ContractNLI for Advanced NLP

Team Transformers (1984)
Members: Sidhi Panda, Patanjali B, Druhan Shah

November 5, 2023

1 Introduction

The task of Natural Language Inference is understandably non-trivial. Inferring whether sentences are entailed, contradicted or unrelated to presented data is a very basic task for humans. Indeed, it is often used as a measure of reading comprehension for young language students. In the context of legal contracts, such comprehension is vital to determine whether certain actions are within the scope of the contract or violate it. However, reviewing contracts and analysing their entailments is both time-consuming and fraught with potential loophole exploitation. A study was conducted (Exigent Group Limited, 2019) which claimed that 60-80% of all business-to-business transactions are governed by some form of written contract or agreement.

As such, the automation of such a task could be monumental from the perspective of legal efficiency. Natural Language Inference as a task handles this quite well, but it faces one main drawback for this particular case. Firstly, the base task uses a large source of data and then makes one singular inference about a statement. Additionally, while it makes the statement, it does not provide any information as to what part of the data reinforces its inference.

The extended task of ContractNLI (Koreeda & Manning, 2021) attempts to bridge this gap and provide a model to infer the entailment or contradiction of multiple hypotheses over multiple contracts independently, and present evidence from the document for each entailment or contradiction inference. This project attempts to analyse the implementation of such a model and compare the performance of multiple different models on this task.

2 Theory and Literature

NLI can fund plentiful use in legal contexts, since legal systems are built on the task of inferring entailment of hypotheses from written law. LegalNLI (Yang, 2022) is a dataset that was constructed specifically with the task of Legal Compliance Inspection in mind, using Chinese legal datasets for other problems.

The paper that we follow for this project will be ContractNLI (Koreeda & Manning, 2021), which introduces Span-NLI-BERT: a model that simplifies the task of span boundary detection and then makes inferences based on those spans, providing said spans as evidence.

Firstly, unlike previous works (Devlin, Chang, Lee, & Toutanova, 2019) which predict start and end tokens and scale it to documents by splitting them into contexts, this approach chooses to insert special tokens to split different spans.

This task is split into two subtasks: NLI and Evidence Identification. For the EI task, each span is concatenated using a separator token with the hypothesis and fed into a transformer-based model, followed by a 3-layer MLP with sigmoid activation. This would generate the likelihood of relevance of the chosen span to the hypothesis. For the loss function of this task, we use Cross-Entropy Loss between the predicted probabilities $\hat{s}_i$ and ground-truth probabilities $s_i \in \{0, 1\}$.

$$l_{\text{EI}} = -\sum_i s_i \log \hat{s}_i + (1 - s_i) \log (1 - \hat{s}_i)$$

For the Inference task, we use a similar setup with the span and hypothesis fed into a Transformer-based encoder follower by an MLP decoder with softmax activation to provide $\hat{y}_E, \hat{y}_C, \hat{y}_N$ as predicted probabilities of the "Entailed", "Contradicted" and "Not Mentioned" labels. Even though the "Not Mentioned" label does not require

evidence, it has been incorporated into the model in the sense that all spans have ground truth negative likelihoods $s_i = 0$.

The loss function of this task is also the Cross-Entropy loss function:

$$l_{\text{NLI}} = -\sum_{L \in \{E,C,N\}} y_L \log \hat{y}_L$$

3 Dataset

We have used the dataset provided at the StanfordNLP site as part of the paper by Koreeda and Manning (2021), which contains 17 hypotheses and 607 contracts which are all Non-Disclosure Agreements. The spans are also provided as start token indices for each document.

The annotated data has label distribution among hypotheses as shown in Figure 1.

