Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4615–4624
Marseille, 11–16 May 2020

A Broad-coverage Corpus for Finnish Named Entity Recognition

Jouni Luoma*, **Miika Oinonen***, **Maria Pyykönen, Veronika Laippala, Sampo Pyysalo**
Turku NLP Group, University of Turku, Turku, Finland
{jouni.a.luoma,mhtoin,maria.h.pyykonen,mavela,sampo.pyysalo}@utu.fi

Abstract

We present a new manually annotated corpus for broad-coverage named entity recognition for Finnish. Building on the original Universal Dependencies Finnish corpus of 754 documents (200,000 tokens) representing ten different genres of text, we introduce annotation marking person, organization, location, product and event names as well as dates. The new annotation identifies in total over 10,000 mentions. An evaluation of inter-annotator agreement indicates that the quality and consistency of annotation are high, at 94.5% F-score for exact match. A comprehensive evaluation using state-of-the-art machine learning methods demonstrates that the new resource maintains compatibility with a previously released single-domain corpus for Finnish NER and makes it possible to recognize named entity mentions in texts drawn from most domains at precision and recall approaching or exceeding 90%. Remaining challenges such as the identification of names in blog posts and transcribed speech are also identified. The newly introduced Turku NER corpus and related resources introduced in this work are released under open licenses via `https://turkunlp.org/turku-ner-corpus`.

Keywords: Named Entity Recognition, Finnish, Corpus, Annotation, Deep Learning

1. Introduction

Named entity recognition (NER) is a fundamental task in natural language processing (NLP), a key part of information extraction, and a prerequisite for many text mining goals. NER has been a major focus of annotation and method development efforts for decades (Grishman and Sundheim, 1996) and well-established reference corpora annotated for named entities are available for many languages (Tjong Kim Sang, 2002; Sang and De Meulder, 2003; Hovy et al., 2006). For languages with corpora of sufficient size and coverage, machine learning methods for NER can achieve very high performance, approaching human annotation quality (Chiu and Nichols, 2016; Devlin et al., 2018; Baevski et al., 2019). Although a degree of success has been demonstrated in multilingual and cross-lingual methods for NER (Al-Rfou et al., 2015), manually annotated corpora for each language remain required to fully realize the benefits of the most recent advances in NER methodology, and many lower-resourced languages still lack broad-coverage corpora of sufficient size and quality to train NER methods comparable to those available for high-resource languages.

In this paper, we focus on Finnish, a Uralic language spoken (nearly exclusively) by approx. 5 million people of the 5.5 million Finnish population. Alongside with the minority language Swedish, Finnish is the official language of the country. Finnish has a long and strong tradition of linguistic research (Setälä, 1880; Hakulinen et al., 2004) and a number of modern language resources and NLP tools are available for the language, such as manually annotated treebanks (Haverinen et al., 2014) and morphological and syntactic analyzers (Pirinen, 2015; Kanerva et al., 2018). However, resources for Finnish named entity recognition are lacking: a manually annotated corpus for Finnish NER, FiNER, was only recently made available (Ruokolainen et al., 2019), and its training data only covers a single specialized text domain, namely technology news.

In this paper, we present a new manually annotated NER corpus that emphasizes broad coverage of different genres, topics, and styles of writing. We draw on the source texts and existing manual annotation of the original Universal Dependencies (Nivre et al., 2016) Finnish treebank and Finnish NER annotation guidelines and tools from the FiNER effort to create an open broad-coverage corpus suitable for training modern NER methods for Finnish. Our results demonstrate that the resulting annotation has high internal consistency, is compatible with existing resources, and can support accurate Finnish NER using deep transfer learning methods.

2. Related work

The direction of NER research has for long been guided by influential programs and shared tasks such as the Message Understanding Conferences (Grishman and Sundheim, 1996), the Automatic Content Extraction program (Doddington et al., 2004) and the CoNLL shared tasks on language-independent NER (Tjong Kim Sang, 2002; Sang and De Meulder, 2003). While these efforts have included also many other targets, they have been instrumental in cementing the recognition of *person*, *organization* and *location* names as a core NER goal, with the identification of time expressions as a frequent associated theme. While early efforts focused largely on English, NER methods have long aimed for language independence, supported by resources such as the comparably annotated corpora introduced for Spanish and Dutch in the CoNLL 2002 shared task and English and German in 2003.

Today, NER corpora covering a range of domains and entity types are available for many major languages and basic NER resources are an expected component of the basic NLP toolkit for any language (e.g. (Hovy et al., 2006; Taulé et al., 2008; Singh et al., 2009; Munkhjargal et al., 2015)). For Nordic and other countries near Finland, NER corpora and models exist for Swedish (Almgren et al., 2016), Norwegian (Johansen, 2019) Danish (Derczynski et al., 2014), Icelandic (Ingólfsdóttir et al., 2019), Latvian and Lithua-

* Equal contribution

Section	Documents	Sentences	Tokens
Wikipedia articles	200	2 269	31 906
Wikinews articles	100	1 120	14 281
University online news	50	942	13 232
Blog entries	77	1 781	22 287
Student magazine articles	23	1 058	14 390
Grammar examples	80	2 002	16 982
Europarl speeches	80	1 082	19 932
JRC-Acquis legislation	29	1 141	23 920
Financial news	50	1 002	12 477
Fiction	65	2 739	32 709
Total	754	15 136	202 116

