Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

FACTRUEVAL 2016: EVALUATION OF NAMED ENTITY RECOGNITION AND FACT EXTRACTION SYSTEMS FOR RUSSIAN

Starostin A. S. (astarostin@abbyy.com)[1],
Bocharov V. V. (bocharov@opencorpora.org)[2,3],
Alexeeva S. V. (sv.bichineva@gmail.com)[2,3],
Bodrova A. A. (anastasie.bodrova@gmail.com)[2,3],
Chuchunkov A. S. (alex.chuchunkov@gmail.com)[2],
Dzhumaev S. S. (sdzhumaev@abbyy.com)[1],
Efimenko I. V. (veassi@mail.ru)[4],
Granovsky D. V. (dima.granovsky@gmail.com)[2],
Khoroshevsky V. F. (v.khor@mail.ru)[5],
Krylova I. V. (krylova93@gmail.com)[2],
Nikolaeva M. A. (mary.nikolaeva@gmail.com)[2],
Smurov I. M. (ismurov@abbyy.com)[1],
Toldova S. Y. (stoldova@hse.ru)[6]

[1]ABBYY, Moscow, Russia
[2]OpenCorpora.org
[3]St. Petersburg State University
[4]Semantic Hub, Moscow, Russia
[5]Dorodnicvn Computing Centre, RAS (CC RAS)
[6]Russian Research University—Higher School of Economics

In this paper, we describe the rules and results of the FactRuEval information extraction competition held in 2016 as part of the Dialogue Evaluation initiative in the run-up to Dialogue 2016. The systems were to extract information from Russian texts and competed in two named entity extraction tracks and one fact extraction track. The paper describes the tasks set before the participants and presents the scores achieved by the contending systems. Additionally, we dwell upon the scoring methods employed for evaluating the results of all the three tracks and provide some preliminary analysis of the state of the art in Information Extraction for Russian texts. We also provide a detailed description of the composition and general organization of the annotated corpus created for the competition by volunteers using the OpenCorpora.org platform. The corpus is publicly available and is expected to evolve in the future.

Key words: information extraction, evaluation, named entity recognition, fact extraction, relation extraction

FACTRUEVAL 2016: ТЕСТИРОВАНИЕ СИСТЕМ ВЫДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ И ФАКТОВ ДЛЯ РУССКОГО ЯЗЫКА

Старостин А. С. (astarostin@abbyy.com)[1],
Бочаров В. В. (bocharov@opencorpora.org)[2,3],
Алексеева С. В. (sv.bichineva@gmail.com)[2,3],
Бодрова А. А. (anastasie.bodrova@gmail.com)[2,3],
Чучунков А. С. (alex.chuchunkov@gmail.com)[2],
Джумаев С. С. (sdzhumaev@abbyy.com)[1],
Ефименко И. В. (veassi@mail.ru)[4],
Грановский Д. В. (dima.granovsky@gmail.com)[2],
Хорошевский В. Ф. (v.khor@mail.ru)[5],
Крылова И. В. (krylova93@gmail.com)[2],
Николаева М. А. (mary.nikolaeva@gmail.com)[2],
Смуров И. М. (ismurov@abbyy.com)[1],
Толдова С. Ю. (stoldova@hse.ru)[6]

[1]ABBYY, Москва, Россия
[2]Проект «Открытый Корпус»
[3]Санкт-Петербургский государственный университет
[4]Semantic Hub, Москва, Россия
[5]Вычислительный центр им. А. А. Дородницына РАН
[6]Высшая школа экономики, Москва, Россия

Статья описывает правила и результаты соревнования по извлечению информации из русскоязычных текстов FactRuEval, проводившегося в 2016 г. в рамках инициативы Dialogue Evaluation, приуроченной к конференции Диалог 2016. Соревнование включало две дорожки по извлечению именованных сущностей и одну дорожку по извлечению фактов. В работе излагаются задачи, ставившиеся перед участниками соревнования, и приводятся оценки качества работы систем участников. Кроме того, обсуждаются методы оценки качества для всех трех дорожек и делаются предварительные выводы о современном положении дел в области извлечения информации для русского языка. Особое внимание в работе уделяется составу и устройству размеченного корпуса, созданного в процессе проведения соревнования усилиями волонтеров в рамках платформы OpenCorpora.org. Этот корпус является общедоступным и его планируется развивать в дальнейшем.

Ключевые слова: извлечение информации, тестирование систем, выделение именованных сущностей, выделение фактов, выделение отношений

1. Introduction

At the beginning of 2016, a competition was held for systems capable of extracting information from Russian texts. The competition was given the name FactRuEval and was part of the Dialogue Evaluation initiative. This paper is a report on the results of the competition. As organizers of the competition, we set ourselves three main goals:

- Create an infrastructure for regular evaluation of information extraction systems
- Hold the first evaluation event to analyze state-of-the-art Russian-language information extraction systems
- Create a publicly available corpus that can be used for evaluation and further development of information extraction systems

All of the above goals were successfully achieved. Using the OpenCorpora.org platform, we created a technology for annotating corpora geared towards information extraction needs. A generalized annotation model was developed, which was then used to create a gold-standard markup for several classes of information extraction tasks. Additionally, comparator software was developed, enabling automated comparisons of test markups with the gold standard. Both the annotation model and the comparator software can be used for future evaluations with multiple domain-specific tracks.

