтаксические группы, которые обозначают вовлеченные в событие сущности (entities). Как участники события размечаются только те релевантные для данного события сущности, упоминания о которых находятся в границах упоминания о событии.

Таким образом, если снова вернуться к примеру 4, исходя из правил разметки ACE, здесь будет два упоминания о событии:

*Владимир Путин назначил нового советника президента. Им стала первый замглавы аппарата правительства Александра Левицкая.*

Событие: *назначение / назначил*

Кто назначил \_\_\_\_\_ *Владимир Путин* \_\_\_\_\_  
 Кого назначил \_\_\_\_\_ \_\_\_\_\_  
 На какую должность \_\_\_\_\_ *советника* \_\_\_\_\_

Событие: *назначение / стала*

Кто назначил \_\_\_\_\_ \_\_\_\_\_  
 Кого назначил \_\_\_\_\_ *Александра Левицкая* \_\_\_\_\_  
 На какую должность \_\_\_\_\_ *им* \_\_\_\_\_

Для оценки модальности, референциального статуса и времени события стандартами ACE предусмотрено определение значений атрибутов маркера события (event trigger). Также определяются и дополнительные участники (например, указывающие на место или время события). Всего в проекте ACE рассматривается 35 видов событий, объединенных в 8 больших групп.

### 3. Современное состояние проблемы извлечения фактографической информации из текстов на русском языке

Для русского языка также проводятся исследования по извлечению информации о событиях. Системы извлечения фактографической информации разрабатываются и поддерживаются коммерческими компаниями Идеограф, RCO (проект RCO Fact Extractor) [7, 8], Интегрум [9], Авикомп-Сервисез (проект ОНТОС), Яндекс (проект «пресс-портреты», [10]). Ведутся и научные исследования различных аспектов задачи извлечения информации о событиях [11–15]. Однако за все время работ по теме извлечения из русскоязычных текстов информации о событиях не было создано ни представительного размеченного корпуса текстов, ни стандартов разметки, как, например, в проекте ACE. В рамках семинара РОМИП в 2005 и 2006 годах проводилась дорожка фактографического поиска [16, 17]. Но задача для этой дорожки была поставлена очень узкая, рассматривалось всего два типа событий (событие работы в организации и событие владения организацией). Кроме того, результаты извлечения оценивались только для определенных экспертами имен собственных, что фактически свело задачу извлечения информации к поиску фрагмента текста, соответствующего фрейму с заранее известными слотами. Работа по развитию разметки текстов и тестированию систем извлечения фактографической информации на семинаре РОМИП в последующие годы продолжена не была.

Конечно, для текстов на русском языке хотелось бы решить самую сложную задачу — построение семантического представления (фрейма) события определенного типа по языковому выражению (выражениям) любого вида. Это общая задача для всех исследователей и разработчиков. В настоящее время в большинстве случаев извлечение фактографической информации для русскоязычных текстов происходит поэтапно, как и на МУС. Первый этап — извлечение именованных сущностей, которые потенциально могут заполнять слоты фрейма события. На втором этапе в границах клаузы с предикатным словом (маркером события) проводится анализ ограниченного набора синтаксических конструкций, которые могут выражать событие определенного семантического типа. Извлечение опирается на шаблоны события (это могут быть семантические сети, регулярные выражения, грамматики, которые порождают описывающие события конструкции). При этом остро стоит проблема составления шаблонов

(экспертом или автоматически) для повышения полноты извлечения фактографической информации [11]. Почти нет исследований, посвященных проблеме разрешения кореферентности событий и их участников, по-разному представленных в тексте. Почти нет данных о полноте и точности современных разработок ввиду отсутствия размеченного корпуса текстов для задачи извлечения фактографической информации.

#### **4. О разметке текстов на русском языке для задачи извлечения информации о событиях (на примере событий назначения и отставки)**

Современные системы извлечения фактографической информации из текстов на русском языке вполне успешно анализируют текстовые упоминания о событиях в следующих случаях:

- (1) слово-маркер события является однозначно интерпретируемым как обозначающее ситуацию заданного типа (такие слова и выражения обычно задаются списком);
- (2) текстовые упоминания участников — отдельные языковые выражения для каждого участника. Эти выражения четко обозначают участника, информация о котором должна заполнить соответствующий слот итогового фрейма (например, слот «кого назначили» во фрейме события «назначение» должен заполняться ФИО, а слот «на какую должность» — названием должности).

