University of Arkansas – CSCE Department
Capstone I – Preliminary Proposal – Fall 2021

NLP-based Contract Analysis

Colten McDaniel, Chandler Hotchkies, Marcus Langston, Sai Elagandula, Thomas Smith

Abstract

Millions of dollars and countless hours are spent reading through legal contracts by junior level attorneys, often error-prone due to stress and overwork. In order to help address this problem, we aim to create a Natural Language Processing (NLP) based contract analysis model using the FlairNLP framework trained using the Contract Understanding Atticus Dataset (CUAD) and legal word embeddings. The model will be tested for accuracy using 10 of the 41 labels in the CUAD dataset.

1.0 Problem

Contract review is a highly important task in enterprises, but it is a task which is mainly manual and costly. Law firms and legal teams will spend a large amount of time analyzing contracts. Without a way to quickly and cheaply analyze contracts these law firms and legal teams will be wasting time and money which could be better allocated.

2.0 Objective

The objective of this project is to develop a natural language processor to sift through large quantities of text in the form of contracts and withdraw the important information to lower the amount of manual analysis needed to create a meaningful contract.

3.0 Background

3.1 Key Concepts

Natural Language Processing (NLP) refers to the branch of computer science concerned with giving computers the ability to understand text and spoken words in the same way humans can[2]. This allows a computer, with its incredibly fast processing speed, to analyze documents and understand enough to relay some important information about the documents to the human

operator. It utilizes rule based modeling of human language with statistical, machine learning, and deep learning models to analyze documents. One example of the use of NLP-based analysis is programs that convert speech to text or one language to another.

Machine learning is a type of artificial intelligence. It is when a basic neural network is created and allowed to train itself by constantly trying over and over on a data set. The data set will have the correct answers included, and the AI will score itself and try again with slightly different parameters and compare the scores to see if it is getting better or not.

3.2 Summary of Study Introducing CUAD

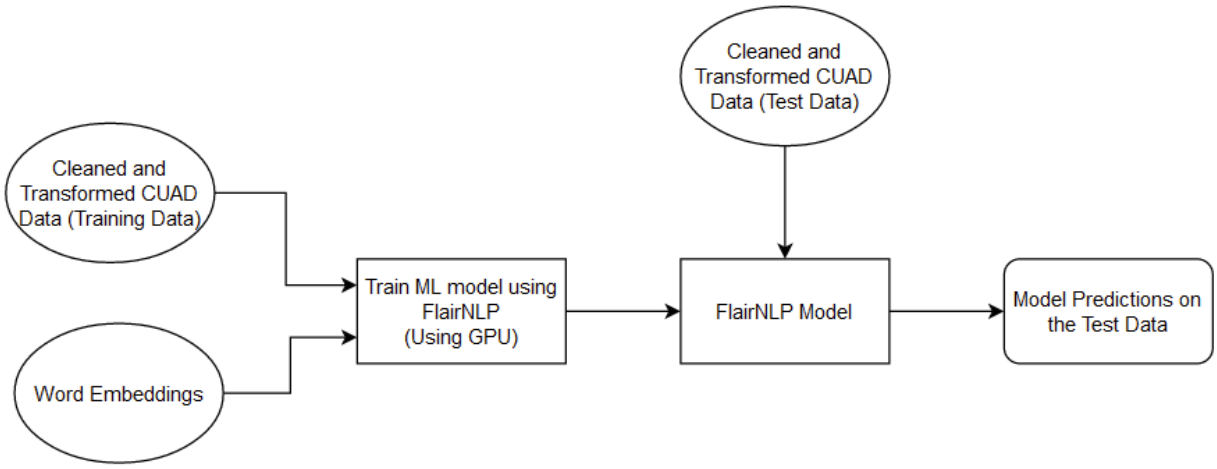
For a long time, the biggest hurdle in contract analysis and identifying is the access to a high quality, large, labeled dataset. Much of the datasets in the legalNLP space thus far have either been in another language outside of English or done for a smaller task which is not applicable to most contract analysis tasks. The researchers that put forth CUAD also evaluated 10 well known pretrained models with further training using CUAD and had promising results, however there was still room for improvement with the best model. Along with this, the researchers found that the amount of data used to train the model had a huge impact which showcased the importance of the dataset [3].

4.0 Design

4.1 Requirements and/or Use Cases and/or Design Goals

The goal is to create a program that would accept a text document and be able to analyze it for certain keywords and phrases. Then relay the surrounding text of said keyword/phrase to the user

4.2 High Level Architecture



The architecture of our project is fairly straightforward. It is the typical ML model architecture which uses training data to train our model. Then, use testing data to test the accuracy of our

model. The test data will have correct labels which will be used to see how accurate the model was in predicting.

