DANSK and DaCy 2.6.0: Domain generalization of Danish named entity recognition

Anonymous ACL submission

Abstract

Named entity recognition is one of the cornerstones of Danish NLP, useful for providing insights within both industry and research. However, the field is inhibited by a lack of available datasets. As a consequence, no models are capable of fine-grained named entity recognition, nor have they been evaluated for potential generalizability issues across datasets and domains. To alleviate these limitations, this paper introduces: 1) DANSK: a named entity dataset providing for high-granularity tagging as well as within-domain evaluation of models across a diverse set of domains; 2) DaCy 2.6.0 that includes three generalizable models with fine-grained annotation; and 3) an evaluation of current state-of-the-art models' ability to generalize across domains. The evaluation of existing and new models revealed notable performance discrepancies across domains, which should be addressed within the field. Shortcomings of the annotation quality of the dataset and its impact on model training and evaluation are also discussed. Despite these limitation, we advocate for the use of the new dataset DANSK alongside further work on the generalizability within Danish NER.

1 Introduction

Danish **A**nnotations for **N**LP **S**pecific Tas**K**s (**DANSK**) version 0.0.1. is a new gold-standard dataset for Danish with named entity annotations for 18 distinct classes. The annotated texts are from 25 text sources that span 7 different domains and have been derived from the Danish Gigaword Corpus (Strømberg-Derczynski et al.). The dataset is publicly accessible[1] and pre-partitioned into a training, validation, and testing set in order to standardize future model evaluations.

The release of DANSK is motivated by present limitations facing Danish NER. The first limitation concerns a lack of generalizability measures of current SOTA models: all have been either fully or partially fine-tuned for the NER task on a single dataset, Danish Named Entities (DaNE) (Hvingelby et al.). Although DaNE features high-quality NER annotations and features texts from a wide array of domains and sources, it has several shortcomings. First, domains such as social media and legal texts are lacking from DaNE entirely and spoken language is severely underrepresented. Moreover, since the texts are from 1883-1992, no contemporary linguistic trends are included. While current Danish models perform quite well on DaNE (Nielsen), their performances is naturally an expression of performance on the texts that are included.

Second, individual domain evaluation is not possible even for domains included in the dataset, as DaNE lacks metadata on the origin of the texts. Information on domain biases is therefore occluded in any evaluations. This is especially problematic because many models' current use cases are on texts that are not represented in DaNE; e.g. on social media data.

Third, DaNE contrains models to the CoNLL-2003 annotation standard consisting of four types, as opposed to more fine-grained NER datasets like OntoNotes 5.0 with 18 entity types.

Danish NLP is in need of more open and free datasets, in part to navigate impediments to generalizability (Kirkedal et al.). Domain shifts in data cause drops in performance, as models are optimized for the training and validation data, making cross-domain evaluation, particularly for tasks like NER, crucial (Plank et al.). A study by Enevoldsen et al., furthermore found generalizability issues for NER in Danish, not across domains, but across different types of data augmentations — further indicating generalizability issues for Danish models.

The DANSK dataset was designed to address these limitations currently facing Danish NER. Based on DANSK, we also introduce three new models of varying sizes incorporated into DaCy

[1] https://anonymous.4open.science/r/dansk-3A03

	Average Cohen's κ
Annotator 1	0.6
Annotator 2	0.52
Annotator 3	0.51
Annotator 4	0.58
Annotator 5	0.54
Annotator 6	0.56
Annotator 7	0.47
Annotator 8	0.51
Annotator 9	0.52
Annotator 10	0.56

Table 1: Table showing the average Cohen's κ scores for each rater for the overlapping data.

(Enevoldsen et al.) that are specifically developed for fine-grained NER on the comprehensive array of domains included in DANSK to ensure generalizability.

Finally, we evaluate the three newly released DaCY models against some of the currently best-performing and most widely-used NLP models within Danish NER using the DANSK dataset, in order to attain estimates of generalizability across domains.

2 Dataset

2.1 Initiatial annotation

The texts in the DANSK dataset were sampled from the Danish Gigaword Corpus (DAGW) (Strømberg-Derczynski et al.), and filtered to exclude texts from prior to 2000 and segmented into sentences. DANSK dataset utilized the annotation standard of OntoNotes 5.0. For NER annotation using Prodigy, texts were first divided up equally for the 10 annotators, with a 10% overlap between the assigned texts. The annotators were ten native speakers of Danish (nine female, one male) between the ages of 22-30 years old, studying in the Masters degree program in English Linguistics at Aarhus University. Instructions provided to the annotators followed the 18 shorthand descriptions of the OntoNotes 5.0 named entity types (Weischedel et al.). Initial annotations suffered from extremely poor intercoder reliability, as measured by Cohen's kappa (κ) scores, calculated by matching each rater pairwise to every other (Table 1). In order to assess the annotation consensus between annotators on a entity type level, additional F1-mean scores were calculated for all annotators (Table 2).

Named-entity type	Mean F1-score	F1 SD
CARDINAL	0.47	0.23
DATE	0.55	0.21
EVENT	0.5	0.34
FACILITY	0.22	0.38
GPE	0.91	0.05
LANGUAGE	0.0	0.0
LAW	0.23	0.32
LOCATION	0.22	0.24
MONEY	0.62	0.49
NORP	0.5	0.39
ORDINAL	0.5	0.27
ORGANIZATION	0.72	0.14
PERCENT	0.0	0.0
PERSON	0.59	0.32
PRODUCT	0.12	0.23
QUANTITY	0.18	0.26
TIME	0.33	0.36
WORK OF ART	0.4	0.29

Table 2: The mean and standard deviation of the F1-scores across the raters for each of the named entity types.

