

**Федеральное государственное бюджетное образовательное  
учреждение высшего профессионального образования  
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА  
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ  
РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**Данилова В.В., Александров М.А., Трусов А.В.**

**Представление событий в новостных лентах  
регулярных изданий интернета**

**Москва– 2016**

**Аннотация.** В работе предложены способы описания (кодирования) событий, связанных с протестными настроениями. Представлен обзор современных методов извлечения протестных событий и указан способ извлечения событий, используемый в лаборатории. Рассмотренные методы и технологии могут быть адаптированы для описания и извлечения событий, не связанных с протестами (природные катаклизмы, реализация новых технологий, военно-политическое противостояние, и т.п.). В работе дается описание простой программной системы для динамического аннотирования новостных лент.

Трусов А.В., заведующий международной лабораторией математических методов исследования социальных сетей ИПЭИ Российской академии народного хозяйства и государственной службы при Президенте РФ

Александров М.А., доцент кафедры системного анализа и информатики экономического факультета Российской академии народного хозяйства и государственной службы при Президенте РФ

Данилова В.В., старший научный сотрудник международной лабораторией математических методов исследования социальных сетей ИПЭИ Российской академии народного хозяйства и государственной службы при Президенте РФ

Данная работа подготовлена на основе материалов научно-исследовательской работы, выполненной в соответствии с Государственным заданием РАНХиГС при Президенте Российской Федерации на 2015 год.

# СОДЕРЖАНИЕ

Введение

Глава 1. Описание событий

1.1 Постановка задачи

1.2 Основные подходы к извлечению событий

1.3 Алгоритмы кодирования социополитических событий (обзор)

1.4 Программные системы кодирования социополитических событий

1.5 Выводы по главе

Глава 2. Мультиязычное извлечение событий

2.1 Постановка задачи извлечения событий

2.2 Системы мультиязычного извлечения событий

2.3 Выводы по главе

Глава 3. Динамическое аннотирование событий

3.1 Постановка задачи и обзор работ

3.2 Динамические словари: алгоритм и программа

3.3 Построение аннотаций: алгоритм и программа

3.4 Эксперимент (украинские события)

3.5 Выводы по главе

Заключение

Литература

## **ВВЕДЕНИЕ**

Доступность Интернета, и, как следствие, повсеместное распространение Интернета обеспечило возможность свободного общения активной части сообщества между собой. Такое общение представляется обширной информацией, циркулирующей в Интернете, которая может явиться основой для выявления скрытых тенденций в социальном развитии общества с одной стороны, и быть инструментом для воздействия на общество с другой.

И для анализа материалов сети, и для воздействия на сеть необходимо, прежде всего, узнать предмет рассмотрения, отраженный в текстах. То есть необходимо уметь каким-либо образом извлекать события из текста. Решение подобной проблемы «на все случаи жизни» оказывается весьма затруднительным, и на сегодняшний день эта проблема успешно решается только в рамках отдельных предметных областей. В настоящей работе рассматривается достаточно узкая, но весьма важная предметная область – протестные настроения.

В работе предложены способы описания (кодирования) событий, связанных с протестными настроениями. Представлен обзор современных методов извлечения протестных событий и указан способ извлечения событий, используемый в лаборатории. Рассмотренные методы и технологии могут быть легко адаптированы для описания и извлечения событий, не связанных с протестами, как например, природные катаклизмы, появление новых технологий, военно-политическое противостояние. В заключительной главе дается описание простой программной системы для динамического аннотирования новостных лент.

# ГЛАВА 1. ОПИСАНИЕ СОБЫТИЙ

## 1.1 Постановка задачи

В рамках извлечения (социополитических или иных) событий решаются три основные подзадачи: (i) отбор, фильтрация, форматирование и занесение в базу подмножества текстов  $T_d [t_{d1}, t_{d2}, \dots, t_{dn}]$ , относящихся к заданной предметной области, множества  $T [t_1, t_2, \dots, t_n]$ , извлеченного из новостных и других источников; (ii) аннотация событийных данных, таких как тип, форма, локация, тип участника, количество участников, и пр. в каждом элементе  $T_d [t_{d1}, t_{d2}, \dots, t_{dn}]$  и сбор сценариев; (iii) слияние сценариев и помещение записей в базу данных.

## 1.2 Основные подходы к извлечению событий

Краткий обзор подходов к извлечению событий сделан в (Hogenboom et al., 2011). Рассматриваются моноязычные подходы за период с 1992 до 2010: ориентированный на знание (knowledge-driven), ориентированный на данные (data-driven) и гибридный (hybrid). Знание-ориентированный подход опирается только на базы данных, такие как онтологии, словари и газетиры (Wikipedia, WordNet, DBpedia, GeoNames, BabelNet, и др.), и шаблоны, созданные вручную на основе результатов разметки (токенизация, сегментация на предложения, тегирование частей речи, синтаксический разбор и пр.).

Использование знание-ориентированного подхода дает высокую точность извлечения и не требует обучающей выборки, однако, необходимы значительные экспертные и временные ресурсы. В рамках данного подхода, *лексико-семантические* шаблоны позволяют локализовать в тексте отношения из заданной онтологии путем обнаружения концептов на базе лингвистической предобработки и словарей, подключенных к онтологии. Действие *лексико-синтаксических* шаблонов опирается на менее абстрактный синтаксический уровень языка, лексический в данном случае представлен газетирями, не привязанными к базе данных. Используя данный подход, сложнее иметь дело с синонимами, парафразом и другими феноменами. Пары шаблон/правило в системах извлечения событий в основном представлены в виде каскадных грамматик на основе конечных автоматов. Грамматика каждого уровня строится с использованием аннотаций и их признаков, внесенных на предыдущих уровнях.

Подход, полностью ориентированный на данные (data-driven), применяет техники машинного обучения без учителя (e.g., кластеризация, EM-алгоритм, метод моментов, анализ главных компонент, сингулярное разложение, самоорганизующиеся карты, и др.), а также другие количественные методы обработки естественного языка (методы вероятностного моделирования, теории информации, линейной алгебры) к неструктурированному тексту, т. е. не основывается на априорном знании, для обнаружения скрытой структуры данных. В рамках данного подхода представление текстового документа основано на подсчете частот слов, значений TF-IDF слов,  $n$ -грамм, и т.д. На сегодняшний день точность извлечения сценариев таким способом не достаточно высока, поэтому системы мониторинга событий и базы данных его не используют. Кроме того, данные системы менее прозрачны по сравнению со знание-ориентированными, поэтому затрудняется интерпретация результатов.

Третий подход относится к гибридным. Он совмещает техники обоих подходов (ориентация на данные и на знания). За последние годы гибридные системы извлечения информации из текста стали наиболее распространенными. Они используют машинное обучение для автоматического пополнения шаблонов (слабо и частично контролируемое), для классификации последовательностей различных единиц текста на основе размеченного корпуса или банка синтаксических деревьев (контролируемое).

В (Hogenboom et al., 2011) подходы сравниваются с точки зрения количества необходимых обучающих данных, экспертных знаний, уровня подготовки пользователя, а также интерпретируемости результатов. По заключению авторов, компромиссным решением для создания системы извлечения событий является использование гибридного подхода, однако, комбинация техник из разных областей знания требует соответствующего уровня подготовленности.

### **1.3 Алгоритмы кодирования социополитических событий**

#### ***1. Экспертное и автоматическое кодирование***

Большинство проектов, использующих масштабный сбор данных по политическим событиям для социологических исследований, мониторинга или иных

целей, рассматривают, в соответствии с онтологией, множество типов событий, подтипы и свойства которых недостаточно детализированы.

В настоящем разделе рассмотрены несколько основных проектов в области. Автоматическое кодирование в них происходит по общей схеме, поэтому особое внимание уделено различиям между системами.

## ***2. Кодирование протестных данных экспертом***

Экспертное кодирование обеспечивает сбор данных для сложных качественных и количественных переменных и, таким образом, создание подробного описания события, однако, возможна субъективная интерпретация, которая может привести к ошибкам. Кроме того, сбор и кодирование вручную требует значительных временных и интеллектуальных ресурсов.

База "Динамика коллективного действия" (Dynamics of Collective Action or DoCa) Стэнфордского Университета содержит данные о протестных акциях, собираемые вручную на базе одного источника ("The New York Times"). Кодовая книга включает 38 позиций, каждая из которых снабжена развернутым описанием переменной. К примеру, для переменной "Цель" ("Target") существует 3 кода, которые являются ответами на следующие вопросы: 18a: "Была ли у протестующих четко определенная цель ?", 18b: "Если да, то сколько именно целей преследовалось протестующими и какие ?" и 18c: "Если да, была ли основная цель и какая ?". Эффективный анализ на основе подобных переменных может быть проведен экспертом, использование большей их части машинными алгоритмами затруднительно и может оказаться невостребованным. Преимущество данной базы в том, что она включает качественно отобранную информацию по ряду основных переменных, характеризующих событие, а именно: *тип, форма, дата, локация, масштаб, актор-участник, актор-инициатор, требование, актор-цель, применение насилия, жертвы*. На данных этих слотов в (Hanna, 2014) основано обучение и тестирование алгоритмов автоматического кодирования для английского языка.

"База местоположений и событий, связанных с вооруженными конфликтами" (Armed Conflict Location and Event Dataset или ACLED), описанная в (Raleigh et al., 2010), в настоящее время функционирует как контролируемая система обработки данных в почти реальном времени. По сравнению с DoCa, она обнаруживает больше

типов событий, относящихся к *политическому насилию*. Политическое насилие определяется как политически мотивированное применение насилия группой людей. Сбор и первоначальное кодирование данных в ACLED осуществляется индивидуальными обученными кодировщиками. Проверка и окончательное принятие решений производится последовательно тремя экспертами. Каждое событие описано следующими переменными: *ID, тип, дата, локация, контекст, акторы (несколько подтипов), жертвы*.

### **3. Автоматическое кодирование данных**

Цель системы автоматического кодирования событий - обеспечить динамически растущую базу качественных и количественных признаков различных типов событий в данной области и минимизировать необходимость экспертного контроля. В области сбора и кодирования политических событий, системы выполняют примерно одну и ту же последовательность действий. События обнаруживаются по лексиконам, собранным вручную и подключенным к онтологии акторов и событий, в связи с чем качество полученной информации напрямую зависит от лежащей в основе онтологии. Событие, представляющее интерес (Event of Interest или EOI), - это диадическая система, составленная из двух сущностей (исходной и целевой), объединенных определенным отношением (тип события). Для данных систем каждое предложение может содержать от одного до нескольких событий, где группа глагола содержит триггер события, взятый из лексиконов, а левая и правая именные группы представляют собой искомые сущности (исходного и целевого акторов). Для достижения своих целей системы используют в основном инструменты обработки естественного языка (Natural Language Processing или NLP) и машинное обучение (Machine Learning или ML). Инструменты NLP включают простые алгоритмы предобработки и поверхностный парсинг именных и глагольных групп (shallow parsing). Гранулярность одних и тех же признаков (например, геотемпоральных характеристик события) для разных систем варьируется. В случае, если система обрабатывает неанглийские тексты, тексты на одном из этапов предобработки переводятся на английский. В данном и следующем разделах рассмотрены следующие проекты, использующие компьютерные методы: KEDS (Schrodt, 1998), El:DIABLO (<http://openeventdata.github.io/eldiablo/>), W-ICEWS (<http://www.lockheedmartin.com/us/products/W-ICEWS/iData.html>), SPEED (Nardulli et

al., upcoming), и GDELT (Schrodt, 2013). KEDS и его нынешняя реинкарнация El:DIABLO (Event/Location: Dataset In A Box, Linux-Option) под патронажем Открытого объединения данных о событиях (Open Event Data Alliance) - это завершенные и регулируемые системы для сбора и кодирования данных о политических событиях, ориентированные на конечного пользователя. Протокол общественной стабильности (Societal Stability Protocol или SSP) создан в рамках проекта "База данных социальных, политических и экономических событий" (Social, Political, and Economic Event Database или SPEED) Института изучения демократии при Университете штата Иллинойс в Урбана Шампейн и покрывает события, связанные с общественными волнениями во всем мире после Второй Мировой Войны. Информационная интегрированная система раннего предупреждения о возникновении кризисных ситуаций (World-Wide Integrated Crisis Early Warning System или W-ICEWS) была разработана для нужд правительства США компанией Локхид Мартин (<http://www.lockheedmartin.com/us/products/W-ICEWS.html>) и включает ряд автоматических средств для сбора данных (iDATA), мониторинга (iTRACE), прогнозирования (iCAST), и анализа мнений из открытых социальных сетей (iSENT). Она выросла из более раннего одноименного проекта, использовавшего кодирование данных экспертом вручную. Доступ к данным SSP и iDATA ограничен. Ранее упомянутая система GDELT появилась в 2013 году и на данный момент является самым крупным открытым проектом, посвященным сбору данных о событиях.

Общая схема работы систем автоматического кодирования представлена на Рис. 1.1. Основные характеристики систем показаны в Табл. 1.1 Представленные системы имеют модульное строение, и показатели качества их работы зависят от уровня инструментов, входящих в их состав.

Рассматриваемые системы машинного кодирования событий действуют по общей схеме. RSS-агрегаторы (Рис.1:1 *Aggregator*) собирают новостные статьи из белого каталога RSS (Рис.1: *Tabloids*), опубликованные за заданный промежуток времени и относящиеся к заданной области. Итоговая коллекция проходит первичную обработку, включающую очищение выдачи от полных дубликатов, форматирование и перевод в том случае, если статьи на неанглийском языке (Рис.1/1: *Filtering, Formatting, Translation*).

Рис.1.1. Общий механизм работы систем автоматического кодирования событий.

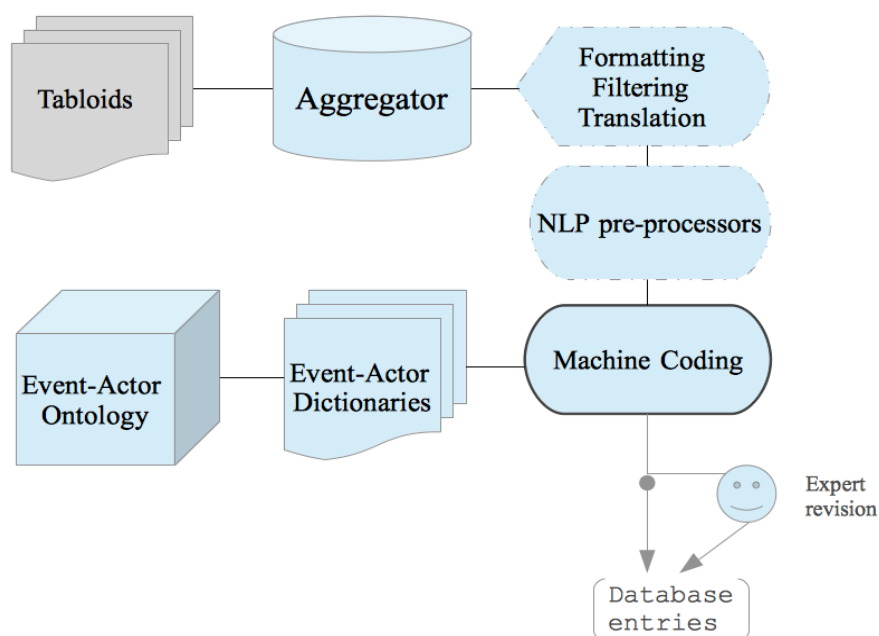


Таблица 1.1 Характеристики систем автоматического кодирования событий.

| Проект                   | KEDS                                   | El:DIABLO                  | W-ICEWS<br>(2001-<br>наст.вр.)             | SPEED<br>(1945-<br>2005)       | GDELT 2.0<br>(1979-<br>наст.вр.) |
|--------------------------|----------------------------------------|----------------------------|--------------------------------------------|--------------------------------|----------------------------------|
| Контроль<br>сбора данных | Неконтролируемый/<br>Контролируемый    | Неконтролируемый           | Неконтролируемый                           | Контролируемый                 | Неконтролируемый                 |
| Возможность<br>настройки | +                                      | +                          | -                                          | -                              | -                                |
| Языковое<br>покрытие     | английский                             | английский                 | английский,<br>испанский,<br>португальский | перевод                        | Google<br>Translate              |
| Область                  | политические<br>отношения              | международные<br>конфликты | Международные<br>конфликты                 | общественные<br>волнения       | международные<br>конфликты       |
| Источники                | Reuters, Agence<br>France Presse       | около 160<br>вебсайтов     | более 6000<br>новостных лент               | более 800<br>новостных<br>лент | тысячи<br>новостных лент         |
| Онтология                | WEIS <sup>2</sup> , CAMEO <sup>3</sup> | CAMEO                      | CAMEO                                      | SSP                            | CAMEO                            |
| Лексиконы/<br>Словари    | KEDS <sup>4</sup> или по<br>выбору     | по выбору<br>пользователя  | собственные<br>словари                     | -                              | WordNet <sup>5</sup>             |
| Поисковая                | Nexis <sup>6</sup>                     | собственная                | Factiva <sup>7</sup> , OSC <sup>8</sup>    | Heritrix <sup>9</sup> ,        | Google                           |

|                                |                               |                                                                                   |                                                   |                                                   |                   |
|--------------------------------|-------------------------------|-----------------------------------------------------------------------------------|---------------------------------------------------|---------------------------------------------------|-------------------|
| система                        |                               |                                                                                   |                                                   | BIN <sup>10</sup>                                 |                   |
| Кодировка местоположения       | CountryInfo.txt <sup>11</sup> | Penn State “GeoVista” project coder <sup>12</sup> , UT/Dallas coder <sup>13</sup> | CountryInfo.txt                                   | GeoNames <sup>14</sup>                            | CountryInfo.txt   |
| Кодировка триггера и сущностей | собственный парсер            | CoreNLP <sup>15</sup> , PETRARCH <sup>16</sup>                                    | BBN Serif <sup>17</sup> , JabariNLP <sup>16</sup> | Apache open NLP <sup>18</sup> , EAT <sup>19</sup> | CoreNLP, PETRARCH |

