E-NER - An Annotated Named Entity Recognition Corpus of Legal Text

Au, Ting Wai Terence; Lampos, Vasileios; Cox, Ingemar J.

Published in:
NLLP 2022 - Natural Legal Language Processing Workshop 2022, Proceedings of the Workshop

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):

Download date: 24. nov., 2023
E-NER — An Annotated Named Entity Recognition Corpus of Legal Text

Ting Wai Terence Au¹, Vasileios Lampos¹ and Ingemar J. Cox¹,²

¹ Centre for Artificial Intelligence, Department of Computer Science, University College London, UK
² Department of Computer Science, University of Copenhagen, Denmark

{ting.au.19, v.lampos}@ucl.ac.uk, ingemar@ieee.org

Abstract

Identifying named entities such as a person, location or organization, in documents can highlight key information to readers. Training Named Entity Recognition (NER) models requires an annotated data set, which can be a time-consuming labour-intensive task. Nevertheless, there are publicly available NER data sets for general English. Recently there has been interest in developing NER for legal text. However, prior work and experimental results reported here indicate that there is a significant degradation in performance when NER methods trained on a general English data set are applied to legal text. We describe a publicly available legal NER data set, called E-NER, based on legal company filings available from the US Securities and Exchange Commission’s EDGAR data set. Training a number of different NER algorithms on the general English CoNLL-2003 corpus but testing on our test collection confirmed significant degradations in accuracy, as measured by the F1-score, of between 29.4% and 60.4%, compared to training and testing on the E-NER collection.

1 Introduction

Named Entity Recognition (NER) aims to identify names of specific objects in the world (mostly nouns with few exceptions), such as the name of a person, location and organization, which indicate possibly important phrases that readers should pay attention to. NER has been used in a variety of downstream tasks such as question answering (Mollá et al., 2006), document de-identification (Stubbs et al., 2015; Catelli et al., 2020), relation extraction (Miwa and Bansal, 2016), and machine translation (Babych and Hartley, 2003). Consequently, there has been considerable work on NER using general language corpora (Yadav and Bethard, 2018; Li et al., 2020a) and a variety of test collections are available. Previous work has examined domain-specific NER, e.g. in finance (Alvarado et al., 2015; Alexander and de Vries, 2021; Zhang and Zhang, 2022), biomedical (Zhou et al., 2004; Wang et al., 2018), online user-generated content (Tran et al., 2015; Li et al., 2014), and legal (Luz de Araujo et al., 2018) applications, and found that the performance of domain-specific NER systems was poor if trained on general language corpora. Constructing test collections for specialist domains can be a time consuming task requiring manual annotation of a corpus. To reduce this effort there has been considerable recent interest in transfer learning, such as pre-trained language models (Brown et al., 2020; Howard and Ruder, 2018). Nevertheless, there remains a need for specialist test collections whether for training or fine-tuning.

Legal text is one specialist domain where NER is of interest, due to its usefulness in assisting other legal tasks such as record linkage (Dozier et al., 2010), court case linkage (Kríž et al., 2014), contract analysis (Chalkidis et al., 2017), prediction of judicial decisions (Aletras et al., 2016), credit risk assessment (Alvarado et al., 2015), and question-answering systems (Jayakumar et al., 2020). However, despite increasing interest in this sub-domain, there is no publicly available corpus for the evaluation of NER methods for legal applications.

This paper describes E-NER, an annotated NER collection of legal documents,¹ based on publicly available legal company filings in the United States Securities and Exchange Commissions’ EDGAR database. Overall, we deployed four NER models to compare classification performance when (i) trained and tested on general English, (ii) trained on general English and tested on E-NER, and (iii) trained and tested on E-NER. The results support insights from earlier work, i.e. we observed significant performance degradation when trained on general English but tested on legal text. Our experiments justify the utility of a domain-specific (legal)

¹E-NER data set, github.com/terenceau2/E-NER-Dataset
NER corpus.

2 Related work

The primary contribution of this paper is a legal-English test collection for NER. We do not propose a new algorithm for NER and consequently restrict our description of NER methods to those used in our experimental work.