Table 1: Universal Dependencies Finnish TDT corpus statistics by genre. The source data for the grammar examples section of the corpus did not originally have document structure, and for *Grammar examples* the given number of documents reflects their somewhat arbitrary grouping for data distribution.

nian (Pinnis, 2012), and Estonian (Tkachenko et al., 2013). It is therefore surprising that the first manually annotated corpus for Finnish NER, FiNER, was introduced only recently (Ruokolainen et al., 2019). The primary texts of this corpus were drawn from a Finnish technology news magazine[1], with training and development sets containing articles from 2014 and the test set articles from 2015. The corpus additionally contains an "out-of-domain" test set of Wikipedia articles. Ruokolainen et al. (2019) evaluated two recently proposed deep learning-based NER methods on the corpus (Güngör et al., 2018; Sohrab and Miwa, 2018) and found that while these methods achieve satisfactory results on the primary in-domain test set (reaching 85% F-score), performance drops precipitously for the out-of-domain data, where the best machine learning-based result is only 61% F-score. This failure of existing resources to support accurate machine learning for Finnish NER is a primary motivator for our work, and a central goal of our effort is to create a corpus that allows state-of-the-art NER methods to be trained for Finnish to achieve a high level of recognition performance across multiple domains.

3. Corpus annotation

We next introduce the source data, the annotation targets and the manual annotation process of the newly created Turku NER corpus and briefly discuss some details of the annotation guidelines.

3.1. Data

We draw the texts for our Finnish NER corpus from the Universal Dependencies (UD) (Nivre et al., 2016) version of the Turku Dependency Treebank (TDT) corpus (Haverinen et al., 2014; Pyysalo et al., 2015). TDT is a broad-coverage corpus spanning a range of text domains including news, blog posts, and legal texts (Table 1) with manual annotation for morphology and dependency syntax. One of the benefits of adding a named entity annotation layer to the treebank data is that its existing annotations can support the annotation effort. First, like all UD treebanks, the TDT corpus annotation includes part-of-speech tagging according to the Universal POS tagset defined by the UD project.

[1] `http://www.digitoday.fi/`

	Documents	Sentences	Tokens
Train	602	12 217	162 746
Dev	76	1 364	18 308
Test	76	1 555	21 062
Total	754	15 136	202 116

Table 2: Universal Dependencies Finnish TDT corpus statistics for training, development and test data.

In this formalism, the tag `PROPN` is used to mark proper nouns, defined as *a noun (or nominal content word) that is the name (or part of the name) of a specific individual, place, or object.*[2] Second, the enhanced dependency annotation layer of the corpus includes `flat:name` dependency annotation, an extension of UD marking a sequence of words as a name. Although the UD definitions of names are not typed and not expected to exactly match the scope of NER annotation, words tagged `PROPN` or spanned by a `flat:name` pseudo-dependency are strong candidates for named entity annotation, and this data was used as a starting point for the annotation. For the training, development and test sets of the new Turku NER corpus, we follow the existing splits of the TDT corpus, summarized in Table 2.

3.2. Annotation targets

To assure compatibility with existing resources and approaches, the annotation aims to follow established conventions for NER corpora. Entity mentions are marked as continuous non-overlapping spans of text where each mention is assigned a single type (person, organization, etc.). The boundaries of mentions are required to align with syntactic words so that each word is either fully included in a mention or not part of one. For the specific definition of *syntactic word*, we follow the Universal Dependencies approach as implemented in the UD Finnish TDT corpus, thus preserving token alignment with the existing resource.

Following the approach of Ruokolainen et al. (2019), the annotation targets six classes of mentions: person (PER), organization (ORG), location (LOC), product (PRO), event (EVENT), and data (DATE). The first three broadly match

[2] `https://universaldependencies.org/u/pos/PROPN.html`

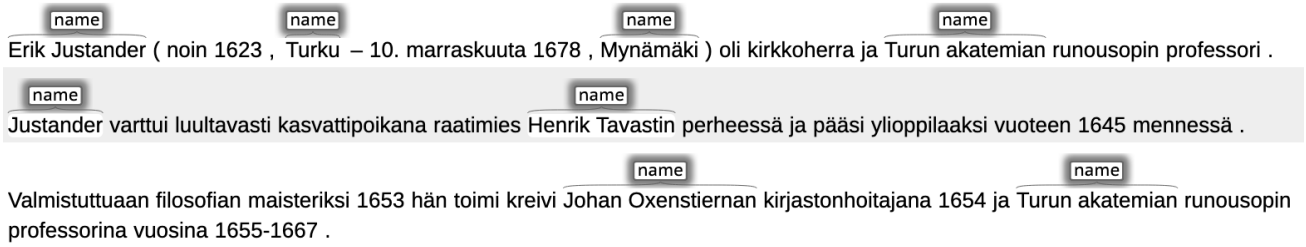

Figure 1: Examples of candidates for annotation derived from existing TDT corpus morphosyntactic annotation.

