MESHCHERYAKOV
LABORATORY of
INFORMATION
TECHNOLOGIES

Practical comparative analysis of named entity recognition methods for JINR digital services

Anna Ilina

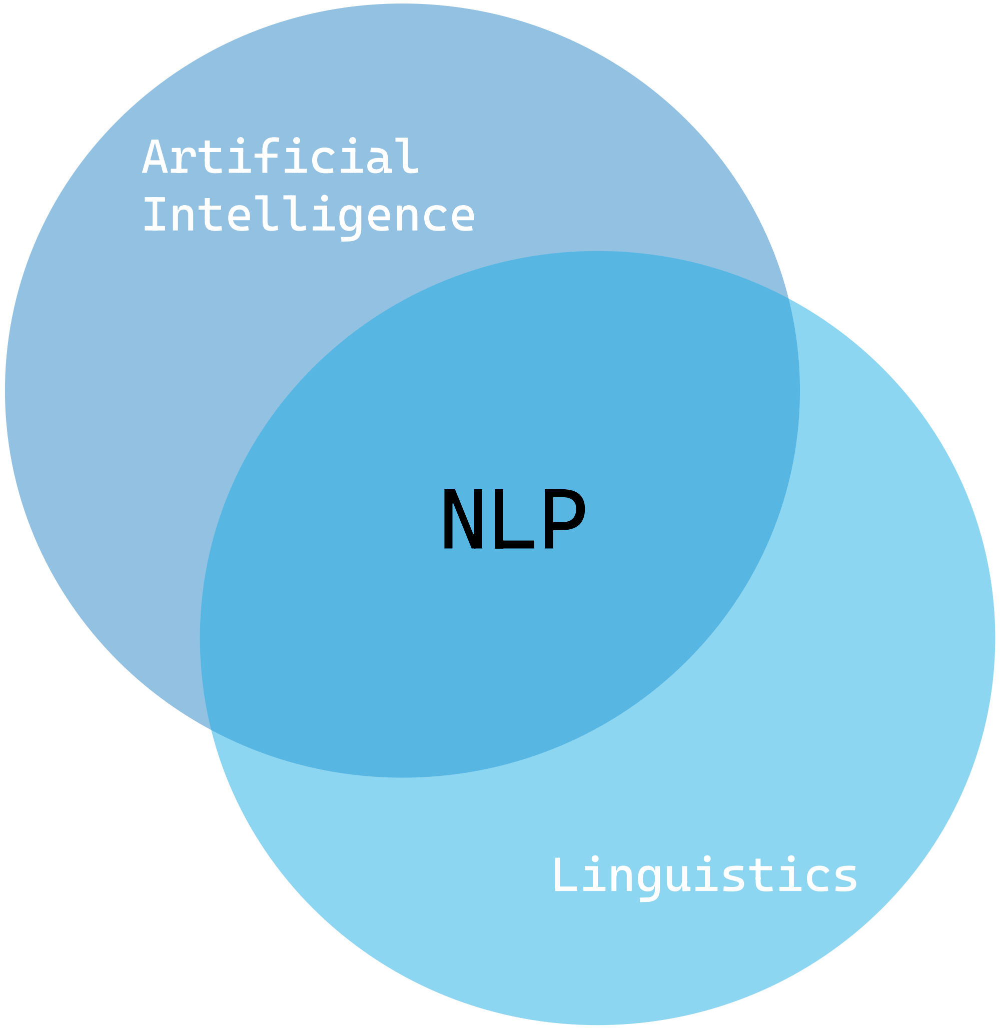
The 28th International Scientific Conference of Young Scientists and Specialists

JINR, Dubna, Russia

28.10.2024

What is NLP?

- **NLP (Natural Language Processing)** is a branch of machine learning dedicated to the recognition, generation, and processing of spoken and written human speech.
- NLP is at the intersection of the disciplines of ***artificial intelligence*** and ***linguistics***.

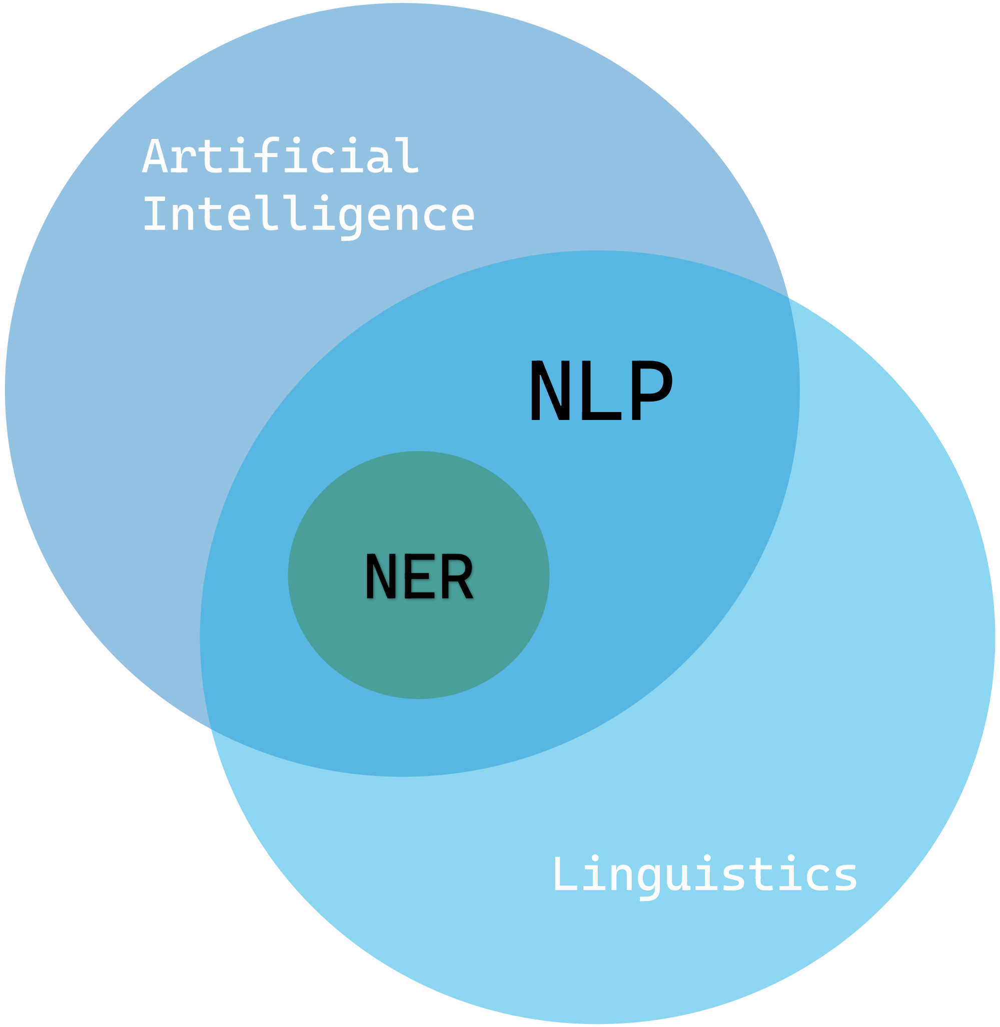


What is NER?