Вернемся снова к примеру 1, в котором информация о событии выражена самым простым для автоматического анализа образом:

*Президент России Владимир Путин назначил Александру Левицкую советником главы государства.*

Как видно из приведенного выше примера, он полностью удовлетворяет этим условиям. Однако такие классические способы выражения информации о событиях далеко не всегда встречаются в текстах. Проанализировав еще раз примеры (2-10), можно убедиться, что очень часто приходится иметь дело с такими выражениями информации о событиях, которые не вписываются в классическую схему анализа. Тем не менее, в таких примерах содержится информация, которая позволяет человеку построить полностью или частично заполненный фрейм данного события. Зачастую такие контексты упоминания о событиях просто не рассматриваются исследователями — они ограничиваются конструкциями, как в примере (1), или конверсивами конструкций такого рода. Естественно, если исходить

из предположения, что нам нужно извлечь информацию о событиях с максимальной точностью и полнотой, необходимо включать в рассмотрение и более сложные контексты упоминаний о событиях.

Для более полного и качественного извлечения информации очень важно правильно выделить те структурные элементы текста, которые подвергаются анализу. При поэтапном подходе к извлечению информации о событиях, который применяется в подавляющем большинстве систем для русского языка (от извлечения именованных сущностей к анализу шаблонов и разрешению кореферентности), возможно, окажется полезным введение некоторого вспомогательного, технического уровня разметки текстов. Этот уровень позволит охватить максимальное количество контекстов, в которых упоминаются события, представить в виде промежуточных фреймовых структур данные, которые можно извлечь только из этого контекста.

Самая крупная единица анализа — это текст целиком (например, в задаче классификации текстов). Самая мелкая — отдельное слово и именная группа (такими единицами удобно оперировать в задаче извлечения именованных сущностей). Понятно, что для задачи извлечения информации о событиях такие структурные элементы не подходят, потому что событие в тексте в подавляющем большинстве случаев описывается сложными синтаксическими конструкциями. Чтобы эффективно извлекать информацию о событиях, нужно, с одной стороны, учитывать все многообразие способов обозначения событий в тексте, а с другой стороны, необходимо, чтобы это многообразие укладывалось в удобный для анализа универсальный формат.

Для тестирования алгоритмов на качество извлечения информации обычно используются размеченные коллекции документов. Размечены они могут быть по-разному. Для объективной оценки работы алгоритмов необходимо выработать единый стандарт разметки. Формат разметки всегда отражает в большей или меньшей степени представление исследователей о том, какие языковые единицы размечаемого текста и каким образом обозначают информацию, которая должна быть извлечена.

Разметка текстов на конференции MUC полностью соответствовала формату фрейма, в котором представлен результат извлечения. Такая разметка никак не отражает особенностей структуры языковых выражений, описывающих событие. В проекте ACE разметка упоминаний о событиях больше приближена к поверхностному синтаксису. Упоминание о событии рассматривается в границах предложения, что делает такую разметку более удобной для анализа.

Для удобства анализа текстов на русском языке (в рамках решения задачи по извлечению информации о событиях) предлагается использовать модель текстового упоминания о событии. Это микроконструкция, состоящая из линейного непрерывного текстового фрагмента, соответствующего точечному упоминанию о событии и соответствующего этому фрагменту фрейма. В слотах этого фрейма содержатся языковые выражения, обозначающие в рамках данного фрагмента само событие и его участников. Модель текстового упоминания события будет играть вспомогательную роль при построении полного фрейма события на основе анализа одного текста или целого кластера текстов на одну тему. Эта модель должна оказаться удобной единицей при разработке алгоритмов отождествления упоминаний об одном и том же событии. Кроме того, она может использоваться для разметки текстов. Оценка размеченной таким образом коллекции позволит выявить качество предварительной обработки текста для задачи полного извлечения информации о событиях.

## 5. Понятия текстового упоминания события и участника события

Мы исходим из предположения, что событие в тексте обозначается с помощью предикатного слова. В настоящей работе не будут рассматриваться случаи, когда информация о событиях выражена имплицитно и не обозначена предикатным словом. Например,

- (11) *Новый губернатор Ярославской области провел сегодня первую встречу с депутатами городской думы.*

Модель упоминания о событии состоит из текстового упоминания и фрейма события. Все примеры, иллюстрирующие модель текстового упоминания события, относятся к области событий назначения и отставки.