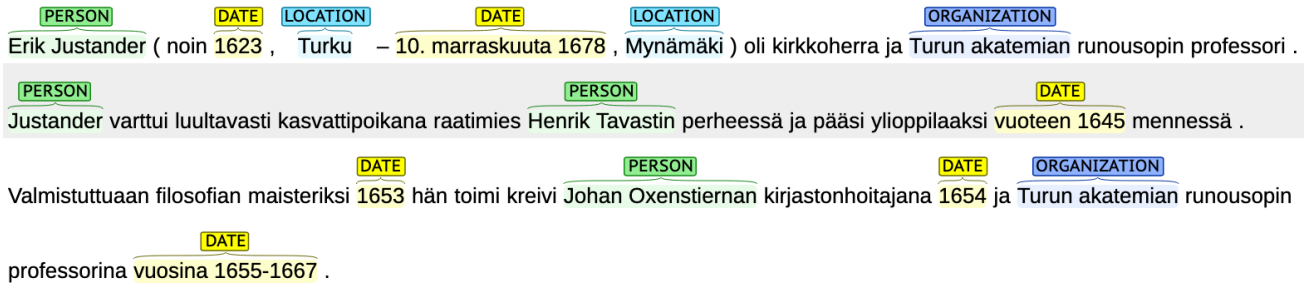

Figure 2: Example annotation. Translation: Erik Justander (approx. 1623, Turku - March 10 1678, Mynämäki) was a vicar and a professor of poetics at the Academy of Turku. Justander grew up as the foster child in magistrate Henrik Tavasti's family and passed his matriculation examination by the year 1645. After graduating university as a Master of Arts in 1653 he acted as count Johan Oxenstiern's librarian in 1654 and as the professor of poetics for the Academy of Turku in the years 1655-1667.

the core MUC/CoNLL types of the same names, while PRO and EVENT capture specific categories that would have been annotated under the broad and diverse MISC type in CoNLL (Sang and De Meulder, 2003). Although dates are not named entities under most definitions of the term, the recognition of time expressions has frequently been considered along with NER in efforts since MUC, and we here follow the associated slightly imprecise usage.

The types and scope of the annotation as well as the specific guidelines for each annotated type were defined following the previously introduced FiNER corpus to allow comparison and combinations of the resources. The FiNER annotation guidelines (Ruokolainen et al., 2019) and associated materials such as the FiNER tagger documentation[3] were used as reference and expanded on to cover specific cases arising during the annotation effort.

3.3. Annotation process

As a starting point for their work, annotators were provided with documents where words tagged `PROPN` or covered by a `flat:name` dependency in the source data were marked generically as `name` candidates,[4] illustrated in Figure 1. Annotators were then required to either identify the correct type and span for each candidate or delete the candidate as out of scope, as well as to complete the annotation by marking all relevant mentions not appearing in the initial set of candidates.

The corpus was annotated using BRAT, a web-based tool for text annotation (Stenetorp et al., 2012). As the tool uses a custom standoff representation for annotations,[5] we created a simple conversion script for initially extracting tokenized texts and name candidates from the TDT corpus CoNLL-U format representation, and adapted tools provided with BRAT to convert the standoff data into the CoNLL IOB2 representation for experiments.

Initial exploratory annotation was created for part of the data by 11 students as part of NLP course projects. This annotation was then made consistent and extended to cover the entire corpus by one primary annotator working with an annotation coordinator. Issues and open questions encountered in the annotation were logged and compiled into annotation guidelines extending and further specifying the application of the FiNER guidelines to phenomena encountered in the new text domains.

To improve consistency with the FiNER annotation, the corpus texts were then tagged using the FiNER tagger (Kettunen and Löfberg, 2017) and differences in annotation examined manually to identify potential divergences from the annotation criteria of the FiNER effort. This use of the tool could be viewed as potentially introducing a bias in favour of the FiNER tagger: for example, annotation errors of omission that are caught by the tagger are more likely to be fixed than ones that are not. However, we consider this risk of bias to be minor and note that the tagger represents a baseline method for our study, and any possible bias introduced by this cross-check would thus work *against* the methods for Finnish NER proposed in this study. Figure 2 shows an example of the corpus annotation.

[3] `https://github.com/Traubert/FiNer-rules/`

[4] We note that as both the parts of speech and the dependencies in the source data had been fully manually annotated, providing these candidates to annotators does not introduce any potential bias toward particular automated methods.

[5] `http://brat.nlplab.org/standoff`

3.4. Annotation guidelines

A set of detailed annotation guidelines extending on those of the FiNER corpus (Ruokolainen et al., 2019) were prepared to support the annotation effort, and the complete Turku NER corpus annotation guidelines are provided along with the corpus data.[6] Due to space constraints, in the following we briefly discuss some of the key corpus annotation guidelines.

