Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4597–4604
Marseille, 11–16 May 2020

DaNE: A Named Entity Resource for Danish

Rasmus Hvingelby[1*], Amalie Brogaard Pauli[1*], Maria Barrett[2], Christina Rosted[1], Lasse Malm Lidegaard[1], Anders Søgaard[2]
[1]Alexandra Institute, Denmark
[2]University of Copenhagen, Denmark

Abstract

We present a named entity annotation for the Danish Universal Dependencies treebank using the CoNLL-2003 annotation scheme: DaNE. It is the largest publicly available, Danish named entity gold annotation. We evaluate the quality of our annotations intrinsically by double annotating the entire treebank and extrinsically by comparing our annotations to a recently released named entity annotation of the validation and test sections of the Danish Universal Dependencies treebank. We benchmark the new resource by training and evaluating competitive architectures for supervised named entity recognition (NER), including FLAIR, monolingual (Danish) BERT and multilingual BERT. We explore cross-lingual transfer in multilingual BERT from five related languages in zero-shot and direct transfer setups, and we show that even with our modestly-sized training set, we improve Danish NER over a recent cross-lingual approach, as well as over zero-shot transfer from five related languages. Using multilingual BERT, we achieve higher performance by fine-tuning on both DaNE and a larger Bokmål (Norwegian) training set compared to only using DaNE. However, the highest performance is achieved by using a Danish BERT fine-tuned on DaNE. Our dataset enables improvements and applicability for Danish NER beyond cross-lingual methods. We employ a thorough error analysis of the predictions of the best models for seen and unseen entities, as well as their robustness on un-capitalized text. The annotated dataset and all the trained models are made publicly available.

Keywords: named entity recognition, resource, Danish, cross-lingual transfer

1. Introduction

Named entity recognition (NER) is a crucial, yet challenging, component of a wide range of natural language processing applications, ranging from knowledge base population and other natural language understanding tasks to privacy protection systems. For Danish, there has not yet been a larger named entity (NE) annotated training set and until very recently; also no test set. Despite the lack of freely available training data, a few NER tools exist for Danish (Bick, 2004; Johannessen et al., 2005; Derczynski, 2019; Al-Rfou et al., 2015) but until Plank (2019), they have not been consistently benchmarked. This study found that Danish NER can benefit from transfer from English.

In this paper, we present a new annotated resource for Danish NE, DaNE, which contains annotations of the Danish Universal Dependencies treebank (Johannsen et al., 2015). We describe the annotation process and evaluate the annotation quality intrinsically by looking at annotator agreement, and extrinsically by comparing to the recent NE annotation of the validation and test split of the same resource (Plank, 2019). We benchmark our new resource with supervised state-of-the-art NE taggers, both off-the-shelf systems and models that are trained or fine-tuned on our train set.

Our train set is smaller than available resources in other North-Germanic languages (Swedish, Norwegian (Bokmål and Nynorsk) and West-Germanic languages (English and Dutch). We, therefore, experiment with different ways of improving Danish NER using cross-lingual transfer from these languages using multilingual BERT (Pires et al., 2019; Devlin et al., 2019).

Contributions We introduce the largest gold-annotated and publicly available Danish NE dataset. We train/fine-tune state-of-the-art NER systems on our train set as well as explore cross-lingual transfer from five related languages using multilingual BERT and benchmark these on our resource. All resources and trained models are made publicly available.[1]

2. Related work

A few tools for Danish NER (Derczynski, 2019; Bick, 2004; Johannessen et al., 2005) have been presented. Plank (2019) recently annotated the Danish Universal Dependencies validation and test sets as well as 10,000 tokens from the train set, benchmarked existing models and explored the potentials for cross-lingual transfer from English to Danish. This study found that neural transfer is possible, and that a small amount of Danish NE training data helps cross-lingual models.