Using the evaluation infrastructure mentioned above we held three tracks—one involved the classic task of named entity recognition in the tradition of MUC (see, for example, [Grishman and Sundheim, 1996]) and CoNLL (see, for example, [Tjong Kim Sang and De Meulder, 2003]), another evaluated recognition of named entities with attributes[1], and the third was a competition of fact extracting systems. In the "Tracks" section below, we provide a brief description of the tasks set before the participants in each of the three tracks. A total of thirteen systems were enrolled, some of them being commercial software and some developed for research purposes. Since the competition was anonymous, the "Participants" section contains only general information about the systems without giving their exact names. The "Results" section lists the scores achieved by the competing systems. These scores were obtained using the comparator software mentioned above. The principles underlying the comparator tool are described in the "Evaluation Methods" section.

An annotated corpus that was used during competition contains a demo (or training) subcorpus and a test subcorpus. Both of them were annotated by volunteers using the OpenCorpora.org platform. The "Corpus and Markup" section describes the generalized model used by the annotators, provides some statistics, and discusses the handling of disagreements that arose among the annotators when dealing with some entities. The corpus is publicly available for download and use. Plans for expanding and improving the corpus are also discussed in "Corpus and Markup" section.

---

[1] By "named entity with attributes recognition" we mean recognition of the entity mentions along with specification of simple string values of particular attributes (for example, surname or first name for the Person entity). This task differs strongly from the relation or fact extraction task because entity attributes may have only string values. Moreover such values are always based on some internal parts of the whole entity mention.

Due to space constraints, we have been unable to provide complete details in some of the sections. Readers requiring more information are invited to visit https://github.com/dialogue-evaluation/factRuEval-2016, where they will find all the competition materials.

2. Related Work

The organization of the competition was based on the experience of the international evaluation events devoted to Named entities recognition (NER), relation detection and fact extraction tasks. The first evaluation event devoted to these tasks was inspired by DARPA and was held at Message Understanding Conference (MUC) in 1987–1997 (see, for example, [Grishman and Sundheim, 1996]). Initially, information extraction tasks focused on military messages and information concerning terrorist activities. Later, the focus shifted to newswire articles, from which not only military but also economic information was extracted. It was at the MUC events that the first evaluation principles were laid down and guidelines were developed for the creation of gold-standard annotated corpora enabling comparisons of different information extraction system. Starting from 1999, these tracks in the evaluation events have become part of the Automatic Content Extraction (ACE) program. Detailed descriptions of the tasks, data, and rules over the years are available at https://www.ldc.upenn.edu/collaborations/past-proiects/ace/annotation-tasks-and-specifications (see, for example, [Doddington et al., 2004]). The ACE datasets have included not only English texts, but also texts in Arabic and Chinese. Similar tasks have also been set by CoNLL (see, for example, [Tjong Kim Sang and De Meulder, 2003]).

From 2009 onwards, named entity, relation, and fact extraction tasks were also set in the Knowledge Base Population section of the Text Analysis Conference (TAC), http://www.nist.gov/tac/publications/ (see, for example, [Surdeanu 2013]). The TAC tasks are formulated somewhat differently from ACE's and require some transformation of extracted information into structured data. Contending systems have either to populate a database with information about the objects they detect in the texts and relations among them (Cold Start KBP), or link mentions in the texts to the relevant objects in a database. The task is known as Tri-Lingual Entity Discovery and Linking. There are also relation detection and event extraction tracks in TAC competitions.

Evaluation methods employed within different contests vary from the simple procedure used by CoNLL to the complicated methods employed by ACE. Under the CoNLL rules, only the exact matches between the text fragments detected by a system and those in the gold standard are considered as true positives, with false positives not penalized. Given the lack of gold-standard corpora for Russian and considerable variation in the standards adopted for the existing Russian-language systems, we decided against using this direct evaluation method in the FactRuEval competition.

ACE uses a more sophisticated algorithm [NIST: ACE08 Evaluation Plan], with different named entities assigned different weights, making the interpretation and comparison of results complicated [Nadeau and Sekine, 2007]. The evaluation method used by FactRuEval is largely similar to that used by ACE, allowing one and the same entity to be marked up in several different ways. Scores are computed by assessing

how close a participant's markup is to that in the gold standard (see the "Evaluation Methods" section for details).

We also took into consideration the experience of evaluation events for named entities recognition and relation and fact extraction for some particular languages. For example, such an event (EVALITA (http://www.evalita.it, see, for example, [Caselli et al., 2014]) is held annually for Italian.

In Russia, the first competitions for fact detection systems were held from 2004 to 2006 as part of the ROMIP workshop (http://romip.ru/). The ROMIP tasks in 2005 and 2006 involved named entity recognition and fact extraction (employers' names, ownership of an organization, see [Nekrestjanov and Nekrestjanova, 2006]). The systems had to select text fragments where a certain event was mentioned. However, the ROMIP events attracted only a small number of participants (with only two systems competing in the 2006 fact detection track). When developing the tasks for the 2016 FactRuEval competition, we also drew on the experience gained from the relevant ROMIP tracks.

The number of systems working with texts in Russian has grown considerably in the past few years. Starting from 2010, a number of RU-EVAL events have been held, where certain automated text processing modules are evaluated, including morphology and syntax modules (see [Toldova et al., 2015], http://ru-eval.ru/new/). The RU-EVAL events have provided an insight into the current state of the art in automated processing of Russian texts at the basic levels of linguistic analysis and resulted in the creation of datasets on which text processing systems can be tested.

Being focused on information extraction tasks, FactRuEval is the next step in the direction provided by RU-EVAL. Due to a lack of generally agreed on standards for detecting named entities and facts in Russian texts, we opted for more flexible annotation and evaluation principles similar to those used by ACE. We developed a software tool for creating corpora with sophisticated annotation and organized annotation work on a gold-standard corpus featuring the very basic entities. Following in the footsteps of CoNLL, we also developed and made publicly available a software tool that automatically compares the results obtained by competing systems against the gold standard.