Условимся, что *текстовым упоминанием события* будет считаться пропозитивная часть клаузы, предикатное слово в которой обозначает событие определенного семантического типа. *Фрейм текстового упоминания о событии* есть структурированная информация, которая может быть извлечена только на основе анализа фрагмента, соответствующего текстовому упоминанию. В слоты фрейма помещается информация (в виде фрагментов текста) о типе и участниках события, встретившихся в границах данного текстового упоминания.

Рассмотрим пример:

- (12) *Президент РФ Владимир Путин подписал указ «О заместителе директора Федеральной службы исполнения наказаний». Согласно данному документу, заместитель директора ФСИН РФ Вячеслав Кузьмин освобожден от занимаемой должности.*

Фрейм события увольнения, описанного в этом тексте, такой:

Событие: *увольнение*

Кто уволил \_\_\_\_\_ *Владимир Путин* \_\_\_\_\_  
 Кого уволил \_\_\_\_\_ *Вячеслав Кузьмин* \_\_\_\_\_  
 С какой должности \_\_\_\_\_ *заместитель директора ФСИН РФ* \_\_\_\_\_

Вот так выглядят фреймы текстовых упоминаний события увольнения:

Событие: *увольнение / подписал указ*

Кто уволил \_\_\_\_\_ *Владимир Путин* \_\_\_\_\_  
 Кого уволил \_\_\_\_\_ \_\_\_\_\_  
 С какой должности \_\_\_\_\_ \_\_\_\_\_

Событие: *увольнение / освобожден от должности*

Кто уволил \_\_\_\_\_ \_\_\_\_\_  
 Кого уволил \_\_\_\_\_ *Вячеслав Кузьмин* \_\_\_\_\_  
 С какой должности \_\_\_\_\_ *заместитель директора ФСИН РФ* \_\_\_\_\_

Текстовые упоминания служат промежуточным этапом анализа текста для извлечения информации о событиях. Используя модели текстовых упоминаний, можно решать задачу отождествления упоминаний о событиях и перехода к полному фрейму события, описанного в тексте.

*Текстовым маркером события* будем называть предикатное слово (или группу слов), которое обозначает событие определенного семантического типа. Необходимо отметить, что текстовый маркер может быть устроен довольно сложным образом. Вообще, предикатное слово клаузы, обозначающей событие, далеко не всегда само по себе однозначно маркирует событие определенного семантического типа. Часто бывает так, что описание события складывается из сочетания предикатного слова и зависимых от него синтаксических групп. Эти группы, в свою очередь, могут обозначать участников события или вместе с предикатным словом формировать устойчивое выражение, обозначающее данное событие. Например,

- (13) *После ухода с поста президента Грузии Михаил Саакашвили планирует заняться виноделием.*

Здесь маркер события — словоформа «ухода» обозначает событие «увольнение» только в сочетании с группой «с поста президента Грузии».

- (14) *Министр энергетики и ЖКХ края Александр Фенев написал заявление об уходе.*

- (15) *Нынешний градоначальник Новосибирска Владимир Городецкий сложил с сегодняшнего дня свои полномочия.*

- (16) *Ушедший вчера в отставку губернатор Ивановской области Михаил Мень в ближайшее время может занять кресло в Совете Федерации.*

- (17) *Владимир Пехтин решил сложить депутатские полномочия.*

Примеры (14-17) иллюстрируют ситуацию, при которой событие маркируется устойчивым выражением из нескольких словоформ, причем они могут быть расположены не непосредственно рядом друг с другом. Кроме того, обозначающие участников события синтаксические группы могут зависеть от разных слов, входящих в состав устойчивого сочетания.

Текстовым упоминанием участника события будем считать именную группу (или ее часть), которая синтаксически является актантом обозначающего событие предикатного слова и выражает информацию об участнике данного события. Предлагается выделять границы текстового упоминания участника события, а также выделять главное слово (группу слов), непосредственно обозначающее участника. Пример,

- (18) *Президент Украины Виктор Янукович отправил в отставку премьер-министра Николая Азарова.*

Событие: *увольнение / отправил в отставку*

Кто уволил	<u>Виктор Янукович</u>
Кого уволил	<u>Николая Азарова</u>
С какой должности	<u>премьер-министра</u>

В этом примере два участника события («кого уволили» и «с какой должности») выражены в рамках одной синтаксической группы. Модель текстового упоминания о событии позволяет учесть это, выделяя упоминания участников в тексте и во фрейме текстового упоминания.

