Estonian Named Entity Recognition: New Datasets and Models

Abstract

This paper describes the annotation of two Estonian named entity recognition datasets. For this purposes, annotation guidelines for labeling eleven types of entities were created. In addition to the common entities of person names, organization names and locations, the annotation scheme includes geopolitical entities, product names, titles (or roles), events, dates, times, monetary values and percents. One annotation task involved the reannotation of an existing Estonian named entity recognition dataset, consisting mostly of news texts, with the new annotation scheme. The second annotated dataset includes new texts taken from both news and social media domains. Transformer-based models were trained on these datasets to establish the baseline predictive performance. The best results were obtained by training a single model on the joined dataset, suggesting that the domain differences between the datasets are relatively small.

1 Introduction

Named entity recognition (NER) is a useful natural language processing task that enables to extract from text information in the form of named entities that can be useful for various downstream tasks, like for instance anonymisation of documents or assigning thematic keywords to texts. Contemporary NER systems are usually trained as supervised tagging models where annotated training data is used to train the model to tag the spans in the text corresponding to named entities in the text.

For Estonian, previous efforts to develop NER systems involves the creation of an annotated corpus labeled with person, organization and location names (Tkachenko et al., 2013), and training CRF- and transformer-based models on this data (Tkachenko et al., 2013; Kittask et al., 2020; Tanvir et al., 2021) This section gives the context, introduces the problem, proposes a solution and outlines the paper contributions. From a less common domain, recently, a corpus of 19th century parish court records annotated with named entities was created (Orasmaa et al., 2022).

This paper describes the efforts to annotate additional data in Estonian with named entities with the purpose of advancing the development of general purpose NER systems for Estonian. The data annotation part of the study involves creating two annotated datasets labelled with a rich annotation scheme developed as part of the study. First, we reannotated the existing NER dataset according to the new annotation scheme and secondly, we annotated about 130K tokens of new texts, mostly from news portals and social media sources.

The second part of the paper describes the experimental results of training transformer-based predictive models on these datasets. Our main goal was first to establish the baseline performance for various entity types based on the new annotation scheme. The second goal was to study the optimal ways of using these two datasets that originate from somewhat different domains. The experimental results showed that the baseline performance on the newly annotated dataset is somewhat lower than was on the less richly annotated Estonian NER dataset, which suggests that the new annotation can be on one hand more noisy and on the other hand more rich and complex. We also found that the domains of the two datasets are similar enough, such that a model trained on the joint dataset performs as well or better than two models trained on each dataset separately.

In short, our contributions in this paper are:

1. Two new Estonian NER datasets annotated with a rich set of entities;

2. Baseline performance assessment of a transformer-based model on these two datasets.

2 Dataset Creation

In this section, we describe the process of creating the two labelled NER datasets for Estonian.

2.1 Data Sources

The first dataset that we subsequently call the Main NER dataset, is the reannotation of the existing Estonian NER dataset (Tkachenko et al., 2013). This dataset consists of ca 220K words of news texts and is generally homogeneus in its domain. Previously, this dataset has been annotated with person, organization and location names. Previous works adopting this dataset have observed errors in its annotation (Tanvir et al., 2021), which was one of the reason why this dataset was chosen for reannotation.

The second dataset that we will subsequently call New NER dataset, is totally new. To compile this dataset, the aim was to choose ca 130K tokens from both news and social media domains, with roughly 100K from news domains and 30K from the social media domain. The underlying texts were sampled from the Estonian Web Corpus 2017 (Jakubíček et al., 2013). The metadata containing the URL and the title of the web page was used to choose the texts from both categories. For selecting the news sources, we looked for the urls and titles referring to the major Estonian news sites, such as *Postimees*, *EPL*, *ERR* and *Delfi*. For extracting social media texts, we searched for keywords that would point to well-known blogging and forum platforms, such as *blogspot* and *foorum*.