2.2 Annotation improvement

Due to the low consensus between annotators, it was deemed necessary for the annotated texts to undergo additional processing before they could be unified into a coherent, high-quality dataset.

Texts with multiple annotators Some curated datasets utilize a single annotator for manual resolvement of conflicts between raters (Weischedel et al.), however this skews the annotations towards the opinion of a single annotator, rather than the general consensus across raters. In order to resolve conflicts while diminishing this skew, an automated procedure was employed.

The procedure was rule-based and followed a decision tree-like structure (Figure 1). It was only applied to texts that had been annotated by a minimum of four raters, ensuring that that an annotation with no consensus was accepted in a text annotated by two annotators. To exemplify the streamlining of the multi-annotated texts, Figure 2 is included.

After employing the automated procedure, the 886 multi-annotated texts went from having 513 (58%) texts with complete rater agreement to 789 (89%). The texts with complete agreement were added to the DANSK dataset, while the remaining 97 (21%) of the multi-annotated texts had remain-

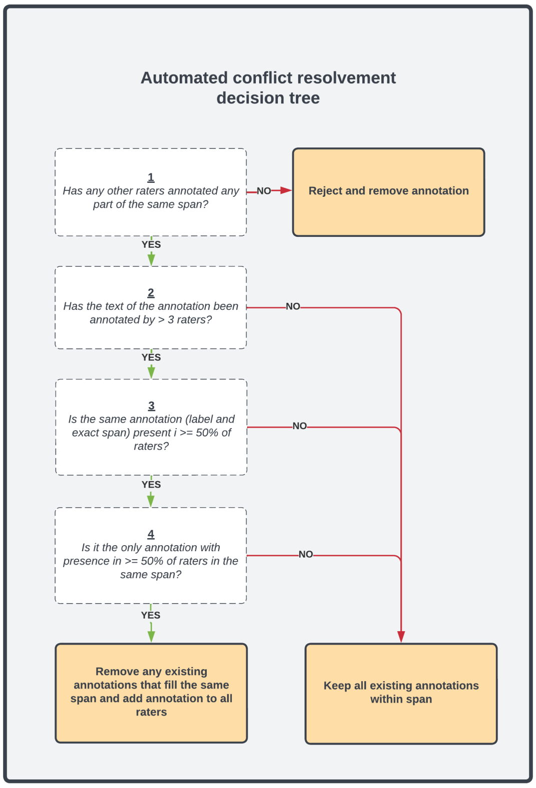

Figure 1: The decision tree for automated conflict resolvement of multi-annotated texts. Each annotation span in a text followed the steps from 1 to 4 on the diagram. The decision tree was only followed for annotation spans found in texts that had been annotated by at least four raters.

ing annotation conflicts. The remaining texts with conflicting annotations were resolved manually by the first author, by changing any annotations that did not comply with the extended OntoNotes annotation guidelines. However, three texts were of such bad quality that they were rejected and excluded. The remaining resolved 94 texts were then added to DANSK.

	Initial annotation	Streamlined annotation
Rater 1	[Mette F.] (PER) er statsminister i [DK] (GPE)	[Mette F.] (PER) er statsminister i [DK] (GPE)
Rater 3	[Mette F.] (PER) er statsminister i [DK] (GPE)	[Mette F.] (PER) er statsminister i [DK] (GPE)
Rater 5	[Mette] (PER) F. er statsminister i DK	[Mette] (PER) F. er statsminister i [DK] (GPE)
Rater 9	[Mette] (PER) F. er [statsminister] (PER) i DK	[Mette] (PER) F. er statsminister i [DK] (GPE)

Figure 2: An example of a text along with its four annotations being processed on the basis of the decision-tree in Figure 1.

Finally, to ensure that any named entities of the type LANGUAGE, PERCENT, and PRODUCT had not been missed by the annotators, an extra measure was taken. The model `TNER/Roberta-Large-OntoNotes5`[2] was used to add these types of annotations to the accepted multi-annotated texts (Ushio and Camacho-Collados). Each text with any predictions by the models was then manually assessed by the first author, to inspect whether the additional model annotations should be included. None of the predictions matched the annotation guidelines and were thus not added to the texts. This step concluded the processing of the multi-annotated texts, which resulted in a total of 883 texts added to the DANSK dataset.

Texts with a single annotator Based on the poor quality and low consensus between multiple raters, it was assumed that the single-annotator texts also suffered from limitations. To refine these annotations, we utilize the existing DANSK annotations to train a model and then manually resolve the discrepancies. The rationale for this process is that it propagates the aggregated annotations across the dataset and can thus be seen as a supervised approach to anomaly detection.

As the preliminary DANSK dataset included relatively few annotations, we explored the effect of enriching our existing datasets using the English subsection of OntoNotes 5.0 (Recchia and Jones). We trained a total of three models using the first 80% of the preliminary DANSK dataset, the second additionally adding English OntoNotes 5.0 and the third duplicating the 80% of the preliminary DANSK to match the size of the English OntoNotes 5.0. For our model we used the multilingual `xlm-roberta-large`[3] to allow for cross-lingual transfer (Conneau et al.). The models was validated on the remaining 20% of the DANSK dataset. The best model (the third) was then applied to the remaining 15062 texts and discrepancies were manually resolved by the first author.

