Natural Language Processing Models

ONTOLOGY BUILDING AND ANNOTATION OF DESTABILIZING EVENTS IN NEWS FEEDS[1]

Vera Danilova

***Abstract**: This paper presents an attempt of elaborating a domain-specific knowledge resource to create semantic annotations of destabilizing events (civil unrest) within the framework of the socio-political event extraction task. The final objective is to reduce the manual effort of sociological researchers by automatically generating structured information on the progress of an event (protest as a verbal expression or an action), its participants, origins and the aftermath, as well as other reported details that can contribute to the analysis. The previous experience of destabilizing event ontology building (SPEED project), as well as several state-of-the-art works on the analysis of protest behaviour, are addressed. Ontology development in Protégé-4.3 and mapping using the GATE Developer 8.0 are described.*

***Keywords**: ontology construction, sociological application, event indexing*

***ACM Classification Keywords**: H.3.1 Content Analysis and Indexing*

Introduction

Protest activity manifests itself in a wide variety of events from verbal expressions, relatively peaceful demonstrations and strikes to civil wars, revolutions, coup d'état, etc. and is being explained by diverse social, economic, political and environmental causes. The main task addressed in the present paper deals with improving the quality of social protest events extracted from news streams by enriching them with semantic data. Russian researchers of social protest point out that the analysis of protest activity causes, motives and factors is insufficient in most state-of-the-art sociological papers within the framework of national studies. They address the formation of social movements (ethnic, feminist, ecological) as a conflict type of collective action, job actions, regional protest activity, modeling of public opinion dynamics, etc.. Although certain phenomena are covered, there is still a room for further research [Dementieva, 2013]. The scope of the international studies ranges from the examination of the intra-state protest expression to the global phenomena. In [Braha, 2012], newspaper reports on civil unrest events that took place in 170 countries during the period 1919-2008 are analyzed in order to model the mechanisms of social unrest contagion, which have proved to be similar to those of natural hazards and epidemics. The long-term dataset includes such action types as anti-government demonstrations, riots, and general strikes.

Ontology is an essential application for knowledge management and research. It is used for semantic search, annotation and linking, etc.. The most common representation of an ontology is a graph, where concepts are stored in the nodes and the edges represent the corresponding relations. The main advantage of this structure is that it can be easily understood by the machine due to the explicit and accurate definition of hierarchical and nonhierarchical relations between concepts. The largest ontology of destabilizing events

[1] Work done under partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

(Societal Stability Protocol) that covers many kinds of human-initiated protests, politically motivated attacks by non-governmental initiators, as well as the reaction of the government was built within the framework of the SPEED project at the Cline Institute for Democracy of Illinois and will be addressed in detail in Section 2. It is a hierarchical domain ontology that generates event data collected across all countries after the Second World War [Hayes and Nardulli, 2011]. The data for the period from 2006 till present was crawled from over 5.000 news feeds in 120 countries several times each day.

The main objective of our work at the present stage is to develop a conceptual representation of protest activity in Russia and other countries (viewpoint of Russian news media), basing on the data crawled from Russian news streams. Concept mapping will allow the ability to automatically register the relationships between the observed characteristics of protest activity, e.g.: participation of certain actors (governmental/non-governmental) depending on the origins of the protest, association between the direct or indirect motivation to/not to participate and the actual involvement in the action, etc..

The rest of the paper is organized as follows. Section 2 describes the Societal Stability Protocol as a detailed conceptualization of the civil unrest domain, the structure of which was taken into account in the course of ontology building. Also, we mention the most common techniques in the field of ontology building. In Section 3, input data acquisition and tools are outlined. In Section 4, we describe the construction of the ontology, lexicons and annotation rules. Section 5 discusses the results and concludes the paper.