Our model will be built using FlairNLP, a python NLP library, and be trained using the CUAD dataset mentioned earlier in the proposal. Along with this, our model will take in established word embeddings to help the model understand legal word pairs. Then, the model will be tested for accuracy after it is run on the test data.

4.3 Risks

Risk	Risk Reduction
Overfitting the model	Splitting the data into training and testing data
Underfitting the model	By allocating a majority of the data to be used for training and a small amount used to testing (possibly a 60/40 or 80/20 split)

4.4 Tasks

Below are the tasks for this project. We are not splitting up the tasks for this project since it is a rather linear process. The entire team will need to work on and understand the different parts of the project in order to progress.

1. Understand FlairNLP
2. Understand the contents of the CUAD Dataset
3. Understand concepts of NLP such as text classification, named-entity recognition, and word embeddings
4. Learn how to use FlairNLP and build a model using FlairNLP by completing their tutorials
5. Establish metrics to evaluate our model and choose 10 features to focus on
6. Outline how to manipulate the CUAD data to be input into a FlairNLP model
7. Clean and Transform CUAD Dataset
8. Research legal word embeddings and learn how to apply them to our model
9. Build and evaluate model for a subset of the data to make sure inputs are correct and model is able to be trained using the data before acquiring a GPU
10. Determine whether or not a GPU is needed to train the data and figure out a way to obtain computing power
11. Build and evaluate model for entire dataset
12. Interpret model results
13. Possibly adjust model for greater accuracy
14. Write final report

4.5 Schedule

Tasks	Dates

1. Understand FlairNLP	11/05-11/12
2. Understand the contents of the CUAD Dataset	11/13-11/19
3. Understand concepts of NLP such as text classification, named-entity recognition, and word embeddings	11/20-11/23
4. Learn how to use FlairNLP and build a model using FlairNLP by completing their tutorials	2/10-2/15
5. Establish metrics to evaluate our model and choose 10 features to focus on	2/10-2/24
6. Outline how to manipulate the CUAD data to be input into a FlairNLP model	1/15-1/22
7. Clean and Transform CUAD Dataset	1/23-1/30
8. Research legal word embeddings and learn how to apply them to our model	1/31-2/6
9. Build and evaluate model for a subset of the data to make sure inputs are correct and model is able to be trained using the data before acquiring a GPU	2/13-2/20
10. Determine whether or not a GPU is needed to train the data and figure out a way to obtain computing power	2/21-2/28
11. Build and evaluate model for entire dataset	3/1-3/8
12. Interpret model results	3/9-3/16
13. Possibly adjust model for greater accuracy	3/17-3/24
14. Write Final Report	3/25-4/1

4.6 Deliverables:

- Data Document - Detailing the data and describing how we transform the data
- Python code for the program
- Data sets used to train the program
- Final Report

5.0 Key Personnel

Colten McDaniel – McDaniel is a senior Computer Engineering major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has completed Artificial Intelligence and Applied Probability and Statistics for Engineers.
Thomas Smith – Smith is a senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has completed Data Mining and Probability and Statistics for Engineers
Sai Elagandula - Elagandula is a Senior Computer Science and Economics major at the University of Arkansas. He has completed Artificial Intelligence, Data Mining, and Applied Probability and Statistics for Engineers. In the past, Elagandula interned at Credera, a boutique consulting firm, in the data and analytics wing, primarily doing data engineering work in the AWS ecosystem. He also interned with STOPWATCH, where he helped build data pipelines and wrote scripts to clean data. In this project, Elagandula will be responsible for
Chandler Hotchkies - Hotchkies is a senior Computer science major in the Computer Science and Computer Engineering Department at the University of Arkansas. Chandler has complete Probability and Statistics for Engineers and is currently in Artificial Intelligence.
Marcus Langston - Langston is a Senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has completed Data Mining and Probability and Statistics for Engineers.

Nathaniel Zinda, Industry champion – *[waiting for response from Nathaniel]*

6.0 Facilities and Equipment

A GPU that is powerful enough to train the model on. Possibly an AWS or Azure account to train the model, if we cannot find a way to train the model on campus.

7.0 References

[1] Contract Understanding Atticus Dataset (CUAD) v1 Overview - YouTube
[2] https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can
[3] https://arxiv.org/pdf/2103.06268.pdf