For the related North-Germanic languages, there are already NE resources. The large Stockholm-Umeå Corpus (SUC) (Nilsson Björkenstam and Byström, 2012) is also annotated with NE in SUC 2.0 (Källgren, 1998) and SUC 3.0 (Östling, 2012). Recently, the Norwegian Dependency Treebank (Solberg et al., 2014) was also annotated with NE (NorNE) (Jørgensen et al., 2020). English, German, and Dutch are established languages in CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) for NER.

Looking at other languages, state-of-the-art performance on NER has been achieved on a set of benchmark tasks, e.g., the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) on English and German by Akbik et al. (2018) using their proposed contextual string embeddings, denoted FLAIR embeddings. The results have been further improved by Akbik et al. (2019b) proposing a pooled version of the embeddings. Also, BERT-based models (Devlin et al., 2019) and Long Short Term Memory models

*First two authors contributed equally

[1]https://github.com/alexandrainst/danlp

	TRAIN	DEV	TEST
# sentences	4383	564	565
# tokens	80,378	10,322	10,023
# sentences w/ NE	2021	272	285
% sentences w/ NE	46.1%	48.2%	50.4%
# entities	4003	480	558
type-token ratio	0.20	0.35	0.34
% unseen entities	-	57.5%	70.7%

Table 1: DaNE dataset statistics.

	LOC	MISC	ORG	PER
COUNT	945	1007	802	1249
%	23.6	25.2	20.0	31.2

Table 2: Distribution of NE classes in the DaNE train set.

(LSTM) with a Conditional Random Field (CRF) layer (LSTM+CRF) (Straková et al., 2019; Lample et al., 2016) are among strong models for NER.

3. Dataset description

The Danish Universal Dependencies treebank is a publicly available resource following Universal Dependencies standards (Johannsen et al., 2015)[2]. It is a conversion of the annotation of the Copenhagen Dependency Treebank (Kromann, 2003)[3]. It consists of 474 texts with 5,512 sentences and approximately 100k words. The source is texts from the Danish PAROLE corpus (Keson, 2000) which comprises a range of textual domains, both written and spoken from the years 1983–1992. We use the established train/validation/test splits from the Danish Universal Dependencies. General statistics of the dataset is presented in Table 1.

4. Annotation

The dataset has been annotated twice; once by a linguist and once by six non-linguist annotators, with no annotation overlap amongst the non-linguist annotators. All are native Danish speakers. Conflicts have been solved manually after calculating the inter-annotator agreement. The vast majority of conflicts were caused by an annotator missing an entity but in Section 4.1. we outline some of the difficult cases. Table 2 shows the the final train set class distribution. We annotate NE using the four classes, LOC(ation), ORG(anisation), PER(son) and MISC(ellaneous), following the guidelines of the CoNLL-2003 NE annotation scheme (Tjong Kim Sang and De Meulder, 2003). Despite being a predecessor to the more extensive CoNLL-2012 NE annotation scheme, the CoNLL-2003 scheme is still actively used for annotation (Darwish, 2013; Fromreide et al., 2014) as well as for evaluation (Peters et al., 2018; Akbik et al., 2018). Below we summarize the annotation guidelines for the four classes.

[2] https://github.com/UniversalDependencies/UD_Danish-DDT

[3] The Danish text in the Copenhagen Dependency Treebank is similar to the Danish Dependency Treebank.

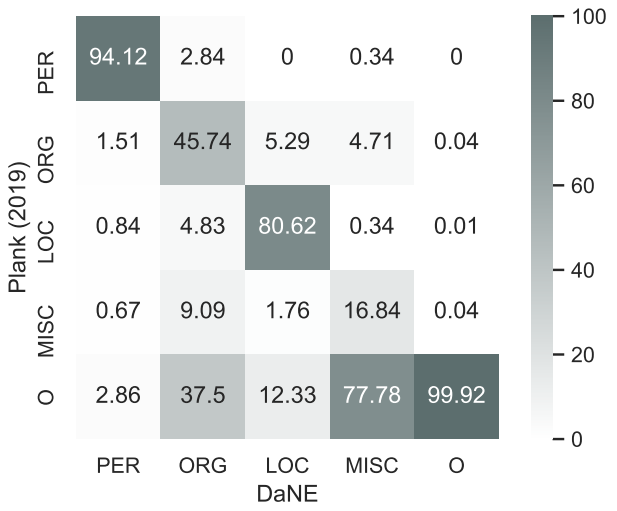


Figure 1: Normalized confusion matrix when comparing our validation and test set annotation to the annotation of Plank (2019).

