Fine-grained Named Entity Annotation for Finnish

Jouni Luoma, Li-Hsin Chang, Filip Ginter, Sampo Pyysalo
TurkuNLP group
Department of Computing,
Faculty of Technology
University of Turku, Finland
{jouni.a.luoma,lhchan,figint,sampo.pyysalo}@utu.fi

Abstract

We introduce a corpus with fine-grained named entity annotation for Finnish, following the OntoNotes guidelines to create a resource that is cross-lingually compatible with existing resources for other languages. We combine and extend two NER corpora recently introduced for Finnish and revise their custom annotation scheme through a combination of automatic and manual processing steps. The resulting corpus consists of nearly 500,000 tokens annotated for over 50,000 mentions categorized into 18 name and numeric entity types. We evaluate this resource and demonstrate its compatibility with the English OntoNotes annotations by training state-of-the-art mono-, bi-, and multilingual deep learning models, finding both that the corpus allows highly accurate tagging at 93% F-score and that a comparable level of performance can be achieved by a bilingual Finnish-English NER model.[1]

1 Introduction

Named Entity Recognition (NER), the identification and typing of text spans referring to entities such as people and organizations in text, is a key task in natural language processing. State of the art NER approaches apply supervised machine learning methods trained on corpora that have been manually annotated for mentions of entity names of interest. While extensive corpora with fine-grained NER annotation have long been available for high-resource languages such as English, NER for many lesser-resourced languages has been limited by smaller, lower-coverage corpora with comparatively coarse annotation.

A degree of language independence has long been a central goal in NER research. One notable example are the CoNLL shared tasks on Language-Independent Named Entity Recognition in 2002 and 2003 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The Spanish, Dutch, English and German datasets introduced in these shared tasks were all annotated for the same types of entity mentions – persons, organizations, locations, and miscellaneous – and the datasets still remain key benchmarks for evaluating NER methods today (e.g. (Devlin et al., 2019)). Nevertheless, until recently most NER methods aimed for language independence only in that they supported training on corpora of more than one language, resulting in multiple separate monolingual models.

In recent years, advances in deep learning have made it possible to create multilingual language models that achieve competitive levels of performance when trained and applied on texts representing more than one language (e.g. Kondratyuk and Straka (2019)). One notable model is the multilingual version of the influential BERT model (Devlin et al., 2019), mBERT, trained on more than 100 languages. mBERT performs well on zero-shot cross-lingual transfer experiments, including NER experiments (Wu and Dredze, 2019). Moon et al. (2019) propose an mBERT-based model trained simultaneously on multiple languages. Training and validating on the OntoNotes v5.0 corpus (see Section 2.3) and the CoNLL datasets, they show that multilingual models outperform models trained on one single language and have cross-lingual zero-shot ability. The zero-shot cross-lingual transfer ability of mBERT also spikes interest in the study of multilingual representations, both on mBERT (Pires et al., 2019; K et al., 2020), and on multilingual encoders in general (Ravishankar et al., 2019; Zhao et al., 2020; Choenni and Shutova, 2020).

[1] The corpus is available under an open license from https://github.com/TurkuNLP/turku-one

Corpus	Language	Tokens	Entities	Domain(s)
OntoNotes	English	2.0M	162K	News, magazines, conversation
FiNER	Finnish	290K	29K	Technology news, Wikipedia
Turku NER	Finnish	200K	11K	News, magazines, blogs, Wikipedia, speech, fiction, etc.

Table 1: Corpus features and statistics. OntoNotes token count only includes sections of the corpus annotated for name mentions. Entity counts include also non-name types such as DATE.

In this paper, we aim to assess and realize the potential benefits from cross- and multi-lingual NER for Finnish, a lesser-resourced language that currently lacks NER resources annotated compatibly with larger similar resources in other languages. Recently, two NER corpora were introduced for Finnish: FiNER (Ruokolainen et al., 2019), focusing on the technology news domain, and the Turku NER corpus (Luoma et al., 2020), covering 10 different text domains. The two corpora are both annotated in the same custom variant of the CoNLL'02 and '03 scheme, making them mutually compatible, but incompatible with resources existing in other languages. This incompatibility has so far made it impossible to directly evaluate the performance of cross- and multi-lingually trained NER methods on manually annotated Finnish resources. To solve this incompatibility issue, we combine and extend these two corpora and adjust the annotations to follow the OntoNotes scheme. The resulting corpus has close to 500,000 tokens annotated for over 50,000 mentions assigned to the 18 OntoNotes name and numeric entity types. We show that our OntoNotes Finnish NER corpus is compatible with the English OntoNotes annotations through training state-of-the-art bi- and multilingual NER models on the combination of these two resources.