Related Work

An ontology can be constructed in a manual, semi-automatic or automatic way using natural language processing, unsupervised machine learning, and other techniques and resources. Ontologies can be generated either directly from text or from dictionaries, thesauri, knowledge bases, semi-structured and relational schemata [Lieto, 2008]. Learning pipeline includes term extraction, disambiguation, concept identification, concept hierarchy construction, identification of relations and rules within the ontology. Term extraction uses either indexing mechanisms from the information retrieval domain or natural language processing. For the context-based term disambiguation, clustering along with the use of association measures to detect statistically correlated pairs is applied. Thesauri and dictionaries are also employed to group terms with similar meanings. For concept identification, unsupervised machine learning techniques are widely used. In some approaches to concept hierarchy construction, lexical relations of hyponyms are extracted from corpus using automatically acquired context-based lexico-syntactic patterns. Also, an approach based on Harris's Distributional Semantics Hypothesis [Harris, 1970] has been applied. It takes into account the correlation between a word and a context. Contexts are encoded in term vectors, clustering is performed and, finally, a distance measure (TF-IDF or chi-square) is applied to separate term senses. The identification of nonhierarchical relations within an ontology involves the use of text mining techniques and linguistic analysis. The automatic rule identification (a rule looks like "X caused Y", "Y is triggered by X", etc.) is a less developed area, where there are no common and well-established approaches [Toledo-Alvarado et al., 2012]. The interaction with the user or expert at different stages, as well as the comparison to the existing ontologies and term hierarchies, contribute to the ontology refinement.

Societal Stability Protocol created within the framework of the SPEED Project (Social, Political, and Economic Event Database) of the Cline Institute for Democracy of Illinois contains a well-developed ontology of destabilizing events. A destabilizing event is a happening that unsettles the routines and expectations of citizens, causes them to be fearful, and raises societal anxiety about the future [Nardulli et al., 2013]. The database encompasses texts collected from the newspapers issued in 165 countries in the Post WWII era. All the articles are translated into English. The final protocol design iteration includes eleven sections

responsible for data processing. The sixth sections deals with the domain ontology of event types consisting of three main Tier 1 categories (political expression events, politically motivated attacks, disruptive state acts). Political expression involves an obligatory presence of such parameters as public articulation, non-governmental actor, threatening or unwelcome political message. The main expression modes are a) verbal or written message, b) symbolic act, c) forming an association and d) mass demonstration or strike with the subsequent subcategories. Politically motivated attacks are violent actions or attempts by non-governmental initiators. Political motives in this case are defined as hatred toward socio-cultural groups or revenge for their prior actions, desire to change or control the government, follow or oppose a political ideology, advance a social cause etc.. Disruptive state acts include extraordinary or repressive acts by governmental initiators.

SPEED personnel constructed the ontology by exploring the literature on political violence, terrorism, political instability, and social movements in search of event categories. Secondly, real data was analyzed and the respective classification was refined. Event-specific information that is relevant to the study of the origins and development of civil unrest was defined. Event attributes, such as geospatial and temporal information (event coverage, latitude/longitude, precise/estimated time and date), participants (governmental/non-governmental initiators and their traits (number, weapons used), messengers, rioters, reactors), consequences (negative/positive to initiators and actual participants), targets and effects (*what* happened to *whom*: damage, injuries etc.), origins (*why* it happened) and event linking, are distributed between the other sections [Hayes and Nardulli, 2011; Nardulli et al., 2013]. The access to SSP is limited.

Our study focuses on the real-time dynamics of the daily happenings and their attributes reported in the news of a specific country. These smaller events are surrounded by various factors that under certain circumstances may affect the status quo. At this stage, a gold-standard ontology of social protest events is created using the Protégé-4.3 tool and verified by an expert in order to provide quality semantic annotation to the data via the GATE 8.0 framework.

Input Data and Tools

Input Data. News titles are selected as the input data for the ontology building and annotation, because they contain short descriptions of a wide variety of events related to the protest activity, most of which are reflected in the SSP ontology. The relevant data on the participants of events, their attributes and triggers can be extracted from the titles. It was also experimentally proved that the noun in the headline is the main argument of an event in 80% of cases [Wunderwald, 2011].

The dataset includes 2000 news titles, extracted from Russian and Ukranian news portals that provide data on socio-political situation in Russia and abroad on a daily basis (ria.ru, lenta.ru, fontanka.ru, forbes.ru, livejournal.com, gazeta.ru, news.rambler.ru, newsru.com, interfax.ru, news.yandex.ru, news.bigmir.net, kommersant.ru, hopesandfears.com/news, kp.ru, mk.ru, ng.ru, gazeta.ua/ru/, pravda.ru, trud.ru). The test set and gold standard for the present experiments include 553 titles each.