<sup>1</sup> World Event/Interaction Survey Ontology: см. (McClelland, 1999)

<sup>2</sup> Наблюдения за конфликтами и их урегулированием (Conflict and Mediation Event Observations): см. (Schrodt, 2012)

<sup>3</sup> Встроенные словари для онтологии WEIS конфликтов на Среднем Востоке (Pevehouse, PANDA). См. (Schrodt, 2012: p. 14)

<sup>4</sup> WordNet тезаурус: <http://wordnet.princeton.edu>

<sup>5</sup> Открытая для использования в исследовательских целях поисковая система Nexis: <http://lexisnexis.com>

<sup>6</sup> Коммерческая поисковая система Factiva: <http://new.dowjones.com/products/factiva/>

<sup>7</sup> Центр открытых источников при правительстве США: [http://en.wikipedia.org/wiki/Open\\_Source\\_Center](http://en.wikipedia.org/wiki/Open_Source_Center)

<sup>8</sup> Открытый поисковой робот Internet Archive's open-source crawling project: <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

<sup>9</sup> Автоматическое распределение больших коллекций документов по тематически неоднозначным классам (Leetaru, 2011)

<sup>10</sup> CountryInfo.txt включает более 32 000 лексем для распознавания названий 240 стран и административных единиц: стран, столиц, регионов, национальных лидеров: <https://github.com/openeventdata/CountryInfo>

<sup>11</sup> Геовизуализация и пространственный анализ медицинских данных: <http://www.geovista.psu.edu/grants/nci-esda/software.html>

<sup>12</sup> Geospatial Information Sciences (GIS) при Университете штата Техас в Далласе: <https://github.com/mriiiron/utd-geocoding-locator>

<sup>13</sup> Открытая географическая база данных GeoNames: <http://www.geonames.org/>

<sup>14</sup> NLP-пакет на базе Java, включающий алгоритм распознавания именованных сущностей на основе случайных условных полей: <https://github.com/stanfordnlp/CoreNLP>

<sup>15</sup> Программное обеспечение Open Event Data Alliance: <https://openeventdata.github.io/>

<sup>16</sup> См. (Boschee et al., 2005)

<sup>17</sup> NLP-пакет на базе Java на основе методов максимальной энтропии и перцептрона: <http://opennlp.apache.org/>

<sup>18</sup> См. (Hayes and Nardulli, 2011)

В рамках проекта SPEED используются архивы "The New York Times", Информационной службы зарубежного вещания (CIA Foreign Broadcast Information Service) и службы всемирного обозрения BBC (Summary of World Broadcasts), включающие материалы локальной прессы разных стран, переведенные на английский язык профессиональными переводчиками. Кроме того, схема процессов SPEED содержит этап предварительной категоризации текстов основной коллекции по темам: политически мотивированные атаки (*politically motivated attacks*), политические выражения (*political expressions*), атаки со стороны власти (*state attacks*). Классификация произведена с использованием инструмента BIN на основе наивного байесовского классификатора.

На следующем этапе (NLP pre-processors) набор данных обрабатывается NLP-пакетами, такими как Apache OpenNLP (функции: *токенизация, сегментация текста на предложения, тегирование частей речи, парсинг* (поверхностный или глубокий), *разрешение кореференции, распознавание именованных сущностей*) или разработкой университета Стэнфорда CoreNLP (функции: *токенизация, лемматизация, сегментация текста на предложения, тегирование частей речи, парсинг зависимостей, разрешение кореференции, распознавание именованных сущностей*).

Следующий шаг - извлечение и кодирование события (Machine Coding) такими программными средствами, как PETRARCH, JABARI-NLP, TABARI и др.. В настоящее время TABARI (Text Analysis By Augmented Replacement Instructions) (Schrodt, 2014b) и его версию на базе Java с NLP дополнениями JABARI-NLP (Schrodt & Van Brackle, 2013) заменяет более современный инструмент PETRARCH (Python Engine for Text Resolution and Related Coding Hierarchy) в системах El:DIABLO и GDELT. Данный кодировщик принимает на вход текст, предобработанный с использованием NLP-пакетов, таких как ранее упомянутые CoreNLP, Apache OpenNLP, а также Penn Treebank (<https://www.cis.upenn.edu/~treebank/>). Собственно кодирование события производится на основе поиска кодов в подключенной онтологии (CAMEO, SSP или

др.) (Рис. 1: *Event-Actor Ontology*) через словари акторов и событий (Рис. 1: *Event-Actor Dictionaries*). Словари акторов W-ICEWS содержат собственные имена с вариантами написания и названия неименованных групп лиц ("военные", "шииты") и их времязависимые аффилиации (<http://thedata.harvard.edu/dvn/dv/icews>). Что касается кодирования данных, проект W-ICEWS использует комплексный подход: JABARI-NLP выполняет поверхностный парсинг, и BBN's Serif (Statistical Entity and Relation Information Finder) извлекает триплеты Сущность-Отношение-Сущность. Serif - ограниченный для доступа инструмент, использующий гибридный подход на основе шаблонов и алгоритмов машинного обучения. Перед загрузкой на сервер в некоторых системах (напр., El:DIABLO) производится контрольная дедупликация событий. GDELT не удаляет дубликаты новостей, поэтому использование данной системы для прогнозирования и ряда других задач затруднено. Наилучшее ее применение - измерение покрытия СМИ тех или иных событий (отмечено во множестве ресурсов, в т. ч.: <http://aaboyles.com/tag/gdelt/>, передовые блоги компьютерных социологов Bad Hessian, Jasper Ginn, Jay Ulfelder, и др.). Высокая частота упоминаний определенного события в СМИ говорит о значении, которое ему хотят придать. В SPEED результаты корректируются вручную экспертами перед помещением в базу данных (Рис.1.1: Expert Revision).

#### ***4. Прототипы систем кодирования протестных событий***

Уже в течение нескольких десятилетий передовые институты социологических исследований, такие как Центр Беркмана по изучению Интернета и общества, Центр изучения демократии при Университете штата Иллинойс в Урбана Шампейн, и др. исследуют закономерности проявления протестной активности, используя как данные одиночных событий, так и совокупности событий, а также, с недавнего времени, большие данные (big data). Кроме того, с появлением социальных медиа стало актуальным изучение взаимовлияния между виртуальными и реальными проявлениями протестной активности, между традиционными средствами массовой информации, зачастую контролируемые тем или иным государством, и дискуссиями в социальных сетях, где пользователь чувствует себя независимым от власти. Основным источником для сбора базовых данных о протестной активности по-прежнему являются периодические издания. Как отмечает (Wueest et al., 2013), с экспоненциальным ростом количества, глобальностью и

доступностью онлайн новостей второе поколение аналитиков протестных данных впервые продвинуло кросснациональные исследования протестной активности в середине девяностых годов. Третье поколение аналитиков протестных данных придает большое значение решению проблемы тенденциозной подачи новостей (news bias). Кроме того, в это время впервые применяются машинные методы сбора политических событий: созданы адаптированные версии KEDS (Kansas Event Data System: <http://eventdata.psu.edu>). Они используют ключевые слова с целью ускорения поиска релевантных текстов в электронных архивах. Лексиконы на базе онтологии служат в таких системах для обнаружения события и его участников. Четвертое поколение рассматривает событие изнутри как совокупность различных действий, а также особое значение придает требованию протестующих, при этом проблема тенденциозной подачи новостей исследуется в меньшей степени.

Несмотря на то, что исследования протеста до сих пор проводятся методом опроса участников в гуще событий и анализа ограниченного объема данных по теме, получение как можно более полноценного представления о таких компонентах как масштабы, причины, последствия и параллели с другими событиями представляет определенные трудности без использования автоматических методов извлечения информации. Сбор этих компонентов для большого объема событий из множества новостных источников на разных языках вручную не возможен, однако, его необходимость подчеркивается в самых последних исследованиях в данной области (Hanna, 2014; Wueest et al., 2013), в связи с чем нами создан прототип соответствующей системы.

Автоматическое кодирование данных имеет следующие преимущества: возможность обработки больших данных за короткий промежуток времени, геотемпоральное покрытие, возможность повторного запуска, модифицируемые словари, отсутствие необходимости оплаты труда обученным или необученным кодировщикам, объективность данных. Кроме того, существующие системы имеют следующие недостатки: невозможность парсинга сложных предложений, образных выражений, время-зависимого текста (Schrodt & Yonamine, 2012).

Основным преимуществом использования новостных источников для анализа протестных событий является открытый доступ и покрытие больших временных интервалов. При этом, главным недостатком остается тенденциозная подача

новостей. Покрытие одних и тех же событий в разных странах неоднородно и зависит от политики высших инстанций: делается упор на событиях локального масштаба; часть событий не освещается (напр., происходящих в определенной стране); фактические данные о событии отличаются от других источников, и т.д.. Эффекты предвзятой подачи новостей могут быть смягчены и устранены, по мнению ученых (Hanna, 2014), через подключение и сравнение множества информационных источников.

Калев Леетару, соавтор Глобальной базы данных событий, языка и тона (Global Database of Events, Language and Tone или GDELT), согласно (Fortuna's Corner, 2014), подчеркивает важность покрытия новостей на национальных языках. Он отмечает, что в ходе автоматической селекции данных 70% событий из португалоязычных источников не упомянуты в англоязычных СМИ. До сотрудничества с Google Translate, GDELT упустила важные факты, касающиеся вспышки вируса Эбола во франкоязычных гвинейских новостях.

Предметом анализа систем, представленных в данном разделе, является протестная активность. Прототипы используют смесь техник машинного обучения и обработки естественного языка для автоматической селекции и кодирования протеста.

В (Wueest et al., 2013) предложен метод постселекции статей для дальнейшего кодирования. Первоначально статьи о протестах отбираются путем поиска в LexisNexis на основе паттернов, состоящих из ряда операторов (логических и других) и ключевых слов. Пример запроса представлен на Рис. 3. Указание стема слова и следующего за ним знака усечения (!) позволяет искать данное слово независимо от его морфологической вариации. I/1 означает, что искомые термины должны быть расположены на расстоянии одного токена друг от друга. В целях экспериментов авторы используют коллекцию новостей из архива "The Guardian" за 2010 год. Как видно из Рис. 1.2, данный запрос покрывает множество форм события: включены не только протесты против (действий) некоторой сущности, но и акции в поддержку, памяти, нападения на людей и имущество и пр.: *signature collection* (сбор подписей), *petition submission* (подача петиции), *protest* (протест), *demonstration* (демонстрация), *manifestation* (манифестация), *march* (марш), *parade* (парад), *rally* (митинг), *picket* (пикет), *riot* (бунт, беспорядки), *human chain* (живая цепь), *affray*

(драка), *letter campaign* (письменное обращение), *festival* (фестиваль), *ceremony* (церемония), *street theater* (уличный театр), *road show* (театральное представление), *student strike* (забастовка студентов), *vigil* (пикетирование), *boycott* (бойкот), *hunger strike* (голодовка), *blockade* (блокада), *sit-in* (сидячая забастовка), *squatter* (сквоттер), *mutiny* (мятеж), *assaults* (налет), *attack* (атака), *bombing* (бомбардировка), *arson* (поджог), *fire raising* (поджог), *sabotage* (саботаж), и т.д.. В связи с омонимией терминов, высока доля ложноположительных результатов (только в 68 из 727 обнаруженных статей речь идет о протестах).

((submission OR submit! AND initiative OR referendum) OR petition! OR (collect! AND signature! AND campaign!) OR protest! OR demonstrat! OR manifest! OR marche! OR marchi! OR parade OR rall! OR picket! OR (human chain) OR riot! OR affray OR (letter! I/1 campaign!) OR parade OR festival OR ceremony OR (street theatre) OR (road show) OR vigil OR (consumer OR lecture OR university OR campus OR college OR school OR pupil! OR student! AND strike!) OR boycott! OR (hunger strike!) OR blockade OR (block! AND street OR traffic OR area OR site) OR sit-in OR (sit! AND strike!) OR squatter! OR (squat! AND house OR building OR area OR property) OR mutin! OR bomb! OR firebomb! OR molotov OR graffiti OR (paint! OR colour OR fire AND assault) OR attack OR arson OR incendiar! OR (fire I/1 raising) OR (set AND ablaze) OR landmine OR sabot! OR hostage! OR assassinat! OR shot OR murdered OR killed)

Рис.1.2. Запрос в LexisNexis для поиска статей о протестах

Собранные тексты проходят полную лингвистическую предобработку: токенизацию, удаление стоп-слов, лемматизацию, тегирование частей речи<sup>1</sup>, сегментацию на предложения и разбор синтаксических зависимостей<sup>2</sup>.

Последнее не улучшило результат классификации, однако, авторы рассматривают применения парсинга зависимостей на этапе кодирования признаков.

В целях постселекции был построен бинарный классификатор с активным обучением, который использует результат лингвистического разбора системой UIMA (<https://uima.apache.org/>) и скрытые тематические модели (*hidden topic models*<sup>1</sup>) в качестве признаков. Активное обучение подразумевает активное участие учителя/аннотатора: первоначально аннотируется небольшая коллекция текстов, на которой алгоритм обучается; далее на вход подается большая коллекция, на базе

которой алгоритм решает, какие примеры из множества кандидатов должны быть аннотированы учителем/аннотатором для увеличения точности.

<sup>1</sup> TreeTagger (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>) и набор тегов Penn Treebank (<http://www.cis.upenn.edu/treebank/>)

<sup>2</sup> mate tools (<http://code.google.com/p/mate-tools/>)

Таким образом, алгоритм отбирает наиболее "полезные" и менее известные примеры, аннотирование которых сократит время обучения и повысит его эффективность. На практике это означает, что для таких примеров распределение вероятностей принадлежности к целевому классу имеет наибольшую энтропию при данных значениях признаков (Piskorski & Yangarber, 2013). Оценка работы классификатора была проведена на обучающей и настроенной выборках. В результате точность (Precision) составила 70.1% и полнота (Recall) - 88.6%.

Процедура кодирования события состоит в извлечении следующих признаков:

тематика (экономика, культура, безопасность, учреждения), соответствующая ей форма протеста (экономика: петиции, письменные обращения, и т.д.; культура: общественные демонстрации, и т.д.; общественная безопасность и порядок: забастовки, бойкоты, оккупации, и т. д.; институциональные мероприятия/кампании: атака, блокада и др.), количество участников, местоположение события, дата, участвующие организации. Извлечение всех признаков, кроме формы и тематики, производится распознавателем именных сущностей Стэнфордского университета (OpenNLP Named Entity Recognizer). Самостоятельное построение газетира потребовалось только для извлечения количества участников. Авторы предлагают распознавать информацию о тематике протеста на основе расстояния между данным текстом и соответствующей моделью скрытой темы (hidden topic). Формы протеста будут идентифицироваться с использованием эвристических правил на результатах построения моделей пространства слов (word space models), состоящих из триплетов синтаксических зависимостей. Пример триплета, извлеченного из предложения "*Workers protest against the Government's proposed public sector pensions cuts*" (Рабочие протестуют против предложения Правительства сократить пенсии в общественном секторе): *protest* + *against* + *cuts* (причина протеста). Оценка данного

подхода к автоматическому аннотированию событий в (Wueest et al., 2013) не представлена.

В (Hanna, 2014) представлены предварительные эксперименты по созданию системы генерации протестных данных на основе обучения с учителем. Задача аннотирования признаков события рассматривается как задача классификации (1) документов: бинарная классификация на документы, посвященные протестным событиям, и другие (не описывающие протестные события) документы, и

<sup>1</sup> Mallet toolkit (<http://mallet.cs.umass.edu/>)

(2) последовательностей символов: частичное кодирование события, классификация причины, цели, доминантных форм протеста и иницирующих групп (многоклассовая классификация). В целях экспериментов два классификатора на основе машин опорных векторов (Support Vector Machines) обучены на протестных данных базы DoCa (по статьям "The New York Times"), используя пакет Python scikit-learn. Автор намечает решение для второго этапа кодирования (определение местоположения, масштаба события и участвующих организаций), однако, алгоритм в разработке, и оценка его работы не представлена. Предлагается обучить для данной задачи классификатор именных сущностей на основе линейных цепей условных случайных полей (OpenNLP)

Предобработка текстов включает XML-форматирование, перевод в нижний регистр, фильтрацию пунктуации и стоп-слов, вычисление меры TF-IDF (Term Frequency - Inverted Document Frequency). Для представления использованы оригинальные формы слов (без стеммирования/лемматизации). В качестве параметров выполнения установлены линейное ядро (linear kernel) и мягкий зазор (soft margins), что, по словам автора, позволяет улучшить результат на несбалансированных классах. Документы представлены векторами значений меры TF-IDF, подсчитанной на униграммах (словах, состоящих из одного токена). Оценка бинарной классификации показывает очень низкую точность (0.06) для класса "protest", в связи с чем автор предлагает в дальнейшем ввести порог (1) для разграничения классов "protest" ( $>1$ ) и "non-protest" ( $<1$ ). В результате многоклассовой классификации получены сбалансированные классы. Точность и

полнота определения причины и цели протеста колеблются около 70% (возможно, вследствие достаточно четкого языкового представления данных концептов), акторов (участников события) - 61 % и 54% соответственно. Классификация формы протеста намного менее успешна: точность и полнота менее 50%. Подгруппы информационного слота цели/целевого актора события (сущность или организация, к которой обращены требования протестующих) обобщены на основе спецификаций DoCa следующим образом: Government/State (правительство/государство), Private/Business (частная организация/бизнес), University/School (университет/школа), Medical Facility/Organization (медицинское учреждение/организация), Other (другое), Ethnic/Racial Group (этническая/расовая группа). Предложенный подход к кодированию протеста естественен в контексте повсеместной интеграции машинного обучения в социологию, однако, при его использовании в данном случае возникает ряд сложностей, связанных с неполноценностью базы DoCa (моноязыковая, один новостной ресурс, и др.). Автор намечает следующие улучшения: 1) аннотировать корпус, построенный на большем количестве новостных ресурсов, для обучения классификатора (необходимо в связи с наличием в DoCa множества "белых пятен", ограниченностью словаря и тенденциозной подачей новостей "The New York Times"), 2) обеспечить извлечение не одного, а нескольких или множества событий на статью, как это сделано в PETRARCH.

