NNE: A Dataset for Nested Named Entity Recognition in English Newswire

Nicky Ringland[1]† **Xiang Dai**[1,2]‡ **Ben Hachey**[1,3] **Sarvnaz Karimi**[2]
Cecile Paris[2] **James R. Curran**[1]
[1]University of Sydney, Sydney, Australia
[2]CSIRO Data61, Sydney, Australia
[3]Digital Health CRC, Sydney, Australia
†nicky.ringland@sydney.edu.au ‡dai.dai@csiro.au

Abstract

Named entity recognition (NER) is widely used in natural language processing applications and downstream tasks. However, most NER tools target flat annotation from popular datasets, eschewing the semantic information available in nested entity mentions. We describe NNE—a fine-grained, nested named entity dataset over the full Wall Street Journal portion of the Penn Treebank (PTB). Our annotation comprises 279,795 mentions of 114 entity types with up to 6 layers of nesting. We hope the public release of this large dataset for English newswire will encourage development of new techniques for nested NER.

1 Introduction

Named entity recognition—the task of identifying and classifying entity mentions in text—plays a crucial role in understanding natural language. It is used for many downstream language processing tasks, e.g., coreference resolution, question answering, summarization, entity linking, relation extraction and knowledge base population. However, most NER tools are designed to capture flat mention structure over coarse entity type schemas, reflecting the available annotated datasets.

Focusing on flat mention structures ignores important information that can be useful for downstream tasks. Figure 1 includes examples of nested named entities illustrating several phenomena:

- Entity-entity relationships can be embedded in nested mentions. For instance, the *location* of the *'Ontario Supreme Court'* is indicated by the embedded STATE mention *'Ontario'*;
- Entity attribute values can be embedded in nested mentions. For instance, the *title* is the embedded ROLE *'Former U.N. Ambassador'*, which also encodes the employment relation between the PERSON *'Jane Kirkpatrick'* and ORG *'U.N.'*;
- Part-whole relationships can be encoded in nested mention structure. For instance, the REGION *'Southern California'* is part of the STATE *'California'*.

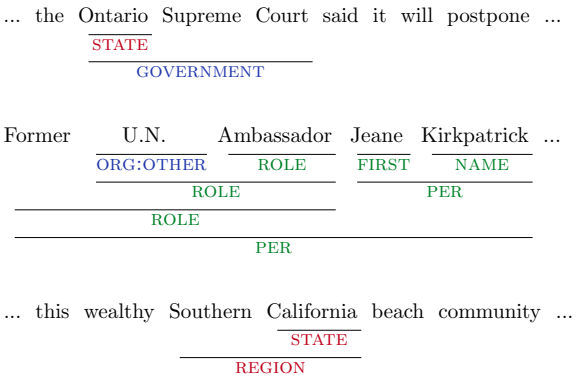

Figure 1: Example nested mentions in NNE.

Recent work has demonstrated increasing interest in nested entity structure, including local approaches (Xu et al., 2017; Sohrab and Miwa, 2018), hypergraph-based approaches (Lu and Roth, 2015; Muis and Lu, 2017; Katiyar and Cardie, 2018; Wang and Lu, 2018), cascaded approaches (Alex et al., 2007; Ju et al., 2018), and parsing approaches (Finkel and Manning, 2009; Wang et al., 2018). See Dai (2018) for a survey. Yet these techniques have seen little translation from the research literature to toolsets or downstream applications.

To facilitate ongoing research on nested NER, we introduce NNE—a large, manually-annotated, nested named entity dataset over English newswire. This new annotation layer over the Wall Street Journal portion of the PTB includes 279,795 mentions. All mentions are annotated, including nested structures with depth as high as

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5176–5181
Florence, Italy, July 28 - August 2, 2019.

six layers. A fine-grained entity type schema is used, extending the flat BBN (Weischedel and Brunstein, 2005) annotation from 64 to 114 entity types.