3. Tracks

There were three tracks in FactRuEval, with the systems competing in named entities recognition, extracting entities with attributes, and extracting facts. Participants could enroll their systems in any of the tracks. The rules for each track are described in detail in a separate document which is available on the FactRuEval website, so here we provide only a brief overview.

3.1. Named entity recognition track

This track posed the classic task of Named Entity Recognition [Grishman and Sundheim, 1996]. The competing systems had to locate the mentions of entities of particular types. Entities of the following three types had to be recognized:

1. Person
2. Organization
3. Location (place names)[2]

3.2. Named entities with attributes recognition track

In this track, the participants had to list unique named entities of specific types detected in the texts. For each entity, certain string fields had to be filled with normalized values if the corresponding information was available in the texts.

Normalization required transforming phrases to their canonical form. In most cases, this meant recasting the corresponding text fragment in the nominative case (preserving the grammatical agreement between the elements).

The types of entities in this track were identical to those in the entity extraction track[3].

An essential requirement was that the lists of named entities generated by the systems should not contain any duplicates. By prohibiting duplicates we could evaluate the systems' ability to locally identify the referents of named entities, which has important practical applications. When scoring the results in this track, both inclusion of duplicates ("underidentification") and collapsing different entities into one ("overidentification") were penalized.

3.3. Fact extraction track

In this track, the participants had to detect facts of specific types in the texts. A fact is a relation between several entities (a mention of a certain type of situation, with participants playing certain roles). Only those facts had to be extracted which were *explicitly* mentioned in the texts. The fact fields had to be filled with string values allowing unambiguous identification of the corresponding named entities. Facts of the following types had to be extracted:

- Occupation (employment of a person by an organization[4])
- Deal (some interaction of economic nature between people or organizations)[5]

---

[2] In one variation of the track, the participants had to distinguish between normal mentions of locations and mentions of locations in what may be termed "organization uses" (for example, *"Russia announces counter-sanctions"*). A similar class of entities was given the tag GPE in CoNLL competitions ([Tjong Kim Sang and De Meulder, 2003]).

[3] In this track, organization uses of locations were treated as ordinary locations.

[4] In this track, all contexts that were allowed for organizations were also allowed for "organization uses" of locations.

[5] We had planned to work out a classification of deals while annotating the corpus and give participants the chance to compete in identifying different subtypes of deal. Unfortunately, the selected texts contained too few mentions of deals and we had to give up the idea. For next year's competition we are planning to collect a specialist corpus containing a sufficient number of mentions of various types of deal.

- Ownership (of an organization by a person or organization)
- Meeting (of two or more people)

It is important to note that the competing systems were expected to extract fact instances at text level rather than at sentence level. Different field fillers could be mentioned in different sentences linked by anaphora. For example, from the text fragment *"Russian Milk Ltd. is a highly profitable company. This Friday, it was bought by the famous entrepreneur J. J. Ivanov"* the participants had to extract the fact of a deal between two parties (Russian Milk Ltd. and J. J. Ivanov)[6]. The most complex variation of this track required that the participants distinguish between actual facts[7] and all other kinds of facts (i.e. facts mentioned in negative, future, conditional, etc. contexts).

4. Corpus and Markup

An important objective of the FactRuEval project was to create an open annotated corpus of Russian texts that could be used for future evaluations. To achieve this, we had to develop an annotation model that would cover the main tasks solved by information extraction systems. Such a model was successfully developed and subsequently used to annotate 255 documents. We will first describe the annotation model and then provide some statistics for the annotated corpus.

4.1. Annotation model

The markup has four layers. The first two layers contain annotated mentions of entities, the third layer contains coreference relations, and the fourth layer groups entities into facts. The entity markup (the first two layers) was used to evaluate named entity extraction systems competing in the first two tracks. The results shown in the second track were additionally evaluated using the third markup layer. The fact extraction systems were evaluated using all four markup layers.

As the FactRuEval 2016 tracks sometimes allowed multiple versions of correct or partially correct markup, it was decided to include several markup variants for some of the objects in the test corpus.

4.1.1. Layer 1 markup: spans

In layer 1, typified spans have been marked up in the texts. These are **continuous chains of words** labelled with one or more predefined tags (e.g. "surname", "org_name", "loc_descr"). It is assumed that each type of marked up object has its own set of tags (i.e. types of spans). For example, in the case of people, we had to distinguish

[6] All facts that required anaphora resolution for their successful extraction were marked as "difficult to extract", resulting in two scores for fact extraction—one for extracting easily detectable facts (i.e. those that were stated in their entirety in one sentence) and one for extracting all facts.

[7] Or, to be more precise, facts represented in the texts as having actually occurred.

first names, surnames, patronymics, and nicknames, while in the case of organizations and locations, we had to distinguish object names (“org_name” and “loc_name”) and the object descriptors (“org_descr” and “loc_descr”)[8].

4.1.2. Layer 2 markup: object mentions

In layer 2, spans are grouped into **object mentions**. Object mentions are also typified. The types of object mentions correspond to the types of entities involved. For example, the following already mentioned types of entity were marked up: people, organizations, locations, and organization uses of locations.

Several mentions may share common spans. This most commonly occurs when annotating coordinated items where two mentions of an object share a common descriptor or where two people are mentioned sharing the same surname (see Fig. 1).

