DAN+: Danish Nested Named Entities and Lexical Normalization

Barbara Plank, Kristian Nørgaard Jensen and Rob van der Goot
Department of Computer Science
ITU Copenhagen, Denmark
bplank@itu.dk, krnj@itu.dk, robv@itu.dk

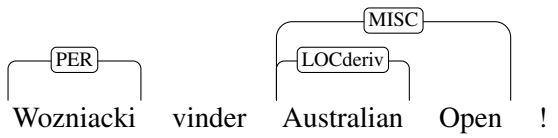
Abstract

This paper introduces DAN+, a multi-domain resource for nested named entities (NEs) and lexical normalization for Danish, a less-resourced language. We empirically assess three strategies to model the two-layer NE annotations, cross-lingual cross-domain transfer from German versus in-language annotation, language-specific versus multilingual BERT, and the effect of lexical normalization on Danish NE. Our results show that the most robust strategy is multi-task learning which is rivaled by multi-label decoding, transfer is successful also for zero-shot, and in-language BERT and lexical normalization works the best on the least canonical data. However, our results also show that out-of-domain remains challenging, while performance on news plateaus quickly. This highlights the importance of cross-domain evaluation of cross-lingual transfer.

1 Introduction

Named Entity Recognition (NER) is the task of finding entities in text, such as locations, organizations, and persons. NER is a key step towards natural language understanding, for instance for question answering and information extraction. The task has received a substantial amount of attention, particularly for English and CoNLL-style entity extraction (with four core entity types: PER, ORG, LOC, MISC). Most research so far, including for Danish, focused on newswire data and flat entities only. It ignores nested entities, like 'Australian Open' (illustrated in Figure 1), which contains both an event and a location-derived sub-entity. There is also little prior work that studies transfer learning for nested NER.

In this paper we introduce DAN+, a novel resource for Danish nested NER and lexical normalization, covering texts from canonical data from newswire and non-canonical social media sources. Danish bears interesting challenges for NER, similar to German (Benikova et al., 2014), which we capture by drawing inspiration from the NOSTA-D (Benikova et al., 2014) NER annotation scheme. In particular, location-derived adjectives like 'dansk' (Danish) or 'hollandske' (Dutch) are not capitalized, and there are phrases which are only partially named entities, like 'Baltica-aktierne' (the Baltica shares). Such entities were mostly ignored so far. Full annotation guidelines are provided in the supplement.



a: *Wozniacki wins the Australian Open.*

Morten	er	bare	ginial	og	fraek	og	sjov	xD	xD	!
Morten	er	bare	genial	og	fræk	og	sjov	xD	xD	!
PER	O	O	O	O	O	O	O	O	O	O

b: *Morten is just brilliant, naughty and fun*

Figure 1: Examples from DAN+ with (a) nested entities and (b) lexical normalization annotation.

Contributions We present 1) DAN+, a new multi-domain dataset for nested NER and lexical normalization; 2) an evaluation of various models for Danish nested NER, including BERT variants and in-language versus cross-language experiments; 3) experiments for lexical normalization on Danish and its impact on NER. All code and data to reproduce the experiments is available as supplement.

2 Related Work

Nested NEs have received less research focus in contrast to flat entities (Grishman and Sundheim, 1996; Grishman, 1998; Tjong Kim Sang and De Meulder, 2003; Baldwin et al., 2015b). This has been attributed to technological complexity (Finkel and Manning, 2009) and limited data availability (Ringland et al., 2019). Existing nested NE data mostly spans newswire and biomedial data for English (Kim et al., 2003; Mitchell et al., 2005) and German news (Benikova et al., 2014), for example. Interest in nested NER is re-emerging (Katiyar and Cardie, 2018), with many new recent neural approaches (Sohrab and Miwa, 2018; Luan et al., 2019; Lin et al., 2019; Zheng et al., 2019). To facilitate research, a fine-grained nested NER annotation on top of the Penn Treebank has been released recently (Ringland et al., 2019).