The crawlers are created within the Scrapy web crawling framework (http://scrapy.org). They extract news titles and, optionally, the text body, date/time and source. The crawler relies on the mutual presence of several keywords from two predefined keyword lists. The keyword lists are based on the previous manual analysis of civil unrest-related titles that mention conceptual components (trigger, actors, time, location, purpose, etc.) and their combinations. A separate module deals with the large amount of duplicates and partial duplicates in the dataset. They are completely removed using embedded python libraries. Levenstein distance algorithm (NLTK package) for string similarity is efficient, but very slow for big datasets. An example of the resulting collection is presented in the Tab. 1.

Tools. GATE Developer 8.0 (General Architecture for Text Engineering: http://gate.ac.uk) is a powerful open source annotation tool. Multiple plugins for the processing of various natural languages can be uploaded or created manually within the framework. In our experiments it is used, firstly, for preprocessing the collection, which includes such steps as tokenizing, sentence splitting, gazetteer lookup (an embedded Russian Gazetteer, a manually populated OntoGazetteer and a Flexible Gazetteer) and morphological analysis (an embedded version of Yandex API Mystem 2.1 (http://api.yandex.ru/mystem/doc/)). Flexible gazetteer matches words in any morphological variant, while standard GATE gazetteers (ANNIE, embedded language-specific gazetteers) provide only exact string match. OntoGazetteer main functionality consists in lexicon mapping to the ontology classes.

Table 1. An example of crawled titles with translation

Митинг в защиту детдома №2 состоится 16 марта.	An action in defence of the orphanage No2 will be held on 16th of March.
В Чите прошли два митинга в поддержку народов Украины.	Two actions were held in Chita in support of the peoples of Ukraine.
Митинг националистов на Марсовом поле завершился без проишествий.	The nationalist protest at the Marsovo field ended without consequences.
В Симферополе прошел митинг против "евромайдана".	In Simferopol a protest was held against "Euromaidan".
Пассажиры задержанного рейса устроили митинг в аэропорту.	The passengers of the delayed flight staged a protest in the airport.

Secondly, it is applied for the gazetteer population, grammar building and ontology mapping. The concept hierarchy itself is represented formally using Protégé-4.3 software of the Stanford University (http://protege.stanford.edu).

Ontology Construction and Text Annotation

Ontology Construction. The gold standard ontology of social protest is constructed manually on the basis of the SSP domain ontology and real data from news feeds, and it is formalized within the Protégé-4.3 framework. It has been revised by one domain expert, and it will be subject to control assessment focused on revealing and highlighting the less studied aspects. The current version spans action types, participant classification and different attributes that are included in the corresponding sections of the SSP: geographical and temporal characteristics, motives, consequences, origins, event nature (pacific/violent), etc.. All these data have been considered within the same ontology in order to visualize the dependencies, if any, between the attributes, participants and events, which can be a powerful application for sociologists.

Classes. Class hierarchy is based on the analysis of 2000 unique news headlines that were crawled using combinations of keywords from two sets. The first set contains words like "protest", "demonstration", "piqueting", "boycott", "march" etc., the second - words like "against", "contra", "in support of", etc.. SSP classes are taken into account, as well as the wide variety of resources on the Web, including the interactive access to DBpedia ontology classes and relations (gFacet tool: http://visualdataweb.org). Firstly, lists of events constituting protest activity, as well as those preceding and following it, have been built, analyzed and

organized into a structure. Secondly, other parameters, such as geospatial, temporal data and event status have been added .

Protest activity is divided into *verbal expression* and actual *action*, which can be a *mass gathering* or a *symbolic act*. These classes are not disjoint, because such events may anticipate or follow one another. Also, an *action* can be *pacific* or *violent* (or it turns out to be violent (*ViolentConsequence*), we have not encountered any example of the opposite). A violent action may involve the use of weapons (*WeaponType* class). *ActionReason* divides actions into the expressions of protest, support, requirement, conmemoration or other. All of these reasons (support, conmemoration, requirement) can imply a protest. The scale of the action is measured by the amount of actions (single/multiple), amount of participants (one/group/many), location coverage (town/province/country/world). The status of the action is planned/in_progess/finished/never_took_place. *ActionType* includes *Strike, HungerStrike, BusinessStrike, March, Concert, Picketing, MassDemonstration, Riot, Rebellion, Revolution,* symbolic acts, etc.. *Motivation/Demotivation* classes are based on the reported data on event support (financing or other) or rejection by governmental or non-governmental actors. The events that precede and follow the action are put in the corresponding classes that describe threats, warnings, authority interventions into the planning process, different kinds of consequences (financial, property damage or other) and reactions of governmental and non-governmental actors. Participants include individuals, unnamed and named groups of people, governing authorities, political parties, church representatives, enterprises, law enforcement, etc., that can be initiators, targets, victims, support or participants of a protest event. We present a screenshot of the ontology that does not cover the entire structure, because the latter is subject to a control revision. The current version includes 13 first-level classes, 71 second-level classes and 102 third-level classes. A screenshot of the ontology visualization in the Protégé-4.3 framework is presented in Fig. 1.