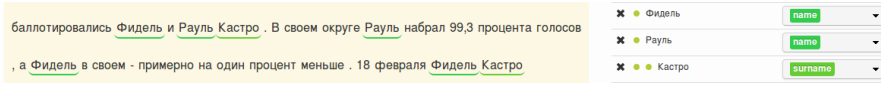

Fig. 1. Two mentions sharing a common descriptor

4.1.3. Layer 3 markup: coreference

In layer 3, object mentions from layer 2 contained in the same text and having the same referent are grouped together (see Fig. 2). Such group is called an identified object. Each group may be linked to an object identifier in an external database (e.g. Wikidata).

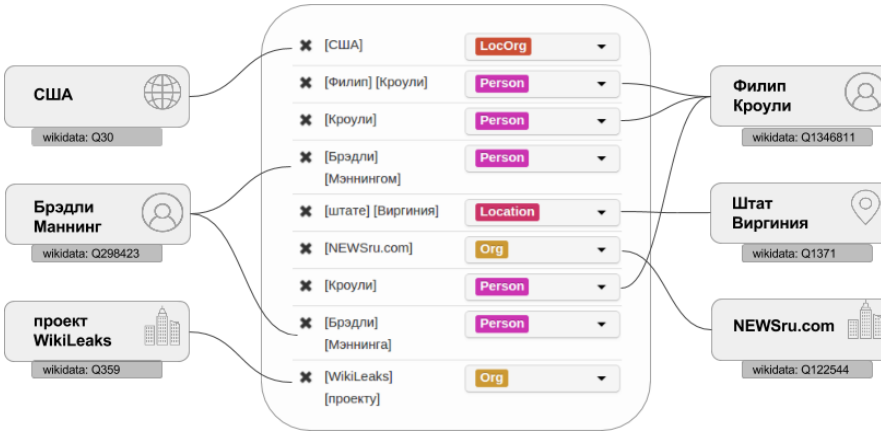

Fig. 2. An example of an identified object

[8] Descriptors are words or word combinations denoting a superordinate concept. For example, “company" is a descriptor in the phrase *“XYZ company"*. Distinguishing between names and descriptors makes it easier to mark up discontinuous mentions of entities. For example, in *“Michurin and Lenin Avenues"* the mention of Michurin Avenue is discontinuous.

4.1.4. Layer 4 markup: facts

A fact is a typified relation between multiple identified objects mentioned in a text. The type of a fact determines what fields it may have. Each field has a name and a list of possible types of object that may fill it. Besides object fields, a fact have string fields. These are filled not by arbitrary strings but by sets of spans (in the general case, by multiple sets of spans, as markup variants are allowed). In a sense, such sets of spans may be considered mentions of virtual objects, i.e. objects that need not be annotated in layers 2 and 3. Fig. 3 illustrates an Occupation fact.

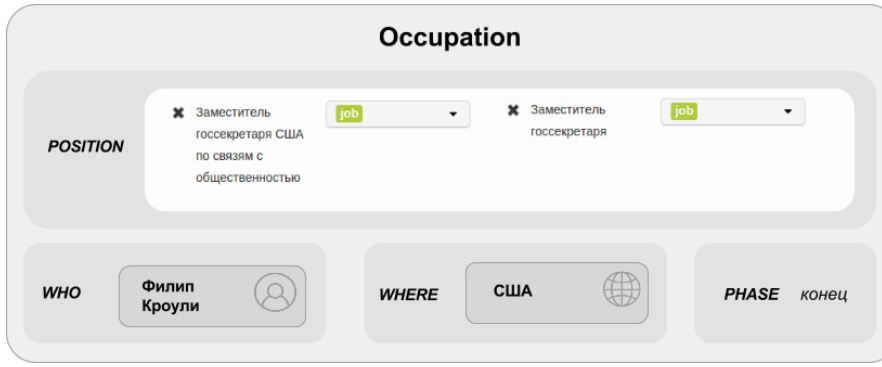

Fig. 3. An example of a fact

It is important that, unlike mentions of objects, facts have no direct links with the original text. They are related to the entire text rather than to a specific text fragment. This approach was adopted because any attempt to link a fact with a specific text fragment often causes disagreement among human annotators and makes it hard to lay down clear requirements for competing systems.

Firstly, it is often the case that a fact is expressed across multiple sentences by means of various anaphoric devices. Secondly, sometimes a fact may logically follow from the text without being explicitly stated[9].

4.2. Corpus characteristics

The FactRuEval corpus consists of newswire and analytical texts in Russian dealing with social and political issues. The texts were gathered from the following sources:

- Private Correspondent (http://www.chaskor.ru/)
- Wikinews (https://ru.wikinews.org)

The corpus was split into two parts—a demo corpus of 122 texts and a test corpus of 133 texts. The demo corpus had been sent out to the participants before the start of the competition. The participants could both test and train their systems on this

[9] This case was excluded from the competition, but obviously we had to keep it in mind when developing the annotation model.

corpus. The test corpus was made available to the participants once the competition ended. During the competition, the participants received a collection of approximately 30,000 documents, which also included documents from the test corpus (of course, the participants did not know which documents came from the test corpus). The text statistics are provided in Table 1. The markup statistics are provided in Table 2.

