Cross-Lingual Cross-Domain Nested Named Entity Evaluation on English Web Texts

Barbara Plank
IT University of Copenhagen
Department of Computer Science
Rued Langgaards Vej 7, 2300 København S
bplank@gmail.com

Abstract

Named Entity Recognition (NER) is a key Natural Language Processing task. However, most existing work on NER targets flat named entities (NEs) and ignores the recognition of nested structures, where entities can be enclosed within other NEs. Moreover, evaluation of Nested Named Entity Recognition (NNER) across domains remains challenging, mainly due to the limited availability of datasets. To address these gaps, we present EWT-NNER, a dataset covering five web domains annotated for nested named entities on top of the English Web Treebank (EWT). We present the corpus and an empirical evaluation, including transfer results from German and Danish. EWT-NNER is annotated for four major entity types, including suffixes for derivational entity markers and partial named entities, spanning a total of 12 classes. We envision the public release of EWT-NNER to encourage further research on nested NER, particularly on cross-lingual cross-domain evaluation.

1 Introduction

Named Entity Recognition (NER) is the task of finding and classifying named entities in text, such as locations, organizations, and person names. It is a key task in Natural Language Processing (NLP), and an important step for downstream applications like relation extraction, co-reference resolution and question answering. The task has received a substantial amount of attention. However, tools and existing benchmarks largely focus on flat, coarse-grained entities and single-domain evaluation.

Flat, coarse-grained entities however eschew semantic distinctions which can be important in downstream applications (Ringland et al., 2019). Examples include embedded locations ('New York Times'), entities formed via derivation ('Italian cuisine') and tokens which are in part named entities ('the Chicago-based company').

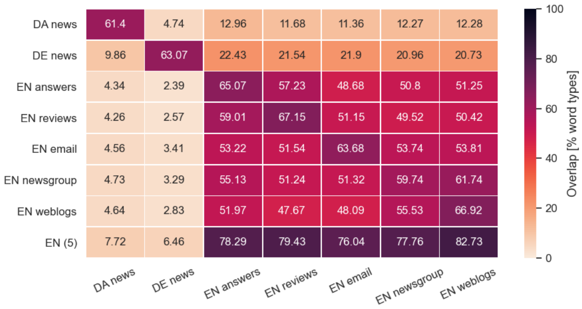

Figure 1: Domain overlap between target (x-axis) and source training (y-axis) domains (DE: German, DA: Danish, EN: proposed dataset EWT-NNER).

Research interest on methods to handle nested entities is increasing (Katiyar and Cardie, 2018). However, there is a lack of datasets, particularly resources which cover multiple target domains.

To facilitate research on cross-domain nested NER, we introduce a new layer on top of the English Web Treebank (EWT), manually annotated for NNER. The corpus spans five web domains, four major named entity types, enriched with suffixes marking derivations and partial NEs. Figure 1 shows the domain overlap in terms of word types. Besides providing in-language benchmark results, EWT-NNER enables research on cross-lingual transfer from German and Danish.

Contributions The main contributions are: i) We introduce EWT-NNER, a corpus for nested NER over five web domains. ii) A report on cross-lingual and in-language baselines. Our results highlight the challenges of processing web texts, and the need for research on cross-lingual cross-domain NNER.

2 Related Work

Nested NER Much research has been devoted to flat Named Entity Recognition, with a long tradition of shared tasks (Grishman and Sundheim, 1996; Grishman, 1998; Tjong Kim Sang and De Meulder, 2003; Baldwin et al., 2015). The problem of nested named entity recognition (NNER) has instead received less attention. This lack of breadth of research has been attributed to practical reasons (Finkel and Manning, 2009), including a lack of annotated corpora (Ringland et al., 2019).

