International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Legal Clauses Prediction Model

Prof. Sadia Patka[1], Mirza Asim Baig[2], Moin Khan Pathan[3], Princy Gundetiwar[4], Adil Ahmed[5], Aaliya Sheikh[6], Shiban Siddiqui[7]

[1]Assistant Professor, Computer Science and Engineering, A.C.E.T. Nagpur, Maharashtra, India
[2]B.E. Student, Computer Science and Engineering, A.C.E.T. Nagpur, Maharashtra, India

---

Abstract – *There is lack of accessible labelled data that has hindered the application of deep learning in various mastered domains. However, a new solution has been introduced to solve this obstacle within the legal domain of this industry. The creators of this solution are The Atticus Project and they have developed the Contract Understanding Atticus Dataset (CUAD), which is a new dataset for legal contract review. CUAD consists of over 13,000 annotations and was created with the help of dozens of legal experts. The aim of this dataset is to highlight the important portions of a contract that require human verification. Whilst initial results of Transformer models have shown promising performance, there is still a significant room for improvement. This dataset is one of the few large, specialized NLP benchmarks annotated by experts, making it a challenging research benchmark for the wider Natural Language Processing community. Thus, CUAD dataset has the potential to address the bottleneck in deep learning within the legal domain.*

Key Words: Labelled Datasets, CUAD, Pre-processing, Legal Contract.

1. INTRODUCTION

The Contract Understanding Atticus Dataset (CUAD) v1 is a specified dataset consisting of more than 13,000 labels among 510 commercial legal contracts. This dataset was manually labeled by experienced lawyers under the supervision of The Atticus Project, Inc. to identify 41 types of legal clauses that are important for contract review in corporate transactions such as mergers and acquisitions. The CUAD dataset is maintained and curated by The Atticus Project, Inc. to support research and development in Natural Language Processing for legal contract supervision.

Contract review is an essential but a time-consuming work that can cost businesses a good amount of money. For example, many law firms have to spend approximate 50% of their time reviewing contracts, and billing rates for lawyers at large firms can cost between $500-$900 per hour in the US. This can result in significant transaction costs for companies, as they have to pay for lawyers to verify that there are no problematic obligations or requirements in the contracts. Furthermore, contract review can be a tedious and monotonous task, which may explain why many individuals and small companies often sign contracts without reading them. This can result in predatory behavior that harms consumers.

By openly releasing better-quality data (i.e. cleaned data) and models for contract review, improvisation of this task could become possible, and access to legal support for smaller businesses and individuals can exponentially increase. Accessibility of this legal support would be affordable for everyone, rather than just being available to wealthy companies.

2. Literature Review

This section includes the related work done in past CUAD dataset.

"CUAD: An Expert-Annotated Natural Language Processing Dataset [1] for Legal Contract Review" presents the Contract Understanding Atticus Dataset (CUAD), which contains over 13k manually annotated labels in 500+ commercial legal contracts. The task is to highlight important legal clauses in the contracts. The study finds that Transformer models have nascent performance and that their performance is influenced by model design and dataset size, leaving room for improvement.

"The Law of Large Documents: Understanding the Format of Legal Contracts Using Visual Cues" proposes a method of segmenting legal contracts based on their structural metadata to improve understanding [2]. This study uses Adam, Attention Dropout, and BERT models and outperforms existing methods on four long-document understanding tasks, as measured on CUAD. This study argues that the segmentation of strategies that are used on longer documents typically miss structural details, leading to poor results on downstream tasks.

"Law Informs Code: A Legal Informatics [3] Approach to Aligning Artificial Intelligence with Humans" describes how data generated by legal processes can facilitate the specification of inherently vague human goals, increasing human-AI alignment and the local usage of AI system. The study proposes a research agenda called "Law Informs Code," which aims to capture complex computational legal processes and embed them in AI, similar to how legal

contracts cannot foresee all potential contingencies, and legislators might not predict all the circumstances under which their proposed bills will be applied.

"Numerical Reasoning over Legal Contracts via Relational Database" [4] study and identify the problem of refining natural language text into a relational database with numerical data structure and querying this database to obtain desired answers. This study however finds that existing approaches are limited in their ability to incorporate domain-specific knowledge and express these mathematical formulas over data structures.