2 Data

In the following, we introduce the corpora used in this study, additional text sources for the new corpus, and the pre-trained models used in our experiments. The properties and key statistics of the corpora are presented in Table 1.

2.1 FiNER corpus

FiNER (Ruokolainen et al., 2019) is a Finnish NER corpus consisting mainly of texts from the Finnish technology news source Digitoday, with an additional test set of Wikipedia documents used to assess cross-domain performance of methods trained on the FiNER training section.

FiNER is annotated for mentions of dates (type DATE) and five entity types: person (PER), organization (ORG), location (LOC), product (PRO) and event (EVENT). Of these, PER, ORG and LOC are broadly compatible with the CoNLL types of the same names. The original corpus includes a small number of nested annotations (under 5% of the total) that were excluded in our work.

2.2 Turku NER corpus

The Turku NER corpus (Luoma et al., 2020) is a Finnish NER corpus initially created on the basis of the Universal Dependencies (Nivre et al., 2016) representation of the manually annotated Turku Dependency Treebank (TDT) (Haverinen et al., 2014; Pyysalo et al., 2015), a multi-domain corpus spanning ten different genres.

The Turku NER annotation follows the types and annotation guidelines of the FiNER corpus. An evaluation by Luoma et al. (2020) demonstrated the compatibility of the two Finnish NER corpora by showing that models trained on the simple concatenation of the two corpora outperformed ones trained on either resource in isolation.

2.3 OntoNotes corpus

OntoNotes (Hovy et al., 2006; Weischedel et al., 2013) is a large, multilingual (English, Chinese, and Arabic), multi-genre corpus annotated with several layers covering text structure as well as shallow semantics. In this work, we focus exclusively on the OntoNotes English language NER annotation and refer to this part of the data simply as OntoNotes for brevity. Specifically, we use the NER annotations of the OntoNotes v5.0 release (Weischedel et al., 2013), cast into CoNLL-like format by Pradhan et al. (2013).[2] Sections of the corpus lacking NER annotation (such as the Old and New Testament texts) are excluded.

The OntoNotes NER annotation uses a superset of the ACE entity annotation representation (LDC,

[2] https://github.com/ontonotes/conll-formatted-ontonotes-5.0

Type	Description	Examples
PERSON	People, including fictional	Keijo Virtanen, Obama
NORP	Nationalities or religious or political groups	suomalainen, kristitty
FAC	Buildings, airports, highways, bridges, etc.	Turun linna, LHC
ORG	Companies, agencies, institutions, etc.	Nokia, EU
GPE	Countries, cities, states	Suomi, Venäjä
LOC	Non-GPE locations, mountains, bodies of water	Välimeri, Ararat
PRODUCT	Objects, vehicles, foods, etc. (Not services.)	Oltermanni, iPhone
EVENT	Named hurricanes, wars, sports events, etc.	toinen maailmansota, CES
WORK_OF_ART	Titles of books, songs, etc.	Raamattu, Kid A
LAW	Named documents made into laws	rikoslaki, Obamacare
LANGUAGE	Any named language	suomi, englanti, C++
DATE	Absolute or relative dates or periods	viime vuonna, 1995
TIME	Times smaller than a day	yö, viisi sekuntia
PERCENT	Percentage, including "%"	seitsemän prosenttia, 12%
MONEY	Monetary values, including unit	sata euroa, 500 dollaria
QUANTITY	Measurements, as of weight or distance	kilometri, 5,1 GHz
ORDINAL	"first", "second"	ensimmäinen, 1.
CARDINAL	Numerals that do not fall under another type	yksi, kaksi, 10

Table 2: OntoNotes name annotation types. Adapted from Weischedel et al. (2013).