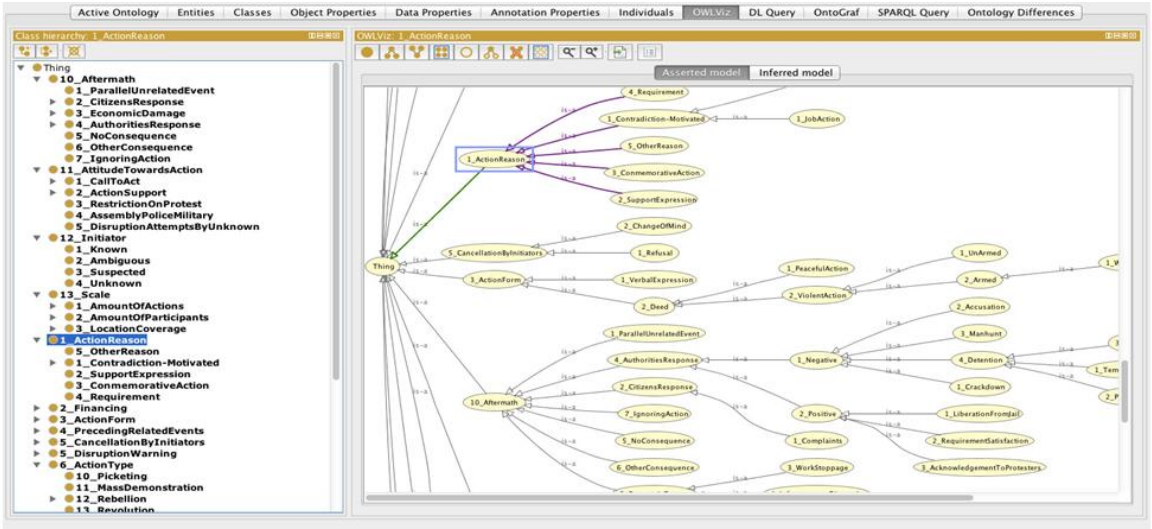

Fig. 1. Class hierarchy implemented in Protégé-4.3

Properties. Most ontology classes are characterized by the mutual influence, which can be manually defined in Protégé by means of specific rules (object property hierarchy). It currently includes a basic set of rules, such as HasParticipant, InitiatedBy, HasProperty, HasMotivation, HasReason, HasConsequence, IsFollowedBy, IsPrecededBy, etc.

Text Annotation. Annotation is the first step to transforming the unstructured text into quantitative event data. In our study we focus on ontology-based annotation and, as the future work, we consider the ontology

population from the annotated data. At the current stage we have a manually constructed social protest gazetteer that is mapped to the ontology within the GATE Developer 8.0 (Fig. 2).

We also use embedded Russian lexicons (Lang_Russian package) that contain lists of geographical terms, date/time expressions, named entities, such as person names, person titles, organization names (government, companies, etc.). The annotation tool relies on JAPE (Java Annotation Pattern Engine) rules. JAPE finite state transducers consist of a left-hand side (LHS) that sets pattern constraints and a right-hand side (RHS) that contains annotation commands. In our experiments we apply cascaded grammars over annotations. Rules take into account the preprocessing results, OntoGazetteer, Russian gazetteer and Flexible gazetteer lookups, PoS tagging, discourse structure, etc.