Hidden Markov models (HMM) (Rabiner and Juang, 1986) can be used to label sequences. Bikel et al. (1997) demonstrated the application of HMM to NER. Conditional Random Fields (CRF) (Lafferty et al., 2001) is another sequence labelling model which improves on HMM, by relaxing the stationarity and the output independence assumptions. McCallum and Li (2003) and Sobhana et al. (2010) demonstrated the application of CRF to NER.

In more recent years, pre-trained language models (Qiu et al., 2020) and prompt-based learning (Liu et al., 2022) have demonstrated superior performance. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a pre-trained language model which is based on transformers (Vaswani et al., 2017). BERT pre-trains on a large corpus of non annotated text, performing self-supervised tasks, namely masked word prediction and next sentence pairing. BERT can facilitate transfer learning: the model parameters from the pre-training step are used during the fine-tuning step, in order for the model to learn downstream tasks such as NER (Souza et al., 2019; Hakala and Pyysalo, 2019; Li et al., 2020b).

There exist publicly available annotated NER data sets for general English text, such as CoNLL-2003 (Sang and De Meulder, 2003), WNUT17 (Derczynski et al., 2017), and the Wikipedia gold standard corpus (Balasuriya et al., 2009), as well as for other languages (Neudecker, 2016; Sang and De Meulder, 2003; Santos et al., 2006; Ševčíková et al., 2007). For legal domain-specific data sets, non annotated legal text is abundant, as detailed in Pontrandolfo (2012). For example, the pre-training of Legal-BERT (Chalkidis et al., 2020) is performed on a corpus of non annotated documents consisting of legislation, court cases, and contracts from the UK, US, and the European Union. However, the fine-tuning of Legal-BERT is based on an annotated data set CONTRACTS-NER that is not publicly available. Alvarado et al. (2015) annotated 8 filings from the US SEC EDGAR data set, the source of documents for our data set. The primary distinction between their work and ours is the size of the data set, 54K tokens in their data set vs. 400K tokens in ours. Furthermore, Alvarado et al. (2015) was focused on NER in the financial (credit risk) rather than legal domain.

Păias et al. (2021) published a Romanian NER data set consisting of 370 legal documents, and Trias et al. (2021) created a data set consisting of header sections of court cases text (the header section will declare the parties involved in a court case). Finally, we also note that the EDGAR database has been used by Loukas et al. (2022) to create an annotated data set, called FiNER, which contains over 1.1 million sentences. However, this data set is tagged with eXtensive Business Reporting Language (XBRL) tags, and it is used for numeric entity recognition.

3 EDGAR-NER (E-NER) data set

We first describe the source documents that constitute the EDGAR-NER (E-NER) data set. We then enumerate the named entity classes, which slightly extend those used by CoNLL-2003 (CoNLL), which is widely used in the NER community.

Financial entities, such as companies, individuals, and funds, that are registered with the United States Securities and Exchange Commission (US SEC) are required by law to submit financial filings to the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR). All filings in the EDGAR data set are publicly available. There is a wide variety of different filings some of which contain almost no text, e.g. Form 3 (Initial statement of beneficial ownership of securities) or Form 4 (Statement of changes in beneficial ownership of securities). We have arbitrarily chosen the year 2010 and downloaded 52 documents.

The 52 EDGAR documents consisted of a variety of different filings. We only selected filings that contain content in the form of sentences, such as Form 10-Q, which are company quarterly reports, or Form 8-K, which are used by companies to announce major events relevant to their shareholders. The 52 documents consist of 24 different types of forms. Please see Appendix A for details.

The filings were downloaded using the index file provided by EDGAR, in the form of HTML.

3This is available at sec.gov/os/accessing-edgar-data.
text. Each document was pre-processed using the Python package “Beautiful Soup” to extract sentences. We remove:

- the SEC filing header, where the filer fills in the information in a designated space. This is indicated by the HTML tag `<SEC-HEADER>`.

- graphical elements, such as company logos or scanned photos. This is indicated by `<TYPE>GRAPHIC`.

- tables with no sentences in them. Tables are indicated by the HTML tag `TABLE`.

- page titles and page numbers.

- figures and plots.

- any XBRL (eXtensible Business Reporting Language) instance.

An illustration of what elements we removed or kept in an example filing is shown in Appendix B. After preprocessing the 52 documents, we split the document into sentences by identifying the line breaks in the document, and using the sentence tokenizer from the Python NLTK package. In total, we identified 11,696 sentences that required annotation.

