Evaluation of Named Entity Recognition Models for Russian News Texts in the Cultural Domain[1]

Mariia Levchenko / DHDK

The 2022 introduction of GPT-4, a cutting-edge large language model developed by OpenAI, has significantly advanced the field of natural language processing (NLP), offering remarkable capabilities for various information extraction tasks, including named entity recognition (NER). However, despite these advancements, challenges remain in ensuring the efficient, effective, and sustainable use of such models, particularly in specialized domains like cultural studies. Issues such as maintaining context and coherence, addressing inherent biases, ensuring accuracy and reliability, understanding nuanced language, and managing the substantial computational resources required underscore the need for ongoing research and optimization. This project explores strategies to harness the potential of large language models for NER in the cultural domain while mitigating these limitations and promoting sustainability.

Specifically, we compare the performance of traditional NER approaches with transformer-based models such as spaCy and RoBERTa, and evaluate the results against those of GPT models. Our findings indicate that even with OpenAI's advanced models, the results are not 100% accurate with the specific dataset, demonstrating the ongoing necessity for refinement and improvement.

The purpose of this evaluation is to select an appropriate Named Entity Recognition (NER) model for analyzing Russian news texts about literary and cultural events and academic seminars. The selection criteria are based on the model's ability to accurately identify and categorize personal names in diverse and complex textual environments.

NER state-of-the-art in humanities

As demonstrated by the survey conducted by Ehrmann (2024), the approach to named entity recognition (NER) has undergone a significant transformation over time. Initially, it evolved from rule-based systems in 2002, subsequently transitioning to the application of traditional machine learning techniques, such as conditional random fields (CRF), before finally adopting deep learning methods, including bidirectional long short-term memory (BiLSTM) CRF and BERT, between 2018 and 2020. This evolution reflects the ongoing quest to improve the accuracy and efficiency of NER across various domains.

[1] The data and code used in this report are available in the NER GitHub repository: https://github.com/mary-lev/NER

Historical newspaper datasets, such as NewsEye and HIPE-2020 (González-Gallardo 2023), are complemented by classical datasets, including ajmc, CoNLL (Todorov & Colavizza 2020), and CoNLL-2003. The following datasets, 003, BioNLP2004, WNUT2017, MIT-Movie, MIT-Restaurant, and BC5CDR, covers domains including news, medical, social media, and reviews (Peng 2024). The datasets provide a comprehensive benchmark for evaluating NER models.

For instance, Boros et al. (2020) introduced the L3i NERC-EL model for multilingual NER. This model employs a hierarchical, multitask learning approach with a fine-tuned BERT encoder and additional Transformer layers. The model demonstrated strong performance on the HIPE historical dataset, highlighting the advantages of combining BERT with transformer-based enhancements for NER tasks in English, French, and German.

In their 2020 study, Todorov and Colavizza presented a modular embedding layer combined with a BiLSTM-CRF model for NER. Their findings indicated that simplifying the architecture by removing the tanh nonlinearity after the LSTM can still yield effective results.

The investigation of OpenAI's large language models (LLMs) for named entity recognition (NER) tasks began in 2023, with a focus on the cultural domain (González-Gallardo 2023) and the clinical domain (Naguiba 2024, Hu 2024). González-Gallardo (2023) specifically evaluated ChatGPT's ability to extract named entities from historical documents using a zero-shot approach. The study revealed that, despite its considerable power, ChatGPT produced relatively poor results compared to traditional transformer-based models such as the Stacked and Temporal NERC. Factors such as inconsistent entity annotation guidelines, entity complexity, multilingualism, and the specificity of prompting contributed to these results. Additionally, the unavailability of historical newspaper datasets to web scrapers, which OpenAI relied on for training data, was a significant limitation.

These studies underscore the ongoing challenges and advancements in NER, demonstrating the necessity for continuous refinement and domain-specific adaptations to fully utilize the potential of modern NLP technologies. The use of LLMs for NER tasks requires a more detailed protocol to efficiently utilize these models. In this context, our study aims to 1) compare the performance of LLMs with traditional deep learning models on the same dataset, and 2) identify the best approaches to fully exploit their potential.