2.2 Annotation Guidelines

Annotation guidelines were developed to label the data. In addition to the commonly used person, organization and location names we wanted to adopt a richer set of labels. First, we decided to separate locations into geopolitical entities and geographical locations. Following similar works in Finnish (Ruokolainen et al., 2020), we added events, products and dates. Finally, we also added titles, times, monetary values and percentages. The short description for each entity as used in the annotation guidelines was as follows:

- Persons (PER): This includes names referring to all kinds of real and fictional persons.

- Organizations (ORG): This includes all kinds of clearly and unambiguously identifiable organizations, for example companies and similar commercial institutions as well as administrative bodies.

- Locations (LOC): This includes all geographical locations that are not associated with a specific political organization such as GPEs.

- Geopolitical entities (GPE): This includes all geographic locations associated with a political organization, such as countries, cities, empires.

- Titles (TITLE): This includes job titles, positions, scientific degrees, etc. Titles must be indicated only if a specific person behind the title can be identified based on the preceding text. The personal name immediately following the title is not part of the TITLE. If the title is preceded by the ORG tag, only the job title must be marked with the TITLE, not the words in the ORG.

- Products (PROD): This includes all clearly identifiable products, objects, works, etc., by name.

- Events (EVENT): This includes events with a specific name.

- Dates (DATE): This includes time expressions, both in day/month/year type, e.g., “October 3rd”, ”in 2020”, ”2019”, ”in September”, as well as as general expressions (yesterday, last month, next year) if the expression has a clear referent. This means that based on the expression it must be possible to determine a specific point in time, i.e., a specific year or month or day. Thus, vague expressions such as a few years from now, a few months ago are not suitable, but more specific five years later, three months ago, the day before yesterday are suitable.

- Times (TIME): This includes time expressions that refer to an entity smaller than a day: times and parts of day that have a referent

(analogous to DATE entities). General expressions without a referent are not marked. Durations are also not marked.

- Monetary values (MONEY): This includes expressions that refer to specific currencies and amounts in those currencies.

- Percentages (PERCENT): This includes entities expressing percentages. A percentage can be expressed both with a percentage mark (%) or verbally.

Similar to Ruokolainen et al. (2020), we decided to include nested entities in our annotation schema. An example of a nested entity is *New York City Government* which itself is an ORG entity and which contains the GPE entity *New York*. We limited the maximum number of levels of nesting to three. Nested labelling of the same types of entities was not permitted, with the exception of ORG. For instance, in the phrase *Republic of Ireland* annotated as GPE, the further annotation of *Ireland* as a nested GPE was not allowed. However, for instance in a phrase such as *UN Department of Economic and Social Affairs* labelled as ORG, the word token *UN* was allowed to be annotated as nested ORG.

2.3 Annotation Process

The annotation process took place for both datasets separately. For the Main NER dataset, three annotators were recruited who were graduate students in general or computational linguistics. All annotators labelled the dataset independently, according to the given guidelines. Two annotators completed the annotations of the full dataset, one annotator completed most of the annotations, except for few documents. The annotation of the Main NER dataset was done with the Label Studio, which is a free open-source data annotation platform.

The New NER dataset was annotated by total of twelve annotators. Two annotators labelled the data in full. One of those was an undergraduate linguistics student and another was a graduate computer science student with undergraduate degree in linguistics. The rest of the ten annotators took part of the graduate level NLP course and each of them annotated ca 12K word tokens as part of their course work. All annotators worked independently without access to any other's persons work, following the given annotation guidelines. Thus, each text received three independent annotations. The annotation of the New NER dataset was done with the currently non-existent DataTurks annotation platform.

2.4 Label Harmonization

The annotations of the New NER dataset were harmonized using both automatic and manual approaches. First, automatic harmonization was applied according to the principle if the annotators A and B had agreed on the annotation but the annotator C had not annotated anything, the final label was set to the annotation of A and B. After that, the entire corpus was manually reviewed by two people, one of whom was the original annotator A and the other an author of this paper, and the labels were disambiguated after discussion. Mostly, the final label was the one that was chosen by at least two annotators. However, in some cases the label was completely changed or a span of words was annotated as an entity that had been left unmarked by all annotators.

The annotations of the Main NER dataset were disambiguated automatically. According to the automatic procedure, the word span was labelled as an entity if at least two annotators had marked it as an entity with the same tag.