"Con Reader: Exploring Implicit Relations in Contracts for Contract Clause Extraction" proposes a method [5] for automatic Contract Clause Extraction (CCE) by modeling implicit relations in legal contracts. This study analyzes the difficult issues of legal contracts and distills three main implicit relations commonly found in these contracts. Existing CCE methods typically treat contracts as plain text, creating a barrier to understanding complex contracts. This paper introduces a new framework for automatic Contract Clause Extraction (CCE), which takes into account of three important relations: Long-range Context Relation, Term-Definition Relation, and Similar Clause Relation. By exploiting these relations, Con Reader improves contract understanding and CCE performance. The results of the experiments show that Con Reader achieves state-of-the-art performance on two CCE tasks, both in conventional and zero-shot settings. The approach is also found to make predictions more interpretable.

3. PROPOSED WORK

This section includes problem statement identified and proposed system.

3.1 PROBLEM STATEMENT

- This legal profession struggles with reviewing very large amounts of legal documents quickly and accurately, which can impede agile and data-driven decision making.

- Reliable and working Artificial Intelligence systems that can understand legal text are urgently needed, but their effectiveness is dependent on the availability, quality, and transparency of annotated legal documents created by legal experts.

Let's understand our main problem with the help of an Example: -
Suppose there is a law student who has to review a bunch of law contract agreements to scan them and extract all the main information such as the document name , name of parties that are involved, the agreement date, the expiration date etc ... It will take him/her a lot of time to read them and extract all the required information.

3.2 PROPOSED SYSTEM

Progression of the dataset in every next layer also passing through the LSTM model. our project works in a very systematic flow that involves very different layer of which is main layer of our chart ending with the final prediction part which is user based.

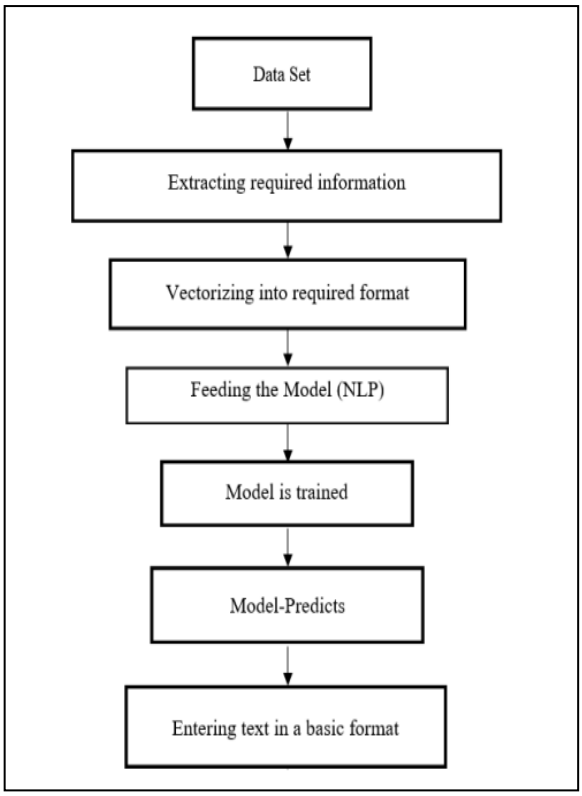


Fig-1 Flow Chart

Module 1 : Data Set

In this module, developers collect data about different fruits and then combine the data to create a new data set. Loading

Dataset in Jupyter Firstly, the dataset is downloaded from its website and then is opened in jupyter.

Module 2 : Extracting Required Information

Then it is loaded and then some functions will be written in python language to process some parts of the dataset.

Module 3 : Vectorizing Into Required Format

The data that is processed is then converted into csv file which is then vectorized through the model which contains all the required information

Module 4 : Feeding The Model

The dataset now is then loaded into the model in the required format in the IOB Tagging method and then is fed into the model.

Module 5 : Model Is Trained

We have to then load pre trained ELMo model from LSTM keras tensorflow library will which will be used to train the data that we fed into the model ELMo works in such that way that it makes easier to process the language for a computer.

Module 6 : Model Predicts

Making Predictions Now that the model will be trained, we will use different text format to test the prediction s of the model and then lastly deploy it on a web server using flask.

4. RESULT AND DISCUSSION

This is the final IOB tagging function and it's Data Frame created for each clause.

Fig-2 IOB Tagger Function

The layering of the modelling is shown below which is loaded through keras function . the summary is displayed with 6 layering of models . Input is taken from the user and is passed onto the next layer with the output as it's input and so on . Total Parameters are mentioned as well.

Fig-3 Model Summary

This is the training and validation graph which shows exact amount of epochs needed for the model to not over train itself. It shows the accuracy and loss we got after the training is completed.

