Named Entity Recognition for Extracting Corporate Governance Information

Christian Millsop, Sam Tosaria
W266: Natural Language Processing
August, 2019

Abstract

The quality of corporate governance is a composite judgement of many factors. Part of that judgement can be inferred from the structure and qualification of a company's board of directors. We developed a Named Entity Recognition (NER) model to extract information about the directors from unstructured DEF14A filings. Our model has a micro f1-score of 91.02% on a novel dataset generated from DEF14A filings. We also introduced a novel NER entity type ('TITLE') and attained a micro f1-score for it of 88.11%. We used a combination of CNN-biLSTM and transfer learning to achieve this. During our work we also performed preliminary examinations of CNN and Transformer architectures.

1. Introduction

All public companies in the United States are required to release information on their finances and operation to the Securities and Exchange Commission (SEC). Most of these documents are structured due to the well-defined nature of the data and are readily consumed by programmatic means for analysis. However, the filing related to the governance structure of the company is released as unstructured text in the DEF14A document. In order to build automated systems that evaluate corporate governance, a method needs to be developed to extract data from these unstructured documents.

The "Director Nominee" section is present in every DEF14A filing and provides biographical data about the qualifications of everyone nominated to the board of directors. We believe that this is a rich source of information, that when collected *en masse* could also illustrate the network of connections between companies and their boards of directors. We propose using NER, a sub-task of Information Extraction (IE) of Natural Language Processing (NLP) to collect this data. The objective described in this paper is to develop an NER model that is successful at labeling useful information in the DEF14A domain-specific corpus. The primary entities that we wish to extract are:

- Person, such as the names of the directors themselves
- Organization, such company names, board committees, etc.
- Title, such as director, trustee, etc.
- Date, such as age, tenure, etc.

2. Background

NER is used in a wide range of contexts, with academic and commercial applications. Examples of extracting data from medical data, financial records, and chemistry research using NER exist **[1,7,11]**. Cloud providers, including Google Cloud Platform, provide NER as a service for accessible IE. It's important to note that each of these NER applications relies on a different set of entity types ('PERSON', 'ADDRESS', 'MEDICATION', etc.) to achieve its domain goal.

The development of models for NER closely follow the models for NLP at large. Early models used CRF, SVM, or simple feed-forward neural networks, which have been supplanted by CNN and biLSTM models **[6]** and most recently by attention-based Transformer models **[9,10]**.

The choice of sequence tagging formatting (BIO1, BIO2, BILOU) has generated considerable contradictory results **[2,12]**. In this paper we aim to assess the impact of sequence tagging format on NER accuracy, along with embedding and model choice for a hand-annotated small dataset of 708 sequences consisting of board of directors biographies submitted to the SEC by publicly listed companies annually in the DEF14A filings.

Common datasets for NER training and model development are CoNLL-2003 **[4]**, Ontonotes 5.0 **[3]**, and various domain specific corpora (eg. i2b2 for biomedical). CoNLL-2003 only has 4 entity types annotated and is relatively short in comparison to Ontonotes 5.0. The CoNLL-2003 dataset is based upon data from the Reuters news service between August 1996 to August 1997. The Ontonotes 5.0 dataset was compiled from a range of television, news, and website text data. **Table 1** below summarizes this comparison.

Table 1. Summary of general NER datasets

Dataset	Entity Types	Tokens	Entities
CoNLL-2003	4	302,811	34,999
Ontonotes 5.0	18	1,388,955	104,151

Transfer learning is a common approach for leveraging generic corpora to train NER models with novel entity types that do not exist in the generic training data. This was demonstrated by Hofer et al **[7]** using the Ontonotes 5.0 corpus to recognize medical entities and Rodriguez et al **[8]** more generally.

3. Methodology

In this work, we explore three different neural network models for our domain-specific task. We explore both training on a small domain-specific corpus as well as transfer learning from Ontonotes-trained models. A comparison of the annotation variants, BIO and BILOU, on our domain-specific task is also performed. We introduce a novel entity type, 'TITLE', on top of the pre-existing 18 entity types found in Ontonotes. The 'TITLE' type designates any specific position held by an individual. We believe that this will support our goal of extracting important information from biographical data.