LOC includes locations like cities, roads and mountains, as well as both public and commercial places like specific buildings or meeting points, but also abstract places.

PER consists of names of people, fictional characters, and animals. The names includes aliases.

ORG can be summarized as all sorts of organizations and collections of people, ranging from companies, brands, political movements, governmental bodies and clubs.

MISC is a broad category of e.g. events, languages, titles and religions, but this tag also includes words derived from one of the four tags as well as words for which one part is from one of the three other tags.

The O tag is used for the remaining tokens.

Raw token accuracy of the double annotation is 98.0%. MISC was the most difficult class for the annotators. We achieve an inter-annotator agreement on 0.87 using Cohens κ when excluding O tags. When also excluding MISC, the Cohens κ is 0.91.

4.1. Solving annotation conflicts

All annotation conflicts are solved manually by two annotators. Homonym ambiguity is solved by looking at the context, e.g., `Brøndby` (the name of a Danish city) would normally be LOC, but when referring to the city's soccer team it is ORG. Similarly, schools and universities, according to the guidelines can act both as ORG and LOC. In most cases, they have been categorized as ORG. For instance, `Københavns Universitet` (University of Copenhagen) rarely imply a specific location, as its campuses are spread all around the city. It is thus most often in the sense of an organisation that `Københavns Universitet` is mentioned. With other schools, the case is not as clear, since most schools have one particular location. Whether to annotate the school as ORG or LOC must be based on an interpretation of each instance. However,

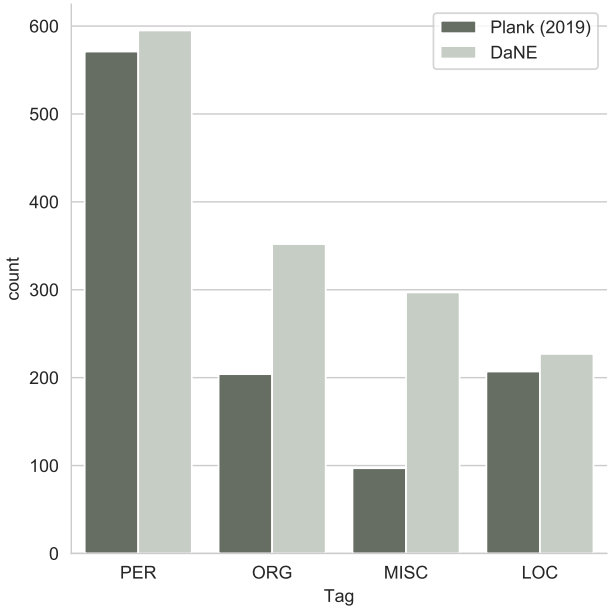


Figure 2: Distribution of non-O NE tags (irrespective of I- and B- status) in DaNE and Plank (2019) of the validation and test splits.

names of municipalities and counties are more likely to be referring to a location, but in some contexts they are ORGs. Other atypical cases are, e.g., names of big ships that can act both as ORG and LOC, ferry routes are LOC, `Vesten` (The West) is vague but nonetheless LOC. Then there are cases such as God and Santa Claus which are, some would say, fictional PERs.