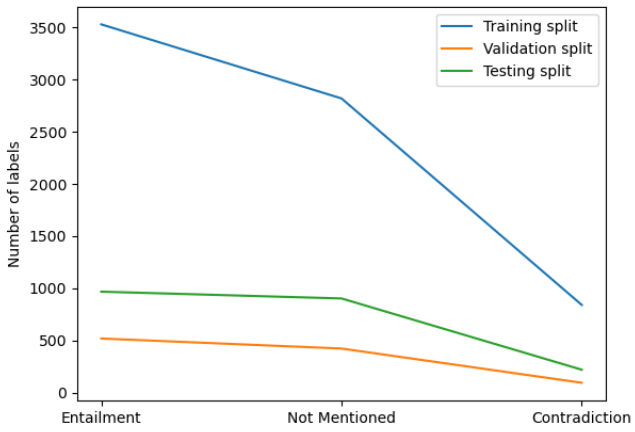
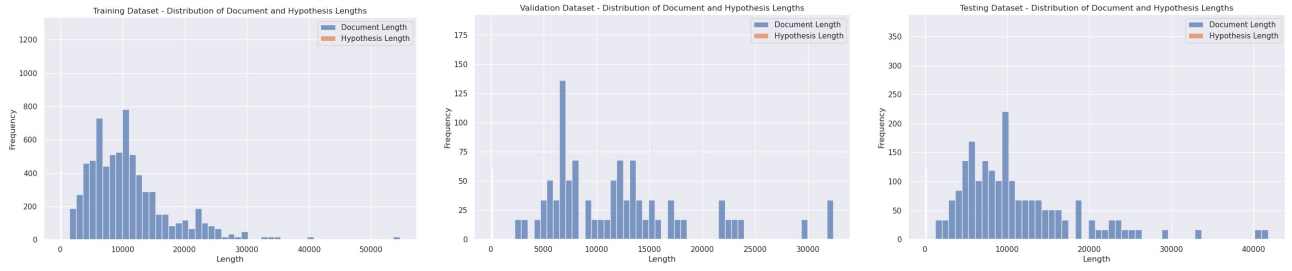


Figure 1: Division of labels among hypotheses

Additional statistics of each dataset are given in Table 1 and Figure 2:

Statistic	Training Split	Validation Split	Testing split
Number of documents	7191	1037	2091
Mean document length (characters)	11049	12095	11218
Mean Hypothesis length (characters)	97	97	97

Table 1: Statistics of dataset splits



(a) Training split (b) Validation split (c) Testing split

Figure 2: Distribution of document length in data splits

4 Methodology

4.1 Model architecture

We have run each subtask using distinct models and obtained distinct accuracy scores for each.

Additionally, for the first subtask (NLI) we have used multiple Transformer-based encoder models as bases and compared the resulting accuracies. For the second subtask (EI), we used two distinct models, one each for identifyig evidence for entailment cases and contradiction cases. In case of "Not Mentioned" inference, the second step is omitted entirely.

The models used for the first subtask are all based on BERT (Devlin et al., 2019) and are listed below:

1. DistilBERT
2. MobileBERT
3. AlBERT

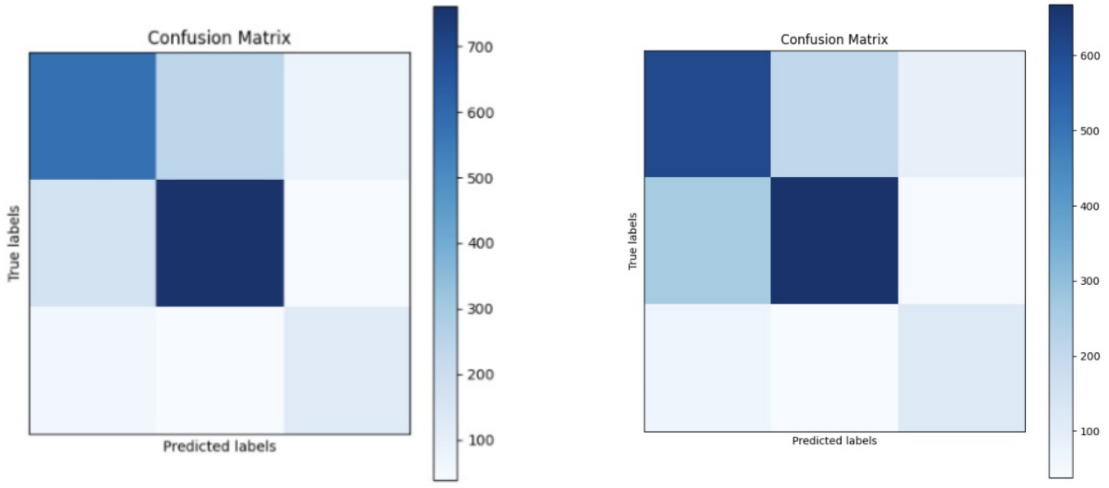
Each model used in the second subtask has the same architecture as described in Section 2, and is used independently of the other. The Transformer-based encoder used is BERT.