To facilitate research on a less-resourced language, namely Danish, Plank (2019) introduced publicly available evaluation data of flat NER on top of Danish UD (Johannsen et al., 2015), providing annotations for approximately 20% of the data. The study also first benchmarked existing NER tools and evaluated the feasibility of transfer for Danish. Hvingelby et al. (2020) recently independently annotated the entire Danish UD data for flat NERs, though with different guidelines, annotating also adjectives, for example. Before these two recent studies, Danish NER data was behind a paywall or available tools were not benchmarked (Bick, 2004; Derczynski et al., 2014; Johannessen et al., 2005; Al-Rfou et al., 2013). To the best of our knowledge, DAN+ is also the first Danish nested NER dataset beyond newswire.

Domain shift is a pressing issue in NLP. One solution is to normalize the input text before detecting NEs, which is a mitigation strategy particularly suitable for social media (Eisenstein, 2013). Previous work has evaluated lexical normalization for a variety of languages—but not for Danish—with varying degrees of success (Schulz et al., 2016; Küçük and Steinberger, 2014; Nguyen et al., 2016; Liu et al., 2013; Li and Liu, 2015; Dugas and Nichols, 2016). Most works do not evaluate the normalization model intrinsically, which is often restricted to a simple rule-based approach which unlikely transfers well.

To the best of our knowledge, there is very little prior work on cross-lingual and cross-domain transfer for nested (or overlapping) entities, except for contemporary work on English-Arabic (Lan et al., 2020).

3 Data and Annotation

This section depicts the data sources and annotation. We refer the reader to the appendix for details on sampling, annotation guidelines and data statement. Table 1 provides an overview of the DAN+ dataset.

Included Data Varieties DAN+ includes canonical data from newswire and three social media varieties: 1) NEWS: annotations on top of the existing treebank (Johannsen et al., 2015); 2) REDDIT: from the `r/Denmark` sub-reddit, sampled from top-voted posts; 3) TWITTER: sample from 2019-2020 based on Danish-specific emotion words (love, pain, surprise); 4) ARTO: blog posts and comments from a Danish social media platform (active from 1988 until 2006).

Variety	German: News	DAN+: News (UD-DDT)			Reddit		Twitter		Arto	
	TRAIN	TRAIN	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Sentences	24,002	4,383	564	565	328	137	120	110	336	336
Tokens	452,853	80,378	10,332	10,023	4,668	5,005	5,346	5,086	5,496	43,90
Types	74,609	16,330	3,640	3,424	1,829	1,829	2,103	2,017	1,648	1,475
Sentences w/ NEs	59%	42%	45%	43%	56%	53%	74%	75%	22%	21%
1st level-NE	29,078	3,497	439	453	303	135	239	248	93	105
2nd level-NE	2,467	170	9	25	27	12	11	15	1	9
Tokens normalized	—	—	—	—	—	—	3.5%	2.3%	16.7%	15.2%

Table 1: Overview of DAN+: **Da**nish **N**ested **N**amed entities and lexical **N**ormalization, which includes news and social media varieties (Reddit, Twitter, Arto). First column: GermEval (Benikova et al., 2014).

Data Annotation We opted for a two-level NER annotation scheme following NoSTA-D (Benikova et al., 2014). *First-level* annotations contain outermost entities (e.g., the company 'Maribo Frø'). *Second-level* annotations are sub-entities (location 'Maribo'). Three annotators were involved, two of which are

native Danish speakers and one is proficient in Danish. For each task, a native speaker annotated the entire dataset after initial training. Inter-annotator agreement was high. For NER on the development sections of the Reddit and Twitter datasets, Cohen's κ on the entities without nesting was 90.97 and 83.08, respectively. With nesting, the κ scores were 87.81 and 80.94. For lexical normalization, 10% of the data was annotated by the two native speakers. For the decision on whether to normalize they reached a κ of 88.66, whereas for the choice of the correct normalization the agreement was 96.30%.