4.2. Extrinsic evaluation

We compare the annotation of our validation and test set with the annotation of Plank (2019). This resource also follows the CoNLL-2003 annotation scheme, but only considers tokens marked as proper nouns in Danish Universal Dependencies as candidates for one of the entity types. In the validation and test split, Plank (2019) annotated 1079 tags to be non-O. In DaNE, there are 1471 non-O tags. The increased non-O annotations are distributed unevenly across all non-O tags as shown in Figure 2. In DaNE, there are more ORG and MISC tags than in the annotations of the former study. Raw accuracy when comparing to Plank (2019) is 97.3%. Cohens κ is 0.84, however, when excluding the MISC tag, we get a much higher agreement 0.90.
Figure 1 shows the confusion matrix when comparing the two annotations. This also shows a high agreement for the three main classes (LOC, PER, and ORG). But 77.78% of the MISCs in DaNE are annotated with O by Plank (2019). This seems to be mainly due to languages, nationalities and adjectives referring, e.g., to nations not being annotated as MISC. Examples are `dansk, dansker, europæiske` (Danish, Dane, European). In Danish, opposed to in English, such words should not be capitalized and are thus easier to miss by annotators. Other systematic differences are compounds consisting of at least one NE that are annotated with O by Plank and with MISC in DaNE. Examples are `Beatles-musik, Volvo-motoren, FN-debatten` (Beatles music, the Volvo engine, the UN debate).
37.5% of PER-tagged words in DaNE are annotated as O by Plank. This can be attributed to the difference in annotation guidelines. E.g. names of political parties and ministries seem to be largely annotated with O by Plank whereas they are annotated as ORG in DaNE.

5. Experiments

This section describes the models used for benchmarking on our dataset. We benchmark off-the-shelf models along with well-known models trained/fine-tune on our train set. The best performing trained/fine-tuned models are also evaluated for comparison on the test set of Plank (2019).

5.1. Off-the-shelf Models

POLYGLOT The POLYGLOT model (Al-Rfou et al., 2015) is trained without any human annotation or language-specific knowledge but by automatic generating a dataset using the link structure from Wikipedia. The model recognizes the three tags: LOC, ORG, and PER, but not MISC. We benchmark the model implemented in the POLYGLOT framework[4]

DANER DANER[5] (Derczynski, 2019) is a wrapper around the Stanford CoreNLP (Manning et al., 2014) using data from Derczynski et al. (2014) (not released). Like POLYGLOT, this model recognizes LOC, ORG, and PER, but not MISC.

5.2. Models trained/fine-tuned on the dataset

For all trained/fine-tuned models, we report average performance over five different random seeds.

5.2.1. biLSTM+CRF

We train a bidirectional LSTM (biLSTM) model (Lample et al., 2016) using a public implementation[6]. The model consists of a single-layer biLSTM (Hochreiter and Schmidhuber, 1997) with 200 hidden dimensions that takes a concatenation of character encoding and word embeddings as input. The character encoding is obtained by applying a biLSTM with 50 hidden dimensions on 25-dimensional character embeddings. The CRF uses the output from the biLSTM to make a prediction. The model is trained for 100 epochs with early stopping and a dropout of 0.5, a batch size of 10 and it uses SGD as optimizer with a learning rate of 0.01.
For pre-trained word embeddings (+EM), we use FastText word embeddings trained on Common Crawl and Wikipedia by Grave et al. (2018). The word embeddings are chosen due to best performance among a list of publicly available Danish embeddings evaluated intrinsically on Danish word similarity datasets by Schneidermann et al. (2020).

[4] `https://github.com/aboSamoor/polyglot`
[5] `https://github.com/ITUnlp/daner`
[6] `https://github.com/allanj/pytorch_lstmcrf`

	nl	no	nn	sv	en	da
Resource	CoNLL-2002	NorNE	NorNE	SUC 3.0	CoNLL-2003	DaNE
# tokens	309,206	310,222	301,353	1,166,593	301,369	100,733
% entities	9.30%	6.49%	6.67%	3.48%	16.84%	7.25%
type-token ratio	0.137	0.123	0.119	0.105	0.116	0.203
% word overlap with DaNE	6.88	25.27	16.26	17.18	6.81	-

Table 3: Overview of resources for cross-lingual experiments compared to DaNE. Measures below the line only concern the train split.