#### **1.4 Программные средства кодирования социо-политических событий**

*Отбор и кодирование события* - две основные подзадачи в области автоматического пополнения баз протестных данных. *Отбор события* - это автоматическая селекция документов, описывающих событие в данной области. Кодирование события определяется как процедура автоматического присвоения кода некоторому элементу информации: к примеру, код для обозначения правительственного актора может выглядеть как [GOV]. Алгоритмы отбора событий в KEDS, El:DIABLO, W-ICEWS, SPEED, и GDELT были кратко рассмотрены в предыдущем разделе. Они основаны на запросах по ключевым словам и стандартных алгоритмах тематической категоризации текстов. Далее описаны алгоритмы и

ресурсы для кодирования событий в рассматриваемых системах: PETRARCH, Serif и EAT (инструменты) и CAMEO и SSP (онтологии).

## ***1. Онтологии***

### ***CAMEO***

CAMEO разрабатывается с 2000 года по настоящее время и описывает отношения в области медиации межгосударственных конфликтов. В связи с тем, что данная онтология выросла из онтологий времен холодной войны (таких как WEIS и другие), она не ориентирована на кодирование некоторых политически релевантных событий, таких как стихийные бедствия, криминальная и финансовая деятельность, миграция, беженцы, нарушение прав человека, предвыборные события, действия парламента, и др. (Schrodt, 2014a). CAMEO включает 20 категорий глагола-действия (action verb) первого уровня (Tier-I): 01: Make Public Statement (публичное заявление), 02: Appeal (обращение с призывом), 03: Express Intent to Cooperate (выражение намерения сотрудничать), 04: Consult (встречи, совещания), 05: Engage in Diplomatic Cooperation (вовлечение в дипломатическое сотрудничество), 06: Engage in Material Cooperation (вовлечение в материальное сотрудничество), 07: Provide Aid (предоставление помощи), 08: Yield (уступки, концессии), 09: Investigate (не тайные расследования), 10: Demand (запросы), 11: Disapprove (жалобы, возражения), 12: Reject (отказы), 13: Threaten (угрозы), 14: Protest (протесты), 15: Exhibit Military Posture (демонстрация готовности к применению вооруженной силы), 16: Reduce Relations (ограничение сотрудничества), 17: Coerce (репрессивные меры, применение насилия против гражданских лиц), 18: Assault (использование нетрадиционных форм насилия, таких как похищение людей, террористические акты и пр. ), 19: Fight (использование традиционных форм насилия: вооруженных сил (армии), ВВС, танков, артиллерии и пр.), 20: Engage in Unconventional Mass Violence (вовлечение в нетрадиционное массовое насилие: массовое изгнание, использование биологического оружия и пр.) (Schrodt, 2012). (Schrodt et al., 2014) отмечает, что CAMEO не ориентирована на кодирование проявлений коллективного поведения. В категории 14: Protest (протесты) перечислены 4 типа протестных событий с основными целями и формами (мануал CAMEO: (Schrodt, 2012)): 140. Engage in political dissent (склонять к политическому инакомыслию), 141. Demonstrate or rally (проводить демонстрацию или митинг) с подкатегориями второго уровня (Tier-II):

проводить демонстрацию или митинг в поддержку смены власти, с требованием реформ, в защиту социальных, экономических или других прав, за изменение режима; 142. Conduct hunger strike (устроить голодовку) (Tier-II как в 141); 143. Conduct strike or boycott (устроить забастовку или бойкот) (Tier-II как в 141); 144. Obstruct passage, block (блокировать движение, оккупировать территорию) (Tier-II как в 141); 145. Protest violently, riot (применять насилие в рамках протеста) (Tier-II как в 141). 140 - общая подкатегория без уточнения формы протеста (протест в любой форме против целевого актора). Демонстрации "в поддержку" будут кодироваться отдельно: они могут быть отнесены к 05: Engage in Diplomatic Cooperation (вовлечение в дипломатическое сотрудничество). Устное выражения протеста также будет принадлежать к другому типу событий, так как считается менее "действенным" в отношении давления на целевого актора, чем массовые собрания. Протесты в военной сфере войдут в категорию 19. Fight (использование традиционных форм насилия: вооруженных сил (армии), ВВС, танков, артиллерии и пр.). Другие события, имеющие отношение к протестам, такие как поддержка (финансовая или иная), угроза, и пр., либо распределены по онтологии, либо не представлены. Временные или иные отношения между событиями не фиксируются, в связи с чем подсобытия не будут ассоциироваться с основным так, как это следует из естественного текста.

Часть онтологии, посвященная акторам, включает различные наименования представителей власти, организаций, движений, аффилированных и неаффилированных групп лиц и пр..

Как отмечено в (Hanna, 2014), подкатегории позиции 14: Protest (протесты) редко извлекаются на практике. Кроме того, спецификация акторов в CAMEO подходит для описания межгосударственных отношений и в меньшей степени применима к взаимодействиям между государством и обществом (общественными движениями).

## **SSP**

Протокол общественной стабильности (Societal Stability Protocol или SSP) -

это система баз данных прямого ввода, состоящая из шести секций ("кто?", "что?", "каким образом?", "где?", "когда?" и "почему?"), каждая из которых

содержит ряд вопросов для эффективного извлечения информации о событии. Протокол содержит подробную онтологию дестабилизирующих событий: 65 категорий Tier-I и Tier-II и более ста категорий в целом. *Дестабилизирующее событие* - это происшествие, нарушающее привычное течение жизни и ожидания граждан, заставляющее их испытывать страх и усиливающее их неуверенность в благополучном будущем (Nardulli et al., upcoming). База данных покрывает тексты, вышедшие в новостных источниках 165 стран со времен Второй Мировой. Данные статьи представляют собой оригинальные тексты на английском языке, либо переводы оригинальных неанглийских статей для англоязычных обозревателей. Окончательная версия дизайна протокола включает одиннадцать секций, отвечающих за обработку данных. Шестая секция описывает доменную онтологию типов событий, состоящую из 3-х категорий Tier-I: political expression events (события политического выражения), politically motivated attacks (политически мотивированные атаки негосударственных лиц), disruptive state acts (проявления агрессии со стороны власти). Политическое выражение - это публичное выражение угрозы или иного нежелательного политического сообщения лицом (лицами), не имеющим(и) отношения к власти. Основные способы выражения следующие: а) письменное или устное сообщение (с подпунктами); b) символическое действие (с подпунктами); c) формирование сообщества (с подпунктами); d) массовая демонстрация или забастовка (с подпунктами). Политически мотивированные атаки - это акты агрессии или попытки их совершить, инициируемые лицом (лицами), не имеющим(и) отношения к власти. Политическая мотивация - это ненависть к социокультурным группам или месть за их предшествующие действия, желание изменить или проконтролировать действия власти, поддержать или выразить протест против какой-либо идеологии, предложить прогрессивную общественную идею, и т.д. Агрессия со стороны власти подразумевает репрессии или экстраординарные меры, инициируемые властью по отношению к общественным группам. Данные о категориях социополитических событий для онтологии SPEED были собраны и организованы в структуру на основе сведений из литературы о политическом насилии, терроризме, политической нестабильности и общественных движениях. Далее полученная структура была улучшена на основе анализа данных корпуса. Были определены характеристики событий, представляющие интерес для изучения источников и развития общественных волнений. Атрибуты события, такие как

геотемпоральные данные (страна, регион, город, широта, долгота, время и дата), акторы (участники и инициаторы: представители власти или группы людей/организации, не имеющие отношения к власти, и их описательные характеристики, такие как количество, используемое оружие, и пр.), последствия (негативные/позитивные для инициаторов/участников), цели, эффекты (ущерб, ранения, и пр.), причины и связки между событиями распределены между другими секциями (Hayes and Nardulli, 2011; Nardulli et al., upcoming; Danilova, 2014).

## ***2. Программная поддержка аннотирования и кодирования событий***

### ***PETRARCH***

PETRARCH (Python Engine for Text Resolution And Related Coding Hierarchy) - усовершенствованная альтернатива кодировщику TABARI, который производит аннотацию триплетов Сущность-Отношение-Сущность на основе поверхностного анализа синтаксической структуры. Метод TABARI робастный, но в результате недостаточности правил получается много идентичных ложноположительных аннотаций. TABARI, в свою очередь, - преемник инструмента KEDS, разработанного в 90-е годы. В настоящее время из упомянутых программ только PETRARCH обновляется: совершенствуется алгоритм и расширяются словари акторов, агентов и глаголов (<https://github.com/openeventdata/Dictionaries>). Отличие PETRARCH от предшественников состоит в подаче на вход результатов полного разбора парсером Penn TreeBank (Schrodt, 2014b) и подключении словаря глагольных шаблонов CAMEO.verbpatterns.140609.txt с большим набором синонимов. Однако, введенные инновационные элементы замедляют обработку. TABARI кодирует от 1000 до 2000 предложений в секунду, в зависимости от словаря, в то время как PETRARCH обрабатывает только 150 предложений в секунду, при этом анализ CoreNLP занимает около 2-5 предложений в 1 сек.. Следующие задачи все еще актуальны для PETRARCH после модернизации: 1) снятие омонимии; 2) распознавание синтаксической компрессии, которая, будучи распространенным явлением в новостном тексте, приводит к частым ошибкам парсера; 3) обработка сложных синтаксических структур. Авторы видят следующие преимущества модернизации: 1) потенциальная возможность обработки мультязычных новостей; 2) идентификация акторов в именных группах независимо

от словаря; 3) улучшение распознавания исходного актора в прямых дополнениях при корректном разборе; 4) разрешение многозначности существительных/глаголов/прилагательных; 5) Язык Python, на котором написан PETRARCH, постоянно развивается, моложе и гибче по сравнению с C/C++ (TABARI). Пример результата кодирования предложения в TABARI и PETRARCH показан на Рис. 2.

BAGHDAD. Iraqi leaders criticized Turkey on Monday for bombing Kurdish militants in northern Iraq with airstrikes that they said had left at least one woman dead.

Event Code: 111  
Source: IRQ GOV  
Target: TUR

Event Code: 223  
Source: TUR  
Target: IRQKRD REB

Рис. 1.3 . Пример кодирования события (TABARI, PETRARCH)

Результатом обработки PETRARCH единицы информации (предложения) производится кодирование типа события, исходного и целевого акторов (в соответствии с CAMEO) и их местонахождения. Рис. 1.3 взят из опубликованной презентации инструментов в (Schrodt, 2014a). REB (англ., "rebel", рус. "мятежник") - код первичной роли, GOV (англ., "government", рус. "правительство") - код первичной роли. Первичные роли (primary roles) используются для внутригосударственных акторов. REB обозначает "вооруженные и агрессивные оппозиционные группы, действующие с нечеткими целями". TUR, IRQ и IRQKRD (спецификация страны и региона для данного актора) - геолокационные атрибуты акторов: Ирак (IRQ), Турция (TUR), Иракский Курдистан (IRQKRD). В доступной версии мануала CAMEO версии 1.1b3 2012 года код "IRQKRD" отсутствует; код "KUR" служит для обозначения соответствующей этнической группы. Событие 111 (критиковать или обвинять, англ. "Criticize or denounce") относится к категории 11: Disapprove (жалобы, возражения). Что касается события "bombing" (бомбардирование), в версии 1.1b3 нет упоминания события 223 в категории 20:

Engage in Unconventional Mass Violence (вовлечение в нетрадиционное массовое насилие), так как данное событие добавлено в расширения за период 2012-2015. PETRARCH последовательно сопоставляет группу глагола в каждом предложении с газетными онтологиями, и, в случае если аннотация акторов на основе онтологии не была произведена, аннотирует акторов, согласно результатам разметки парсера: исходный актор - в первой именной группе от начала предложения, целевой актор - в первой именной группе справа от глагола. В случае, если в правой части от глагола актор не обнаружен, принимается, что целевой актор располагается в левой от глагола именной группе.

### ***EAT***

Инструмент для автоматического аннотирования событий (Automated Event Annotation Tool или EAT) разработан Quang Xuan Do в группе когнитивных методов программирования Университета штата Иллинойс и встроен в SSP в 2014 году. EAT использует техники машинного обучения на базе десятков тысяч экспертных аннотаций типа события, даты, локации и акторов для построения когнитивных NLP-моделей для автоматизации разметки компонентов события в новых неструктурированных текстовых данных. Авторы отмечают, что данный инструмент позволяет избежать ошибок при отборе и дедупликации событий (CCD, 2014; Hayes & Nardulli, 2011). В изученных публикациях проекта SPEED мы не обнаружили детального описания алгоритмов машинного обучения и процедуры их применения.

### ***Serif***

Serif - инструмент, разработанный компанией BBN ("Болт, Беранек и Ньюман"), для автоматического высокоточного языконезависимого извлечения триплетов Сущность - Отношение - Сущность на основе онтологии. Система использует экспериментально обнаруженную эффективную комбинацию статистических алгоритмов для обучения моделей на базе экспертных аннотаций. Отличительной чертой подхода является использование промежуточного представления текста в форме пропозиции как альтернативы деревьям разбора, одновременное корректное построение которых для многих языков представляет в настоящее время большую сложность. Пропозиции формально описываются следующим образом: *predicate (role<sub>1</sub>:arg<sub>1</sub>, role<sub>2</sub>:arg<sub>2</sub> ... role<sub>n</sub>:arg<sub>n</sub>)*, где *predicate* - глагол или существительное (тип события). Оно отличается от стандартного

представления события *predicate* ( $arg_1, arg_2 \dots arg_n$ ) присвоением ролей аргументам. Из фразы "government was attacked by the party" ("партия атаковала правительство") модель узнает о том, что "партия" - это логический субъект пропозиции, а "правительство" - логический объект. Аргументы могут быть представлены не только именованными сущностями, но и другими пропозициями. Среди наиболее распространенных ролей можно выделить логический субъект, логический объект, премодификатор именной группы, модификатор глагола (дополнение предлога в предложной группе).

Serif использует комбинацию моделей обучения для повышения точности извлечения триплетов. Результат наиболее эффективной комбинации учитывается в выводе. Первая комбинация включает генеративную (generative) и признаковую (feature-based) модели. Генеративная модель рассчитывает вероятность того, что та или иная пропозиционная структура представляет определенное отношение путем вычисления совместной вероятности отношения и структуры. Соответствующие вероятности рассчитываются как сглаженная смесь оценок максимального правдоподобия. Модель вектора признаков представляет триплет Сущность-Отношение-Сущность, используя 5 признаков: предикат, обрамляющие его именные сущности и их синтаксические роли. Наиболее подходящий тип отношения определяется на основе метрики максимального правдоподобия. Следующая комбинация включает дискриминативные модели (перцептрон с голосованием) и учитывает взаимную зависимость признаков. Используется 11 признаковых шаблонов, составленных из атомарных признаков: 1) типы сущностей или их упоминания, 2) вид аргументно-предикатной структуры, 3) корень предиката, 4) строка между упоминаниями именных сущностей, 5) WordNet-синсет предиката, 6) словарный кластер предиката (Boschee et al., 2005). Признаковые шаблоны 4 и 5 позволили улучшить на 3.2% (метрика не указана) качество извлечения триплетов по сравнению с предыдущими комбинациями. Авторы не уточняют типы отношений, выбранные для экспериментов.

## 1.5 Выводы по главе

В главе 1 проанализированы подходы к извлечению политических событий (1.1, 1.2), и протестных событий как более специфических (1.3). Данные системы

применяют следующие технологии: релевантные тексты отбираются поисковым роботом, использующим эвристики на списках ключевых слов, и/или путем обучения моделей (напр., машины опорных векторов или наивного байесовского классификатора). Процедура кодирования базируется на просмотре словарей и сопоставлении с шаблонами (в PETRARCH) или вероятностных методах (Serif, проект А. Нанна) на уровне предложения или целого текста. С точки зрения исследования социального протеста, использование онтологии SSP более выгодно, так как она ориентирована на анализ отношений власть-общество, в отличие от CAMEO. Необходимо отметить следующие недостатки данных систем:

- невозможность обработки неанглийских текстов;
- необходимость больших объемов аннотированных данных для классификации;
- ограниченное описание событий кодерами (набор стандартных признаков включает исходного актора, целевого актора, триггер события, дату и локацию) и онтологиями (4 типа протестных событий в CAMEO);
- интенсивность событий в CAMEO измеряется на основе заранее установленных значений шкалы Гольдштейна (<http://web.pdx.edu/~kinsella/jgyscale.html>), изначально разработанной для онтологии WEIS. В силу специфики данная шкала подходит не для всех типов событий (в т. ч. связанных с протестами).

