Information Extraction on Hong Kong Court Cases

Author:

Yeung Tsz Lok (3035366788) **Yu Tung Chuen** (3035377127) **Yang Lingqin** (3035330404)

Department of Computer Science, The University of Hong Kong

{u3536678, eddiyu98, yanglq}@connect.hku.hk

Supervisor: Prof. **Benjamin Kao**

Date of Submission: 2 May 2020

Abstract

Studying past rulings by tribunals plays an essential role in the daily routine of legal professionals. However, such a task is expensive in the effort for comprehension and is deemed to be tedious. This project aims to study the applicability of Natural Language Processing (NLP) technologies for automatic information extraction in the context of Hong Kong court cases, to reduce the prohibitive cost of reviewing legal judgments. The experiment results display the possibility of information extraction. Moreover, we classified the class of tasks fall within the circle of competence of Question Answering model and tasks which are still immature. Last but not least, we also outline the potential of question answering being extended to other domains and, the future works to be done.

Acknowledgment

We would like to thank our advisor Professor Benjamin Kao for guidance and support. Explaining the rationale and importance of this project, providing high-level research and development directions.

Last but not least, we would like to thank our families and friends for their support and love.

Table of Contents

Lists of Equations

List of Figures

List of Tables

List of Abbreviations

CRF	Conditional Random Field
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
IR	Information Retrieval
NER	Named Entity Recognition
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
Q&A System	Question Answering System
RNN	Recurrent Neural Network

1. Introduction

The theory of Stare Decisis (Latin for “Let the decision stand”), which the Common Law based on, stipulates that the rulings of judges are bounded by similar prior decisions [1]. Thus, scrutinizing prior rulings is an essential research process for legal professions. The research routine involves processing a sizeable amount of text and synthesizes the information court rulings entailed. Due to the dull and cumbersome nature of reviewing court cases, the common practice would narrow down the focus to a finite number of precedents, hoping the most relevant findings could be found in the process and would be accepted in court. Advancements in the field of legal research were rare and limited by the technological constraints of that time, thus, the human-centric researching approach endures.

With the recent advancement in machine learning, especially in Natural Language Processing (NLP). The research group believes that current machine learning techniques have attained an adequate level to assist lawyers in the research process and could potentially automate the extraction process.

With the task of analyzing court rulings automated, legal professionals would liberate from low value-added paperwork and their endeavors could focus on delivering value from high-level thinking, which could enhance productivity in general. Beyond such, legal practitioners could gain better insight, from the full landscape of past rulings, instead of only a selected review of court cases due to limited time and resources.

Most existing works in legal information extraction are statistical-based and rule-based. These works will be applicable to the data preprocessing stage in this project.

The Question Answering affiliated NLP technologies reviewed are statistical methods and deep learning enabled techniques, including vector representation of words, and machine comprehension model. We have concluded that the performance and flexibility of statistical methods are comparatively worse than deep learning enabled techniques. Moving on to deep learning enabled techniques, we elucidated the notion of word vectors and the need for contextualized word embeddings. Introducing transformer-based word vectors, which are contextualized and equip with multi-head attention span, thus, triumphs in the metrics of machine comprehension. Last but not least, we briefly lay out the landscape of machine comprehension models. Explained the mechanism of Bi-Directional Attention Flow (BiDAF), pinpointing three key implementation details, (1) it’s an LSTM, (2) it is bi-directional, (3) it has attention flow to focus on the main points in a paragraph. Thus, giving a succinct idea of the broad picture of the field of question and answering.

A deep neural network is being used in the current project, unlike prior works of legal information extraction. We first scrutinize the nature of our data, by comparing the SQuAD dataset, drug trafficking cases dataset, and free text. The drug trafficking dataset is substantially lengthier than the rest and contains errors due to data preprocessing, but with fewer language variations and language errors within. The SQuAD dataset is much shorter in length but contains more language variations, also of a high standard of quality in terms of language. Free text is also comparatively short and is of a low standard of quality in all other aspects compared to the other two kinds of texts. Then, we discuss the details of the QA models in use, namely ELMo + BiDAF and ALBERT QA. Last but not least, listing out all the training parameters configured.

The performance of the two models was a close match, where ALBERT is slightly better than ELMo + BiDAF. ALBERT attained an exact match of 71.6% and an F1 score of 86.1, where ELMo + BiDAF achieved an exact match of 70.6% and an F1 score of 85.5. The performance increment between ALBERT and ELMo BiDAF is small, compare to the results of the SQuAD dataset. We believe QA models were still inadequate to answer subjective questions, for instance, the motive of the defendant, the personality of the defendant. Thus, due to the nature of the questions itself. 5 categories of questions which the QA model fails to answer are identified, namely false negatives, arbitrarily nature of labeling, incorrect answer due to preprocessing and labeling errors and errors in preprocessing, inappropriate type of questions for QA model and, coreference dependent questions. We evaluate the reasons behind the failure and supplement with examples.

After identifying the strengths and weaknesses of the QA model, there are three applications which QA model could be applied to. They are (1) legal preliminary research: since the key information is extracted, lawyers could read the most relevant judgment instead of skimming through possible related rulings; (2) legal documents summarization: since the most important elements are identified, the task of summarization should be relatively easy. If the technology of natural language generation is mature enough, judgments could be summarized easily; (3) recommendation engine: recommending and ranking judgments based on the percentage of text matching or standard searching algorithms, is far from perfect. With the key factors extracted, a recommendation could be made by nature, instead of text matching. Nevertheless, there are still works needed to enhance the performance of information extraction.

Though the performance of information extraction is acceptable, there are fine details to be optimized. (1) hyperparameters could be tuned for less training time, for instance, if we turn the learning rate up, we could use computational power. (2) Personal Injury cases share a similar level of difficulty with drug trafficking; the success of drug trafficking should be able to replicate in this case. (3) New embeddings have been rolled out since we conduct our research, some produce superior performance over ALBERT, some require less training resources. Conduct research on the new embeddings might have interesting findings. (4) there exists auxiliary dataset applicable to question answering and might be beneficial to legal QA. Conduct transfer learning on these types of data might have a positive impact on the performance of information extraction. (5) Judges refer cases and defendants in different forms, either by case number, its name, or referring as the first defendant. Unifying the form of coreferences would enhance the performance of information extraction. (6) As data preprocessing has caused a noticeable number of errors, it worth spending time devising the structure of judgments and produce a cleaner reference text. (7) Narrowing down the location of answers would accelerate the speed of question answering. We have summarized 7 potential future works that might have a positive impact on legal information extraction.

The main contributions of this paper are as follows:

- Fine-tuned two novel Question Answring Models, namely ELMo + BiDAF and ALBERT QA on the drug trafficking dataset. (Section 3)
- We evaluated the results and categorize the failing cases into 5 categories, give a hypothesize each regarding to the reason they fail. (Section 4)
- Analysison the factors of data convergence. In particular, we investigate whether data or computational power is the main contributor to the performance of the model.
- Last but not least, a list of applications and the future works of the current work is listed, conveying our vision towards the future.

2. Related Work

This section is dedicated to reviewing the prior works of information extraction in the legal field, and NLP technologies that have an affinity with information extraction, namely, vector representation of words, Named Entity Recognition, Question Answering System.

2.1 Existing work on legal research and automated document reviews

Existing publications and projects of information extraction on court cases are based on rule-based systems, often referred to as expert systems. Relying on a fixed set of predefined logic to perform extraction and pattern recognition[2]. The main focus of these models mainly revolving in mining metadata, for instance, date, name of judges, etc. A prominent technique for extracting legal citations would be regular expression[3]. Nevertheless, the emphasis of this project intends to extract information embedded within the unstructured text in court cases. Hence, software developed to accomplish such a goal must be able to adapt to the variability of the language used within the rulings. Though, the prior work could still serve as a reference, when we pre-process the data and perform metadata extraction, a more sophisticated method is needed.

2.2 Question Answering Affiliated NLP Technologies

This section gives an introduction to the affiliated technologies relating to question answering, namely statistical methods and deep learning enabled methods.

2.2.1 Statistical Methods

Statistical methods were the mainstream approach for solving NLP problems, their drawbacks are significant, hence, became less prominent in recent years. Methodologies such as Naïve Bayes (NB)[4], Hidden Markov Model (HMM), Conditional Random Field (CRF), logistic regression, etc. Are all able to solve a wide range of tasks, for instance, CRF was known to prevail in the task of named entity recognition (NER), a task of identifying mentions of named entities in text, subsequently labeling its type. Logistic regression was directly related to Question Answering, serving as the baseline of SQuAD[5]. But the use of the searching algorithm is the main technique for Question Answering in the era of statistical methods, by applying TF-IDF, and BM25 search algorithms[6], etc. Statistical models attained a fair level of performance and were the best at its time. Nonetheless, the emergence of deep learning catapulted the performance of most NLP tasks, leaving statistical methods behind. Furthermore, most statistical methods are limited to a narrow domain, lacking flexibility, unlike the deep learning approach, like Question Answering, which could be applied to a wider range task.

2.2.2 Deep Learning Enabled Techniques

2.2.2.1 Vector Representation of Words

The vector representation of words, or embeddings, is a cornerstone of modern natural language processing, referring to words are being represented in real-valued vectors. The notion of representing words as vectors is that vectors could be processed by mathematical operators, which are the building blocks of computers. The conversion of words to vector mainly stems from the distributional hypothesis: linguistic items with similar distributions have similar meanings[7]. It was famously articulated by the leading figure in linguistics, J. R. Firth, “You shall know a word by the company it keeps”[8]. In short, words that have similar meanings are likely to share a similar context. As a result, words with similar context will be assigned with similar vector values by word embedding models.

The similarity between the two words could be easily obtained numerically from word vectors. The cosine similarity is a prevalent measurement for gauging the semantic distance between words. For two words w_1 and w_2, the cosine similarity is given as follows:

Equation 1 Cosine Similarity

Traditional skip-gram based word embeddings, such as GloVe and Word2Vec, can capture semantics analogy in general. The semantic similarity could be compared at a scientific and data-driven base with the help of word vectors. Nevertheless, word vectors have an apparent pitfall — unable to identify ambiguity of words. For instance, “bank” in a sentence could be referred to as financial institutions that provide credits, in the meantime, point to the river “bank”. The static representation of words is unable to capture the true meaning in context.