#### **б. О разметке текстов на русском языке для задачи извлечения информации о событиях (на примере событий назначения и отставки)**

Как уже упоминалось при разборе примеров в разделе 5, в процессе решения задачи извлечения фактографической информации из текстов на русском языке исследователи сталкиваются с большим количеством контекстов, которые не укладываются в стандартный шаблон языкового выражения, описывающего ситуацию. В текстовом упоминании ситуации могут встретиться местоимения (личные, указательные и возвратные) — примеры (19-20) ниже. Упоминания участников могут сложным образом сочетаться друг с другом (21-23). Кроме того, участники в текстовом упоминании о событии могут быть названы так, что для понимания, кто же собственно является участником ситуации, могут понадобиться дополнительные процедуры анализа текста (24).

- (19) *Глава News International Т. Мокридж покидает свой пост.*

- (20) *В свете постоянного нахождения Ерлана Арына в «зоне риска» рейтинга глав регионов его отставка стала закономерной.*
- И.о. главы Дагестана Рамазан Абдулатипов в понедельник*  
(21) *провел первое назначение после отставки правительства — главой Минздрава стал директор московской стоматологической клиники Танка Ибрагимов.*
- Владимир Путин назначил министра промышленности и торговли РФ Дениса Мантурова председателем наблюдательного совета «Ростехнологии».*  
(22)
- (23) *Владимир Путин назначил на эту должность министра промышленности и торговли РФ Дениса Мантурова.*
- (24) *Коллега Александра Реймера лишился поста.*

Во всех этих случаях инструмент, предложенный в настоящей работе, поможет структурировать и описать единообразным способом языковой материал и представить его в виде, удобном для последующей обработки.

## **Заключение**

Очевидно, что для достижения высокого качества извлечения информации о событиях необходимо, чтобы при поэтапном подходе к анализу текста был достигнут хороший уровень решения каждой из промежуточных задач.

На этапе извлечения фактографической информации нужно уметь правильно выделить границы текстовых упоминаний события для максимально широкого круга контекстов, а внутри этих границ — границы маркера и участников события. Это позволит создавать фреймы событий, описываемых в рамках одного предложения, а в дальнейшем перейти к следующим этапам анализа текста — отождествлению разных упоминаний о событиях и их участниках.

Хочется надеяться, что предложенный инструмент разметки текстов окажется полезным для исследователей и разработчиков. С использованием модели текстового упоминания о событиях отставки и назначения размечена тестовая коллекция текстов из 1000 новостных документов на русском языке [18].

## Список литературы

- [1] Grishman R., “Information extraction: Techniques and challenges”, *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, International Summer School, SCIE-97 (Frascati, Italy, July 14–18, 1997), Lecture Notes in Computer Science, **1299**, ed. Maria Teresa Paziienza, Springer-Verlag, 1997, pp. 10–27 ↑ 67.
- [2] Appelt D. E., “Introduction to information extraction”, *Journal AI Communications*, **12**:3 (1999), pp. 161–172 ↑ 70.
- [3] Grishman R., Sundheim B., “Message Understanding Conference-6: A Brief History”, *Proceedings of the 16th International Conference on Computational Linguistics*. v.I, COLING '96 (Kopenhagen, 1996), pp. 466–471 ↑ 70.
- [4] Chinchor N. A., *Overview of MUC-7/MET-2*, [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/proceedings/muc.7\\_proceedings/overview.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc.7_proceedings/overview.html), 1998 ↑ 70.
- [5] Ahn D., “The stages of event extraction, Annotating and Reasoning about time and events”, *ARTE '06 Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 2006, pp. 1–8 ↑ 71.
- [6] *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, version 5.4.3 2005.07.01 edition, Linguistic Data Consortium, 2005 ↑ 71.
- [7] Ермаков А. Е., Плешко В. В., «Компьютерный анализ текста при сборе информации к досье из открытых источников», Доклад на 3-ей конференции «Конкурентная разведка в металлургии» (Москва, 2005), URL [http://rco.ru/article.asp?ob\\_no=1562](http://rco.ru/article.asp?ob_no=1562) ↑ 72.
- [8] Ермаков А. Е., «Автоматическое извлечение фактов из текстов досье. Опыт установления анафорических связей», *Компьютерная лингвистика и интеллектуальные технологии*, По материалам ежегодной Международной конференции «Диалог», 2007, URL <http://www.dialog-21.ru/digests/dialog2007/materials/html/26.htm> ↑ 72.
- [9] Гершензон Л. М., Ножов И. М., Панкратов Д. В., «Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности», *Компьютерная лингвистика и интеллектуальные технологии*, По материалам ежегодной Международной конференции «Диалог», 2005, URL [http://www.dialog21.ru/Archive/2005/Gershenson%20Nozhov%20Pankratov/Gershenson\\_Nozhov\\_Pankratov.htm](http://www.dialog21.ru/Archive/2005/Gershenson%20Nozhov%20Pankratov/Gershenson_Nozhov_Pankratov.htm) ↑ 72.
- [10] *Семинар: Natural Language Processing (автоматическая обработка естественного языка)*, <http://nlpseminar.ru/archive/lecture32>, 2010 ↑ 72.