2.5 Interannotator Agreement

In order to assess the reliability of the annotations, interannotator agreements were computed for the Main NER dataset, which are shown in Table 1. We computed the Fleiss κ, which is an extension of the Cohen's κ to more than two annotators. We followed the procedure described by (Ruokolainen et al., 2020), where each entity in the running text is treated as an instance of a positive class. The exact match annotation of this entity was checked for each annotator. If the annotator had marked this exact entity with the same label then it was recorded as an instance of the positive class, otherwise it was recorded as an instance of the negative class.

The overall agreement of the 1st level entities is in the range of substantial agreement, while the annotations on the second and third level do not agree considerably as illustrated by the low or even negative Fleiss κ values. In the first level, person names, geopolitical entities, and percentages obtain almost perfect agreement ($\kappa > 0.8$). Most other entities show substantial agreement ($\kappa > 0.6$). The lowest agreement scores were

	1st level	2nd level	3rd level
	0.65	0.23	-0.16
PER	0.95	0.27	0.66
ORG	0.76	0.33	0.19
LOC	0.65	0.35	0.18
GPE	0.84	0.47	-0.08
TITLE	0.63	0.21	0.00
PROD	0.48	0.02	–
EVENT	0.43	0.53	–
DATE	0.72	0.06	–
TIME	0.53	0.00	–
MONEY	0.78	0.00	–
PERCENT	0.90	–	–

Table 1: Interannotator agreement of the Main NER dataset as measured with the Fleiss κ.

found for products and events that still obtained moderate agreement ($\kappa > 0.4$).

2.6 Final Datasets

After the label unification process, the both final datasets were divided into train, validation and test splits. The datasets will be distributed with these prepared splits to allow for future comparison of developed models. The statistics of the final datasets are shown in Table 2.

Previously, the Main NER dataset was annotated only with PER, ORG and LOC entities (Tkachenko et al., 2013). PER and ORG labels are still among the most frequent ones while the bulk of the LOC annotations have changed to GPE according to the new annotation guidelines. The Main NER dataset also features relatively large number of titles, dates and products. The event entity is the least frequent in this dataset.

Almost similar prevalence patterns can be observed also in the New NER dataset. The PER, ORG and GPE entities are also the most frequent, followed by a relatively large number of titles, dates and products. Compared to the Main NER dataset, the New NER dataset also contains considerably more event entities. The time, percent and money entities are the least frequent in this New NER dataset.

3 Experiments

We had two main goals when conducting the experiments. The first goal was to establish the baseline performance on both datasets. Although several previous results have been published on the old annotation of the Main NER dataset (Tkachenko et al., 2013; Kittask et al., 2020; Tanvir et al., 2021), the new annotations are much richer and were collected without looking at the old annotations. Therefore the baseline performance of the new annotations might be different than was for the old annotations. As the New NER dataset contains new material, its baseline performance also had to be evaluated.

The second goal was related to the potential domain difference of the two datasets—the average document length of the New NER corpus (1281 word tokens) was more than three times higher than the average document length of the Main NER corpus (373 word tokens). Also, the New NER corpus contains at least 30K tokens from social media domain. Moreover, the documents from the news sources are not all formal news texts but contain also less formal opinion pieces. Thus, our goal was to determine how to use these datasets—whether two models should be trained, one for each dataset separately, or if joining the data and training a single model would be more beneficial.

We only used the first level annotations to train the models because as shown in Table 2, much less entities were labelled on the 2nd and on the 3rd level and as shown in Table 1, the inter-annotator agreements of the second and third level entities are lacking.

We adopted the transformer-based token classification model that for each word token, assigns a label in the commonly-used BIO format, where the B-tag denotes the start of an entity, I-tag denotes the continuation of an entity and the O tag is assigned to all word tokens that are not part of any named entity. We used the TokenClassification implementation from the Huggingface `transfomers` library (Wolf et al., 2020). The EstBERT model with 128 sequence length (Tanvir et al., 2021) was used as a base model that was fine-tuned on the NER datasets.