According to the evaluation principles formulated by Biderman (2024)—which include sharing code and prompts, providing model outputs, performing qualitative analysis, and measuring and reporting uncertainties to ensure reproducibility and transparency—we have shared all the data and model outputs. Additionally, we propose to formulate best practices, grounded in the current consensus within the computer science community, for using LLMs (including open-source

models) in both production and research contexts. These best practices aim to enhance the reliability and overall effectiveness of LLMs in NER and other NLP tasks.

Dataset description

In this study, we evaluate several Named Entity Recognition (NER) models using a dataset created by parsing and cleaning raw text from electronic newsletters sent in the SPbLitGuide (Saint Petersburg Literary Guide) project between 1999 and 2019. These newsletters detailed upcoming cultural events in Saint Petersburg, providing a robust source of information for analyzing the city's literary landscape over two decades.

Cleaned and structured, this dataset has been published on Zenodo: Dataset Link. It contains 15,012 records of events from 1999 to 2019, detailed with attributes such as event ID, event description, date/time, location, address, and geographic coordinates (latitude/longitude).

While parsing and cleaning the geodata was relatively straightforward – successfully extracting addresses and geocodes for each event location – extracting personal names from the event descriptions proved much more challenging due to the complexity of the Russian language, with its multiple forms and cases, combined with the unique characteristics of the news genre.

With over 15,000 events and approximately 15,000 individual names mentioned in the event descriptions, automating Named Entity Recognition (NER) became a critical requirement. To automate the extraction of person names and streamline the analysis process, I need to evaluate and identify the most appropriate NER model for this task. The goal of this evaluation is to determine which NER model provides the best results for extracting personal names accurately and efficiently, allowing for a comprehensive analysis of the cultural dynamics in St. Petersburg during this period.

The intended use of this dataset, with successfully recognized and matched names, is to conduct an in-depth analysis of the cultural dynamics and literary landscape in St. Petersburg from 1999 to 2019, allowing researchers to track key figures and their connections within this literary community, map event locations, and identify trends that have influenced the evolution of the city's cultural scene over two decades.

Sample Description and Annotation

The dataset sample of 1000 records was randomly selected based on event date and event description length from the entire dataset of 15,012 records. This sample was manually annotated using the Doccano text annotation framework provided online for this project (login: demo, password: demouser). This annotation process produced a JSONL file containing text with

associated labels, where each label has a start and end index indicating its position within the text.

We used only the "PERSON" label, recognizing that named entities can occur multiple times within a single text. In this sample dataset of 1,000 texts, we identified a total of 5,611 "PERSON" labels. This focused approach aims to investigate the accuracy and efficiency of person name extraction, which is critical for literary network analysis. By focusing on individual names and excluding names from organizational titles or works of art, this method ensures that the dataset accurately reflects the real participants, without ambiguity or misclassification caused by different types of entities.

Some of the models explored include very simple tags or labels such as PER (person), LOC (location), and ORG (organization), while others distinguish additional details such as "WORK_OF_ART" and others. In keeping with the goal of our project and to keep the models consistent, we use only the "PERSON" label. This approach avoids confusion and ensures consistency across the evaluated models, focusing solely on extracting person names to facilitate accurate analysis of cultural dynamics and literary networks.

NER dataset challenges

The main challenges in this dataset are due to the diverse and varied nature of the source material, which includes a wide range of names from different cultures, complex name structures, context variability, and language variations. These characteristics have a significant impact on the effectiveness of Named Entity Recognition (NER) models.

1. **Diversity in Names**: The dataset contains a wide range of names from different cultures (Russian, post-Soviet, English, Finnish, Italian, Japanese) and their various spellings.
2. **Complex Name Structures**: The data include official names with initials or patronyms (*Елена Николаевна Долгих, М. Ю. Герман*), and more informal or playful representations as short names or complex pseudonyms (*По, Люба Визуальновпорядке*).
3. **Context Variability**: Names appear within diverse contexts ranging from academic settings to avant-garde art events, impacting the consistency of name representation.
4. **Language and Case Variations**: Names are presented in different forms and cases, sometimes within the same sentence, complicating entity extraction and classification: *Илья Стогоff, Мари Stell & Денис Acid_C, Евгений "Ес" Соя, Мурад Гаухман-s, Мария D'Espoir, ДАНИИЛ «dakins» ВЯТКИН, Lena Smirno.* This can be explained by the contemporary cultural context in which the actors represent themselves and their artistic names, but also by the inclusion of world cultures in the St. Petersburg literary scene.
5. **Disambiguating Names in Complex Contexts.** In Named Entity Recognition (NER) tasks, particularly within datasets involving cultural, literary, and public event texts, one