## ГЛАВА 2. МУЛЬТИЯЗЫЧНОЕ ИЗВЛЕЧЕНИЕ СОБЫТИЙ

### 2.1 Постановка задачи извлечения событий

В рамках мультязычного извлечения (социополитических или иных) событий выполняются в общем виде три основные подзадачи: (i) отбор, фильтрация, форматирование и занесение в базу подмножества текстов  $T_d [t_{d1}, t_{d2}, \dots, t_{dn}]$ , относящихся к заданной предметной области, множества мультязычных текстов  $T [t_1, t_2, \dots, t_n]$ , извлеченного из новостных и других источников; (ii) аннотация событийных данных, таких как тип, форма, локация, тип участника, количество участников, и пр. в каждом элементе  $T_d [t_{d1}, t_{d2}, \dots, t_{dn}]$  и сбор сценариев; (iii) слияние мультязычных сценариев (создание единой записи на английском языке) и помещение записей в базу данных.

### 2.2 Системы мультязычного извлечения событий

Современные системы мультязычного извлечения событий преимущественно модульные и задействуют множество готовых технологий, априори мультязычных. Обработка новостных текстов на разных языках в целях извлечения событий - это ключевая функция, позволяющая скорректировать сценарий события и таким образом получить более объективные данные (Steinberger, 2012). Кроме того, объединение мультязычных сценариев улучшает качество извлечения событий, согласно (Piskorski et al., 2011). Системы мониторинга событий в последние годы расширяют языковое покрытие, чтобы отслеживать возникновение искомых событий в целевых локациях, где данные публикуются исключительно на местных языках. Глобальная система перевода Google Translate также задействуется для приведения текста на любом языке к английскому варианту, однако, перевод не работает одинаково хорошо на разных направлениях. В 90-е годы уже предпринимались попытки использовать нейтральное языковое представление мультязычных текстов или *интерлингва* для генерации сценариев событий в таких системах, как MURASAKI (Aone et al., 1994), MIETTA (1998-2000) (Buitelaar et al, 1998) и др., однако на тот момент качество машинного перевода не позволяло получить высокоточный результат.

На настоящий момент системы мультязычного извлечения событий обрабатывают различные источники информации, снабженные или не снабженные метаданными: электронные СМИ (новости, блоги, твиттер), отчеты радиотехнической и агентурной разведки, библиотеку Wikipedia, персональные данные пользователей (профили в социальных сетях, CV и иные документы), предметно-специфические страницы сети (для рекомендательных систем и др.) и пр.. Коммерческая система NetOwl (<http://netowl.com>) отслеживает киберпреступления на 7 языках. Проект Biographe (<http://biographe.org>) извлекает факты о людях из профилей в социальных сетях и других данных на 4 языках (английский, датский, немецкий, французский). Проект MULTISENSOR (Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation), находящийся в разработке, создан для поддержки принятия бизнес-решений в разных областях знания на основе анализа мультязычного информационного контента (ТВ, радио, WWW) на 4-х языках (английский, французский, испанский, немецкий). Основные технологии мультязычного извлечения событий в MULTISENSOR аналогичны изложенным далее: пайплайн (от англ. "*pipeline*" - цепочка процессов обработки) распознавания сущностей, распознавание концептов на основе множества лексических ресурсов (BabelNet<sup>1</sup>, Unified Verb Index<sup>2</sup>, Eurovoc<sup>3</sup>, Reegle Glossary<sup>4</sup>), пайплайн для извлечения отношений между концептами (глубокий разбор зависимостей и аннотирование семантических ролей). Распознавание именных сущностей (имплементация в Java) идет в несколько этапов: сегментация текста на предложения, токенизация и собственно распознавание сущностей (локальный парсер, текстовый анализатор (разрешение кореференции) и генератор вывода). Для аннотации именных сущностей используется сетевой инструмент Brat<sup>5</sup> (Aleksić et al., 2014).

Наиболее распространенный пайплайн систем мультязычного извлечения событий представлен на Рис. 2.1. В целом, он аналогичен пайплайну ранее рассмотренных систем извлечения социополитических событий. Информация (заголовок, основной текст статьи, дата публикации, автор, и т.д.) из новостных источников (**News Feeds**) собирается поисковыми роботами (**Crawlers**) и подается в базу данных (**Storage**). Далее тексты статей проходят предобработку (**Pre-processing**). Тип события (**Event Type Detection**) обнаруживается на основе

<sup>1</sup> <http://babelnet.org/search.jsp>

<sup>2</sup> <http://verbs.colorado.edu/verb-index/search.php>

<sup>3</sup> <http://eurovoc.europa.eu/drupal/>

<sup>4</sup> <http://www.reegle.info/glossary>

<sup>5</sup> <http://brat.nlplab.org/index.html>

газетиров (**Gazetteers**), связанных с базой знаний (**Ontology**). Затем производится заполнение слотов сценария (**Template Filling**). Наиболее типичные слоты включают

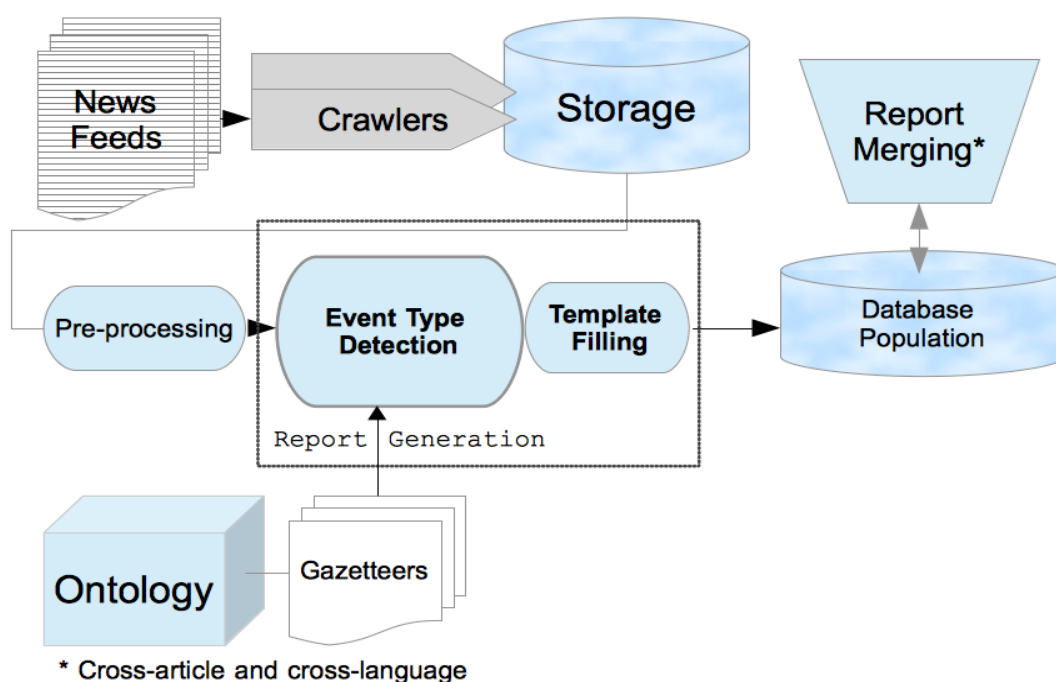


Рис. 2.1. Общий пайплайн мультязычного мониторинга событий.

количество участников, наименование участников, участвующие организации, количество пострадавших, наименование пострадавших, географическое местоположение с точностью до широты и долготы, время и дату, указанные в основном тексте, оружие, и пр., в зависимости от типа события.

В данном разделе подробно рассмотрены следующие системы извлечения событий: NEXUS, NewsReader, PULS, Xlike, ZENON. Выбор систем обусловлен тем, что основные детали архитектуры находятся в свободном доступе. Общее описание

систем показано в Табл. 2.1: языковое покрытие, предметная область, источники данных и единица анализа.

Таблица 2.1 Общее описание систем мультязычного мониторинга событий.

| Проект             | NEXUS                                                                                               | NewsReader                                           | PULS                                                                           | Xlike                                                                                                               | ZENON                                   |
|--------------------|-----------------------------------------------------------------------------------------------------|------------------------------------------------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|-----------------------------------------|
| Языковое покрытие  | английский,<br>итальянский,<br>испанский,<br>французский,<br>португальский,<br>русский,<br>арабский | английский,<br>датский,<br>итальянский,<br>испанский | английский,<br>французский,<br>русский                                         | английский,<br>немецкий,<br>испанский,<br>китайский,<br>французский,<br>итальянский,<br>португальский,<br>, датский | английский,<br>дари,<br>таджикский      |
| Предметная область | нелегальная деятельность на пограничных территориях Евросоюза, стихийные бедствия                   | экономика и финансы                                  | угрозы эпидемий, нелегальная деятельность на пограничных территориях Евросоюза | общая                                                                                                               | взаимоотношения в военных частях        |
| Источники данных   | ленты новостей                                                                                      | ленты новостей                                       | ленты новостей                                                                 | ленты новостей, блоги                                                                                               | корпус отчетов KFOR                     |
| Единица анализа    | заголовок + первые 1-2 предложения центроида кластера                                               | полный текст статьи                                  | полный текст статьи                                                            | цетроид кластера                                                                                                    | полный текст отчета агентурной разведки |

## NEXUS

NEXUS (Atkinson et al., 2011; Piskorski et al., 2011), проект Объединенного исследовательского центра при Комиссии Евросоюза (1999 - наст. вр.), - это

инструмент, разработанный для Frontex (Европейская служба пограничного контроля), с целью наблюдения за событиями на границах Евросоюза и в некоторых странах Третьего мира и интеллектуального анализа рисков. Алгоритм извлечения событий был разработан в сотрудничестве с проектом PULS, который описан подробно далее в настоящем разделе.

Системы NEXUS и PULS отличаются языковым покрытием и тем, как они "видят" событие. Они дополняют друг друга, так как NEXUS производит поверхностный анализ центроидов кластеров, а PULS извлекает события на основе глубокого синтаксического разбора, что позволяет выделить группы частично пересекающихся событий. NEXUS охватывает события, относящиеся к нелегальной миграции, пограничной преступности и стихийным бедствиям на границах Евросоюза и в странах Третьего мира. Активное языковое покрытие включает английский, итальянский, испанский, французский, португальский, русский и арабский языки. По словам авторов, адаптация системы к новому языку занимает около 3-х месяцев. На вход NEXUS и PULS подаются коллекции новостей, собранные новостным агрегатором EMM (European Media Monitor: <http://press.jrc.it>) и кластеризованные в соответствии с их содержанием, при этом каждый кластер геолоцируется и получает метатеги. Кластеры статей, относящихся к предметной области, выделены с использованием эвристик на базе ключевых слов. В (Atkinson et al., 2011) указаны следующие составляющие предобработки EMM и их функции:

- Поисковой робот (проверка обновлений RSS);
- Grabber: создание нового RSS-канала для каждого вебсайта, содержащего заголовок, ссылку, описание и основной текст статей;
- Определитель языка: идентификация языка статьи на базе автоматически пополняемых таблиц частот слов;
- Распознавание именных сущностей: обнаружение людей и организаций с использованием собственной автоматически пополняемой мультязычной информационной базы;
- Геокодирование: определение местоположения на базе GeoNames и собственного газетира с предшествующим и последующим снятием омонимии;
- Тематическая категоризация: автоматическая классификация

документов по темам на основе лингвоспецифических регулярных выражений и булевой логики;

- Фильтрация по мета-тегам: фильтрация статей с использованием булевых комбинаций мета-тегов;
- Кластеризация: иерархическая аггломеративная кластеризация методом средней групповой связи; вектор признаков строится на основе представления текста как "мешка слов" и некоторых правил (игнорировать 100 самых частотных слов длиной в два символа);

В результате предварительной обработки основной коллекции системой ЕММ выявлены кластеры статей, относящиеся к тематике Frontex. Данные статьи подаются на вход NEXUS и PULS для извлечения сценариев событий.

NEXUS ориентирован на сканирование заголовков и первых 1-2 предложений из каждой статьи-центроида кластера и создание одного сценария события на кластер. С этой целью система производит поверхностный парсинг, обработку каскадом грамматик на базе конечных автоматов и слияние событий на уровне кластеров. Для поверхностного парсинга используется пакет инструментов собственной разработки CORLEONE (Core Linguistic Entity Online Extraction), включающий токенайзер, сплиттер предложений, просмотр словаря (численные значения, квантификаторы, формы обращения), алгоритм извлечения неименованных групп лиц и морфологический анализатор. Каскадные грамматики построены и скомпилированы в системе собственной разработки ExPRESS (Extraction Pattern Recognition Engine and Specification Suite). Шаблоны представляют собой регулярные выражения с использованием плоских признаков структур (нерекурсивные со строковыми атрибутами). Целевой сценарий содержит следующие слоты: TYPE (тип), SUBTYPE (подтип), SNIPPET (фрагмент текста, содержащий триггер), PUB\_DATE (дата публикации), DATE (указание даты в основном тексте), LOCATION (локация), CONFIDENCE (автоматически определяемая уверенность, с которой система выделяет событие), RELEVANCE (автоматически определяемая релевантность события для конечного пользователя), SEVERITY (тяжесть события, в зависимости от наличия жертв, пострадавших, и пр.), SOURCE (источник), PERPETRATOR (правонарушитель), VICTIM (жертвы), ITEM (продукция, провезенная контрабандным путем), MEANS (транспорт, на

котором была пересечена граница), NUM\_AFFECTED (количество потерпевших), NUM\_INJURED (количество раненых), NUM\_ARRESTED (количество арестованных), WOMEN/MINORS\_INVOLVED (участие женщин и детей) и т.п.. ExPRESS создан по аналогии с JAPE (GATE язык шаблонов на основе Java) и XTDL, формализмом платформы SProUT (<http://sprout.dfki.de/>), основанных на схожих принципах. Правило состоит из левой (LHS) и правой частей (RHS). В LHS в линейном порядке, в соответствии с расположением элементов в тексте, указаны ограничения по шаблонам. В RHS содержатся манипуляции на аннотациях. Шаблоны NEXUS включают 1/2 слота (семантические роли со строковыми атрибутами) и соответствуют простым синтаксическим конструкциям.

**NewsReader** (<http://www.newsreader-project.eu/>) (2013-2016) - готовящийся к выпуску инструмент для мультязычного извлечения, мониторинга и анализа событий в предметной области "экономика и финансы". Система обрабатывает текст на 4 языках (английский, датский, итальянский и испанский), используя обученные и готовые средства предобработки и мультязычные базы знаний. Пайплайн предобработки включает следующие этапы: сегментация на предложения, токенизация, тегирование частей речи, лемматизация, распознавание сущностей, синтаксический разбор и разрешение кореференции. Для испанского и английского в целях предобработки используется готовый пайплайн IXA (на базе Java) (<https://github.com/ixa-ehu>); для датского используется готовый парсер Alpino (<http://www.let.rug.nl/vannoord/alp/Alpino/>), компоненты для итальянского языка обучены на корпусе текстов. В рамках NewsReader также производится аннотирование семантических ролей (для английского и испанского), привязывание именных сущностей к концептам в Wikipedia и DBpedia (<http://dbpedia.org>), аннотирование временных отношений между событиями с использованием разработки Джеймса Пустейовского TimeML (Pustejovsky et al., 2005), классификация событий, анализ мнений, дискурса и *действительности (Factuality)* фактов. Упомянутые компоненты высокоуровневой обработки в полной мере реализованы только для английского языка, что является частым случаем в подобных системах. Поддержка остальных языков достаточно ограничена из-за недостатка размеченных данных и готовых ресурсов. *Factuality* - это индикатор того, имело ли в действительности место данное событие, рассчитываемый на основе текстовых и структурных индикаторов (модальные глаголы и др.). Кроме того,

авторы вводят индикаторы для авторитетности статьи (*Authority*) и степени доверия к ней (*Trust*), измеряемые на основе метаданных страницы и стилистических особенностей текста.

Ключевая стратегия извлечения события в NewsReader состоит в интеграции знаний из банка синонимов WordNet и банков английских предикатов VerbNet (и аналогичных баз, таких как FrameNet и PropBank) (<http://verbs.colorado.edu/semlink/>) для разметки семантических ролей предикатов. Однако адаптация банков предикатов к новым языкам требует значительных временных и экспертных ресурсов. В условиях отсутствия разметки семантических ролей, датский и итальянский модули полагаются на данные синтаксического разбора.

NewsReader использует систему аннотирования NLP Annotation Framework (NAF) (<https://github.com/ixa-ehu/NAF>), которая является потомком KYOTO/Knowledge Annotation Framework (KAF) (<http://kyoto-project.eu>), инструмента для мультязычного семантического аннотирования. В эти системы по умолчанию встроены мультязычные базы знаний, такие как WordNet, онтология (KYOTO-DOLCE), и поддержка генерации фактов. Поиск события производится на уровне предложения на базе трех классов предикатов ("Коммуникация", "Познание" и "Другое") с соответствующими ролями. В класс "Другое" входят предметно-специфические действия. Семантические фреймы предикатов (карты семантических ролей) берутся из *матрицы предикатов*, специального ресурса, интегрирующего предикатные данные из FrameNet, VerbNet, PropBank и WordNet (Agerri et al., 2013).

