Modern
NER, EL & RE

Automatic knowledge graph creation

RESEARCH TOPICS **TIMELINE**

NER

Locating and classifying entities in unstructured text

EL

Assigning a unique identity to entities from knowledge base

RE

Detection and classification of semantic relationships

cIE

Combination of all previous topics in one algorithm

01

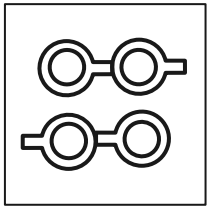
NER

Named Entity Recognition

WHAT IS **NER**?

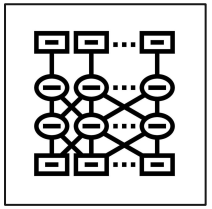
In December 1903 DATE the Royal Swedish Academy of Sciences ORG awarded Marie PERSON and Pierre Curie PERSON , along with Henri Becquerel PERSON , the Nobel Prize in Physics WORK_OF_ART .

NER SOLUTIONS



LSTM

Bi-directional LSTM models using TensorFlow and PyTorch



BERT

Fine-tuned Vanilla DistilBERT with LoRA

LSTM FOR **NER**

TensorFlow Model:

- Layers: Embedding, Dropout, two Bidirectional LSTM, Time Distributed
- Initial accuracy: 91.12%
- Issue: Accuracy dropped below 60% after updating TensorFlow

PyTorch Model:

- Simplified architecture
- Achieved validation accuracy: 97.8%

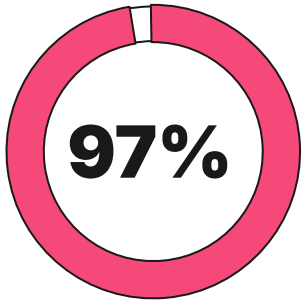
BERT FOR **NER**

We used ClearML to track metrics and logs for BERT experiments on the *“wnut_17”* dataset.

Approaches:

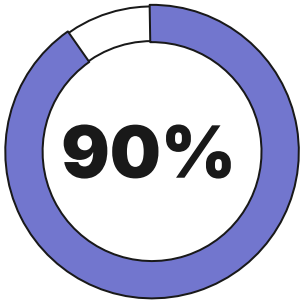
- Vanilla DistilBERT model: 90% accuracy without fine-tuning
- Fine-tuned via LoRA: Accuracy improved to 96%

