Master's Thesis

A study on extracting Cause-Effect relations and these application for Why-question answering

DANG HOANG ANH

Supervisor Prof.NGUYEN LE MINH

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

January, 2021

Abstract

From the early of the study about machine learning, natural language processing is one of the main major focusing that all the researcher want to focus on. Not only to understand more of human language and culture, it is also to help improve the computer's understanding and comments and requests from human. Since the 1960s, scientists have been interested in developing a question answering system to help people find knowledge as well as questions that need to be addressed. Most successfully, two early question answering systems during this time period were BASEBALL [1] and LUNAR [2]. Both question answering systems were very successful in their own domain. And thanks to these advancements, we can now rely on the computer to get answers for variety kinds of questions.

The question answering system is split into two types of domain system: opening and closing. The Automatic Response Generating System focuses primarily on factual questions such as who, what, when, where, and when. The reasons behind that is these answers can be extracted directly from the answer-passages based on the relevant main words. Why question answering task is mostly ignored because of the techniques that applied for factoid questions are not suitable and the frequency of why-question is normally lower than the others.

There are a variety of earlier methods suggested for improving the efficiency of answering questions such as concentrating on the causality of linked terms. In the other hand, it also has downside when the answer passage is not clear about the connection and the answer passages are scratched widely in the data.

In this study, I concentrate mainly on two pathways: the use of a new embedding method that is useful for keyword or search expansion, semantic search and information retrieval for learning causality from annotated data. After that, and maybe more important, it perfects the causal relations between the cause and effect pieces, which can benefit greatly from downstream models like LSTM [3] or CNN's [4] or BM25[5], sentence BERT [6] which require numerical inputs in order to provide us with a good idea for answering the model questions.

Keywords: Deep Learning, Question Answering, cause-effect relations, why-question answering, news articles.

Acknowledgement

I thank you for contributing to my accomplishments in academia. My parents, who helped me first of all with affection and empathy. Without you, I could never have reached that degree of success. Secondly, Professor Nguyen Le Minh and my Committee members, all of whom offered patient support during the research process. In conclusion, I would like to pay homage to all the representatives of the Nguyen sensei and Tojo sensi laboratory who accompanied me on this marvelous day. Thank you all for your continuing support.

Author
Dang Hoang Anh

Contents

1 Introduction **1**
- 1.1 Problem Statement . . . 1
- 1.2 Objectives . . . 1
- 1.3 Originality . . . 2
- 1.4 Thesis Outline . . . 2

2 Related works and Background **3**
- 2.1 Related works . . . 3
  - 2.1.1 Extracting causality knowledge method . . . 3
  - 2.1.2 Sentence embeddings . . . 3
- 2.2 Background Knowledge . . . 4
  - 2.2.1 Sequence-to-sequence model . . . 4
  - 2.2.2 Attention Mechanism . . . 6
  - 2.2.3 BERT embeddings . . . 8
  - 2.2.4 Transformer . . . 9
  - 2.2.5 Sentence BERT . . . 11
  - 2.2.6 BERT-base-uncased . . . 13
  - 2.2.7 RoBERTa . . . 14
  - 2.2.8 ALBERT . . . 15
  - 2.2.9 BM25 . . . 15
- 2.3 Data-set . . . 18

3 Methodology **22**
- 3.1 Extracting Cause-Effect pairs from news using causality model 22
  - 3.1.1 Data annotation . . . 24
  - 3.1.2 Cause-Effect extraction . . . 27
- 3.2 Detecting cause through most similar effect . . . 28
  - 3.2.1 Why-question preparation . . . 28
  - 3.2.2 Answer for why-question . . . 28

4 Experiments and Results **30**
4.0.1 Experiments 30
4.1 Experimental Results 31
4.1.1 Phase one 31
4.1.2 Phase two 33
4.2 Result Analysis 33
4.2.1 The first Phase 33
4.2.2 The second phase 35

5 Conclusion **37**

List of Figures

2.1 Illustration of Sequence-to-Sequence model 4
2.2 Bidirectional recurrent network with Attention Mechanism by Bahdanau et al., 2015 6
2.3 Sample is obtained by alignment model 7
2.4 Process of calculating self-attention score 10
2.5 Multi-head attention illustration 11
2.6 Source: http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/ 12
2.7 SBERT architecture with classification objective function . . . 13
2.8 SBERT architecture at inference, to compute similary scores or used with the regression objective function 13
2.9 One article from business section 20
2.10 cause-effect relation pair example 21

3.1 language modeling sample 23
3.2 Phase two model 23
3.3 Article annotation example 25
3.4 Annotated data representation 25