of the significant challenges is disambiguating names that could refer to multiple entities depending on context. This complexity is exacerbated when names are not just personal identifiers but also part of addresses, titles, or event names.
Additionally, names of music groups or pseudonyms used in artistic contexts often mimic standard personal names, further complicating entity classification.

6. **Recognition of Fictional Characters**: News texts may reference fictional characters from literature, films, or other media, which often bear names that could be confused with those of real-life individuals or historical figures. For instance, a mention of "*Sherlock Holmes*" or "*Robinson Crusoe*" in a text might need to be identified as a fictional character rather than as a person. This requires NER systems to differentiate based on contextual clues and perhaps additional data sources that can confirm the fictional nature of the character within the given text.

My next goal is to automatically extract the names of the real attendees of each event, and this can only be achieved with Large Language Models (LLMs) using appropriate prompts. Therefore, this evaluation focuses on extracting the names of people mentioned in event descriptions. This approach avoids including names from organization titles or works of art, highlighting individual names, and is consistent with the settings of the selected models, which include predefined labels such as "WORK OF ART" and "ORG". By focusing on person names, this evaluation ensures that the resulting dataset accurately reflects real participants, without ambiguity or misclassification caused by different types of entities.

Models to evaluate

In the evaluation of Named Entity Recognition (NER) models suitable for processing Russian news texts, several leading models have been selected based on their architectural strengths, training datasets, and expected performance in handling the complexity of the Russian language and the specific challenges outlined earlier. Here's a detailed description of each model and the motivation for its inclusion in this study:

1. **DeepPavlov** (ner_collection3_bert). This model utilizes the BERT architecture, which has demonstrated substantial success in NLP tasks due to its deep bidirectional training and powerful language modeling. It is specifically trained on a large and diverse dataset (2.1 GB) focusing on Russian texts. Labels: PER, ORG, others.

2. **DeepPavlov** (ner_ontonotes_bert_mult). Also based on the BERT architecture, this model is trained on the OntoNotes dataset, which covers multiple languages and includes a wide range of annotations for syntax, semantics, and discourse information. Its multilanguage training allows it to recognize entities across various languages and linguistic structures, providing flexibility in diverse datasets. Labels: PERSON, GPE, WORK_OF_ART, ORG, DATE and others.

3. **RoBerta Large NER Russian** (RoBerta Large NER Russian). This model is a Russian-specific adaptation of the RoBERTa model, which refines BERT's methodology by training the model longer, on more data, and with more rigorously optimized hyperparameters. Labels: PER, LOC, ORG.

4. **SpaCy Russian Pipeline** (ru_spacy_ru_updated on Hugging Face). This model is a Russian-optimized NLP pipeline developed for SpaCy, designed specifically to perform efficiently on CPUs. It integrates multiple core components to handle various aspects of linguistic analysis: tok2vec, morphologizer, parser, senter, ner, attribute_ruler, lemmatizer. This model leverages SpaCy's efficient architecture to provide a comprehensive toolset for processing Russian texts, making it ideal for applications requiring robust linguistic analysis on systems with limited computational resources. Labels: BOMB, DATE, LOC, MONEY, ORG, PER, PRODUCT, VIRUS.

5. **OpenAI models** (gpt-3.5-turbo-0125, gpt-4-turbo-2024-04-09, gpt-4o-2024-05-13). Several of OpenAI's transformer-based models represent advanced iterations of the Generative Pre-trained Transformer series, using large-scale transformer architectures optimized for complex language understanding and generation tasks. They excel in applications such as summarization, translation, and named entity recognition (NER) due to their extensive training on diverse datasets. To evaluate the latest version of this model (gpt-4o-2024-05-13), we used two few-shot methods: 1) an API request with a prompt requesting a JSON output format, and 2) an API request using the LangChain package with Pydantic for parsing the output.