Excepting for DATE, annotated mentions typically involve one or more proper nouns that specifically identify an entity of a targeted type (person, location, etc.). Common head nouns (e.g. *katedraali*, "cathedral") are included in the span of annotations when they further specify such an entity (e.g. *Uspenskin katedraali*, "Uspenski Cathedral"). In cases where a noun phrase does not refer to a specific targeted entity, common head nouns are excluded from the span of annotations: for example, for *Suomen talous* ("Finland's economy"), only the proper noun *Suomen* ("Finland's") is annotated. Inflectional affixes are very common in Finnish (e.g. *-ssa* in *Turussa*, "in Turku"), and included in the span of annotations as part of a syntactic word. Affixes are also included in when separated from the noun due to for example tokenization (e.g. *NBA :ssa*, "in the NBA"). Similarly, hyphenated compounds such as *Youtube-sivustolla* ("on the Youtube website") are either annotated as a whole, when the compound refers to the same entity as the proper name, or not at all. As for inflectional affixes, this rule applies also when there is space separating the parts of the compound. As an extension of this rule, head nouns following a hyphenated compound with *-niminen* ("-named") or comparable expressions are included in the span of annotations (e.g. *Accenture -niminen firma*, "a company named Accenture").

Following the FiNER guidelines, we also include quotation marks in the span of annotation in cases in which the entity or part of it is enclosed in quotes (e.g. *"Simpsonit", "Maailman vahvin"-turnauksessa*) in order to ensure that all cases involving quotes are treated consistently in the annotation. When two or more entities are mentioned in a coordinate construction in which ellipsis is used to avoid repetition, as in *Spotify Free ja Open -tileille* ("Spotify Free [accounts] and Spotify Open accounts"), the whole expression is annotated as one mention. Finally, any abbreviations or acronyms appearing immediately after an entity mention are marked as part of a single annotation covering both the full form and the abbreviation, as in for example *Turku Centre for Computer Science (TUCS)*.

4. Methods

In this section, we present the data preprocessing, introduce the NER methods used in the experiments, and detail the experimental setup and evaluation criteria.

4.1. Data preprocessing

For NER experiments, the corpus annotation is cast from the source standoff format into the simple IOB2 representation used in the CoNLL shared tasks, where each word is tagged as either beginning a mention (B), in a mention (I), or out (O), i.e. not part of a mention, with the relevant mention type affixed to the B and I tags. The representation is illustrated in Table 3. Finnish NER can then be addressed as a standard sequence labeling task using a broad range of existing tools supporting the representation.

Erik	B-PER
Justander	I-PER
oli	O
Turun	B-ORG
akatemian	I-ORG
professori	O
.	O

Table 3: Example IOB2 named entity annotation. Translation: Erik Justander was a professor for the Academy of Turku.

4.2. NER methods

We apply a number of machine learning approaches as well as a previously introduced rule-based system to assess the corpus and the NER task it represents.

FiNER tagger (Kettunen and Löfberg, 2017; Ruokolainen et al., 2019) is a dictionary- and rule-based tagger for Finnish NER that has been developed together with the FiNER corpus. The system is based on a combination of morphological analysis and tagging tools (Pirinen, 2015) and an extensive dictionary of known names together with pattern-matching rules (Hardwick et al., 2015) to detect and classify targeted mentions.

Simple CRF We apply a simple baseline tagger using explicitly defined features derived from the surface forms of words using the CRFsuite (Okazaki, 2007) implementation of first-order linear chain conditional random fields (CRFs) (Lafferty et al., 2001), a probabilistic sequence labeling model underlying many NER methods. Specifically, we use the features defined by the CRFsuite NER feature extraction example, including focus and context word prefixes, suffixes, shape features, and combinations of these. We refer to the documentation and implementation distributed with CRFsuite[7] for the full details of the feature representation. We note that this baseline is intentionally knowledge-poor, not incorporating e.g. dictionary or word vector features.

BiLSTM-CNN-CRF We use the NCRF++ neural sequence labeling toolkit (Yang and Zhang, 2018) implementation of the NER model proposed by Ma and Hovy (2016), which uses a concatenation of word vectors and character representations computed using convolutional neural networks (CNNs) to represent words as input to a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with a CRF output layer. This class of models represented the state of the art in neural NER methods prior to the introduction of deep transfer learning methods such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018).

[6] https://turkunlp.org/turku-ner-corpus

[7] http://www.chokkan.org/software/crfsuite/

BERT (Devlin et al., 2018) is a state-of-the-art deep transfer learning approach based on the transformer model (Vaswani et al., 2017) and pre-training on large unannotated corpora. BERT can be readily applied to sequence labeling tasks such as NER by attaching a time-distributed dense layer on top of the model output layer and fine-tuning the model on data with named entity annotation. In this work, we apply the recently introduced FinBERT model[8] pre-trained from scratch on Finnish data (Virtanen et al., 2019). As the official BERT implementation does not directly support NER, we apply a custom tool based on the NER approach described by Devlin et al. (2018), implemented using Keras[9] and Keras-BERT[10] and available from `https://github.com/jouniluoma/keras-bert-ner/`.

4.3. Experimental setup

The hyperparameters and other settings for all NER methods were selected by evaluating alternative configurations on the development subset of the data. The test data was held out throughout parameter selection and only used in the final experiments.

When applying the **FiNER tagger**, we run the system with fixed tokenization and map the fine-grained tags assigned by the tagger (e.g. `OrgPlt` for *political organization*) to their corresponding top-level categories (e.g. `ORG`) using a simple postprocessing script. Based on evaluation on development data, we discard any generated numeric expression annotations (money and units) as well as the subcategories `PrsTit` (titles) and `TmeHrm` (times of day) as out of scope with respect to the FiNER corpus guidelines.