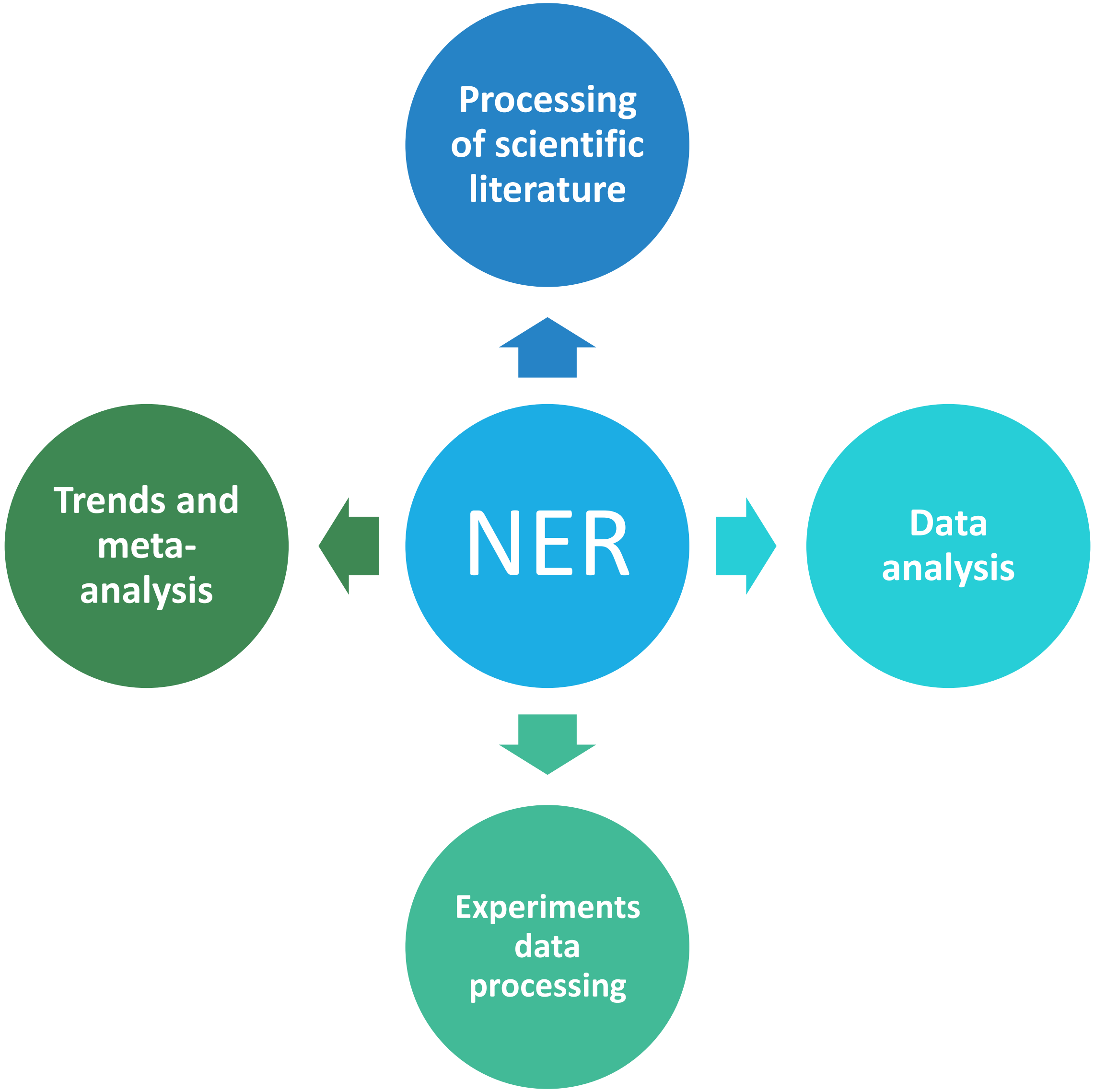
- There are specific approaches from the NLP domain for the task of **Named Entity Retrieval (NER)**.
- Named entities include *names of people, countries, cities, continents, organizations*, etc.

"In sunny Tokyo, the capital of Japan, there lived a young man named Hiroshi. He worked for Green Planet, an international organization that focused on ecology and sustainable development."

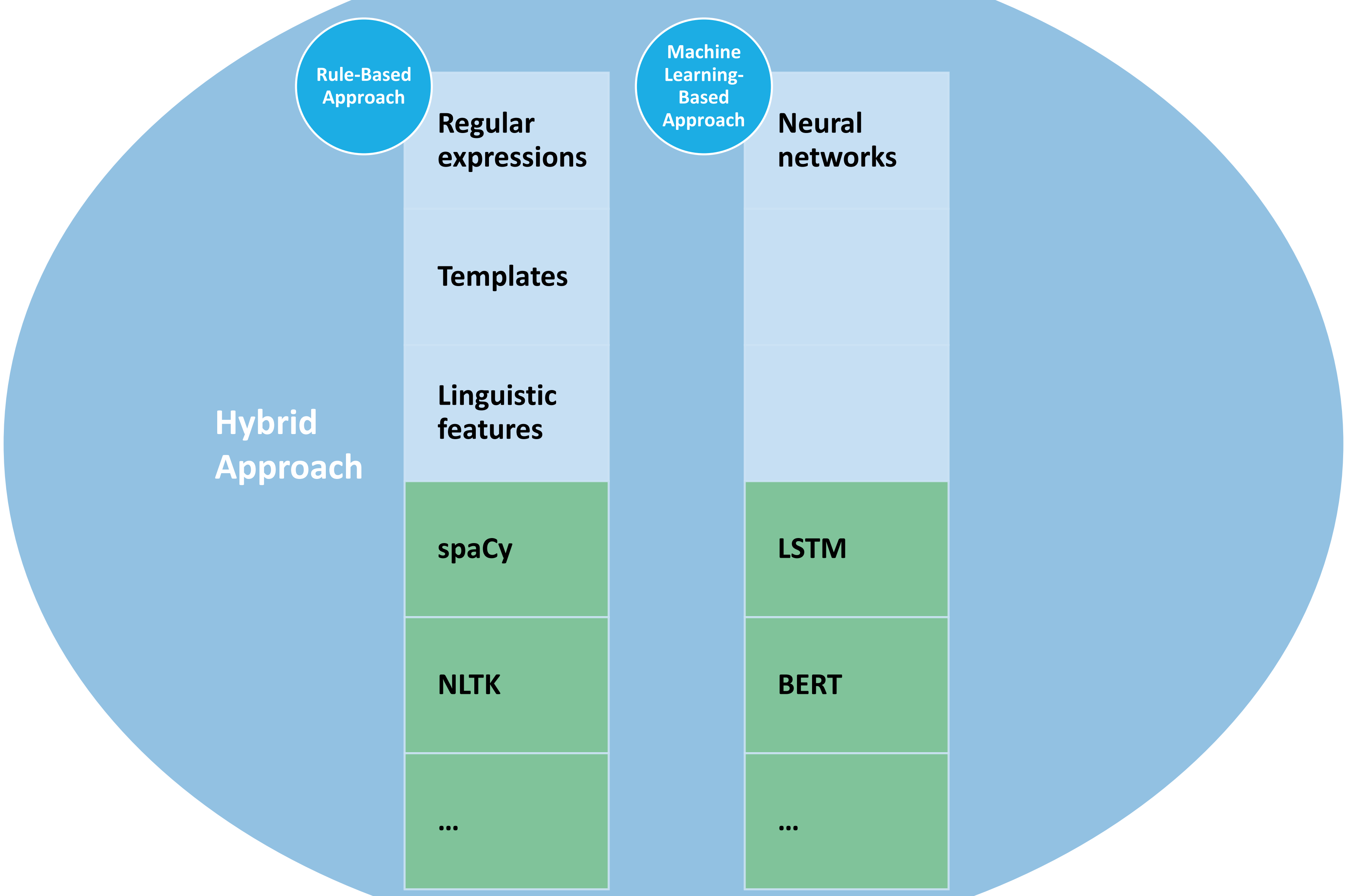
LOC location | PER person | ORG organization



In what areas of science can NER be applied?



Basic NER Approaches[1]

