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 new multi-domain corpus and annotation guidelines for Danish nested named entities (NEs) and lexical normalization to support research on cross-lingual cross-domain learning for a less-resourced language. We empirically assess three strategies to model the two-layer Named Entity Recognition (NER) task. We compare transfer capabilities from German versus in-language annotation from scratch. We examine language-specific versus multilingual BERT, and study the effect of lexical normalization on NER. Our results show that 1) the most robust strategy is multi-task learning which is rivaled by multi-label decoding, 2) BERT-based NER models are sensitive to domain shifts, and 3) in-language BERT and lexical normalization are the most beneficial on the least canonical data. Our results also show that an out-of-domain setup 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. Most research so far, including for Danish, focused on newswire data and flat entities. It ignores nested entities, like 'Australian Open' (illustrated in Figure 1) being both an event and a location-derived entity. There is also little prior work on transfer learning for nested NER.

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

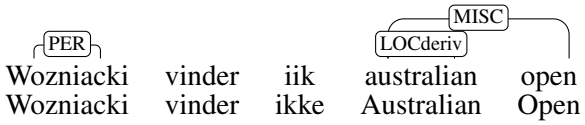

Figure 1: Example with (a) nested entities and (b) lexical normalization.

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) first experiments of lexical normalization on Danish and its downstream impact on NER.[1]

[1] All code and data to reproduce the experiments is available at https://github.com/bplank/DaNplus

Proceedings of the 28th International Conference on Computational Linguistics, pages 6649–6662
Barcelona, Spain (Online), December 8-13, 2020

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 biomedical 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. To the best of our knowledge, DAN+ is 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 is unlikely to transfer well. DAN+ provides also data to study lexical normalization for Danish.

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

3 Data and Annotation

This section depicts the data sources and annotation. Table 1 provides an overview of the DAN+ dataset. For normalization, as opposed to earlier annotation efforts in other languages, we included correction of capitalization. We refer to the appendix for details on annotation guidelines and data statement.

3.1 Data varieties

DAN+ includes canonical data from newswire and three social media varieties:

News The Danish DDT UD treebank (Johannsen et al., 2015; Kromann et al., 2003), which consists of news texts from PAROLE-DK (Bilgram and Keson, 1998). We use the canonical train/dev/test split.

Reddit Sampled from the `r/Denmark` sub-reddit, in particular the top voted posts.[2] The collected posts all span a single date (November 28th 2019) and the data contains some non-Danish tokens (842 English tokens, 101 Swedish and 5 Norwegian).

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

Arto Arto was Denmark's first large-scale 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 from the Wayback Machine.[4] Similar to the Twitter data, we sample a subset and filter the data to contain some normalization density.

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

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

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

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	326	126	120	110	336	337
Tokens	452,853	80,378	10,332	10,023	4,547	4,497	5,347	5,086	5,496	4,389
Types	74,609	16,330	3,640	3,424	1,807	1,616	2,103	2,017	1,648	1,474
Sentences w/ NEs	59%	45%	47%	48%	60%	56%	80%	77%	21%	20%
1st level-NE	29,078	3,800	468	525	319	128	279	284	93	103
2nd level-NE	2,467	235	36	41	36	20	13	32	1	12
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).

3.2 Annotation

We opted for a two-level NER annotation scheme following largely the annotation scheme provided by 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'). Four annotators were involved, three 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 entity tokens 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).[5] We evaluate three decoding strategies:

- `single task-merged`: both annotation layers are merged into a single flat entity.
- `multi-task`: the encoder is shared and each layer of annotation has its own decoder.
- `multi-label`: treats nested NER as multi-label problem, where a label i is predicted if $P(l_i|\cdot) \geq \tau$ (Bekoulis, 2019) further illustrated in Ramponi et al. (2020).