For the **Simple CRF** model based on explicitly defined features, we selected the L_2 regularization parameter by evaluating performance for c_2 values $\{2^{-20}, 2^{-19}, \ldots 2^{10}\}$ on the development data, selecting 2^{-13} for the final experiments. Other hyperparameters were left at their CRFsuite default values, using L-BFGS optimization without L_1 regularization or limiting the number of iterations.

For the **BiLSTM-CNN-CRF**, we initialize the model word vectors using the 100-dimensional word2vec skip-gram embeddings (Mikolov et al., 2013) introduced for the CoNLL 2017 shared task (Ginter et al., 2017). Other than the word vector dimension, other parameters are left at the defaults defined in the NCRF++ toolkit, including 30-dimensional character embeddings, hidden layer dimension 200, and training with stochastic gradient descent for a maximum of 100 epochs with early stopping, selecting the best-performing model on the development set.

To set **BERT** hyperparameters, we followed Devlin et al. (2018), selecting the batch size, learning rate and epochs using an exhaustive grid search, otherwise using the values suggested in the BERT manuscript but skipping batch size 32 due to GPU memory limitations. Other parameters were left at their defaults, using sequence length 512, an Adam optimizer with warmup, linear training rate decay and weight decay with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-6$ and weight decay rate 0.01.

[8] `https://turkunlp.org/finbert`

[9] `https://keras.io/`

[10] `https://github.com/CyberZHG/keras-bert`

	Train	Dev	Test	Total
LOC	2,694	288	287	3,269
PER	2,477	298	310	3,085
ORG	2,154	239	208	2,601
DATE	1,099	119	114	1,332
PRO	799	102	79	980
EVENT	157	17	7	181
Total	9,380	1,063	1,005	11,448

Table 4: Turku NER corpus annotation statistics

The task is cast as sentence-level sequence labeling for all methods except BERT, for which we aim to replicate the use of broader context as done by Devlin et al. (2018) for CoNLL English data. Specifically, for each input sentence, as many of the following sentences as fit into the BERT window of 512 tokens are concatenated as context. In prediction, only the labels of the initial sentence of each sequence are used. We note that this approach only approximates true document context and will in cases join together sentences from unrelated documents.

For cross-corpus experiments, the hyperparameter selection is done separately for both corpora (Turku NER and FiNER) as well as their combination, selecting the parameters providing the best performance on the corresponding development data. For the neural methods with random initialization and non-convex optimization problems, we repeat each experiment five times and report averages for each evaluation metric.

4.4. Evaluation criteria

We evaluate NER performance in terms of exact mention-level precision, recall and F-score as implemented in the standard `conlleval` script, requiring both the type and span of predicted mentions to match a gold standard mention and summarizing results as microaverages.

5. Results

We next present key statistics for the newly introduced Turku NER corpus annotation, an evaluation of the annotation quality, and the results of the evaluation of the various NER methods on the corpus.

5.1. Corpus statistics

Table 4 summarizes key statistics for the Turku NER corpus annotation. Throughout the corpus, the three most prominent entity types are LOC, PER and ORG, which comprise 28.6%, 26.9%, and 22.7% of all annotations, respectively. DATE and PRO are fairly frequent as well, comprising 11.6% and 8.6% of all annotations (resp.), while EVENT mentions are rather infrequent in the corpus, with only 1.5% of the total categorized as an event.

The entity mentions in the Turku NER corpus are more evenly distributed than they are for example in the FiNER corpus: in the latter, the most frequent entity class was, by far, ORG (48.4% of all top-level entities), followed by PRO (23.6%). This is most likely due to the fact that the FiNER corpus consists of technology news articles, in which organizations and products naturally play an integral part, while

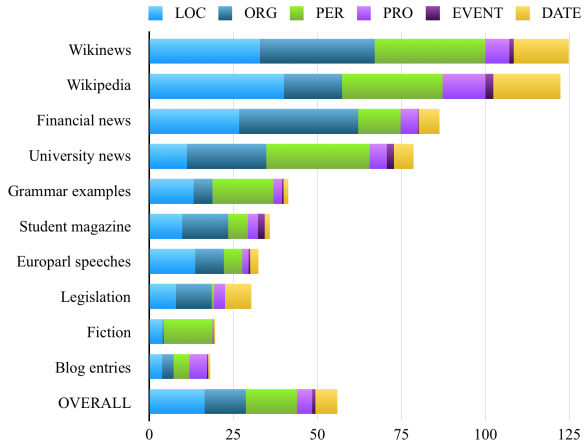

Figure 3: Average number of annotations of each type per 1000 tokens

the Turku NER corpus includes a noticeably more wide-ranging variety of different text types (Table 1).
Figure 3 illustrates the annotation density per genre. There are clear differences between the different subcorpora: while Wiki documents average over 120 annotations for 1000 tokens, e.g. fiction and blog entries average below 20. Overall, the average density is 56 annotations per 1000 tokens, a comparatively low number: as one point of comparison, the CoNLL 2003 English corpus averages 116 annotations per 1000 tokens.