4 Experimental Setup

For nested NER, we use BERT (Devlin et al., 2019) with fine-tuning implemented in MaChAmp (van der Goot et al., 2020). We evaluate three decoders: 1) `single task-merged`, where we simply merge both annotation layers into a single flat entity, 2) `multi-task`, where the encoding is shared and each layer of annotation has its own decoder, 3) `multi-label`, which treats it as multi-label problem, where a label i is predicted if $P(l_i|\cdot) \geq \tau$ (Bekoulis, 2019).

We first evaluate all NER models on Danish, both within news and on out-of-domain (OOD) varieties. We further compare to transfer from German: 1) `zero-shot` transfer, fine-tuning only on German; and 2) `union` of the Danish and German data for fine-tuning. We compare multilingual BERT (`ml`, which includes Danish) versus training with Danish BERT (`da`). For MaChAmp, we use the proposed default parameters (van der Goot et al., 2020) shown to work well across tasks. We tune early stopping and τ on Danish news dev data. We compare our final model to the `boundary-aware` model (Zheng et al., 2019), a state-of-the-art nested NER which also evaluated on GermEval 2014. We train it with bilingual Danish and German Polyglot embeddings obtained via procrustes alignment (Conneau et al., 2018). We always use the offical GermEval evaluation script (Benikova et al., 2014) with strict span-based F1.

For normalization, we choose to use MoNoise (van der Goot, 2019), because it is open-source and is the only model that has shown to reach competitive performance across multiple languages. Intrinsic normalization results are reported as word-level accuracy on a 10-fold setup of dev+test, whereas results on NE tagging are span-f1 score on the development data. More details are provided in the appendix.

5 Results

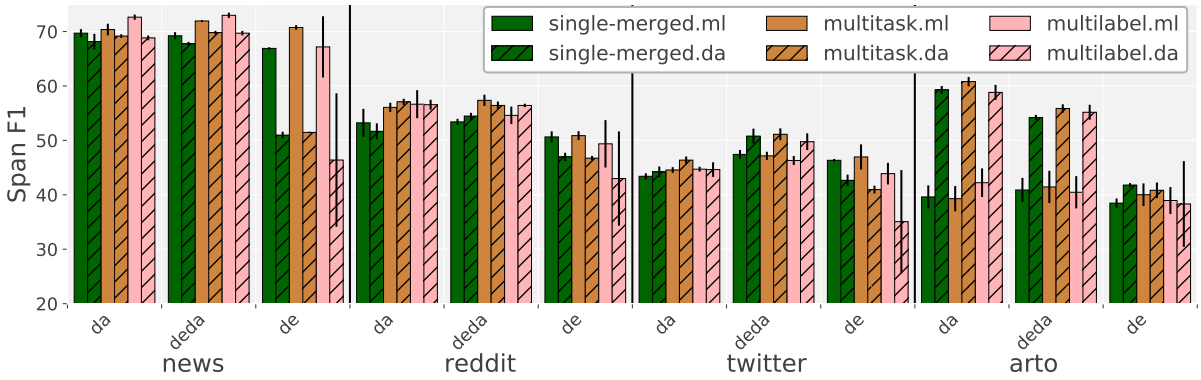
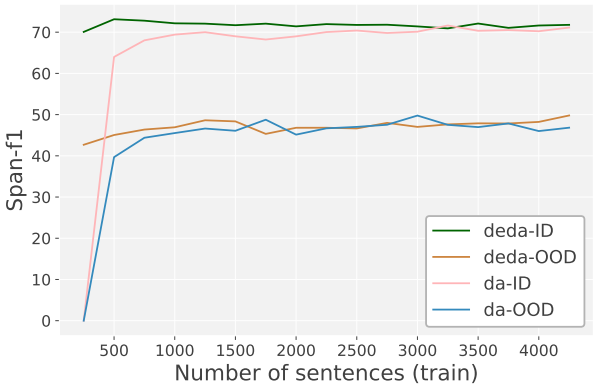


Figure 2: Nested NER results trained on German (de), Danish (da) or both (deda). Average over 3 runs.