Existing nested NE corpora span only a handful of languages and text domains. This is in stark contrast to resources for flat NER, which are available for at least up to 282 languages (Pan et al., 2017) and multiple domains, including a very recent effort (Liu et al., 2021). Existing NNER resources for English cover newswire (e.g., ACE, WSJ) (Mitchell et al., 2005; Ringland et al., 2019) and biomedical data (e.g., GENIA) (Kim et al., 2003; Alex et al., 2007; Pyysalo et al., 2007). Beyond English, there exist free and publicly available nested NER datasets. These include the GermEval 2014 dataset (Benikova et al., 2014a), which is one of the largest existing German NER resources covering largely news articles (Benikova et al., 2014b). Recently, the GermEval annotation guidelines inspired the creation of a Danish corpus (Plank et al., 2020). They added a layer of nested NER on top of the existing Danish Universal Dependency treebank (Johannsen et al., 2015). Both German and Danish corpora derive their annotation guidelines from the NoStA-D annotation scheme (Benikova et al., 2014b), which we adopt for EWT-NNER (Section 3.1). To facilitate research, a fine-grained nested NER annotation on top of the Penn Treebank WSJ has been released recently (Ringland et al., 2019). In contrast to ours, the WSJ NNER corpus spans 114 entity types and 6 layers, and includes numericals and time expressions beyond named entities. We instead focus on NEs with a total of 12 classes and 2 layers.

As outlined by Katiyar and Cardie (2018), nested named entities are attracting more research attention. Modeling solutions opt for diverse strategies, from hierarchical systems to graph-based methods and models based on linearization (Alex et al., 2007; Finkel and Manning, 2009; Sohrab and Miwa, 2018; Luan et al., 2019; Lin et al., 2019; Zheng et al., 2019; Straková et al., 2019; Shibuya and Hovy, 2020). The current top-performing neural systems use typically either a linearization, a multi-task learning or a graph-based approach (Straková et al., 2019; Plank et al., 2020; Yu et al., 2020). We evaluate two such methods.

English Web Treebank The English Web Treebank (EN-EWT) (Bies et al., 2012; Petrov and McDonald, 2012; Silveira et al., 2014) is a dataset introduced as part of the first workshop on Syntactic Analysis of Non-Canonical Language (SANCL). The advantage of EWT is that it spans over 200k tokens of texts from five web domains: Yahoo! answers, newsgroups, weblogs, local business reviews from Google and Enron emails. Gold annotations are available for several NLP tasks. The corpus was originally annotated for part-of-speech tags and constituency structure in Penn Treebank style (Bies et al., 2012). Gold standard dependency structures were annotated on EWT via the Universal Dependencies project (Silveira et al., 2014). Recently, efforts extend EWT (or parts thereof) to further semantic (Abend et al., 2020) and temporal layers (Vashishtha et al., 2019). We contribute a novel nested NER layer on top of the freely available UD EN-EWT corpus split (Silveira et al., 2014).

3 The EWT-NNER corpus

This section describes the corpus and annotation.

3.1 Annotation Scheme and Process

We depart from the NoSTA-D named entity annotation scheme (Benikova et al., 2014b), introduced in the GermEval 2014 shared task and adopted for Danish (Plank et al., 2020). The entity labels span a total of 12 classes, distributed over four major entities (Tjong Kim Sang and De Meulder, 2003): location (LOC), organization (ORG), person (PER) and miscellaneous (MISC). There are two further sub-types: '-part' and '-deriv'. Entities are annotated using a two-level scheme. First-level annotations contain largest entity spans (e.g., the 'Alaskan Knight'). Second-level annotations are nested entities. In particular:

- We annotate named entities with two layers. The outermost layer 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 further potential readings in favor of keeping two layers. Example: '[[UNSC]*ORG* Resolution 1559]*MISC*'[1]

[1] Benikova et al. (2014b) report a few cases (around 1 in 1,000 sentences) where the 2 levels do not suffice, but opted for the 2-layer scheme for simplicity. We follow this, observing a similar pattern (2 cases every 1,000 sentences). We kept notes of these cases, yet leave an investigation to future work.

- Full NEs are annotated as LOC (location), ORG (organization), PER (person) or MISC (miscellaneous other).

- Only full nominal phrases are potential full NEs. Pronouns and all other phrases are ignored. National holidays or religious events (*Christmas, Ramadan*) are also not annotated. Determiners and titles are not part of NEs.

- Named entities can also be part of tokens and are annotated as such with the suffix "-part". Example: '[Thailand-based]*LOCpart*', '[Nintendo-inspired]*ORGpart* costume'

- Derivations of NEs are marked via the suffix *deriv*, e.g., 'the [Alaskan]*LOCderiv* movie'.