5.2.2. FLAIR

The sequence tagging architecture implemented in the FLAIR framework (Akbik et al., 2019a) is a biLSTM+CRF based on Huang et al. (2015) with the option of passing concatenated embeddings of different types. Using contextual embeddings (FLAIR embeddings) in this architecture has shown state-of-the-art performance on NER tasks in other languages (Akbik et al., 2018; Akbik et al., 2019b). Therefore, we have pre-trained Danish FLAIR embeddings on Danish text from Wikipedia[7] and EuroParl (Tiedemann, 2012). The FLAIR embeddings are the extracted 1024-dimensional hidden states of a biLSTM character-level language model. We train the FLAIR embeddings for five epochs with a batch size of 50. The FLAIR embeddings are concatenated with FastText embeddings (described in Section 5.2.1.). The models are trained for 150 epochs and a batch size of 32 with SGD optimization and an annealing learning rate starting at 0.1. We also train a FLAIR model where we concatenate the FLAIR embeddings and FastText with byte pair encoding (BPE) embeddings computed on Danish Wikipedia by Heinzerling and Strube (2018).

5.2.3. Danish BERT

BERT is a transformer-based architecture that can be pre-trained on a large corpus of raw text. Devlin et al. (2019) show that it requires only a small amount of fine-tuning of pre-trained BERT representations to obtain high performance on, e.g., NER. Recently, a Danish pre-trained BERT was made publicly available[8]. The model is pre-trained on Danish lowercased text from Common Crawl, Danish Wikipedia, OpenSubtitles (Lison and Tiedemann, 2016) and various online forums. For our BERT experiments, we use a public implementation of BERT[9] along with the pre-trained weights for the Danish BERT (DA-BERT). We fine-tune the BERT model on the DaNE train set for 5 epochs with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 8. Checkpoints are evaluated on the validation set every 50 iterations.

5.2.4. Multilingual BERT

Using a BERT pretrained on text in 104 languages, has shown promising results for cross-lingual transfer for NER (Pires et al., 2019). We experiment with cross-lingual transfer using the pre-trained multilingual BERT[10] (M-BERT). The model is pre-trained on texts from the 104 largest language-specific Wikipedias. We use the same BERT implementation and fine-tuning settings as in our experiments with DA-BERT.

For the monolingual supervised setting, we fine-tune only on the DaNE train set. For cross-lingual transfer experiments, we try different combinations of transfer, first exploring the transfer from one language to Danish (zero-shot). We also explore neural transfer by fine-tuning on one other language in combination with the DaNE train set. Plank (2019) find that the best neural transfer from English to Danish happens when the size gap between the amount of English train data and the amount of Danish train data was smallest. To lower the gap between the amount of Danish data compared to the amount of the data from the transfer language, we experimented with over-sampling the Danish data by a factor of three. This makes the amount of Danish roughly equal to the amount of nl, no, nn and en. However, we find that oversampling Danish does not yield better results for any of the languages compared to the equivalent experiment without oversampling and these results are therefore not reported.

Cross-lingual data We selected five West- and North-Germanic languages for which large NE resources exist. When choosing transfer language for cross-lingual learning, Lin et al. (2019) found that word overlap unsurprisingly correlates well with best transfer languages. We provide these figures, as well as dataset sizes in Table 3.

As we use the same annotation scheme as the Dutch (Tjong Kim Sang, 2002) and the English (Tjong Kim Sang and De Meulder, 2003) datasets, they can be directly used for cross-lingual training. However, for the Swedish SUC 3.0 (Östling, 2012) we map WRK (work of art), EVN (event) and OBJ (miscellaneous) to MISC to make the classes compatible. As the SUC 3.0 dataset does not come with predefined splits we define the first 70% of the data as train, the next 15% as validation and the last 15% as test set. Due to the large difference in size between SUC 3.0 and DaNE, we do not run any experiment with a combination of the two.