List of Tables

2.1 cause-effect relation examples 19

3.1 Tagging explanation 25
3.2 Coordinating conjunction example 26
3.3 Semantic meaning recognition 27
3.4 Data annotation statistics 28
3.5 why-question generated sample 28
3.6 Phase 2 data statistic 29

4.1 The detail setting of the pre-trained model 30
4.2 Cause-effect relation extraction model accuracy 32
4.3 Sample extracted cause-effect pair 32
4.4 why-question answering experimental results 33
4.5 answer extracted sample 34

Chapter 1

Introduction

1.1 Problem Statement

Providing acceptable and accurate responses to why question answering might be difficult but demanding achievement for natural language processing. Cause-effect relations and also causality are the essential aspect of semantic knowledge for why question answering task in order to retrieve answers from a given data or knowledge based. The question-answering challenge comprises two main questions: factoid, which offers succinct information . The wide open domain has traditionally been moving these fields of study forward.The Standford Question Answering Data-set (SQuAD) [7] is one of the most popular data-sets for question answering tasks at present with over 100 000 examples for factoid question answering (triple of meaning, question, answer). On the other hand the Whyset [8] data set was used for why question answering is 17,000 non-factoide question answering with 850 Japanese why-question and its top-20 answers. The creation of machine learning models in recent years has been fantastic thanks to these data sets and modern hardware. In contrast, Why-question answering still need more elegant approach in research and experiments to archive its finest.

1.2 Objectives

The main objective of my research is make a simple approach to why question answering by creating a cause-effect extraction model bring out accurate responses to why-question which is the highest similar cause to an effect from a text data with improvement from the input model which is applied new embedding tokens by BERT transformer model. This is the most effective embedding and exploit as much information as possible from

the cause-effect relations, which could applied to an information retrieval model.

1.3 Originality

The ideal of focusing on cause-effect relations was implemented along almost every research of why-question answering. While a considerate number of study focusing on popular data-set as mentioned above, the accurate score of them also increase based on simple of advance modifications on original model. I want to reach on a method and test them on my own built data that fit for further research in the future. The answer for Why-question in this result is going to be extracted based on the cause-effect extraction model and in the same hand, being justified by the efficiency of widely recognized information retrieval, similarity measurement methods which are BM25 and sentence BERT.

1.4 Thesis Outline

The major content of this thesis is described as follow:

- **Chapter 2: Related works and Background:** I bring up several characteristics of related data-set. From there I could bring up idea to take advantages of them and built annotated data-set. On the same hand, thanks to that I could come up with several methods to handle the data at its best.
- **Chapter 3: Methodology:** I will display the approach to optimize the data and the training strategy that I applied in the experiments.
- **Chapter 4: Experiments:** This chapter shows experimental conditions, methods of measurement, hyper-parameter tuning and performance.
- **Chapter 5: Discussion and future works:** We analyze the drawback of current methods the current research stream. In the other hand, state several ways to advance for future works.

Chapter 2

Related works and Background

2.1 Related works

There are great number of research in question answering and in why-question answering specifically, but I only mention two of the most influencing approaches to my research:

2.1.1 Extracting causality knowledge method

I want to mention above an approach that closely to my work on cause-effect relations that is the work about weakly supervised multilingual causality extraction from Wikipedia. On this research, they proposed method for extracting causality knowledge from Wikipedia such as *war* to *extermination*. Therefor they can comprehend on knowledge cause-effect pair that could be correct in almost of the cases. In this research, to archive the most capability, they must have a large scale dataset that could applied in to then thanks to that, SQUAD data-set or Wikidump data-base is useful in this situation. Nevertheless, this kind of cause-effect relation could not be accurate in the why question that based on the moment or different circumstance will come up with different cause-effect relations.

2.1.2 Sentence embeddings

This work is an advancement of BERT [9] and RoBERTa [10] which is not requires a massive computational overheard. This modification of the pre-trained-BERT network could have found the most similar pair in the collection. It is well suited for searching for semantic similarity as well as for unsupervised tasks such as feature learning or clustering. Depending on