Model	Language(s)	Vocab. size	Reference
BERT (original)	English	30K	Devlin et al. (2019)
FinBERT	Finnish	50K	Virtanen et al. (2019)
mBERT	104 languages	120K	Devlin et al. (2019)
biBERT	Finnish and English	80K	Chang et al. (2020)

Table 3: Pre-trained models. Cased base variants of all models are used.

2008), applying the 18 types summarized in Table 2. We note that while OntoNotes PERSON, EVENT and DATE largely correspond one-to-one to types annotated in the Finnish NER corpora, the great majority of the types either require a more complex mapping or need to be annotated without support from existing data to create OntoNotes annotation for Finnish.

2.4 Additional texts

During annotation, we noted that the FiNER and Turku NER corpora contained relatively few mentions of laws, which could potentially lead to methods trained on the combined revised corpus performing poorly on the recognition of LAW entity mentions. To address this issue, we augmented the combined texts of the two corpora with a random selection of 60 current acts and decrees of Finnish Acts of Parliament,[3] totaling approximately 24K tokens.

2.5 Pre-trained models

We perform NER tagging experiments by fine-tuning monolingual and multilingual BERT models. Specifically, for monolingual models, we tested English and Finnish (FinBERT) models, and for multilingual models, we tested the mBERT model trained on 104 languages, and a bilingual model trained on only English and Finnish (biBERT). Devlin et al. (2019) trained the original English BERT on the BooksCorpus (Zhu et al., 2015) and English Wikipedia. FinBERT is trained on an internet crawl, news, as well as online forum discussions (Virtanen et al., 2019). The bilingual BERT is trained on English Wikipedia and a reconstructed BooksCorpus, as well as the data used to train FinBERT (Chang et al., 2020). The multilingual BERT is trained on the Wikipedia dump for languages with the largest Wikipedias. The pre-trained models and their key statistics are summarized in Table 3.

We note that while a number of variations and improvements to the pre-training of transformer-

[3] Available from https://finlex.fi/fi/laki/ajantasa/

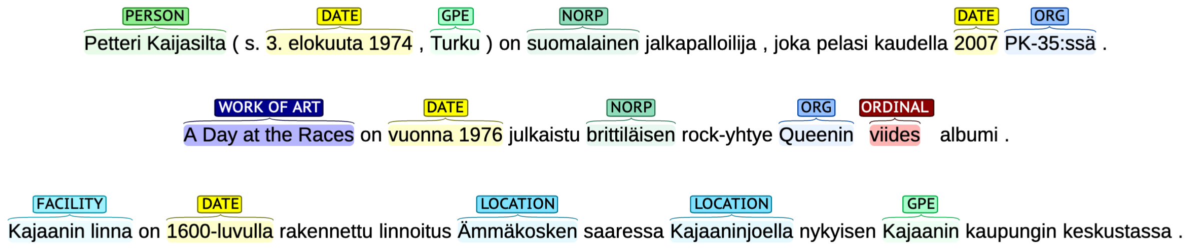
PERSON
DATE
GPE
NORP
DATE
ORG
Petteri Kaijasilta (s. 3. elokuuta 1974 , Turku) on suomalainen jalkapalloilija , joka pelasi kaudella 2007 PK-35:ssä .
WORK OF ART
DATE
NORP
ORG
ORDINAL
A Day at the Races on vuonna 1976 julkaistu brittiläisen rock-yhtye Queenin viides albumi .
FACILITY
DATE
LOCATION
LOCATION
GPE
Kajaanin linna on 1600-luvulla rakennettu linnoitus Ämmäkosken saaressa Kajaaninjoella nykyisen Kajaanin kaupungin keskustassa .

Figure 1: Example annotations

based deep language models have been proposed since the introduction of BERT (e.g. Conneau et al. (2019); Xue et al. (2020)), BERT remains by far the most popular choice for training monolingual deep language models and an important benchmark for evaluating methods for tasks such as NER. As the focus of our evaluation is more on assessing the quality and compatibility of corpora through the application of comparable models rather than optimizing absolute performance, we have here opted to use exclusively BERT models. For the same reason, we only consider BERT base models instead of a mix of base and large models.