The inability of traditional word embeddings in comprehend context gives rise to contextualized word embedding. The first contextualized word representation — ELMo, look into the context, and assign a value to words according to their relevant meaning, by applying a bi-directional Long Short-Term Memory (LSTM) Network. The implementation of bi-LSTM empowers ELMo to comprehend the connections between words in either direction, hence, it could capture the meaning of a word holistically. Results showed that ELMo is ably eschewing the loophole that trapped previous word representation models[9].

The invention of the transformer has taken the performance of word vectors into the next level. The transformer is an encoder-decoder stack with a multi-head attention stack. The exemplar of transformer-based word embeddings is Bi-Directional Encoder Representation from Transformer (BERT). The transformer architecture allows the model to capture the meaning precisely. The multi-head attention mechanism, on the other hand, could model a complex relationship between words. Nevertheless, a known caveat is bidirectional conditioning that would allow each word to see itself indirectly, thus, defeating the purpose of training the embedding. BERT was pre-trained by masking the word itself. Forcing BERT to learn the language structure instead of memorizing the words. As a result, transformer-based word vectors have achieved superior performance in most NLP tasks, such as Question Answering.

Fine-tuning is relatively easy for BERT. The self-attention mechanism in BERT could encode and concatenated text pairs effectively, unlike previous models require encoding text pairs independently before applying bidirectional cross attention. BERT model takes a sentence pair as input, namely sentence A and sentence B, see Figure 1. For fine-tuning a specific task, a sentence pair could consist of the passage and answer in the case of question answering, see figure 2, SQuAD section. Simply plugging the input and fine-tune all the parameters end-to-end would produce a meaningful result. Though fine-tuning is relatively inexpensive when compared to training, the fine-tuning cost is still higher than other less complex word vectors.

There are various transformer-based models rolled out after BERT, A Lite BERT has the best result on the SQuAD leaderboard when we conducted preliminary research[10]. Achieving an F1 score of 90.2 and an exact match (EM) of 83.2% on the SQuAD 1.1, an F1 score of 91.3 and an exact match of 88.6%. Two breakthroughs in the design of ALBERT are the factorization of embedding parameters and sharing parameters between transformers layers. By factorizing the embedding matrix into two smaller matrices, input layers embeddings, and hidden-layer embeddings. Input layer embeddings are responsible for context-independent representations and the hidden layer, context-dependent representations. The separation of work has reduced the number of parameters of the projection block significantly. ALBERT requires less training time but achieved state of the art (SOTA) performance.

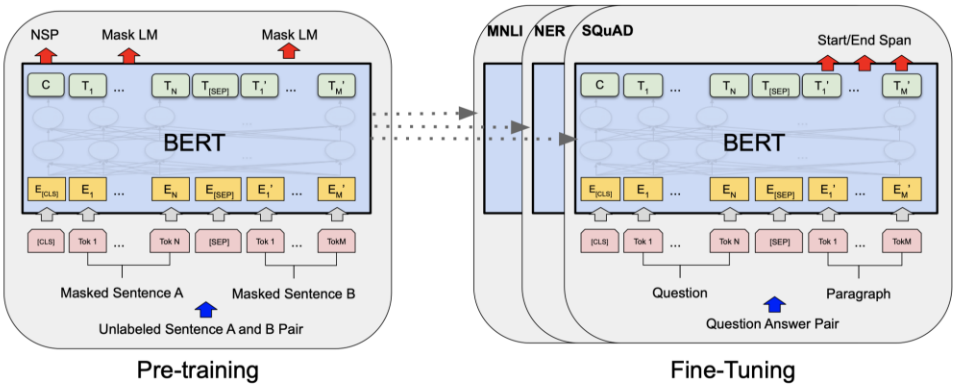

Figure 1 Overall pre-training and fine-tuning procedures for BERT, taken from [9].

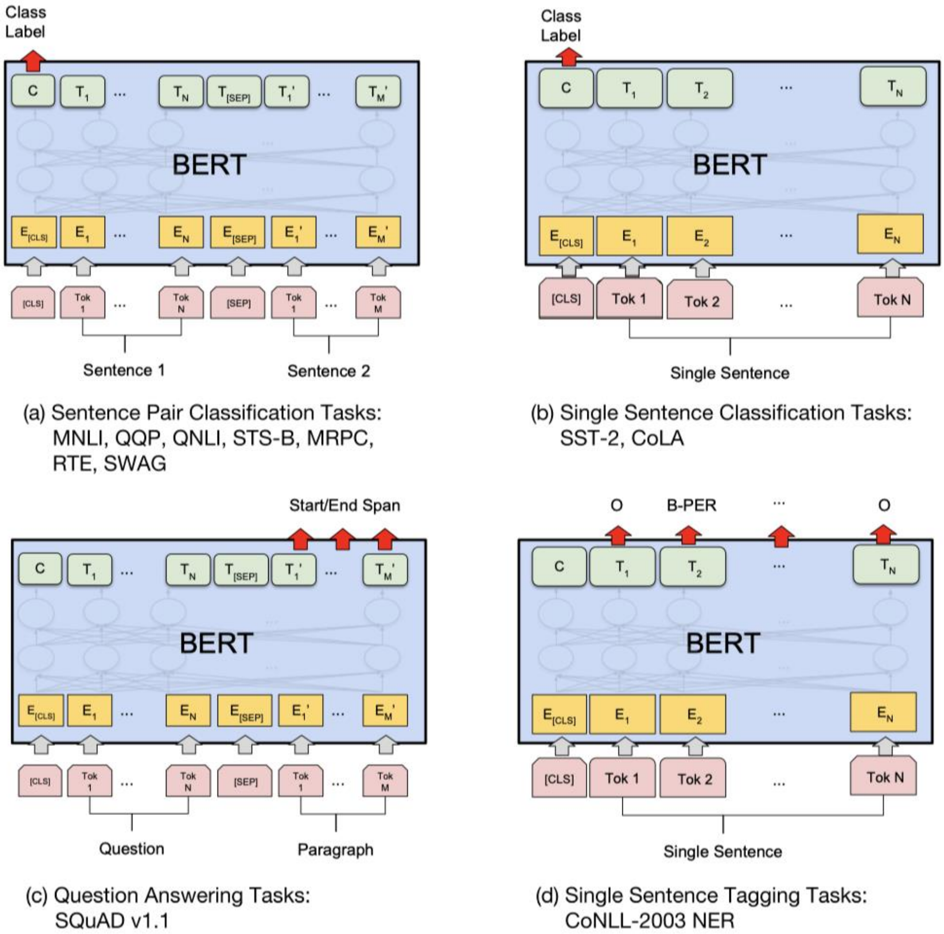

Figure 2 Illustrations of Fine-Tuning BERT on Different Tasks. Taken from [9].

2.2.2.2 Machine Comprehension Models

Machine Comprehension (MRC), is a task attempts to teach machines to answer user proposed questions, either open domain or domain-specific. The approach to this task could be generalized into 2 genres, 1.) Information Retrieval-based factoid Question Answering, and, 2.) Knowledge-based Question Answering. Information retrieval methods are analogous to search, the IR method searches for relevant corpora and paragraphs from a large number of documents, then, extracting an answer from the text retrieved. Knowledge-based methods work like a database query, questions are converted from a semantic format into a structured query, as a result, the answer could be simply retrieved from a structured database[11]. Among all machine comprehension specific models, BiDAF has achieved the best performance.

Bi-Directional Attention Flow Model (BiDAF)

Bidirectional Attention Flow (BiDAF) Model [12] is being considered as the state of the art of machine comprehension models, outperforming other models substantially when the model was debuted.

BiDAF has a complex structure, consist of 5 layers: Embedding Layer, Encoder Layer, Attention Layer, Modeling Layer, and Outputting Layer. In the following, we have summarized three important high-level structure to illustrate the nature of BiDAF.

First, it is a Long Short-Term Memory (LSTM) architecture. Secondly, it has a bi-directional stricture. Last but not least, its name implies that it contains an attention flow mechanism. The features aforementioned above are all designed for improving the performance and avoiding certain pitfalls exists in previously designed models. The details of the model will be elaborated as follows.

LSTM is a special kind of RNNs, a RNN that would discard irrelevant information and learn key features. RNNs are sequence models, their structures are genuinely flexible. They could take each word as input dynamically, adjusting the structure to fit the length of sentences, or even passage. The formation of LSTM has an idiosyncratic nature compare with other RNNs. As shown in figure 4, the previous output of a state is passed on to the next state; hence, each state is taking all previous states as inputs. Overloading with a massive amount of data, the beauty of LSTM lies in its ability to forget irrelevant information. Equip with different gates within the network, LSTM could learn to divide relevant information into long term and short term, discarding the rest[13]. As a result, LSTM could be trained without being overloaded by a vast amount of data and being enforced to capture the most relevant features.

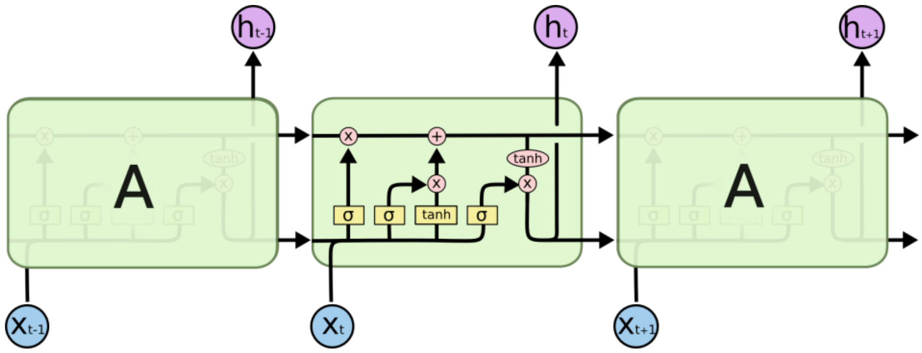

Figure 3 The structure of LSTM

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs

Bidirectional structure in LSTM is another breakthrough that aims to maximize the amount of data a model could process. Most neural networks are forward structured, in fact, from a technical point of view, all NNs are designed in a forward structure. The Achilles' heel of such a design is its inability to capture backward relationships. For example, referring to a previously mentioned concepts in a passage. The aforementioned referral creates a connection between the current sentence and the concept previously appeared sentence. Such sentence structures are eminently common to occur in writings. The incapability of forwarding neural networks substantially degrades the performance of MRC models. As a remedy, two RNNs are in use, as shown in figure 5. One matching relevant answer from left to right (L2R), another matching relevant answer from right to left (R2L). The key to merging the two outputs of these RNNs is a new algorithm "Synchronous Bidirectional Beam Search" [12]. The details of the algorithm are not the focus of our research, hence, interested readers might want to refer to the journal article cited below. In summary, the bidirectional structure in LSTM empowers the network to recognize connections between words in both forward and backward directions.