3.1 Dataset Generation

The Ontonotes data was extracted from the official release by the Linguistic Data Consortium. It comes pre-annotated with entity types following the BIO method.

The domain-specific corpus was manually generated for this project. The DEF14A filings are publicly available on the SEC's EDGAR tool in PDF and HTML formats. The "Director Nominee" section was extracted into per director biographies. This data was tokenized using the NLTK sentence and word tokenizing subroutines (nltk.tokenize.punkt). It was important for us to use a common tokenizing method so that the generation process is reproducible and extensible to new data. After tokenization, each token was annotated according to the BIO format.

The manual annotation process was useful for understanding the variety of formats that entities take as well as the unique sentence structure found in biographical summaries. We did not clean up the extracted data beyond automated tokenization, and preferred to leave it in this state for reproducibility and extensibility. During annotation, it became apparent that long company names, hierarchical company organizations, and company abbreviations would pose a challenge unique to this domain. Some of the unique structures included lists of committee memberships, dense descriptions of work histories (eg. 'Company X (1999 to 2010)'), and out-of-sentence description of name, age, and position (eg. 'John Doe - age 65 - independent director').

The domain-specific dataset totalled 708 sentences. These sentences represent 79 directors across 12 companies. We split the data into train: 300, development: 100, and test: 308

sentences. Since each company is allowed to format the DEF14A independently, we expect there to be variation between company filings. For this reason, we prioritized having a large test set to ensure that our models generalize across companies. The key entity types for our task include: 'ORG','TITLE', 'DATE', and 'PERSON'. The other entity types occur infrequently in the corpus. The overall token and entity counts divided by set are in **Table 2** below.

Table 2. Summary of domain dataset

	Train	Development	Test
Tokens	78,848	25,600	76,800
Entities	2,879	938	2,832

In order to compare performance between BIO and BILOU annotation formats, we used code from AllenNLP to translate BIO annotations into BILOU format. This operation was performed on all of the Ontonotes and domain-specific data.

Table A1 in the Appendix provides further detail on the entity breakdown of the test set with examples. It includes the entity count by type and position (BILOU) and examples for each entity type found in the corpus. To translate those counts to BIO, combine B+U for B and I+L for I.

3.2 Model Descriptions

The three models used were (1) CNN, (2) CNN-biLSTM, and (3) BERT.

3.2.1 Model 1: CNN

The baseline model consists of varying number of convolutional layers and dense layers. This model was chosen for its simplicity and used to determine the path of future work. The embedding matrix uses the cased Google-News 300 dimensional vectors trained on 3 million words and phrases.

3.2.2 Model 2: CNN-biLSTM

The second model is a CNN-biLSTM with three feature inputs: word-level tokens, word-level casings, and character-level tokens. This model was inspired by Chiu and Nichols **[6]** and the follow-up work by Hofer et al **[7]**. Chiu and Nichols demonstrated that a CNN-biLSTM model out-performed state-of-the-art non-biLSTM models. The full model architecture is in the Appendix **Figure A1**.

The model inputs were processed to the specifications in **Table 3** below. They were truncated and padded accordingly. Truncation only occurred on three sentences in the Ontonotes development data. We do not expect this to impact the performance since it is a small fraction of the 15,680 total development samples. The development data is also only used for early-stopping in training and not evaluation.

Table 3. CNN-biLSTM Input Details

Input	Dimensions	Vocabulary
Words	256 per sentence	6 billion tokens in GloVe
Word Cases	1 per word (one-hot)	Upper, Lower, Title, Numeric, Other
Characters	52 per word	100 tokens in Python's string.printable

The Stanford GloVe 6 billion token word embeddings were used **[5]**. We compared the performance of the 50 and 300 dimensional vectors. Chiu and Nichols **[6]** reported no difference in performance between the two on the Ontonotes corpus.

Out-of-vocabulary terms for both words and characters were replaced with a special token <UNK> when they were encountered for both model 1 and model 2. Padding to the dimension size was done with the token <PAD>.

3.2.3 Model 3: BERT

The final model is a transformer model with ten trainable layers. We used the cased BERT embeddings for this model. We did not expose the DEF14A data to the BERT model.