NorNE (Jørgensen et al., 2020) is a balanced dataset of both Norwegian languages, Nynorsk (nn) and Bokmål (no). Bokmål is closer to Danish and we, therefore, split NorNE into two parts for these experiments. As NorNE also uses other NE classes, we map GPE_LOC to LOC, GPE_ORG to

[7] https://dumps.wikimedia.org/dawiki/latest/

[8] https://github.com/botxo/nordic_bert

[9] https://github.com/huggingface/transformers

[10] https://github.com/google-research/bert/blob/master/multilingual.md

Model	Train set	Micro F1		LOC	MISC	ORG	PER
		–MISC	+MISC				
Polyglot	-	64.18	-	64.95	-	39.3	78.74
Daner	-	54.57	-	58.51	-	25.64	68.14
biLSTM+CRF	DaNE	67.17	64.44	68.28	53.21	47.98	80.82
biLSTM+CRF+EM	DaNE	77.16	72.68	78.84	54.12	59.77	89.05
Flair	DaNE	82.08	79.70	84.62	70.73	68.70	90.96
Flair+BPE	DaNE	80.33	78.05	82.66	69.99	67.46	88.59
da-BERT	DaNE	**86.61**	**83.76**	**87.30**	74.72	**78.27**	**93.52**
m-BERT	DaNE	83.93	81.73	82.49	73.79	73.27	93.52
m-BERT	en	81.31	76.99	80.10	57.75	70.62	91.83
m-BERT	nl	77.14	69.60	78.54	46.13	62.13	87.37
m-BERT	no	81.09	74.77	78.09	45.99	71.23	90.98
m-BERT	nn	78.85	72.46	80.08	43.31	65.41	89.73
m-BERT	sv	69.08	59.61	79.00	9.37	33.98	83.77
m-BERT	en + DaNE	82.26	80.80	79.04	**75.44**	71.00	93.45
m-BERT	nl + DaNE	83.65	81.60	81.56	74.11	74.20	92.76
m-BERT	no + DaNE	85.54	83.29	84.20	75.11	76.79	93.37
m-BERT	nn + DaNE	82.42	80.70	82.89	74.43	71.58	87.51

Table 4: F1 scores average over five runs of models on our test set. Best result per class is boldfaced.

	LOC	MISC	ORG	PER
Plank (2019)	63.6	**24.8**	42.5	86.6
biLSTM+CRF+EM	72.44	20.40	47.12	91.23
Flair	79.85	13.35	53.33	90.66
m-BERT no+DaNE	80.31	13.19	57.68	92.31
da-BERT	**80.33**	11.48	**59.96**	**92.44**

Table 5: F1 score per class label on the test set from Plank (2019). The results for models trained on our train set is averaged over five runs. The results for Plank (2019) are taken directly from this paper.

ORG and PROD (product), EVT (event), DRV (derived) to MISC for compatibility with the CoNLL-2003 annotation scheme.

6. Results

The models are evaluated on the dataset using the CoNLL evaluation script to get entity-level F1 scores per class. We report micro average F1 score both with and without the MISC class to be able to compare it to models which do not predict MISC. The results on our test set using existing off-the-shelf models and trained models for Danish are reported in Table 4. We observe that all trained or fine-tuned models work better than the off-the-shelf tools, Polyglot and Daner. For the biLSTM+CRF, we observe that using pretrained embeddings substantially helps the performance. The best model on average is da-BERT fine-tuned on DaNE. We find that the best performance with m-BERT comes from fine-tuning on DaNE and no (Bokmål) which also has the largest word overlap (Table 3) with DaNE.

For the zero-shot models, we observe a significant performance drop on the F1 scores on especially the MISC and ORG classes but a decent performance on PER and LOC.