Resolving remaining inconsistencies Because of the large number of annotation reviews, we were able to identify common annotation mistakes. To further enhance the quality of the annotations, all texts were screened for common errors using a list of regex patterns. This resulted in flagged matches in 449 texts which were re-annotated in accordance with the OntoNotes 5.0 extended annotation guidelines (Weischedel et al.) and the newly developed

[2] https://huggingface.co/tner/roberta-large-ontonotes5
[3] https://huggingface.co/xlm-roberta-large

Danish Addendum designed to clarify ambiguities and issues specific to Danish texts, as described in the dataset card (Appendix A).

3 Final dataset: DANSK

3.1 DANSK quality assessment

Finally, upon finalizing the dataset, the quality of DANSK was assessed.

	Average Cohen's κ
Annotator 1	0.92
Annotator 3	0.93
Annotator 4	0.93
Annotator 5	0.91
Annotator 6	0.93
Annotator 7	0.93
Annotator 8	0.89
Annotator 9	0.92
Preliminary DANSK	0.92

Table 3: Table showing the average Cohen's κ scores for each of the non-discarded raters for the overlapping data after the automated streamlining process.

Average Cohen's κ scores were calculated on the processed, finalized versions of texts with multiple annotators. All of the non-removed raters' texts were included, as well as the preliminary version of DANSK with the conflicts resolved. As expected, the average scores of the processed texts saw a great increase, ultimately ranging between 0.93 and 0.89, compared with scores of the original annotated texts which ranged from 0.47 to 0.60 (Table 1 and Table 3).

To assess which inconsistencies still remained between the DANSK dataset and the raters' annotations, a confusion matrix between the annotations of DANSK and the accumulated annotations of the processed rater texts was assessed. As can be seen in Figure 3, the majority of differences are cases in which a token or a span of tokens was considered a named entity by one party, but not by the other. In other words, no unequivocal systematic patterns between a pair of named entities existed.

3.2 DANSK descriptive statistics

To provide complete transparency about the dataset distributions, descriptive statistics are reported in the dataset card in Appendix A with regard to source, domain, and named entities.

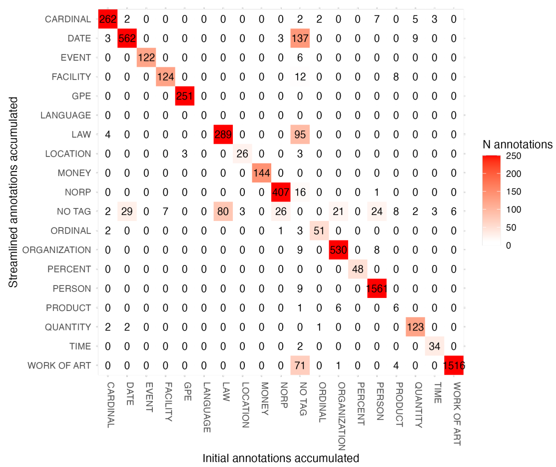

Figure 3: Confusion matrix across the annotations before and after the automated streamlining.

4 DaCy model curation

4.0.1 Model Specifications

In order to contribute to Danish NLP with both fine-grained tagging as well as non-domain specific performance, three new models were fine-tuned to the newly developed DANSK dataset. The three models differed in size and included a large, medium, and small model as they were fine-tuned versions of `dfm-encoder-large-v1`[4], `DanskBERT`[5] and `electra-small-nordic`[6] (Snæbjarnarson et al., 2023). These models contain 355, 278, and 22 million trainable parameters, respectively. They were chosen based on their ranking among the best-performing Danish language models within their size class, according to the ScandEval benchmark scores current as of the 7th of March, 2023 (Nielsen).

The models were all fine-tuned on the training partition of the DANSK dataset using *Python*, *Jupyter*, and the Python package *spaCy 3.5.0* (Honnibal et al.; Van Rossum and Drake Jr, 1995). The fine-tuning was performed on an NVIDIA T4 GPU through the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark. To get an overview of the training procedure, some of the hyperparameter settings are listed in this section. For brevity, the impact and nature of these settings will not be

[4]https://huggingface.co/chcaa/dfm-encoder-large-v1
[5]https://huggingface.co/vesteinn/DanskBERT
[6]https://huggingface.co/jonfd/electra-small-nordic

	DaCy fine-grained model		
	Large	**Medium**	**Small**
F1-score	**0.823**	*0.806*	0.776
Recall	**0.834**	*0.818*	0.77
Precision	**0.813**	*0.794*	0.781

Table 4: Table reporting the overall DaCy fine-grained model performances in macro F1-scores. Bold and italics are used to represent the best and second-best scores, respectively.

explicated. An exhaustive list of all configurations of the system as well as hyperparameter settings is provided in the GitHub repository [7].

The three models shared the same hyperparameter settings for the training with the exception that the large model utilized an accumulate gradient of 3. They employed a batch size of 2048 and applied Adam as the optimizer with $\beta1$ = 0.9 and $\beta2$ = 0.999 and an initial learning rate of 0.0005. It used L2 normalization with weighted decay, α = 0.01, and gradient clipping with c-parameter = 1.0. For the NER head of the transformer we used a transition-based parser with a hidden width of 64. The models were trained for 20 000 steps with an early stopping patience of 1600. During training the model had a dropout rate of 0.1 and an initial learning rate of 0.0005.

For the progression of the training loss of the NER head, loss of the transformer, NER performance measured in recall, precision, and F1-score, refer to Figure 7 in Appendix B.

4.1 Results

This section presents the results of the performance evaluation. A crude overview of the general performance of the three fine-grained models is reported in Table 4. Domain-level performance can be seen in Table 6. To account for the differences in domain size, Figure 4 is further included as it adds an additional dimension of information through the depiction of the size of the domains. Insights into performance within named entity categories are provided in Table 5.