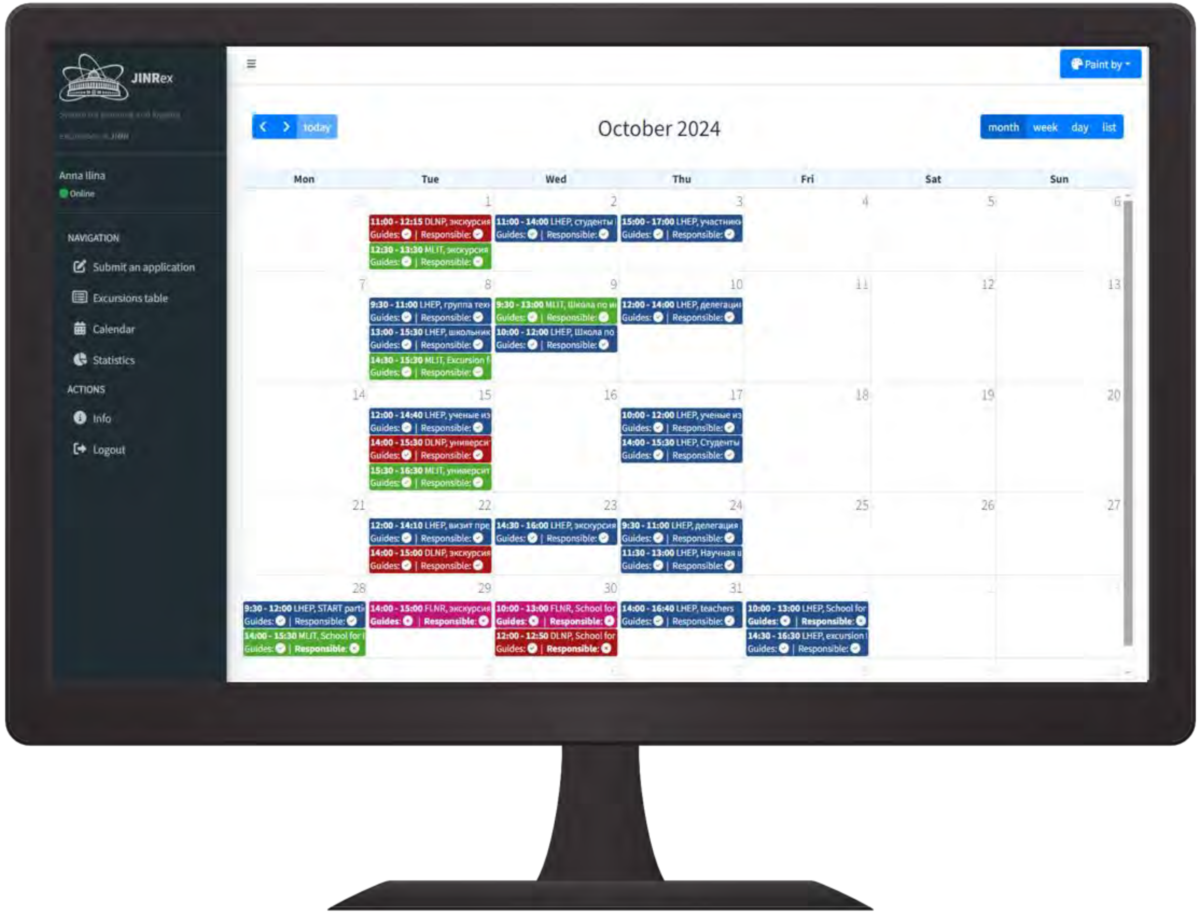


[1] **What are NER services and how they are used in business from A to Z (practice)** [Electronic resource]. URL: https://habr.com/ru/articles/763542/ (Accessed 27.10.2024).

Overview of NER capabilities on the example of analyzing the Institute's internal service data

System for planning and logging excursions at JINR

- The system for planning and logging excursions at JINR is an internal service of the Institute, which has been functioning for more than 2 years.
- The core of the system is a database that allows entering and storing information on excursions.
- Currently, **over 600 excursions** have already been accumulated in the database since 2022.



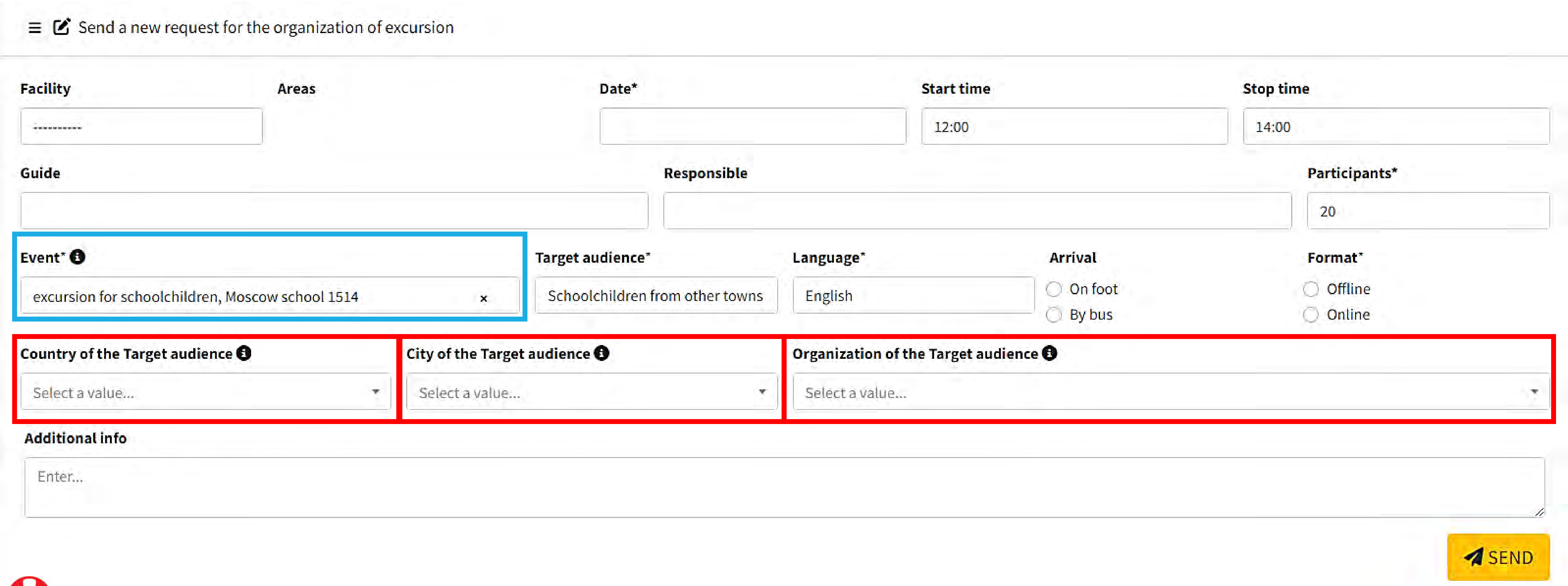
Counting statistics

The system has developed functionality for obtaining statistical data on conducted excursions such as:

- Number **by target audience**
- Number **by visited laboratories**
- Number **by visited areas**
- Number **by language**
- Number **by country of target audience**
- Number **by city of target audience**
- Number **by organizations of target audience**

since 2023

Problem: there were no special fields until 2023



Until 2023: all information was contained only in the **Event** field, which is filled out by the organizer **in a free form**

Nuances of free-form texts

Event (*free-form*)	**Language**
Экскурсия уч-ся Предуниверситария НИЯУ МИФИ/Excursion to the Pre-University of the NRU MEPhI	MIXED (*Russian & English*)
TV "Kazakhstan"	ENGLISH
экскурсия для студентов Университета «Дубна»	RUSSIAN
excursion for students Dubna University	ENGLISH
сотрудники Института астрономических исследований (Сербия)	RUSSIAN
…	

- **Ambiguity**
- **Structural differences**
- **Language mixing**
- **Errors and misprints**
- **Contextual nuances**

The aim of the research

The aim of the research is **to find a tool that allows for the automatic extraction of location and organization names as accurately as possible, in order to fill in the gaps in the data for over 600 conducted excursions**.

The obtained results will allow **to form a correct statistical picture** of all excursions conducted at the Institute using the JINRex system.

Choosing an approach to solve the task

Since *the Event title is a free-form text* and *locations and organizations of the target audience is not limited to a predefined list*, it doesn’t seem possible to create a universal algorithm based on a rule-based approach to extract information about countries, cities and organizations (such as regular expressions).

This is the reason for **choosing machine learning based tools**.

Specifics of texts in Russian

The Russian language has a certain unique specificity, different from a number of other languages. In particular, unlike English, in Russian words are inflected in cases.

Thus, the name of the same organization can be represented by different variants:

1. Экскурсия для Московского университета.
2. Московский университет .

Ignoring these features in computerized text processing can distort the resulting statistics.