NER IS EASY



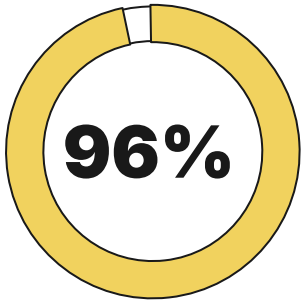
LSTM

Pytorch solution with
one bi-directional layer



DistillBERT

Vanilla DistillBERT
without fine-tuning



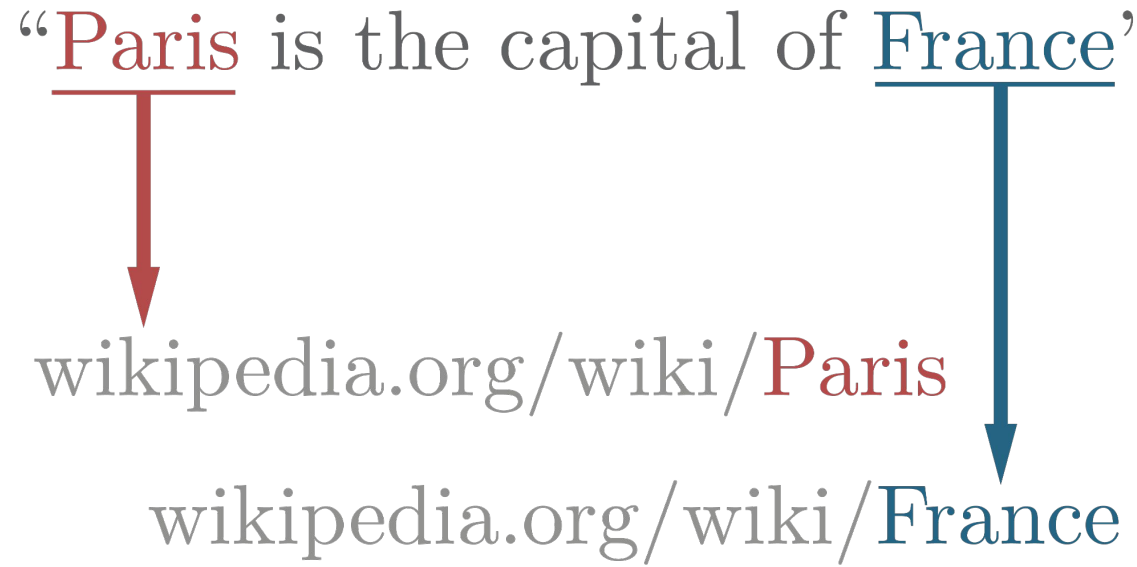
LoRA

Fine-tune of DistillBERT
on wnut_17 dataset

02

EL

Entity Linking

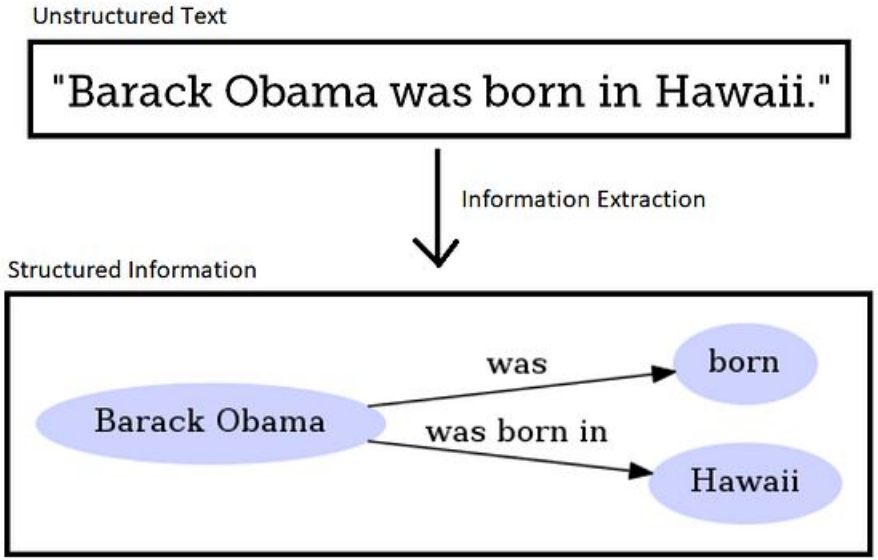

“Paris is the capital of France”
wikipedia.org/wiki/Paris
wikipedia.org/wiki/France

03

RE

Relation Extraction

WHAT IS **RE**?


Unstructured Text
"Barack Obama was born in Hawaii."
Information Extraction
Structured Information
Barack Obama
was
born
was born in
Hawaii

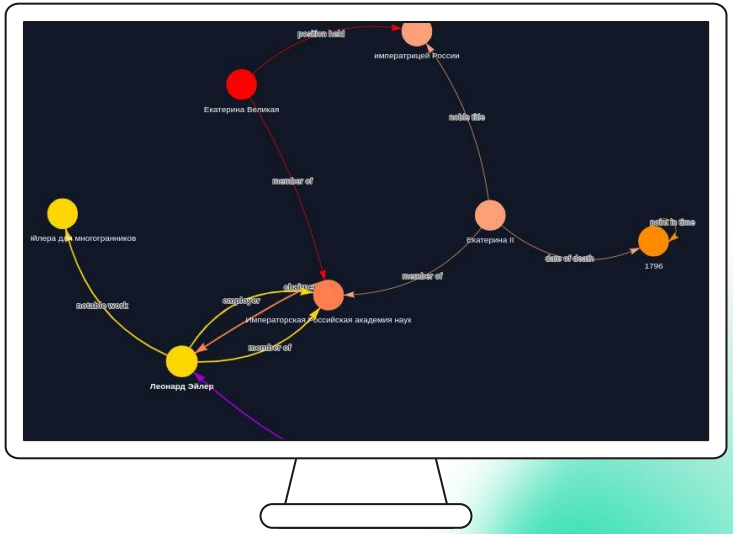
mREBEL

Multi-lingual REBEL (Relation Extraction By End-to-end Language generation) is a sequence-to-sequence model based on the BART, designed to perform end-to-end relation extraction for over 200 types of relationships