5.2. Annotation quality

To assess the quality and consistency of the annotation, an additional annotator was trained after the initial annotation process (Section 3.3.) and the 76 documents constituting the test set of the corpus were independently re-annotated. The two sets of annotations were then compared to each other, and both overall and entity-wise F-scores were calculated to establish inter-annotator agreement (IAA). In addition, a qualitative comparison of the points of disagreement was conducted in order to correct the final gold standard annotation where necessary.
Overall agreement reached an F-score of 94.40% using the standard exact conlleval maching criterion. Under a relaxed *overlap* matching criterion where differences in annotated entity spans (but not types) were permitted, the F-score was 96.35%. This indicates that the two independent sets of annotations were, in general, very well aligned, and also reflects well on the quality and consistency of annotation in the remaining dataset. In the IAA experiment, the highest entity-wise agreement rates were measured for the categories PER and DATE, which yielded F-scores of 98.55% (overlap 98.87%) and 98.25% (both exact and overlap), respectively. Agreement was also at a high level with regard to LOC mentions, the F-score being 94.65% (overlap 96.55%). For ORG, while the overlap match F-score was also quite high (94.62%), issues regarding annotation boundaries lowered the exact F-score to 91.44%. The most prominent point of disagreement were cases in which it was not completely clear whether a common noun following a proper noun should be included in the span of annotations or not, such as *ensemble* in *Raatikon ensemble*. There were also cases in which it was difficult to distinguish whether a mention refers to a single entity or two separate ones, such as *Sanoma Television Oy / Nelonen Median*, which refers to a single entity despite the separating slash.
The class PRO, which elicited an exact F-score of 82.19% and 88.37% for overlap, also suffered from some issues with boundaries. For example in the case of *Mikael Agricolan Psalttari* (lit. "Mikael Agricola's Book of Psalms", referring to The Book of Psalms translated by Mikael Agricola), it was not completely clear whether person name (*Mikael Agricolan*) should be annotated as part of the name of the book or a separate PER mention. The class EVENT was found particularly challenging, with a noticeably lower exact match F-score of 66.67% (80.00% for overlap) than for the other entity categories. This is most likely due in part to EVENT mentions being very rare in the data used for the IAA experiment: the first annotator had tagged six mentions of the class, and the second nine. For EVENT, the differences between the annotators not related to boundary issues were limited to two events: *Kulttuuripääkaupunkivuosi 2011* ("Culture Capital year 2011") and *Moskovan sisäratojen MM-kisoissa 2006* ("in the Moscow's inside track Word Championship 2006").

Method	Prec.	Rec.	F-score
FinBERT	**90.87**	**92.44**	**91.65**
BiLSTM-CNN-CRF	82.92	80.20	81.54
FiNER tagger	77.16	71.24	74.08
Simple CRF	74.53	63.18	68.39

Table 5: Comparison of NER methods on Turku NER corpus

5.3. NER method comparison

Table 5 presents the results for the various methods on the Turku NER corpus test data.
The simple CRF method making use only of explicitly defined word surface form features performs poorly, with the low recall (63%) in particular indicating that the corpus represents a difficult challenge for the knowledge-poor machine learning approach. As expected for a method based on dictionary matching and manually crafted rules, the FiNER tagger likewise has comparatively high precision and low recall. Its overall performance on the new corpus, 74% F-score, is lower but broadly comparable to the 78% F-score reported by Ruokolainen et al. (2019) for the tagger on their out-of-domain Wikipedia test data. Given the broader scope of the Turku NER corpus and its inclusion of texts written in informal Finnish, we find the FiNER tagger performance here a positive indication of compatibility with FiNER resources.
The best performance is achieved by the two neural methods, with the BiLSTM-CNN-CRF achieving largely balanced precision and recall above the 80% level. In addition to the neural feature learning framework, this model has the advantage of a stronger context model via the bidirectional LSTM as well as the ability to incorporate information from pretraining on a large unannotated corpus through its word vector initialization. The best results by a clear margin are nevertheless achieved by the language-

Type	Prec.	Rec.	F-score
DATE	95.88	97.90	96.87
PER	93.69	96.77	95.21
LOC	93.80	95.75	94.76
ORG	89.29	91.25	90.25
PRO	70.16	62.28	65.86
EVENT	37.94	51.43	43.52

Table 6: BERT performance for different entity types

specific BERT model, which exceeds 90% performance in terms of both precision and recall, approximately halving the error rate of the BiLSTM-CNN-CRF. This result is in line with recent findings indicating that deep transformer-based transfer learning methods such as BERT represent a substantial advance also for sequence labeling tasks such as NER (Devlin et al., 2018).

Given the clear advantage of FinBERT over the other methods considered in this comparison, we chose to focus the performance on the BERT model in the remainder of our analysis.

5.4. Analysis of NER performance

Table 6 details the performance of the FinBERT tagger on the Turku NER corpus test data by entity type. Interestingly, we find that the ordering of the entity types from most to least reliably tagged mirrors that of the human IAA experiment: DATE, PER and LOC recognition performance approaches or exceeds 95% F-score, ORG performs somewhat below, around 90%, with PRO and EVENT showing notably poorer performance. The very low result for EVENT may be explained in part by annotation sparsity in the training data and isolated difficult cases in the test. The low (66%) result for PRO, for which approx. 800 training examples were available (Table 4) is more surprising, and may warrant further examination of the definition and annotation of this class.