- [11] Котельников Д. С., Лукашевич Н. В., «Итерационное извлечение шаблонов описания событий по новостным кластерам», *Труды XIV Всероссийской научной конференции RCDL'2012 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»* (Переславль-Залесский, 2012), с. 353–359, URL <http://rcdl.ru/doc/2012/paper45.pdf> ↑ 72, 73.
- [12] Кузнецов И. П., «Семантические методы извлечения имплицитной информации», *Системы и средства информатики*, **21:2** (2011), с. 116–138 ↑ 72.
- [13] Пивоварова Л. М., «Фактографический анализ текста в системе поддержки принятия решений», *Вестник Санкт-Петербургского университета, серия Филология, востоковедение, журналистика*, 2010, №4, с. 190–197, URL <http://www.hse.ru/data/2010/10/14/1223101084/Faktograficheskiy%20analiz%20teksta.pdf> ↑ 72.
- [14] Власова Н. А., «Извлечение информации о ситуациях отставок-назначений в новостных текстах. Опыт разметки коллекции. Результаты тестирования», *Труды XV Всероссийской научной конференции RCDL'2013 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»* (Ярославль, 2013), с. 145–154, URL [http://rcdl2013.uniyar.ac.ru/doc/full\\_text/s4\\_2.pdf](http://rcdl2013.uniyar.ac.ru/doc/full_text/s4_2.pdf) ↑ 72.
- [15] Загорулько М. Ю., Кононенко И. С., Сидорова Е. А., «Система семантической разметки корпуса текстов в ограниченной предметной области», *Компьютерная лингвистика и интеллектуальные технологии*, По материалам ежегодной Международной конференции «Диалог». т. 1 (Бекасово, 2012), 674–683, URL <http://www.dialog-21.ru/digests/dialog2012/materials/pdf/94.pdf> ↑ 72.
- [16] И. С. Некрестьянов (ред.), *Труды третьего российского семинара по оценке методов информационного поиска*, НИИ Химии СПбГУ, Санкт-Петербург, 2005, 226 с., URL <http://romip.ru/romip2005> ↑ 72.
- [17] *Труды четвертого российского семинара РОМИП'2006* (Суздаль, 19 октября 2006 г.), НУ ЦСИ, Санкт-Петербург, 2006, URL <http://romip.ru/romip2006> ↑ 72.
- [18] Situations-1000, *Размеченная коллекция новостных текстов на русском языке, содержащих информацию о назначениях и отставках*, <http://ai-center.botik.ru/Airec/index.php/ru/collections/33situations-1000>, Исследовательский центр искусственного интеллекта, ИПС им. А. К. Айламазяна РАН, 2014 ↑ 79.

*Об авторе:*



### **Наталья Александровна Власова**

Младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.

*e-mail:* (nathalie.vlassova@gmail.com)

*Образец ссылки на эту публикацию:*

Н. А. Власова. *К проблеме разметки текстов на русском языке для задачи извлечения фактографической информации* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. 5, № 4(22), с. 67–82. URL [http://psta.psiras.ru/read/psta2014\\_4\\_67-82.pdf](http://psta.psiras.ru/read/psta2014_4_67-82.pdf)

N. Vlasova. *On annotating Russian texts for information extraction task.*

ABSTRACT. In this paper we give a brief overview of the state of the art in information extraction from Russian-language texts. We analyze MUC and ACE experience in event annotation. We introduce and give the definition of a model of event mention. Event mention is a syntactically connected text fragment referring to a target event of a pre-specified type. Information about the target event extracted from an event mention is used to populate an intermediate-level structure. This is assumed to be a helpful way of dealing with a great variety of textual references to the same target event. Extracting information on retirements and appointments is taken as example to discuss the challenges of fact extraction from Russian-language text. (*In Russian*).

*Key Words and Phrases:* automatic information extraction, factual information, test corpora, markup.