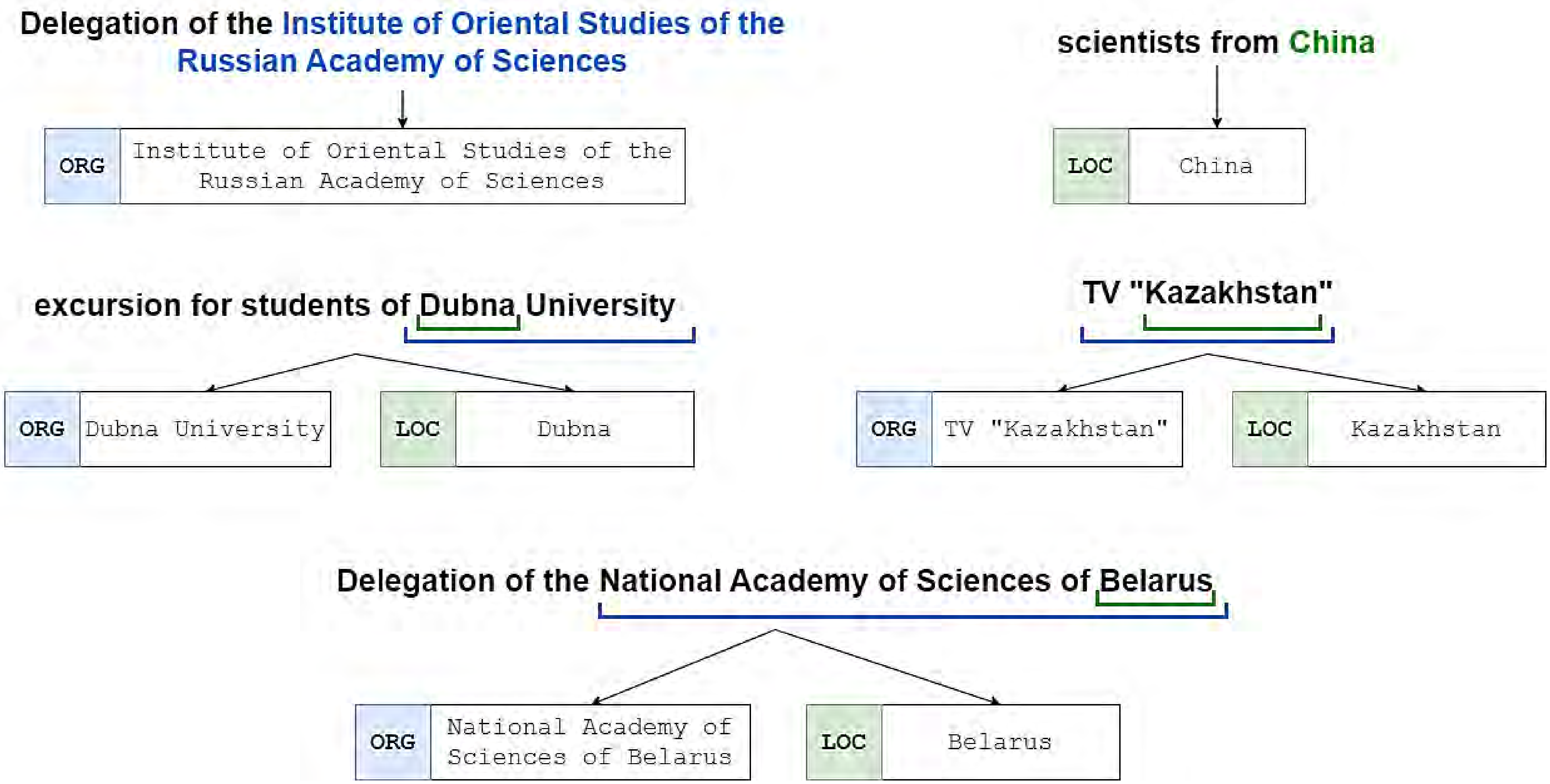
However, when translated into English, these specifics are leveled out:

1. Excursion for Moscow University.
2. Moscow University.

The general idea

1. Obtain a sample of text data of the Event titles.
2. Translate all texts into Russian and English. Thus we get three datasets:
   1. Texts “as is” (mixed-language texts).
   2. Texts translated into Russian.
   3. Texts translated into English.
3. For each variant of the obtained datasets, manually mark up named entities and their groups (*tags* or *classes*).
4. For each dataset, obtain the results of the markup with the appropriate tool.
5. Compare the obtained results with the manual markup and quantify the matches.

Manual tagging



Existing tools

Four machine learning tools were investigated for the task of named entity extraction: the **NER module of Natasha library** (*for texts in Russian*) and **three different pre-trained neural network language models** (*for mixed-language texts*).

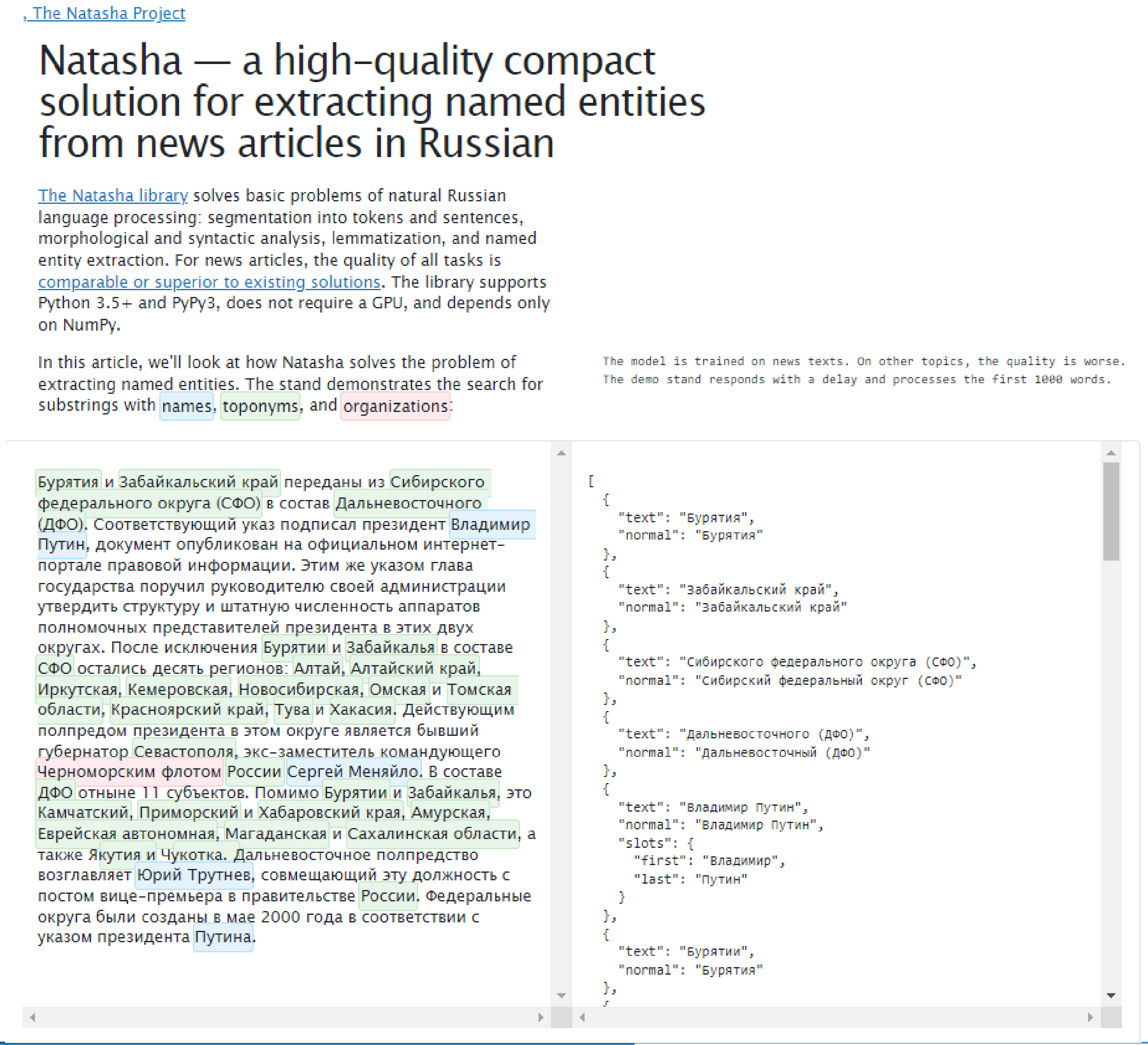
Natasha[2]

Natasha[2] solves basic NLP tasks for Russian language:

- tokenization,
- sentence segmentation,
- word embedding,
- morphology tagging,
- lemmatization,
- phrase normalization,
- syntax parsing,
- NER tagging[3],
- fact extraction.

NER module uses **Slovnet NER model**[4] internally.

Available entity groups (tags): **PER**, **LOC**, **ORG**.


, The Natasha Project

Natasha — a high-quality compact solution for extracting named entities from news articles in Russian

The Natasha library solves basic problems of natural Russian language processing: segmentation into tokens and sentences, morphological and syntactic analysis, lemmatization, and named entity extraction. For news articles, the quality of all tasks is comparable or superior to existing solutions. The library supports Python 3.5+ and PyPy3, does not require a GPU, and depends only on NumPy.

In this article, we'll look at how Natasha solves the problem of extracting named entities. The stand demonstrates the search for substrings with names, toponyms, and organizations:

The model is trained on news texts. On other topics, the quality is worse.
The demo stand responds with a delay and processes the first 1000 words.