6.1. Evaluating on existing evaluation set

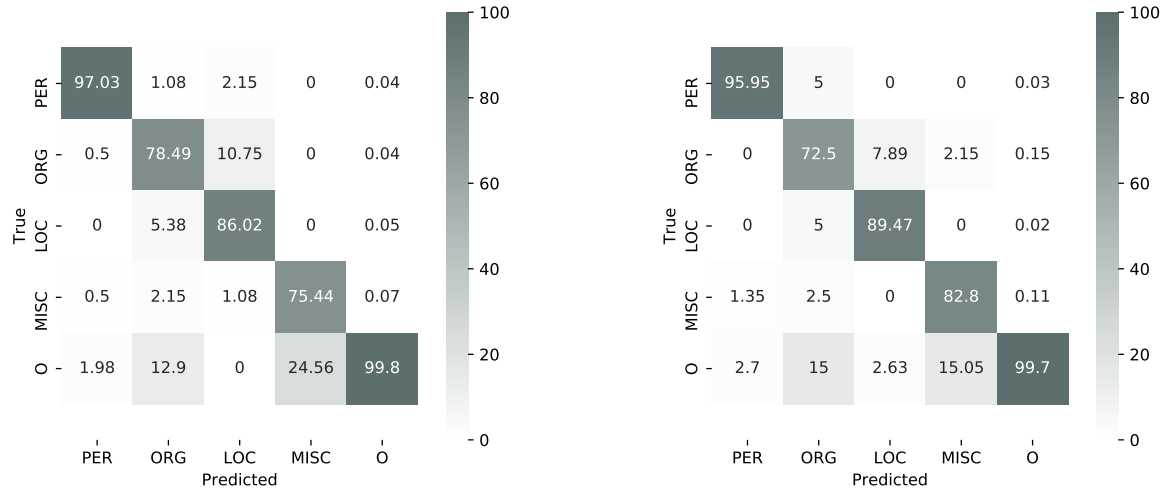
Results on the test set provided by Plank (2019) for the best m-BERT, biLSTM+CRF, da-BERT, and Flair are reported in Table 5 along with figures for the neural transfer model by Plank (2019). On this dataset, the best model on LOC, ORG, and PER is the da-BERT. We observe that the MISC class is difficult for all models including the cross-lingual model of Plank (2019). The neural transfer model by Plank (2019) is best on the MISC class.

7. Robustness test and error analysis

The robustness is tested on both da-BERT and the cross-lingual transfer m-BERT models. Like Augenstein et al. (2017), we calculate the F1 score on unseen and seen entities. The error analysis digs deeper into this for the single best performing da-BERT model.

Table 6 show performance on the validation splits across seen and unseen entities during training. Table 1 shows the number of unseen entities per data split. The m-BERT fine-tuned on nn+DaNE is best for unseen entities and the m-BERT fine-tuned only on DaNE is best for seen entities. We observe a notable drop from seen to unseen entities e.g. the model with the best performance the unseen data drops more than seven percentage points compared to the performance on seen data.

To get more insight into where the model fails, we choose the single da-BERT model performing the best on the validation set. The model achieves a F1 score of 87.60 on AVG(+MISC) on the validation set. Figure 3 shows the confusion matrix for unseen and seen entities, respectively.



(a) Seen entities

(b) Unseen entities

Figure 3: Normalized confusion matrix for predictions on the validation set from the best DA-BERT trained on DaNE.

TRAIN SET	VALIDATION			TEST
	All	Unseen	Seen	Lowercase
DaNE	90.08	82.19	**92.64**	22.15
DaNE lower	84.06	74.84	86.09	79.51
en	84.50	71.79	87.18	8.02
nl	80.66	70.80	83.50	2.84
nn	84.67	74.84	85.12	7.90
no	83.18	76.33	85.16	2.30
nl+DaNE	89.32	82.04	91.67	19.73
en+DaNE	89.79	82.70	91.99	17.00
no+DaNE	88.95	82.87	90.90	14.57
nn+DaNE	**90.59**	**84.52**	92.34	16.91
DA-BERT	88.34	80.89	90.59	**86.61**

Table 6: Robustness results of M-BERT and DA-BERT models on subsets of entities in the validation set and on a lowercased version of test set. F1 score averaged over five runs and calculated without MISC. Best result per class is boldfaced.

