PERO OCR NER

Documentation for PERO OCR NER (PONER) 1.0 dataset.

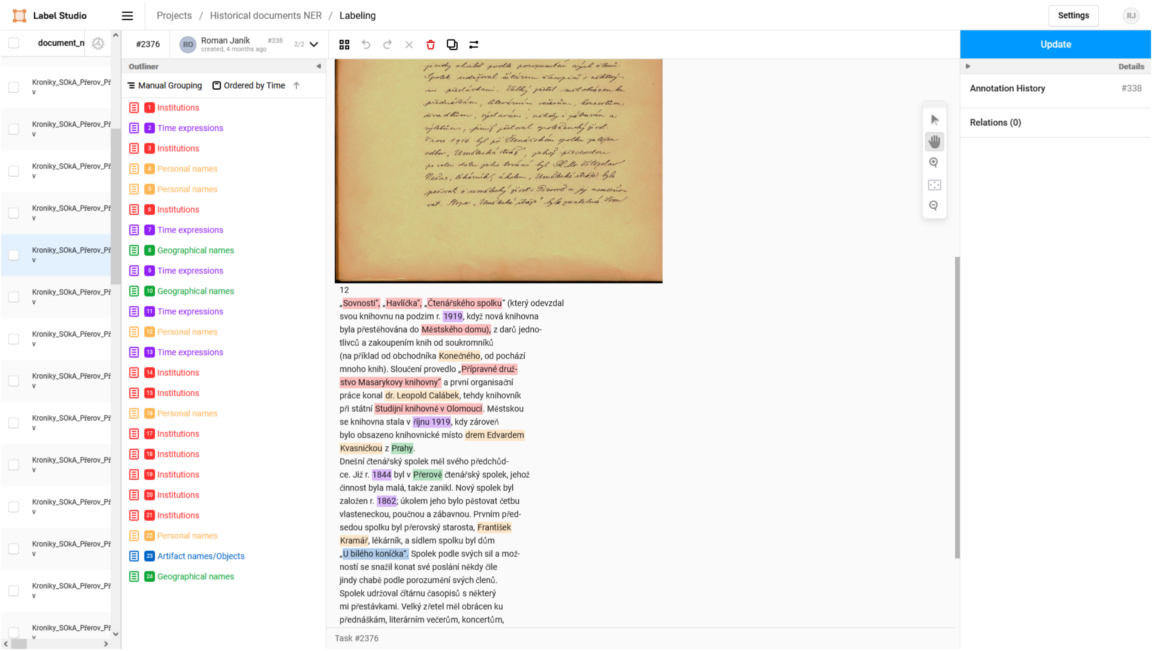
Source data

Document pages from chronicles are the source data because chronicles are rich in named entities placed in various contexts. PONER consists of 400 document pages, of which 250 are from rural and 150 from urban chronicles. Out of the 250 rural chronicles, 180 are from *central Moravia*, especially from *Šumperk* and *Přerov regions*. The rest 70 are from *České Budějovice region*. The urban chronicles are from *Přerov* city chronicle. This data distribution is important in order to include text from various environments, dialects (e.g. not only central Moravian dialect), places, and language styles. Scans of document pages were first processed by PERO OCR, and the resulting text transcriptions were annotated. The quality of OCR transcriptions is very high, errors occur very rarely. The document pages are mostly from the first half of the 20th century, less from the post-war period. The oldest document page is from the year 1771 describing the last widespread famine in the Czech kingdom, the most recent document page is from the year 1993 describing the introduction of the telephone connection. Rural chronicles often describe village inhabitants, agricultural works like harvests, prices of agricultural commodities, weather, and cultural events. Unlike rural chronicles, urban chronicles describe industrial information, city development, local politics, and many more cultural events. Nationwide or even international events (e.g. *Munich Agreement*) occur in both types. Pages from the post-war period are ideologically affected, collectivization of the Czech countryside is a major theme.

