

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

Ting Wai Terence Au¹, Vasileios Lamos¹ and Ingemar J. Cox^{1,2}

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.lamos}@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 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 (Kříž et al., 2014), contract analysis (Chalkidis et al., 2017), prediction of judicial decisions (Aletas 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.

Păiş 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),² 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 file³ provided by EDGAR, in the form of HTML

²CoNLL-2003, [clips.uantwerpen.be/conll2003/ner/](https://.clips.uantwerpen.be/conll2003/ner/)

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

CoNLL	E-NER
Location	Location
Person	Person
Organization	Business
	Government
	Court
Miscellaneous	Legislation/Act
	Miscellaneous

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

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.

⁴See cmts.ua.ac.be/conll2003/ner/annotation.txt

Data set	Tokens	Sentences	Avg. words / sentence	NE phrases	Avg. tokens / NE
CoNLL train	204,563	14,986	13.7	23,498	1.45
CoNLL val.	51,578	3,466	14.9	5,942	1.45
CoNLL test	46,666	3,684	12.7	5,648	1.44
E-NER	403,673	11,696	34.5	8,821	2.68

Table 2: Basic statistics of the CoNLL and E-NER data sets

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 generalisation). 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.

Model	G to G	G to L	L to L
Baseline	.596	.136	.491
HMM	.622	.148	.401
CRF	.820	.216	.902
BERT	.905	.611	.961

Table 3: F1-scores of different models when trained (or fine-tuned) and tested on different data sets. In the column headers, the first entry is the training data set (or data set to fine-tune on), and the second is the test data set. **G** denotes a general data set for NER (here CoNLL), and **L** denotes a legal data set (here E-NER). For the column **L to L**, we perform 5-fold cross-validation.

4.4 BERT

We used a pre-trained version of BERT.⁵ In our experiments, we fine-tuned BERT using Hugging Face’s transformer package.⁶

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

⁵BERT, huggingface.co/bert-base-uncased

⁶Available at github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification.

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

- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoŕiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.
- Daria Alexander and Arjen P de Vries. 2021. “This research is funded by...”: Named Entity Recognition of Financial Information in Research Papers. In *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval*, pages 102–110.
- Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of Named Entity Recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 84–90.
- Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.
- Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in Wikipedia. In *Proceedings of the 2009 workshop on the people’s web meets NLP: Collaboratively constructed semantic resources (People’s Web)*, pages 10–18.
- Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing*, pages 194–201.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901.
- Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pages 9–18.
- Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2020. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Applied Soft Computing*, 97:106779.
- Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pages 19–28.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.
- 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.
- Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named Entity Recognition and Resolution in Legal Text*, pages 27–43. Springer.
- Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.
- Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.
- Hariharan Jayakumar, Madhav Sankar Krishnakumar, Vishal Veda Vyas Peddagopu, and Rajeswari Sridhar. 2020. RNN based question answer generation and ranking for financial documents using financial NER. *Sādhanā*, 45(1):1–10.
- Vincent Kríž, Barbora Hladká, Jan Dědek, and Martin Nečaský. 2014. Statistical recognition of references in Czech court decisions. In *Mexican International Conference on Artificial Intelligence*, pages 51–61.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.
- Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287.
- 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.
- Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020b. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics*, 107:103422.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*. In Press.
- Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431.
- Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a rabiner for named entity recognition in Brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323.
- Andrew McCallum and Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.
- Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116.
- Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop*, pages 51–58.
- Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named Entity Recognition using Hidden Markov Model (HMM). *International Journal on Natural Language Computing (IJNLC)*, 1(4):15–23.
- Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*, pages 4348–4352.
- Gianluca Pontrandolfo. 2012. Legal corpora: An overview. Technical report, University of Trieste.
- Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.
- Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.
- Lawrence Rabiner and Biinghwang Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16.

- Erik 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*, pages 142–147.
- Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1986–1991.
- Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195.
- N Sobhana, Pabitra Mitra, and SK Ghosh. 2010. Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):143–147.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.
- Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58:S11–S19.
- Van Cuong Tran, Dosam Hwang, and Jason J Jung. 2015. Semi-supervised approach based on co-occurrence coefficient for named entity recognition on Twitter. In *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pages 141–146.
- Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. 2021. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 172–179.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, volume 30.
- Xu Wang, Chen Yang, and Renchu Guan. 2018. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, 9(3):373–382.
- Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.
- Yuzhe Zhang and Hong Zhang. 2022. FinBERT-MRC: financial named entity recognition using BERT under the machine reading comprehension paradigm. *arXiv preprint arXiv:2205.15485*.
- Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

A Tables

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

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.

The diagram illustrates the process of restoring a table from an HTML document. It is divided into three horizontal sections:

- Top Section (Original HTML):** Shows a snippet of HTML code. It includes a logo with the text "BODCORG 'WHITE'" and a table with a single row and one column. The table content is "JAMES LONG-SHORT FUND".
- Middle Section (Preprocessed HTML):** Shows the HTML after the first preprocessing step. The logo is replaced by a placeholder "JPEG picture" (indicated by a red arrow). The table is replaced by a placeholder "Tables (removed during preprocessing)" (indicated by a red arrow). The text "Content text (we keep and annotate)" is highlighted in a red box.
- Bottom Section (Restored HTML):** Shows the HTML after the second preprocessing step. The table is restored, and the content text is annotated with a red box around it. The table content is "JAMES LONG-SHORT FUND".

Figure 1: Raw HTML of an example filing, downloaded from the EDGAR database.

JAMES LONG-SHORT FUND
 Ticker: JAZZX

MAY 23, 2011

Before you invest, you may want to review the Fund's prospectus, which contains more information about the Fund and its risks. The Fund's prospectus and Statement of Additional Information, both dated May 23, 2011, are incorporated by reference into this Summary Prospectus. For a free paper or electronic copy of the Fund's prospectus and other information, go to www.jamesfunds.com/Prospectus.aspx; call 1-800-699-JAMES (1-800-699-2637); email a request to Info@jamesfunds.com or ask any financial intermediary who offers shares of the Fund.

JPEG picture
(removed during preprocessing)

Tables
(removed during preprocessing)

2

JAMES LONG-SHORT FUND
 Ticker: JAZZX

INVESTMENT OBJECTIVE
 James Long-Short Fund seeks to provide long-term capital appreciation.

FEES AND EXPENSES OF THE FUND
 This table describes the fees and expenses that you may pay if you buy and hold shares of the Fund.

ANNUAL FUND OPERATING EXPENSES
(expenses that you pay each year as a percentage of the value of your investment)

Management Fee	1.25%
Distribution (12b-1) Fees	0.25%
Other Expenses*	
Dividend Expenses on Short Sales	0.11%
Remainder of Other Expenses	0.04%
Total Other Expenses*	0.15%
Acquired Fund Fees and Expenses	0.10%
Total Annual Fund Operating Expenses	1.70%

* Other expenses are based on estimated amounts for the current fiscal year.

Content text (we keep and annotate)

EXAMPLE
 The Example is intended to help you compare the cost of investing in the Fund with the cost of investing in other mutual funds. The Example assumes that you invest \$10,000 in the Fund for the time period indicated and then redeem all of your shares at the end of those periods. The Example also assumes that your investment has a 5% return each year and that the Fund's operating expenses remain the same. Although your actual costs may be higher or lower, based on these assumptions your costs would be:

Figure 2: The rendered version of the filing.