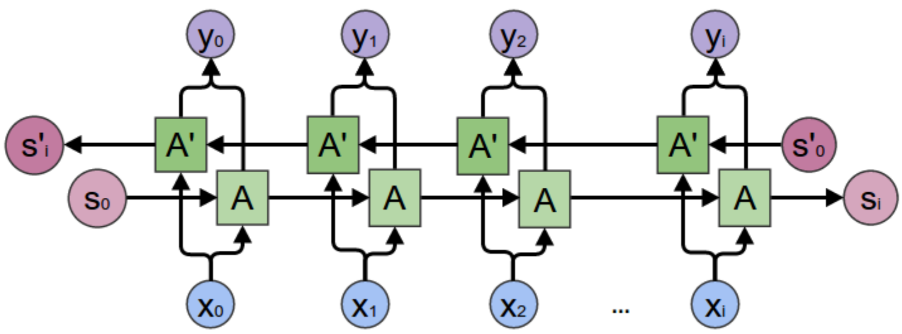

Figure 4 The architecture of bi-LSTM

Source: https://towardsdatascience.com/understanding-bidirectional-rnn-in-pytorch-5bd25a5dd66

Attention flow mechanism improves accuracy and robustness of the task machine comprehension, by mimicking human attention – focuses on relevant key points in sentences. RNNs, in contrast, scan through the whole passage given for identifying the answer in a passage. Even LSTM needs to read through the whole passage, though it discards irrelevant information. In comparison, human works differently, we read only chunks of the passage and digest the read chunks, then move on to another chunk and connect the two. The reading approach of human consumes much less memory and requires processing a significantly less amount of data. This observation sparks the idea of attention flow mechanism in LSTM in particular.

Not all words worth equal attention, or equal weights in a model. Attention is trained to help models focus on the most relevant section of the text to render an answer from the question received. The attention flow mechanism of the BiDAF model has two components, namely, Context-to-query attention (C2Q), Query-to-context (Q2C) attention. C2Q computes the relevance of each query word to each context word. Similarly, Q2C computes the relevance of each context word to each query word. The two results are in matrix form and will be merged by a multilayer perceptron[12]. Then feed into the modelling layer, as shown in figure 6, for yielding the final result. By focusing on the main points in sentences, models are less likely to be confused by an enormous load of data, resulting in higher robustness. Attention flow essentially assigns higher weights to relevant words; thus, better performance could be achieved.

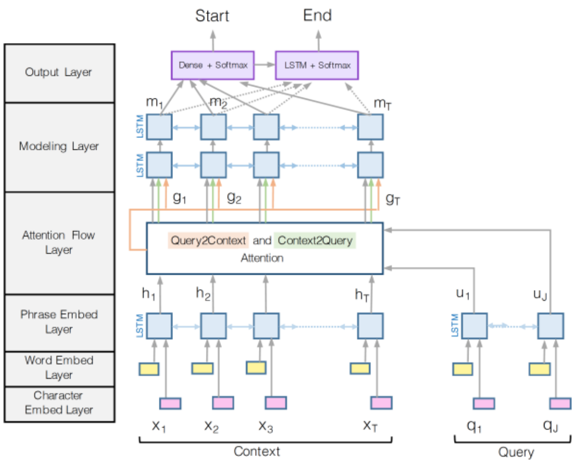

Figure 5 BiDAF model Architecture (Figure adapted from [12])

3. Methodology

Questions and answering is a fundamental block of assessing the ability of reasoning and understanding. Harvard Law School (HLS) holds a similar belief. Harvard Law School is an early adopter of the Socratic method. A form of teaching consisting of asking and answering formulated questions, which originated from Plato's Theaetetus (dialog)[14]. Question and answering play a central role in teaching law; hence, applying the machine comprehension model to legal-related problems would be analogous to The Socratic Method.

3.1 Problem Definition

Question Answering (QA) models are systems that can answer questions formulated in natural language automatically, either from a collection of documents written in natural language or structured information retrieval system. Before diving further into the topic, we must define the terminology associated with the topic. The passage p is the context text the model aimed to search for, significantly longer in length than the rest input in general. The query q is a short text written as a question aimed for an answer within the text. The answer is a text span extracted from the context text. Interested readers in lieu of a detailed description could take the book "An Introduction to Neural Information Retrieval" (2018)[15] as reference.

The task we dedicated to solving is framed as follows, given a Hong Kong court judgment in an unstructured natural language form, the question and answering (QA) model would retrieve an answer span from the correspondence judgment following a natural language query. The output of the model includes the text extracted from the context, the start and end position, probabilities of the answer ––the confidence score. In summary, the inputs of QA models consist of a passage, question, and answer.

The objective function is straightforward and intuitive, the model aims for answering all answers "correctly" within the passage. A right answer in life might be ambivalent, nevertheless, researchers have the subjective discretionary power for deciding the correctness of an answer in the realm of question and answering system. In practice, questions and passage pairs are associated with ground truth labels — the definitive answer assigned or trusted by researchers. Thus, the objective function of the QA model could be simplified as optimizing the accuracy for generating the same pre-defined answer span, when given the same query and passage.

3.2 The Nature and Characteristics of Training Data

	Legal Judgments	SQuAD	Free Text
Length	Longest	Longer in general	Shortest
Out-of-Vocabulary	The least	Slightly more than legal judgments	The most
Grammatical Error	Negligible	Negligible	Noticeable
Variations in languages	The least	Less	Considerably more
Formality	Strictest formality	Formal	Informal
Data Purity	Least pure	Highest	Moderate

Table 1 Comparison of the nature of datasets.

3.2.1 Stanford Question Answering Dataset (SQuAD)

The SQuAD is the first and de facto large QA dataset (over 100k QA pairs)[5], it has been recognized as the golden standard in the task of Question Answering. The SQuAd has two versions, v1.1 and v2.0. Their main distinction lies in 2.0 contains a larger dataset and unanswerable questions. The collection of SQuAD is crafted by college grade crowd-workers from Wikipedia articles; hence, the data is of high quality. Anecdotally, SQuAD has few grammatical errors, Wikipedia paragraphs are generally within a few hundred characters in length, the spectrum of words and the formality align with a typical magazine article. The SQuAD team has good management on the dataset, hence, the SQuAD is the purest dataset among all others. The SQuAD is an exemplar dataset that suits QA training the most.

Reasoning	Description	Example
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes <u>called</u>? Sentence: The Rankine cycle is sometimes referred to as a <u>practical Carnot cycle</u>.
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which governing bodies have veto power? Sen.: <u>The European Parliament and the Council of the European Union</u> have powers of amendment and veto during the legislative process.
Syntactic variation	After the question is paraphrased into declarative form, its syntactic	Q: What Shakespeare scholar is currently on the faculty?

	dependency structure does not match that of the answer sentence even after local modifications.	Sen.: Current faculty include the anthropologist Marshall Sahlins, ..., Shakespeare scholar <u>David Bevington</u>.
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does the V&A Theatre & Performance galleries hold? Sen.: The V&A Theatre & Performance galleries opened in March 2009. ... They hold the UK's biggest national collection of <u>material about live performance</u>.
Ambiguous	We don't agree with the crowd-workers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: Achieving crime control via <u>incapacitation</u> and deterrence is a major goal of criminal punishment.

Table 2: SQuAD questions type. Words in bold are corresponding to reasoning type. The underlined part is the answer. (Table adapted from [5]).

The question and answer pairs of SQuAD are crafted to test models for their ability to reasoning and adaptation to variations. There are 5 types of answering question pairs, namely lexical variation (synonymy), lexical variation (world knowledge), Syntactic variation, multiple sentence reasoning, ambiguous[5]. Some of the questions are designated to assess the adaptability of the model to language variations, some on inductive power on sentences. Reasoning and robustness to a diverse form of text are both key parts in the reading comprehension task, in which SQuAD has considerable coverage.

Reasoning	Description	Example	Question
Negation	Negation word inserted or removed.	Sentence: "Several hospital pharmacies have decided to outsource high risk preparations . . ."	Question: "What types of pharmacy functions have never been outsourced?"
Antonym	Antonym used.	S: "the extinction of the dinosaurs. . . allowed the tropical rainforest to spread out across the continent."	Q: "The extinction of what led to the decline of rainforests?"
Entity Swap	Entity, number, or date replaced with other entity, number, or date.	S: "These values are much greater than the 9–88 cm as projected . . . in its Third Assessment Report."	Q: "What was the projection of sea level increases in the fourth assessment report?"
Mutual Exclusion	Word or phrase is mutually exclusive with something for which an answer is present.	S: "BSkyB. . . waiv[ed] the charge for subscribers whose package included two or more premium channels."	Q: "What service did BSkyB give away for free unconditionally?"

Impossible Condition	Asks for condition that is not satisfied by anything in the paragraph.	S: "Union forces left Jacksonville and confronted a Confederate Army at the Battle of Olustee. . . Union forces then retreated to Jacksonville and held the city for the remainder of the war."	Q: "After what battle did Union forces leave Jacksonville for good?"
Other Neutral	Other cases where the paragraph does not imply any answer.	S: "Schuenemann et al. concluded in 2011 that the Black Death . . . was caused by a variant of Y. pestis. . ."	Q: "Who discovered Y. pestis?"

Table 3: Genres of negative examples in SQuAD 2.0. The bold-faced words are relevant to the reason of unanswerable. (Table adapted from [16])

The unanswerable questions are designed to be challenging in a sense that they are relevant to the passage and exist plausible answers. Questions that come out of the blue with no relevance with the context itself, could be filtered by simple heuristics and TF-IDF, thus, only questions sharing the same or similar topic are selected into the dataset. Aside from context matching, type matching is also crucial. Without a matching type in the context, distinguishing unanswerable questions would be trivial by utilizing type-matching heuristics[16]. Nevertheless, unanswerable questions could be further divided into 7 more categories, negation, antonym, entity swap, mutual exclusion, impossible condition, other neutral, answerable. Each corresponds to a question type that could easily mislead models believing there exists an answer.