This choice implies that we want to investigate the effectiveness of a number of Named Entity Recognition (NER) models for processing Russian news texts. Our goal is to determine whether any of the models demonstrate state-of-the-art performance, thereby potentially reducing the need for manual verification.

We selected several models for comparison: DeepPavlov's *ner_collection3_bert* and *ner_ontonotes_bert_mult*, which use advanced BERT architectures for deep contextual learning, and *Roberta Large NER Russian*, which is optimized for the Russian language. In addition, we have included *SpaCy Russian Pipeline* for its efficiency on CPUs, providing a baseline with its conventional NLP approach. The advanced capabilities of OpenAI's *GPT-3.5, GPT-4* and *GPT-4o* are also evaluated to explore the limits of generative language models in NER tasks.

The tokenizers from the respective models were used for processing the text, resulting in different token counts for the same sample. For instance, the GPT-4-2024-05-13 model produced 642,006 tokens, while the GPT-4-turbo-2024-04-09 model generated 880,126 tokens. This variation in token count highlights the differences in how each model handles tokenization, which can impact the performance and efficiency of Named Entity Recognition tasks. By using

the tokenizers from the models, we ensure that each model processes the text in its native format, allowing for a fair comparison of their NER capabilities.

We test these models against standard performance metrics – precision, recall, and F1 score – to determine their reliability and accuracy. The results will help determine whether automated NER systems can meet or exceed the accuracy required for practical applications, or whether manual review remains critical in scenarios requiring high accuracy.

Workflow Description for Evaluating NER Models

The evaluation of Named Entity Recognition (NER) models on Russian news texts involves a systematic workflow designed to assess each model's performance against manually annotated benchmarks. Here's how the process is structured:

1. Dataset Preparation.
2. Model Deployment. Each selected model is run across the texts in the dataset.
3. Adjusting Model Outputs. Not all models output data in a format that is directly comparable to the manually annotated labels. Part of the workflow involves transforming each model's output to match the format of the labels. This step is critical to ensuring that subsequent analysis is accurate and meaningful.
4. Performance Evaluation. With the model outputs adjusted and aligned with the annotation format, we proceed to compare these results against the manual labels. This comparison involves calculating: True Positives (TP): Entities correctly identified by the model as present in the text; False Positives (FP): Entities that the model incorrectly identified, i.e., marks that are not present in the manual annotations, False Negatives (FN): Entities that are present in the manual annotations but were missed by the model.
5. Metrics Calculation. From the counts of TP, FP, and FN, we calculate the key metrics for each model:
    - Precision: The ratio of true positives to the sum of true positives and false positives (TP / (TP + FP)). This metric indicates the accuracy of the positive predictions.
    - Recall: The ratio of true positives to the sum of true positives and false negatives (TP / (TP + FN)). This measures the model's ability to identify all relevant instances.
    - F1 Score: The harmonic mean of precision and recall, providing a single metric to assess the balance between precision and recall.
6. Analysis and Reporting.

This workflow not only helps in evaluating existing NER models but also sets a benchmark for developing future models that are more aligned with the practical needs of processing Russian news texts. The main stages of the workflow implemented with Python is represented in the Jupyter notebook in Google Colab environment.

Results

Our objective was to determine which model most effectively identifies and classifies named entities with a focus on accuracy (precision), completeness (recall), and their harmonic mean (F1 score). Here are the summarized performance results for each model:

Model	Precision	Recall	F1 Score
rus_ner_model	0.96	0.65	0.78
mult_model	0.94	0.71	0.81
roberta_large	0.92	0.77	0.84
spacy	0.84	**0.81**	0.83
gpt-3.5	0.95	0.71	0.81
gpt-4	**0.99**	0.69	0.81
gpt-4o	**0.96**	0.86	**0.91**
gpt4o (json)	**0.96**	0.90	**0.93**