We are publicly releasing the standoff annotations along with detailed annotation guidelines and scripts for knitting annotations onto the underlying PTB corpus.[1] Benchmark results using recent state-of-the-art approaches demonstrate that good accuracy is possible, but complexity and run time are open challenges. As a new layer over the already rich collection of PTB annotations, NNE provides an opportunity to explore joint modelling of nested NER and other tasks at an unprecedented scale and detail.

2 The NNE dataset

Annotation Scheme: BBN (Weischedel and Brunstein, 2005) is a pronoun coreference and entity type corpus, annotated with 64 types of entities, numerical and time expressions. We use its flat entity schema as a starting point to design our schema. We analyzed existing BBN annotations to develop and automatically apply structured pre-annotation for predictable entity types. Additional fine-grained categories and further structural elements of entities, inspired by Sekine et al. (2002) and Nothman et al. (2013), are used to augment the BBN schema. We adhere to the following general principles when annotating nested named entities in the corpus:

- Annotate all named entities, all time and date (TIMEX) and numerical (NUMEX) entities, including all non-sentence initial words in title case, and instances of proper noun mentions that are not capitalized.

- Annotate all structural elements of entities. These elements could be other entities, such as *'Ontario'* (STATE) in *'Ontario Supreme Court'* (GOVERNMENT), or structural components such as *'40'* (CARDINAL) and *'miles'* (UNIT) in *'40 miles'* (QUANTITY:1D), as well as the internal structure induced by syntactic elements, such as coordination.

- Add consistent substructure to avoid spurious ambiguity. For example, the token *'Toronto'*, which is a CITY, would be labeled as part of an ORG:EDU organization span *'University of Toronto'*. We add layers of annotations to allow each token to be annotated as consistently as possible, e.g., [University of [Toronto]$_{\text{CITY}}$]$_{\text{ORG:EDU}}$.

- Add additional categories to avoid category confusion. Some entities are easy to identify, but difficult to categorize consistently. For instance, a hotel (or any business at a fixed location) has both organizational and locative qualities, or is at least treated metonymously as a location. Rather than requiring annotators to make an ambiguous decision, we elect to add category HOTEL to simplify the individual annotation decision. We also apply this principle when adding MEDIA, FUND, and BUILDING categories.

- Pragmatic annotation. Many annotation decisions are ambiguous and difficult, thus may require substantial research. For instance, knowing that *'The Boeing Company'* was named after founder *'William E. Boeing'* would allow us to annotate *'Boeing'* with an embedded PERSON entity. However, this does not apply for other companies, such as *'Sony Corporation'*. To let annotation decisions be made without reference to external knowledge, we label all tokens that seem to be the names of people as NAME, regardless of whether they are actually a person's name.

Entity types and mention frequencies can be found in Appendix A. See Ringland (2016) for annotation guidelines and extended discussion of annotation decisions.

Annotation Process: Although some existing annotation tools allow nested structures (e.g., Brat (Stenetorp et al., 2012)), we built a custom tool that allowed us to create a simple and fast way to add layers of entities, and suggest reusing existing structured annotations for the same span.

Using the annotations from BBN as underlying annotations, the annotator is shown a screen with the target sentence, as well as the previous and next sentences, if any. A view of the whole article is also possible to help the annotator with contextual cues. When annotators select a span, they are prompted with suggestions based on their own previous annotations, and common entities. Some entities are repeated frequently in an article,

[1] https://github.com/nickyringland/nested_named_entities

Depth	Number	%	Three most frequent categories
1	118,525	45.5	CORP (22,752), DATE (15,927), PER (13,460)
2	106,144	40.8	CARDINAL (19,834), NAME (18,640), UNIT (14,871)
3	31,573	12.1	CARDINAL (11,697), MULT (5,859), NAME (3,450)
4	3,813	1.5	CARDINAL (1,650), MULT (1,041), UNIT (400)
5	327	0.1	CARDINAL (154), MULT (96), UNIT (51)
6	4	0.0	UNIT (1), CITY-STATE (1), MULT (1)

Table 1: Number of spans at each layer of nesting with their most frequent categories.

or over many articles in the corpus. The annotation tool allows a user to add a specified annotation to all strings matching those tokens in the same article, or in all articles.