We first evaluate all NER models on Danish, both within news and on the three 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 (mlBERT) versus training with Danish BERT (danishBERT).[6] Even though both are trained on Danish data, for mlBERT this is Wikipedia data, whereas danishBERT is trained on Wikipedia, Common Crawl, Danish debate forums, and Danish subtitles. 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, and set $\tau = 0.9$. We compare our final model to the `boundary-aware` model (Zheng et al., 2019), a state-of-the-art nested NER model which was also evaluated on GermEval 2014. We train it with bilingual Danish and German Polyglot embeddings obtained via Procrustes alignment (Conneau et al., 2018). For evaluation we use the official GermEval script (Benikova et al., 2014) with strict span-based F1 over both entity levels.[7]

For normalization, we choose to use MoNoise (van der Goot, 2019), since it is open-source and is the only model that has shown to reach competitive performance across multiple languages. MoNoise

[5] MaChAmp v0.2, included in the repository.

[6] version 2 from: `https://github.com/botxo/nordic_bert`

[7] Compared to the evaluation of Zheng et al. (2019), this script is more strict. The scores we report are thus slightly lower compared to the ones reported in Zheng et al. (2019).

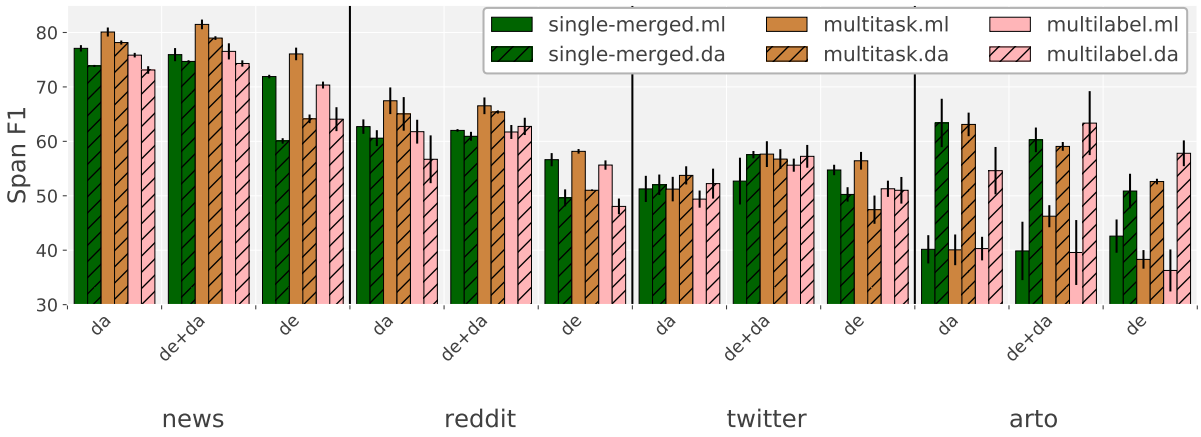

Figure 2: Nested NER results. Models trained on German (de), Danish (da) or both (de+da) and with mlBERT (ml) vs danishBERT (da). Average over 3 runs. Standard deviation indicated as error line.

requires n-grams and word embeddings to reach a good performance. We use 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). Intrinsic normalization results are reported as capitalization sensitive word-level accuracy over all words (including words which are not normalized). Because we have no external training data for normalization, we use a 10-fold setup of dev+test.

5 Results

5.1 NER

Figure 2 depicts the main results for nested NER on the dev set, while detailed results are given in Table 4 and 5 in the appendix. First, we note that the model performs well within Danish newswire, reaching an F1 score in the 80ies (left bars). However, we observe a domain shift, as performances drop to 38-67% on the three non-canonical social media datasets, with Twitter and Arto reaching lowest scores.

Our Danish training dataset is of modest size, hence the question arises whether existing German data is beneficial. The German multi-task model performs remarkably well on Danish news in zero-shot setups with mlBERT, reaching an F1 of 76%. This can be explained by the closeness of the languages, the annotations and the large training data (Table 1).

For a model trained on the union of the German and Danish data (`de+da`), we observe that performance is overall close to the model trained on Danish only, which is five times smaller. The average F1 over all Danish datasets (News, Reddit, Twitter, Arto) for the two best models (using multi-task learning) is 65.01 with `da.da.multitask` and 65.05 with `de+da.da.multitask`. Interestingly, danishBERT is the best for the least non-canonical domain (Arto), in contrast to mlBERT which fares best on news. This is likely due to forum data included for pre-training danishBERT,[8] while mlBERT is based on Wikipedia data, which is less fit for non-canonical data. This suggests that adaptive pre-training could yield better results (Han and Eisenstein, 2019; Ramponi and Plank, 2020).