Recommendations

1. **Best Overall Performance**: **GPT-4o json** (with format instructions provided) demonstrated the highest F1 Score of 0.93, indicating a balanced and strong performance in both precision and recall. This model should be preferred for tasks requiring high accuracy and reliability in entity recognition.
2. **High Precision**: **GPT-4** achieved the highest precision of 0.99, making it suitable for applications where minimizing false positives is critical. However, its lower recall suggests it might miss some entities, which should be considered based on the application requirements.
3. **Balanced Performance**: **roberta_large** and **spacy** provided balanced results with F1 Scores of 0.84 and 0.83, respectively. These models are reliable for general use where both precision and recall are important, and spacy, in particular, offers efficiency on CPUs.
4. **High Recall**: **GPT-4o** (with simple API request) showed a high recall of 0.86, making it effective for scenarios where capturing all possible entities is more important than precision.
5. **Efficiency Considerations**: **SpaCy** is noted for its efficiency and CPU performance, making it a practical choice for environments with limited computational resources.

Specific Use Cases

- **Detailed Analysis and Research**: Use **GPT-4o json** due to its superior F1 Score, ensuring comprehensive and accurate entity recognition.
- **Critical Applications with Low Tolerance for Errors**: Use **GPT-4** for its high precision, reducing the risk of false positives.
- **General NER Tasks**: Consider **roberta_large** or **spacy** for balanced performance across precision and recall, especially where computational efficiency is also a factor. They can be used fot tasks where missing entity is costly.
- **High Recall Needs**: Choose **GPT-4o** to ensure most entities are captured, suitable for exploratory analyses or initial data gathering.

Since my primary goal was to identify exactly which persons are mentioned in a text with high accuracy, choosing a model with the highest precision is usually the best approach. This is because high accuracy means that the model makes very few false positives — that is, when it predicts an entity to be a person, it is very likely to be correct.

Summary

The choice between traditional and modern NER methods should be guided by the task's specific requirements, including the nature of the dataset, required accuracy, and available resources. In some cases, a hybrid approach that combines the strengths of both can offer the best solution, ensuring high accuracy while maintaining efficiency and explainability.
While additional training or prompt adjustments could improve model performance, this project evaluated the models as they are to understand their baseline capabilities. For the effective use of large language models (LLMs), it can be beneficial to modify querying methods, which may enhance the models' performance and adaptability to specific tasks.

Perspectives

Future efforts could explore automatic alignment of person mentions across the dataset, linking different writings for each individual, thus increasing the utility of the dataset for more complex analytical tasks. This could potentially reduce or eliminate the need for manual verification, thereby streamlining processes in journalistic and scholarly environments.

This evaluation underscores the importance of selecting an appropriate NER model based on specific needs and challenges, and highlights the advances and limitations of current technologies in handling the complexity of Russian linguistic data.

References

Biderman, S., et al. (2024). Lessons from the Trenches on Reproducible Evaluation of Language Models. *arXiv preprint arXiv:2405.14782*.

Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N., & Doucet, A. (2020, July 17). Robust Named Entity Recognition and Linking on Historical Multilingual Documents. *Conference and Labs of the Evaluation Forum (CLEF 2020)*. https://ceur-ws.org/Vol-2696/paper_171.pdf

Ehrmann, M., Hamdi, A., Linhares Pontes, E., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Comput. Surv.*, 56(2), Article 27. https://doi.org/10.1145/3604931

González-Gallardo, C. E., Boros, E., Girdhar, N., Hamdi, A., Moreno, J. G., & Doucet, A. (2023, June). Yes but.. Can ChatGPT identify entities in historical documents?. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 184-189). IEEE. https://arxiv.org/pdf/2303.17322

Hu, Y., Chen, Q., Du, J., Peng, X., Kuttichi Keloth, V., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., & Xu, H. (2024). Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. https://arxiv.org/pdf/2303.16416

Levchenko, M. (2023). Literary Events in Saint Petersburg (1999-2019) from SPbLitGuide Newsletters (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10086515

Naguiba, M., Tannier, X., & Névéola, A. (2024). Few shot clinical entity recognition in three languages: Masked language models outperform LLM prompting. https://arxiv.org/pdf/2402.12801

Peng, L., Wang, Z., Yao, F., Wang, Z., & Shang, J. (2024). MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks. https://arxiv.org/pdf/2404.00457

Todorov, K., & Colavizza, G. (2020). Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition. *Workshop on Computational Humanities Research*. https://ceur-ws.org/Vol-2723/long32.pdf