Results on nested NER are reported in Figure 2. First, we observe the expected domain shift, with F1 in the 70ies on news, which drops to 40-60% on the non-canonical data. Training on the union of German and Danish (`deda`) is mostly beneficial for the Twitter data, but harms performance on the Arto, where training on Danish alone is best. Overall, results of `deda` are close to training on Danish only, which is five times smaller. The German model performs remarkably well on Danish in zero-shot setups. This can be explained by the closeness of the languages, the annotations and the relatively large training data (Table 1).[1]

[1]We leave training on sub-samples of German for future work.



(a) Learning curve for multitask, ml-BERT on in-domain (ID) news and average over all out-of-domain datasets (OOD).

	Normalization		NE tagging	
	Twitter	Arto	Twitter	Arto
Baseline	97.17	83.93	47.16	41.45
MoNoise	97.17	92.52	47.37	52.61
Gold	100.00	100.00	47.85	65.17

	German	News	Reddit	Twitter	Arto
boundary-aware	57.89	56.89	16.48	21.37	13.77
Raw (ml)	83.93	71.68	56.18	63.57	53.26
Norm'ed (ml)	—	—	—	64.36	57.40
Raw (da)	65.10	68.68	45.35	61.07	57.97
Norm'ed (da)	—	—	—	60.56	58.60

(b) Top: Normalization accuracy, and its downstream effect on NER. Bottom: Nested NER F1 score on the test sets.

Figure 3: Learning curves, normalization evaluation and impact of normalization on NER.

Regarding BERT, interestingly, the Danish BERT seem to be mostly beneficial for the non-canonical domains, in contrast to multilingual BERT which clearly fares best on news. This is likely due to forum data which is included for pre-training Danish BERT.[2] In contrast ml-BERT is trained on Wikipedia, and makes it less fit for out-of-domain data. This suggests that adaptive pre-training could yield even better results (Han and Eisenstein, 2019). The learning curve in Figure 3a further shows that in-domain performance plateaus surprisingly quickly (both for `da` and `dade`). Instead, the gap to the non-canonical domains is large, calling for more out-of-domain evaluation of NER models.

What is the accuracy and effect of normalization? We take a straighforward baseline which always copies the original token, and evaluate the impact of automatic vs gold normalization on NER (Table 3b).

The results in Table 3b (top) show that normalization results on the Arto data are in a similar range as previously reported state-of-the-art results on other languages (van der Goot, 2019).[3] On Twitter, which contains fewer anomalies, improvements are marginal compared to the baseline.

Performance of NER increases when text is normalized first, in all settings, both on dev (top) in Table 3b and on the final test set (bottom). As expected, the performance gain on the Twitter data is marginal as there is not much to normalize. On the Arto data however, we can observe a substantial improvement when applying MoNoise, and an even larger gain when using gold normalization.

Test set results We evaluate the best model (`deda-multitask`) with Danish and ml-BERT on the test sets. Table 3b (bottom) shows that our model outperforms the `boundary-aware` method, which turns out to be brittle to domain shifts. Overall, the results confirm that normalization helps (except on Twitter), and multilingual BERT is better than Danish BERT on canonical news data, whereas on the least standard data (Arto) it is the other way around.

6 Conclusions

This paper contributes to the limited prior work on cross-lingual cross-domain transfer of nested NER. We provide a new resource for Danish, DAN+, with baselines on nested NER and lexical normalization, using two BERT variants and training on Danish, German or both. Our results show that BERT-based variants are sensitive to domain shift for cross-domain nested NER, whereas they can cope relatively well with missing in-language data. Results on normalization show that it helps in case of very non-standard data only, for which automatic normalization improves Danish nested NER performance.

[2] It should be noted that it is trained on lower-cased texts, which is suboptimal for NER yet works surprisingly well.

[3] van der Goot (2019) use Error Reduction Rate (ERR) for evaluation, which is accuracy normalized for the amount of words that need to be normalized; ERR in our setup would be 52.05, (van der Goot, 2019) report ERR's between 29 and 77.