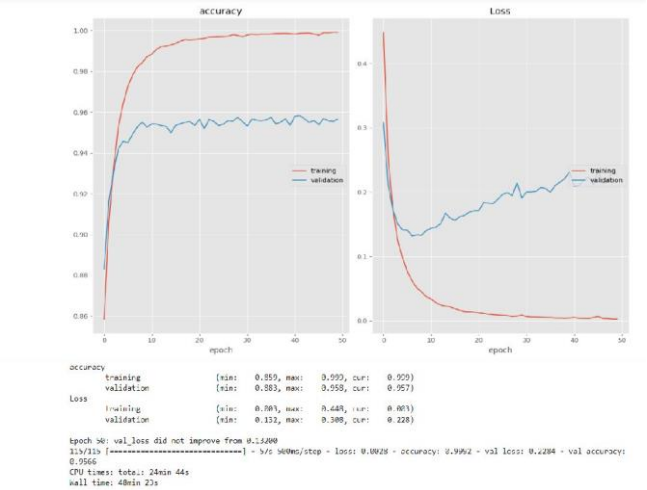
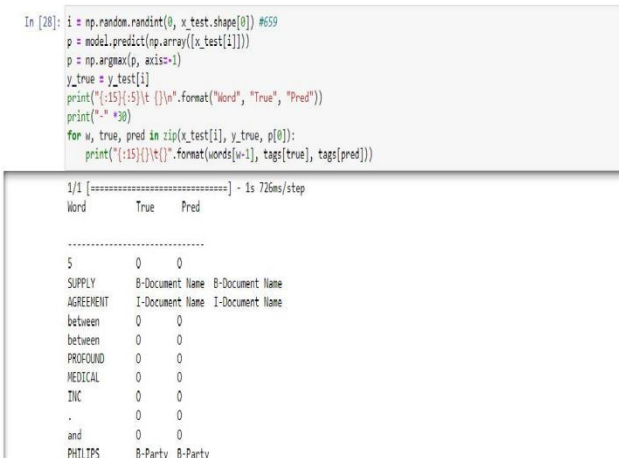


Fig-4 Training and Validation Graph

Finally is the prediction part where the trained model is directly loaded using keras function and then the text is fed into it for the prediction



```
In [28]: i = np.random.randint(0, x_test.shape[0]) #659
         p = model.predict(np.array([x_test[i]]))
         p = np.argmax(p, axis=-1)
         y_true = y_test[i]
         print("{:15}{:5}\t {}\n".format("Word", "True", "Pred"))
         print("-" *30)
         for w, true, pred in zip(x_test[i], y_true, p[0]):
             print("{:15}{}\t{}".format(words[w-1], tags[true], tags[pred]))

1/1 [==============================] - 1s 726ms/step
Word           True     Pred

------------------------------
5              O        O
SUPPLY         B-Document Name  B-Document Name
AGREEMENT      I-Document Name  I-Document Name
between        O        O
between        O        O
PROFOUND       O        O
MEDICAL        O        O
INC            O        O
.              O        O
and            O        O
PHILIPS        B-Partv  B-Partv
```

Fig-5 Final Prediction

5. CONCLUSION

By far CUAD Dataset is one of the most difficult datasets that can be handled compared to those in the past considering the amount of information that is to be extracted and vectorized into another specific form at required by the LSTM model. The training of the model actually gave a brief idea about how important it is to maintain the accuracy if the model even when changing some of its hyper-parameters. Tuning the model also has a deep effect on the amount of data that is pre-processed into the model. Lastly, the ability to differentiate between the actual pre trained datasets and also the ones which have been improvised.

REFERENCES

[1] Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In ACL/IJCNLP 2021, pages 7307–7318.

[2] Ilias Chalkidis and Ion Androutsopoulos. 2017. A deep learning approach to contract element extraction. In JURIX, pages 155–164.

[3] Afra Nawar, Mohammed Rakib ,Salma Abdul Haq. LATERAISSE (LREC) 2022.

[4] DBAI 2021

[5] 17 Oct 2022 · Weiwen Xu, Yang Deng, Wenqiang Lei, Wenlong Zhao, Tat-Seng Chua, Wai Lam Fountalis, May.

[6] Dan Hendrycks, Collin Burns, AnyaChen, SpencrBall CUAD Encodings 10 Mar 2021

[7] Neel Guha, Daniel E.Ho,Julian Nyarko, Christ opher Rg. Legal Bench. 13 Sep 2022

[8] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In ICAIL 2017, pages 19–28.

[9] https://www.youtube.com/watch?v=hFUSdgryXyU

[10] https://analyticsindiamag.com/explained-cuad-the-dataset-for-legal-nlp/

[11] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, ´ and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. ArXiv, abs/2004.10964, 2020.