- Geopolitical entities deserves special attention. We opted for annotating its first reading as ORG, with a secondary LOC reading, to reduce ambiguity. This is the same as in the Danish guidelines. The original German NoStA-D guidelines did not provide detailed guidelines for this case, yet mentions some categories were conflated (LOC and geopolicitcal entities). They seem most frequently annotated as LOC, yet we find also similar annotations in the German data (especially for multi-token NEs like *Borussia Dortmund*).

The full annotation guidelines for EWT-NNER with examples, annotation decisions and difficult cases can be found in the accompanying repository.[2] Two annotators were involved in the process, both contributed to the earlier Danish corpus. One annotator is an expert annotator with a degree in linguistics; the second annotator is a computer scientist. A data statement is provided in the appendix. Inter-annotator agreement (IAA) is measured on a random sample of 100 sentences drawn from the development data. This resulted in the following agreement statistics: raw token-level agreement of 97.8%, Cohen's kappa over all tokens 89.1% and Cohen's kappa of 83.1% for tokens taking part in an entity as marked by at least one annotator. The final dataset was annotated by the professional linguist annotator. The annotation took around 3 working days per 25,000 tokens.

	Train	Dev	Test
answers 💬	2,631	419	438
reviews 📝	2,724	554	535
email ✉	3,770	524	606
newsgroup 👥	1,833	274	284
weblogs 📰	1,585	231	214
total	12,543	2,002	2,077

Table 1: Number of sentences in the UD EWT split.

	All	Nest.	💬	📝	✉	👥	📰
Location	3,553	901	20.6%	9.7%	15.2%	22.2%	32.3%
LOC deriv	1,094	161	18.1%	7.7%	3.4%	21.7%	49.1%
LOC part	76	29	7.9%	5.3%	14.5%	18.4%	53.9%
Person	4,202	353	4.5%	9.0%	45.0%	18.2%	23.3%
PER deriv	22	2	0%	4.8%	38.1%	19%	38.1%
PER part	24	2	0%	4.2%	16.7%	58.3%	20.8%
Organization	3,309	133	9.8%	15.2%	25.1%	21.0%	29.0%
ORG deriv	32	0	15.2%	9.1%	27.3%	12.1%	36.4%
ORG part	45	45	10.3%	0%	62.1%	0%	27.6%
Miscellaneous	1,576	11	21.2%	8.4%	29.7%	29.0%	11.8%
MISC deriv	3	0	0%	33.3%	33.3%	33.3%	0%
MISC part	9	2	0%	0%	57.1%	28.6%	14.3%
total	13,945	1,622	% of all entities per domain				

Table 2: Distribution of entities in all of EWT-NNER (16k sentences): All, Nested, and % of All.

3.2 Data statistics

Statistics over the data split and distribution of the web texts are provided in Table 1. A comparison to German and Danish on coarse-level statistics are provided in Table 3. Details of the entity distibution in EWT-NNER are given in Table 2. The entire EWT-NNER contains a total of over 16,000 sentences and over 13,000 entities. Around 42% of the sentences contain NEs. Over 11.6% are nested NEs, 8.3% are derivations and 1.1% are parts of names. Compared to GermEval 2014, this is a higher density of nested entities (11.6% vs 7.7% in the German data), yet a lower percentage of derivations and partial NEs. The data is provided in CoNLL tabular format with BIO entity encoding.

4 Experimental Setup

We are interested in a set of benchmark results to provide: a) zero-shot transfer results from Danish and German; b) in-language results (training on all 5 EN domains vs per-domain models); and c) results on cross-lingual cross-domain evaluation when training on multiple languages jointly.

For the experiments, we use fine-tuning of con-

[2] Data and tech report (Plank and Dam Sonniks, 2021) are available at: `http://github.com/bplank/nested-ner`

	Training			Development			Test
Language	German	Danish	English	German	Danish	English	English
Domain	news	news	web (5)	news	news	web (5)	web (5)
Sentences	24,002	4,383	12,538	2,200	564	2,002	2,077
Tokens	452,853	80,378	204,609	41,653	10,332	25,150	25,097
# Entities	31,545	3,035	10,673	2,886	504	1,549	1,711

Table 3: Overview of EWT-NNER (this paper) and comparison to existing nested NER datasets adopting the NoSTA-D 2-level NER annotation scheme: German (Benikova et al., 2014b) and Danish (Plank et al., 2020).

textualized embeddings with AllenNLP (Gardner et al., 2018) using the MaChAmp toolkit (van der Goot et al., 2021). We use the proposed default parameters. Whenever we train on English data, we took the smallest weblogs domain (231 sentences) as model selection set and assume no further in-domain dev set. For German and Danish, we use the provided news dev sets. For all experiments we report the average performance over 3 runs.