Бурятия и Забайкальский край переданы из Сибирского федерального округа (СФО) в состав Дальневосточного (ДФО). Соответствующий указ подписал президент Владимир Путин, документ опубликован на официальном интернет-портале правовой информации. Этим же указом глава государства поручил руководителю своей администрации утвердить структуру и штатную численность аппаратов полномочных представителей президента в этих двух округах. После исключения Бурятии и Забайкалья в составе СФО остались десять регионов: Алтай, Алтайский край, Иркутская, Кемеровская, Новосибирская, Омская и Томская области, Красноярский край, Тува и Хакасия. Действующим полпредом президента в этом округе является бывший губернатор Севастополя, экс-заместитель командующего Черноморским флотом России Сергей Меняйло. В составе ДФО отныне 11 субъектов. Помимо Бурятии и Забайкалья, это Камчатский, Приморский и Хабаровский края, Амурская, Еврейская автономная, Магаданская и Сахалинская области, а также Якутия и Чукотка. Дальневосточное полпредство возглавляет Юрий Трутнев, совмещающий эту должность с постом вице-премьера в правительстве России. Федеральные округа были созданы в мае 2000 года в соответствии с указом президента Путина.


```
[
  {
    "text": "Бурятия",
    "normal": "Бурятия"
  },
  {
    "text": "Забайкальский край",
    "normal": "Забайкальский край"
  },
  {
    "text": "Сибирского федерального округа (СФО)",
    "normal": "Сибирский федеральный округ (СФО)"
  },
  {
    "text": "Дальневосточного (ДФО)",
    "normal": "Дальневосточный (ДФО)"
  },
  {
    "text": "Владимир Путин",
    "normal": "Владимир Путин",
    "slots": {
      "first": "Владимир",
      "last": "Путин"
    }
  },
  {
    "text": "Бурятии",
    "normal": "Бурятия"
  },
```


---

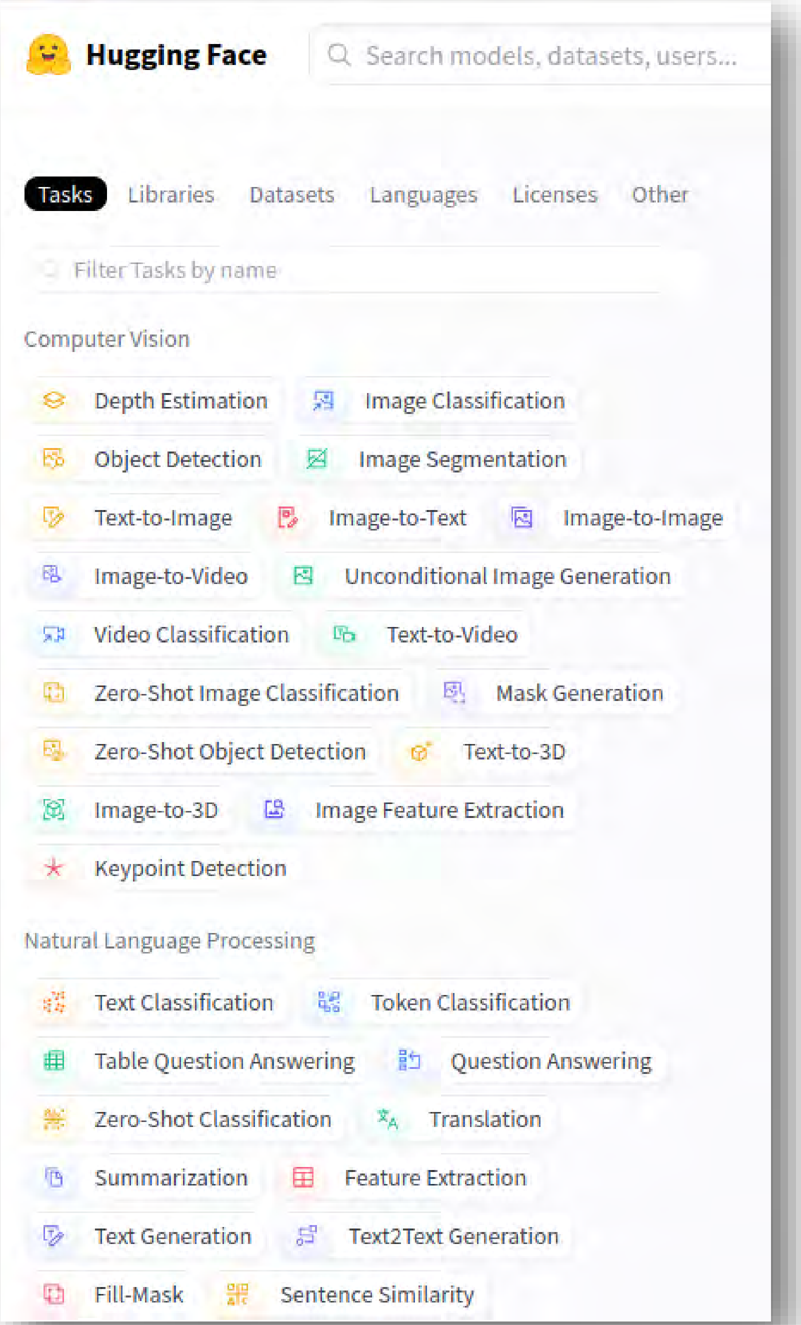
[2] **GitHub - natasha/natasha: Solves basic Russian NLP tasks, API for lower level Natasha projects** [Electronic resource]. URL: https://github.com/natasha/natasha (Accessed: 27.10.2024).

[3] **Natasha — a high-quality compact solution for extracting named entities from news articles in Russian** [Electronic resource]. URL: https://natasha.github.io/ner/ (Accessed: 27.10.2024).

[4] **GitHub - natasha/slovnet: Deep Learning based NLP modeling for Russian language** [Electronic resource]. URL: https://github.com/natasha/slovnet#ner (Accessed: 27.10.2024).

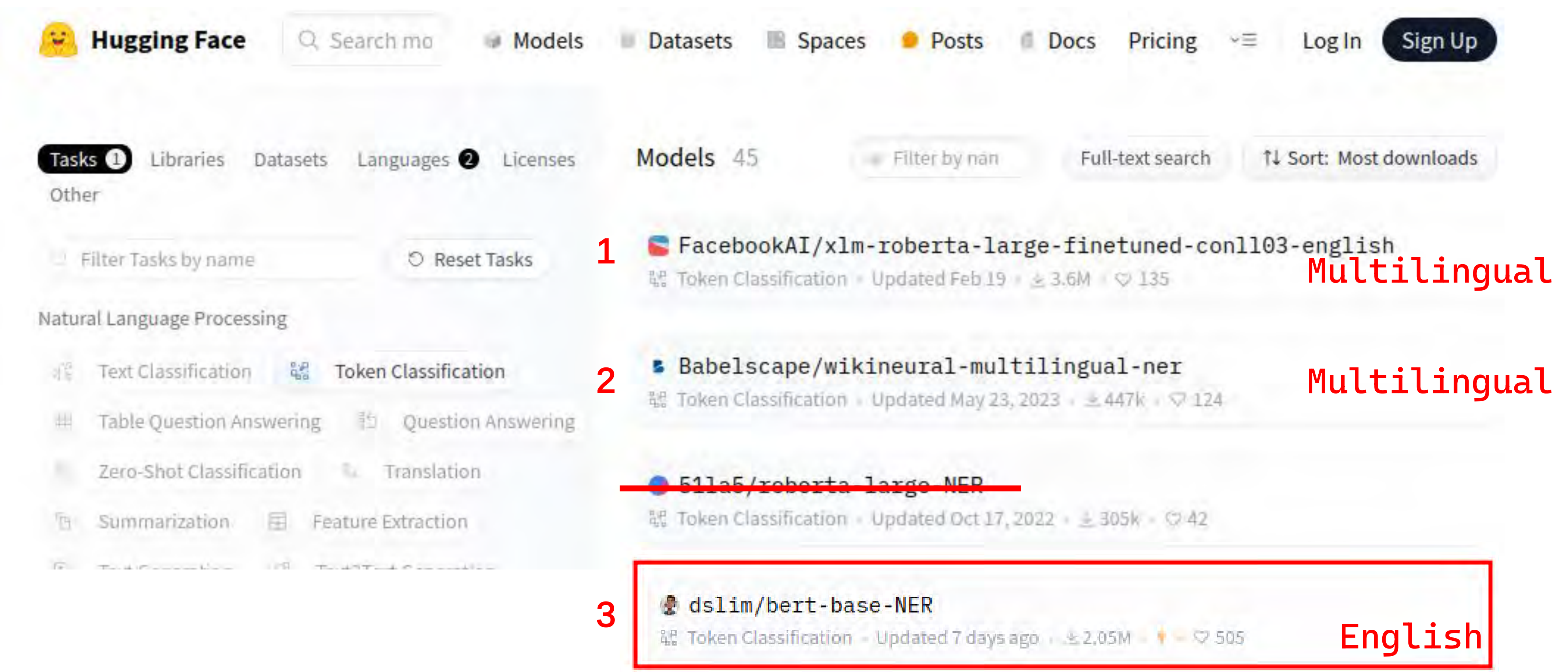
HuggingFace[5]

HuggingFace[5] — the platform where the machine learning community collaborates on models, datasets, and applications.
It has a repository of pre-trained models for a wide range of tasks, including computer vision, auditory processing, and natural language processing.