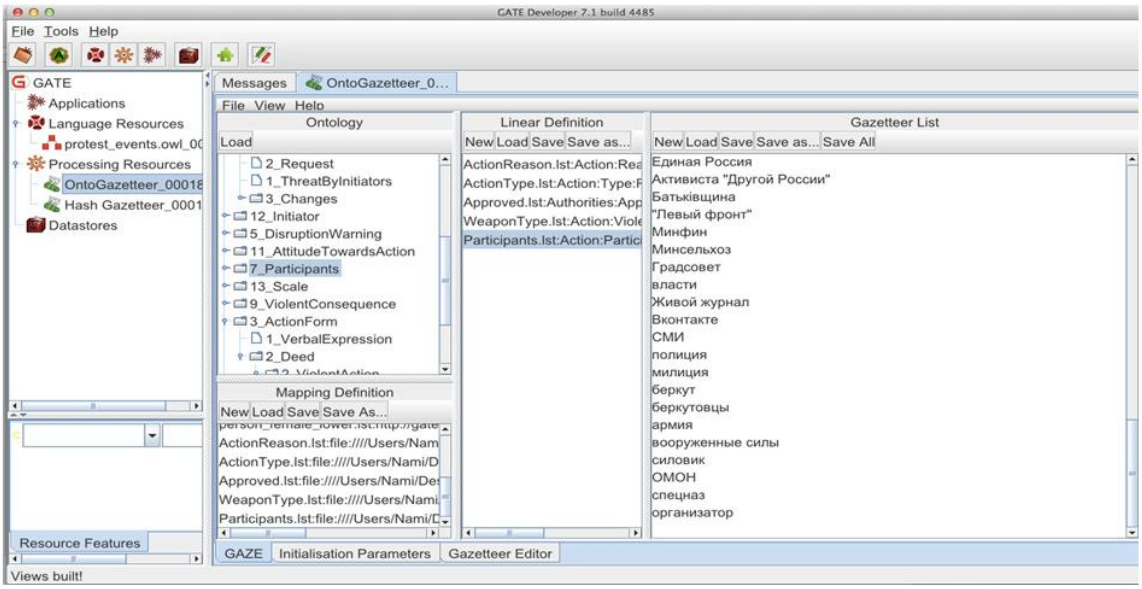

Fig. 2. OntoGazetteer

Annotation Experiments

Annotation rules. Within the present work, we have carried out several experiments on annotating word sequences that characterize the origins of protest activity. The *ActionReason* class representation commonly consists of two components: OntoGazetteer component that defines the action nature (protest, support, conmemoration, requirement or other) and a word sequence of variable length that contains new information, which needs to be categorized for automatic ontology population. Also, the data on the origins of protest activity can be represented as a prepositional or postpositional adjective to the protest type substantive: "антивоенный протест" ("*a protest against war*"), "митинг неонационалистов" ("demonstration of neonationalists"), etc..

Within the framework of the present experiments, we performed the annotation using two sets of rules and OntoGazetteer lookup. As it turned out, *ActionReason* occupies the final position in the headline in a half of the dataset: <*ActionReason*><End Point> - 46%, that is why the first set includes simple and robust rules that rely on the positional properties of this information block. The main pattern constraints are as follows: a sequence is annotated if it is preceded by [ActionReason Lookup], contains any number of tokens and is followed by the sentence end, a verb in indicative mood or a coma. [ActionReason Lookup] includes words denoting protest, support, conmemoration, requirement: "в поддержку" ("in support of"), "в защиту" ("in

defence of"), "против" ("against"), etc.. A screenshot of the annotation based on the first set of rules (ARverb and ARpunkt) is presented in Fig. 3.

The second set are rules represented in cascaded grammar phases. The first phase filters the tokenizer results. The subsequent phases include pattern/rule pairs for sequential processing of the following: [ActionType Lookup] + [Participant rule]/[Random token sequence rule] + [ActionReason Lookup] + [Noun Phrase rule]. [ActionType Lookup] includes words like "протест" ("protest"), "марш" ("march"), "пикет" ("piqueting"), etc.. [Participant rule] defines the annotation of event participants and relies on the OntoGazetteer lookup and morphological analyzer results. [Random token sequence rule] extracts a random word sequence between ActionType and ActionReason Lookup annotations. [Noun Phrase rule] extracts the sought-for data on event origins that is commonly represented as a complex noun phrase. These rules do not take into account the position of information blocks and rely solely on the gazetteer lookups, tokenizer and PoS tagger results.