The batch size was fixed to 16, the Adam optimizer was used with betas 0.9 and 0.98 and epsilon 1e-6. The models were trained 150 epochs in maximum, by stopping early if the overall F1-score on the validation set did not improve in 20 epochs for more than 0.0001 F1-score points. The evaluations during the training and final testing were

	Main NER dataset				New NER dataset			
	Train	**Val**	**Test**	**Total**	**Train**	**Val**	**Test**	**Total**
Documents	525	18	39	582	78	16	15	109
Sentences	9965	2415	1907	14287	7001	882	890	8773
Tokens	155983	32890	28370	217243	111858	13130	14686	139674
1st lvl entities	14944	2808	2522	20274	8078	541	1002	9594
2nd lvl entities	987	223	122	1332	571	44	59	674
3rd lvl entities	40	14	4	58	27	0	1	28
PER	3563	642	722	4927	2601	109	299	3009
ORG	3215	504	541	4260	1177	85	150	1412
LOC	328	118	61	507	449	31	35	515
GPE	3377	714	479	4570	1253	129	231	1613
TITLE	1302	171	209	1682	702	19	59	772
PROD	874	161	66	1101	624	60	117	801
EVENT	56	13	17	86	230	15	26	271
DATE	1346	308	186	1840	746	64	77	887
TIME	456	39	30	525	103	6	6	115
PERCENT	137	62	58	257	75	11	1	87
MONEY	291	76	153	520	118	12	1	131

Table 2: Statistics of the two new Estonian NER datasets.

done with the `seqeval` package.[1] The learning rate was optimized on the validation set using the grid of {5e-6, 1e-5, 3e-5, 5e-5, 1e-4}. Each model was trained ten times with different random seeds and their mean values with standard deviations are reported.

4 Results

We first trained and evaluated models on both datasets separately to assess the overall modeling performance on each dataset. Then, we also trained another model on the joint dataset and compared its performance on the evaluation sets of both datasets.

4.1 Separate Models

First, we trained predictive models on both datasets separately. The results of these experiments on the respective validation sets are shown in Table 3. The overall performance (bottom row) is on the same level on both datasets, showing that the annotation and modeling difficulty is comparable in the two datasets.

The most accurately predicted entities in both datasets are PER, GPE and PERCENT. The lowest accuracy is obtained when predicting LOC, EVENT and TIME for the Reannotated Main NER datset and LOC, EVENT and PROD for the New NER dataset. The prediction of the EVENT names is especially poor in the Main NER dataset, probably because there are only 56 of the EVENT instances in the respective train set.

Comparing the results of the Reannotated Main dataset with the old annotations of the Main NER dataset (see Table 4, taken from Tanvir et al. (2021), Table 11) shows that the performance on all three entities (PER, ORG, LOC) used in the old annotations has decreased. Although the modeling results are not directly comparable because Table 3 shows results on the validation set and Table 4 on the test set, the differences are large enough to suggest that the new annotation is considerably more complex for the models to learn.

4.2 Joint Model

The Joint model is trained on the concatenated train sets of both the Main NER and New NER datasets. Table 5 shows the F1-scores of the Joint model on the joined validation set as well as on the validation sets of both datasets separately. The overall F1-scores on each validation set are somewhat higher than for the separate models (0.766 vs 0.747 for the Main dataset and 0.752 vs 0.735 for