3.2.2 Labeled Drug Trafficking Cases

To facilitate information extraction and evaluate the performance of such a task on drug trafficking judgments, a large amount of relevant English judgments is being tagged, consisting of a total of more than 3,000 judgments. Instead of random sampling, cases are selected in ascending order to ensuring its relevancy to the near future; thus, the documents span from 1998 to 2019. The data tagging process was conducted by law students and supervised by legal post-doctorate students. All labeling is cross-validated, hence, the data tagged is of high quality. Legal experts at HKU has conducted research on this matter and identified the ontologies of this problem. The dominating factors of legal could be subdivided into 6 categories, including (1) charge information, (2) drug information, (3) defendant background, (4) mitigating factors, (5) aggravating factors, (6) sentence, the details of the labels are illustrated in the latter table.

Category	Description	Example

Charge information	Relating to the charge, e.g.: the name of the defendant, ordinance involved, charge, etc.	Label: "First charge of defendant 1" Points: "Trafficking in a dangerous drug"
Drug Information	Relevant to the value, type, quantity of the drug involved.	Label: Drug type of the first charge of first defendant Points: "Ketamine"
Defendant Background	The background of the defendant, e.g.: Gender, marital status, health status etc.	Label: The health status of the defendant Points: "At the time of the arrest she was pregnant, but that pregnancy was terminated with the child stillborn."
Mitigating factors	Mitigating factors considered by the judge to reduce sentencing term, e.g.: good character, showed remorse, etc.	Label: Defendant displayed good character Points: "positive good character."
Aggravating factors	Aggravating factors considered by the judge for adding sentence, e.g. commission of offences on bail, trafficking in drugs on the streets, etc.	Label: Commission of offences on bail Points: "on bail awaiting trial for this case in the High Court, you were unwise enough to commit a further offence."
Sentence	The sentence of the defendant, Training Centre Order, Detention Centre Order, imprisonment, etc.	Label: The penalty of defendant 1 Points: "2 years and 8 months"

Table 4 Categories of labels

The labels we included are the salient factors in drug trafficking cases and labels that are sensible to be answered by a QA model. There are altogether 82 features, nevertheless, legal experts pointed out that a selected 12 features out of 82 are the dominating compare to the rest. The 12 features are (1) the weights and (2) types of drug involved; (3) does the defendant plead guilty; mitigating factors, such as (4) defendant shows remorse, (5) drugs are mostly self-consumed, (6) defendant assists in controlled delivery, (7) the defendant gives testimony in court, (8) defendant has a good character; and aggravating factors, including (9) defendant is a refugee claimant, (10) defendant is on bail, (11) the defendant is a persistent offender, (12) drugs are trafficked internationally. The inclusion of such factors in the training and validation dataset is beyond doubt.

Most of the remaining features are being selected, though, the insertion of some features is inappropriate, for instance, are there multiple charges, this could be accomplished with HTML parsing or rule-based system. The recognition of citations is also an ineffective use of a neural network, other means such as regular expression are much more efficient. Nevertheless, for experimental purposes, we included some features not effective for QA system, which proved to be nor effective nor efficient, interested readers might refer to the evaluation section.

The characteristics of legal judgments are distinct in its type, consistent and formal in the language choices, rarely contains grammatical mistakes; nonetheless, the challenges lie in the data purity after preprocessing and its length. Legal professionals are known to be the cream of society, holding a high, if not the top standard of English within the society of Hong Kong. Therefore, it is not a surprise that the mistakes contained in the dataset are syntactical errors, for instance, minor misspellings. Hence, we mainly concern about the length of judgments, a significant portion of them contains more than thousands of words. QA models are designed and train on a much succinct version of corpora. Aside from such, there consists of minor flaws within the dataset due to preprocessing. An example would be an unexpected newline, separating a connected paragraph into two, reducing the available context information to the QA model. In summary, the legal judgments data is in high quality, but there are still imperfections due to its length and minor defects.

3.2.3 Free Text

Free text is generally being regarded as daily text, it is not being used in this project, yet, serves the purpose of drawing a comparison with the dataset in use. Corpora of free text are collected from forums, e-commerce sites, social media[17]. The text of these datasets is much shorter in size. For the reason they are user-generated data, the language choices are much more diverse than the other two datasets. The spectrum of the writers' background is considerably wider than the two aforementioned datasets, consequently, the text generated collected are mostly informal, potentially contain grammatical mistakes jeopardizing interpretability. Social media posts and e-commerce reviews occasionally appear to be scammers, further, degrading the credibility of the dataset[18, 19]. Nevertheless, the structure of the text is relatively simple, thus, the dataset gathered is typically well organized. The data size of such a class of datasets is substantially larger than researcher crafted datasets, for instance, the ELI5 dataset contains 270k questions, in comparison SQuAD 1.1 only contains 100k questions[19]. The benefits and pitfalls of free-text data are apparent, it is large in size, favorable to accuracy in general; meanwhile, generated in daily usage, capturing the most natural human usage. Nevertheless, the reliability of such data remains questionable.

3.2.4 Concluding Remarks

After comparing the SQuAD dataset, drug trafficking dataset, and free text datasets, we have briefly outlined the nature of data and the possible challenges that QA models might encounter in the process of answering these questions. Due to the limitations of technology, datasets must be further transformed into the format which fits the QA model.

3.3 Data Preprocessing

Due to memory constraints, a sliding window is applied to truncate the paragraph to fit the size of the model. Unlike conventional neural networks, such as CNN, sequence models, such as LSTM models, store all the states of each token for backpropagation. As a result, the memory consumption of an LSTM-liked model is not fixed, but linearly proportional to the length of text as input. As a remedy, we use the sliding window method, splitting articles into non-overlapping passages. When converting drug trafficking labels to SQuAD 1.1, near boundary answer spans are being discarded, while converting to SQuAD 2.0 answer spans cross the boundary is being further split into different sliding windows. As we will see in the evaluation section, the process may truncate useful information for question answering, such as context-dependent information.

Negative examples are generated by pairing questions with paragraphs that do not contain the labels that are added, a standard machine negative examples generation approach[20]. The tagging process is not exhaustive; thus, negative examples might exist valid answers within the passage, with a low probability though. The dataset also inflated into a drastic size, random sampling is needed to reduce the size. The paragraphs of these examples are mostly not relevant to the questions, making them easy by nature for the model to distinguish.

3.4 Approaches

Two approaches of the Question Answering model are being experimented, namely, ELMo embeddings joining with a Bi-Directional Attention Flow network, and ALBERT with a fully connected linear layer with softmax function as logit. These two models are chosen from other state of the art (SOTA) models, because ELMo + BiDAF is the first to implement attention flow, and ALBERT is the QA with the highest exact match (EM) and F1 score at the time we were conducting planning. The implementations are as follows.

3.4.1 ELMo + BiDAF

The rationale of ELMo and BiDAF has been discussed in the previous section. The set-up of this project follows closely with the question-answering section of the ELMo dissertation [21]. Except in the dissertation, ELMo embeddings were concatenated both in the character embed layer (x_1, x_2, …) and the hidden state from phrase Embed Layer (h_1, h_2, …). This project only differs in feeding ELMo at the bottom of the graph, but not other layers. The passage is converted into ELMo and feed into BiDAF, the model outputs are the start position and end position, with logit value associated. The set-up aligns closely with its original paper; thus, we could expect homogeneous performance on the same nature of questions.

3.4.2 ALBERT QA model

ALBERT is a bidirectional encoder representation from transformers without task-specific architecture in its design, the pre-trained embedding could be fine-tuned with additional output layer to perform a wide range of task —— Question Answering in this case.

The input of ALBERT is a concatenated sequence consisting of the question and the passage. ALBERT is in use of a technique of segment embedding, to differentiate the question from the passage. The pair is separated with the separation token [SEP] for the model to distinguish the two. The question is named segment A and the passage is named segment B. A special token [CLS] always appears at the front, as the ground truth of unanswerable questions. In this case, the predicted answer start, and end position will both point to 0. The details are shown in the following figure.

ALBERT Input Format:	[CLS]	Who	Is	…	?	[SEP]	[PAD]	involving	1	,	754	.	44	grammes	…
	Question							Passage							

Figure 6 ALBERT input format.

Extracted Question Answering is equivalent to span prediction task; hence, the model consists of two classifiers —— the start classifiers and end classifiers. The tokens of the packed single sequence consisting of the question and the passage go through each transformer layers, the hidden vector of the last layer is passed to a fully connected linear layer, figure 2 illustrates the flow of the model. The objectives for us to optimize are the start vector S and end vector E, both are position classifiers. The probability of the word *i* being the start of the answer could be calculated by the dot product between the hidden vector T from ALBERT and the start vector S, as in equation 2. Then the softmax output will convert the final results to log-likelihood of the start and end position. The end position could be done by replacing the start vector with the end vector.

Equation 2 The probabilities of the start and end position

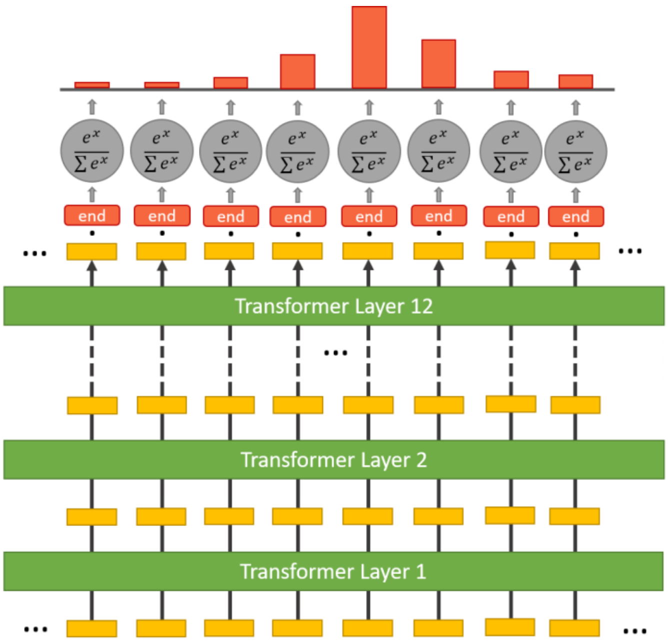

Figure 7 Span prediction architecture of ALBERT

3.5 Parameters of Fine-tuning and the Set-up

The ALBERT QA model was fine-tuned on the Nvidia 12GB P100, with a half-precision performance of 18.7 Teraflops. The hardware set up was at the top of its class, of its time. Nevertheless, memory is still the major drawback constraining the model for longer length of passages.

