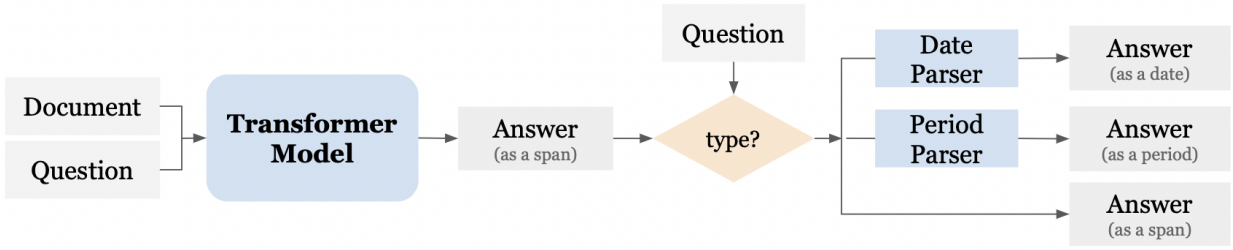
Project MS2

Ahmed Ahmed, Danqi Luo, Hannah Gonzalez, Jiani Huang, Karan Jaisingh

1. Weak Baseline: Transformer + PostProcessing

Our weak baseline utilized the question answering transformer model in hugging face, then we post-process the output into a structural format, such as date and time. The first part is the baseline described in the original CUAD dataset paper, while the second part is for measuring the real answer accuracy rather than just measuring the explanation extraction accuracy. The pipeline is shown below.



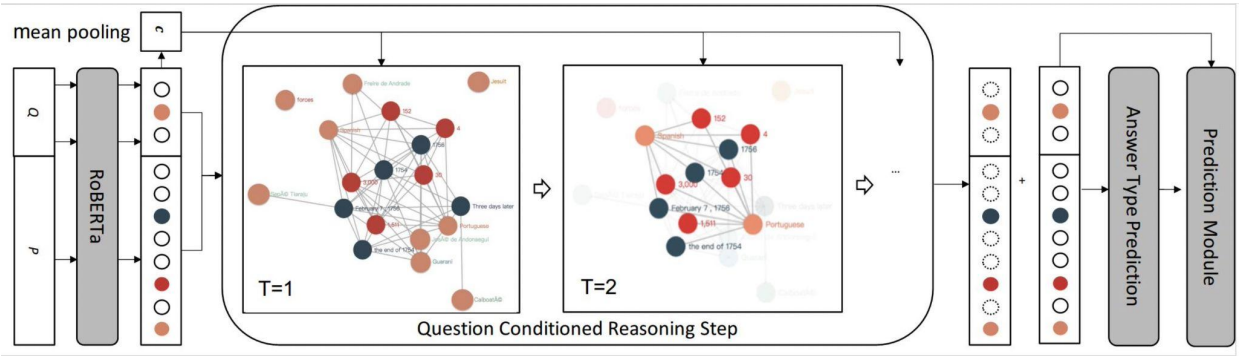
There are two components we add into the pipeline, the date parser, and the period parser. Depending on the question, we choose which parser to use. For example, the "effective date" expects an answer in date type, while the "term period" expects an answer in period type. Both the date parser and the period parser rely on the off-the-shelf date parsing library and regex extraction to obtain the month, day, year information. Further, the period parser also relies on a unit conversion library, so that it can unify different units into the same measurement, such as 1 month to 30.44 days.

The transformer output is far from ideal. Especially in the date-related questions, hardly can this pipeline capture the numerical related information. Here are a few sample answers:

Effective Date	Expiration Date	Renewal Term	Notice to Terminate Renewal	Governing Law
without the prior written consent of the Company	without the prior written consent of the Company	without the prior written consent of the Company	consent shall not be unreasonably withheld	consent shall not be unreasonably withheld
advertisements provided by your Application	continuation in force of the remainder of the term	advertisements provided by your Application	in respect of the affected country or territory only	good faith on the form and content of the disclosure (each acting reasonably
Schedule 13G or Schedule 13D and any future amendments	March 27, 2020	Chinese version	Chinese version	Chinese version

2. Strong Baseline: QDGAT

As described in previous literature, Question directed graph attention network (QDGAT) is a heterogeneous typed graph network model which encodes relationships between entities and numeric values as well as relationships among numeric values themselves. Our strong baseline utilized the model from QDGAT and we applied the model on the eight questions that require numeric reasoning and span extraction from CUAD. In order to do this, since the QDGAT model requires data in DROP format and annotation before processing, we add a data preprocessing and annotation step to the pipeline as well as a prediction generation process.



According to QDGAT, we are only getting around 0.2 F1 score. However, from the prediction results, we can see that the model performs exceptionally well on 'Document Name' extraction and some Date predictions. Although it cannot give the correct answer to many of the questions, it gives the correct types most of the time, i.e. numbers for date and spans for other questions. Here are some examples:

Document Name	Agreement Date	Effective Date	Expiration Date	Renewal Term
WEB HOSTING AGREEMENT	9/9/1997	9/9/1997	9/9/1997	correct any technical problems (Server being down or inaccessible) 24 hours per day , 7 days per week ….
BROKER DEALER MARKETING AND SERVICING AGREEMENT	2013	2013	2013	2013
GLOBAL MAINTENANCE AGREEMENT	3/3/2017	3/3/2017	3/3/2017	BRASILEIRAS S / A (as Company) and AVIONS DE TRANSPORT REGIONAL , G.I.E . (as Repairer) 2015 03 09 AZUL - ATR Global Maintenance Master Agreement DS / CS - 3957 / 14 / Issue 7 Page 1/110 Source : AZUL SA , F - 1 / A , 3/3/2017

Some reasons that are causing the low performance could be the length of passage and question and answer format. For example, the model cannot extract enough context information from the questions being asked since they are short and do not include any entities. In addition, for questions like Renewal Term which expects answers like 'successive 1 year', it is difficult for the model to extract such information from the span.

3. Evaluation Metric

The selected evaluation metric for the given problem is a Macro-Averaged F1 Score. For each question in a given document, an F-1 based scoring metric is computed based on the ground truth answer and the predicted answer. For questions involving dates, we first parse the date string to obtain numeric representations of day, month and year, which are then compared. For questions involving time periods, we parse the time period to obtain numeric representations of time and unit, which are then compared. This metric is then averaged over all questions for the given document to obtain a per-document F-1 score, which is then again averaged over all documents to obtain an overall F-1 score for the model itself. A more detailed breakdown of the evaluation metric can be found in 'scoring.md'.

4. Evaluation Result

The Macro-Averaged F1 score for the baseline model was found to be 0.1538, while the Macro-Averaged F1 score for the QDGAT model was found to be 0.8501613735595093.