[5] **Hugging Face – The AI community building the future.** [Electronic resource]. URL: https://huggingface.co/ (Accessed: 27.10.2024).

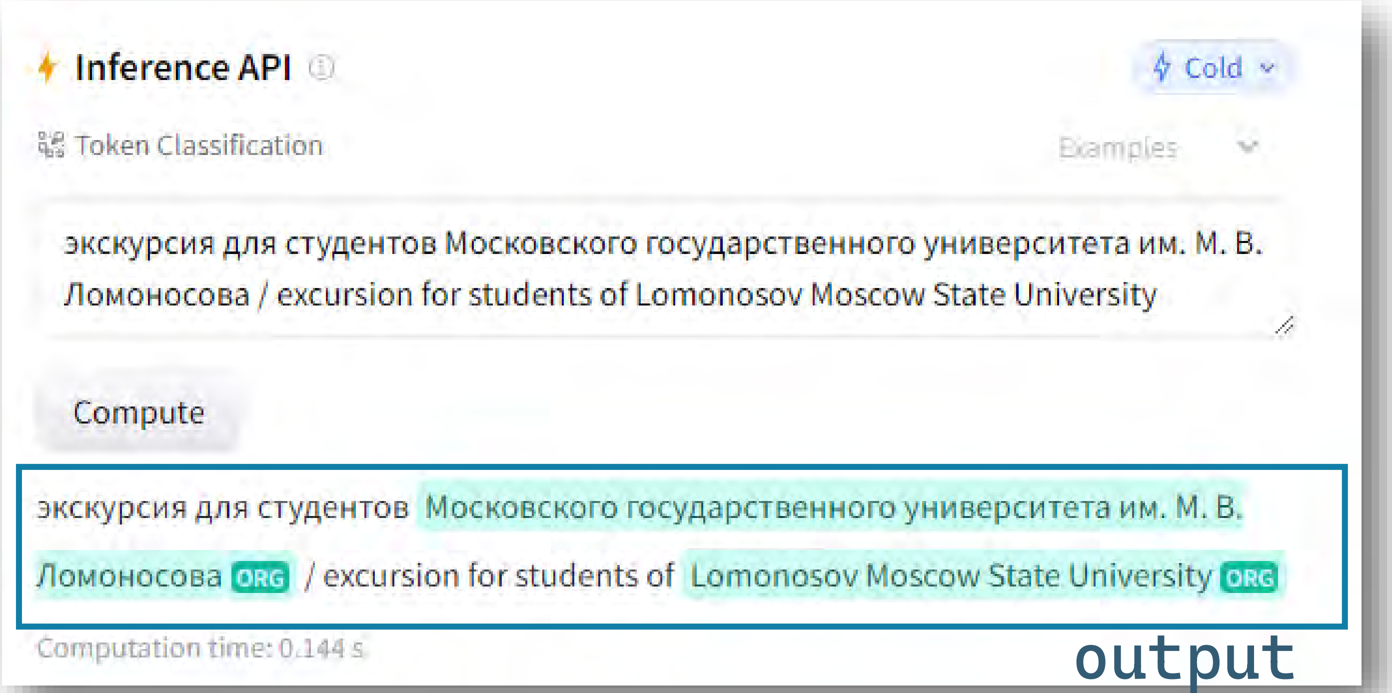
Choosing models from HuggingFace by Most downloads



FacebookAI/xlm-roberta-large-finetuned-conll03-english[6]

XLM-RoBERTa[6] is a large multilingual language model pre-trained on 2.5TB of filtered CommonCrawl* data containing 100 languages including Russian and English.

- Based on Facebook's RoBERTa model, released in 2019.
- **Model size**: 560M params.
- **Entity groups (tags)**: **PER**, **LOC**, **ORG**, **MISC**.



***CommonCrawl**[7] is a corpus of web data consisting of 250 billion pages in different languages over 17 years. It contains raw web page data, metadata and text extracts. A free and open corpus since 2007. Cited in over 10,000 scholarly articles. 3-5 billion new pages are added every month.

[6] **FacebookAI/xlm-roberta-large-finetuned-conll03-english · Hugging Face** [Electronic resource]. URL: https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-english (Accessed: 27.10.2024).
[7] **Common Crawl - Open Repository of Web Crawl Data** [Electronic resource]. URL: https://commoncrawl.org/ (Accessed: 27.10.2024).

Babelscape/Wikineural[8]

Babelscape/wikineural-multilingual-ner is a multilingual language model supporting 9 languages (Dutch, English, French, German, Italic, Polish, Portugese, Russian, Spanish).

- Pre-trained on the **Babelscape/WikiNEuRal corpus**[9] for Named Entity Recognition (NER).
- **Model size**: 177M params.
- **Entity groups (tags)**: PER, LOC, ORG, MISC.

Dataset Version	Sentences	Tokens	PER	ORG	LOC	MISC	OTHER
WikiNEuRal EN	116k	2.73M	51k	31k	67k	45k	2.40M
WikiNEuRal ES	95k	2.33M	43k	17k	68k	25k	2.04M
WikiNEuRal NL	107k	1.91M	46k	22k	61k	24k	1.64M
WikiNEuRal DE	124k	2.19M	60k	32k	59k	25k	1.87M
WikiNEuRal RU	123k	2.39M	40k	26k	89k	25k	2.13M
WikiNEuRal IT	111k	2.99M	67k	22k	97k	26k	2.62M
WikiNEuRal FR	127k	3.24M	76k	25k	101k	29k	2.83M
WikiNEuRal PL	141k	2.29M	59k	34k	118k	22k	1.91M
WikiNEuRal PT	106k	2.53M	44k	17k	112k	25k	2.20M