3.3 Transfer Learning Method:

After training the model on Ontonotes, we transferred the pre-trained weights to a new model of the same architecture and fine-tuned the model using various sizes of training data. In addition to training size, we also explored the impact of transferring various layers and whether freezing the weights of a transferred layer during fine-tuning affected performance.

We test training sizes of 10, 25, 50, 100, 200, and 300. This allows us to develop training curves, which inform us on how the model might behave with more training data. Our method for increasing the training set size is to append additional samples instead of re-sampling from all training data. This ensures that our tests are comparable, especially because there is considerable variation in the data depending on what company the biography was taken from.

In Hofer et al **[7]**, the authors demonstrate that the best results using the Ontonotes pre-trained model were had when only using the pre-trained weights from the biLSTM layer. All other layers were randomly initialized to the un-trained state.

Freezing the layer weights after transferring is similar to using pre-compiled word embeddings, except at a specific layer in the network. We test freezing the biLSTM layer in our model and explore how the training rate and generalizability of the model changes.

4. Results and discussion

4.1 CNN

The CNN model accuracy was highly sensitive to the length of the sequence fed into the model. A sequence of 32 tokens generated model accuracy of 54%. Shortening the token sequence to 26 tokens, which represents 82% of all sequences in the Ontonotes training data, improves the accuracy to 61.5%.
Improving the loss rates per epoch required corresponding and consistent reduction in learning rates along with early stopping. Adding more convolutional and dense layers did not lead to meaningful performance gains as the unstable gradient **[13]** on longer sequences may have lead to accuracy reduction.

The final model with two convolutional layers and two dense layers performed better than deeper more complicated models. The final baseline model used a learning rate of 0.0004 and beta-one of .85.

The BIO and BILOU formats produced similar accuracy scores. BIO produced lower loss rates on both Ontonotes test set and DEF14A dataset. Moreover, the high degree of negative correlation of the model to the length of the input sequences implies gradient calculation issues and a need for attention layer in the following models.

4.2 CNN-biLSTM

Although the model was trained to recognize a total of 39 (BIO) or 77 (BILOU) tags, many of the tags are not substantially represented in the domain data. The most important tags for our task are: ‘PERSON’, ‘ORG’, ‘TITLE’, ‘DATE’. The ‘TITLE’ tag is novel and we expected it to perform the worst.

The most common tag by far in our test set is ‘O’ (‘outside’), which accounts for 96% of the tags. All f1-scores reported for the CNN-biLSTM are the micro score excluding the ‘outside’ tag.

In total, for this model we generated 16 different predictions against the domain specific data. Four of the configurations were CNN-biLSTM baseline models without transfer learning. The remaining 12 were variations on the two fine-tuning parameters (pre-trained layers, and freezing biLSTM layer) and two model-level parameters (word embedding dimensions and annotation method).

The result on the Ontonotes corpus achieved by Chiu and Nichols was 86.25% using the 50d

GloVe embeddings and BILOU annotation **[6]**. With our model we achieved a best score of 80.62% using the 300d embeddings and BIO annotation. This is notably worse, which we believe is primarily due to the additional feature extraction and pre- and post-processing that they did on the results.

The top result for each of our configurations and the top CNN-biLSTM baseline on the domain corpus are below. The top model in **Table 4** had a micro f1-score on TITLE (B+I) of 88.11%.

Table 4. Summary of CNN-biLSTM results

Embedding Dimensions	Pre-Trained Weights	Annotation	biLSTM Frozen	F1-Score
300	biLSTM, softmax	BIO	No	91.02%
300	All Layers	BILOU	No	88.36%
50	All Layers	BIO	No	88.15%
50	All Layers	BIO	Yes	87.11%
300	None (Baseline)	BIO	No	84.15%

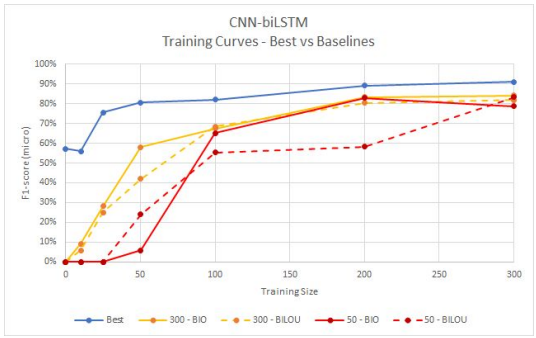
The models with frozen biLSTM weights generally performed poorly. This result implies that the biLSTM layer is being tuned to understand the structure of our domain sentences.