Acknowledgments

We thank Amanda Jørgensen for help with data annotation. We also thank NVIDIA, Google cloud computing and the ITU High-performance Computing cluster for computing resources. This research is supported in part by the Independent Research Fund Denmark (DFF) grant 9131-00019A MultiSkill, a DFF Sapere Aude grant 9063-00077B and by an Amazon Faculty Research Award.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015a. Guideline for English lexical normalisation shared task. Technical report, Workshop on Noisy User-generated Text.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015b. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Ioannis Bekoulis. 2019. *Neural Approaches to Sequence Labeling for Information Extraction*. Ph.D. thesis, Ghent University.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Eckhard Bick. 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).

Thomas Bilgram and Britt Keson. 1998. The construction of a tagged Danish corpus. In *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, pages 129–139.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Sixth International Conference on Learning Representations*.

Leon Derczynski, Camilla Vilhelmsen Field, and Kenneth S Bøgh. 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.

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, June. Association for Computational Linguistics.

Fabrice Dugas and Eric Nichols. 2016. DeepNNNER: Applying BLSTM-CNNs and extended lexicons to named entity recognition in tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 178–187, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore, August. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Ralph Grishman. 1998. Research in information extraction: 1996-98. In *Proceedings of the TIPSTER Text Program: Phase III*, pages 57–60, Baltimore, Maryland, USA, October. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China, November. Association for Computational Linguistics.

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, Marseille, France, May. European Language Resources Association.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana, June. Association for Computational Linguistics.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Matthias T Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish dependency treebank. In *International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 217–220.

Dilek Küçük and Ralf Steinberger. 2014. Experiments to improve named entity recognition on Turkish tweets. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 71–78, Gothenburg, Sweden, April. Association for Computational Linguistics.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. A focused study to compare arabic pre-training models on newswire ie tasks. In *arXiv 2004.14519*.

Chen Li and Yang Liu. 2015. Improving named entity recognition in tweets via detecting non-standard words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 929–938, Beijing, China, July. Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5182–5192, Florence, Italy, July. Association for Computational Linguistics.

Xiaohua Liu, Furu Wei, Shaodian Zhang, and Ming Zhou. 2013. Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):1–15.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Vu H Nguyen, Hien T Nguyen, and Vaclav Snasel. 2016. Text normalization for named entity recognition in Vietnamese tweets. *Computational social networks*, 3(1):10.

Barbara Plank. 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*, pages 370–375, Turku, Finland, September–October. Linköping University Electronic Press.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In *Workshop Proceedings of the 12th KONVENS 2014*.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R. Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy, July. Association for Computational Linguistics.

Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. 2016. Multimodular text normalization of Dutch user-generated content. *ACM Transactions on Intelligent Systems Technology*, 7(4):1–22, July.

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

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*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, and Barbara Plank. 2020. Massive Choice, Ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. *arXiv*.

Rob van der Goot. 2019. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy, July. Association for Computational Linguistics.

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. 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 357–366, Hong Kong, China, November. Association for Computational Linguistics.

A Evaluation Metrics and Setups

Nested NER For evaluation of NER we report strict F1 for both first and second level entities, using the official GermEval evaluation script (Reimers et al., 2014).

Normalization Because of the limited data size, we evaluate the normalization model in a 10-fold setup. For evaluation we use capitalization-sensitive accuracy on the word-level (including all words, also words not in need of normalization).

We based the n-grams used by the system on a Wikipedia dump from 01-01-2020 and Twitter data collected throughout 2012 and 2018, filtered with the FastText language classifier (Joulin et al., 2017), and trained skip-gram embeddings with default settings (Mikolov et al., 2013) on the same Twitter data.

Bert variants For Danish BERT we use the model trained by Botxo (`https://github.com/botxo/nordic_bert`), which is pre-trained on Wikipedia, Common Crawl, Danish debate forums and Danish open subtitles. For multilingual BERT, we use the following pretrained embeddings: `multi_cased_L-12_H-768_A-12`.

Results on GermEval It should be noted that the results for German we report in Table 3b are lower compared to the ones reported in (Zheng et al., 2019), this is partly because of the different embeddings, and partly a difference in implementation of the span-f1 metric. We here strictly follow the evaluation setup of the GermEval 2014 shared task and use the official strict F1 evaluation setup.