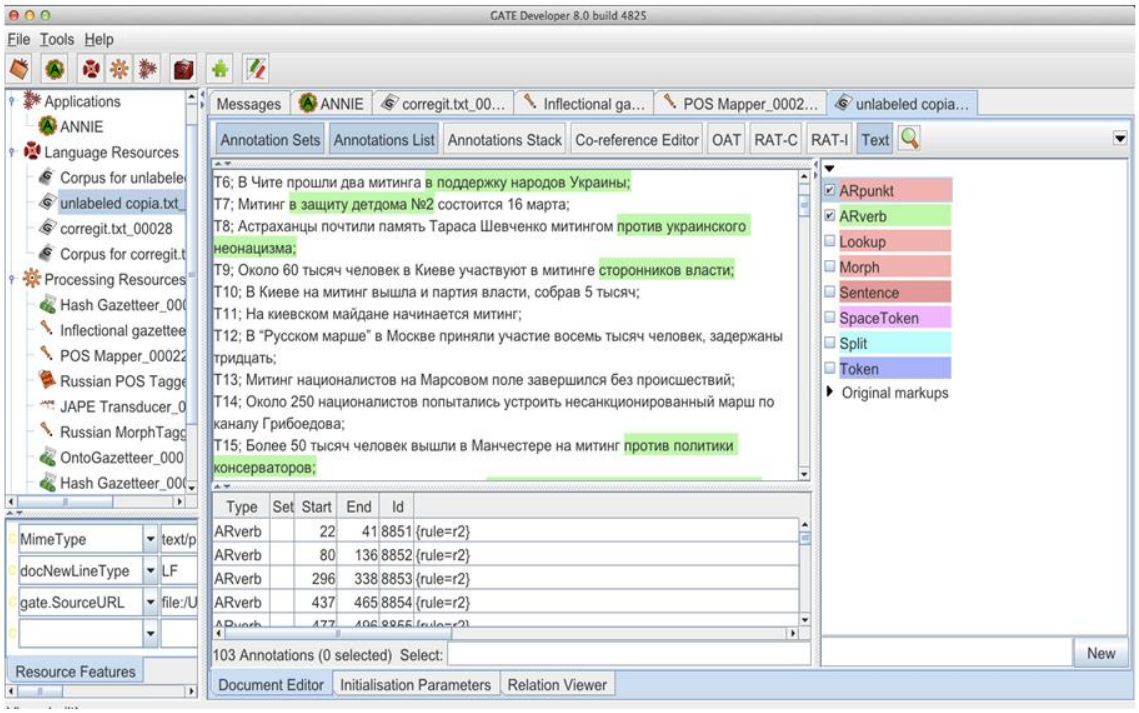

Fig. 3. "ActionReason" class annotation

Gold standard. A gold standard (553 headlines) is built to check the performance of the *ActionReason* class annotation rules at the present stage. In the Fig. 4 an example of the gold standard is given, where T[1...*n*] denotes the number of the sentence, ARverb and ARgazetteer are rules, which must trigger the class annotation, N/A is set if no rule is applicable. The meaning of the annotated strings is as follows: "в поддержку народов Украины" ("*in support of Ukranian peoples*"); "в защиту детдома №2" ("*in defense of the orphanage No2*"); "против украинского неонацизма" ("*against Ukranian neonationalism*"); "сторонников власти" ("*of the pro-governmental activists*"); "националист" ("*nationalists*"); "против политики консерваторов" ("*against the conservative politics*").

```
T6; ARverb: в поддержку народов Украины;
T7; ARverb: в защиту детдома №2;
T8; ARverb: против украинского неонацизма;
T9; ARverb: сторонников власти;
T10; N/A;
T11; N/A;
T12; N/A;
T13; ARgazetteer: националист;
T14; ARgazetteer: националист;
T15; ARverb: против политики консерваторов;
```

Fig. 4. Gold standard

Evaluation. Within the framework of the present experiments, F_1 score (standard harmonic mean of Precision and Recall) of *ActionReason* class annotation has been calculated for two sets of rules. The test set (553 headlines), as well as the gold standard, are divided into five subsets.

$$Precision = \frac{|G \cap C|}{|G|},$$

where G is the number of sequences that were extracted from all the headlines for the *ActionReason* class, C is the amount of strings that coincide with the expert annotation of the same slot.

$$Recall = \frac{|G \cap C|}{|E|},$$

where E is the total amount of sequences that are relevant to the *ActionReason* class within a given test set, according to the expert annotation.

$$F_1 = 2\frac{PR}{P + R},$$

where P is the resulting Precision value for a given test set, and R is the resulting Recall value for a given test set.