As contextualized embeddings, we investigate BERT (Devlin et al., 2019), multilingual BERT and XLM-R (Conneau et al., 2020). We evaluated two decoding strategies: the first takes the Cartesian product of inner and outer NER layer and treats it as a standard single-label decoding strategy. An advantage of this strategy is that any sequence tagging framework can be used; a disadvantage is the increased label space. To tackle this we use a two-headed multi-task decoder, one for each entity layer, as found effective (Plank et al., 2020). Initial experiments confirmed that the single-label decoding is less accurate, confirming earlier findings (Straková et al., 2019; Plank et al., 2020). We report results with the two-headed decoder only, and further results in the appendix.

Evaluation is based on the official GermEval 2014 (Benikova et al., 2014b) metric and script, i.e., strict span-based F1 over both entity levels.

5 Results

Table 4 shows the results of training models on German (DE: GermEval 2014), Danish (DA: DaN+), and their union (+), for zero-shot transfer (top rows). It provides further results of training on all English EWT-NNER training data (from all five web domains) both for multilingual models (using multilingual BERT or XLM-R) and monolingual models (English BERT and Roberta). Figure 2 provides cross-domain results of training only on English subdomains.

Language	DE	DA	EN (EWT-NNER)				
Domain	news	news	💬	📝	✉	👥	📰
# Entities	2,886	504	285	174	340	354	396
			zero-shot:				
DA	68.47	80.95	52.43	53.09	54.99	48.92	62.66
DE	84.41	76.67	59.04	55.93	60.53	59.11	63.23
DA+DE	**85.32**	82.39	61.12	56.86	61.98	60.48	64.43
			full in-language data:				
EN (all 5)	69.98	75.22	71.00	73.35	79.33	81.20	86.80
EN+DA	71.53	82.77	73.04	74.05	80.48	80.95	86.69
EN+DE	84.91	77.71	71.79	73.70	80.53	79.64	85.48
EN+DA+DE	84.80	82.43	72.61	74.45	79.39	**81.92**	86.59
EN+DA+DE$_X$	83.47	**83.07**	**73.86**	74.35	79.60	79.03	86.56
			full in-language, monolingual embeds:				
EN BERT	26.59	27.54	70.53	73.25	**81.09**	81.31	**87.04**
EN Roberta	34.35	33.37	70.01	**78.77**	81.02	77.44	84.58

Table 4: F1 scores on dev sets with mBERT/X[LM-R] (upper columns) and monolingual models for English.

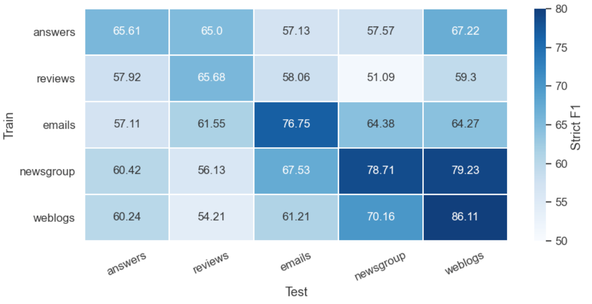

Figure 2: In-language cross-domain evaluation.

Take-aways While zero-shot transfer between news (on German and Danish) is around 70 F1 (68.5 and 76.7), zero-shot transfer to the EWT-NNER web domains is low, particularly for answers (💬), reviews (📝), emails (✉) and newsgroups (👥). Training on both Danish and German improves zero-shot performance over all domains.

For English cross-domain evaluation, we observe a large variation across domains in Figure 2. Here, we train models on the EWT-NNER training portion of a single web domain, and evaluate the resulting model across all five web domains (in-domain and out-domain). The heatmap confirms that training within domain is the most beneficial (results on the diagonal), but large drops can be observed across domains. Reviews (📝) and Yahoo! answers (💬) remain the most challenging with the lowest F1 scores. Weblogs (📰) shows the highest results. We tentatively attribute this to the good coverage of weblogs over all entity classes (see Table 2) and the well-edited style of the text (by inspection many posts are about politics and military events). If we compare the results to the model trained on all English data in Table 4 (EN all 5), we observe that training on all web training data improves over the single web texts.