Four annotators, each with a background in linguistics and/or computational linguistics were selected and briefed on the annotation task and purpose. The WSJ portion of the PTB consists of 25 sections (00–24). Each annotator started with a subset of section 00 as annotation training, and was given feedback before moving on to other sections. Weekly meetings were held with all annotators to discuss ambiguities in the guidelines, gaps in the annotation categories, edge cases and ambiguous entities and to resolve discrepancies.

Total annotation time for the corpus was 270 hours, split between the four annotators. Sections 00 and 23 were doubly annotated, and section 02 was annotated by all four annotators. An additional 17 hours was used for adjudicating these sections annotated by multiple annotators.

Dataset Analysis: The resulting NNE dataset includes a large number of entity mentions of substantial depth, with more than half of mentions occurring inside another mentions. Of the 118,525 top-level entity mentions, 47,020 (39.6%) do not have any nested structure embedded. The remaining 71,505 mentions contain 161,270 mentions, averaging 2.25 structural mentions per each of these top-layer entity mentions. Note that one span can be assigned multiple entity types. For example, the span *'1993'* can be annotated as both DATE and YEAR. In NNE, 19,144 out of 260,386 total spans are assigned multiple types. Table 1 lists the number of spans occurring at each depth. To measure how clearly the annotation guidelines delineate each category, and how reliable our annotations are, inter-annotator agreement was calculated using annotations on Section 02, which was annotated by all four annotators. An adjudicated version was created by deciding a correct existing candidate label from within the four possibilities, or by adjusting one of them on a token level. For the purposes of inter-annotator agreement, a *tag stack* is calculated for each word, essentially flattening each token's nested annotation structure into one label. For example, the tag of token *'California'* in the third sentence of Figure 1 is STATE_REGION, while *'beach'* is *O_O*. Agreement using Fleiss' kappa over all tokens is 0.907. Considering only tokens that are part of at least one mention according to at least one annotator, Fleiss' kappa is 0.832. Both results are above the 0.8 threshold for good reliability (Carletta, 1996). Average precision, recall and F_1 score across four annotators with respect to the adjudicated gold standard are 94.3, 91.8 and 93.0.

3 Benchmark results

We evaluate three existing NER models on our dataset: (1) the standard BiLSTM-CRF model which can handle only flat entities (Lample et al., 2016); (2) hypergraph-based (Wang and Lu, 2018); and, (3) transition-based (Wang et al., 2018) models. The latter two models were proposed to recognize nested mentions. We follow CoNLL evaluation schema in requiring an exact match of mention start, end and entity type (Sang and Meulder, 2003). We use sections 02 as development set, sections 23 and 24 as test set, and the remaining sections as training set. The model that performs best on the development set is evaluated on the test set for the final result. Since the standard BiLSTM-CRF model cannot handle nested entities, we use either the outermost (BiLSTM-CRF-TOP in Table 2) or the innermost mentions (BiLSTM-CRF-BOTTOM) for training. We also combine the outputs from these two flat NER models, and denote the result as BiLSTM-CRF-BOTH.

From Table 2, we can see that single flat NER models can achieve high precision but suffer from low recall. For example, the model pretrained on outermost (top) mentions has 38.0 recall, as

	P	R	F_1
BiLSTM-CRF-TOP	89.9	38.0	53.5
BiLSTM-CRF-BOTTOM	93.8	62.0	74.7
BiLSTM-CRF-BOTH	92.2	85.8	88.9
Hypergraph	91.8	91.0	91.4
Transition	77.4	70.1	73.6

Table 2: NER results on NNE using different methods.

around 60% of mentions are nested within others. The hypergraph-based model performs best on our dataset, presumably because it can capture mentions from different levels and does not suffer from issues of *structural ambiguity* during inference (Muis and Lu, 2017; Wang and Lu, 2018). However, its decoding speed of 9 words per second is slow due to the large number of entity categories of our dataset.[2] The transition-based method has a higher decode speed of 57 words per second, but has much lower precision than flat NER models.