To evaluate performance on text representing different genres, the test set was split into ten different parts, each containing only documents from a specific section of the corpus such as financial news and Wikipedia articles. A model was trained on the full training set using hyperparameters selected for the full development set and then evaluated on each of the test subsets. The results of this evaluation are shown in Table 7. The highest performance is achieved for sections representing news of various types, with F-scores over 95% for both Financial news and Wikinews. This performance could potentially be explained with news language being highly formal and standardised and thus easy for the model to learn. By contrast, the lowest-performing sections, blog entries and transcribed speech (Europarl), are informal and non-standard in their own ways. Blog entries can contain variations such as nicknames that could prove difficult for generalisation, while transcribed speech may contain its own idiosyncrasies and irregularities.

Overall, the analysis indicates that the Turku NER corpus allows a BERT model to be trained to achieve satisfactory or high performance for most named entity types and domains, with remaining issues in the recognition of product and event names and some specialized genres of text.

Section	Prec.	Rec.	F1-score
Financial news	98.84	94.44	96.59
Wikinews	95.91	94.80	95.35
Student magazine	91.49	97.73	94.51
Grammar example	94.83	90.16	92.44
Fiction	88.54	96.59	92.39
Wikipedia	90.88	92.17	91.52
University news	87.39	92.38	89.81
Legistlation	89.23	86.57	87.88
Europarl	91.67	80.88	85.94
Blog entry	73.53	89.29	80.65

Table 7: BERT performance for different corpus sections

	Test			
Train	FiNER/ news	FiNER/ wiki	Turku NER	Comb.
FiNER	92.98	82.62	84.77	91.28
Turku NER	88.98	89.25	91.64	89.30
Combined	**93.26**	**89.88**	**92.09**	**93.11**

Table 8: Results for cross-corpus evaluation

5.5. Cross-corpus evaluation

To assess the compatibility of the newly introduced annotations with the previously released FiNER corpus, we performed cross-corpus experiments where the FinBERT model was variously fine-tuned on the Turku NER corpus or FiNER corpus training data or their combination, and tested on the three test sets defined by these resources as well as their combination. Table 8 summarizes the results of the cross-corpus evaluation. We first note that training on the Turku NER corpus data rather than on the FiNER technology news training set gives better results not only for the Turku NER corpus test data but also for the FiNER out-of-domain Wikipedia test set, suggesting that the new corpus succeeds in representing also this domain. Performance on the FiNER technology news test set is nevertheless better when training with FiNER data, indicating that dedicated in-domain training data has additional value for this specialized domain.

We further note that the highest performance on each of the four test sets is achieved when combining the FiNER and Turku NER corpus training data, indicating that our efforts to maintain compatibility with the previously released corpus were successful and that the two resources can be pooled together for training without compromising performance on the specific target domains of either resource.

Overall, the best results achieved here are high, with the performance of the tagger trained on the combined training data falling only 2.5% points below the human inter-annotator results on the on the Turku NER corpus test set. Given the difficulties for training coherent models that frequently arise from corpus combinations and the challenges of replicating an annotation process even with good documentation, we find the positive results of this cross-corpus evaluation and the high combination performance in particular a very positive indication of the quality of both Finnish NER corpora.

6. Discussion

We next discuss some challenges encountered in the corpus annotation that may be instructive at least to readers specifically interested in Finnish NER.

The Finnish conventions of capitalization posed some challenges for annotation, as capitalization is not as reliable as an indicator of names in Finnish as it is in English. For example, while English tends to capitalize all nouns in multiword names, in Finnish, only the proper nouns are capitalized (e.g. *Euroopan unionin virallinen lehti* ("Official Journal of the European Union"). This eliminates one simple heuristic for distinguishing between names and nominal references. In cases in which it was difficult to distinguish between the two based on the context offered by the surrounding text alone, external sources were used to determine whether a multi-word expression is an established name for a specific entity, or rather a descriptive nominal construction. As a result, for example *Tiedeakatemian vuotuinen kunniapalkinto* ("The annual honorary award of the Finnish Academy of Science and Letters") is not annotated as it is not the official name of the prize in question, while *Valtioneuvoston asetus seulonnoista* ("The Government's Decree on Screenings") is annotated as one entity as it is an official act referred to by that name in an official Finnish database for legislative information (the Finlex Data Bank).

Another point of discussion were mentions in which it was not clear whether an expression enclosed in quotes constituted a product or not, as, following the FiNER guidelines, slogans were included in the category of products. This is made particularly challenging by the various forms that slogans take, as they vary from single word mentions to fully formed sentences. In the case of for example *"vähemmän kovaa rasvaa"* ("less saturated fat") it was debatable whether the expression was a slogan or just a quote describing the product in question – here, the mention was ruled to be a slogan (and thus annotated as PRO), as it appeared in a list with other slogans used in the marketing of the product in question.

In the fiction parts of the corpus, there were also some difficulties to identify the line between animate and inanimate entities with respect to the category PER. For example, in one of the texts, the name *Karri* is used multiple times in reference to an old vehicle but discussed in a distinctly animate manner, as in *Karriki on joskus ollut ihan oikeissa töissä, taksina tai jotain* ("Karri has also had a real job once, as a taxi or something"). In some cases, context several sentences removed from some mentions was required even for a human annotator to determine what is being referred to, representing obvious challenges for NER approaches operating on the basis of sentence contexts.