3 Annotation

We next summarize the primary steps performed to revise and extend the annotation of the two source corpora to conform with the OntoNotes NER guidelines (Weischedel et al., 2013). Figure 1 shows visualizations of the annotation for selected sentences.

Trivial mappings Of the mentions annotated in the existing Finnish NER corpora, effectively all annotations with the type PER are valid OntoNotes PERSON annotations. Similarly, most EVENT and DATE annotations were valid as-is as OntoNotes annotations of the same names. These annotations were carried over into the initial revised data, changing only the type name when required.

Conditional mappings By contrast to the types allowing trivial mapping from existing to revised annotation, LOC, ORG and PRO required more complex mapping rules. For example, the existing annotations mark both geo-political entities (GPEs) and other locations with the type LOC without distinguishing between the two. To create OntoNotes-compatible annotation, source LOC annotations were mapped to either LOC or GPE annotations on the basis of the annotated text using manually created rules. For example, Suomi/LOC ("Finland") was mapped to Suomi/GPE and Välimeri/LOC ("Mediterranean") to Välimeri/LOC. Similar rules were implemented to distinguish e.g. FAC from ORG and LOC as well as WORK_OF_ART and LAW from PRO.

Dictionary-based tagging Not all mentions in scope of the OntoNotes annotation guidelines are in scope of the FiNER annotation guidelines applied to mark the previously introduced Finnish NER corpora. In addition to most OntoNotes numeric types (see below), in particular nationalities, religious and political groups (NORP in OntoNotes) and languages (LANGUAGE) were not annotated in the source corpora. To create initial OntoNotes annotation for these semi-closed categories of mentions, we performed dictionary-based tagging using lists compiled from sources such as Wikipedia and manually translated OntoNotes English terms tagged with the relevant types.[4]

Numeric types To annotate OntoNotes numeric types (CARDINAL, ORDINAL, etc.) in the Turku NER corpus section of the data, we mapped the manual part-of-speech and feature annotation of the source corpus (TDT) to initial annotations that were then manually revised to identify the more specific types such as PERCENT, QUANTITY and MONEY based on context. For the FiNER texts, annotation for these types followed a similar process with the exception that automatic part-of-speech and feature annotation created by the Turku neural parser (Kanerva et al., 2018) was used as a starting point as no manual syntactic annotation was available for the texts.

Fine-grained tokenization The FiNER annotation guidelines specify that annotated name men-

[4] The accuracy of this initial dictionary-based tagging step was not evaluated separately.

Language	Model	Train data	Development data	Test data
Finnish	FinBERT	Finnish	Finnish	Finnish
Finnish	mBERT	Combined (Fi+En)	Finnish	Finnish
Finnish	biBERT	Combined (Fi+En)	Finnish	Finnish
English	BERT	English	English	English
English	mBERT	Combined (Fi+En)	English	English
English	biBERT	Combined (Fi+En)	English	English

Table 4: Combinations of models, training and evaluation data included in the experiments.

tions must start and end on the boundaries of syntactic words. As hyphenated compound words that include names as part, such as *Suomi-fani* ("fan of Finland"), are comparatively common in Finnish, the FiNER guidelines have a somewhat complex set of rules for the annotation of such compound words (we refer to Ruokolainen et al. (2019) and the relevant guidelines for details). In the revised corpus, we chose to apply a fine-grained tokenization where punctuation characters (including hyphens) are separate tokens, eliminating most of the issues with names as part of hyphenated compounds. To map FiNER-style annotation to the fine-grained version, we wrote a custom tool using regular expressions and manually compiled white- and blacklists of suffixes that can and cannot be dropped from name mention spans.[5]

Semi-automatic and manual revision After initial automatic revisions, a series of semi-automatic and manual revision rounds were performed using the BRAT annotation tool (Stenetorp et al., 2012). In particular, the consistency of mention annotation and typing was checked using the search functionality of the tool[6] and all cases where a string was inconsistently marked or typed were revisited and manually corrected when in error. Additionally, the automatically created pre-annotation for the newly added text (Section 2.4) was revised and corrected in a full, manual annotation pass. All manual revisions of the data were performed by a single annotator familiar with the corpora as well as the FiNER and OntoNotes guidelines. While the single-annotator setting regrettably precludes us from reporting inter-annotator agreement, our monolingual and cross-lingual results below suggest that the consistency of the annotation has not decreased from that of the source corpora.