4 Related Work

Other corpora with nested entities: We briefly compare existing annotated English corpora involving nested entities. A comparison of statistics between our dataset and two widely used benchmark datasets is shown in Table 3. The ACE corpora (Mitchell et al., 2004; Walker et al., 2005) consist of data of various types annotated for entities, relations and events. The entity component of ACE is framed in terms of nominal modification, and nested mentions are only annotated in nominal mentions, not inside other named entity mentions. For example, in ACE2005, *'Secretary of Homeland Security Tom Ridge'* is annotated as a PERSON, containing two other PERSON annotations: *'Secretary'* and *'Secretary of Homeland Security'*. In contrast, our annotations capture more interactions between different semantic spans: PERSON consisting of ROLE and NAME, and ROLE containing GOVERNMENT.

The GENIA corpus (Kim et al., 2003) is a richly-annotated corpus for bio-text mining that has 36 entity types among 2,000 MEDLINE abstracts. Due to the biomedical domain's specialized terminology and complex naming conventions, entities of interest, such as genes, proteins or disease names, often nest. For example, the RNA *'CIITA mRNA'* contains a DNA mention *'CIITA'*.

[2]The decoding time complexity of the method proposed by Wang and Lu (2018) is $O(cmn)$, where m is the number of entity types, n is the sentence length, and c is the maximal mention length.

Item	NNE	GENIA	ACE2005
Documents	2,312	2,000	464
Sentences	49,208	18,546	12,548
Sentences w. nesting	32,387	9,533	4,266
Tokens	1.1M	0.5M	0.3M
Mentions	279,795	92,681	30,966
Entity types	114	36	7
Mentions per sentence	5.69	4.99	2.46
Top-level mentions	118,525	76,582	23,464
Maximum depth	6	4	6

Table 3: A comparison between NNE and two commonly used corpora with nested entities.

In addition to these two commonly used nested entity corpora, Byrne (2007) and Alex et al. (2007) introduced datasets with nested entities in historical archive and biomedical domains, respectively. However, their datasets are not publicly available. Four percent of entity mentions annotated in the English entity discovery and linking task in TAC-KBP track include nesting (Ji et al., 2014).

Resources built on the PTB: A lots of effort has been made on adding syntactic and semantic information to the PTB (Marcus et al., 1993). PropBank (Kingsbury et al., 2002) extended the PTB with the predicate argument relationships between verbs and their arguments. NomBank (Meyers et al., 2004) extended the argument structure for instances of common nouns. Vadas and Curran (2007), and Ficler and Goldberg (2016) extended the PTB with noun phrase and coordination annotations, respectively.

Our dataset is built on top of the PTB and enriches the full ecosystem of resources and systems that stem from it.

5 Summary

We present NNE, a large-scale, nested, fine-grained named entity dataset. We are optimistic that NNE will encourage the development of new NER models that recognize structural information within entities, and therefore understand fine-grained semantic information captured. Additionally, our annotations are built on top of the PTB, so that the NNE dataset will allow joint learning models to take advantage of semantic and syntactic annotations, and ultimately to understand and exploit the true structure of named entities.

Acknowledgments

We would like to thank annotators for their excellent work: Kellie Webster, Vivian Li, Joanne Yang and Kristy Hughes. We also thank three anonymous reviewers for their insightful comments.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *BioNLP*, pages 65–72.

Kate Byrne. 2007. Nested named entity recognition in historical archive text. In *ICSC*, pages 589–596.

Jean Carletta. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *ACL-SRW*, pages 37–44.