4.2 Evaluation

We have used Precision, Recall, Accuracy and F1-scores on the test split for each model as our metrics for the task. The scores for each model are as in Table 2 and corresponding confusion matrices are in Figure 3:

Model	Precision	Recall	Accuracy	F1-score
DistilBERT	0.7019	0.7001	0.7001	0.6982
MobileBERT	0.6889	0.6882	0.6882	0.6865
AlBERT	0.7263	0.7207	0.7207	0.7215

Table 2: Metrics for NLI subtask



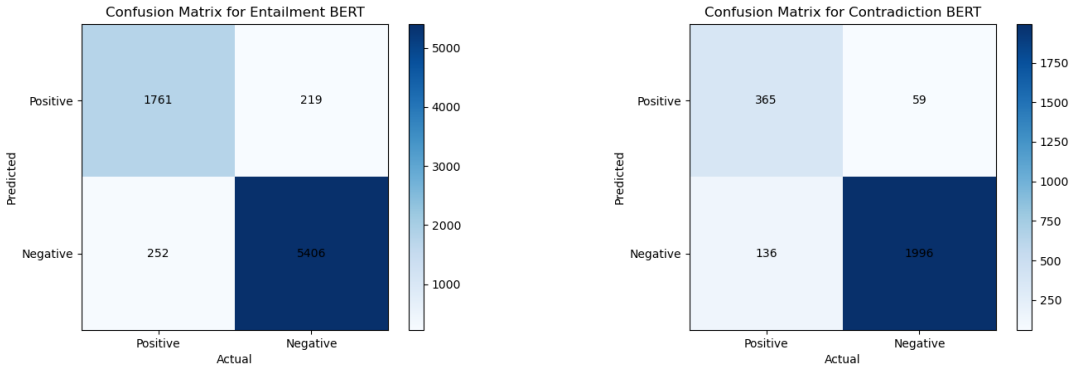
(a) DistilBERT (b) MobileBERT

Figure 3: Confuion matrix heatmaps of models used in first subtask

For the second subtask the results we obtained are as in Table 3 and corresponding confusion matrices are in Figure 4:

Model	Precision	Recall	Accuracy	F1-score
EntailmentEI	0.8748	0.8894	0.9383	0.8821
ContradictionEI	0.7285	0.8608	0.9237	0.7892

Table 3: Metrics for EI subtask



(a) EntailmentEI

(b) ContradictionEI

Figure 4: Confuion matrix heatmaps of models used in second subtask

5 Discussion and Conclusion

Koreeda and Manning (2021) had quoted similar scores for their model architecture. This has been verified through slightly modified code that we ran (obtained from code that was attached in the paper).

This allows for the observation that our modified model, which involved independent models for EntailmentEI and ContradictionEI. The net metrics will evidently be slightly lower than quoted, when run and evaluated as one single model instead of independently, but the conclusion still holds. Additionally, as verified in the Literature review, the architecture proposed by Koreeda and Manning (2021) is the best that was available, our approach provides results very close to the State-of-the-Art (SOTA).

Future Work These results, while close to SOTA, are not perfect. Now without a doubt, larger models (or Large Models) can significantly improve the performance as well as the results. Additionally, we cnojecture that changes in architectures like using Longformers or SentenceTransformers instead of Transformer-based encoders could possibly yield better results.

6 Acknowledgements

We would like to acknowledge the contribution provided by Ada, the high-performance-compute cluster of IIIT Hyderabad, since all models here were trained and run on its compute.

Also, we would like to acknowledge the contribution of the StanfordNLP github account for uploading code related to the paper for public access, since the baseline could be verified using modified code from their repository.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Exigent Group Limited. (2019). *How GCs can thrive, not just survive.*

Koreeda, Y., & Manning, C. D. (2021). *ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts.*

Yang, Z. (2022). LegalNLI: natural language inference for legal compliance inspection..