[1] `https://github.com/chakki-works/seqeval`

	Reannoated Main NER				New NER			
	#	Precision	Recall	F1-score	#	Precision	Recall	F1-score
PER	642	.827 (.012)	.871 (.009)	.848 (.005)	109	.809 (.044)	.816 (.023)	.811 (.019)
ORG	504	.654 (.016)	.666 (.014)	.660 (.013)	85	.580 (.027)	.585 (.052)	.581 (.024)
LOC	118	.643 (.036)	.478 (.028)	.547 (.016)	31	.600 (.065)	.560 (.060)	.576 (.044)
GPE	714	.821 (.012)	.831 (.021)	.826 (.008)	129	.900 (.017)	.879 (.030)	.889 (.014)
TITLE	171	.676 (.023)	.814 (.014)	.739 (.011)	19	.750 (.062)	.718 (.064)	.731 (.048)
PROD	161	.572 (.033)	.628 (.026)	.598 (.024)	60	.509 (.043)	.474 (.052)	.488 (.029)
EVENT	13	.069 (.029)	.077 (.034)	.072 (.031)	16	.518 (.104)	.558 (.104)	.525 (.070)
DATE	308	.682 (.020)	.720 (.017)	.700 (.007)	64	.816 (.027)	.824 (.024)	.820 (.021)
TIME	39	.553 (.066)	.555 (.045)	.553 (.053)	6	.812 (.041)	.788 (.108)	.797 (.074)
PERCENT	62	.985 (.016)	.867 (.032)	.922 (.019)	11	.895 (.126)	1 (–)	.940 (.074)
MONEY	76	.636 (.040)	.568 (.030)	.600 (.030)	12	.659 (.085)	.742 (.126)	.693 (.083)
Overall	2571	.737 (.010)	.757 (.009)	.747 (.004)	497	.736 (.014)	.734 (.017)	.735 (.006)

Table 3: Predictive performance of models trained on both two datasets, evaluated on the respective validation set.

	Precision	Recall	F1-score
PER	.948	.958	.953
ORG	.784	.826	.805
LOC	.899	.914	.907
Overall	.891	.912	.901

Table 4: Results of the old annotations of the Main NER test set. Adapted from Table 11 (Tanvir et al., 2021).

the New dataset, compare with the bottom row of Table 3).

Figure 1 shows the entity-level comparison of the Joint models and the separate models on the respective validation sets. Figure 1a, which depicts the comparison on the validation set of the Main NER dataset, shows that the Joint model performs the same or better on all entities, except the TIME entity, which performance already was among the lowest one Main dataset and which with the Joint model drops even more from 0.553 to 0.433. On the other hand, the prediction accuracy of the EVENT entity, while still remaining quite low, improves considerably, from 0.072 to 0.310.

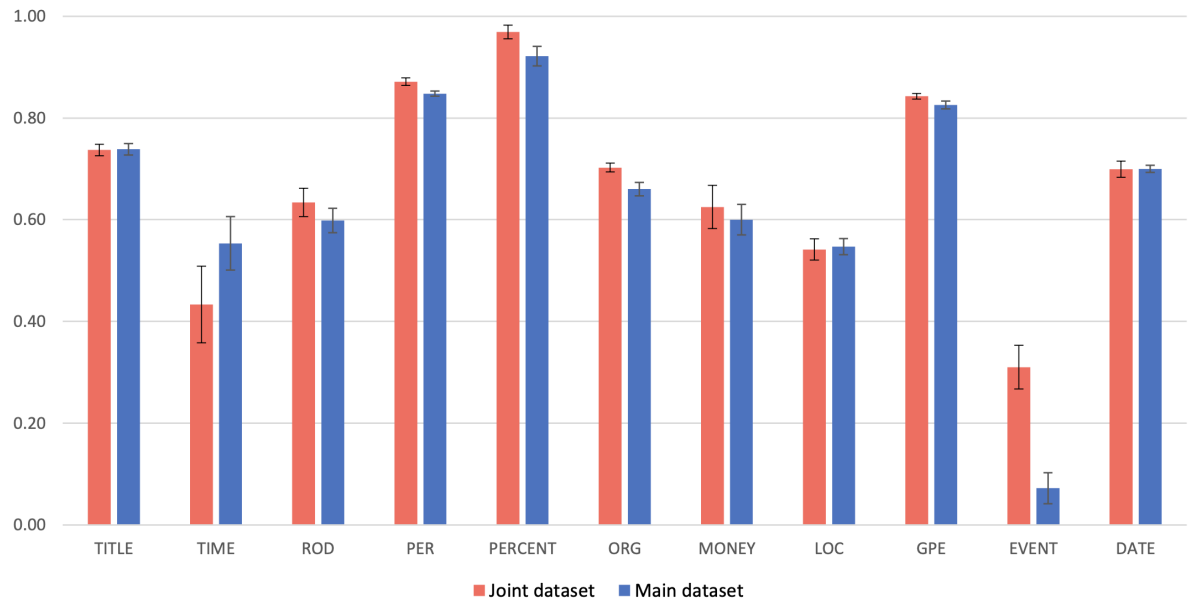
When comparing the Joint and separate model results on the New NER dataset (see Figure 1b, we observe that the Joint model performs the same or better on some entity types (PER, ORG, GPE, LOC, PROD, PERCENT, MONEY) and somewhat worse on the rest. The largest drop occurs again on the TIME entity, which drops from 0.797 to 0.627.