Annotation of the collection was performed by the first author. Note that we did attempt to outsource the annotation to a commercial crowdsourcing platform. We provided instructions, including the definitions of the named entity classes and the tagging guidelines. Each document was assigned to 3 crowd workers to independently label so as to ensure the correctness of the labels. However, we found that there were significant discrepancies in the labels provided. While we acknowledge that this variation may have been due to our instructions being poor, it is our opinion that the task has a significant difficulty for a non-expert.

The CoNLL-2003 data set defines 4 classes of named entities (and the class “Outside” for non-named entities)⁴ as enumerated in Table 1. For our data set we annotated the filings with 7 named entity classes as shown in Table 1. We note that there is no consensus on the appropriate labeling of named entities for the legal domain, with various authors (Dozier et al., 2010; Cardellino et al., 2017; Leitner et al., 2019) proposing related but different classifications. Our class labels were chosen in consultation with a legal company (Clifford Chance LLP). Note, however, that for the experimental results reported in Section 4, we used the same categories as CoNLL-2003, merging and matching categories as shown in Table 1. E-NER follows the same file format conventions as CoNLL.

Table 2 provides a statistical comparison between the E-NER and CoNLL-2003 data sets. We see that while the number of tokens in the E-NER data set exceeds that of CoNLL (by combining the training, validation, and test sets), the number of NE phrases is considerably smaller (8,821 for E-NER, compared to 35,088 CoNLL combined). We also observe that the CoNLL data set has considerably more sentences (22,136 vs. 11,696) and that these sentences are much shorter (13.7 words vs. 34.5 words per sentence). The number of tokens constituting a NE is also shorter in CoNLL (1.45 vs. 2.68).

### 4 Experiments

To verify the need for a legal NER collection, we evaluated the performance of four NER methods by (i) training and testing on a general English collection (CoNLL), (ii) training on general English, but testing on our legal collection (E-NER), and (iii) training and testing on our E-NER collection.

The CoNLL collection is subdivided into train, validation, and test partitions, as indicated in Table 2. When training and testing using E-NER, we performed k-fold cross-validation. Since the size of the train and test data sets in CoNLL-2003 has a ratio of approximately 4:1, we chose \( k = 5 \). We report the micro-F1 score.

<table>
<thead>
<tr>
<th>CoNLL</th>
<th>E-NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Location</td>
</tr>
<tr>
<td>Person</td>
<td>Person</td>
</tr>
<tr>
<td>Organization</td>
<td>Business</td>
</tr>
<tr>
<td>Government</td>
<td>Court</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Legislation/Act</td>
</tr>
</tbody>
</table>

Table 1: Named entities used in the CoNLL and E-NER data sets and their pairing in the two classification frameworks

---

⁴See cnls.ua.ac.be/conll2003/ner/annotation.txt
### 4.1 CoNLL-2003 workshop baseline model

The baseline model records all the NE phrases in the training set. During testing, phrases are matched against these learned NE phrases and labeled accordingly (i.e., there is no generalization). If a phrase in the dictionary has multiple NE classes associated to it, the one with the highest frequency is used.

### 4.2 Hidden Markov Model

Our HMM implementation follows the same approach as proposed by Morwal et al. (2012). The NE tags are treated as the hidden states, and the tokens are treated as the observations.

### 4.3 Conditional Random Fields

Our CRF implementation is similar to the one proposed by McCallum and Li (2003). However, we did not use lexicons or other reference corpora to assist our CRF models to identify names of countries, companies, and surnames. Our choice of feature functions is hand-crafted, and consists of (i) the current word, (ii) the first and last 2 letters of the current word, (iii) the capitalization of the word, and (iv) the above 3 features for the word to the left and to the right of the current word.

### 4.4 BERT

We used a pre-trained version of BERT. In our experiments, we fine-tuned BERT using Hugging Face’s transformer package.

### 4.5 Results

In Table 3, we present the F1-score for the aforementioned NER models when we train and test them on different data sets. In the columns, the first entry in the brackets shows the data set used for training (or fine-tuning), and the second entry shows the test data sets.