Table 1. Text statistics

Total texts		**Total characters**		**Total tokens**		**Total sentences**	
Demo Set	**Test Set**	**Demo Set**	**Test Set**	**Demo Set**	**Test Set**	**Demo Set**	**Test Set**
122	133	189,893	460,636	30,940	59,382	1,769	3,138

Table 2. Markup statistics

Spans		**name**		**surname**		**patronymic**		**nickname**		**loc_name**	
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
4,084	**7,670**	531	810	691	1,268	15	27	12	56	1,067	1,367
		loc_descr		org_name		org_descr		job		Other	
		Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
		127	194	637	1,628	497	1187	471	915	36	218
Objects		Person		Location		Org		LocOrg		Other	
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test	Demo	Test
2,611	**5,019**	741	1,388	529	728	787	2,034	553	846	1	23
Facts		Occupation		Deal		Ownership		Meeting			
Demo	**Test**	Demo	Test	Demo	Test	Demo	Test	Demo	Test		
273	**786**	211	444	30	114	17	176	5	51		

The following approach was used to annotate the data for the first two layers. First, the organizers drew up guidelines for annotating each type of entity (these guidelines are available on the FactRuEval website). Next, each paragraph was independently annotated by four volunteer annotators. Any disagreements were resolved through moderation by an expert appointed by the organizers. Layer 2 and 3 annotation was carried out by experts appointed by the organizers. For the demo portion of the corpus, the markup for all the four layers was made available to the participants prior to the competition. The participants had the opportunity (of which they made frequent use) to discuss on the FactRuEval website any problems that they encountered. Following up on these discussions, we made every effort to make the necessary changes to the markup. The markup of the test portion of the corpus was disclosed after the competition ended. The participants had the opportunity to lodge an appeal and state their case for correcting the markup via the FactRuEval website. Some corrections were made to the markup after such appeals, following which the scores were recalculated and finally declared.

The distributed iterative annotation process described above resulted in highly accurate markup. Some problem cases persisted, but for the overwhelming majority of instances a consensus was reached among the annotators, the organizers, and the participants. LocOrg was the most contentious type of entity. There was a lot of debate

as to when it should be treated as an organization and when purely as a location. We had expected LocOrg to cause some controversy, but still decided to have this entity marked up by way of experiment. To account for this uncertainty, we provided a scoring mode in the first track which treated LocOrg as location proper. In this evaluation mode, the three entities have their usual interpretations and the first track is no different from similar evaluations previously conducted in other competitions. The results for the two evaluation modes are provided in the "Results" section.

Unlike the entity markup, the factual markup has certain significant defects, which we hope to rectify before next year's competition. Firstly, too few facts were marked up due to time and labour constraints. In the demo corpus, the fact of employment (Occupation) is the best marked up fact. The other facts are few and far between. The test corpus is slightly better in this respect, but the marked up facts are still far from being representative. Secondly, the demo and test corpora are out of sync in terms of the number of facts they contain. Finally, all facts were marked up by only one (albeit a highly skilful one) annotator. In view of the above, the current factual markup should be treated as a preliminary version which enabled us to carry out a test run of the fact extraction track. For next year's competition, we are planning to resolve these issues and re-run the fact extraction track.

5. Participants

Initially, over sixty teams had expressed their interest in FactRuEval. However, only thirteen actually took part in the competition (not all of them participated in all three tracks).The competition attracted commercial developers, research teams, and news agencies (Fig. 4).

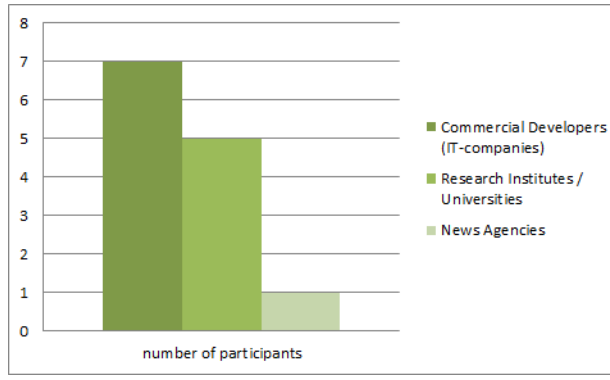

Fig. 4. Types of participants

Many of the participants use the hybrid approach to NLP, i.e. employ both linguistic methods and machine learning. However, when asked by the organizers, most of them described their system either as rule-based or as using machine learning. The rule-based systems prevailed (eight systems were rule-based and five systems used machine learning).

6. Evaluation Methods

This section describes the principles behind the comparator software, which enabled near-automated comparisons of the results. The software was made available to the participants during the competition, allowing them to compare their results against the demo markup. The comparator software is now publicly available. We are planning to improve it and use in future competitions.

The comparator tool examines all possible correspondences between a test markup and the gold-standard markup and chooses the best matches. A good match is a correspondence between objects (i.e. mentions, identified entities or facts) in the gold standard and test markup which meets the following criteria:

- Correspondence must be established between objects of compatible types
- Each object in the test markup must have only one corresponding object in the gold-standard markup
- Each object in the gold standard-markup:
  - Must have only one corresponding object in the test markup (for mentions and entities)
  - May have any number of corresponding objects in the test markup (for facts)

For each pair (or group) of objects, its quality Q, is calculated using a certain formula. Extraction quality of each individual object in the pair (or in the group) is also considered equal $Q(s_i) = Q(t_i) = Q$. The extraction quality of unmatched objects is considered to be zero.

- The obtained values are then used to calculate precision, recall, and F-measure:
- $Precision = \frac{\sum Q(t_i)}{|T'|}, where\ T' = T \backslash T_i$
- $Recall = \frac{\sum Q(s_i)}{|S'|}, where\ S' = S \backslash S_i$
- $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

In each track F-measure is chosen as the final score awarded to participants. Below we describe quality estimation principles for each track.

6.1. Entity recognition quality metric

The quality of a matching pair is calculated using this formula:

$$Q(s,t) = \frac{TP}{TP + FP + FN}$$

where

TP is the total number of tokens that belong to both mentions,

FP is the total number of tokens that belong to the test mention, but do not belong to the gold-standard mention,

FN is the total number of tokens that belong to the gold-standard mention, but do not belong to the test mention[10].