B Full results

Table 2 contains the exact scores which Figure 2 is based on. We also report the scores on only the nested entities in Table 3; the multitask approach clearly outperforms the other models for this category.

	German	News	Reddit	Twitter	Arto
da.ml.single-merged	65.78	69.70	53.22	43.40	39.61
da.ml.multitask	68.22	70.37	56.06	44.56	39.29
da.ml.multilabel	67.38	72.64	56.65	44.71	42.22
da.da.single-merged	24.74	68.17	51.67	44.22	59.27
da.da.multitask	28.07	69.16	57.08	46.37	**60.80**
da.da.multilabel	26.14	68.83	56.55	44.64	58.82
deda.ml.single-merged	76.68	69.22	53.40	47.42	40.89
deda.ml.multitask	84.71	71.94	**57.36**	47.16	41.45
deda.ml.multilabel	79.93	**72.98**	54.60	46.31	40.47
deda.da.single-merged	61.49	67.75	54.47	50.79	54.20
deda.da.multitask	66.47	69.81	56.42	**51.12**	55.86
deda.da.multilabel	64.48	69.70	56.42	49.76	55.17
de.ml.single-merged	76.74	66.90	50.64	46.34	38.47
de.ml.multitask	**84.92**	70.74	50.87	46.96	40.02
de.ml.multilabel	77.40	67.18	49.37	43.89	38.95
de.da.single-merged	61.68	50.97	46.97	42.65	41.79
de.da.multitask	66.00	51.47	46.74	40.96	40.83
de.da.multilabel	59.86	46.39	42.99	35.07	38.31

Table 2: Span-f1 scores on all development sets for all out proposed models (single-merged, multi(task), multilabel), having two types of embeddings (da/ml), and all our training data combinations (da, deda, de).

	German	News	Reddit	Twitter	Arto
da.ml.single-merged	1.43	0.00	0.00	0.00	0.00
da.ml.multitask	4.80	**63.89**	21.05	15.38	0.00
da.ml.multilabel	1.82	13.97	3.70	4.76	0.00
da.da.single-merged	0.00	0.00	0.00	0.00	0.00
da.da.multitask	0.00	56.23	19.35	0.00	0.00
da.da.multilabel	0.00	0.00	0.00	0.00	0.00
deda.ml.single-merged	10.33	0.00	5.44	13.97	0.00
deda.ml.multitask	66.02	51.45	37.17	**21.37**	0.00
deda.ml.multilabel	21.69	9.72	7.02	5.13	0.00
deda.da.single-merged	3.69	0.00	3.34	4.76	0.00
deda.da.multitask	42.01	48.28	21.91	0.00	0.00
deda.da.multilabel	4.79	0.00	0.00	0.00	0.00
de.ml.single-merged	10.71	0.00	8.84	10.68	0.00
de.ml.multitask	**67.13**	27.22	**38.06**	20.51	0.00
de.ml.multilabel	22.53	0.73	7.69	2.99	0.00
de.da.single-merged	3.18	0.00	3.81	4.76	0.00
de.da.multitask	40.39	23.09	2.15	0.00	0.00
de.da.multilabel	5.61	12.39	2.00	1.35	**1.15**

Table 3: Span-f1 scores on all development sets for only the nested entities.

C DaN+ Data Statement

Following (Bender and Friedman, 2018), the following outlines the data statement for DaN+:

A. CURATION RATIONALE Collection of examples of Danish language for identification of named entities in different text domains, complemented with lexical normalization annotation to study the impact of it on NER.

B. LANGUAGE VARIETY The non-canonical data was collected via the Twitter search API, the Reddit API and the Wayback archive.

Danish (da-DK) and some US (en-US) mainstream English, Swedish (se-SE) and Norwegian (no-NO) in the Reddit sample.

C. SPEAKER DEMOGRAPHIC For the newswire data this is unknown. For the social media samples it is Danish and Scandinavian Reddit, Twitter and Arto users. Gender, age, race-ethnicity, socioeconomic status are unknown.