Jessica Ficler and Yoav Goldberg. 2016. Coordination annotation extension in the Penn tree bank. In *ACL*, pages 834–842.

Jenny Rose Finkel and Christopher Manning. 2009. Nested named entity recognition. In *EMNLP*, pages 141–150.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 entity discovery and linking tasks. In *TAC*, pages 1333–1339.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *NAACL*, pages 1446–1459.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *NAACL*, pages 861–871.

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

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding predicate argument structure to the Penn Treebank. In *HLT*, pages 252–256.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*, pages 260–270.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *EMNLP*, pages 857–867.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.*, 19.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2004. ACE 2004 multilingual training corpus. *LDC*.

Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *EMNLP*, pages 2608–2618.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.*, 194:151–175.

Nicky Ringland. 2016. *Structured Named Entities*. Ph.D. thesis, University of Sydney.

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

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *LREC*.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *EMNLP*, pages 2843–2849.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *EACL*, pages 102–107.

David Vadas and James Curran. 2007. Adding noun phrase structure to the Penn Treebank. In *ACL*, pages 240–247.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. ACE 2005 multilingual training corpus. *LDC*.

Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *EMNLP*, pages 204–214.

Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018. A neural transition-based model for nested mention recognition. In *EMNLP*, pages 1011–1017.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. *LDC*.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *ACL*, pages 1237–1247.

A Full annotation scheme

Category	Frequency	Category	Frequency	Category	Frequency
CARDINAL	43873	STREET	475	QUANTITY2D	81
NAME	28537	GRPORG	437	PRODUCTFOOD	80
ORGCORP	23339	ORGPOLITICAL	436	SUBURB	78
UNIT	19289	VEHICLE	432	GRPLOC	63
DATE	17381	LAW	419	HOTEL	55
PER	14960	ORGEDU	411	QUANTITYOTHER	55
DURATION	13655	CONTINENT	354	FUND	54
MONEY	12640	BUILDING	346	SONG	54
MULT	7851	SEASON	337	SPACE	53
FIRST	6797	GPE	333	RIVER	52
CITY	6723	FOLD	313	WAR	51
PERCENT	6542	MIDDLE	313	CHEMICAL	45
REL	6170	TIME	296	BRIDGE	44
CORPJARGON	5560	WEIGHT	293	PLAY	42
HON	5524	OCEAN	291	STADIUM	37
NATIONALITY	5193	LOCATIONOTHER	261	AWARD	36
GOVERNMENT	4674	EVENT	260	ORGRELIGIOUS	35
COUNTRY	4047	DISEASE	246	AIRPORT	32
QUAL	3903	QUANTITY1D	220	ANIMATE	29
YEAR	3421	CITYSTATE	220	GOD	29
MONTH	3385	WOA	207	HOSPITAL	25
STATE	3245	TVSHOW	172	ATTRACTION	24
ORDINAL	2590	ELECTRONICS	167	WEAPON	23
IPOINTS	2395	SPORTSTEAM	166	MUSEUM	17
ROLE	2368	DATEOTHER	164	ENERGY	17
RATE	2141	QUANTITY3D	156	SPEED	14
MEDIA	1712	NAMEMOD	155	PAINTING	13
DAY	1631	GRPPER	154	BAND	10
NUMDAY	1495	BOOK	149	SPORTSSEASON	8
INI	1445	ARMY	139	SCINAME	7
NORPOTHER	1247	FACILITY	129	ADDRESSNON	3
ORGOTHER	1099	PRODUCTDRUG	116	ALBUM	3
PERIODIC	1066	HURRICANE	107	TEMPERATURE	2
REGION	864	SPORTSEVENT	100	NATURALDISASTER	2
NORPPOLITICAL	731	RELIGION	99	CONCERT	2
AGE	661	NICKNAME	96	STATION	1
INDEX	657	LANGUAGE	92	BORDER	1
PRODUCTOTHER	656	FILM	89	CHANNEL	1