Dataset creation

Data were annotated in Label Studio [4] tool. Label Studio is a web server application, which is run locally. Data are annotated in a provided webpage. The tool is organized in a project manner, project has its settings defining a labeling interface. Labeling interface sets GUI of annotation webpage, controls, and annotation schema (here marking text span by its class). An annotation item is called a “Task”, here one task is the annotation of one document page. Tasks are defined in a JSON task file containing paths/urls to annotated files and other task information. Document page text transcription files are imported together with their respective JPEG images in order to provide a way to check the original data in case of OCR errors. JSON task file creation is implemented in script `create_label_studio_import_file.py`. A task webpage is shown in Figure 1.

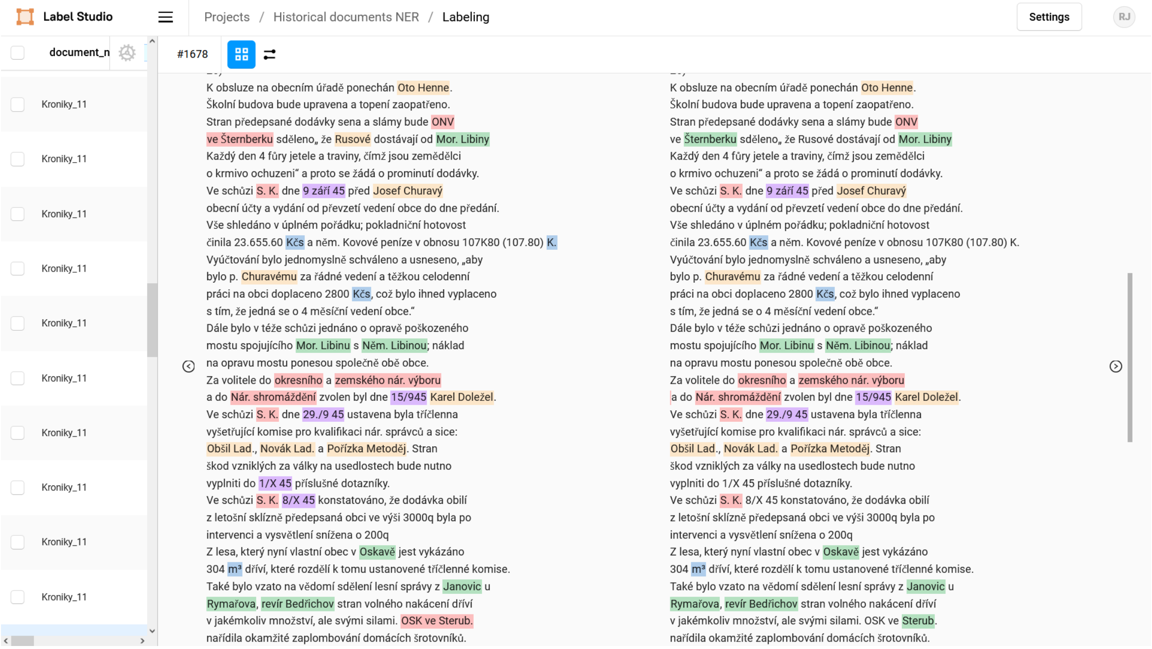
The manual annotation is a lengthy process. To accelerate the annotation, tasks were pre-annotated by a trained NER model. The best model from the first round of experiments was used. Pre-annotations can be considered as high-quality. A comparison of pre-annotations and final manual annotation is shown in Figure 2. The model often marked a