We observe that for unseen entities, MISC is more often predicted to be O than for seen entities. Also, ORG is more often predicted to be LOC for both seen and unseen. The model also often confuses O tags to be ORG.

The Danish Universal Dependencies treebank contains well-edited text meaning that entities are capitalized. Note, however, that writing conventions in Danish dictate that names of languages, religions, and nationalities (in the MISC class) are not capitalized. We report results on a lowercased test set in the last column in Table 6. We observe that the M-BERT models fine-tuned on capitalized text heavily relies on capitalization and we observe that this is especially true for the zero-shot models. To lower the gap in performance we also fine-tune a M-BERT model on a lowercased DaNE training set (DaNE lower). This gives a F1 score of 79.51 on the lowercased test set which is closer to to the F1 score of 83.93 for the M-BERT trained and evaluated on the capitalized train and test sets. However, the DA-BERT which is trained purely on lowercased text both in pre-training and fine-tuning does not rely on capitalization and thus has the highest performance on the lowercased test set. In order to use capitalization information without overfitting to these features, further data augmentation strategies (Bodapati et al., 2019) would be a promising direction for NER on less well-edited text such as user-generated text or speech transcripts.

8. Conclusion

We have presented the largest Danish NE resource to date: DaNE. The data source is the Danish Universal Dependencies treebank. The resource is publicly available and will enable the training and evaluation of NER models and allows for tracking of progress for Danish NER. We have trained/fine-tuned models on DaNE and benchmarked the resource using state-of-the-art models including multilingual BERT and a recent Danish BERT model.

9. Acknowledgements

This research is partly supported by a performance contract allocated to the Alexandra Insitute by the Danish Ministry of Higher Education and Science.

10. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019a). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019b). Pooled contextualized embeddings for named entity recognition.

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 724–728, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015*, April.

Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Bick, E. (2004). A named entity recognizer for Danish. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Bodapati, S., Yun, H., and Al-Onaizan, Y. (2019). Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China, November. Association for Computational Linguistics.

Darwish, K. (2013). Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Derczynski, L., Field, C. V., and Bøgh, K. S. (2014). DKIE: Open source information extraction for Danish. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 61–64.

Derczynski, L. (2019). Simple natural language processing tools for Danish. *arXiv preprint arXiv:1906.11608*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Fromreide, H., Hovy, D., and Søgaard, A. (2014). Crowdsourcing and annotating NER for twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2544–2547, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Heinzerling, B. and Strube, M. (2018). BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Johannessen, J. B., Hagen, K., Haaland, Å., Jónsdottir, A. B., Nøklestad, A., Kokkinakis, D., Meurer, P., Bick, E., and Haltrup, D. (2005). Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Johannsen, A., Alonso, H. M., and Plank, B. (2015). Universal dependencies for Danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Jørgensen, F., Aasmoe, T., Husevåg, A.-S. R., Øvrelid, L., and Velldal, E. (2020). NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseilles, France, May. European Language Resources Association (ELRA).

Källgren, G. (1998). Documentation of the stockholm-umeå corpus. *Department of Linguistics, Stockholm University*.

Keson, B. (2000). Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus. Technical report, Det Danske Sprog- og Litteraturselskab (DSL).

Kromann, M. T. (2003). The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 217–220, Växjö, Sweeden.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, July. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Nilsson Björkenstam, K. and Byström, E. (2012). SUC-CORE: SUC 2.0 annotated with np coreference. In *Swedish Language Technology Conference*.

Östling, R. (2012). Stagger: A modern pos tagger for swedish. In *The Fourth Swedish Language Technology Conference*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Plank, B. (2019). Neural cross-lingual transfer and limited annotated data for named entity recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Schneidermann, N. S., Hvingelby, R., and Pedersen, B. S. (2020). Towards a gold standard for evaluating Danish word embeddings. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseilles, France, May. European Language Resources Association (ELRA).

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. B. (2014). The Norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Straková, J., Straka, M., and Hajic, J. (2019). Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy, July. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Tjong Kim Sang, E. F. and De Meulder, F. (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.

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)*.