The two models were both fine-tuned with 3 epochs, a learning rate of 1e-5, and a batch size of 1 unless specified otherwise. The reason behind such arrangements is because, according to our anecdotal observation not reported in this report, the performance of the model peaked at running with 3 epochs. The EM accuracy and F1 score slide slightly for more than 3 epochs. The learning rate of 1e-5 is the golden standard adopted by related works and is widely regarded to be a small increment that could prevent oscillation. Due to the memory constraint, the batch size is limited to 1. When fine-tuning the ALBERT QA with negative questions, the model was only fine-tuned with 1 epoch, because of the enormous size of the negative examples filled dataset.

4. Performance & Evaluation

In this section, we investigate the performance of each QA model extracting key factors from drug trafficking court cases. The evaluation could be divided into two approaches: quantitative, reporting the metrics of the performance, and qualitative, reporting on the reasons for failing and supplemented with examples.

4.1 Evaluation Method

We apply two metrics to evaluate the performance of each QA model: Exact Match (EM) score and F1 Score.

4.1.1 Exact Match (EM)

The EM metric is a binary measure (True/False), which measures the percentage of output string of the model and the answer string matching exactly.

4.1.2 F1 Score (F1)

The F1 score measures the mean of precision and recall. More specifically, the overlap between the prediction and ground truth answer. For example, if the prediction is a subset of the answer (3 words out of 6 words answer), it would have 100% precision, but only 50% recall. And the F1 score would be 0.666.

Equation 3 Equation of F1.

When unanswerable questions are encountered, both the F1 and EM would score 1, if the model predicts no-answer, otherwise, 0.

4.2 Analysis of F1 and Exact Match

ALBERT has better performance on both SQuAD and drug trafficking cases in general, but the superiority of ALBERT is less apparent in drug trafficking cases. ALBERT contains more parameters available for tuning, hence, it is not a surprise that ALBERT could perform with high accuracy (See table 4). However, the disparity in the performance of ALBERT and ELMo BiDAF on drug trafficking cases is noticeably less than the SQuAD 1.1 dataset (See table 3 and 4). Implying QA models might have reached a technical bottleneck on drug trafficking cases. The questions that the QA model unable to answer are either lost useful context information or questionable data purity, more details could be seen in the later section.

Model	Fine-tuned ELMo + BiDAF	Fine-Tuned ALBERT
Exact Match (EM)	70.6%	71.6%
F1 Score	85.5	86.1

Table 5 Evaluation results of the fine-tuned model on the drug trafficking dataset.

Model	ELMo	ALBERT
Exact Match (EM)	78.6%	88.3%
F1 Score	85.8	94.1

Table 6 Evaluation results of QA models on SQuAD 1.1

The reference text and question pairs are irrelevant in general; hence, ALBERT could differentiate unanswerable questions easily, resulting in high accuracy. Even there might exist questions and reference text pair with a correct answer, nevertheless, the number of such cases are rare and overwhelmed by the number of irrelevant questions. As a result, this shows ALBERT is robust to unanswerable reference text.

Model	NoAns_exact	NoAns_F1	NoAns_total
ALBERT Fine-tuned	99.9	99.9	30872

Table 7 Evaluation results of ALBERT on unanswerable dataset of drug trafficking cases.

ALBERT could handle questions regarding facts with a higher accuracy. Questions could be divided into subjective questions and factual questions. For instance, questions regarding gender, penalty, etc. ALBERT could extract the facts with near perfection. In contrast, subjective questions, such as motive, personality, are performing noticeably less accurate. These questions have a considerably less F1 score and a substantial reduction in EM, only an approximately average of 30% of exact match. It is suspected the discrepancy is due to the fact that objective questions have less degree of freedom for labeling. Each data taggers may have different style of labeling, varying in length and the yardstick of appropriate coverage. Confusing the model when learning from such examples. Supplementing cases are provided in the case studies subsection.

ALBERT QA is less adaptive to questions with high variability in language usage. From the table of F1 distribution by question types, starting form plead guilty to personality, there are less exact match, and an increasing number of answers fall into the lower F1 score slots. The language variations of describing plead guilty, motive, personality, are generally much richer than narrating personality, motive, or plead guilty. Hence there is ground to believe that language variation is the main culprit of the inferiority of question clusters consisting of personality to plead guilty.

question type	avg F1	total count	Exact Match	% of EM
personality	0.64	30	13	43.33%
motive	0.67	169	45	26.63%
criminal role	0.68	216	64	29.63%
health status	0.68	43	14	32.56%
family background	0.68	342	110	32.16%
previous criminal record	0.70	433	179	41.34%
education level	0.79	172	78	45.35%
other cases cited	0.79	374	241	64.44%
drug addict	0.79	146	82	56.16%
age	0.82	418	278	66.51%
relationship status	0.82	200	142	71.00%
occupation	0.85	261	196	75.10%
plead guilty	0.86	369	248	67.21%
salary	0.87	76	44	57.89%
nationality	0.88	90	66	73.33%
drugs	0.88	1488	1242	83.47%
sentenced reduced	0.91	361	267	73.96%
starting tariff	0.91	632	521	82.44%
date	0.95	346	321	92.77%
mitigatin factors	0.96	14	13	92.86%
charge of the defendant	0.96	611	475	77.74%
penalty	0.96	459	416	90.63%
sentenced to	0.96	179	159	88.83%
ordinace the charge	0.96	227	195	85.90%
gender	0.98	291	278	95.53%

Table 8 Average F1 and EM grouped by question types.

question type	F1:0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-0.10	Exact Match
personality	4	1	2	1	3	1	5	0	0	0	13
motive	9	10	7	15	6	11	19	14	19	14	45
criminal role	8	15	16	11	14	21	13	9	28	17	64
health status	1	1	2	3	3	8	5	1	4	1	14
family background	23	18	23	14	21	25	23	29	26	30	110
previous criminal record	20	26	38	32	21	26	19	17	20	35	179
education level	2	1	2	2	12	30	13	6	19	7	78
other cases cited	12	13	26	21	14	16	9	7	9	6	241
drug addict	4	4	5	6	6	8	16	6	7	2	82
age	10	1	7	3	51	35	30	1	1	1	278

relationship status	16	4	3	4	5	9	8	5	1	3	142
occupation	9	8	9	6	9	7	8	2	5	2	196
plead guilty	4	2	14	24	7	12	14	6	28	10	248
salary	1	0	1	0	3	1	8	1	15	2	44
nationality	1	0	1	2	3	6	6	1	4	0	66
drugs	99	5	10	1	6	108	13	0	4	0	1242
sentenced reduced	8	0	0	0	3	11	34	0	38	0	267
starting tariff	13	0	9	2	21	26	18	8	12	2	521
date	10	0	0	0	0	7	7	0	1	0	321
mitigating factors	0	0	0	1	0	0	0	0	0	0	13
charge of the defendant	3	1	2	1	1	3	2	28	93	2	475
penalty	2	0	2	2	14	3	3	8	8	1	416
sentence	0	0	1	1	4	5	0	0	5	4	159
ordinance of the charge	1	1	0	2	0	2	4	2	10	10	195
gender	3	0	1	0	0	2	1	0	6	0	278

Table 9 Distributions of F1 score grouped by question types

4.3 Analysis on Data Convergence

The size of data and number of training epochs are the two most important factors contributing to QA model performance. Experiments are conducted to investigate which factor plays as a major contributor or bottleneck to model performance.

Experiment Set-up

The dataset is divided into 10 equal data size randomly. In order to investigate on the impact of 2 dimensions independently. 2 experiments were conducted.

Experiment (1): We trained the ALBERT QA model on each slice of the dataset with 5 epochs.

Experiment (2): we trained the ALBERT QA model on only 1 dataset, 5 epochs each iteration.

4.3.1 Experiment Results

The results in hand suggested that the majority of accuracy improvement comes from computational power, much less data was needed compared to our expectation.

In both cases, over 80 F1 score was attained for one iteration of training and improve gradually in each iteration following. A steady performance increment of F1 and EM could be observed when more data is present. In comparison, the performance increment is much more fluctuating when only 1 dataset was is present. The model performance of the more dataset case is slightly better than experiment 2, nevertheless, only accounting to approximately 1% of EM accuracy, falling within the range of statistical error. As a result, there is evidence showing both cases have achieved similar performance. Most of the parameters are kept constant, except the data size fed into the model. Therefore, it is conclusive that with 10% of data, the QA model could have converged with the same accuracy with the scenario of feeding ten times as much data.

The adaptability of ALBERT QA is beyond the expectation when we conducted planning. More experiments are needed to scrutinize the model performance in a scarce resource scenario. Specifically, the dataset should be further divided into smaller subsets randomly. The lack of such data, any remark on the minimum threshold of data required would be inconclusive and unconvincing. Therefore, the focus of discussion made beyond this point regarding convergence issues will mainly revolve with the future works section.

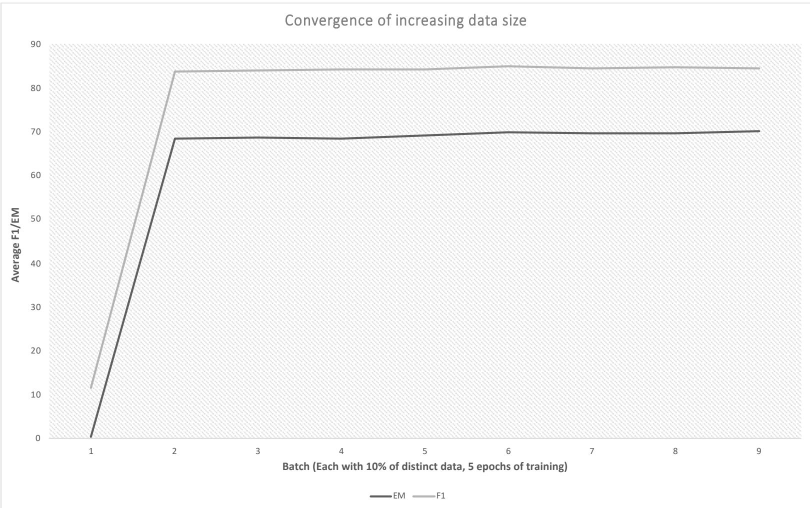

Figure 8 Result of experiment 1 on data convergence of increasing data size.