12
„Sovnosti“, „Havlíčka“, „Čtenářského spolku“ (který odevzdal
svou knihovnu na podzim r. 1919, když nová knihovna
byla přestěhována do Městského domu), z darů jedno-
tlivců a zakoupením knih od soukromníků
(na příklad od obchodníka Konečného, od pochází
mnoho knih). Sloučení provedlo „Přípravné druž-
stvo Masarykovy knihovny“ a první organisační
práce konal dr. Leopold Calábek, tehdy knihovník
při státní Studijní knihovně v Olomouci. Městskou
se knihovna stala v říjnu 1919, kdy zároveň
bylo obsazeno knihovnické místo drem Edvardem
Kvasničkou z Prahy.
Dnešní čtenářský spolek měl svého předchůd-
ce. Již r. 1844 byl v Přerově čtenářský spolek, jehož
činnost byla malá, takže zanikl. Nový spolek byl
založen r. 1862; úkolem jeho bylo pěstovat četbu
vlasteneckou, poučnou a zábavnou. Prvním před-
sedou spolku byl přerovský starosta, František
Kramář, lékárník, a sídlem spolku byl dům
„U bílého koníčka“. Spolek podle svých sil a mož-
ností se snažil konat své poslání někdy čile
jindy chabě podle porozumění svých členů.
Spolek udržoval čítárnu časopisů s některý
mi přestávkami. Velký zřetel měl obrácen ku
přednáškám, literárním večerům, koncertům,

Figure 1: An annotation task in Label Studio application.

phantom named entity in front of a real one containing only a newline character. Longer named entities were usually also marked wrongly. Some named entities are not marked by the model at all, notably groups of people as *Personal names* (e.g. nationalities "*Němci*"). Nevertheless, the annotation took several weeks of effort.



K obsluze na obecním úřadě ponechán Oto Henne.
Školní budova bude upravena a topení zaopatřeno.
Stran předepsané dodávky sena a slámy bude ONV
ve Šternberku sděleno„ že Rusové dostávají od Mor. Libiny
Každý den 4 fůry jetele a traviny, čímž jsou zemědělci
o krmivo ochuzeni“ a proto se žádá o prominutí dodávky.
Ve schůzi S. K. dne 9 září 45 před Josef Churavý
obecní účty a vydání od převzetí vedení obce do dne předání.
Vše shledáno v úplném pořádku; pokladniční hotovost
činila 23.655.60 Kčs a něm. Kovové peníze v obnosu 107K80 (107.80) K.
Vyúčtování bylo jednomyslně schváleno a usneseno, „aby
bylo p. Churavému za řádné vedení a těžkou celodenní
práci na obci doplaceno 2800 Kčs, což bylo ihned vyplaceno
s tím, že jedná se o 4 měsíční vedení obce.“
Dále bylo v téže schůzi jednáno o opravě poškozeného
mostu spojujícího Mor. Libinu s Něm. Libinou; náklad
na opravu mostu ponesou společně obě obce.
Za volitele do okresního a zemského nár. výboru
a do Nár. shromáždění zvolen byl dne 15/945 Karel Doležel.
Ve schůzi S. K. dne 29./9 45 ustavena byla tříčlenna
vyšetřující komise pro kvalifikaci nár. správců a sice:
Obšil Lad., Novák Lad. a Pořízka Metoděj. Stran
škod vzniklých za války na usedlostech bude nutno
vyplniti do 1/X 45 příslušné dotazníky.
Ve schůzi S. K. 8/X 45 konstatováno, že dodávka obilí
z letošní sklízně předepsaná obci ve výši 3000q byla po
intervenci a vysvětlení snížena o 200q
Z lesa, který nyní vlastní obec v Oskavě jest vykázáno
304 m^3 dříví, které rozdělí k tomu ustanovené tříčlenné komise.
Také bylo vzato na vědomí sdělení lesní správy z Janovic u
Rymařova, revír Bedřichov stran volného nakácení dříví
v jakémkoliv množství, ale svými silami. OSK ve Sterub.
nařídila okamžité zaplombování domácích šrotovníků.