Demo

mREBEL

Gradio + Pyvis



04

clE

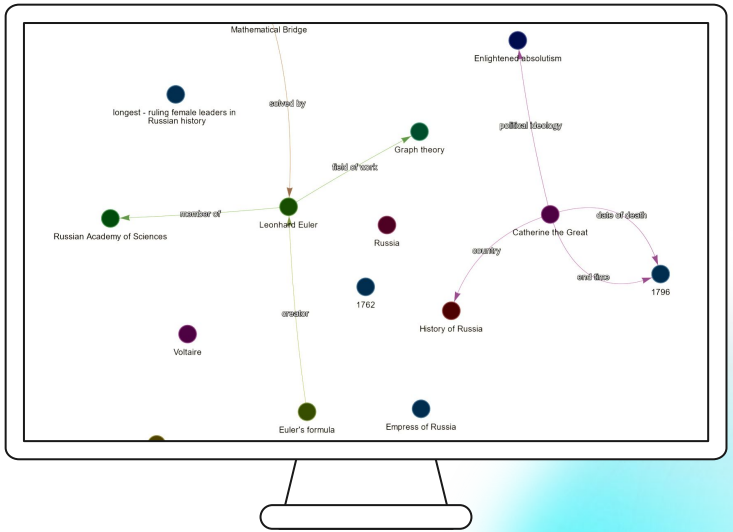
closed Information Extraction

ReLiK

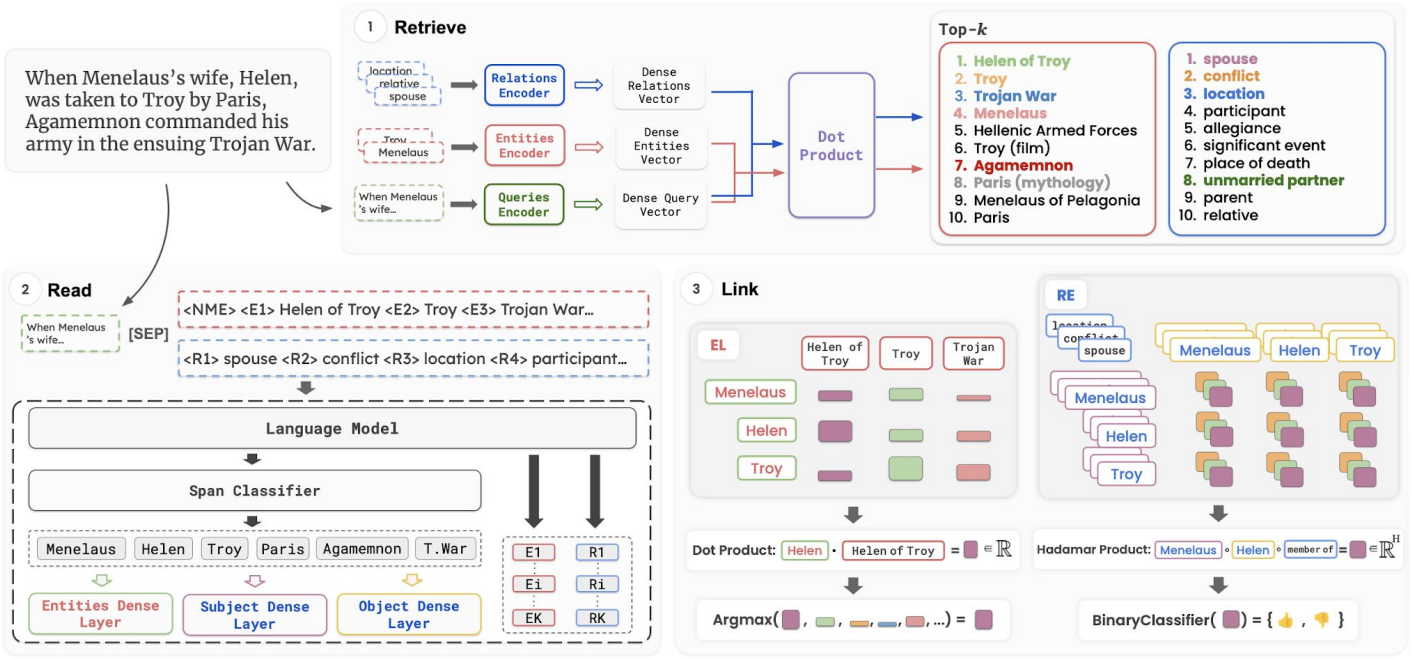
Retrieve and LinK, a blazing fast and lightweight Information Extraction model for Entity Linking and Relation Extraction