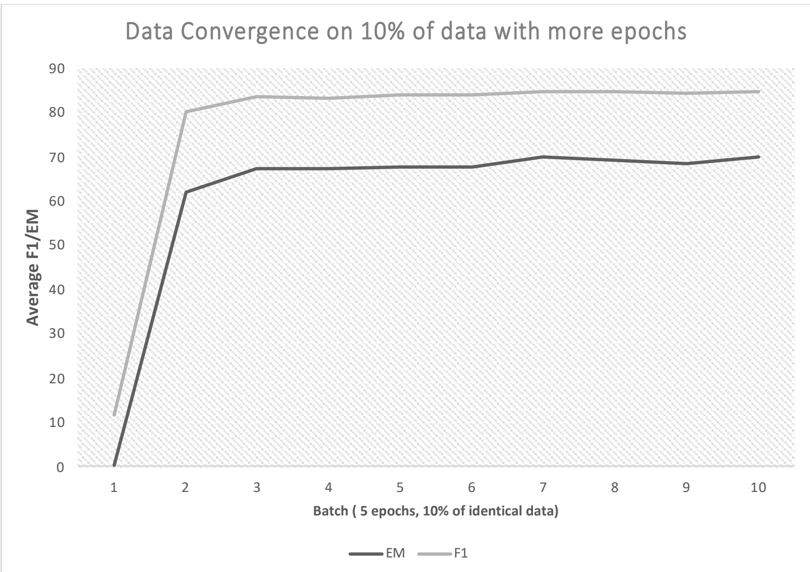

Figure 9 Result of experiment 2 on data convergence on 10% of data with more epochs.

4.4 Case Studies on the performance of the model

The reasons of failing on extracting relevant information vary, but the research team has identified five major categories of failing examples, namely false negatives, arbitrarily nature of labeling, incorrect answer due to preprocessing and labeling errors and errors in preprocessing, inappropriate type of questions for QA model and last but not least, coreference dependent questions.

4.4.1 False Negative Answers

Due to the variability in language and the non-exhaustive nature of labeling, there may exist multiple valid answers within the reference text. Extracting a correct answer different from the ground truth label is indistinguishable from being wrong by the evaluation script. For instance, referring to the following table. The ground truth answer is “2 years 2 months’ imprisonment”. On the contrary, the QA answered “26 months”. The two answers would be counted as identical to a reasonable man. Nevertheless, the evaluation script fails to recognize it as a correct answer.

There also exists some non-identical answer, but still should be counted as correct. Taking example 2 as a reference, when being asked the motive of the defendant, the QA model highlighted a much more detailed answer. “obtain money as you could not find work after returning from the mainland” compared to the ground truth label of “obtain money”. From our point of view, the predicted answer of the ALBERT QA model is equally good, if not a more detailed version of the ground truth answer. As a result, it is reasonable to believe that there is a portion of answers dimmed to be wrong, are indeed correct in a practical sense.

Questions	what is the penalty on the second defendant of the first charge?
QA answer	26 months
Ground truth label	2 years 2 months’ imprisonment
Reference Text	16. Taking all matters into account that I have mentioned, I take as a starting point 3 years and 3 months’ imprisonment, that is 39 months, reduced to 26 months, that is 2 years 2 months’ imprisonment.

Table 10 Example 1 of a false negative.

Questions	What is the motive of the second defendant committing first charge?
QA answer	obtain money as you could not find work after returning from the mainland
Ground truth label	obtain money

Reference Text	8. Mr Chan entered mitigation on your behalf. He told me that you are now 21 years of age, 20 at the time of the commission of the offence. The offence had been committed to obtain money as you could not find work after returning from the mainland. He stressed that you were extremely co-operative with the police on your arrest and that you pleaded guilty at the first available opportunity. When in work you had provided for your family as best you could.

Table 11 Example 2 of a false negative.

4.4.2 Arbitrarily Nature of Labels

As we have seen in the last section, false negative, differentiating the correctness of an answer is discretionary. Thus, those who failed to capture the most important message of the key factors qualitatively should be counted as true negative.

The team has also found answers that are relevant to the questions, but the answer spans extracted do not contain information that assumed to be useful for law practitioners. Taking example 1 as an example, the answer of the QA model, “She lives with her family” could certainly count as a correct without the context of the reference text is a judgment. Nevertheless, the purpose of extraction is to inform legal professionals for their convenience. Living with her family is not a consideration in court. The ground truth label, “She has a younger sister in court today to show her support", clearly outlines the relationship of the defendant with her family, pinpointing the bondage within the family. There is no monopoly of truth in our world, but the ground truth label would be widely regarded as more informative than the predicted answer of the QA model. These relevant but unimportant answers are all classified by the evaluation script as true negatives, hence, it does not affect the accuracy.

The reason for such failure relates to the fact that co-occurrence of the same word appearing in both the question and reference text embeddings. The same word “family” appeared both in the reference text and the question. The linkage of the same word is very strong in attention mechanism; thus, a dominating amount of weights was added to the probability of the candidate answer span surrounding the token “family”. Besides the co-occurrence of the same word, the sparse nature of the question type of family background is also a technical difficulty. The family background type of questions relates to either marital status or number of children, example 2 would be an epitome. However, there are plenty of exceptions, distinct from its own type. For instance, the family background concerned on the defendant’s mother was passed away in example 4 and mentioned the defendant’s parents in example 3. These examples are listed out to illustrate that the labeling of family background is arbitrarily in nature. Even humans could hardly label homogeneously. Thus, the model faced great difficulty when encountering such kind of task.

Questions	What is the family background of the second defendant?
QA answer	She lives with her family
Ground truth label	She has a younger sister in court today to show her support
Reference Text	5. The defendant's best mitigation today is her plea of guilty, she gave an indication of this plea earlier last month. This has saved time and shown her remorse. The defendant herself is only 31, single but been on drugs for the past 10-odd years. She lives with her family and has been supporting herself as a waitress. She has a younger sister in court today to show her support. The defendant has been in custody since April and is convinced she no longer has a drug addiction. She promises herself and her family that she will not take up the addiction again and turn over a new leaf.

Table 12 Example 1 of arbitrarily nature of labels

Questions	What is the family background of the second defendant?
Ground truth label	has a 6-year-old daughter with her ex-husband
Reference Text	The defendant is now 43 years of age. She has no criminal conviction in Hong Kong. Ms. D Crebbin, mitigating on behalf of the defendant, informed me that the defendant was married and has a 6-year-old daughter with her ex-husband. The man who travelled with her, whom she told the police was her husband, was in fact a common-law husband. Since her arrest, the man had left her and cut all ties with her.

Table 13 Example 2 of arbitrarily nature of labels

Questions	What is the family background of the second defendant?
Ground truth label	the defendant has a mother who is 53 years of age who is a saleslady, and a father who is also in his mid-50s who is, sadly, not in the best of health. He is suffering from a serious heart condition
Reference Text	4. In mitigation, I was told that the defendant has a mother who is 53 years of age who is a saleslady, and a father who is also in his mid-50s who is, sadly, not in the best of health. He is suffering from a serious heart condition. The defendant, who was born in Hong Kong, is now aged 26 and had been working as a lorry attendant.

Table 14 Example 3 of arbitrarily nature of labels

Questions	What is the family background of the second defendant?
Ground truth label	Your mother sadly passed away earlier this year in January 2010

Reference Text	14. Your mother sadly passed away earlier this year in January 2010. You had incurred $20,000 of expenses and wanted to repay your debtors quickly and got involved in this wrongful means of repayment.

Table 15 Example 4 of arbitrarily nature of labels

The label concerning family background is being scrutinized specifically in the previous section, nevertheless, such a phenomenon is prevalent among other labels arbitrarily in nature. For instance, we have mentioned personality, motives, health status, etc. The observations in this section are applicable to their cases, in general.

4.4.3 Labeling Errors and Errors in Preprocessing

There exists impurity within the labeling data. For instance, we found that "/" was labeled as sentence reduced for pleading guilty in the table "example 1 of the inconsistent label. This label is nonsensible in any circumstance. Though, cross-validation is applied to our labels. Various errors still exist in our labeling data. This would inevitably confuse the model and degrade the performance of the predicted value.

Case	HKSAR v. LAI HON MAN [2013] HKDC 1012; DCCC 485/2013 (30 July 2013)
Label	Sentence reduced for pleading guilty of charge 1 and charge 2 of defendant 1
Start	51
End	51
Text	/
Reference Text	HKSAR v. LAI HON MAN [2013] HKDC 1012; DCCC 485/2013 (30 July 2013) DCCC 485/2013 IN THE DISTRICT COURT OF THE

Table 16 Example 1 of the inconsistent labeling.

The ground truth answer and QA answer from the table "Example 1 of Errors in preprocessing" seems like an exact match from our first sight, nevertheless, it scored an F1 score of 0. It is possible that there are a few new lines following the value (64,260), confuses the evaluation script. Thus, some of the extracted texts that are flagged as incorrect, indeed are also false negative.

Questions	What is the value of the drugs involved in the first charge
QA answer	HK$64 260
Ground truth label	$64,260
Reference Text	The drugs were subsequently analysed, and the estimated retail value of the ketamine at the time of seizure is HK$64,260.

Table 17 Example 1 of Errors in preprocessing.

4.4.4 Inappropriate Type of Questions for QA model

QA models are designed to extract answer span base on induction and exist a unique correct answer, an answer span with a specific format and multiple correct answers are not the designated type of questions QA models suit to answer.