We investigate cross-lingual cross-domain results, to evaluate whether a model trained on English data alone can be improved by further cross-lingual transfer. Table 4 shows that this is the case. There is positive transfer from German and Danish data, with the mBERT model (EN+DA+DE) boosting performance (on most domains). The larger XLM-R model helps on specific domains, but it is not consistently better than mBERT.

So far we focused on multilingual contextualized embeddings. The last rows in Table 4 compares the multilingual models to monolingual ones. Interestingly, in this domain a monolingual model does not consistently outperform the multlingual model. While for some domains the EN model is substantially better, this is not the case overall. On average over the 5 web domains, the tri-lingual model with mBERT reaches a slightly overall F1 (average of 78.99), followed by both the monolingual BERT model (78.64) and XLM-R (78.63).

Language	DE	DA	EN EWT-NNER					
Domain	news	news	💬	📝	✉	👥	📰	Avg
Entities	6,693	566	316	167	465	314	449	
EN	70.10	72.68	73.47	68.47	75.93	82.66	88.87	77.88
EN+DA+DE	84.36	79.94	74.28	66.54	77.44	84.10	87.71	**78.01**

Table 5: F1 scores on the test sets with mBERT.

Test sets Finally, we run the best model on the test sets and compare to training on English alone. Table 5 confirm the overall trends. There is a positive transfer across languages for cross-domain evaluation, with improvements on the majority of domains. The best model reaches an average F1 score of 78.01 on the five web domains. Compared to results within newswire, there is room to improve NNER over domains.

Analysis We perform a qualitative analysis of the best model (EN+DA+DE) on the dev sets. Detailed scores are in Table 7 in the appendix. The overall F1 of 72% on 💬 is largely due to low recall on person names (recall 63%) (e.g., peculiar names such as 'Crazy Horse', a Dakota leader) and missed lower-cased product names ('ipod'). On ✉, recall on ORG and LOC is low (55% and 65%), as organizations and locations are missed also due to unconventional spelling in emails. In reviews (📝), the model reaches its lowest F1 on ORG (67%) as it mixes up people names with organizations and lacks recall. Newsgroup (👥) is broad (e.g., discussions from astronomy to cat albums) with the lowest per-entity F1 of 75% for MISC. Newsgroup and weblogs are the domains with the most LOCderiv entities, which the model easily identifies (F1 of 93% and 99% in 👥 and 📰, respectively). Overall, weblogs (📰) has the highest per-entity F1 scores, all above 75%, with the highest overall F1 on LOC (92 F1; in comparison to 57% on ✉ and 79% on 💬). This high result on weblogs can be further attributed to smaller distance to the training sources (as indicated in the overlap plot in Figure 1) and to some degree of using this domain for tuning. From a qualitative look, we note that the weblogs sample is rather clean text, often in reporting style about political events similar to edited news texts, which we believe is part of the reason for the high performance compared to the other domains in EWT-NNER.

6 Conclusions

We present EWT-NNER, a nested NER dataset for English web texts, to contribute to a limiting nested NER resource landscape. We outline the dataset, annotation guidelines and benchmark results. The results show that NNER remains challenging on web texts, and cross-lingual transfer helps. We hope this dataset encourages research on cross-lingual cross-domain NNER. There are many avenues for future research, which include e.g., alternative decoding (Yu et al., 2020), pre-training models and adaptation (Gururangan et al., 2020).

Acknowledgments

The author would like to thank the three anonymous reviewers for their valuable comments to improve this paper. This research is supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B and 9063-00077B.

References

Omri Abend, Dotan Dvir, Daniel Hershcovich, Jakob Prange, and Nathan Schneider. 2020. Cross-lingual semantic representation for NLP with UCCA. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 1–9, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. 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.