D. ANNOTATOR DEMOGRAPHIC Two students and one faculty (age: 25-40), gender: male and female. White European. Native language: Danish, German. Socioeconomic status: higher-education student and university faculty.

D. SPEECH SITUATION Both standard and colloquial Danish, i.e., edited and spontaneous speech. Time frame of data between 1988 and 2020.

D. TEXT CHARACTERISTICS Sentences from journalistic edited articles and from social media discussions and postings.

PROVENANCE APPENDIX The news data originates from the Danish UD DDT data, GNU Public License, version 2 OR CC BY-SA 4.0: `https://github.com/UniversalDependencies/UD_Danish-DDT/blob/master/README.md`

D Details of data collection

The NEWS data is from the publicly available Danish Universal Dependencies data. The Danish DDT UD is a conversion of the Danish Dependency Treebank (DDT) (Kromann et al., 2003) with news texts from PAROLE-DK (Bilgram and Keson, 1998). It consists of 5,512 sentences and 100k tokens. We annotate the entire dataset for nested NERs. For most of the experiments, we use the canonical train/dev/test split. For the non-canonical datasets, we sample approximately 5,000 tokens to create development and test sets. We use three different online sources: Twitter, Reddit and Arto. An overview of the collected data with statistics is provided in Table 1.

We collected posts from REDDIT from the `r/Denmark` sub-reddit, in particular the top voted posts.[4] The collected posts all span from a single data (November 28th 2019) and contain tokens other than Danish (842 English tokens, 101 Swedish and 5 Norwegian).

We sample a set of Danish tweets collected from TWITTER collected over 2019-2020 based on a list of Danish-specific emotion-carrying words (like love, pain, surprise), to avoid having mainly news articles. To make sure the data contains some phenomena interesting for normalization, we filtered the data to contain at least 3 words not present in the Aspell dictionary.[5]

ARTO was Denmark's first social media platform and operated from 1988 till 2006. Because the website is not accessible anymore, we scraped all blog pages (where 'blogs' can also consist of only a few words) and their corresponding comments available from the Wayback Machine.[6] Similar to the Twitter data, we sample a subset and filter the data to contain some normalization density based on the ratio of out-of-vocabulary words (> 0.12).

After publication, we will make all data freely available, with scrapers for parts which cannot be directly released due to licensing (i.e., Reddit).

[4] Using a universal Reddit scraper `https://github.com/JosephLai241/Universal-Reddit-Scraper`

[5] We complemented the Aspell dictionary with some common named entities and interjections for this purpose.

[6] `https://archive.org/web/`

E Annotation guidelines for NER

This section describes the annotation guidelines which we used for our DAN+ NER corpus. Our guidelines were adopted from the German NoSta-D guidelines (Benikova et al., 2014).

We stick to a two layer annotation, where the outermost embraces the longer span and is the most prominent entity reading, and the inner span contains secondary or sub-entity readings. If there would be more than 2 layers, we drop the second potential reading in favor of keeping two layers (e.g., Australian Open is both an event and hence MISC but also an ORG; as Australian is a LOCderiv, we here keep only MISC for the event and LOCderiv for Australian).

Step 1: Named entities are nominal phrases that determine specific people, organizations, locations or miscellaneous specific objects like film titles or products. National holidays or religious events (*Jul, Ramadan*) are not annotated. Given the following example:

[Leila] bought [the house]

There are two nominal phrases. Only one of them is a named entity (Leila), the second nominal is a common noun.

Step 2: Potential NEs Only full nominal phrases are potential full NEs. Pronouns and all other phrases should be ignored. Derivations of NEs, i.e., words which are derived through morphological derivation processes, are marked (e.g., *danske*). NEderiv do not need to be nominal phrases. Declination (e.g., possessives) are not considered derivations and are directly annotated as NEs. For mediums such as social media we do mark user names and hashtags as potential NEs.