## **PULS**

Система **PULS** (Pattern Understanding and Learning System) разрабатывается в Университете Хельсинки (2006 - наст.вр.) (<http://puls.cs.helsinki.fi/static/index.html>). Изначально она создавалась для поддержки эпидемиологического контроля. В отличие от NEXUS, данная система не является модульной, то есть языкозависимые модули не встроены в один общий пайплайн. PULS состоит из 3 отдельных компонентов для обработки английского, русского и французского текстов. В то время как английский и русский компоненты имеют аналогичное строение и опираются на лексические ресурсы (доменная онтология и газетиры) и шаблоны (Pivovarova, 2013), французский компонент использует робастные правила, основанные на особенностях дискурса, и небольшие газетиры локаций и

наименований болезней (Lejeune, 2009). Наиболее завершенным в системе является английский компонент; русский компонент интенсивно развивался в последние годы. Согласно (Du et al., 2011), архитектура PULS аналогична GATE (General Architecture for Text Engineering: <http://gate.ac.uk>) во многих отношениях. Она представляет собой многоступенчатый пайплайн, включающий токенизацию, обработку пунктуации, просмотр лексических ресурсов, морфологический анализ, поверхностный синтаксический анализ, аннотацию именованных сущностей, разрешение анафоры для полных текстов статей, сопоставление с шаблонами, логический вывод, заполнение сценария, нормализацию и генерацию вывода, где работа составляющих каждого уровня базируется на результатах обработки на предыдущих уровнях.

В связи с тем, что PULS рассматривает каждое предложение как потенциальное событие, необходим коэффициент релевантности для фильтрации посторонних данных. Релевантность предсказывается алгоритмами машинного обучения (SVM и наивным байесовским классификатором) на основе предопределенных признаков, таких как (i) относительная позиция предложения в тексте, (ii) отличие между датой публикации и датой события, указанной в основном тексте статьи, (iii) длина документа, (iv) лексические признаки. Стоит отметить, что система различает одиночные и периодические события, а также нераспознанные болезни.

Русский и английский модули PULS применяют стандартные каскадные грамматики на основе конечных автоматов и словари на базе онтологии. В целях мониторинга вспышек заболеваний строками из текста заполняются следующие информационные слоты: EVENT TYPE (наименование болезни или типа пограничного преступления), EVENT\_DATE (дата события), LOCATION (наименование населенного пункта), VICTIM\_DESCRIPTOR (наименование пострадавших), VICTIM\_COUNT (число пострадавших), VICTIM\_STATUS (статус пострадавших: инфицированы и пр.), VICTIM\_TYPE (тип пострадавших: человек, животное, растение). Кроме того, система добавляет к полученному описанию номер документа (DOCUMENT\_NUMBER), нормализованную дату (NORMALIZED\_DATE), порождающее событие (PARENT\_INCIDENT) и тип включения (INCLUSION\_TYPE). Слоты PARENT\_INCIDENT и INCLUSION\_TYPE

служат для восстановления данных о прецедентах (Grishman et al., 2003). В целях мониторинга пограничных преступлений заполняются следующие информационные слоты: DATE (дата), LOCATION (местонахождение), ITEM (предмет контрабанды), VICTIM\_DESCRIPTOR (описание жертвы), PERPETRATOR\_DESCRIPTOR (описание нарушителя). Эпидемиологическая онтология содержит около 4000 концептов; онтология предметной области "Безопасность пограничной территории ЕС" включает 1190 концептов (Pivovarova, 2013; Du et al., 2011).

Шаблоны PULS включают следующие комбинации на синтаксическом уровне: сказуемое/дополнение и сказуемое/подлежащее. Шаблоны русскоязычного компонента соответствуют пяти синтаксическим конструкциям. Морфологический и синтаксический анализ выполняются адаптированными инструментами AOT (Automatic Processing of Text: <http://aot.ru>). Заполнение сценария запускается при ассоциации шаблона с концептом в доменной онтологии. Такие свойства русского языка, как флективность и свободный порядок слов, составляют основную трудность при конструировании шаблонов (из обзора состояние дел в области в (Danilova & Popova, 2014)).

### **Xlike**

**Xlike** (2012-2014) (<http://xlike.org>) - система, генерирующая данные о событиях из новостей и блогов на английском, испанском, каталанском, китайском, немецком и словенском языках с использованием агрегатора Newsfeed ([newsfeed.ijs.si](http://newsfeed.ijs.si)) и свободной таксономии веб сайтов dmoz (<http://dmoz.org>). Основные категории таксономии отображены на Рис. 2.2. Данная база постоянно пополняется методом краудсорсинга и включает огромное количество вложенных категории из всевозможных областей знания, в связи с чем разработчики Xlike используют только 3 уровня вложений (около 1000 категорий).

```
<Topic r:id=""> <catid>1</catid>

<d:Title></d:Title>
<lastUpdate>2010-03-24 01:16:43</lastUpdate>
<d:Description></d:Description>
<narrow r:resource="Top"></narrow>
<narrow r:resource="Kids_and_Teens"></narrow>
```

```
<narrow r:resource="Bookmarks"></narrow>
<narrow r:resource="Test"></narrow>
<narrow r:resource="Adult"></narrow>
</Topic>
<Topic r:id="Top">
  <catid>2</catid>
  <d:Title>Top</d:Title>
  <lastUpdate>2010-02-16 08:43:34</lastUpdate>
  <d:Description></d:Description>
  <narrow r:resource="Top/Shopping"></narrow>
  <narrow r:resource="Top/Society"></narrow>
  <narrow r:resource="Top/News"></narrow>
  <narrow r:resource="Top/Science"></narrow>
  <narrow r:resource="Top/Business"></narrow>
  <narrow r:resource="Top/Health"></narrow>
  <narrow r:resource="Top/AOL"></narrow>
  <narrow r:resource="Top/Computers"></narrow>
  <narrow r:resource="Top/Home"></narrow>
  <narrow r:resource="Top/Sports"></narrow>
  <narrow r:resource="Top/Arts"></narrow>
  <narrow r:resource="Top/Regional"></narrow>
  <narrow r:resource="Top/Netscape"></narrow>
  <narrow r:resource="Top/Reference"></narrow>
  <narrow r:resource="Top/Recreation"></narrow>
  <narrow r:resource="Top/Games"></narrow>
  <narrow r:resource="Top/World"></narrow>
</Topic>
```

Рис. 2.2. Фрагмент веб-таксономии dmoz

Xlike кластеризует статьи на основе ключевых слов и затем производит кросс-языковое ассоциирование документов путем латентно-семантического индексирования (Latent Semantic Indexing) и обобщенной версии канонического корреляционного анализа (Canonical Correlation Analysis) на основе параллельного

мультиязычного корпуса. Данная техника также используется в кросс-языковом обнаружении плагиата (Danilova, 2013). В системе Xlike кросс-языковое сопоставление кластеров событий предшествует помещению в базу данных (реестр событий) в отличие от таких систем, где сценарии сравниваются внутри базы данных (напр., NEXUS).

Пайплайн системы включает следующие этапы: (i) предобработка (аннотирование семантических ролей, извлечение дат, кросс-языковое сопоставление, удаление дубликатов); (ii) формирование события (кластеризация статей, кросс-языковое сопоставление кластеров, заполнение слотов сценария, распознавание связанных событий); (iii) помещение события в базу данных.

На уровне кластера собираются только самые общие данные о событии: заголовок и первый параграф статьи-центроида, дата (наиболее частотное из упоминаний в кластере события), местонахождение (в соответствии с GeoNames), категория в соответствии с таксономией, основные именованные сущности и ключевые слова (наиболее частотные термины). Xlike разграничивает сущности и ключевые слова, используя данные об отношениях из инфобоксов библиотеки Википедия. Связанные события идентифицируются путем сравнения описывающих их концептов. Для этого каждое событие представляется как набор слов с весами TF-IDF, и сходство между векторами вычисляется с помощью коэффициента Охай (Leban and Brank, 2011).

Поверхностный лингвистический анализ для каждого из языков включает определение языка, сегментацию на предложения, токенизацию, лемматизацию, тегирование частей речи и/или морфологическое описание (обогащение аннотации морфологическими признаками), распознавание сущностей, преобразование из .conll в .xml. Для обработки английского, испанского и каталанского языков используется библиотека Freeling (Padró and Stanilovsky, 2012). Модуль немецкого языка использует Apache OpenNLP для токенизации и разметки границ предложений, Stanford PoS tagger для тегирования частей речи, TreeTagger для лемматизации и Stanford NER для распознавания именованных сущностей. Модуль словенского языка использует лемматизатор и морфологический теггер, обученные на корпусе текстов, и классификатор именованных сущностей на основе условных случайных полей. Модель обучена с использованием Mallet Toolkit на корпусе из 2173

размеченных предложений. Модуль китайского языка состоит из токенайзера, теггера частей речи и классификатора именованных сущностей (лиц, локаций и названий организаций) на основе алгоритма скрытых моделей Маркова (Tadić et al., 2011).

Глубокий синтаксический и семантический разбор производится на этапе предобработки (семантическая аннотация статьи). Он основан на использовании обученных моделей для синтаксического анализа и извлечения семантических фреймов. Обучение синтаксических парсеров осуществляется на основе алгоритмов библиотек анализа зависимостей (e.g., **Treeler**: <http://treeler.lsi.upc.edu>) и других банков. Для некоторых языков ресурсы очень ограничены: напр., банк деревьев зависимостей для словенского языка содержит только 10 типов зависимостей. Аннотаторы семантических ролей также опираются на модели, построенные на основе Treeler. Классификаторы первого типа разграничивают предикаты и непредикаты, а классификаторы второго типа предсказывают аргументную структуру, в зависимости от типа предиката. Эта задача не была выполнена для словенского языка в силу нехватки ресурсов. Конструктор семантической рамки действует по следующей схеме: каждый аргумент (аннотация семантической роли в тексте) предиката ассоциируется при помощи индекса VerbIndex (<http://verbs.colorado.edu/verb-index>) с соответствующим компонентом семантической рамки на основе данных о семантике предиката (WordNet) и его семантических фреймах (VerbNet). На данный момент в отчетах по проекту Xlike недостаточно данных об апробации используемого метода. Система функционирует как онлайн-сервис по мониторингу событий (<http://eventregistry.org>).

Система **ZENON** (2001-2011) была разработана для поддержки анализа отчетов агентурной разведки. Предметная область покрывает конфликты (между этническими группами, политическими сторонами и др.), инфраструктурные проблемы и человеческие взаимоотношения (браки, совещания, встречи и пр.) в областях размещения военных частей Германии в Афганистане. Система обрабатывает тексты отчетов на английском, дари и таджикском языках, входящих в состав корпуса KFOR (Hecking & Sarmina-Baneviciene, 2010). Лингвистическая обработка включает, помимо стандартной предобработки, морфологический анализ, тегирование частей речи и выделение группы глагола. В условиях отсутствия

обучающих корпусов для дари и таджикского, авторы системы разработали соответствующие программные средства и газетиры именных существительных и интегрировали их в пайплайн GATE.

Система использует следующий пайплайн. Тексты отчетов подаются на вход цепочки модулей GATE, и производится разметка групп существительного и глагола. Авторы принимают, что группа глагола определяет тип события. Далее, на основе полученной разметки именных существительных, группы глагола и тегов частей речи, FrameNet (<http://framenet.icsi.berkeley.edu/fndrupal/>) аннотирует составляющие семантических фреймов. Данная функция не реализована для дари и таджикского по причине отсутствия ресурсов. Планировалось объединение языковых представлений концептов с использованием онтологии собственной разработки, однако, данные об этой онтологии не были опубликованы. В статьях упоминаются только некоторые типы событий: напр. REPORT (сообщать), KNOW (знать), COMMAND (давать распоряжение), PROPOSE (предлагать), EXPLODE (взрываться), и др.. Авторы не уточняют, какие именно сценарии использованы, и каким образом производится слияние сценариев, описывающих одно и то же событие. Полученные данные преобразуются с помощью языка XSLT (Extensible Stylesheet Language Transformation) (<http://www.w3schools.com/xsl/>) и Системы представления извлеченной информации (IEPS) собственной разработки (Elvira & Hecking, 2005).

Согласно изложенной информации, ключевые характеристики систем мультязычного извлечения событий резюмированы в Табл. 2.2, 2.3, и 2.4.

Табл. 2.2. Базы знаний для геолоцирования и определения триггера события

| Базы знаний |                              |                 |                     |       |                                              |
|-------------|------------------------------|-----------------|---------------------|-------|----------------------------------------------|
| Проект      | NEXUS                        | NewsReader      | PULS                | Xlike | ZENON                                        |
| Онтология   | собст. иерархия<br>концептов | KYOTO-<br>DOLCE | собст.<br>онтологии | Dmoz  | собст.<br>онтология                          |
| Словари     | собст.<br>газетиры           | N/A             | собст.<br>газетиры  | N/A   | собст. и<br>встроенные<br>газетиры<br>(GATE) |

|                |                                                   |                         |                               |                     |                                              |
|----------------|---------------------------------------------------|-------------------------|-------------------------------|---------------------|----------------------------------------------|
| База локаций   | GeoNames и<br>собст. газетир<br>(ок. 600 000 сл.) | GeoNames,<br>GoogleMaps | GeoNames и<br>собст. газетиры | GeoNames            | собст. и<br>встроенные<br>газетиры<br>(GATE) |
| Другие ресурсы | -                                                 | WordNet,<br>SemLink     | -                             | WordNet,<br>VerbNet | FrameNet                                     |

Табл. 2.3. Программное обеспечение для сбора данных, лингвистической  
предобработки и конструирования шаблонов

| Инструменты                      |                                  |             |                                           |                                                                                               |       |
|----------------------------------|----------------------------------|-------------|-------------------------------------------|-----------------------------------------------------------------------------------------------|-------|
| Проект                           | NEXUS                            | NewsReader  | PULS                                      | Xlike                                                                                         | ZENON |
| Сбор данных                      | EMM                              | LexisNexis  | EMM                                       | Newsfeed                                                                                      | -     |
| Лингвистическая<br>предобработка | собст.<br>разработка<br>CORLEONE | IXA, Alpino | AOT, Proteus                              | Freeling, Apache<br>OpenNLP, Stanford<br>NLP, TreeTagger,<br>Mallet toolkit,<br>Treeler, etc. | GATE  |
| Язык шаблонов                    | собст.<br>разработка<br>ExPRESS  | NAF         | собст.<br>разработка (на<br>базе Proteus) | -                                                                                             | JAPE  |

Табл. 2.4. Алгоритмы пополнения и слияния сценариев

| Ключевые алгоритмы               |                                                                                                                                      |                                                                 |                                                                    |                                                                            |                                     |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------------------|----------------------------------------------------------------------------|-------------------------------------|
| Проекты                          | NEXUS                                                                                                                                | NewsReader                                                      | PULS                                                               | Xlike                                                                      | ZENON                               |
| конструирование шаблонов         | слабо контролируемый бустреп                                                                                                         | -                                                               | частично контролируемый бустреп (английский),<br>вручную (русский) | -                                                                          | конструирование вручную             |
| Обнаружение типа события         | кластеризация, сопоставление с шаблонами                                                                                             | обнаружение предикатов на базе онтологии                        | сопоставление с шаблонами                                          | кластеризация*, обнаружение предикатов на базе онтологии                   | аннотирование группы глагола        |
| Заполнение слотов                | сопоставление с шаблонами                                                                                                            | создание семантических фреймов                                  | сопоставление с шаблонами                                          | мэппинг семантических фреймов                                              | мэппинг семантических фреймов       |
| кросс-языковое слияние сценариев | машинный перевод сценариев и удаление дубликатов: разрешение омонимии семантических ролей, подсчет жертв, классификация типа события | разрешение кореференции (контролируемая попарная классификация) | машинный перевод сценариев и удаление дубликатов                   | латентно-семантическое индексирование и канонический корреляционный анализ | объединение в онтологии, XSLT, IEPS |

\* иерархический дивизимный (bisecting) *k*-means

### 2.3 Выводы по главе

Алгоритмы рассмотренных систем извлечения событий основываются на использовании шаблонов, построенных вручную (и методом бустрепа), и лексико-семантических баз (онтологий, баз предикатов, словарей, газетиров и пр.). В

наиболее современных системах есть тенденция к обогащению аннотаций информацией из множества баз знаний, обучению более абстрактных моделей представления текста, основанных на лексико-семантических базах, использованию кластеризации на разных этапах обработки. Можно выделить 4 основных этапа обработки текста в таких системах: (1) поиск и сбор информации; (2) предобработка (форматирование, кластеризация, и/или многоступенчатая лингвистическая обработка, начинающаяся токенизацией и завершающаяся классификацией именованных сущностей и отношений); (3) извлечение события (триггера и аргументов) на основе онтологии с использованием грамматик, мэппинга семантических фреймов и/или техник машинного обучения (в случае доступности банка соответствующей информации); (4) склеивание событий и сохранение финального варианта записи.