For full information on distributions for named entities and domains within the partitions, refer to Appendix A.

[7]https://anonymous.4open.science/r/DaCy-1BAF

	DaCy Fine-grained NER		
Named-entity type	**Large**	**Medium**	**Small**
CARDINAL	*0.87*	0.78	**0.89**
DATE	0.85	*0.86*	**0.87**
EVENT	**0.61**	*0.57*	0.4
FACILITY	**0.55**	*0.53*	0.47
GPE	**0.89**	*0.84*	0.80
LANGUAGE	**0.90**	*0.49*	0.19
LAW	**0.69**	*0.63*	0.61
LOCATION	*0.63*	**0.74**	0.58
MONEY	*0.99*	**1**	0.94
NORP	0.78	**0.89**	*0.79*
ORDINAL	0.70	*0.7*	**0.73**
ORGANIZATION	**0.86**	*0.85*	0.78
PERCENT	*0.92*	**0.96**	**0.96**
PERSON	*0.87*	**0.87**	0.83
PRODUCT	**0.67**	*0.64*	0.53
QUANTITY	0.39	*0.65*	**0.71**
TIME	*0.64*	0.57	**0.71**
WORK OF ART	*0.49*	**0.64**	0.49
AVERAGE	**0.82**	*0.81*	0.78

Table 5: Table reporting the DaCy fine-grained model performances in F1-scores within each named entity type. Bold and italics are used to represent the best and second-best scores, respectively.

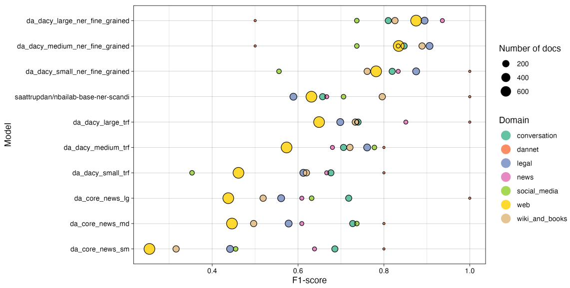

Figure 4: Figure displaying the domain performance in macro F1-scores of the three models on the test partition of DANSK. The size of the circles represents the size of the domains, and thus their relative weighted impact on the overall scores. See Appendix A for scores.

	DaCy fine-grained model		
Domain	**Large**	**Medium**	**Small**
All domains combined	**0.82**	*0.81*	0.78
Conversation	*0.80*	0.72	**0.82**
Dannet	*0.75*	0.667	**1**
Legal	*0.85*	*0.85*	**0.87**
News	*0.84*	0.76	**0.86**
Social Media	0.79	**0.85**	*0.8*
Web	**0.83**	*0.80*	0.76
Wiki and Books	*0.78*	**0.84**	0.71

Table 6: Table reporting the DaCy fine-grained model performances in macro F1-scores within each domain. Bold and italics are used to represent the best and second-best scores, respectively.

5 State-of-the-art model generalizability

5.1 Methods

5.1.1 Models

To assess whether there exists a generalizability issue for Danish language models, a number of SOTA models were chosen for evaluation on the test partition of the newly developed DANSK dataset. The field of Danish NLP and NER is evolving rapidly, making it hard to establish an overview of the most important models for Danish NER. However, the models for the evaluation were chosen on the basis of two factors; namely prominence of use, and performance. The latter was gauged on the basis of ScandEval, the NLU framework for benchmarking (Nielsen).

At the time of the model search, the model `saattrupdan/nbailab-base-ner-scandi` [8] ranked amongst the best-performing models for Danish (and scandinavian) NER.[9] It was trained on the combined dataset of DaNE, NorNE, SUC 3.0, and the Icelandic and Faroese part of the WikiANN (Hvingelby et al., 2020; Gustafson-Capková and Hartmann; Ejerhed et al.; Jørgensen et al.; Pan et al.). Because of the wide palette of different datasets, texts from more domains are represented. It was thus conjectured that the model might not suffer from the generalizability issues outlined in the introduction section of the paper.

Apart from this model, the three v0.1.0 DaCy models large, medium, and small were also included. Note that these are the existing non-fine-grained models that were already in DaCy prior to the development of the fine-grained DaCy models presented in this paper. The models are fine-tuned versions of 1) Danish Ælæctra[10], Danish BERT[11], and the XLM-R (Conneau et al.). The model are fine-tuned on DaNE (Hvingelby et al., 2020) and DDT (Johannsen et al., 2015) for multitask prediction for multiple task including named-entity recognition and at the time of publication achieved state-of-the-art performance for Danish NER (Enevoldsen et al.).

We also include the NLP framework *spaCy* (Explosion AI, Berlin, Germany), to explore the generalization of production systems. SpaCy features three Danish models (small, medium, and large[12]) which similarly to the DaCy models are multi-task models with NER capabilities. Although spaCy also includes a Danish transformer model, it was not incorporated in the generalizability analysis. The reason for this is that DaCy medium v.0.1.0 is already included and the two models are almost identical. Both are based on the model `Maltehb/danish-bert-botxo`[13] and fine-tuned on DaNE, and thus only deviate on minor differences in hyperparameter settings.