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, Max Kisselew, and Sebastian Pado. 2014a. Germeval 2014 named entity recognition shared task: Companion paper.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ann Bies, Justing Mott, Colin Warner, and Seth Kulick. 2012. English web treebank ldc2012t13. In *Web Download. Philadelphia: Linguistic Data Consortium, 2012*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Ralph Grishman. 1998. Research in information extraction: 1996-98. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 57–60, Baltimore, Maryland, USA. 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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

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. 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.

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. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating cross-domain named entity recognition. In *To appear at AAAI*.

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. 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.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Proceedings of the SANCL shared task workshop*.

Barbara Plank and Sif Dam Sonniks. 2021. Annotation guidelines for Nested Named Entities for English. Technical report, IT University of Copenhagen.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sampo Pyysalo, Filip Ginter, Veronika Laippala, Katri Haverinen, Juho Heimonen, and Tapio Salakoski. 2007. On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. In *Biological, translational, and clinical language processing*, pages 25–32, Prague, Czech Republic. 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. Association for Computational Linguistics.

Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Jana Straková, Milan Straka, and Jan Hajic. 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. Association for Computational Linguistics.

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*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. 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. Association for Computational Linguistics.

A Data Statement

This following data statement (Bender and Friedman, 2018) documents the origin of the data annotations and provenance of the original English Web Treebank (EWT) data.

CURATION RATIONALE Annotation of nested named entities (NNE) in web text domains to study the impact of domain gap on cross-lingual transfer.

LANGUAGE VARIETY Mostly US (en-US) mainstream English as target. Transfer from Danish (da-DK) and German (de-DE).

SPEAKER DEMOGRAPHIC Unknown.

ANNOTATOR DEMOGRAPHIC Native languages: Danish, German. Socioeconomic status: higher-education student and university faculty.

SPEECH SITUATION Scripted, spontaneous.

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

PROVENANCE APPENDIX The data originates from the English Web Treebank (EN-EWT) (Bies et al., 2012; Petrov and McDonald, 2012; Silveira et al., 2014) and data split available at: `https://github.com/UniversalDependencies/UD_English-EWT/`

B Additional results

Table 6 provides additional results for both decoding strategies. It shows that single-label decoding is outperformed by the two-head decoder, confirming similar results on Danish (Plank et al., 2020).

Language	DE	DA	EN EWT-NNER				
Domain	news	news	answ	revs	email	nwsgrp	weblg
# Entities	6,693	566	321	168	468	319	447
	german	news	answers	reviews	email	newsgroup	weblogs
en.bert.single-merged	**29.33**	24.68	67.16	71.57	79.23	75.12	79.81
en.bert.multitask	26.59	**27.54**	**70.53**	**73.25**	**81.09**	**81.31**	**87.04**
en.roberta.single-merged	31.41	24.86	67.09	72.07	81.01	72.35	79.34
en.roberta.multitask	**34.35**	**33.37**	**70.01**	**78.77**	**81.02**	**77.44**	**84.58**

Table 6: F1 scores on the dev set with monolingual models both decoding strategies.

	answers 💬	reviews 📝	email ✉	newsgroups 👥	weblogs 📰
Location	79.01 (72.73)	85.37 (89.74)	57.14 (55.17)	83.76 (92.45)	92.31 (93.10)
LOC deriv	74.42 (84.21)	100.00 (100.00)	0.00 (0.00)	93.33 (100.00)	99.07 (98.15)
LOC part	–	100.00 (100.00)	–	–	100.00 (100.00)
Person	72.73 (63.16)	83.33 (87.50)	90.25 (86.63)	92.09 (98.46)	91.46 (88.24)
PER deriv	–	–	100.00 (100.00)	0.00 (0.00)	0.00 (0.00)
PER part	–	–	–	0.00 (0.00)	–
Organization	64.58 (70.45)	67.42 (60.00)	64.86 (65.45)	83.64 (84.15)	85.83 (83.06)
ORG deriv	0.00 (0.00)	–	–	–	0.00 (0.00)
ORG part	–	–	–	0.00 (0.00)	–
Miscellaneous	75.76 (67.57)	85.71 (81.82)	80.00 (83.64)	75.82 (65.91)	75.00 (84.00)
MISC deriv	–	–	–	–	–
MISC part	–	–	–	–	0.00 (0.00)

Table 7: Per-entity evaluation of outer level strict FB1 score (and recall) of the best model EN+DE+DA with mBERT on the dev sets.