When we train and test the models on the CoNLL corpus, F1-scores range from 59.6% to 90.5%. However, when we train on CoNLL and test on E-NER, F1-scores degrade significantly, ranging from 13.6% to 61.1%. When training and testing using the E-NER collection the F1-scores range from 49.1% to 96.1% which constitutes a significant improvement over training using the CoNLL data set. Interestingly, we observe that the dictionary baseline and HMM models perform similarly or worse on legal text compared to their performance on general English. Conversely, for the more advanced CRF and BERT models, performance on legal text exceeds that for general English. It is unclear whether this is principally due to differences in the models, or differences in the test collections. Nevertheless, experimental results support earlier work indicating degradation in performance when NER methods are trained on general English but applied to the legal domain.

### 5 Conclusions and future work

This paper describes the publicly available E-NER data set, derived from company filings from the US SEC EDGAR data set. The collection contains over 400,000 tokens, and as such, is of similar size to the CoNLL-2003 collection. However, the number
of NE phrases (almost 9,000) is only about 25% of the number of NE phrases in the CoNLL corpus. In part, this reflects the statistical differences between general and legal English, where we observed that the sentence length for legal English (34.5 words) is much larger than for general English (13.7), and that the token length of a NE in legal text is longer (2.68 tokens compared to 1.45). In addition, the fact that E-NER encompasses only 52 documents from EDGAR might also contribute to this discrepancy.

Our experimental results compared the performance of four NER methods when trained and tested on combinations of general and legal English. Our results reaffirm that for legal NER in-domain performance is significantly degraded when training without using specific in-domain data.

There is a number of potential future research directions. First, there is a variety of legal specialties, e.g. finance, civil litigation, and criminal law. Further work is needed to investigate how NER models perform in various legal sub-domains – how diverge and large should annotated corpora be for legal NER? To this end, we plan to create annotated datasets for other types of legal documents, such as court proceedings or contracts. In addition, the evaluation of NER models using state-of-the-art methods and language models in legal NLP might unveil more informative results and drive potential methodological improvements.

Acknowledgements

T.W.T.A. and I.J.C. would like to thank Clifford Chance LLP for the financial support and for providing guidance with respect to requirements from the legal community.

References


the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.


Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He. 2014. Tweet segmentation and its application to named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 27(2):558–570.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1):50–70.


A Tables

<table>
<thead>
<tr>
<th>Form types</th>
<th>Count</th>
<th>Form types</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>497K</td>
<td>6</td>
<td>DEFA14A</td>
<td>1</td>
</tr>
<tr>
<td>8-K</td>
<td>6</td>
<td>N-CSR</td>
<td>1</td>
</tr>
<tr>
<td>10-Q</td>
<td>5</td>
<td>POSASR</td>
<td>1</td>
</tr>
<tr>
<td>425</td>
<td>4</td>
<td>PRE 14C</td>
<td>1</td>
</tr>
<tr>
<td>N-Q</td>
<td>3</td>
<td>SC 13D</td>
<td>1</td>
</tr>
<tr>
<td>11-K</td>
<td>3</td>
<td>SC 13DA</td>
<td>1</td>
</tr>
<tr>
<td>424B3</td>
<td>3</td>
<td>S-3</td>
<td>1</td>
</tr>
<tr>
<td>CORRESP</td>
<td>2</td>
<td>S-4</td>
<td>1</td>
</tr>
<tr>
<td>DEF 14A</td>
<td>2</td>
<td>S-8</td>
<td>1</td>
</tr>
<tr>
<td>10-K</td>
<td>2</td>
<td>10-KA</td>
<td>1</td>
</tr>
<tr>
<td>40-17G</td>
<td>2</td>
<td>424B5</td>
<td>1</td>
</tr>
<tr>
<td>497</td>
<td>2</td>
<td>40-APPA</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Type of forms in the E-NER data set

B Example filing

An example filing in the E-NER data set, in the form of the HTML and its rendered version, is shown in Figure 1 and 2. Figure 3 shows an image element in this filing, which we remove during preprocessing. This filing’s CIK number is 0001045487. The accession number is 000119312511147903. The URL to this filing is sec.gov/Archives/edgar/data/1045487/000119312511147903.
Figure 1: Raw HTML of an example filing, downloaded from the EDGAR database.

Figure 2: The rendered version of the filing.
Figure 3: Raw HTML of an example filing, where one of the documents uploaded is an image.