The category of the question "Other Cited Cases" is an exemplar of the class that is not appropriate for the QA model to answer, since it has a specific format and there might exist multiple correct answers within a single paragraph. For instance, in example 1, "[5] [2012] 2 HKLRD 1121.", "[6] [2015] 1 HKLRD 450.", and "[7] At §33. Also, see HKSAR v Wong Hon Chiu CACC 137/2015" are both other cases cited. But the model could only learn to predict one consistent answer, the result is the other two being marked as incorrect (See example 1, 2, 3).

There existing alternatives to legal citations extraction, mostly involving pattern recognition[3, 22]. These designs could extract legal citations in an accurate, exhaustive, and efficient manner. Unlike the QA model. Thus, conclude the disappointing performance of legal citation QA extraction.

Questions	What is the family background of the second defendant?
QA answer	HKSAR v Wong Hon Chiu CACC 137/2015
Ground truth label	[5] [2012] 2 HKLRD 1121
Reference Text	[5] [2012] 2 HKLRD 1121. [6] [2015] 1 HKLRD 450. [7] At §33. Also see HKSAR v Wong Hon Chiu CACC 137/2015

Table 18 Example 1 of inappropriate question types

Questions	What is the family background of the second defendant?
QA answer	HKSAR v Wong Hon Chiu CACC 137/2015
Ground truth label	[6] [2015] 1 HKLRD 450
Reference Text	[5] [2012] 2 HKLRD 1121. [6] [2015] 1 HKLRD 450. [7] At §33. Also see HKSAR v Wong Hon Chiu CACC 137/2015

Table 19 Example 2 of inappropriate question types

Questions	What is the type of drugs involved in the first charge?
QA answer	HKSAR v Wong Hon Chiu CACC 137/2015
Ground truth label	HKSAR v Wong Hon Chiu CACC 137/2015
Reference Text	[5] [2012] 2 HKLRD 1121.

	[6] [2015] 1 HKLRD 450. [7] At §33. Also see HKSAR v Wong Hon Chiu CACC 137/2015

Table 20 Example 3 of inappropriate question types

Another example would be the type of drugs. Though, this type of questions, involve induction. Nevertheless, multiple answers corresponding to the same question might still confuse the model. Taking example 4 and 5 as examples, this case involved multiple drugs. Hence, existing multiple answers within the same paragraph and the same charge. The model was confused and only highlighted nimetazepam, and the rests are all marked as wrong.

Questions	What is the type of drugs involved in the first charge?
QA answer	nimetazepam
Ground truth label	methamphetamine hydrochloride
Reference Text	C&E officers entered the premises with the defendant and found the following: a plastic bag containing 110 grammes of ketamine; a plastic bag containing 1.86 grammes of methamphetamine hydrochloride, that is the "Ice"; 13 plastic bags containing 1950.87 grammes of cannabis in herbal form; and 13000 tablets containing 87.86 grammes of nimetazepam. And C&E officers also found one set of electronic scales; 207 empty plastic bags; two heat sealers and two vacuum sealers.

Table 21 Example 4 of inappropriate question types

Questions	What is the type of drugs involved in the first charge?
QA answer	nimetazepam
Ground truth label	ketamine
Reference Text	C&E officers entered the premises with the defendant and found the following: a plastic bag containing 110 grammes of ketamine; a plastic bag containing 1.86 grammes of methamphetamine hydrochloride, that is the "Ice"; 13 plastic bags containing 1950.87 grammes of cannabis in herbal form; and 13000 tablets containing 87.86 grammes of nimetazepam. And C&E officers also found one set of electronic scales; 207 empty plastic bags; two heat sealers and two vacuum sealers.

Table 22 Example 5 of inappropriate question types

4.4.5 Coreference Dependent Questions

Coreference resolution is a task of identifying mentions that refer to the same real-world entity. We found some examples that require coreference resolution before the model could render an answer. In example 1, 208, 335, are referring to the case number. This essential piece of information is not supplied to the model. Hence, it is impossible for the QA model to resolve the referral of the mention. Thus, extracting the wrong answer span. This type of problem, lacking sufficient information, is impossible in theory for the QA model to solve.

Questions	What is the starting tariff of the third defendant committing the first charge?
QA answer	21 years
Ground truth label	10 years' imprisonment
Reference Text	The following sentences will be imposed in respect of 208, 10 years' imprisonment. In respect of 335, taking all matters into account, a starting point of 21 years, reduced to 14 years for your plea of guilty. Applying the principle of totality, 8 years will be consecutive and 6 years concurrent. That is a total of 18 years.

Table 23 Example1 of coreference dependent questions

5. Application

In this section, we will briefly discuss the applications of the information extraction on court judgments.

5.1 Legal Research

Legal research at present heavily relies on the full-text search and skimming through the document by well-trained law students. The routine process for a lawyer to conduct preliminary legal research is as follows, the lawyer searches a legal database with relevant keywords, and skim through the judgment for an idea. After grasping the idea of a handful of judgments, the lawyer will drill on a particular judgment for further analysis. The advancement in legal information extraction could assist lawyers, finding the most relevant judgment by nature, instead of the documents that contain matching keywords. Thus, lawyers could directly dive into relevant judgment, skipping the intermediate steps. For instance, in the drug trafficking cases, lawyers could search base on the amount of drug and the drug type involved in the judgment. This would greatly increase the effectiveness and productivity of legal research.

5.2 Legal Judgment Summarization

Legal judgment summary is very time consuming, automatic information extraction could list out the key factors out, hence shorten the time needed. Without the aid of technology, human lawyers or practitioners must skim through the judgment line by line before writing. If empowered with automatic information extraction, the summarizer could have a glimpse over the overall picture of the judgment. The summarizer could better grasp the idea of the judgment with the key factors and the broader picture in mind. Automatic summarization might also be feasible, if the technology of natural language generation (NLG) has significant improvement. At inception, template base NLG conveys little value to the legal community.

5.3 Recommendation System

As aforementioned in the previous section, full-text search, such as TF-IDF is the major recommendation technique in the legal field, nevertheless, the emergence of embeddings coupled with information extraction could bring new potential to the field. The

Airbnb engineering team proposed a search ranking method base on the structured fields to train a new embedding that would have a positive impact on the performance of ranking and searching[23]. Consequently, recommending better or more relevant judgments to users.

6. Future Works

In this section, we discuss the works that are not completed but dimmed to be meaningful in our perspective.

6.1 Hyperparameter Tuning

The increment of model performance hinted sigh of convergence, nevertheless, there are a few parameters available for further tuning for optimizing performance, training time, computation, parameters including epoch, learning rate, batch size, maximum sequence length, maximum query size, doc stride.

The model has shown signs of convergence, nevertheless, there are still chances that the model might achieve higher accuracy with more epochs. The drug trafficking data is considerably less than the SQuAD dataset in size. Thus, more epochs of training might be able to compensate for the imbalanced amount of training data, enabling the model to adapt the format of law documents. The SQuAD dataset has 3 times of labels than our dataset in terms of question and answer pairs. Under the assumption that the performance of our QA model preserves linearity, doubling the number of epochs is likely to tilt the QA model to accommodate the peculiarities of judgments, meanwhile, still leaving a margin of safety from the risk of overfitting the QA model. Nevertheless, more epochs of training are likely to be futile and the implication of increasing the number of epochs inevitably comes with the cost of higher demand for computational power. Thus, leading to a much less efficient training setup.

For the remedy of inefficient training of tuning up the number of epochs, adjusting the learning rate might be served as a balanced approach. Emphasizing on the annotated drug trafficking cases, in the meantime, comparatively fewer are consumed. With the supposition of the performance of model conserving linearity, doubling the learning rate is expected to improve performance, without bearing the cost of overfitting the model and suffering from oscillation. In addition to learning rate adjustment, the batch size is another parameter could save computational power meanwhile preserving accuracy and still keeping overfitting at bay. The batch size was set to 2, due to the limitation of memory constraint of 12 GBs of graphics memory. Nevertheless, when equipped with higher available memory GPU, tuning up the batch size could reduce training time, only a moderate performance reduction in theory and from experience[24]. Hence, tuning on learning rates and batch size possibly could save training time with fewer resources along with little downside.

Aside from concerning training time and resource efficiency, performance optimization also has an important role. Restrained by memory limitation, the of passage the model could skim through, the maximum sequence length (max seq len) was set to be 384 tokens of embeddings. The fixed window limited the amount of context information able to flow into the model for comprehension. Thus, degrading the performance on context-dependent queries especially queries requires coreferencing. A wider scope of visibility possibly could enhance the model's ability to coping with longer paragraphs. Consequently, strengthening it for its current weakness —– coreferencing dependent queries. Doc stride shares a similar circumstance with the case of max seq len, hence, increasing the value of doc stride is presumed would boost the performance of the model. Nevertheless, the length of maximum query length for training should be shrunken for the reason that the team has purposely limited the length of queries, no question exceeds half of the length of the maximum (64 tokens of embeddings). Therefore, shortening the value of max length by 25% could preserve flexibility for future needs, but retaining the benefit of saving memory demands. Thus, concluding the future works of hyperparameter tuning.

6.2 Extension to Personal Injuries Cases (PSLA)

The success of the information extraction of this project could be extended to cases with similar nature. The characteristics of drug trafficking cases are not alone, personal injury cases share some similar nature key factors with drug trafficking cases. For instance, they both require the background of the defendant, namely the age, gender, occupation, etc. QA model performed an impressive job in these tasks since these tasks are context-dependent and have a homogenous quality with the SQuAD dataset, the model has attained a close match with human performance. Thus, there is ground to believe the QA model will perform information extraction on alike nature attributes, with an accuracy at the vicinity of previously reported figures.

Nevertheless, there are features distinct inherently which are less confident to succeed. For instance, within personal injury cases, compensation for pain, suffering, and loss of amenities (PSLA) is prevalent among judgments. The calculation of this term is an aggregate of sub-components, embedded within the text. The judge might cite other cases, casting extra burden to the QA model, sometimes confuses paralegal and law students. As a result, the extraction of the PSLA is expected to witness a performance dip.

In summary, legal natural language processing is an untapped field with ample opportunities, challenging but worthwhile to explore on. The work we have accomplished was far from the full glory of the capabilities of modern computational linguistic tools. The potential for the intersection of two disciplines of law and artificial intelligence is on a long and winding journey, and yet to complete.