6.2. Entity with attributes recognition quality metric

A gold-standard entity *t* may be represented as a set of pairs $< a_i, v_1^i | \dots | v_{k_i}^i >$ and a test entity *s*—as a set of pairs $<a_i, v_i>$. The quality of a matching pair is calculated using this formula:

$$Q(s,t) = \frac{TP}{TP + FP + FN}$$

where

TP is the number of such pairs $<a, v_1|\dots|v_N>$ of the gold-standard entity *s* that the entity *t* has at least one pair of type $<a, v_i>$, where $i = 1 \dots N$,

FN is the number of such pairs $<a, v_1|\dots|v_N>$ of the gold-standard entity *s* that the entity *t* has no pairs of type $<a, v_i>$, where $i = 1 \dots N$,

FP is the number of such pairs $<a,v>$ of the test entity *t* that the entity *s* has no pairs of type $<a, v_1|\dots|v|\dots|v_N>$[11].

6.3. Fact extraction quality metric

The quality of a group is calculated using two metrics. First, we evaluate the test markup to see how well the system has detected values of fields, using the following metric:

$$ArgQuality(s, \{t_1, \dots, t_k\}) = JaccardIndex(S.T) = \frac{|S \cap T|}{|S| + |T| - |S \cap T|}$$

where

S is the set of all pairs <field,value > in the gold-standard fact of group (*s*),

T is the set of all pairs < field,value > in all the test facts of group (t_i),

$S \cap T$ is the set of all pairs < field,value > in all the test facts of group (t_i), that have been deemed correct, i.e. those for which a match has been found in *S*.

According to the rules of the third track all values of all fields in the test markup are strings, whereas for the gold-standard markup, all object fields are filled with links to already marked up named entities. To find matches for string fields, we simply look for exact matches[12]. To find matches for object fields, the following rule is used:

---

[10] It should be noted that tokens forming spans of certain types were ignored. For more information, please refer to the instruction manual to the comparator tool, which is available on the FactRuEval website.

[11] To evaluate extraction of entities with attributes, a light comparison mode was available, where FP was always assumed to be 0, i.e. redundant attribute values were not penalized

[12] The only exception is job titles. For them, like for object fields, we allow (i.e. include in the gold standard) variants of normalized names (provided they occur in the text), as well as normalized strings matching exactly a job title mention in the text.

A pair <field R, value X> from the test fact is matched with a pair <field R,link to named entity E> from the gold-standard fact if and only if X is either contained in the allowed names of entity E in the gold-standard markup[13], or exactly (letter by letter, without normalization) matches one of the continuous mentions of the entity in the text.

The second metric shows how well the test markup reflects the co-occurrence of various field values of a fact, or, to put it another way, how well the system has joined together the fragments of the detected facts. To count the quality two graphs are examined, one for the gold-standard and one for the test markup. Let the set of nodes in each graph correspond to the set of correctly detected <field,value> pairs that have been grouped together (see above)—$S \cap T$, $|S \cap T| = n$. Let us assume that a pair of nodes in the graph is linked by an arc if and only if there is an instance of the fact in the corresponding markup which simultaneously contains the <field,value> pairs, corresponding to these nodes.

Obviously, due to the constraints imposed on a group, the graph corresponding to the gold-standard markup will be complete. On the other hand, in the second graph there will appear connected components v_1, ..., v_m of the sizes n_1, ..., n_m, and $\sum n_i = n$. Using the sizes of these connected components, we calculate the second metric as a ratio of the numbers of arcs in the two graphs:

$$IdQuality(s, \{t_1, \dots, t_k\}) = \frac{\sum n_i(n_i - 1)}{n(n-1)}$$

The quality of the entire group is calculated as:

$$Q(s, \{t_1, \dots, t_k\}) = \frac{ArgQuality * (1 + IdQuality)}{2}$$

7. Results

In this section, we present the performance scores of the systems for each of the three tracks. All the participating systems are listed under their code names assigned to them upon enrolling the competition. If a participant sent in more than one run for a track, we give only the scores for the run with the highest F-measure.

7.1. Entity extraction scores

As we mentioned earlier, the results in the first track were scored using two different evaluation modes. The first mode required that the participating systems distinguish between mentions of proper locations and organization uses of locations. The

[13] Here, too, there is one exception. For Occupation facts, the comparator tool automatically adds the names of superordinate organizations (if mentioned in the text) to the allowed names of entities that indicate organizations where people work. For example, if, analyzing the sentence "Ivanov teaches at the Department of History at Moscow University", a system extracts the fact of working at Moscow University rather than at the Department of History at Moscow University, this answer will be treated as correct.

second mode ignored this distinction. The first evaluation mode gave rise to a lot of objections from participants, as there was no generically agreed on definition of "organization use". Despite the fact that a lot of effort had been put into working out a precise definition for this use of locations, some of the participants decided against detecting LocOrg entities. In Table 3 below, we provide scores only for those systems whose output included LocOrg entities. The scores were computed using the evaluation mode that distinguished between proper locations and entities of type LocOrg. The best result (0.809) was shown by the system known as *pink*, with *aquamarine* and *crimson* close behind.

Table 4 lists the scores computed without making a distinction between proper locations and entities of type LocOrg. In this evaluation mode, the best result was shown by the system known to the organizers as *violet*. Very close to *violet* were *pink* and *beige*. A case apart is the system known as *grey*, whose developers only sent in the results for entities of type Person, making comparisons with the other systems difficult. The only thing we can say about *grey* is that it is in the top five systems when it comes to detection of people.