Datasets: Babelscape/**wikineural** like 30 Dataset card

Split (27)
test_en · 11.6k rows

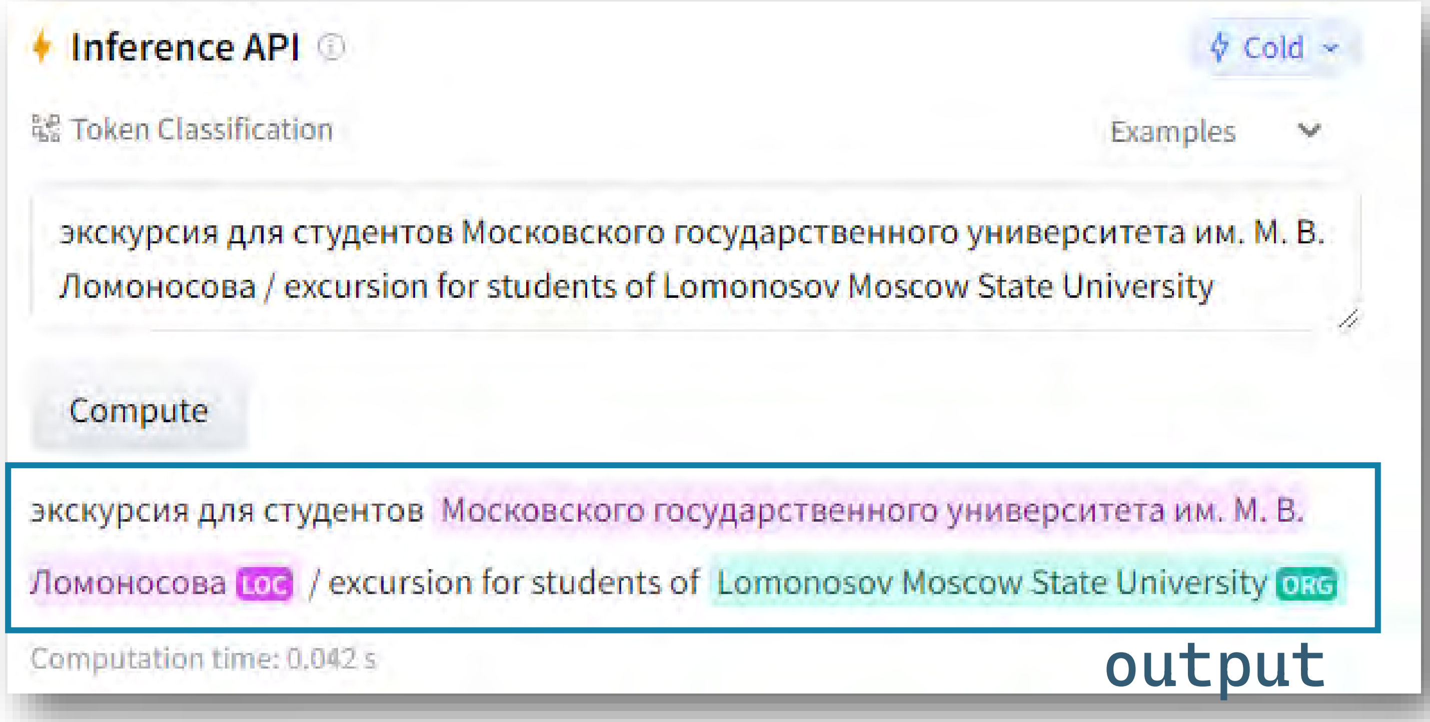
Search this dataset

tokens sequence	ner_tags sequence
["On", "this", "occasion", "he", "failed", "to", "gain", "the", "support", "of", "the", "South", "Wales", "Miners", "'", "Federation", "and", "had", "to", "stand", "down", "."]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 4, 4, 4, 0, 0, 0, 0, 0, 0]
["On", "both", "these", "occasions", "he", "was", "backed", "by", "the", "South", "Wales", "Miners", "'", "Federation", ",", "but", "he", "was", "not", "successful", "."]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0]
["He", "also", "appeared", "as", "himself", "in", "the", "1996", "film", "\"", "Eddie", "\"", "."]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0]
["The", "Colorado", "Rockies", "were", "created", "as", "an", "expansion", "franchise", "in", "1993", "and", "Coors", "Field", "opened", "in", "1995", "."]	[0, 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 6, 0, 0, 0, 0]

Babelscape/Wikineural[8]

Babelscape/wikineural-multilingual-ner is a multilingual language model supporting 9 languages (Dutch, English, French, German, Italic, Polish, Portugese, Russian, Spanish).

- Pre-trained on the **Babelscape/WikiNEuRal corpus**[9] for Named Entity Recognition (NER).
- **Model size**: 177M params.
- **Entity groups (tags)**: **PER**, **LOC**, **ORG**, **MISC**.

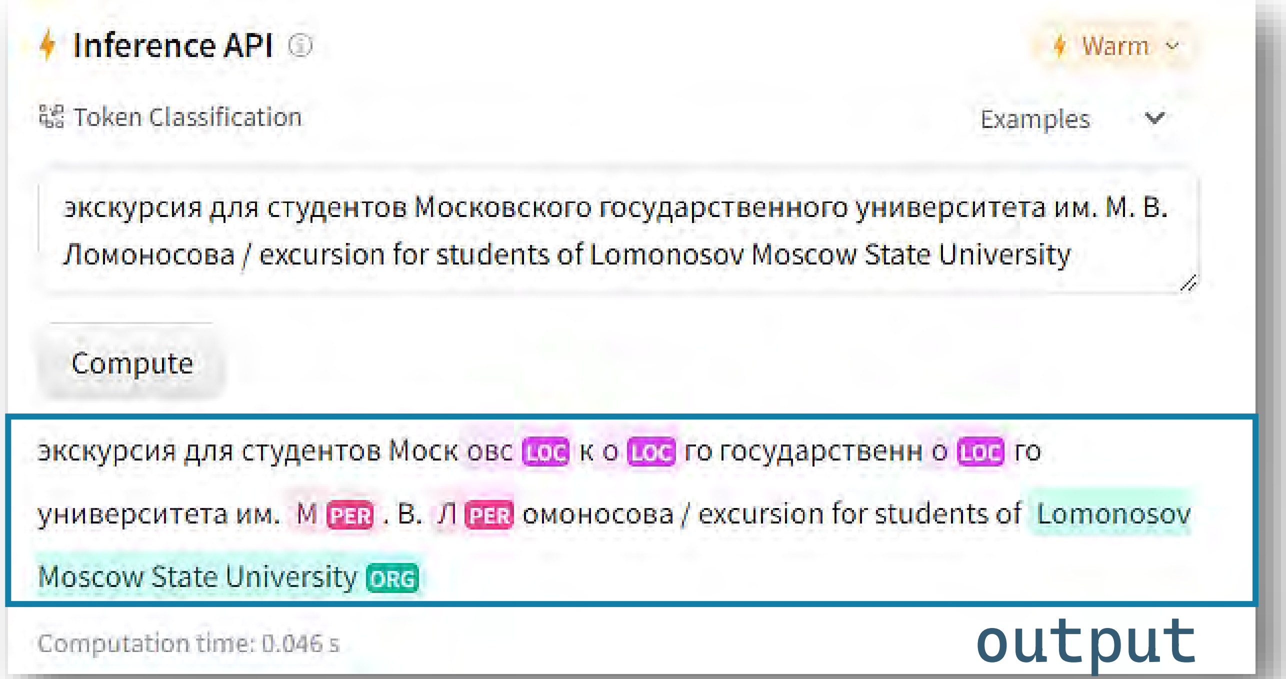


[8] Babelscape/wikineural-multilingual-ner · Hugging Face [Electronic resource]. URL: https://huggingface.co/Babelscape/wikineural-multilingual-ner (Accessed: 27.10.2024).
[9] Babelscape/wikineural · Datasets at Hugging Face [Electronic resource]. URL: https://huggingface.co/datasets/Babelscape/wikineural (Accessed: 27.10.2024).

dslim/bert-base-NER[10]

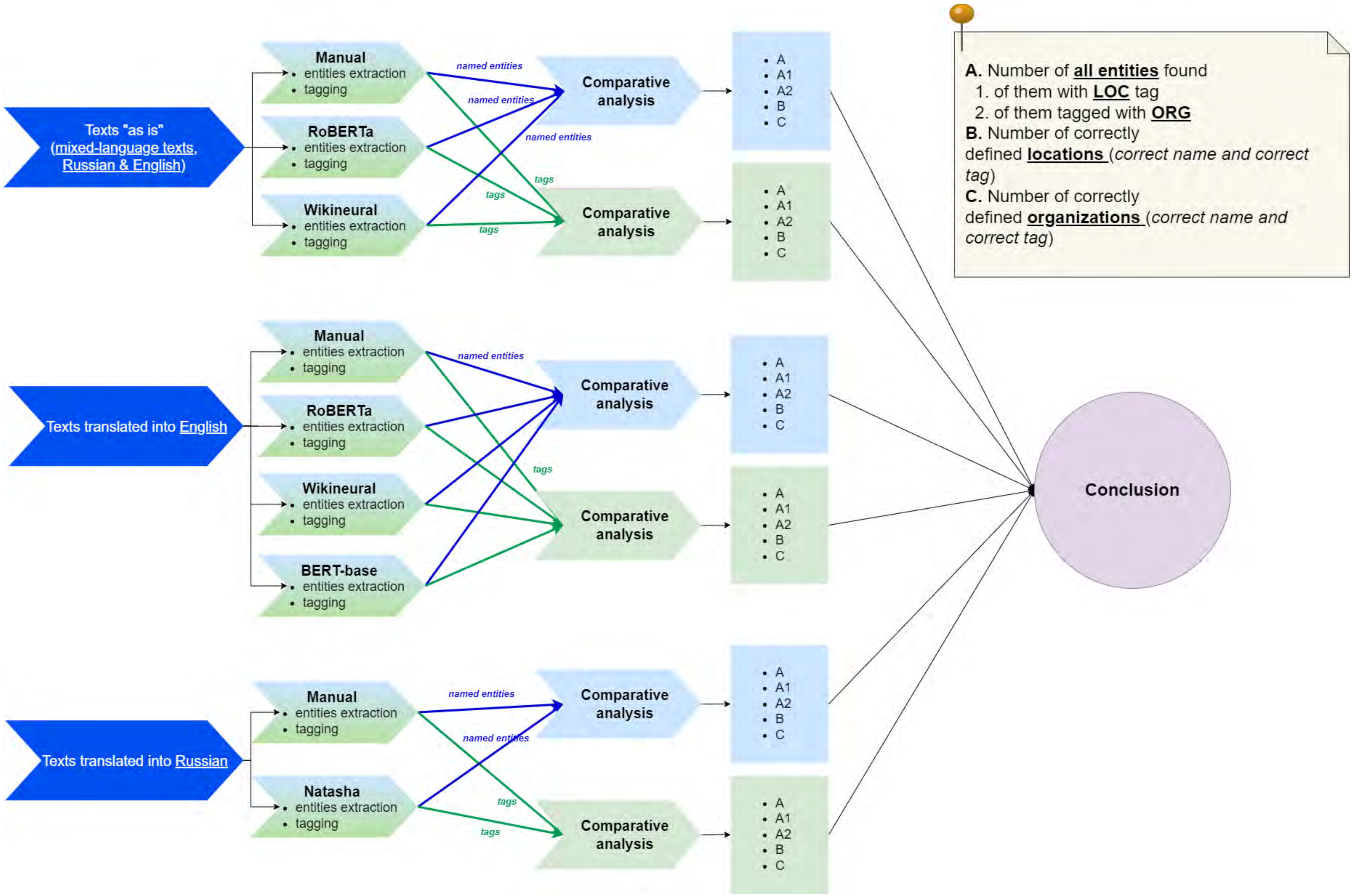
bert-base-NER[10] is a BERT-base model fine-tuned on English version of the standard CoNLL-2003 Named Entity Recognition dataset[11].