В данной главе описываются подходы и системы мультязычного извлечения событий. Инструменты мониторинга мультязычных ресурсов - это также многоуровневые и модульные системы. Функции некоторых из них включают не только извлечение сценария события, но также и интеллектуальное отслеживание истории и взаимоотношений с другими событиями. На каждом уровне обработки признаки аккумулируются и используются в дальнейшем грамматиками и конструкторами семантических фреймов. NEXUS и PULS представлены как дополняющие друг друга системы, однако, их языковое и предметное покрытие не совпадают. PULS способен анализировать пограничную преступность только на русском и английском языках, французский модуль ориентирован на мониторинг эпидемиологических угроз. NEXUS обладает широким языковым покрытием, однако, согласно опубликованным оценкам, он справляется с пограничными преступлениями хуже, чем со стихийными бедствиями. Данная система переходит на использование более абстрактных шаблонов, чтобы избавиться от ошибок при парсинге сложных глагольных структур, вложенных конструкций и пр., что должно значительно улучшить показатели точности. Xlike ориентирован на пополнение глобального реестра событий. Главный недостаток системы, по словам авторов, - это используемая веб-таксономия, которая по сути является коллекцией разбитых по категориям вебсайтов и не подходит для классификации событий. Кроме того, записи в реестре снабжают пользователя только самыми общими данными о событиях.

По сравнению с системами извлечения социополитических событий, системы многоязычного мониторинга СМИ генерируют более детальные сценарии и осуществляют качественное слияние данных, что исключает появление дубликатов в итоговой базе. Можно выделить следующие недостатки данных систем:

- зависимость от множества внешних инструментов;
- невозможность обучения вследствие нерепрезентативности или недоступности банков обучающих данных для малых языков;
- недостаточное покрытие социополитической тематики;
- закрытый доступ к технологиям;
- невозможность настройки.

В рамках исследования разработан мультязычный пайплайн, шаблоны, вспомогательные программные средства и лексико-семантические ресурсы для отбора, извлечения, кодирования и структурирования протестных данных из новостных источников. Кодирование традиционно включает этапы аннотации и присвоения кода событию на основе кодовой книги. В данной работе реализован этап аннотации значимых признаков протеста, а в качестве кода событию присваивается имя класса онтологии, и таким образом объединяются в одно многообразные языковые представления. Процесс обработки новостей состоит в следующем. Новостные статьи собираются краулерами из множества источников на 6-ти языках (испанский, французский, польский, русский, болгарский и шведский), форматируются, фильтруются и тегируются (шведский) для дальнейшей обработки в системе GATE. Вывод производится в CSV-файл. Кроме того, специальная грамматика пополняет неизвестными значениями слотов соответствующие классы онтологии для облегчения дальнейшего аннотирования. Преимущество данной системы в следующем:

- обнаружение протестного события на уровне предложения (заголовка) и обогащение признаками;
- определение масштабов события (по таким текстовым индикаторам как размах, длительность, применение насилия, итеративный характер, интенсивность);

- наличие онтологии протестных событий;
- применение мультязычных шаблонов (нет необходимости в аннотированных данных для обучения);
- быстрая обработка.

Построение прототипа протестной онтологии первоначально опиралось только на реальные данные, однако, впоследствии структура была улучшена на основе опыта других онтологий. Что касается среды аннотирования, выбор был сделан в пользу GATE по ряду причин: (i) это open source; (ii) прозрачная архитектура, т.к. большая часть кода находится в свободном доступе; (iii) обеспечивает быструю обработку каскадных грамматик при расширении объемов памяти; (iv) комбинирует поверхностную и глубокую текстовую обработку; (v) позволяет подключать внешние инструменты и базы знаний.

В данной имплементации система обрабатывает заголовки, обычно состоящие из одного-двух предложений, поэтому нет необходимости применять алгоритмы разрешения кореференции. Кроме того, настоящая версия не включает слияние событий и формирование реестра, т. к. целью работы является не построение полноценной системы с нуля, а апробация отдельных модулей для параллельного мультязычного парсинга событий.

## ГЛАВА 3. ДИНАМИЧЕСКОЕ АННОТИРОВАНИЕ ТЕКСТОВ

### 3.1 Постановка задачи и обзор работ

Динамика событий, отражающих различные социально-экономические и социально-политические процессы, является объектом изучения экономистов, социологов и политиков. Изучение такой динамики предполагает просмотр множества материалов, связанных с изучаемым событием и распределенных во времени. Эти усилия могут быть существенно облегчены, если исследователь имеет начальные знания (информацию) относительно рассматриваемого события. В таком случае, ему следует только представить некоторые маркеры, отражающие динамику события, по которым он смог бы представить себе процесс в целом. Именно для такого случая предлагается метод, основанный на так называемых «динамических словарях». Этими словарями являются списки слов, извлеченные из корпуса документов и скомпонованные определенным образом.

Динамические словари были предложены в работах [Alexandrov, 2001; Makagonov, 2006]. В настоящей работе предложена реализация динамических словарей более простая для интерпретации. Здесь также представлены программные средства, позволяющие создавать такие словари конечному пользователю. Первая версия динамических словарей отражена в [Береснева, 2015].

Представление динамики событий в форме последовательности словарей может быть полезным для специалистов уже достаточно знакомым с изучаемым явлением. Они могут дополнить такое представление своими знаниями о дополнительных характеристиках персон и организаций, отраженных в словарях, и их действиях в связи с изучаемыми событиями. Они также могут дополнить словари своими знаниями о тональности публикаций, где отражены эти события – позитивными или негативными. Таким образом, динамические словари – это инструмент для напоминания специалисту о динамике уже прошедших событий. Однако, пользователю, слабо знакомому с предметом рассмотрения, гораздо удобнее читать литературно связанный текст. Такой текст может быть представлен в форме динамических аннотаций. Такие динамические аннотации представляют собой предложения, обогащенные словами из динамических словарей и относящиеся к определенному интервалу времени. Таким образом, для представления динамики

событий необходимо иметь не одну аннотацию, а их последовательность. Собственно, эту последовательность мы называем динамическими аннотациями. При этом могут быть построены аннотации, учитывающие только вновь появившиеся на данном временном интервале события, или новые и повторяющиеся с предыдущего временного шага события. Все определяется соответствующими динамическими словарями.

Проблема аннотирования не является новой. Аннотирование, которое еще называется в литературе – суммаризацией, уже имеет давнюю историю со своими. Разработаны алгоритмы и программы, позволяющие проводить отбор значимых участков текста. Можно указать здесь одну известную монографию [Das, 2007], где отражены все наиболее популярные сейчас методы аннотирования. Предложенный в нашей работе подход традиционный и связан с отбором предложений или групп предложений, в котором доля ключевых слов превышает заданный порог. Новизна состоит в введении концепции динамических аннотаций на основе динамических словарей. Мы также представляем программы, рассчитанные на конечного пользователя, позволяющие строить такие аннотации.

Для проверки предложенных алгоритмов и функционирования разработанных программ мы рассматриваем публикации, посвященные Евроинтергации Украины на начальном периоде протестных настроений (2013 г). В перспективе предполагается рассмотреть материалы, связанные с публикациями на сайтах региональных администраций.

## **3.2 Динамические словари: алгоритм и программа**

### ***1. Подготовка данных***

На первом этапе алгоритма фиксируется несколько надежных источников информации, которые не должны меняться на протяжении всего исследуемого периода. Это необходимо для сохранения однородности данных. В качестве источников информации мы использовали Интернет версии известных Российских газет.

Собираются документы из указанных источников в примерно равной пропорции, которые распределяются по интервалам времени, которые также заранее

фиксируются. Это могут быть дни, недели, месяцы. Все зависит от интенсивности публикаций и масштаба рассмотрения.

Все документы, относящиеся к одному интервалу времени, объединяются в один интегральный документ. Таким образом, если мы имеем  $n$ - интервалов времени, то в результате подготовительного этапа мы должны иметь  $n$ - документов. Будем называть эти интегральные документы – дневными, недельными, месячными документами.

На втором этапе алгоритма для каждого недельного документа строится список ключевых слов, отражающий содержание этого документа. Такие списки естественно назвать недельными словарями. Понятно, что если квант времени день или месяц, то мы будем иметь, соответственно, дневные или месячные словари.

Отбор ключевых слов, или индексация, является самой важной процедурой алгоритма. Известны десятки способов индексации, каждый из которых определяется решаемой задачей: требуется ли дать описание одиночному документу, или провести группировку корпуса документов, и т.п. Нами выбрана индексация по критерию специфичности, при котором из текста отбираются только те слова, чья относительная частота в документе, превышает относительную частоту в некоторой базовой лексике в заданное  $k$ - число раз. Это  $k$  – называется уровнем специфичности слова, его задает эксперт. В качестве базовой лексики мы рассматривали Генеральную лексику русского языка [Шаров, [http](#)].

Поясним на примере: пусть мы имеем слово  $w$ ='протест' и пусть это слово имеет относительную частоту в документе  $f(w)=0.8*10^{-6}$  и в Генеральной лексике  $F(w)=0.5*10^{-8}$ . Примем уровень специфичности  $k=100$ . Легко видеть, что  $f(w) > k * F(w)$ , то есть условие критерия выполняется, и тогда данное слово будет отобрано.

Мы остановились на данном способе индексации по 2-м причинам: робастность (простота настройки), наличие простой программы. Первая версия программы введена в [Lopez, 2012]. Российский вариант описан в [Александров, 2013].

Оптимальное значение  $k$  определяется опытным путем как компромисс между Recall (полнота) и Precision (точность). Очевидно, что при малых  $k$  словарь будет

наряду с нужными словами содержать много лишних слов. При больших  $k$  словарь будет содержать нужные слова, но много нужных слов будет пропущено. В нашей работе экспериментально было установлено лучшее значение  $k=50$ . При этом в словарях оставалось около 5% лишних слов, которые устранялись вручную.

Итогом второго этапа является  $n$ -недельных словарей по числу выделенных интервалов времени анализа данного события.

## 2. Построение динамических словарей

На последнем этапе содержание недельных словарей комбинируется так, чтобы каждой недели получить 3 списка ключевых слов

1. Список новых слов. Список содержит слова данной недели, которые отсутствовали в предыдущей;
2. Список оставшихся слов. Список содержит слова данной недели, которые входили в словарь предыдущей недели;
3. Список ушедших слов. Список содержит слова предыдущей недели, которые не содержатся в словаре данной недели.

В совокупности эти три списка составляют динамический словарь недели.

Отдельно формируется общий словарь для всего рассматриваемого периода. Он содержит слова, которые входят в не менее, чем заданное число недельных словарей.

На рис. 3.1 схематически представлены динамические словари.



Рис. 3.1 Динамические словари

Здесь: толстой линией обозначены «новые слова», пунктиром — «оставшиеся слова», тонкой линией — «ушедшие слова». Точечной линией отображены общие слова.

### 3. Программные средства

Построение динамических словарей обеспечивается набором программных инструментов, разработанных в нашей научной группе.

1. Программа LexisTerm-I. Это программа индексации по критерию специфичности. Подробно описана в [Александров, 2013]. Эта программа является модифицированным вариантом программы LexisTerm [Lopez, 2012].

2. Программа DynVoc. Это программа построения динамических словарей. Программа описана в [Alexandrov, 2014]

Обе программы рассчитаны на конечного пользователя и являются free-share. На рисунках 3.2, 3.3 и 3.4 показан интерфейс программ, которые упомянуты выше. Все программы рассчитаны на конечного пользователя и написаны в среде C++

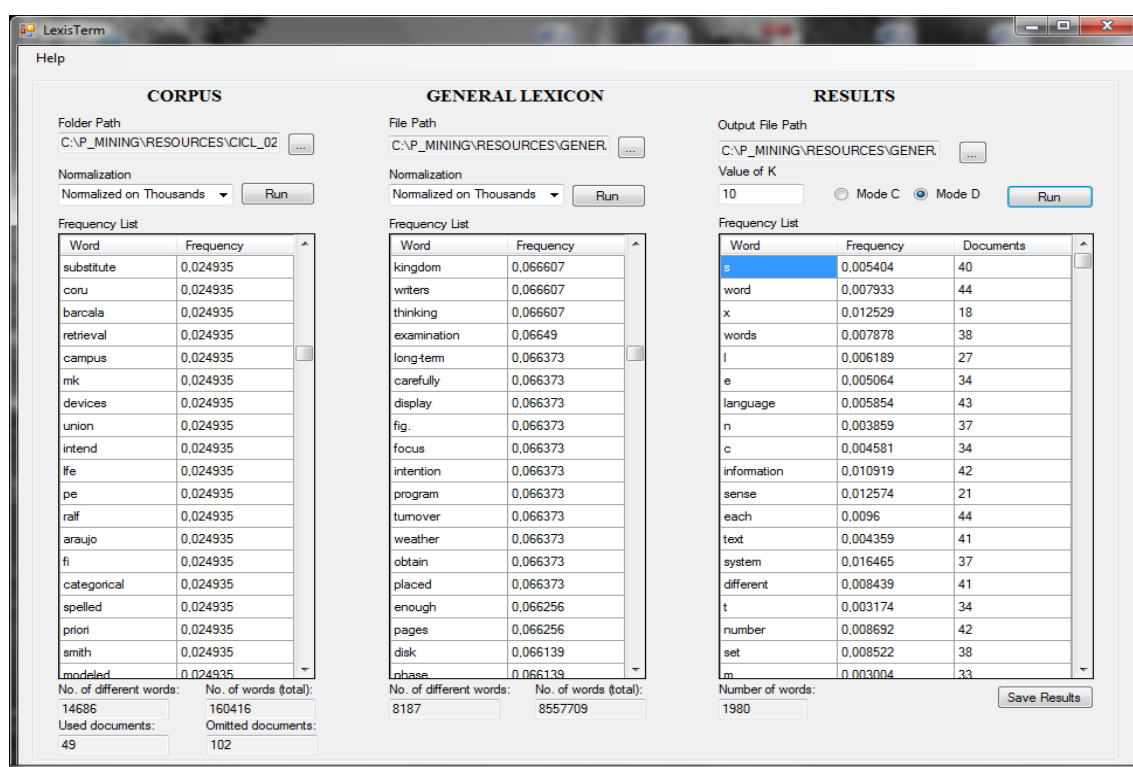


Рис. 3.2 Интерфейс программы LexisTerm

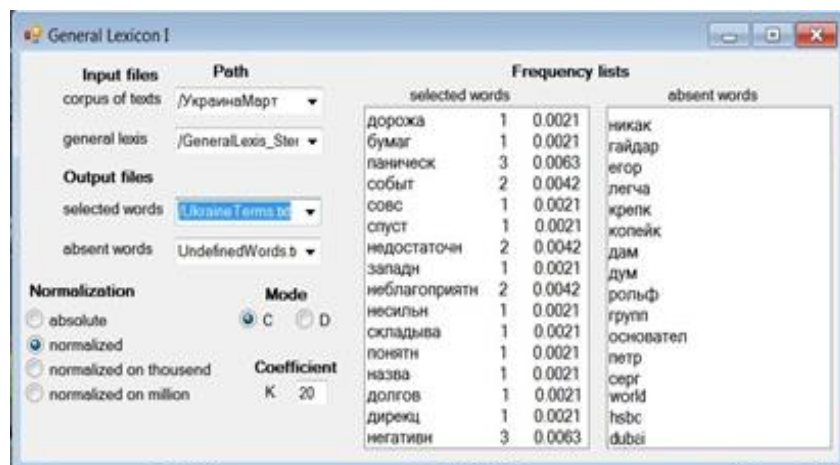


Рис. 3.3 Интерфейс программы LexisTerm-I

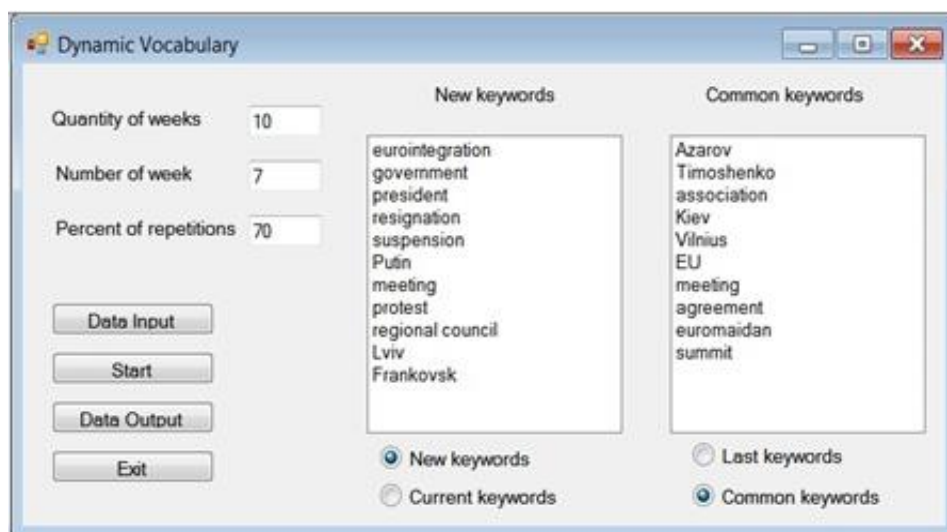


Рис. 3.4 Интерфейс программы DynVoc

### 3.3 Построение аннотаций: алгоритм и программа

#### 1. Основной алгоритм

Как указывалось выше, под динамическим аннотированием здесь мы подразумеваем набор резюме (сжатых изложений) событий, относящихся к определенному промежутку времени. Для формирования абстрактов используются два хорошо известных подхода [Das, 2007]:

1. Формирование (генерирование) фраз на основе выбранных на предыдущем этапе ключевых слов по определенным правилам.
2. Выбор фрагментов (фраз) на основе ранее выбранными ключевыми словами по заданным критериям отбора.

В данной работе используется второй подход, где ключевые слова берутся из динамических словарей. С помощью этих словарей можно использовать различные способы аннотирования. Наиболее интересные случаи представлены в таблице 3.1. Для упорядоченного отбора фраз мы используем параметр  $S$ , который равен отношению доли ключевых слов в предложении. А именно, предложение с  $N_k$  ключевых слов из  $N_w$  слов (кроме стоп-слов) будет отобрано, если  $N_k/N_w > S$ .