In summary, the models included in the final evaluation were:

1. `Base-ner-scandi (nbailab-base-ner-scandi)`
2. `DaCy large (da_dacy_large_trf-0.1.0)`
3. `DaCy medium (da_dacy_medium_trf-0.1.0)`
4. `DaCy small (da_dacy_small_trf-0.1.0)`
5. `spaCy large (da_core_news_lg v. 3.5.0)`
6. `spaCy medium (da_core_news_md v. 3.5.0)`
7. `spaCy small (da_core_news_sm v. 3.5.0)`

5.1.2 Named Entity Label Transfer

A fine-grained NER dataset with 18 labels following the OntoNotes guidelines has not been publicly available for Danish until now. The aforementioned models have thus naturally only been fine-tuned to the classic, more coarse-grained DaNE dataset that follows the CoNLL-2003 named entity annotation scheme (Sang and De Meulder; Hvingelby et al.). This includes the four named entity types PER (person), LOC (location), ORG (organization), and MISC (miscellaneous). This annotation scheme is radically different from the DANSK annotations that match the OntoNotes 5.0 standards. To enable an evaluation of the models, the DANSK named entity labels were coerced into the CoNLL-2003 standard in order to match the nature of the models.

As the description of both ORG and PER in CoNLL-2003 largely matches that of the extended OntoNotes, these named entity types could be used in the evaluation with a 1-to-1 mapping without further handling. However, in CoNLL-2003, LOC includes cities, roads, mountains, abstract places, specific buildings, and meeting points (Hvingelby et al.; Sang and De Meulder). As the extended

[8] https://huggingface.co/saattrupdan/nbailab-base-ner-scandi

[9] https://paperswithcode.com/sota/named-entity-recognition-on-dane

[10] https://huggingface.co/Maltehb/aelaectra-danish-electra-small-cased

[11] https://huggingface.co/Maltehb/danish-bert-botxo

[12] Note that a model size of spaCy are not comparable to model sizes of transformer encoders

[13] https://huggingface.co/Maltehb/danish-bert-botxo

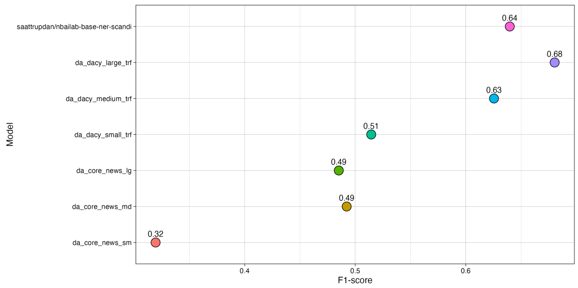

Figure 5: Figure displaying the domain performance in macro F1-scores of the on the test partition of DANSK. The size of the circles represents the size of the domains, and thus their relative weighted impact on the overall scores.

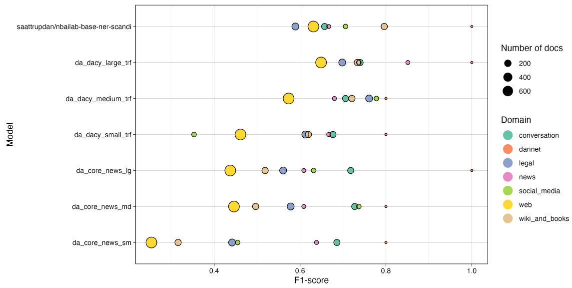

Figure 6: Figure displaying the domain performance in macro F1-scores of the on the test partition of DANSK. The size of the circles represents the size of the domains, and thus their relative weighted impact on the overall scores.

OntoNotes guidelines use both GPE and LOCATION, DANSK GPE annotations were mapped to LOC in an attempt to make the test more accurate. Predictions for the CoNLL-2003 MISC category, intended for names not captured by other categories (e.g. events and adjectives such as "*2004 World Cup*" and "*Italian*"), were excluded.

5.1.3 Evaluation

SOTA models were evaluated using macro average F1-statistics at a general level, a domain level, and finally F1-scores at the level of named entity types.

5.2 Results

A quick overview of the F1-scores can be inspected in Figure 5, while Table 7 elaborates with recall and precision statistics. The performance across domains and across named entity types are reported in Table 8 and Table 9. Finally, Figure 6 is included, in an attempt to provide an easily readable overview of the domain scores.

Model	F1	Recall	Precision
Base-ner-scandi	*0.64*	0.59	**0.70**
DaCy large (0.1.0)	**0.68**	**0.67**	*0.69*
DaCy medium (0.1.0)	0.63	*0.64*	0.61
DaCy small (0.1.0)	0.51	0.48	0.56
spaCy large (3.5.0)	0.49	0.45	0.53
spaCy medium (3.5.0)	0.49	0.47	0.52
spaCy small (3.5.0)	0.32	0.32	0.32

Table 7: Table showing the overall performance in macro F1-scores on the DANSK test set. Bold and italic represent the best and next best scores.

Model	Across	Convo	Dannet	Legal	News	SoMe	Web	Wiki
base-ner-scandi	*0.64*	0.66	**1**	0.59	*0.67*	0.71	*0.63*	**0.80**
DaCy Large (0.1.0)	**0.68**	**0.74**	**1**	*0.70*	**0.85**	*0.74*	**0.65**	*0.73*
DaCy Medium (0.1.0)	0.63	0.71	*0.8*	**0.76**	0.68	**0.78**	0.57	0.72
DaCy Small (0.1.0)	0.51	0.68	*0.8*	0.61	*0.67*	0.35	0.46	0.62
spaCy Large (3.5.0)	0.49	0.72	**1**	0.56	0.61	0.63	0.44	0.52
spaCy Medium (3.5.0)	0.49	*0.73*	*0.8*	0.58	0.61	*0.74*	0.45	0.50
spaCy small (3.5.0)	0.32	0.69	*0.8*	0.44	0.64	0.46	0.25	0.32

Table 8: Table showing the domain performances in macro F1-scores of the models on the DANSK test set. Bold and italic represent the best and next best scores.