We also see that saw that the 300 dimensions embedding model performed better than the model with 50 dimension embedding layer. This is interesting in the context of named entities, since most of our classifications will be on unique arrangements of characters and words that might not have a reference in the embeddings. It suggests that the additional dimensionality encodes for useful, alternative meanings. Comparing similar configurations where the only difference is word embedding dimensions, the 300d set performs much better on 'TITLE' tag recognition and slightly better on 'ORG' and 'PERSON' tags..

There was a notable difference between BIO and BILOU results, with the BILOU annotations generally performing worse than their BIO counterparts. A comparison of the 300d, all weights, BIO versus BILOU results demonstrates that BILOU was generally worse at all tags. The only surprising observation is that BILOU was better at predicting I-PERSON (I+L) than BIO. See **Table A3** in the Appendix for more details.

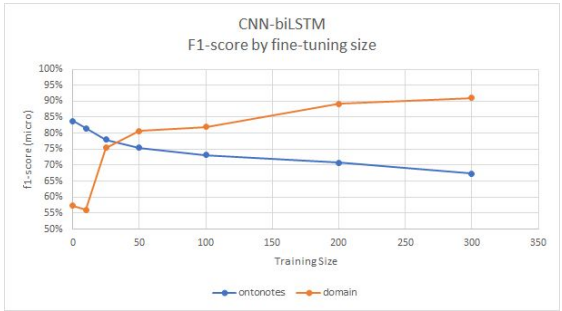
We were surprised by how well the baseline training approaches worked on CNN-biLSTM with very little data and no pre-training. A closer examination of the training curves for the best model versus the baselines is below. The baseline curves follow an approximately log-shaped progression. We experienced diminishing returns on all models past the 200-sample point and we would expect sample sizes beyond 300 to have even more of a reduced effect on improving the model.

Figure 1. CNN-biLSTM Training Curves



We also observed that the generalizability of the model decreased as fine-tuning proceeded. This is evaluated in terms of the Ontonotes f1-score versus the domain f1-score. Please see Figure 2 below for details.

Figure 2. CNN-biLSTM Generalizability



4.3 BERT

The DEF14A dataset consists of long sentences with references to a person's professional experience, tenure and timelines. Intuitively, the Transformer models ability to effectively process information over long sentences may increase accuracy. Conversely, the dense noun tags with names of persons and entities is very different from the training data represented in the Ontonotes dataset.

The model achieved an accuracy score of 92.45% on BIO-Ontonotes test dataset and 82.52% on the DEF14A domain specific data. The accuracy difference may be attributed to the data structure and the model's unfamiliarity with the data.

5. Conclusion

We built three models progressively to perform NER on a novel corpus. Our examination of the CNN-biLSTM provided insight on how transfer learning could be successfully applied in this context. We achieved good results on a small training corpus using the CNN-biLSTM that we believe are good enough to enable automated information extraction.

We believe that there is room for improvement on the CNN-biLSTM based upon comparison of our baseline Ontonotes work to existing publications **[6]**, however it is unclear whether improvements on baseline Ontonotes would translate to improvements in domain accuracy.

Our results on the comparative outperformance of the BIO tags over the BILOU tags verifies the result observed by Konkol and Miloslav Konop ́ık **[12]**, for English language data, in their multilingual NER tag schemes study.