Table 3. Entity extraction scores. Location and LocOrg are treated as two different types of entity

	Overall			Person			Location			Organization			LocOrg		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.84	0.78	0.807	0.91	0.92	0.91	0.74	0.70	0.72	0.80	0.73	0.76	0.87	0.67	**0.76**
crimson	0.86	0.74	0.80	0.96	0.88	0.92	0.73	0.60	0.66	0.84	0.69	0.76	0.79	0.72	0.75
orange	0.83	0.74	0.78	0.93	0.87	0.90	0.73	0.69	0.71	0.78	0.66	0.72	0.79	0.73	0.75
pink	0.86	0.76	**0.809**	0.96	0.87	0.91	0.80	0.67	**0.73**	0.86	0.71	0.78	0.74	0.75	0.75
violet	0.79	0.75	0.77	0.94	0.92	**0.93**	0.52	0.86	0.65	0.82	0.76	**0.79**	0.89	0.31	0.46
white	0.74	0.47	0.58	0.95	0.74	0.83	0.43	0.70	0.53	0.87	0.36	0.51	0.06	0.00	0.00

Table 4. Entity extraction. Location and LocOrg are treated as the same type of entity

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Aquamarine	0.88	0.82	0.85	0.91	0.92	0.91	0.96	0.81	0.88	0.80	0.73	0.76
Beige	0.88	0.83	0.86	0.90	0.90	0.90	0.96	0.86	**0.91**	0.81	0.74	0.77
Black	0.86	0.83	0.85	0.91	0.92	0.92	0.96	0.86	0.90	0.74	0.73	0.74
Brown	0.89	0.69	0.78	0.96	0.84	0.90	0.91	0.72	0.80	0.78	0.54	0.64
Crimson	0.92	0.79	0.85	0.96	0.88	0.92	0.96	0.81	0.88	0.84	0.69	0.76
Green	0.90	0.73	0.81	0.93	0.84	0.88	0.95	0.84	0.89	0.82	0.55	0.66
Grey	—	—	—	0.96	0.87	0.91	—	—	—	—	—	—
Orange	0.87	0.78	0.82	0.93	0.87	0.90	0.91	0.84	0.87	0.78	0.66	0.72
Pink	0.92	0.80	0.86	0.96	0.87	0.91	0.94	0.85	0.89	0.86	0.71	0.78
Purple	0.85	0.79	0.82	0.90	0.88	0.89	0.92	0.84	0.88	0.76	0.68	0.71
Ruby	0.88	0.54	0.67	0.92	0.73	0.81	0.89	0.67	0.76	0.78	0.26	0.39
Violet	0.89	0.84	**0.87**	0.94	0.92	**0.93**	0.93	0.87	0.90	0.82	0.76	**0.79**
White	0.93	0.58	0.71	0.95	0.74	0.83	0.93	0.68	0.79	0.87	0.36	0.51

7.2. Scores for extracting entities with attributes

Two evaluation modes were used for this track. The first mode treated redundant attribute values (i.e. those not found in the gold standard) as errors, while the second mode allowed redundancies. The idea behind was to give a fair chance to those systems that made extensive use of encyclopedic information and that, for various reasons, could not remove this information from their output. As it turned out, switching between the two evaluation modes had almost no impact on the scores. Nevertheless, we provide the scores for both modes (Tables 5 and 6). In both cases, the highest F-measure (0.80) was achieved by *pink*. It should be noted, however, that the scores for *violet*, *crimson*, and *aquamarine* were very close to those of *pink*.

Table 5. Scores for extraction of entities with attributes. Redundant attribute values are penalized

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.79	0.79	0.79	0.81	0.86	0.83	0.94	0.82	**0.87**	0.68	0.71	0.70
crimson	0.83	0.73	0.78	0.86	0.82	0.84	0.93	0.75	0.83	0.73	0.64	0.68
green	0.76	0.72	0.74	0.79	0.83	0.81	0.90	0.79	0.84	0.64	0.58	0.61
pink	0.84	0.76	**0.80**	0.89	0.82	0.85	0.90	0.81	0.86	0.76	0.66	**0.71**
violet	0.79	0.78	0.78	0.88	0.86	**0.87**	0.84	0.81	0.83	0.68	0.69	0.68

Table 6. Scores for xtraction of entities with attributes. Redundant attribute values are not penalized

	Overall			Person			Location			Organization		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
aquamarine	0.79	0.79	0.79	0.81	0.87	0.84	0.94	0.82	**0.87**	0.68	0.71	0.70
crimson	0.83	0.73	0.78	0.87	0.83	0.85	0.93	0.75	0.83	0.73	0.64	0.68
green	0.77	0.73	0.75	0.80	0.84	0.82	0.90	0.79	0.84	0.64	0.58	0.61
pink	0.86	0.77	**0.81**	0.91	0.84	0.87	0.91	0.82	0.86	0.78	0.68	**0.73**
violet	0.79	0.78	0.79	0.88	0.86	**0.87**	0.84	0.81	0.83	0.68	0.69	0.69

7.3. Fact extraction scores

Only two systems participated in the fact extraction track, *violet* and *green*. We can think of three possible reasons for this lack of participants. Firstly, fact extraction is a very complicated task, with only a handful of teams working on it. Secondly, some of the teams that could potentially participate in the track do not share some of the organizers' ideas (e.g. the idea that facts should be detected at text level). Thirdly, and, perhaps, most importantly, prospective participants were not satisfied with the corpus of facts offered by the organizers (the known problems associated

with the factual aspect of the corpus are described in the "Corpus features" section). The small size of the demo corpus shut out systems that relied on machine learning and made it difficult to fine-tune rule-based systems. We hope that for next year's competition we will have a larger corpus of facts that will attract more participants.