Results and Future Work

Results. The annotation of 553 headlines divided into 5 test sets with the information on protest origins (*ActionReason* ontology class) has been performed on the basis of two rule sets. The first set uses OntoGazetteer and Flexible Gazetteer Lookups and position-related rule together with few punctuation and morphological constraints. The second set uses more sophisticated rules taking into account data from all available gazetteers, as well as tokenizer, PoS tagger and NP-chunker output. The results are presented in the Tables 2 and 3. The number of headlines per test set is shown in square brackets. A sequence annotation is considered relevant if it corresponds exactly to the expert annotation in the gold standard.

Table 2. Annotation results for the rule set 1

RuleSet_1	TestSet_1 [100]	TestSet_2 [100]	TestSet_3 [100]	TestSet_4 [100]	TestSet_5 [153]	Total [553]
Retrieved & Relevant	69	60	63	56	87	335
All Relevant	76	71	68	67	97	379
All Retrieved	80	78	77	70	96	401

Table 3. Annotation results for the rule set 2

RuleSet_2	TestSet_1 [100]	TestSet_2 [100]	TestSet_3 [100]	TestSet_4 [100]	TestSet_5 [153]	Total [553]
Retrieved & Relevant	71	66	66	65	91	359
All Relevant	76	71	68	67	97	379
All Retrieved	75	69	67	66	96	373

The results show that RuleSet_1 annotates more irrelevant sequences, while the number of correctly labeled instances is rather high. RuleSet_2 uses many constraints, which allows to annotate more exact sequences, however, in terms of the runtime it performs slightly slower. Manual checking shows that in case of RuleSet_1 most "incorrect" sequences are noisy, but relevant. Evaluation of annotation results is presented in Tab. 4.

The obtained results suggest finding a compromise between RuleSet_1 and RuleSet_2, so that we can reduce the number of constraints and increase the number of correctly annotated sequences. Also, the annotations of other classes that overlap with the *ActionReason* class annotation should be taken into account.

Table 4. Evaluation

Measure/Rule Set	RuleSet_1	RuleSet_2
Precision	0.83	0.96
Recall	0.88	0.94
F1 score	0.85	0.94

Future Work. The present paper describes the creation of knowledge resources for the automatic annotation of events related to the protest activity. The following tasks are proposed as future work: 1) control checking of the ontology by a domain expert; 2) automatic ontology building on the same data; 3) improvement and evaluation of patterns for the corresponding ontology classes; 4) automatic annotation-based ontology population; 5) language coverage improvement for comparison issues: Spanish.

Bibliography

[Braha, 2012] D.Braha. A Universal Model of Global Civil Unrest, PLoS ONE 7(10): e48596, 2012.

[Dementieva, 2013] I.N.Dementieva. Theory and methodology of social protest study, Journal of Public Opinion Monitoring, Vol. 4 (116), 3-12, 2013.

[Harris, 1970] Z. Harris. Papers in Structural and Transformational Linguistics, Dordrecht/ Holland: D. Reidel., x, 850 pp., 1970

[Hayes and Nardulli, 1949] M. Hayes, P. F. Nardulli. SPEEDs Societal Stability Protocol and the Study of Civil Unrest: an Overview and Comparison with Other Event Data Projects (white paper). Cline Center for Democracy, University of Illinois at Urbana-Champaign, 2011.

[Lieto, 2008] A. Lieto. Manual and semi-automatic domain-specific ontology building (master thesis), Università degli studi di Salerno, 2008.

[Nardulli et al., 2013] P. F. Nardulli, M. Hayes, J. Bajjalieh. The SPEED Projects Societal Stability Protocol: An Overview (white paper), Cline Center for Democracy, University of Illinois at Urbana-Champaign, 2013.

[Toledo-Alvarado et al., 2012] J. I. Toledo-Alvarado, A. Guzmán-Arenas, G. L. Martínez-Luna. Automatic Building of an Ontology from a Corpus of Text Documents Using Data Mining Tools, Journal of Applied Research and Technology, Vol.10, No. 3, 398-404, 2012.

[Wunderwald, 2011] M.Wunderwald. *Event Extraction from News Articles (Diploma Thesis)*, Dresden University of Technology. Dept. of Computer Science, 2011.

Authors' Information

Vera Danilova – PhD student at the Autonomous University of Barcelona, Dept. of Romance Languages; Junior research fellow at the Russian Presidential Academy of National Economy and Public Administration. E-mail: maolve@gmail.com

Major Fields of Scientific Research: Multilingual Event Extraction, Ontology Building, Sociological Applications