- Full NEs are annotated as LOC (location), ORG (organization), PER (person) or MISC (miscellaneous other)
- Derivations of NEs are marked as such by appending *deriv*, e.g., den [danske]*LOCderiv* midtbanespiller

Examples:

- Location: *[København]LOC*, *[Kastrup]LOC*
- *[Carsten Jensen]PER*
- *[IKEA]ORG*
- *[Parken]LOC* (Stadium)
- *[The Shining]MISC*, *[Jojo]MISC* (product name, song titles etc)
- Location adjectives: *De [københavnske]LOCderiv gader*
- Person adjectives: *[Freudiansk]PERderiv litteratur*
- BUT possessives: *[[Denmarks]LOC Radio]ORG*, *[Københavns]LOC kommune*, *[Johannsons]PER hus*

Examples:

- Organizations: *[Twitter]ORG*, *[TV2]ORG*, på min [FB]ORG
- BUT: reference to specific Reddit channels *[/r/all]MISC*
- *at være [dansker]LOCderiv på [reddit]ORG*

Step 3: Titles, owners Determiners and titles are not part of NEs. But owners can be NEs by itself.
Examples:

- *dronning [Margareth]PER*, *dronning [Margareth II]PER* (numbers are kept as part of the name)
- *[Vivaldis]PER [Vier Jahreszeiten]MISC*

Step 4: Multi-word tokens NEs often consist of multiple tokens.
Examples:

- person names: *[Terry Hatcher]PER*
- film titles (MISC): *[Breaking Bad]MISC*

Step 5: Nesting NEs can be nested.
Examples:

- locations in organization names: *[[Allerød]LOC Gymnasium]ORG*
  ([[Nordjyllands]LOC politi]ORG)
- organization names in product names: *[[Google]ORG Translate]MISC*

Step 6: Parts Named entities can also be parts of tokens and are annotated as such with the suffix "part".
Examples:

- *[pro-hongkong]LOCpart*
- *[Hverdags-Lars]PERpart*

Step 6: Medium-specific potential NEs Named entities can also be parts of special medium-specific tokens, like user names and hashtags in Twitter. We do annotate them as such.
Examples:

- *[@hik_fodbold]ORG*
- *[#ToppenAfPoppen]MISC*
- *[@realDonaldTrump]PER*

todo: list tables with examples

F Annotation guidelines for lexical normalization

The guidelines are based on (Baldwin et al., 2015a), all the cases were we diverged from these guidelines, or when we believed clarification was necessary are described below.

Systematic miss-spellings

Since the data was taken from social media some words were systematically spelled wrong. This is especially seen on Arto, where many words were spelled using q instead of g. Here q was replaced with g:

jeq ↦ jeg (I) muliqe ↦ mulige

As it is also common to write words without the last one or two letters, there were also many words missing one or multiple letters in the end. Here the missing letters were inserted:

ik ↦ ikke hva ↦ hvad

Capitalization

Capitalization was corrected in names, first letter in a post and after periods, question marks and other signs that require capitalization in the first letter of the following word. Capitalized words that illustrate yelling or emphasis have been decapitalized, acronyms that are capitalized have been kept capitalized.

TILLYKKE ↦ tillykke (congratulations) DR ↦ DR (Denmark's Radio)

Splitting and merging

Words that were incorrectly split or incorrectly merged into one word were corrected.

ar bej der ↦ arbejder (works) istedet ↦ i stedet (instead)

Phrasal abbreviations

There was no correction of phrasal abbreviations because the written-out form does not correspond to the intended meaning of the phrase. The only ones found were in English.

lol ↦ lol omg ↦ omg

Hashtags

Hashtags and usernames were not corrected, even if they were misspelled or if they contained multiple words.

#sundhedforalle ↦ #sundhedforalle (health for all)

Corrections of the letters æ, ø, å

The Danish alphabet contains the three letters æ, ø and å. If these are not available at the used keyboard they are often replaced by other vowels:

ae ↦ æ o ↦ ø aa ↦ å

In words where the replacement vowels are used they have been replaced with the appropriate letter. In some data æ, ø and å were left out entirely, here the letters were inserted. As the missing letter in some cases results in multiple options, the word was determined using the context:

har ↦ har (to have) or hår (hair) fler ↦ flere (more) or føler (feels)