Despite of the shortcomings described above, the fact extraction results obtained by the two participating systems in this year's competition are of some interest. Particularly important are the results of extracting the Occupation fact. The corpus contains a sufficiently large number of instances of this type of fact, so the results shown by *violet*, the winner of this track, are meaningful and can be considered as today's baseline for fact extraction systems working with Russian texts. It would be extremely interesting to analyze the errors made by *violet* to get an insight into what presents the most difficulty to text analysis systems and to have some sort of typology for such problem cases. Additionally, the fact that two systems participated in the fact extraction track and delivered meaningful results shows that the fact markup mechanisms and the comparator tool can be successfully used for future evaluations.

Tables 7 and 8 list the scores awarded for the fact extraction track. The scores in Table 8 were obtained with additional constraints imposed by the comparator tool, penalizing the systems' failure to detect facts expressed using anaphoric devices and detection of "non-facts" mentioned in future tenses or in negative or conditional contexts.

Table 7. Fact extraction scores in standard mode

	Overall			Ownership			Occupation		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
green	0.54	0.23	0.33	0.35	0.09	0.14	0.65	0.36	0.46
violet	0.75	0.38	**0.51**	0.54	0.17	**0.26**	0.80	0.56	**0.66**
				Meeting			Deal		
System				**P**	**R**	**F1**	**P**	**R**	**F1**
green				0.67	0.15	**0.24**	0.18	0.06	0.09
violet				0.87	0.14	0.23	0.68	0.19	**0.30**

Table 8. Fact extraction scores in advanced mode

	Overall			Ownership			Occupation		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
green	0.52	0.22	0.31	0.37	0.09	0.15	0.62	0.30	0.41
violet	0.70	0.36	**0.47**	0.54	0.15	**0.24**	0.74	0.49	**0.59**
				Meeting			Deal		
System				**P**	**R**	**F1**	**P**	**R**	**F1**
green				0.50	0.12	**0.20**	0.15	0.06	0.08
violet				0.58	0.10	0.17	0.63	0.22	**0.32**

8. Conclusion

FactRuEval 2016 gave us some idea of the current capabilities of information extraction systems working with Russian texts. It is clear from the results that the quality of named entity extraction for Russian is comparable to that of systems working with English texts.

We were unable to fully evaluate the quality of fact extraction for two reasons. Firstly, the number of systems that took part in the fact extraction track was too small. Secondly, the corpus offered to the participants did not contain a sufficient number of facts of different types. Of the five types of declared facts, only Occupation had enough marked up instances for meaningful evaluation[14]. Therefore, this year's fact extraction evaluation must be treated as preliminary. Next year we are planning to organize work to gather and mark up a sufficiently large number of fact mentions to have a representative corpus of facts.

An important result of FactRuEval has been the creation of an infrastructure based on the OpenCorpora.org platform that will enable future evaluations of information extraction systems working with Russian texts. We now have a technology in place enabling volunteers to annotate corpora for mentions, objects, and facts. We have also developed a handy comparator tool that compares markups produced by competing systems against the gold standard.

Equally important is the fact that we now have a publicly available gold-standard corpus annotated by volunteers. We hope that its development will continue and invite all interested parties to use this corpus and contribute to it. It would be of great value to all working in information extraction to have various specialized corpora available from OpenCorpora.org containing domain-specific entities and facts from the fields of law, medicine, etc. The existing corpus can also be expanded with averaged and moderated results obtained by the competing systems[15] on 30,000 documents fed to them in the course of the competition.

9. Acknowledgments

We would like to thank all the participants and all those who helped us in organizing FactRuEval 2016. In particular we want to thank Ekaterina Protopopova for her help with gathering documents for markup. We also want to thank her and Liliya Volkova for the careful proofreading of this article.

And of course FactRuEval 2016 would not be possible without all of the volunteers who took part in the annotation effort.

The reported study was partially supported by RFBR, research project No. 15-07-09306 "Evaluation benchmark for information retrieval".

---

[14] And even Occupation instances were often in the form of simple continuous groups of type "*job title-organization-person*".

[15] Only the results from those systems will be used whose developers agreed to such use during registration.

References

1. *Caselli T., Sprugnoli R., Speranza, M., Monachini M.* (2014), EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014 (pp. 27–34). Pisa University Press.
2. *Doddington G. R., Mitchell A., Przybocki M. A., Ramshaw L. A., Strassel S., Weischedel R. M.* (2004), The Automatic Content Extraction (ACE) Program-Tasks, Data and Evaluation, LREC, Vol. 2, p. 1.
3. *Grishman R., Sundheim B.* (1996), Message Understanding Conference-6: A Brief History, COLING, Vol. 96, pp. 466–471.
4. *Nadeau D., Sekine S.* (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26.
5. *Nekrestjanov I., Nekrestjanova M.* (2006), ROMIP'2006: Organizers' report. Proceedings on the 4th russian workshop ROMIP'2006 (pp. 7–29), Saint Petersburg.
6. *NIST:* ACE08 Evaluation Plan. http://www.nist.gov/speech/tests/ace/2008/doc/ace08- evalplan.v1.2d.pdf (2008)
7. *Surdeanu M.* (2013), Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling, Proceedings of the Sixth Text Analysis Conference (TAC 2013).
8. *Tjong Kim Sang E., De Meulder F.* (2003), Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 142–147, Association for Computational Linguistics.
9. *Toldova S., Lyashevskaya O., Bonch-Osmolovskaya A. Ionov M.* (2015), Evaluation for morphologically rich language: Russian NLP. Proceedings on the International Conference on Artificial Intelligence (ICAI) (p. 300). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).