4 Methods

We next present the applied NER method and detail the experimental setup.

4.1 NER method

We use the BERT-based named entity tagger introduced by Luoma and Pyysalo (2020). In brief, the method is based on adding a simple time-distributed dense layer on top of BERT to predict IOB2 named entity tags in a locally greedy manner. The model is both trained and applied with examples consisting of sentences catenated with their context sentences, resulting in multiple predictions for each token (appearing in both "focus" and context sentences). These predictions are then summarized using majority voting. For brevity, we refer to Luoma and Pyysalo (2020) for further details.[7] Here, we do not use the wrapping of data in documentwise manner as in (Luoma and Pyysalo, 2020), but in bilingual experiments the Finnish and English data are separated with a document boundary token (`-DOCSTART-`) to avoid constructing examples where one input would contain sentences in two languages.

4.2 Experimental setup

The corpora are divided into training, development and test subsets following the subdivisions defined by Pradhan et al. (2013) for OntoNotes, Ruokolainen et al. (2019) for FiNER, and Luoma et al. (2020) for the Turku NER corpus. The newly annotated Finnish law texts are divided chronologically on the document level, placing the earliest-published 48 documents (80%) into training, the latest 6 (10%) into test, and the remaining 6 (10%) into development data. For bilingual experiments, combined training, development and test sets are created by concatenating the corresponding files

[5] The implementation is available from `https://github.com/spyysalo/finer-postprocessing`

[6] `search.py -cm` and `-ct` options.

[7] The implementation is available from `https://github.com/jouniluoma/bert-ner-cmv`

Type	Train	Dev	Test
ORG	11597	866	2345
PRODUCT	5278	246	1237
DATE	4937	412	889
CARDINAL	4668	439	866
PERSON	4635	488	737
GPE	4127	501	674
ORDINAL	1274	107	190
NORP	1252	115	192
MONEY	909	47	169
LAW	749	154	86
LOC	776	54	120
QUANTITY	611	25	145
PERCENT	642	22	101
TIME	455	35	74
EVENT	326	32	37
WORK_OF_ART	305	56	30
LANGUAGE	219	34	28
FAC	173	20	30

Table 5: Corpus annotation statistics

in each corpus, separating the data for the two languages with a document boundary token.

The hyperparameters are selected based on a grid search following the setup in Luoma and Pyysalo (2020) with the exception that batch size 2 is omitted. The reason for this is that the large combined dataset with a small batch size is too time-consuming on the computational resources available. The parameter selection grid is therefore the following:

- Learning rate: 2e-5, 3e-5, 5e-5
- Batch size: 4, 8, 16
- Epochs: 1, 2, 3, 4

The size of the OntoNotes training set is considerably larger than e.g. that of the previously introduced Finnish corpora, and due to resource limitations (especially GPU computation time), we set the BERT maximum sequence length to 128 WordPiece tokens for all of our experiments.

Parameter selection is performed by evaluating on the development subsets of the corpora. The test sets are held out during preliminary experiments and parameter selection, and are only used to evaluate performance in the final experiments. All of the experiments are repeated 5 times, both for hyperparameter selection and the final test results. The reported results are means and standard deviations calculated from these repetitions. The hyperparameters for different final models are selected based on their performance on the target language development set as shown in Table 4.

Lang.	Prec.	Rec.	F-score
Finnish	92.58 (0.18)	93.41 (0.13)	92.99 (0.14)
English	87.92 (0.20)	89.57 (0.25)	88.74 (0.22)

Table 6: Monolingual NER evaluation results (percentages; standard deviation in parentheses)

For testing the zero-shot cross-lingual performance on Finnish, we train the mBERT and biBERT models only on the English OntoNotes data and evaluate performance on the Finnish test set. The hyperparameters providing the best results on the English OntoNotes data are used in these experiments, thus reflecting a setting where no annotated Finnish data is available.

5 Results

We next present summary statistics of the newly introduced corpus and then present the results of the machine learning experiments.