Model	LOC	ORG	PERSON
Base-ner-scandi	*0.79*	0.46	0.70
DaCy large (0.1.0)	**0.84**	**0.50**	**0.74**
DaCy medium (0.1.0)	0.74	*0.48*	*0.70*
DaCy small (0.1.0)	0.67	0.38	0.52
spaCy large (3.5.0)	0.63	0.28	0.61
spaCy medium (3.5.0)	0.65	0.31	0.58
spaCy small (3.5.0)	0.44	0.23	0.31

Table 9: Table showing the performance in F1-scores within each of the named entity classes on the DANSK test set. Bold and italic represent the best and next best scores.

6 Discussion

6.1 DANSK dataset

The DANSK dataset enhances Danish NER by focusing on fine-grained named entity labels and diverse domains like conversations, legal matters, and web sources, but omits some domains in DaNE, such as magazines (Norling-Christensen; Hvingelby et al.). Entity distribution varies, influencing model performance for specific types.

DANSK's quality was benchmarked using models trained on different OntoNotes 5.0 annotated datasets (Luoma et al.). Despite the dataset size disparity, performances for English and Finnish models were between F1-scores of .89 and .93 (Luoma et al.; Li et al.), notably higher than DANSK. Given the smaller size of DANSK (15062 texts) compared to English OntoNotes (600000 texts) (Weischedel et al.), performance for models trained on DANSK is expectedly lower, irrespective of annotation quality (Russakovsky et al.).

Annotation quality issues were tackled, improving Cohen's κ values from ~0.5 to ~0.9 (Table 1 and Table 3). Initial difficulties arose from suboptimal sampling from DAGW and insufficient annotator training. Future improvements include initial quality screening and comprehensive training with the OntoNotes 5.0 annotation scheme (Plank; Uma et al.). In the release of the DANSK dataset, we include raw (per annotator) annotations to allow for transparency and further analysis of annotator disagreement.

6.2 DaCy models

New fine-grained models of varying sizes attained macro F1-scores of 0.82, 0.81, and 0.78 respectively. Larger models generally performed better as would be expected. However, due to DANSK's domain imbalance, these scores should be treated carefully. Domains like web, conversation, and legal heavily influenced the F1-scores due to their larger text volume. Performance comparisons are based on OntoNotes 5.0 standard datasets due to the unique annotation scheme of DANSK.

Minor performance variation was found within each domain. The small models excelled in underrepresented domains like news, possibly leading to volatile results. Legal texts were easiest to classify with F1-scores of 0.85 and 0.87.

Classification performance varied with named entity types. Facilities, artworks, and quantities were difficult to predict, whereas entities like money, dates, percentages, GPEs, organizations, and cardinals were easier to classify. This can be attributed to the quantity and context of named entities in the training data. Some entity types might appear in similar contexts or have similar structures, hence easier to distinguish. Variance in performance may arise from differences in text quality and context. Given the observed performance differences across domains and named entity types, it's crucial to understand the strengths and limitations of the new models within the DaCy framework.

6.3 SOTA models and generalizability

The new fine-grained DaCy models demonstrate higher performance on the DANSK dataset, compared to existing SOTA models (refer to Tables 7 and 4). However, due to annotation scheme discrepancies, a direct comparison is challenging.

Performance analysis is two-fold: evaluation across domains for each model, and comparison between models, both following the CoNLL-2003 annotation scheme.

Significant domain performance disparities were observed (see Table 8 and Figure 6). For instance, `base-ner-scandi` scored F1-scores of 0.59 and 0.8 for legal and Wikipedia texts, respectively. Actual model accuracy may vary by domain, contrary to performance reported on DaNE. The models performed best on conversation and news texts, with web and wiki sources performing poorly.

Larger models generally outperformed, with `base-ner-scandi` and `DaCy large` scoring 0.68 and 0.64 F1-scores respectively. Smaller spaCy models underperformed, suggesting their usage for news or conversation texts. The DaCy models, easily accessible via the DaCy framework, performed comparably or better than the `base-ner-scandi` model, hence DaCy is the preferred library for Danish NER.

Despite the insights, the evaluation is hampered by the chosen models, annotation scheme differences, and DANSK dataset quality. Thus, the findings primarily highlight generalizability issues and the impact of annotation schemes.

7 Conclusion

Danish NER suffers from limited dataset availability, lack of cross-validation, domain-specific evaluations, and fine-grained NER annotations. This paper introduces DANSK, a high-granularity named entity dataset for within-domain evaluation, DaCy 2.6.0 with three generalizable, fine-grained models, and an evaluation of contemporary Danish models. DANSK, annotated following OntoNotes 5.0 and including metadata on text origin, facilitates across-domain evaluations but still falls short of quality standards of other languages' datasets. DaCy models, trained on DANSK, achieve up to 0.82 macro F1-score, offering NER on 18 categories, although their performance is slightly lower than models for other languages. Performance discrepancies exist between domains in current Danish models, exemplified by `base-ner-scandi`, scoring 0.8 F1-score on Wikipedia texts but dropping to 0.59 on legal texts. While work remains to be done to augment the size and quality of fine-gained named entity annotation in Danish, the release of DANSK and DaCy will assist in addressing generalizability issues in the field.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. The linguistic annotation system of the stockholm-umeå project.