References:

1. Francis, Sumam, Jordy Van Landeghem, and Marie-Francine Moens. "Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents." *Information* 10.8 (2019): 248.
2. Ratinov, Lev, and Dan Roth. "Design challenges and misconceptions in named entity recognition." *Proceedings of the thirteenth conference on computational natural language learning*. Association for Computational Linguistics, 2009.
3. Hovy, Eduard, et al. "OntoNotes: the 90% solution." *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*. 2006.
4. Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *arXiv preprint cs/0306050* (2003).
5. Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
6. Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." *Transactions of the Association for Computational Linguistics* 4 (2016): 357-370.
7. Hofer, Maximilian, et al. "Few-shot Learning for Named Entity Recognition in Medical Text." *arXiv preprint arXiv:1811.05468*(2018).
8. Rodriguez, Juan Diego, Adam Caldwell, and Alexander Liu. "Transfer learning for entity recognition of novel classes." *Proceedings of the 27th International Conference on Computational Linguistics*. 2018.
9. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
10. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
11. Corbett, Peter, and John Boyle. "Chemlistem: chemical named entity recognition using recurrent neural networks." *Journal of cheminformatics* 10.1 (2018): 59
12. Michal Konkol and Miloslav Konop´ık. "Segment Representations in Named Entity Recognition"(2018)
13. Pascanu, Mikolov, Bengio. "On the difficulty of training RNNs", 2012. arXiv:1211.5063

Appendix

Table A1. Detailed breakdown of the test domain set with examples.

Label	B	I	L	U	Total	Contextual Example
ORG	381	590	381	106	1458	… a general manager with **Microsoft Corporation** ...
TITLE	136	108	136	196	576	… a **general manager** with Microsoft Corporation ...
DATE	145	118	145	96	504	… Hindustan Unilever Limited (**2005 to 2018**) ...
PERSON	43	34	43	86	206	**Jeffrey P. Bezos**, age 55, ...
PRODUCT	7	11	7	12	37	… held a **Guggenheim Fellowship** at the ...
GPE	6	3	6	19	34	… practice in the **United Kingdom** and ...
MONEY	3	2	3	0	8	… an 8,000 person **multi-billion dollar** business ...
LAW	2	1	2	0	5	… reorganization under **Chapter 11** of the ...
NORP				3	3	… Zurich Insurance's **North American** subsidiary.
CARDINAL				2	2	… an **8,000** person multi-billion dollar business ...
ORDINAL				2	2	… the **second** highest-ranking military member ...
O					73968	
Total					76800	

Other tags with no support: WORK_OF_ART, LOC, EVENT, FAC, QUANTITY, TIME, PERCENT, LANGUAGE

Figure A1. CNN-biLSTM Architecture

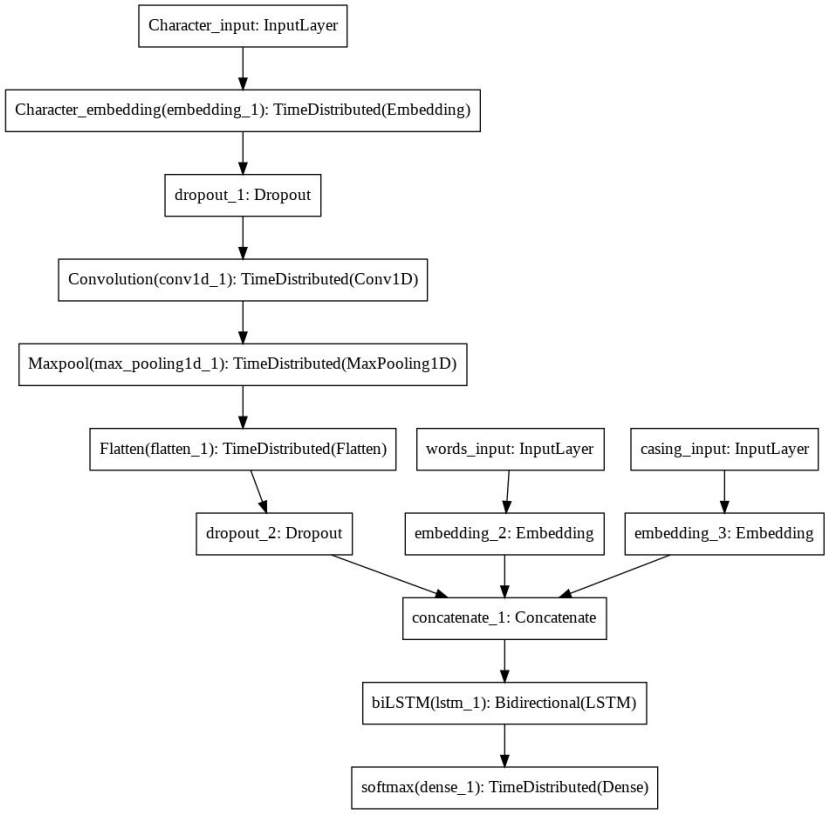


Table A2. All baselines on Ontonotes with CNN-biLSTM

Embedding Dimensions	Annotation	F1-score
50	BIO	79.32%
50	BILOU	75.25%
300	BIO	80.62%
300	BILOU	78.90%