5.1 Corpus statistics

Table 5 summarizes the statistics of the new annotation. The combined, extended corpus with the revised OntoNotes-like annotation contains in total nearly 500,000 tokens of text annotated for approximately 55,000 mentions of names and numeric types. While the corpus represents a substantial increase in size and number of annotations over either of the two previously released Finnish NER corpora, the name-annotated subset of the English OntoNotes corpus remains four times larger in terms of token count and over three times larger in terms of the number of annotated entities (Table 1), motivating our exploration of training bilingual models with combined Finnish and English data.

5.2 Monolingual results

Table 6 summarizes the results of monolingual training and evaluation for the FinBERT model on the newly introduced Finnish NER corpus, with results for the original English BERT model on the English OntoNotes results for reference.

For English OntoNotes, the applied method achieves an F-score of 88.74%, comparable to results for similar implementations reported in the literature: for example, Li et al. (2020) re-

Language	Model	Prec.	Rec.	F-score
Finnish	mBERT	89.81 (0.20)	90.76 (0.22)	90.28 (0.17)
Finnish	biBERT	92.47 (0.22)	93.13 (0.11)	92.80 (0.16)
English	mBERT	88.15 (0.20)	89.62 (0.14)	88.88 (0.16)
English	biBERT	88.57 (0.06)	90.03 (0.11)	89.29 (0.07)

Table 7: Bilingual NER model evaluation results (percentages; standard deviation in parentheses)

	Monolingual			Bilingual		
Type	Prec.	Rec.	F-score	Prec.	Rec.	F-score
PERSON	94.12	97.15	95.60	94.92	96.20	95.55
NORP	94.63	96.15	95.36	97.47	96.15	96.80
FAC	67.83	40.00	50.23	70.10	47.33	56.40
ORG	94.14	94.06	94.10	93.97	93.61	93.79
GPE	95.33	97.36	96.33	94.87	97.06	95.95
LOC	87.12	86.50	86.78	86.11	83.67	84.82
PRODUCT	87.53	88.08	87.81	87.11	88.34	87.72
EVENT	72.17	79.46	75.59	69.46	77.84	73.36
WORK_OF_ART	75.00	77.33	75.97	67.52	79.33	72.84
LAW	90.83	96.74	93.69	91.67	94.65	93.13
LANGUAGE	93.05	95.00	94.01	94.95	93.57	94.25
DATE	94.70	94.78	94.74	94.98	95.32	95.15
TIME	81.70	84.32	82.98	78.01	81.35	79.64
PERCENT	95.60	98.61	97.08	100.00	100.00	100.00
MONEY	95.36	94.79	95.08	95.80	91.60	93.65
QUANTITY	87.18	90.90	89.00	86.61	90.07	88.30
ORDINAL	90.33	91.37	90.84	89.56	90.21	89.88
CARDINAL	94.01	95.36	94.68	93.54	95.64	94.58

Table 8: Result details for Finnish data in monolingual setting using FinBERT and bilingual setting using biBERT (percentages)

port 89.16% F-score for *BERT-Tagger* on English OntoNotes 5.0; an approx. 0.4% point difference. While more involved state-of-the-art methods building on BERT have been reported to outperform this result (e.g. 91.11% F-score for the BERT-MRC method of Li et al. (2020)), we are satisfied that the implementation used here is broadly representative of BERT used for NER in a standard sequence tagging setting.

For Finnish, we note that Luoma and Pyysalo (2020) performed an evaluation of the combination of the FiNER and Turku NER corpora with the comparatively coarse-grained six FiNER corpus NE types, reporting an F-score of 93.66% on the combined test set. While not perfectly comparable, the training and evaluation texts of that experiment are strict subsets of the Finnish training and evaluation data here, and we find the F-score of 92.99% on the 18 fine-grained OntoNotes-like annotation a very positive sign of its quality and consistency: using the newly introduced dataset, we can train models to recognize mentions of *three times as many* name and numeric entity types as previously with only a modest decrease in overall tagging performance.