Типичные значения  $S$  от 0,1 до 0,3.

Таблица 3.1 Типичные аннотации и словари поддержки

|  | <i>Тип резюме</i> | <i>Словари</i>                                                               |
|--|-------------------|------------------------------------------------------------------------------|
|  | Новые события     | Только новые ключевые слова для каждой недели                                |
|  | Все события       | Новые и повторяющиеся слова каждой недели, а также общие слова всего корпуса |
|  | Ушедшие события   | Ушедшие ключевые слова для каждой недели, и общие слова всего корпуса        |

Предложенный метод отбора фраз довольно прост, но легко интерпретируем. Далее предполагается использовать больше характеристик, как, скажем, положение предложения в тексте, взаимное расположение и т.д.

## **2. Программные средства**

В настоящее время разрабатывается программа DynSum (Dynamic Summarization), которая будет формировать динамические аннотации на основе динамических словарей и корпуса текстов. Таблица 3.2 представляет параметры программы с точки зрения данных, необходимых для функционирования программы:

Таблица 3.2 Типичные аннотации и словари поддержки

| <i><b>Входные данные</b></i>                    | <i><b>Выходные данные</b></i>     | <i><b>Управляющие параметры</b></i>              |
|-------------------------------------------------|-----------------------------------|--------------------------------------------------|
| Корпус текстов по неделям                       | Динамические аннотации по неделям | Порог отбора для каждого типа аннотации          |
| Динамические словари по неделям и общий словарь | Общая аннотация корпуса           | Размер группы фраз для которых применяется порог |

### 3.4 Эксперимент (украинские события)

#### 1. Данные

В качестве экспериментальных данных были взяты публикации в СМИ, посвященные событиям, связанным с Евроинтеграцией Украины. За рассматриваемый период был взят интервал с 1 октября по 30 ноября 2013 года. Источниками информации послужили такие Интернет-издания как «Аргументы и Факты», «РИА Новости», «Россия Сегодня» и «Газета.ру». В среднем на каждую неделю приходится порядка 15 публикаций.

#### 2. События 2-й недели

Содержание новостей второй недели следующее:

*Украина провозгласила курс на Евроинтеграцию и собирается в ноябре подписать в Вильнюсе на саммите "Восточного партнерства" соглашение об ассоциации с ЕС, частью которого является положение о зоне свободной торговли. Дело Юлии Тимошенко остается проблемой в отношениях Украины с Евросоюзом, с которым Киев надеется осенью подписать соглашение об ассоциации. Необходимым условием для подписания ассоциации является освобождение заключенной в больнице Харькова Тимошенко.*

В таблице 3.3 показаны словари седьмой недели, отражающие, как и ранее новые ключевые слова, повторяющиеся ключевые слова, ушедшие ключевые слова, а также общие слова набора.

Таблица 3.3. Динамические словари второй недели (Октябрь 7-14, 2013)

| <b>Новые</b>   | <b>Оставшиеся</b> | <b>Старые</b>      | <b>Общие</b> |
|----------------|-------------------|--------------------|--------------|
| демократия     | судебный          | авиапромышленность | брюссель     |
| договор        | власть            | администрация      | азаров       |
| интеграция     | евроинтеграция    | вступление         | ассоциация   |
| комиссия       | европарламент     | газовая            | белоруссия   |
| литва          | оппозиция         | германия           | вильнюс      |
| нато           | осужденная        | государство        | власть       |
| оборона        | правосудие        | доллар             | восточный    |
| партнерство    | харьков           | избирательная      | евросоюз     |
| переговор      |                   | казахстан          | киев         |
| политика       |                   | кремлевский        | подписание   |
| россия         |                   | лидер              | саммит       |
| сотрудничество |                   | полномочия         | соглашение   |
|                |                   | помилование        | союз         |
|                |                   | преследованне      | тимошенко    |
|                |                   | провозгласить      | украина      |
|                |                   | республика         | членство     |
|                |                   | суверенитет        | янукович     |
|                |                   | судостроение       |              |
|                |                   | трехсторонний      |              |
|                |                   | уголовная          |              |
|                |                   | экспортер          |              |

Словари использовались при построении аннотаций для второй недели. Аннотации представлены в таблице 3.4. Аннотации без стоп слов. Красным цветом отмечены слова из словаря общих слов, синим цветом – новые слова недели, зеленым цветом – повторяющиеся слова недели. В столбцах таблицы показаны значения нагрузки предложений. Здесь под нагрузкой понимается доля ключевых слов в предложении без стоп слов.

Таблица 3.4 Аннотации для второй недели

|  |                                                                                   | <b>Новые</b>  | <b>Новые</b>  | <b>Новые</b> |
|--|-----------------------------------------------------------------------------------|---------------|---------------|--------------|
|  |                                                                                   | +             | +             |              |
|  |                                                                                   | <b>Старые</b> | <b>Старые</b> |              |
|  |                                                                                   | +             |               |              |
|  |                                                                                   | <b>Общие</b>  |               |              |
|  | МИД <b>Литвы</b> : Украина <b>ЕС</b> подпишут <b>соглашение ассоциации</b> ноября | 0.625         | 0.1           | 0.1          |

|   |                                                                                                                                                                                                                                                                                                                      |       |      |      |
|---|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------|------|
|   | Украина провозгласила курс <b>евроинтеграцию</b> собирается в ноябре подписать <b>Вильнюс</b> саммите <b>"Восточного партнерства"</b> <b>соглашение ассоциации ЕС</b> , частью которого является положение в зоне свободной торговли.                                                                                | 0.428 | 0.1  | 0.04 |
|   | Подписание <b>соглашения ассоциации</b> в зоне свободной торговли между <b>Европейским союзом</b> и Украиной запланировано в ноябре, заявил в Новости глава департамента <b>Восточного партнерства</b> МИДа <b>Литвы</b> Эдминас Багдонас.                                                                           | 0.304 | 0.08 | 0.08 |
|   | Ранее Украина заявляла, планирует подписать <b>соглашение ЕС</b> саммите <b>Восточного партнерства</b> , который пройдет <b>Вильнюс</b> в ноябре.                                                                                                                                                                    | 0.5   | 0.07 | 0.07 |
|   | Подписание <b>соглашения</b> между <b>ЕС</b> и Украиной запланировано на первую половину дня в ноябре", — сказал Багдонас                                                                                                                                                                                            | 0.25  | -    | -    |
|   | в очередь, глава представительства <b>ЕС России</b> Вигаудас Ушацкас отметил, ЕС рассчитывает большую <b>интеграцию</b> также с Молдавией и Грузией.                                                                                                                                                                 | 0.25  | 0.14 | 0.14 |
|   | "надеюсь, <b>договор</b> о свободной торговле <b>ассоциации</b> между <b>ЕС</b> и Украиной будет подписан.                                                                                                                                                                                                           | 0.4   | 0.1  | 0.1  |
|   | Если <b>Вильнюс</b> <b>соглашения</b> будут парафироваться, будут подписаны в ближайшем будущем", — сказал Ушацкас.                                                                                                                                                                                                  | 0.18  | -    | -    |
|   | Двери <b>ЕС НАТО</b> открыты для Украины, заявил Расмуссен.                                                                                                                                                                                                                                                          | 0.375 | 0.14 | 0.14 |
| 0 | Украина рассчитывает в ноябре в <b>вильнюсском саммите "Восточного партнерства"</b> подписать <b>соглашение об ассоциации ЕС</b> , частью которого является создание Двери <b>Евросоюза НАТО</b> остаются открытыми для Украины, заявил в четверг генеральный секретарь <b>НАТО</b> Андерс Фог Расмуссен.            | 0.333 | 0.06 | 0.06 |
| 1 | Украине открыты двери углубления отношений с <b>Европейским союзом</b> , равно <b>НАТО</b> ", — сказал журналистам в <b>Брюсселе</b> .                                                                                                                                                                               | 0.333 | 0.07 | 0.07 |
| 2 | словам Расмуссена, подвергает сомнению "евроатлантическое" будущее Украины.                                                                                                                                                                                                                                          | -     | -    | -    |
| 3 | в июле 2010 года Верховная Рада Украины приняла закон, определяющий "внеблоковый статус" страны.                                                                                                                                                                                                                     | -     | -    | -    |
| 4 | в феврале этого года, в завершении <b>Брюсселе</b> очередного заседания <b>комиссии</b> Украина- <b>НАТО</b> на уровне министров <b>обороны</b> , Расмуссен подтвердил, принятое <b>НАТО</b> <b>саммите</b> в Бухаресте в году решение, что Грузия и Украина имеют право <b>членство</b> в альянсе, остается в силе. | 0.206 | 0.11 | 0.11 |

|   |                                                                                                                                                                                                                                                                                                                                                       |       |      |      |
|---|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------|------|
| 5 | этом генсек <b>НАТО</b> указал, Украине надлежит самой определить направление <b>сотрудничества НАТО</b> ближайшие годы.                                                                                                                                                                                                                              | 0.153 | 0.16 | 0.16 |
| 6 | Украина уже продемонстрировала серьезный прогресс решении проблем, которых ранее говорил <b>Евросоюз</b> , время еще некоторые вопросы остаются открытыми, заявил депутат <b>Европарламента</b> Яцек Сариуш-Вольски                                                                                                                                   | 0.09  | 0.04 | -    |
| 7 | Освобождение <b>осужденной</b> экс-премьера Юлии <b>Тимошенко</b> по-прежнему остается основным условием подписания <b>соглашения ассоциации</b> между Украиной <b>Европейским союзом саммите</b><br><br>"Восточного партнерства" <b>Вильнюсе</b> конце ноября, заявил пятницу депутат <b>Европарламента</b> Яцек Сариуш-Вольски                      | 0.392 | 0.07 | 0.03 |
| 8 | словом, "никаких иллюзий быть должно, потому оставлять политических <b>оппонентов</b> тюрьме — это атака на <b>демократию</b> ".                                                                                                                                                                                                                      | 0.142 | 0.14 | 0.07 |
| 9 | сможем принять <b>члены</b> такую страну, которая уважает принципы <b>демократии</b> ", — добавил депутат, отметив, решении вопроса президент Украины Виктор Янукович должен сыграть особенную роль.                                                                                                                                                  | 0.09  | 0.04 | 0.04 |
| 0 | <b>Тимошенко</b> была <b>осуждена</b> году превышение полномочий подписании газового контракта <b>Россией</b> , более года находится лечении пределами тюрьмы — городской больнице <b>Харькова</b>                                                                                                                                                    | 0.21  | 0.16 | 0.05 |
| 1 | <b>ЕС</b> считает судебное преследование <b>оппонентов</b> власти Украине политическим требует <b>Киева</b> прекратить практику избирательного применения <b>правосудия</b> .                                                                                                                                                                         | 0.25  | 0.06 | -    |
| 2 | вторник ходе десятого заседания комитета вопросам экономического сотрудничества <b>российско-украинской межгосударственной комиссии</b> состоялась встреча первого замминистра иностранных дел Украины Руслана Демченко статс-секретарем - заместителем министра иностранных дел России Григорием Карасиным.                                          | 0.06  | 0.06 | 0.06 |
| 3 | шла речь на встрече первого замминистра иностранных дел Украины Руслана Демченко статс-секретарем — заместителем министра иностранных дел <b>России</b> Григорием Карасиным, которая состоялась ходе десятого заседания комитета вопросам экономического <b>сотрудничества</b> российско-украинской межгосударственной <b>комиссии</b> вторник Калуге | 0.09  | 0.09 | 0.09 |
| 4 | время встречи была констатирована важность поддержания высокой динамики двустороннего диалога, прежде всего широкому кругу вопросов двусторонних торгово-экономических отношений, которые должны соответствовать характеру стратегического <b>партнерства</b> между Украиной                                                                          | 0.06  | 0.06 | 0.06 |

|   |                                                                                                                                                                                                                           |       |       |       |
|---|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|-------|
|   | Российской Федерацией", — говорится в сообщении                                                                                                                                                                           |       |       |       |
| 5 | Медведев: РФ примет меры в случае вступления Украины в ассоциацию ЕС                                                                                                                                                      | 0.272 | 0.07  | 0.07  |
| 6 | Более того, в консультациях с нашими европейскими партнерами мы говорим о том, что имеют место определенные расхождения по техническим параметрам, в вопросах понимания того иного соглашения.                            | 0.1   | 0.05  | 0.05  |
| 7 | хотели думать, такому стратегическому партнеру, Украина, Россия может применять какие-то меры политического давления.                                                                                                     | 0.076 | 0.076 | 0.076 |
| 8 | исключаем", — сказал, отвечая на вопрос, расценивают ли правительство Украины ситуацию с компанией Roshen как политическое давление РФ на Украину с целью допустить подписания соглашения между Украиной и ЕС ассоциации. | 0.208 | 0.04  | 0.04  |
| 9 | указал Азаров, политическое давление будет противоречить духу договора о дружбе и сотрудничестве между РФ и Украиной.                                                                                                     | 0.307 | 0.230 | 0.230 |
| 0 | "Напомню, в году такой договор был подписан о дружбе и сотрудничестве между Украиной и Россией.                                                                                                                           | 0.272 | 0.272 | 0.272 |
| 1 | определил стратегические аспекты сотрудничества, определил понятие стратегического партнерства и исключил из нашего сотрудничества такие формы, например, политическое давление.                                          | 0.125 | 0.125 | 0.125 |
| 2 | этими документами, которые ратифицированы нашими парламентами, стоят подписи руководителей, поэтому хотели только думать, предполагать, такие меры могут быть приняты", — добавил Азаров.                                 | 0.04  | -     | -     |

### 3. События 7-й недели:

Содержание новостей седьмой недели следующее:

Давление со стороны ЕС на Киев по вопросу Евроинтеграции подстрекает оппозиционно настроенную часть украинского общества на протестные и противоправные действия против законной власти. Акции протестов по поводу приостановления Евроинтеграции прошли в Киеве, Львове, и других городах Украины. Протестующие требуют отправить в отставку премьер-министра Украины Азарова.

В таблице 3.5 показаны словари седьмой недели, отражающие, как и ранее новые ключевые слова, повторяющиеся ключевые слова, ушедшие ключевые слова, а также общие слова набора. Это последний столбец таблицы совпадает с последним столбцом таблицы 2.2

Таблица 3.5 Динамические словари седьмой недели (Ноябрь 8-15, 2013)

| <b>Новые</b>    | <b>Оставшиеся</b> | <b>Старые</b>   | <b>Общие</b> |
|-----------------|-------------------|-----------------|--------------|
| государство     | азаров            | армения         | брюссель     |
| деполитизирован | соглашение        | дума            | азаров       |
| евроинтеграция  | сотрудничество    | евросаммит      | ассоциация   |
| евромайдан      |                   | казахстан       | белоруссия   |
| компенсация     |                   | канцлер         | вильнюс      |
| кредитный       |                   | конфликт        | власть       |
| львовский       |                   | москва          | восточный    |
| международный   |                   | нечестные       | евросоюз     |
| митинг          |                   | обязательные    | киев         |
| облсовет        |                   | партнерство     | подписание   |
| оппозиционно    |                   | перенаправлен   | саммит       |
| отставка        |                   | правосудие      | соглашение   |
| полномочия      |                   | противостоят    | союз         |
| потребовали     |                   | ратификация     | тимошенко    |
| правительство   |                   | рогозин         | украина      |
| президент       |                   | российский      | членство     |
| премьер         |                   | сбыт            | янукович     |
| приостановить   |                   | спекуляция      |              |
| протестные      |                   | таможенные      |              |
| противоправные  |                   | технологии      |              |
| путин           |                   | уголовная       |              |
| резолюция       |                   | экономический   |              |
| суверенитет     |                   | энергоснабжение |              |
| технически      |                   | эштон           |              |
| франковский     |                   |                 |              |

Словари использовались при построении аннотаций для седьмой недели. Аннотации представлены в таблице 3.6. Аннотации без стоп слов. Красным цветом отмечены слова из словаря общих слов, синим цветом – новые слова недели, зеленым цветом – повторяющиеся слова недели. В столбцах таблицы показаны значения нагрузки предложений.