K obsluze na obecním úřadě ponechán Oto Henne.
Školní budova bude upravena a topení zaopatřeno.
Stran předepsané dodávky sena a slámy bude ONV
ve Šternberku sděleno„ že Rusové dostávají od Mor. Libiny
Každý den 4 fůry jetele a traviny, čímž jsou zemědělci
o krmivo ochuzeni“ a proto se žádá o prominutí dodávky.
Ve schůzi S. K. dne 9 září 45 před Josef Churavý
obecní účty a vydání od převzetí vedení obce do dne předání.
Vše shledáno v úplném pořádku; pokladniční hotovost
činila 23.655.60 Kčs a něm. Kovové peníze v obnosu 107K80 (107.80) K.
Vyúčtování bylo jednomyslně schváleno a usneseno, „aby
bylo p. Churavému za řádné vedení a těžkou celodenní
práci na obci doplaceno 2800 Kčs, což bylo ihned vyplaceno
s tím, že jedná se o 4 měsíční vedení obce.“
Dále bylo v téže schůzi jednáno o opravě poškozeného
mostu spojujícího Mor. Libinu s Něm. Libinou; náklad
na opravu mostu ponesou společně obě obce.
Za volitele do okresního a zemského nár. výboru
a do Nár. shromáždění zvolen byl dne 15/945 Karel Doležel.
Ve schůzi S. K. dne 29./9 45 ustavena byla tříčlenna
vyšetřující komise pro kvalifikaci nár. správců a sice:
Obšil Lad., Novák Lad. a Pořízka Metoděj. Stran
škod vzniklých za války na usedlostech bude nutno
vyplniti do 1/X 45 příslušné dotazníky.
Ve schůzi S. K. 8/X 45 konstatováno, že dodávka obilí
z letošní sklízně předepsaná obci ve výši 3000q byla po
intervenci a vysvětlení snížena o 200q
Z lesa, který nyní vlastní obec v Oskavě jest vykázáno
304 m^3 dříví, které rozdělí k tomu ustanovené tříčlenné komise.
Také bylo vzato na vědomí sdělení lesní správy z Janovic u
Rymařova, revír Bedřichov stran volného nakácení dříví
v jakémkoliv množství, ale svými silami. OSK ve Sterub.
nařídila okamžité zaplombování domácích šrotovníků.

Figure 2: Comparison of pre-annotation by NER model (on the right) and final manual annotation (on the left).

PONER dataset uses the same named entity types/classes as CHNEC. Named entities found in page text were marked with respect to annotation manuals for CNEC [3] and CHNEC [2] datasets that can be found in the dataset's archives. However, these manuals have several conflicts listed below:

1. CNEC has animal names included in *Personal names*.
2. CNEC has a subclass "personal names of unspecified type / unclassifiable in other types" in *Personal names* with example "Slované". CHNEC has a subclass "designation of collectives" in *Intitutions* with examples "benediktini, husité, republikané, atd.".
3. CNEC has castles and palaces names in subclass "municipalities, castles and palaces" in *Geographical names*. CHNEC has a subclass "specific building objects" in *Artifact names / Objects* with examples "věž u svatých, kostel sv. Bartoloměje, zámek Kozel, klášter benediktinský u Davle, hrad Domažlický, atd.".
4. CNEC has a subclass "lectures, conferences, competitions, ... " with the example "Stanley Cup" in *Institutions*. CHNEC is missing such a subclass.
5. CNEC have subclasses "measurement units (written is shortcuts)", "names of unspecified types / unclassifiable in other types", "regulations, standards,..., their collections" and "names of chemicals, chemical formulas" in *Artifact names / Objects*. CHNEC is missing such subclasses.
6. CHNEC have subclasses "names of historical events" with examples "bitvě na Bílé hoře, Pražská defenestrace , bitva u Slavkova, atd." and "names of official recurring events" with examples "Mezinárodní filmový festival Karlovy Vary, Všesokolský slet, atd." in *Time expressions*. CNEC is missing such subclasses.
7. CHNEC has a subclass "books, magazines, editions, etc. printed matter" in *Artifact names / Objects*. CNEC has a media class.
8. CHNEC has a subclass "currency names" in *Artifact names / Objects*. Both currency abbreviations with examples "Kč, zl, zl. r. č." and full names with examples "zlatých, tolar, krejcarů" are marked. In CNEC, only currency abbreviations are marked.