Demo
ReLiK

Gradio



HOW IT **WORKS**?



When Menelaus 's wife , Helen , was taken to Troy by Paris , Agamemnon commanded his army in the ensuing Trojan War

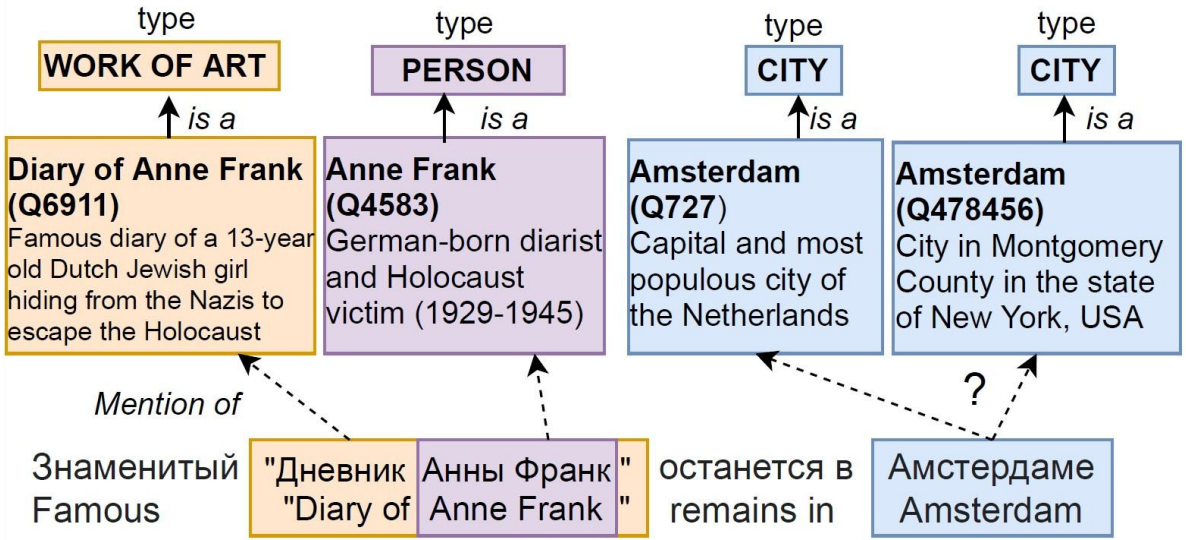
0

Russian cIE solutions

NEREL

A Russian dataset with nested named entities, relations, events and linked entities