The agentive role played by the entities was also a point of frequent discussion during annotation, especially with regard to the distinction between an organization and its product in cases in which the two share a name (e.g. *Google, Facebook*). In these cases the decision was made based on the agentive or non-agentive role played by the entity in context: for example, in *Google onkin hyvä hakukone* ("Google is a good search engine"), *Google* is assigned the tag PRO, while in *Sekä Yandex että Google painottavat hakujensa puolueettomuutta* ("Both Google and Yandex underline the impartiality of their queries"), *Google* is annotated as ORG. For geopolitical entities, especially country names (e.g. *Suomi*), the annotation LOC is generally used also in cases in which the entity is used in an agentive manner, as in *Suomi julistautui itsenäiseksi* ("Finland declared its independence"), but the ORG label is used in the specific case where a geopolitical entity name is used to refer to a sports team representing that entity. These ambiguities are likely to represent challenges also for NER methods, and for applications where the resolution of the ambiguity is not critical, there may be merit to the adoption of types that subsume the ambiguity (cf. GPE).

7. Conclusions and Future Work

We have presented the Turku NER corpus, a manually annotated corpus for Finnish named entity recognition annotated with high internal consistency for six types of entity mentions and covering 200,000 tokens in 754 documents representing ten different genres of text. An evaluation using four NER methods showed that a Finnish BERT model trained on the corpus can achieve performance approaching or exceeding 90% precision and recall on most text domains. Analysis of performance identified potential remaining issues in the recognition of product and event names as well as in NER addressing blog posts and transcribed speech. Finally, a cross-corpus evaluation demonstrated the new corpus annotation to have high compatibility with a previously released Finnish technology news corpus, further showing that these resources can be straightforwardly combined to create training data allowing broader coverage and higher NER performance than either resource alone.

While the Turku NER corpus includes a broad range of domains and types of text, its coverage is far from perfect. In particular, there are no texts representing interactive social media such as Twitter or Suomi24, which represent highly relevant application areas for NER with distinctive, frequently very informal sublanguages that pose challenges for many NER systems. A natural future continuation of our effort would be to extend the annotation to cover additional domains such as these.

The newly annotated corpus and all supporting resources introduced in this study are available under open licenses from `https://turkunlp.org/turku-ner-corpus`.

8. Acknowledgements

We wish to thank the students of the University of Turku Textual Data Analysis 2019 course who contributed to the corpus annotation in their NER projects as well as Miikka Silfverberg and Teemu Ruokolainen for supporting our adaptation of the FiNER guidelines and helping resolve challenges in annotation. The work was partially funded the Foundation of Emil Aaltonen. Computational resources for this work were provided by CSC – Finnish IT Center for Science.

9. Bibliographical References

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594.

Almgren, S., Pavlov, S., and Mogren, O. (2016). Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 30–39.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*.

Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Derczynski, L., Field, C. V., and Bøgh, K. S. (2014). Dkie: Open source information extraction for danish. In *Proceedings of EACL*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC*, volume 2, page 1.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at ÚFAL.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Güngör, O., Üsküdarli, S., and Güngör, T. (2018). Improving named entity recognition by jointly learning to disambiguate morphological tags. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2082–2092.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., and Alho, I. (2004). *Iso suomen kielioppi [The reference grammar of Finnish]*. Suomalaisen Kirjallisuuden Seura.

Hardwick, S., Silfverberg, M., and Linden, K. (2015). Extracting semantic frames using hfst-pmatch. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 305–308.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Ingólfsdóttir, S. L., Þorsteinsson, S., and Loftsson, H. (2019). Towards high accuracy named entity recognition for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Johansen, B. (2019). Named-entity recognition for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 222–231.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Kettunen, K. and Löfberg, L. (2017). Tagging named entities in 19th century and modern Finnish newspaper material with a Finnish semantic tagger. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, number 131, pages 29–36.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pages 1064–1074.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Munkhjargal, Z., Bella, G., Chagnaa, A., and Giunchiglia, F. (2015). Named entity recognition for mongolian language. In *International Conference on Text, Speech, and Dialogue*, pages 243–251. Springer.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pinnis, M. (2012). Latvian and Lithuanian named entity recognition with TildeNER. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1258–1265.

Pirinen, T. A. (2015). Omorfi—free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315.

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., and Ginter, F. (2015). Universal dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–172.

Ruokolainen, T., Kauppinen, P., Silfverberg, M., and Lindén, K. (2019). A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Setälä, E. N. (1880). *Suomen kielen lause-oppi. [The syntax of Finnish.]*. K. E. Holm.

Singh, T. D., Nongmeikapam, K., Ekbal, A., and Bandyopadhyay, S. (2009). Named entity recognition for manipuri using support vector machine. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 811–818.

Sohrab, M. G. and Miwa, M. (2018). Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at EACL*, pages 102–107.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of LREC*.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tkachenko, A., Petmanson, T., and Laur, S. (2013). Named entity recognition in Estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for finnish.

Yang, J. and Zhang, Y. (2018). NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL*.