Kenneth Enevoldsen, Lasse Hansen, and Kristoffer Nielbo. DaCy: A unified framework for danish NLP.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sofia Gustafson-Capková and Britt Hartmann. Manual of the stockholm umeå corpus version 2.0. pages 5–85.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in python. Publisher: Zenodo, Honolulu, HI, USA.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. Dane: A named entity resource for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. DaNE: A named entity resource for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. NorNE: Annotating named entities for norwegian.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. The lacunae of danish natural language processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition.

Jouni Luoma, Li-Hsin Chang, Filip Ginter, and Sampo Pyysalo. Fine-grained named entity annotation for finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 135–144.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Dan Saattrup Nielsen. ScandEval: A benchmark for scandinavian natural language processing.

Ole & Asmussen Jorg Norling-Christensen. The corpus of the danish dictionary. 8(8):223–242. Publisher: Bureau of the WAT.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Barbara Plank. The 'problem' of human label variation: On ground truth in data, modeling and evaluation.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. DaN+: Danish nested named entities and lexical normalization.

Gabriel Recchia and Michael N. Jones. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. 41:647–656. Publisher: Springer.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. 115:211–252. Publisher: Springer.

Erik F. Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands. Linköping University Electronic Press, Sweden.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. The danish gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421. Linköping University Electronic Press, Sweden.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347. Association for Computational Linguistics.

Asahi Ushio and Jose Camacho-Collados. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62. Association for Computational Linguistics.

Guido Van Rossum and Fred L Drake Jr. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Ralph Weischedel, Pradeer Sameer, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, and Mohammed El-Bachouti. OntoNotes release 5.0.

A Dataset card

Following work by Mitchell et al. (2019) and (Gebru et al., 2021), we provide a dataset card for DANSK following the format proposed in Lhoest et al. (2021), which can be accessed here: https://anonymous.4open.science/r/dansk-3A03

A.1 Dataset Summary

DANSK: Danish Annotations for NLP Specific TasKs is a dataset consisting of texts from multiple domains, sampled from the Danish GigaWord Corpus (DAGW). The dataset was created to fill in the gap of Danish NLP datasets from different domains, that are required for training models that generalize across domains. The Named-Entity annotations are moreover fine-grained and have a similar form to that of OntoNotes v5, which significantly broadens the use cases of the dataset. The domains include Web, News, Wiki & Books, Legal, Dannet, Conversation and Social Media. For a more in-depth understanding of the domains, please refer to DAGW.

The distribution of texts and Named Entities within each domain can be seen in the table below:

A.1.1 Update log

- 2023-05-26: Added individual annotations for each annotator to allow for analysis of inter-annotator agreement

A.1.2 Supported Tasks

The DANSK dataset currently only supports Named-Entity Recognition, but additional version releases will contain data for more tasks.

A.1.3 Languages

All texts in the dataset are in Danish. Slang from various platforms or dialects may appear, consistent with the domains from which the texts originally have been sampled - e.g. Social Media.

A.2 Dataset Structure

A.2.1 Data Instances

The JSON-formatted data is in the form seen below:


```
{
    "text": "Aborrer over 2 kg er en uhyre sj\u00e6lden fangst.",
    "ents": [{"start": 13, "end": 17, "label": "QUANTITY"}],
    "sents": [{"start": 0, "end": 45}],
    "tokens": [
        {"id": 0, "start": 0, "end": 7},
        {"id": 1, "start": 8, "end": 12},
        {"id": 2, "start": 13, "end": 14},
        {"id": 3, "start": 15, "end": 17},
        {"id": 4, "start": 18, "end": 20},
        {"id": 5, "start": 21, "end": 23},
```

```
        {"id": 6, "start": 24, "end": 29},
        {"id": 7, "start": 30, "end": 37},
        {"id": 8, "start": 38, "end": 44},
        {"id": 9, "start": 44, "end": 45},
    ],
    "spans": {"incorrect_spans": []},
    "dagw_source": "wiki",
    "dagw_domain": "Wiki & Books",
    "dagw_source_full": "Wikipedia",
}
```

A.2.2 Data Fields

- `text`: The text
- `ents`: The annotated entities
- `sents`: The sentences of the text
- `dagw_source`: Shorthand name of the source from which the text has been sampled in the Danish Gigaword Corpus
- `dagw_source_full`: Full name of the source from which the text has been sampled in the Danish Gigaword Corpus
- `dagw_domain`: Name of the domain to which the source adheres to

A.2.3 Data Splits

The data was randomly split up into three distinct partitions; train, dev, as well as a test partition. The splits come from the same pool, and there are thus no underlying differences between the sets. To see the distribution of named entities, and domains of the different partitions, please refer to the paper, or read the superficial statistics provided in the Dataset composition section.

A.3 Descriptive Statistics

A.3.1 Dataset Composition

Named entity annotation composition across partitions is provided in Table 10.

A.3.2 Domain distribution

Domain and source distribution across partitions is provided in Table 11.