- **Model size**: 108M params.
- **Entity groups (tags)**: **PER**, **LOC**, **ORG**, **MISC**.



[10] dslim/bert-base-NER · Hugging Face [Electronic resource]. URL: https://huggingface.co/dslim/bert-base-NER (Accessed: 27.10.2024).

[11] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.


Texts "as is" (mixed-language texts, Russian & English)
Manual
entities extraction
tagging
RoBERTa
entities extraction
tagging
Wikineural
entities extraction
tagging
named entities
named entities
named entities
tags
tags
tags
Comparative analysis
Comparative analysis
A
A1
A2
B
C
Texts translated into English
BERT-base
entities extraction
tagging
Texts translated into Russian
Natasha
entities extraction
tagging
Conclusion
A. Number of all entities found
1. of them with LOC tag
2. of them tagged with ORG
B. Number of correctly defined locations (correct name and correct tag)
C. Number of correctly defined organizations (correct name and correct tag)

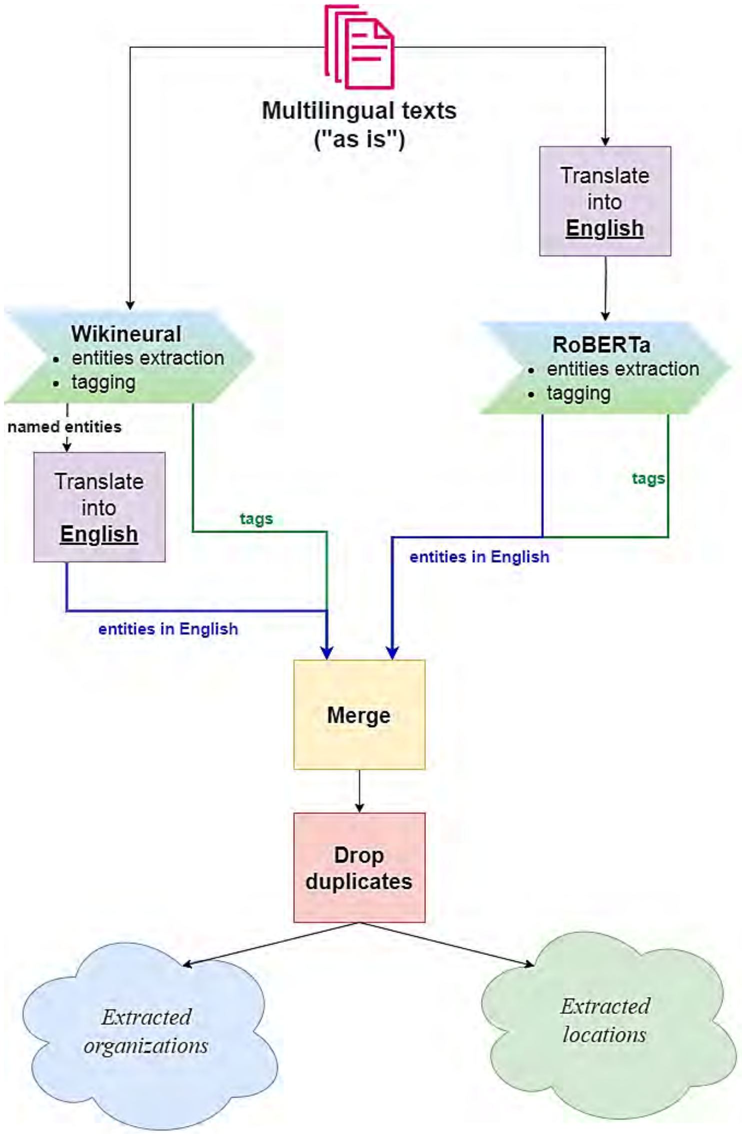
Analysis of the results

	Number of all locations found (percentage)	Number of all organizations found (percentage)	Number of locations found correctly (percentage)	Number of organizations found correctly (percentage)
Natasha (RUS)	61,82%	83,33%	52,73%	67,86%
RoBERTa (MIX)	70,91%	81,18%	65,45%	70,59%
RoBERTa (ENG)	65,52%	91,57%	62,07%	78,31%
Wikineural (MIX)	114,55%	40,00%	80,00%	30,59%
Wikineural (ENG)	63,79%	83,13%	53,45%	59,04%
BERT-base (ENG)	65,52%	91,57%	46,55%	60,24%

Findings

1. The best extraction of organizations is when using RoBERTa model on English texts.
2. The best extraction of locations - when using Wikineural model on mixed-language texts (“as is”).
3. Wikineural has a significant “skew” towards locations. It was noticed that some organizations are marked as locations.

The idea of the final algorithm for named entity extraction using selected models



Conclusion

1. This work is a first step towards understanding the capabilities of NER-tools to process and analyze wide range of data generated during various tasks of the Institute activities.
2. The RoBERTa and Wikineural models seem to work well with the extraction of named entities, which was shown by applying these models to text data processing for one of the JINR internal services - the JINRex excursion planning and logging system.
3. The obtained results will allow to form a correct statistical picture of all excursions conducted at the Institute using the JINRex system.
4. Such tools can be used on any text data for which it is necessary to solve the NER problem, including data from internal services of JINR, such as the analysis of scientific publications, network traffic, etc.

Thank you for your attention!

Additional slides

Why we can't use standard text preprocessing techniques?

	Real texts from database	Standard preprocessing technique	Result
1	online excursion for schoolchildren, lyceum 6 Dubna city	**Tokenization**	[“online”, “excursion”, “for”, “schoolchildren”, “lyceum”, “6”, “Dubna”, “city”]
2	Учащиеся Президентского физико-математического лицея № 239 г.Санкт-Петербург		[“Учащиеся”, “Президентского”, “физико-математического”, “лицея”, “№”, “239”, “г.Санкт-Петербург”]
1	Экскурсия уч-ся Предуниверситария НИЯУ МИФИ/Excursion to the Pre-University of the NRU MEPhI	**Remove punctuation and special characters**	[“Экскурсия”, “учся”, “Предуниверситария”, “НИЯУ”, “МИФИExcursion”, “to”, “the”, “PreUniversity”, “of”, “the”, “NRU”, “MEPhI”]
2	TV "Kazakhstan"		[“TV”, “Kazakhstan”]
1	представители Хэфэйского института физических наук и Института физики плазмы (КНР)	**Lemmatization** *for Russian texts: we lose cases*	представитель Хэфэйский институт физический наука и Институт физика плазма (КНР)
1	представители Хэфэйского института физических наук и Института физики плазмы (КНР)	**Remove stop-words**	представители Хэфэйского института физических наук ~~и~~ Института физики плазмы (КНР)
2	Science School for Students of the Children’s University of the Egyptian Academy of Scientific Research and Technology		Children’s University ~~of the~~ Egyptian Academy of Scientific Research ~~and~~ Technology