Таблица 3.6 Аннотации для седьмой недели

|    |                                                                                                                                                                                                                                        | Новые<br>+<br>Повтор.<br>+<br>Общие | Новые<br>+<br>Повтор. | Новые |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|-----------------------|-------|
| 1  | <b>Янукович</b> условиях финансовой поддержки <b>ЕС</b> : не надо унижать                                                                                                                                                              | 0.333                               | -                     | -     |
| 2  | <b>соглашении</b> <b>ЕС</b> была предусмотрена финансовая <b>техническая</b> помощь Украине размере миллионов евро.                                                                                                                    | 0.333                               | 0.166                 | 0.08  |
| 3  | "Три года подряд конфетку красивой обертке нам показывали и говорили: получите, когда МВФ подпишете <b>соглашение</b> .                                                                                                                | 0.06                                | 0.06                  | -     |
| 4  | хочу грубо выражаться, надо унижать", — сказал <b>Янукович</b> .                                                                                                                                                                       | 0.142                               | -                     | -     |
| 5  | <b>Президент</b> Украины Виктор <b>Янукович</b> считает унижительными условия <b>ЕС</b> выделению стране финансовой помощи.                                                                                                            | 0.307                               | 0.07                  | 0.07  |
| 6  | <b>Янукович</b> : Украина - бедный родственник, ходящий с протянутой рукой                                                                                                                                                             | 0.111                               | -                     | -     |
| 7  | " <b>соглашение</b> <b>ассоциации</b> , миллионов евро заложена финансовая <b>техническая</b> помощь Украине.                                                                                                                          | 0.333                               | 0.222                 | 0.111 |
| 10 | серьезная страна, европейская" сказал <b>Янукович</b> фрагменте интервью, который был показан каналу.                                                                                                                                  | 0.09                                | -                     | -     |
| 11 | Украина рассчитывала возобновление <b>сотрудничества</b> МВФ привлечение рамках новой программы финансирование миллиардов долларов.                                                                                                    | 0.08                                | 0.08                  | -     |
| 12 | Срок предыдущей <b>кредитной</b> программы фонда Украины, начатой году, истек декабре года.                                                                                                                                            | 0.09                                | 0.09                  | 0.09  |
| 13 | Ивано-Франковский <b>облсовет</b> требует подписать <b>ассоциацию</b> с <b>Евросоюзом</b> .                                                                                                                                            | 0.428                               | 0.142                 | 0.142 |
| 14 | Согласно информации кабмина, ноября текущего года было получено письмо фонда, которым МВФ для предоставления <b>кредита</b> <b>потребовал</b> повысить тарифы газ населения, заморозить заработные платы, сократить бюджетные расходы. | 0.07                                | 0.07                  | 0.07  |
| 15 | словам <b>Януковича</b> , <b>власти</b> никогда пойдут условия МВФ.                                                                                                                                                                    | 0.285                               | -                     | -     |

|    |                                                                                                                                                                                                                                                                                                                                                                        |       |       |       |
|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|-------|
| 16 | Митинги центре <b>Киева</b> , также других городах Украины продолжаются вечера четверга, когда <b>правительство</b> объявило своем решении <b>приостановить</b> процессподготовки подписанию <b>соглашения ассоциации ЕС</b> .                                                                                                                                         | 0.363 | 0.181 | 0.136 |
| 17 | Ивано-Франковский <b>областной</b> совет своем заседании вторник <b>потребовал</b> президента Украины Виктора <b>Януковича</b> подписать <b>саммите Вильнюсе</b> ноября <b>соглашение ассоциации ЕС</b> , также отправить <b>отставку</b> премьер-министра Николая <b>Азарова</b> решение <b>приостановить евроинтеграцию</b> страны, сообщает сайт <b>облсовета</b> . | 0.437 | 0.25  | 0.218 |
| 18 | Заседание <b>облсовета</b> прошло местном " <b>евромайдане</b> ", где собрались сторонники <b>евроинтеграции</b> Украины.                                                                                                                                                                                                                                              | 0.3   | 0.3   | 0.3   |
| 19 | Сторонник <b>евроинтеграции</b> Украины сотрудники милиции здания <b>правительства</b> страны <b>Киеве</b> .                                                                                                                                                                                                                                                           | 0.333 | 0.222 | 0.222 |
| 21 | "Требовать <b>президента</b> Украины <b>Януковича</b> Верховной Рады Украины срочно изучить необходимые действия подписания <b>саммите Вильнюсе</b> ноября <b>соглашения ассоциации</b> между Украиной Европейским союзом", — говорится <b>резолуции</b> заседания Ивано-Франковского <b>облсовета</b> .                                                               | 0.307 | 0.153 | 0.115 |
| 22 | Ранее решение с призывом <b>президенту</b> подписать <b>ассоциацию</b> уволить премьера <b>Азарова</b> принял <b>международный</b> горсовет.                                                                                                                                                                                                                           | 0.307 | 0.23  | 0.153 |
| 23 | МИД РФ: давление <b>ЕС Киев</b> подстрекает украинцев нарушению закона.                                                                                                                                                                                                                                                                                                | 0.222 | -     | -     |
| 24 | <b>Протестные</b> акции <b>Киеве</b> других городах Украины начались ноября, когда было обнародовано решение <b>правительства</b> , решившего взять паузу отношениях <b>ЕС</b> .                                                                                                                                                                                       | 0.263 | 0.105 | 0.105 |
| 25 | Давление стороны <b>ЕС Киев</b> вопросу <b>евроинтеграции</b> подстрекает <b>оппозиционно</b> настроенную часть украинского общества <b>протестные противоправные</b> действия против законной <b>власти</b> , говорится распространенном вторник комментарии МИД РФ.                                                                                                  | 0.32  | 0.16  | 0.16  |
| 26 | <b>Путин</b> призвал <b>Евросоюз</b> <b>деполитизировать</b> тему <b>евроинтеграции Киева</b> .                                                                                                                                                                                                                                                                        | 0.714 | 0.428 | 0.428 |
| 27 | " понятны причины, побудившие украинское <b>правительство</b> "взять паузу" процессе <b>евроинтеграции</b> , продумать механизмы <b>компенсации</b> потерь украинской экономики вступления силу <b>соглашения ассоциации ЕС</b> .                                                                                                                                      | 0.333 | 0.238 | 0.19  |
| 28 | <b>Евросоюзу</b> России было адресовано предложение совместно изучить данный вопрос", — указано сообщении.                                                                                                                                                                                                                                                             | 0.09  | -     | -     |

|    |                                                                                                                                                                                                                                                                                                           |       |       |       |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|-------|
| 29 | "Брюссель" ответ такой шаг <b>Киева</b> начал лишь еще жестче давить украинское руководство, любой ценой пытаясь убедить дать согласие подписание <b>соглашения</b> .                                                                                                                                     | 0.15  | 0.05  | -     |
| 30 | этом <b>ЕС</b> могут отдавать себе отчет том, подобное вмешательство внутренние дела <b>суверенного государства</b> подстрекает <b>оппозиционно</b> настроенную часть украинского общества <b>протестные противоправные</b> действия против законной украинской <b>власти</b> ", — говорится комментарии. | 0.275 | 0.172 | 0.172 |

Эксперт, знакомый с событиями на Украине, может согласиться, что динамические словари отражают события, произошедшие в указанный интервал времени.

### 3.5 Выводы по главе

В главе описан простой метод изучения динамики событий, представленных в Интернет публикациях, с помощью специальным образом агрегированных списков ключевых слов (динамических словарей). Для автоматического аннотирования разработан программный комплекс, который ориентирован на применение указанных словарей. Был проведен эксперимент с реальными данными, связанными с событиями на Украине. Результаты эксперимента подтвердили соответствие аннотаций полному тексту, для которого эти аннотации составлялись.

## **ЗАКЛЮЧЕНИЕ**

Содержание данной работы следует отнести к области, которую можно назвать Интернет социологией. Данная предметная область в настоящее время активно формируется, и ее открытыми проблемами являются: модели представления документов (атрибутация), технологии извлечения атрибутов из неструктурированных данных, экономная параметризация документов, преобразование информации в форму, доступную для анализа социологами, экономистами, политиками.

В настоящей работе предлагаются решения для первых двух проблем из указанного списка. Их решение позволяет резко сократить объем информации в базах данных протестов, и привести к единой унифицированной форме содержание баз данных протестов, представленных на разных языках.

## ЛИТЕРАТУРА

- (Agerri et al., 2013) Agerri, R., Aldabe, I., Beloki, Z., Laparra, E., Lopez de Lacalle, M., Rigau, G., Soroa, A., van Erp, M., Vossen, P., Girardi, C. and Tonelli, S.: Event Detection. In: NewsReader. version 1. Deliverable D 4.2.1. Version FINAL (2013)
- (Aleksić et al., 2014) Aleksić, V., Thurmair, G., Casamayor, G., Moumtzidou, A., Vrochidis, S.: Basic techniques for speech recognition, text analysis and concept detection. In: Mining and Understanding of multilingual content for Intelligent Sentiment Enriched context and Social Oriented interpretation, D 2.2 (2014)
- (Alexandrov, 2001) Alexandrov, M.: Dynamic domain-oriented dictionaries as a tool for revealing tendencies in development of some scientific and technological disciplines and interaction between them. Mexican National Council on Sciences and Technologies (CONACyT), Reg.No. 39011-a. Technical report, 2001
- (Alexandrov, 2014) Alexandrov M., Beresneva D., Makarov A.: Dynamic vocabularies as a tool for studying social processes. In: Proc. of 6'th Intern. Conf. on Intelligent Information and Engineering Systems, ITHEA Publ., 2014, p.88-92
- (Aone et al., 1994) Aone, C., Blejer, H., Okurowski, M.E., Van Ess-Dykema, C.: A Hybrid Approach to Multilingual Text Processing: Information Extraction and Machine Translation. In: Proceedings of the First Conference of the AMTA (1994)
- (Atkinson et al., 2011) Atkinson, M., Piskorski, J., Van der Goot, E., Yangarber, Y.: Multilingual Real-time Event Extraction for Border Security Intelligence Gathering. In: U.K. Wiil (ed.), Counterterrorism and Open Source Intelligence, Lecture Notes in Social 355 Networks 2, Springer-Verlag/Wien. (2011)
- (Boschee et al., 2005) Boschee, E., Weischedel, R., and Zamanian, A.: Automatic information extraction. In: Proceedings of the International Conference on Intelligence Analysis (2005)
- (Buitelaar et al, 1998) Buitelaar, P., Netter, K., Xu, F.: Integrating Different Strategies for Cross-Language Information Retrieval in the MIETTA Project. In: Proceedings of the 14th Twente Workshop on Language Technology. (1998)

- (CCD, 2014) Media Data and Social Science Research: Problems and Approaches. Cline Center for Democracy University of Illinois at Urbana-Champaign, February (2014)
- (Danilova, 2013) Danilova, V.: Cross-Language Plagiarism Detection Methods. In: Proceedings of the Student Research Workshop associated with RANLP 2013, pages 5157, Hissar, Bulgaria, 9-11 September (2013)
- (Danilova, 2014) Danilova V.: Ontology Building and Annotation of Destabilizing Events in News Feeds. In: Artif. Intell. Applications to Business and Engineering Domains (Monograph), ITHEA Publ., pp.137-146 (2014)
- (Danilova & Popova, 2014) Danilova V., Popova S.: Socio-Political Event Extraction Using a Rule-Based Approach. In: Proc. of the 13th International Conference on Ontologies, DataBases and Applications of Semantics (ODBASE'2014), Springer, Volume 8842, 2014, pp. 537-546 (2014)
- (Das, 2007) Das,D., Martins, A. A Survey on automatic text summarization. Springer, LNCS, 2007, 8 pp.
- (Du et al., 2011) Du, M., Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building Support Tools for Russian-Language Information Extraction. In: I. Habernal and V. Matousek (Eds.): TSD 2011, LNAI 6836, pp. 380387 (2011)
- (Elvira & Hecking, 2005) Casals Elvira, X., Hecking, M.. IEPS: A Framework to Manage and to Visualize Information Extraction Results. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V . (FGAN), Technischer Bericht FKIE/ITF/2005/2, September 2005.
- (Fortuna's Corner, 2014) Fortuna's Corner blog: Why Big Data Mining Missed the Early Warning Signs of Ebola; over-reliance at the alter of big data mining will have some unpleasant shortfalls: <http://fortunascorner.com/2014/09/29/why-big-data-mining-missed-the-early-warning-signs-of-ebola-over-reliance-at-the-alter-of-big-data-mining-will-have-some-unpleasant-shortfalls/>
- (Hanna, 2014) Hanna, A.: Developing a System for the Automated Coding of Protest Event Data - Available at SSRN 2425232 (2014)

- (Hayes and Nardulli, 2011) Hayes, M., Nardulli, P. F.: SPEED's Societal Stability Protocol and the Study of Civil Unrest: an Overview and Comparison with Other Event Data Projects, October 2011, Cline Center for Democracy University of Illinois at Urbana-Champaign (2011)
- (Hecking & Sarmina-Baneviciene, 2010) Hecking, M., Sarmina-Baneviciene, T.: A Tajik Extension of the Multilingual Information Extraction System ZENON. In: Washington, DC: CCRP C4ISR Cooperative Research Program, 14 pp. (2010)
- (Hogenboom et al., 2011) Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F.: An Overview of Event Extraction from Text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), volume 779 of CEUR Workshop Proceedings, pages 48-57. (2011)
- (Leban and Brank, 2011) Lebal, G., and Brank, J.: Early Event Extraction Prototype. Deliverable D 4.3.1, Version 1.0, (2011)
- (Leetaru, 2011) Leetaru, K.: Automatic Document Categorization for Highly Nuanced Topics in Massive-Scale Document Collections: The SPEED BIN Program. Cline Center for Democracy University of Illinois at Urbana-Champaign, March (2011)
- (Lejeune, 2009) Lejeune, G.: Structure Patterns in Information Extraction: a Multilingual Solution? In: Advances in Method of Information and Communication Technology AMICT09, Volume 11 p. 105-111, Petrozavodsk, Russia. (2009)
- (Lopez, 2012) Lopez, R., Alexandrov, M.:Tejada, J: LexisTerm - the Program for term selection by the criterion of specificity, In: Proc. of 4-th Intern. Conf. on Intelligent Inform and Engineering Systems, ITHEA Publ, 2011, p. 8-15
- (Makagonov, 2006) Makagonov. P., Figueroa, A., Gelbukh, A.: Studying evolution of a branch of knowledge by constructing and analyzing its ontology. Springer, LNCS, 2006, 10 pp.
- (McClelland, 1999) McClelland, C.: World Event/Interaction Survey. ICPSR Inter-university Consortium for Political and Social Research, 1966-1978, ICPSR 5211, January (1999)

- (Nardulli et al., upcoming) Nardulli, P. F., Althaus, S. L., and Hayes, M.: A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data (Forthcoming in Sociological Methodology)
- (Padró and Stanilovsky, 2012) Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.Istanbul, Turkey. May, 2012.
- (Piskorski et al., 2011) Piskorski, J., Belayeva, J., Atkinson, M.: Exploring the Usefulness of Cross-lingual Information Fusion for Refining Real-time News Event Extraction: A Preliminary Study. RANLP 2011: 210-217 (2011)
- (Piskorski & Yangarber, 2013) Piskorski, J., Yangarber, R.: Information Extraction: Past, Present and Future. Survey Chapter in "Multi-source, Multilingual Information Extraction and Summarization", Theory and Applications of Natural Language Processing (T. Poibeau et al., eds.). Springer-Verlag, pp. 23-49 (2012)
- (Pivovarov, 2013) Pivovarov, L., Du, M., Yangarber, R.: Adapting the PULS Event Extraction Framework to Analyze Russian Text. At ACL: 4th Biennial Workshop on Balto-Slavic Natural Language Processing. Sofia, Bulgaria (2013)
- (Pustejovsky et al., 2005) Pustejovsky, J., Knippen, R., Littman, J., Saurí, R.: Temporal and Event Information in Natural Language Text. In: Language Resources and Evaluation.39 (2-3):123-164 (2005)
- (Raleigh et al., 2010) Raleigh, C., Linke, A., Hegre, H., and Karlsen, J.: "Introducing ACLED-Armed Conflict Location and Event Data." Journal of Peace Research 47, no. 5: 1-10 (2010)
- (Schrodt, 1998) Schrodt, P. A.: KEDS: Kansas Event Data System. Version 1.0. (1998)
- (Schrodt, 2012) Schrodt, P. A.: CAMEO: Conflict and Mediation Event Observations. Event and Actor Codebook. Department of Political Science, Pennsylvania State University, Version: 1.1b3, March (2012)
- (Schrodt, 2013) Schrodt, P. A.: GDELT: Global Data on Events, Location and Tone. Workshop at the Conflict Research Society, Essex University, 17 September (2013)

- (Schrodt, 2014a) Schrodt, P. A.: Open Event Data: Opportunities and Challenges. Workshop on “Big Data and Death”, University of Wisconsin, 7 November (2014)
- (Schrodt, 2014b) Schrodt, P. A.: TABARI Textual Analysis by Augmented Replacement Instructions, Version 0.8.4, Parus Analytical Systems. Charlottesville, VA 22901, VERSION 0.8.4B3: August 30 (2014)
- (Schrodt et al., 2014) Schrodt, P. A., Beielser, J., and Idris, M. : Three’s a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. Presented at the International Studies Association, Toronto, March (2014)
- (Schrodt & Van Brackle, 2013) Schrodt, P. A. and Van Brackle, D.: Automated coding of political event data. In: Subrahmanian V. S. (ed.), Handbook of computational approaches to counterterrorism. Springer, New York, pp. 23-49 (2013)
- (Schrodt & Yonamine, 2012) Schrodt, P. A., and Yonamine, J. : Automated Coding of Very Large Scale Political Event Data. Armed Conflict Location & Event Data Project (ACLED). Guide to Dataset Use for Humanitarian and Development Practitioners, pp. 3-9, January (2015)
- (Steinberger, 2012) Steinberger, R.: A Survey of Methods to Ease the Development of Highly Multilingual Text Mining Applications. In: Language Resources & Evaluation 46: 155-176 (2012)
- (Tadić et al., 2011) Tadić, M.: Early Machine Translation Based Semantic Annotation Prototype. Deliverable D 3.3.1, Version 1.0 (2011)
- (Wueest et al., 2013) Wueest, B., Rothenhäusler, K., Hutter, S.: Using Computational Linguistics to Enhance Protest Event Analysis - Available at SSRN 2286769 (2013)
- (Александров, 2013) Александров, М.: Разработка общей методологии анализа общественного мнения Интернет-сообщества и ее приложение к заданным темам (власть, экономика, коррупция и пр.) на основе инструментов Data/Text Mining. Отчет по Гос.заданию-84, РАНХГС, 2013
- (Шаров, http) Шаров, С.: Генеральная лексика русского языка <http://www.artint.ru/projects/frqlist.php>