A.3.3 Entity Distribution across partitions

Domain and named entity distributions for the training, testing, and validation sets can be found in the full dataset card accompanying DANSK: https://anonymous.4open.science/r/dansk-3A03

Table 10: Named entity annotation composition across partitions

	Full	Train	Validation	Test
Texts	15062	12062 (80%)	1500 (10%)	1500 (10%)
Named entities	14462	11638 (80.47%)	1327 (9.18%)	1497 (10.25%)
CARDINAL	2069	1702 (82.26%)	168 (8.12%)	226 (10.92%)
DATE	1756	1411 (80.35%)	182 (10.36%)	163 (9.28%)
EVENT	211	175 (82.94%)	19 (9.00%)	17 (8.06%)
FACILITY	246	200 (81.30%)	25 (10.16%)	21 (8.54%)
GPE	1604	1276 (79.55%)	135 (8.42%)	193 (12.03%)
LANGUAGE	126	53 (42.06%)	17 (13.49%)	56 (44.44%)
LAW	183	148 (80.87%)	17 (9.29%)	18 (9.84%)
LOCATION	424	351 (82.78%)	46 (10.85%)	27 (6.37%)
MONEY	714	566 (79.27%)	72 (10.08%)	76 (10.64%)
NORP	495	405 (81.82%)	41 (8.28%)	49 (9.90%)
ORDINAL	127	105 (82.68%)	11 (8.66%)	11 (8.66%)
ORGANIZATION	2507	1960 (78.18%)	249 (9.93%)	298 (11.87%)
PERCENT	148	123 (83.11%)	13 (8.78%)	12 (8.11%)
PERSON	2133	1767 (82.84%)	191 (8.95%)	175 (8.20%)
PRODUCT	763	634 (83.09%)	57 (7.47%)	72 (9.44%)
QUANTITY	292	242 (82.88%)	28 (9.59%)	22 (7.53%)
TIME	218	185 (84.86%)	18 (8.26%)	15 (6.88%)
WORK OF ART	419	335 (79.95%)	38 (9.07%)	46 (10.98%)

Table 11: Domain and source distribution across partitions

Domain	Source	Full	Train	Dev	Test
Conversation	Europa Parlamentet	206	173	17	16
Conversation	Folketinget	23	21	1	1
Conversation	NAAT	554	431	50	73
Conversation	OpenSubtitles	377	300	39	38
Conversation	Spontaneous speech	489	395	54	40
Dannet	Dannet	25	18	4	3
Legal	Retsinformation.dk	965	747	105	113
Legal	Skat.dk	471	364	53	54
Legal	Retspraktis	727	579	76	72
News	DanAvis	283	236	20	27
News	TV2R	138	110	16	12
Social Media	hestenettet.dk	554	439	51	64
Web	Common Crawl	8270	6661	826	783
Wiki & Books	adl	640	517	57	66
Wiki & Books	Wikipedia	279	208	30	41
Wiki & Books	WikiBooks	335	265	36	34
Wiki & Books	WikiSource	455	371	43	41

A.4 Dataset Creation

A.4.1 Curation Rationale

The dataset is meant to fill in the gap of Danish NLP that up until now has been missing a dataset with 1) fine-grained named entity recognition labels, and 2) high variance in domain origin of texts. As such, it is the intention that DANSK should be employed in training by anyone who wishes to create models for NER that are both generalizable across domains and fine-grained in their predictions. It may also be utilized to assess across-domain evaluations, in order to unfold any potential domain biases. While the dataset currently only entails annotations for named entities, it is the intention that future versions of the dataset will feature dependency Parsing, pos tagging, and possibly revised NER annotations.

A.4.2 Source Data

The data collection, annotation, and normalization steps of the data were extensive. As the description is too long for this readme, please refer to the associated paper upon its publication for a full description.

Initial Data Collection and Normalization

A.4.3 Annotations

Annotation process To afford high granularity, the DANSK dataset utilized the annotation standard of OntoNotes 5.0, featuring 18 different named entity types. The full description can be seen in the associated paper.

Annotators 10 English Linguistics Master's program students from Aarhus University were recruited through announcements in classrooms. They worked 10 hours/week for six weeks from October 11, 2021, to November 22, 2021. Their annotation tasks included part-of-speech tagging, dependency parsing, and NER annotation. Annotators were compensated at the standard rate for students, as determined by the collective agreement of the Danish Ministry of Finance and the Central Organization of Teachers and the CO10 Central Organization of 2010 (the CO10 joint agreement), which is 140DKK/hour. Named entity annotations and dependency parsing was done from scratch, while the POS tagging consisted of corrections of silver-standard predictions by an NLP model.

A.4.4 Automatic correction

During the manual correction of the annotation a series of consistent errors were found. These were corrected using Regex patterns which can be view in full with the DANSK release, along with the Danish Addendum to the Ontonotes annotation guidelines: https://anonymous.4open.science/r/dansk-3A03

A.4.5 Licensing Information

Creative Commons Attribution-Share Alike 4.0 International license

B Training progression

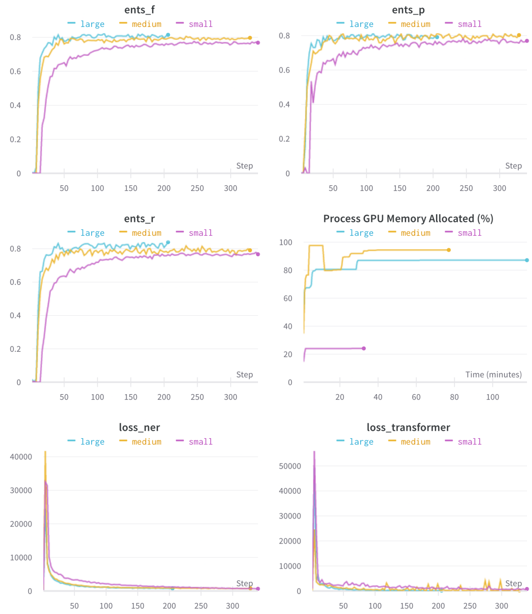

Figure 7: The epoch training progression of loss of the NER head (loss_ner), loss of the transformer (loss_transformer), NER performance measured in recall (ents_r), precision (ents_p), F1-score (ents_f) and GPU-allocation percentage.