5.3 Bilingual results

Table 7 summarizes the results of the bi- and multilingual models trained on the combined Finnish and English data and evaluated on the two monolingual corpora. We first observe that the bilingual biBERT model achieves better results that the multilingual mBERT model, providing further support for the findings of Chang et al. (2020) indicating that multilingual training processes produce notably better models when only two languages are targeted. In the remaining, we focus on the results for the biBERT model. For Finnish, we find that the bilingual model fine-tuned on the combined bilingual training data falls just 0.2%

Language	Model	Prec.	Rec.	F-score
Finnish	mBERT	71.00 (0.81)	69.99 (0.47)	70.49 (0.50)
Finnish	biBERT	77.01 (0.47)	77.01 (0.46)	77.01 (0.19)

Table 9: Zero-shot cross-lingual evaluation results from English to Finnish (percentages; standard deviation in parentheses)

points in F-score below the monolingual FinBERT model fine-tuned with monolingual data. For English, we unexpectedly find that the bilingually trained model *outperforms* the monolingual English model with an approx. 0.5% point absolute difference. These results indicate that the annotations of the English OntoNotes NER dataset and the newly introduced Finnish NER dataset are highly compatible, allowing bi- or multilingual methods trained on a bilingual dataset created by their simple concatenation to perform competitively with or even potentially outperform monolingual NER models.

The detailed results presented in Table 8 further show that the performance of the monolingual and bilingual models track very closely, with the monolingual Finnish model slightly outperforming the bilingual for most mention types. An exception to this pattern is seen for `NORP`, `FAC`, `LANGUAGE`, `DATE` and `PERCENT`, where the bilingual model shows better performance. These results further suggest that there are no notable annotation inconsistencies in individual types, and that multilingual training may still hold benefit for some entity types.

5.4 Zero-shot cross-lingual results

Finally, Table 9 provides the results of zero-shot cross-lingual transfer from English to Finnish, where a bi- or multilingual model is trained exclusively on English data but then evaluated on Finnish data. We again find that the biBERT model considerably outperforms the mBERT model. While the model performance at 77% falls far behind the over 90% F-scores achieved by the monolingual and bilingual models, it is nevertheless interesting to note that this level of performance can be achieved without any target language data. This cross-lingual transfer approach could potentially be applied e.g. to bootstrap initial annotations for manual revision when creating named entity annotation for languages lacking a corpus annotated with OntoNotes types.

6 Discussion and conclusions

We have introduced a new corpus for Finnish NER created by combining and extending two previously released corpora, FiNER and the Turku NER corpus, and by mapping their custom annotations into the fine-grained OntoNotes representation through a combination of automatic and manual processing steps. The resulting corpus consists of over 50,000 annotations for nearly 500,000 tokens of text representing a broad selection of genres, topics and text types, and is not only the largest resource for Finnish NER created to date, but also identifies three times as many distinct name and numeric entity mention types as the previously introduced Finnish NER corpora.

To assess the internal consistency of the newly created annotation and to provide a baseline for further experiments on the data, we evaluated the performance of a BERT-based NER system initialized with the FinBERT model and fine-tuned on the new Finnish data. These experiments indicated that the annotations of the new corpus can be automatically recognized at nearly 93% F-score, effectively matching previous results with much coarser-grained entity types. To further assess the compatibility of the newly introduced annotation with the original English OntoNotes corpus v5.0 name annotation, we fine-tuned bi- and multi-lingual BERT models on the combination of the Finnish and English corpora, finding that bilingual models can effectively match or potentially even outperform monolingual ones, thus confirming the compatibility of the newly created annotation with existing OntoNotes resources.

All resources introduced in the paper are available under open licenses from `https://github.com/TurkuNLP/turku-one`

Acknowledgments

This work was funded in part by the Academy of Finland. We wish to thank CSC – IT Center for Science, Finland, for computational resources.

References

Li-Hsin Chang, Sampo Pyysalo, Jenna Kanerva, and Filip Ginter. 2020. Towards fully bilingual deep language modeling. *arXiv preprint arXiv:2010.11639*.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. *arXiv preprint arXiv:2009.12862*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 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*, pages 133–142.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795.

LDC. 2008. Ace english annotation guidelines for entities. Technical report, Technical report, Linguistic Data Consortium.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 904–914.

Taesun Moon, Parul Awasthy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with bert. *arXiv preprint arXiv:1912.01389*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–172.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Erik F. Tjong Kim Sang. 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)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing language-agnostic multilingual representations. *arXiv preprint arXiv:2008.09112*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.