We also compare transfer learning from German with increasing amounts of in-language Danish data. The learning curve in Figure 3 shows that transfer helps for low amounts of data, and in-domain performance plateaus surprisingly quickly (especially for the `da+de` setup), and in-language data remains the best in-domain (ID). Instead, the gap to the non-canonical domains remains large for both in-language and cross-language setups, and performance on OOD is less stable throughout, calling for more out-of-domain evaluation of NER models.

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

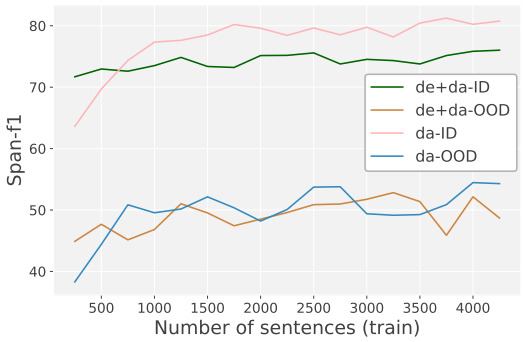

Figure 3: Learning curve for multi-task, mlBERT on in-domain (ID) news and average over all out-of-domain datasets (OOD).

	Normalization		NE tagging	
	Twitter	Arto	Twitter	Arto
Baseline	97.17	83.93	57.65	46.25
MoNoise	97.17	92.52	58.59	55.83
Gold	100.00	100.00	59.18	65.71

Table 2: Normalization accuracy, and its downstream effect on NER. For NER, multitask, mlBERT trained on de+da is used.

	German	News	Reddit	Twitter	Arto
boundary-aware	57.89	56.89	16.48	21.37	13.77
Raw (ml)	83.31	80.73	57.99	60.87	54.88
Norm'ed (ml)	—	—	—	61.74	56.38
Raw (da)	72.80	80.13	50.99	62.88	55.48
Norm'ed (da)	—	—	—	62.91	56.59

Table 3: Nested NER F1 score on the test sets for models with mlBERT (ml) vs danishBERT (da).

5.2 Lexical Normalization

We take a straightforward baseline for normalization, which always copies the original token, and we evaluate the impact of automatic versus gold normalization on NER. In other words, the accuracy of this baseline is equal to the percentage of not-normalized words. The results in Table 2 show that MoNoise is performing well for the less canonical Arto data in contrast to the Twitter data. On the Arto data, MoNoise reaches scores in a similar range compared to state-of-the-art results on other languages (van der Goot, 2019).[9]

In the downstream evaluation (right part of Table 2), we see that normalization is most beneficial when the data is less canonical (Arto), but even on Twitter normalization is beneficial. Furthermore, from the GOLD results, we can conclude that there is still space for improvement for automatic normalization.

5.3 Test Data

We evaluate the model that fares overall best on in-domain source news (`de+da-multitask`) with danishBERT and mlBERT on the test sets. Table 3 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 the most on the least canonical data (i.e. Arto), and mlBERT is better than danishBERT 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.

[9] 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 53.45, (van der Goot, 2019) report ERR's between 29 and 77.

Acknowledgments

We thank Amanda Jørgensen for help with data annotation for lexical normalization. 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 and 9063-00077B and an Amazon Faculty Research Award.

References

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.

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*.

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.

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

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.

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.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *COLING*.

Alan Ramponi, Rob van der Goot, Rosario Lombardi, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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 Full results

Table 4 contains the exact scores which Figure 2 is based on. We also report the scores on only the nested entities in Table 5; the multitask approach clearly outperforms the other models for this category.