Table A3. CNN-biLSTM: 300d, all weights, unfrozen - BIO versus BILOU on select entity types

Annotation	Label	Support	Precision	Recall	F1
BILOU	B-ORG	381	88.07%	91.08%	89.55%
BILOU	U-ORG	106	79.49%	87.74%	83.41%
BILOU	Micro	487	86.20%	90.35%	88.21%
BIO	B-ORG	487	88.29%	91.38%	89.81%
BILOU	I-ORG	590	90.36%	81.02%	85.43%
BILOU	L-ORG	378	87.90%	93.44%	90.59%
BILOU	Micro	968	89.40%	85.87%	87.45%
BIO	I-ORG	968	94.10%	87.40%	90.63%
BILOU	B-DATE	145	97.20%	95.86%	96.53%
BILOU	U-DATE	96	97.87%	95.83%	96.84%

BILOU	Micro	241	97.47%	95.85%	96.65%
BIO	B-DATE	241	98.29%	95.44%	96.84%
BILOU	I-DATE	118	92.06%	98.31%	95.08%
BILOU	L-DATE	145	96.43%	93.10%	94.74%
BILOU	Micro	263	94.47%	95.44%	94.89%
BIO	I-DATE	263	95.54%	97.72%	96.62%
BILOU	B-PERSON	43	90.91%	93.02%	91.95%
BILOU	U-PERSON	86	93.18%	95.35%	94.25%
BILOU	Micro	129	92.42%	94.57%	93.49%
BIO	B-PERSON	129	93.23%	96.12%	94.66%
BILOU	I-PERSON	34	100.00%	94.12%	96.97%
BILOU	L-PERSON	43	97.67%	97.67%	97.67%
BILOU	Micro	77	98.70%	96.10%	97.36%
BIO	I-PERSON	77	94.74%	93.51%	94.12%
BILOU	B-TITLE	136	95.50%	77.94%	85.83%
BILOU	U-TITLE	196	92.18%	84.18%	88.00%
BILOU	Micro	332	93.54%	81.63%	87.11%
BIO	B-TITLE	332	92.09%	87.65%	89.81%
BILOU	L-TITLE	136	93.64%	75.74%	83.74%
BILOU	I-TITLE	108	95.51%	78.70%	86.29%
BILOU	Micro	244	94.46%	77.05%	84.87%
BIO	I-TITLE	244	90.99%	82.79%	86.70%

Figure A2. CNN-biLSTM: 300d, biLSTM + softmax weights, unfrozen - Generalizability

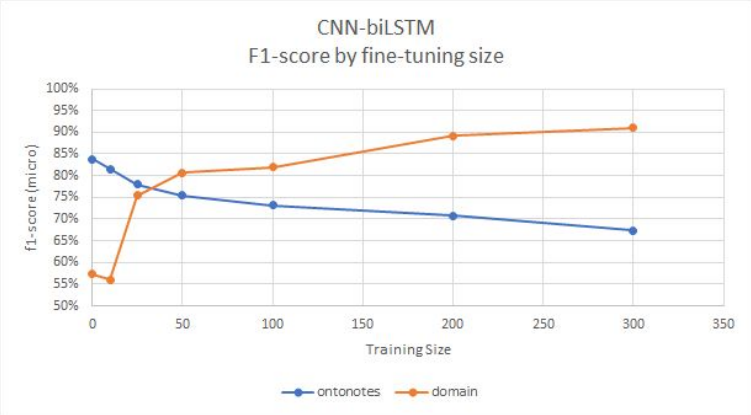


Figure A3. Example NER on domain corpus using Stanford CoreNLP:

Figure A4. Example NER on domain corpus using SpaCy