Overall, we conclude that training a Joint model instead of two separate models is justified. Although with the Joint model, the prediction performance dropped for some entities, especially on the New NER dataset, the overall F1-score was better on the validation sets of both datasets than for the separate models. Therefore we conduct the final evaluations on the test set with the Joint model.

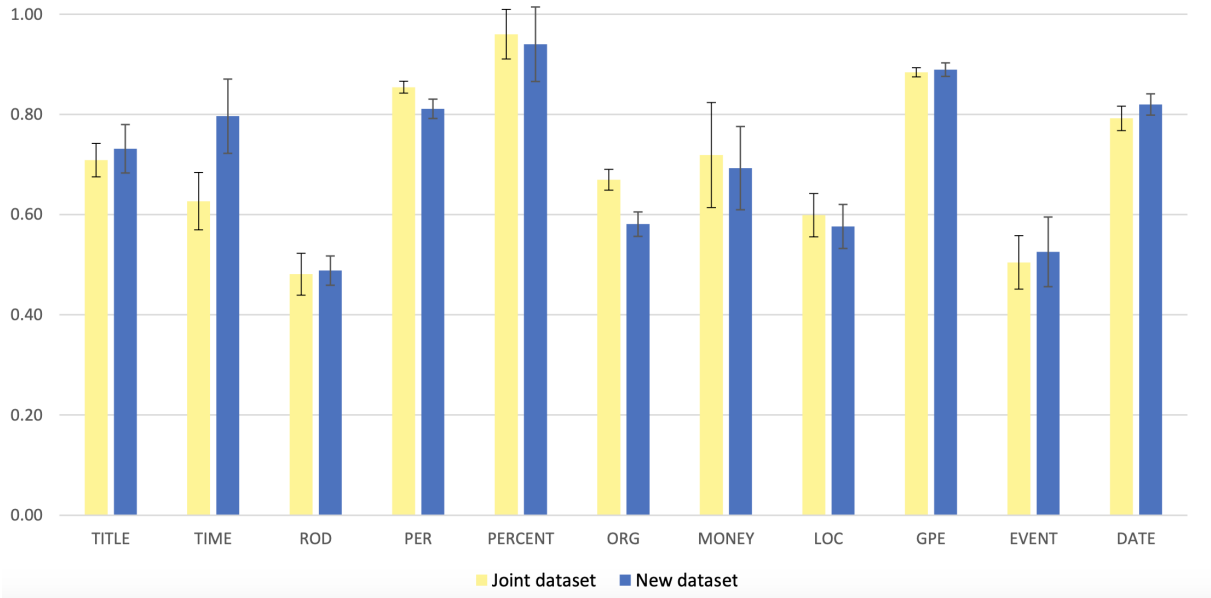
4.3 Test Results

The test results of the Joint model on the joint validation set are shown in the fourth column of the Table 5. The overall F1-score is somewhat higher on the test set than on the validation set. For some entities (PER, ORG, TITLE, DATE, TIME, MONEY), the test score is higher than the validation score and for others it is somewhat lower. The test F1-score drops the most for the EVENT entity (from 0.370 to 0.264).

All previous results were presented as averages over ten different runs. Finally we also picked one model to make it publicly available. We chose the Joint model with the highest overall F1-score on the validation set. The test scores of this models are shown in the right-most block of the Table 5. The overall F1-score of this best model is in line with the mean F1-score, which means that it was not the model with the highest F1-score. However, as the standard deviations are small, all the results



(a) Evaluation on the Main NER validation set.



(b) Evaluation on the New NER validation set.