	German	News	Reddit	Twitter	Arto
da.ml.single-merged	66.32	77.09	62.71	51.26	40.16
da.ml.multitask	67.94	80.09	**67.46**	51.23	40.07
da.ml.multilabel	64.82	75.84	61.78	49.39	40.30
da.da.single-merged	30.35	73.85	60.59	52.02	**63.37**
da.da.multitask	33.94	78.15	65.04	53.75	63.11
da.da.multilabel	25.65	73.11	56.71	52.23	54.61
de+da.ml.single-merged	75.77	75.93	62.01	52.70	39.87
de+da.ml.multitask	83.88	**81.48**	66.53	**57.65**	46.25
de+da.ml.multilabel	76.36	76.53	61.72	55.62	39.58
de+da.da.single-merged	67.76	74.69	60.93	57.56	60.27
de+da.da.multitask	74.35	78.98	65.39	56.74	59.07
de+da.da.multilabel	66.47	74.31	62.75	57.25	63.35
de.ml.single-merged	75.50	71.90	56.63	54.74	42.59
de.ml.multitask	**84.91**	76.06	58.16	56.43	38.31
de.ml.multilabel	72.62	70.34	55.65	51.29	36.28
de.da.single-merged	67.21	60.08	49.62	50.25	50.87
de.da.multitask	74.20	64.14	51.01	47.45	52.61
de.da.multilabel	67.12	64.07	48.05	51.01	57.81

Table 4: 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, de+da, de). Average over all Danish datasets (News, Reddit, Twitter, Arto) for the two best models are 65.01 for `da.da.multitask` and 65.05 for `de+da.da.multitask`. The latter is trained on 5 times more data while performing similarly to the model trained on Danish only.

	German	News	Reddit	Twitter	Arto
da.ml.single-merged	4.15	10.94	4.94	3.17	0.00
da.ml.multitask	22.67	43.60	42.32	21.43	0.00
da.ml.multilabel	0.00	0.00	1.63	0.00	0.00
da.da.single-merged	2.30	10.38	5.31	0.00	0.00
da.da.multitask	6.30	39.44	47.23	22.16	0.00
da.da.multilabel	0.00	0.00	0.00	0.00	0.00
de+da.ml.single-merged	14.90	11.60	14.93	13.56	0.00
de+da.ml.multitask	65.80	**47.61**	39.10	15.74	0.00
de+da.ml.multilabel	2.56	0.00	0.00	0.00	0.00
de+da.da.single-merged	11.62	7.37	20.38	23.15	0.00
de+da.da.multitask	55.32	46.12	**54.10**	**31.21**	0.00
de+da.da.multilabel	0.84	0.00	0.00	0.00	0.00
de.ml.single-merged	15.77	8.20	18.98	19.17	0.00
de.ml.multitask	**68.93**	43.74	40.66	29.72	0.00
de.ml.multilabel	0.00	0.00	0.00	0.00	0.00
de.da.single-merged	11.88	7.14	17.19	21.60	0.00
de.da.multitask	56.42	30.21	25.93	24.02	0.00
de.da.multilabel	3.29	0.00	0.00	0.00	0.00

Table 5: Span-f1 scores on all development sets for only the nested entities.

B 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 Three students and one faculty (age range: 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`

C 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., genitive forms) 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. Note that we diverge from the German NoSta-D guidelines by annotating the names of languages (e.g., *dansk, swahili*) as LOCderiv.

- 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*
- BUT when the location acts as an organized entity (e.g. country, municipality, sports club), it is tagged as ORG with LOC as inner layer: *[[Danmark]LOC]ORG indfører grænsekontrol*
- *[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 genitive forms: *[[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

D 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 $\mapsto$ jeg (I) muliqe $\mapsto$ 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 $\mapsto$ ikke hva $\mapsto$ 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 $\mapsto$ tillykke (congratulations) DR $\mapsto$ DR (Denmark's Radio)

Splitting and merging

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

ar bej der $\mapsto$ arbejder (works) istedet $\mapsto$ 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 $\mapsto$ lol omg $\mapsto$ omg

Hashtags

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

#sundhedforalle $\mapsto$ #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 $\mapsto$ æ o $\mapsto$ ø aa $\mapsto$ 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 $\mapsto$ har (to have) or hår (hair) fler $\mapsto$ flere (more) or føler (feels)