CS 6120 - NLP Course Project

Grade Contract - Contract Question Answering Using CUAD

Team Members: Anastasia Sakhamuru, Hemanth Sai Madadapu, Khushal Bharatkumar Trivedi, Varun Vikram Sha

Project summary
We are building and evaluating systems for question answering over legal contracts using the CUAD v1.1 dataset. CUAD consists of ~510 real-world contracts and ~13k expert-labeled QA pairs over 41 clause categories (e.g., Governing Law, Payment Terms). Our models will answer natural-language questions by extracting answer spans from long, domain-specific legal text. We will compare several architectures (BiLSTM with attention, fine-tuned BERT-style transformer, and a retrieval-augmented approach) and study their robustness to more conversational, paraphrased questions.

Milestones for a B

To receive a **B**, we will successfully complete the **data + annotation** part of the project and set up at least one basic modeling pipeline.

1. **Data acquisition and basic preparation:** Load the CUAD v1.1 dataset into a usable format, including questions, contexts, and answer spans. Perform lightweight data "cleaning" appropriate for a curated benchmark: verify that answer span indices are consistent with the context text for the portions of the dataset we use. Split the dataset into train/test (or adopt an existing split) and clearly document our splitting strategy.
2. **Data Preprocessing Pipeline:** We will implement a preprocessing pipeline to convert the CUAD dataset into model-ready formats. This includes implementing a **sliding-window mechanism** to handle long contract documents that exceed standard token limits.
3. **Annotation (Paraphrasing):** Create **paraphrased questions** in plain, non-legal English, covering **at least 10 different CUAD clause categories** (e.g., Governing Law, Parties, Agreement Date, Payment Terms). Each paraphrased question will be aligned to an existing CUAD question so that it shares the same answer span. These paraphrased questions will be reserved primarily for **held-out evaluation** (test) to measure generalization to user-friendly phrasing, rather than used for training.

Milestones for a B+

To receive a **B+**, in addition to all B milestones, we will **build and train our baseline model** and perform a first round of comparative evaluation on the original legal questions.

1. **BiLSTM with attention (baseline neural model):** We will implement and train a recurrent neural network baseline (specifically a BiLSTM with Attention)

2. **Baseline Evaluation:** We will report the EM and F1 scores of the baseline model on the original CUAD validation set.

Completion of B and B+ milestones will satisfy the **B+** level.

Milestones for an A-

To receive an **A-**, in addition to all B and B+ milestones, we will **extend the modeling and evaluation**.

1. **Fine-tuned BERT-style transformer.** Fine-tune an appropriate pre-trained encoder (e.g., legal-bert-base-uncased) for span-based QA on CUAD. Compare performance before and after fine-tuning.
2. **Advanced Architecture (RAG):** We will implement a Retrieval-Augmented Generation (RAG) pipeline (or a comparable Generative QA approach) that retrieves relevant contract chunks and generates answers.
3. **Evaluation on paraphrased questions (generalization):** Use the **paraphrased** questions (created for B) as a **held-out test set** to evaluate how well our models generalize to user-facing, non-professional phrasing. For **each implemented model** (BiLSTM, BERT, retrieval-augmented):

    - Report EM and F1 on:
        - Original legal-style questions.
        - Paraphrased, plain-English questions.
    - Compare performance drops or gains between these two settings.

If all of the above are done and documented, including paraphrase-based evaluation, we consider the requirements for an **A-** met.

Milestones for an A

To receive an **A**, in addition to all milestones for B, B+, and A-, we will **systematically analyze and compare our models, relate them to existing work, and provide a thorough written analysis**.

1. **Comparison to existing publications / baselines.** Identify one or more published baselines or reported results on CUAD. Compare our model(s) to these baselines.

2. **Deeper performance analysis.** Perform **at least one structured breakdown** of performance, such as:
    - Performance by clause category (e.g., Governing Law vs. Payment Terms).
    - Performance by question type (original legal phrasing vs paraphrased user phrasing).
    - Performance by answer length or by contract length.

3. **Critical discussion of model trade-offs and Reflection**

    - Strengths and weaknesses of BiLSTM vs BERT vs retrieval-augmented systems.
    - Trade-offs in **accuracy vs computational cost** (training time, GPU memory).
    - Summarizes what we learned about contract QA and paraphrase robustness.

If we complete and document all the above - including a literature comparison, structured analysis, and critical reflection - we consider the requirements for an **A** to be met.