These conflicts were solved the following way: Since the media class/entity type is not used in the dataset or during the training of NER models, the CHNEC method is selected. Also for conflicts 6 and 8 CHNEC method is selected. For all other conflicts, the CNEC method is selected.

Several other problems also occurred. Measurement units are sometimes put together with their values by the OCR algorithm (e.g. "25ha", "4q"). However, Label Studio allows only the marking of whole words. These entities were not marked. The "whole word only" policy also caused the marking of special characters such as dots and commas at the end of the entity, even though it is not a part of it (e.g. "Jan Novák."). In cases, where the decision if a text span is a named entity or not and what type it belongs to, CNEC and CHNEC datasets were searched. The resulting annotation by Label Studio is a list of annotations consisting of: the start character index, end character index, and entity type. Annotations can be exported from the Label Studio application in JSON file. Some of the above-mentioned shortcomings were filtered out. However, some could not be easily filtered and required writing a set of rules. Still, several thousand problematic named entities/text spans needed to be looked at manually. This lengthy process took several days. Basic filtering is implemented in script `remove_start_whitespace.py`, and semi-automatic adjustment is implemented in `adjust_annotation_end.py`.

In order to be comparable to CNEC and CHNEC datasets, Label Studio JSON export files needed to be converted to CoNLL format. This conversion is implemented in `create_my_dataset_conll.py` script. In CoNLL format, text is split into sentences separated by an empty line and sentences are split into words, each on a separate line. Natural language toolkit (NLTK)[1] library was utilized for the task of sentence and word tokenization (the task of text splitting into sentences or words). NLTK provides a statistical model *Punkt*, supporting several languages including Czech. Although the tokenization results are relatively good, the Punkt model failed to tokenize more complicated sentences correctly. Most frequent errors originated in separators (dot, comma) inside of real sentences. This problem needed to be solved by manual correction of sentences in the CoNLL file. This process took several days. Since the dataset contains corrected sentences, it can be used for training a new model for sentence tokenization based on Transformers architecture.

The final dataset consists of 1.268 kB, 9,310 sentences, and 14,639 named entities (numbers are taken from split files). Average number of named entities per document page is 36.5975. The dataset is split into three sub-sets: 45% for training, 5% for validation, and 50% for testing. The testing split is 50% because an extensive test evaluation is needed. The split ratio can be changed and the dataset split again. Split statistics are described in Table 1. The distribution of named entity types is shown in Table 2.

Split	Sentences	Entities
train	4,189	6,641
validation	465	707
test	4,655	7,291
Total	9,310	14,639

Table 1: PONER split statistics.

Named entity type	Tag	Entities
Personal names	p	4,009
Institutions	i	2,901
Geographical names	g	2,964
Time expressions	t	2,720
Artifact names/Objects	o	2,045
Total		14,639

Table 2: PONER named entity type distribution.

Bibliography

[1] Bird, S., Klein, E. and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit.* „ O'Reilly Media, Inc.“, 2009.

[2] Hubková, H., Kral, P. and Pettersson, E. Czech Historical Named Entity Corpus v 1.0. In: *Proceedings of the 12th Language Resources and Evaluation Conference.* Marseille, France: European Language Resources Association, May 2020, p. 4458–4465. ISBN 979-10-95546-34-4. Available at: https://aclanthology.org/2020.lrec-1.549.

[3] Ševčíková, M., Žabokrtský, Z., Straková, J. and Straka, M. *Czech Named Entity Corpus 2.0.* 2014. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available at: http://hdl.handle.net/11858/00-097C-0000-0023-1B22-8.

[4] Tkachenko, M., Malyuk, M., Holmanyuk, A. and Liubimov, N. *Label Studio: Data labeling software.* 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio. Available at: https://github.com/heartexlabs/label-studio.