Figure 1: Entity-level comparison of the Joint model with models trained on each dataset separately.

	Main+New Val F1	Main Val F1	New Val F1	Main+New Test F1	Main+New Test Prec	Rec	F1
PER	.868 (.007)	.872 (.008)	.854 (.012)	.879 (.007)	.840	.927	.882
ORG	.690 (.010)	.702 (.009)	.669 (.021)	.700 (.016)	.698	.693	.696
LOC	.549 (.019)	.541 (.021)	.599 (.043)	.526 (.025)	.478	.563	.517
GPE	.849 (.005)	.843 (.005)	.884 (.009)	.826 (.004)	.827	.830	.828
TITLE	.733 (.013)	.737 (.011)	.709 (.034)	.777 (.017)	.788	.758	.773
PROD	.598 (.018)	.634 (.028)	.481 (.042)	.568 (.020)	.576	.579	.578
EVENT	.370 (.053)	.310 (.043)	.504 (.053)	.264 (.034)	.306	.256	.278
DATE	.708 (.013)	.699 (.016)	.792 (.024)	.740 (.010)	.727	.768	.747
TIME	.451 (.065)	.433 (.075)	.627 (.057)	.463 (.043)	.548	.472	.507
PERCENT	.969 (.019)	.969 (.013)	.960 (.049)	.958 (.013)	.967	.983	.975
MONEY	.622 (.032)	.625 (.042)	.719 (.105)	.699 (.014)	.789	.614	.690
Overall	.761 (.004)	.766 (.002)	.752 (.010)	.773 (.006)	.766	.783	.774

Table 5: Evaluations of the Joint models trained on the joined train sets of the both datasets. Left block: F1-scores on the different parts of the validation sets. Middle block: F1-scores on the joined test set. Right block: test scores of the best Joint model.

of all models are in close range, with the highest F1-score obtained on the test set was 0.785.

5 Discussion

Although NER datasets have been created before for the Estonian language, this study presents the first attempt to annotate a richer set of entities beyond the most common person, organization and location names. As can be seen from inter-annotator agreements, there were some entities (PER, GPE, PERCENT) that the annotators labelled with high consistency, while the reliability is lower for the other entities. The annotations of the EVENT entities had to lowest inter-annotator agreement, which suggests that the inconsistencies in annotation in this and other entities could be analysed more thoroughly to understand the sources of confusion and improve the annotation guidelines.

Following previous attempts in other languages (notably Finnish), we decided to annotate also nested entities, by allowing up to three levels of nesting. The data statistics showed that very few entities were annotated on the 3rd level and even when considerable number of entities were labelled on the 2nd level, their reliability in terms of inter-annotator agreements is not high enough and thus using these labels for training predictive models might not be productive.

The experimental results with the BERT-based model showed that while there seems to be some domain shift between the two datasets on the level on some entities, overall training a single joint model on both datasets is justified. We only trained baseline models based on EstBERT and according to previous studies (Kittask et al., 2020; Tanvir et al., 2021), adopting other base models like Estonian WikiBERT (Pyysalo et al., 2021) or XLM-RoBERTa might lead to higher results.

6 Conclusions

We described the annotation process of the two Estonian NER datasets, labelled with a rich annotation scheme involving eleven different entity types. The datasets also include nested annotations of up two three levels, although the reliability of the nested annotations proved to be much less reliable than the first level entities. To establish the baseline predictive accuracy, we experimented with two modeling scenarios on these newly annotated datasets, by training two models, one for each dataset separately, and by training a joint model on the joined dataset. Overall, the joint model performed better than the separate models, except for few entity types, suggesting that the domain differences between these datasets are relatively small. Therefore we suggest using these two datasets jointly as a single, more varied, dataset.

References

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Lancaster University.

Claudia Kittask, Kirill Milintsevich, and Kairit Sirts. 2020. Evaluating multilingual bert for estonian. In *Baltic HLT*, pages 19–26.

Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. 2022. Named entity recognition in estonian 19th century parish court records. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. Wikibert models: Deep transfer learning for many languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54:247–272.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. Estbert: A pretrained language-specific bert for estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.