6.3 Experiment on new embeddings

The new embeddings introduced in recent terms require less training time, resources, and occasionally better performance. The arrival of the era of transformers catapulted the performance of the machine comprehension model to near-human performance. However, the performance leap has been much less dramatic in comparison to the first transformer model – BERT. The research direction of the QA model now concentrated on the construction of lean models, models with noticeable fewer parameters. Thus, require less training for convergence, with fewer memory demands, and higher speed. Indeed, ALBERT is a lite model of BERT by nature, its name "A Lite BERT" has suggested that. The paradigm of academia has shifted from performance-centric to cost efficiency centric, new models are less likely with drastic improvement on performance, instead, a more compact model is expected. Furthermore, the technique of distillation is gaining traction and has better performance in a smaller model in the inference problem in comparison to the previously dominating pretraining + fine-tuning approach. For instance, the newly released ELECTRA embedding is an exemplar of pre-trained distillation, the result released in SQuAD is promising, attained an F1 score of 90.6, prevailing all existing models. Hence, training time and memory consumption could be reduced significantly.

The trend of efficiency centric and the usage of the technique distillation are likely to be secular and consistent, the training time and resources reduction are material, hence, consideration of deploying such a model in the future is a self-explanatory choice for researchers in the related field.

6.4 Auxiliary Dataset for Transfer Learning

The SQuAD is indisputably the de facto dataset for question answering task, but not the only dataset. The notion of expanding the scope of the dataset has a high chance of enhancing model performance is prevalent among both academia and industry. It was famously summarized and quipped "We don't have better algorithms. We just have more data”, by the prominent researcher Peter Norvig in “The Unreasonable Effectiveness of Data”[25]. Norvig is not the only researcher had such observation and endorse such a belief. Various known researchers including Andrew Ng, have shown similar results and relationships in numerous other researches[26-28]. Hence, learning from more data is advantageous in general.

There exists dataset dimmed to be useful, but due to limitations of time and resources, the plan for performing training them was dropped. For instance, the NewsQA dataset organized and developed by Microsoft consists of questions drafted by humans mainly revolving on news, ensuring the quality of data. The structure is analogous to the SQuAD, given a story (analogous to the context in SQuAD) and a Question, pairing up with the ground truth answer span from the corresponding article[29]. This type of question-answer pairs could help the QA model strengthen its ability on coreferencing dependent problems, which currently inadequate to attain high accuracy. Aside from NewsQA, HotpotQA is another potential training dataset candidate. HotpotQA contains multi-hop problems, which appear occasionally in judgments. Feeding HotpotQA dataset to QA model would strengthen the robustness for a QA model in response to this kind of questions types. In short, many more datasets are available for our disposable ever before, utilizing these resources would have a positive contribution to this subject.

6.5 Preservation of Coreference Consistency within Context Data

A wide variety of mentions to the same entity appear in legal judgment. For instance, the judge may refer to the case, the defendant, or the charge. Though different mention is used, the difficulty of coreference resolution differs. Case number mention almost always mislead the model, because the QA model would have mistaken it is a number, instead of a noun. Therefore, we might want to replace the case number with an appropriate type matching mention to preserve consistency on coreference. Enhancing the performance of the model when encountering different scenarios of mentions.

6.6 Data Preprocessing

In this project, the reference text is converted through simple html2text conversion. Hence, does not involve any domain knowledge and suffers from data cleanness. The HTML judgments are well organized in general; the charges and defendant names are in corresponding tags. Hence, more data preprocessing could ameliorate data impurity with the dataset and life model performance in this regard.

6.7 Paragraph Selection & Ranking

Document question answering remains an open challenge, a key element is narrow downing the passage for extraction. Various techniques are invented for better accuracy of passage ranking. One of the most important techniques is globally normalizing the passage ranker, published by AWS AI Labs[30]. Combining with standard search techniques experimented[31], the time for inquiry on full-document question answering could potentially shorten significantly. Commercialization would be possible in such a scenario.

7. Conclusion

Reviewing court cases is a routine task of legal practitioners that consumes a fair amount of effort. Automating such a task would greatly improve the productivity of legal professionals. After experimenting on two QA models, namely ELMo + BiDAF and ALBERT QA, a satisfactory performance has been achieved. We also concluded 5 classes of questions, which QA models fail to answer in general. Posted our speculation on the reason behind. Apart from reviewing court cases, legal information extraction could also apply to automatic legal summarization and legal documents recommendation. Last but not least, we also pinpointed 7 tasks that require refinement that have a positive impact on the performance of legal information extraction.

Division of Work

Data preprocessing	Yeung Tsz Lok
Fine-tuning	Yeung Tsz Lok
Data convergence	Yu Tung Chuen & Yeung Tsz Lok
Evaluation	Yeung Tsz Lok & Yu Tung Chuen
Report Writing & Presentation	Yeung Tsz Lok, Yu Tung Chuen, Yang Lingqin

Reference

1 WELEY-SMrrH, P.: 'THE SOURCES OF HONG KONG LAW', in Editor (Ed.)^(Eds.): 'Book THE SOURCES OF HONG KONG LAW' (1994, edn.), pp.
2 Cheng, T.T., Cua, J.L., Tan, M.D., Yao, K.G., and Roxas, R.E.O.: 'Information extraction from legal documents', 2009 Eighth International Symposium on Natural Language Processing, 2009, pp. 157-162
3 Kalyanasundaram, R.: 'Exploring the Knowledge of Citations in Legal Information Retrieval', International Institute of Information Technology Hyderabad, 2017
4 Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., and McClosky, D.: 'The Stanford CoreNLP Natural Language Processing Toolkit', in Editor (Ed.)^(Eds.): 'Book The Stanford CoreNLP Natural Language Processing Toolkit' (2014, edn.), pp.
5 Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P.: 'SQuAD: 100, 000+ Questions for Machine Comprehension of Text', in Editor (Ed.)^(Eds.): 'Book SQuAD: 100, 000+ Questions for Machine Comprehension of Text' (2016, edn.), pp.
6 Ittycheriah, A.: 'A Statistical Approach For Open Domain Question Answering', in Editor (Ed.)^(Eds.): 'Book A Statistical Approach For Open Domain Question Answering' (2008, edn.), pp.
7 Faruqui, M.: 'Proposal : Beyond the Distributional Hypothesis' (2015. 2015)
8 Sahlgren, M.: 'The distributional hypothesis', Italian Journal of Disability Studies, 2008, 20, pp. 33-53
9 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in Editor (Ed.)^(Eds.): 'Book BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding' (2018, edn.), pp.
10 Lan, Z.-Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R.: 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations', ArXiv, 2019, abs/1909.11942
11 Martin, J.H., and Jurafsky, D.: 'Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition' (Pearson/Prentice Hall Upper Saddle River, 2009. 2009)
12 Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H.: 'Bidirectional attention flow for machine comprehension', arXiv preprint arXiv:1611.01603, 2016
13 Gers, F.A., Schmidhuber, J., and Cummins, F.: 'Learning to forget: Continual prediction with LSTM', 1999
14 Gersen, J.S.: 'The Socratic Method in the Age of Trauma', Harv. L. Rev., 2016, 130, pp. 2320
15 Mitra, B., and Craswell, N.: 'An Introduction to Neural Information Retrieval' (Now Publishers, 2018. 2018)
16 Rajpurkar, P., Jia, R., and Liang, P.: 'Know What You Don't Know: Unanswerable Questions for SQuAD', in Editor (Ed.)^(Eds.): 'Book Know What You Don't Know: Unanswerable Questions for SQuAD' (2018, edn.), pp.
17 Reddy, S., Chen, D., and Manning, C.D.: 'Coqa: A conversational question answering challenge', Transactions of the Association for Computational Linguistics, 2019, 7, pp. 249-266

18 Srujan, K., Nikhil, S., Rao, H.R., Karthik, K., Harish, B., and Kumar, H.K.: 'Classification of amazon book reviews based on sentiment analysis': 'Information Systems Design and Intelligent Applications' (Springer, 2018), pp. 401-411
19 Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M.: 'Eli5: Long form question answering', arXiv preprint arXiv:1907.09190, 2019
20 Joshi, M., Choi, E., Weld, D.S., and Zettlemoyer, L.: 'TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension', in Editor (Ed.)^(Eds.): 'Book TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension' (2017, edn.), pp.
21 Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.: 'Deep contextualized word representations', arXiv preprint arXiv:1802.05365, 2018
22 Martínez-González, M.M., Fuente, P.d.l., and Vicente, D.-J.: 'Reference Extraction and Resolution for Legal Texts', in Editor (Ed.)^(Eds.): 'Book Reference Extraction and Resolution for Legal Texts' (2005, edn.), pp.
23 Grbovic, M., and Cheng, H.: 'Real-time Personalization using Embeddings for Search Ranking at Airbnb', Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018
24 LeCun, Y., Bengio, Y., and Hinton, G.: 'Deep learning', Nature, 2015, 521, (7553), pp. 436-444
25 Halevy, A.Y., Norvig, P., and Pereira, F.C.: 'The Unreasonable Effectiveness of Data', IEEE Intelligent Systems, 2009, 24, pp. 8-12
26 Ng, A.Y., and Jordan, M.I.: 'On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes', in Editor (Ed.)^(Eds.): 'Book On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes' (2001, edn.), pp.
27 Pilászy, I., and Tikk, D.: 'Recommending new movies: even a few ratings are more valuable than metadata', in Editor (Ed.)^(Eds.): 'Book Recommending new movies: even a few ratings are more valuable than metadata' (2009, edn.), pp.
28 Banko, M., and Brill, E.: 'Scaling to Very Very Large Corpora for Natural Language Disambiguation', in Editor (Ed.)^(Eds.): 'Book Scaling to Very Very Large Corpora for Natural Language Disambiguation' (2001, edn.), pp.
29 Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K.: 'NewsQA: A Machine Comprehension Dataset', in Editor (Ed.)^(Eds.): 'Book NewsQA: A Machine Comprehension Dataset' (2017, edn.), pp.
30 Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B.: 'Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering', in Editor (Ed.)^(Eds.): 'Book Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering' (2019, edn.), pp.
31 Yang, W., Zhang, H., and Lin, J.: 'Simple Applications of BERT for Ad Hoc Document Retrieval', ArXiv, 2019, abs/1903.10972