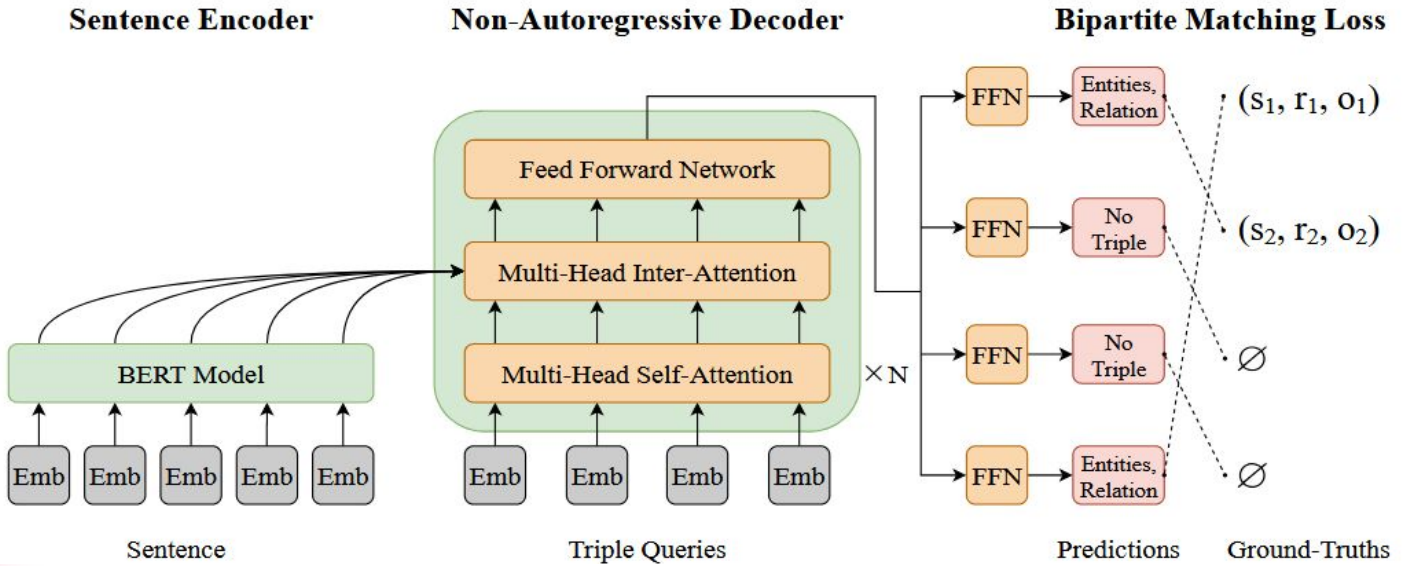
NEREL


type
WORK OF ART
is a
Diary of Anne Frank (Q6911)
Famous diary of a 13-year old Dutch Jewish girl hiding from the Nazis to escape the Holocaust
type
PERSON
is a
Anne Frank (Q4583)
German-born diarist and Holocaust victim (1929-1945)
type
CITY
is a
Amsterdam (Q727)
Capital and most populous city of the Netherlands
type
CITY
is a
Amsterdam (Q478456)
City in Montgomery County in the state of New York, USA
Mention of
?
Знаменитый "Дневник Анны Франк" останется в Амстердаме
Famous "Diary of Anne Frank" remains in Amsterdam

SPN4RE

Joint entity and relation extraction model

HOW IT **WORKS**?


Sentence Encoder
Non-Autoregressive Decoder
Bipartite Matching Loss
Feed Forward Network
Multi-Head Inter-Attention
Multi-Head Self-Attention
×N
BERT Model
Emb
FFN
Entities, Relation
No Triple
(s1, r1, o1)
(s2, r2, o2)
∅
Sentence
Triple Queries
Predictions
Ground-Truths

SPN4RE RESULTS

We fine-tuned SPN4RE on NEREL:

- Almost completely rewrote the source code *(>3500 lines of code)*
- Adapted the model architecture for a very different language
- Achieved F1 accuracy of *0.25* due to very small training dataset

ROLE **DISTRIBUTION**

ANTON KUDRYAVTSEV

SPN4RE mechanic, mathematicians' code fixer

SOFIA TKACHENKO

LSTM majesty, TensorFlow guru

ANATOLY SOLDATOV

BERTologist, InnoDataHub enjoyer

FUTURE **WORK**

STEP 1

Extend training capabilities for **SPN4RE** on **NEREL**

STEP 2

Find approaches to train **ReLiK** on **NEREL**

STEP 3

Replace BERT-based encoder in **ReLiK** on **RuBERTa**

STEP 4

Explore (fix) **PURE** and **PL-Marker** alternatives to **ReLiK**

STEP 5

Find and adapt more **Russian datasets**

STEP 6

Publish a **versatile solution** for Russian language