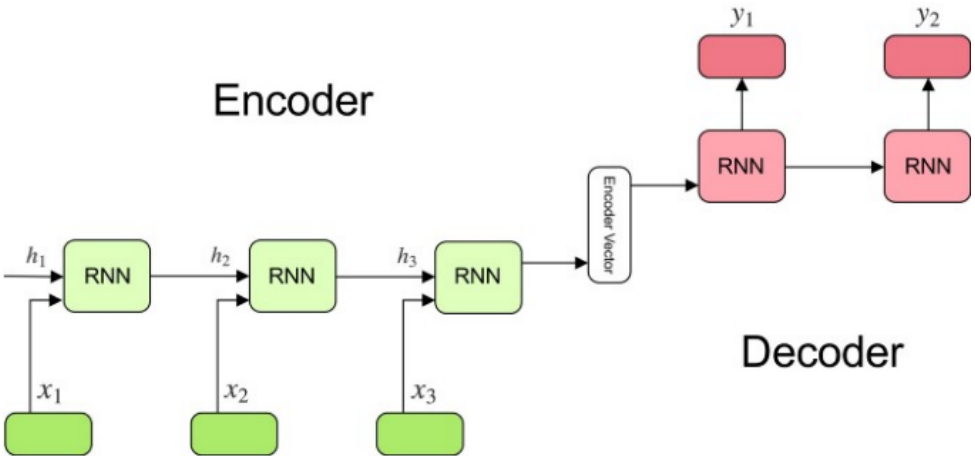

Figure 2.1: Illustration of Sequence-to-Sequence model

this research, I could make a significant improvement on sentence or word embedding with the most computationally efficient.

2.2 Background Knowledge

In order to understand why I move on from sequence to sequence model which is using attention mechanism to Transformer model using BERT embeddings, I will briefly go through all the basic characteristic of them to explain the method chosen later on.

2.2.1 Sequence-to-sequence model

The 2014 Ilya Sutskever implements sequence-to-sequence (S2S) learning [11]. The main objective of this model is mapping the inputs and outputs if not the same length. For instance, 4 words and the output sentence of the input sentence "How old are you?" "年はお幾つですか？" has 8 charaters. The model consists of three parts: encoder, vector and decoder background (encoder). We can verify S2S' high standard look by figure 2.1

1. **Encoder**

    - List of some common recurrent units, such as Long-term memory (LSTM) [12] or Gated recurrent units (GRUS) [13] for effective numbers. Where each cell gets a piece of the input sentence and acquires the information of that piece and transfers it forward to in the next cell.

- The input phrase is the list of tokens of the question when answering the question.
- The following formulation tells us how the secret condition can be calculated:

$$h_t = f(W^{(hh)}h_{t-1} + W^{hx}x_t) \tag{2.1}$$

  where:
  - W^{hh} is weight of recurrent cell.
  - W^x is weight of input cell.
  - h_t is current state.
  - h_{t-1} is previous state.
  - x_t is current input state.

2. Context Vector
   - The last cell in encoder output vector is a context vector that represents the entire phrase.
   - This context vector passes into the decoder's first cell.
3. **Decoder:**
   - Much like encoder, the decoder is provided a set of recurring loads for y_t in the current state t.
   - When answering the question, each decoder cell takes the output vector of the previous state and the word is extracted from the corresponding response.
   - the current hidden state is calculated as formual below:

$$h_t = f(W^{hh}h_{t-1}) \tag{2.2}$$

   - And the output of the current state is calculated as following:

$$y_t = softmax(W^{hy}h_t) \tag{2.3}$$

     where:
     - $W^h y$ is weight at output state.
     - softmax is take a vector as input and give a probability as output.

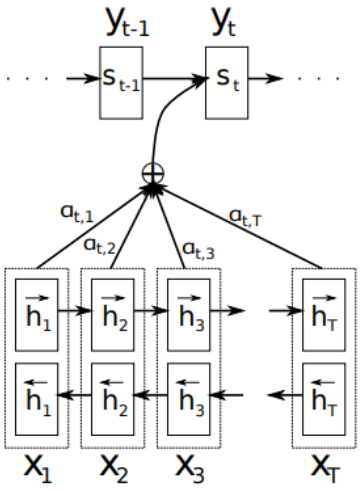

Figure 2.2: Bidirectional recurrent network with Attention Mechanism by Bahdanau et al., 2015

2.2.2 Attention Mechanism

The big problem of sequence-to-sequence model is when it has to handle to the long sentence. The model can not keep the information of very first work in the sentence. It means that the context vector as the input of the decoder may miss some information. This causes the system to make a false prediction.

The attention mechanism [14] is introduced in 2014. This method helps the S2S model hold the information in the context vector better. As content in Figure 2.2, before passing to the decoder, the context vector seems to be linked to all the words in the input. So the decoder can have better information for predicting the result.

Demonstration Firstly, we assume that the input is a sentence x has n words and the output is a sentence y has m words.

$$\mathbf{x} = [x_1, x_2, x_3, ..., x_n] \tag{2.4}$$

$$\mathbf{y} = [y_1, y_2, y_3, ..., y_n] \tag{2.5}$$

Second, for the model S2S, we use bidirectional recurrent network such as LSTM, GRU, etc. Figure 2.2 show that this bidirectional recurrent network has 2 hidden state vector $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$. The most simple way to keep information from two vector is concatenate them together. The ideal of this method is to keep information of the next word and previous word at the current

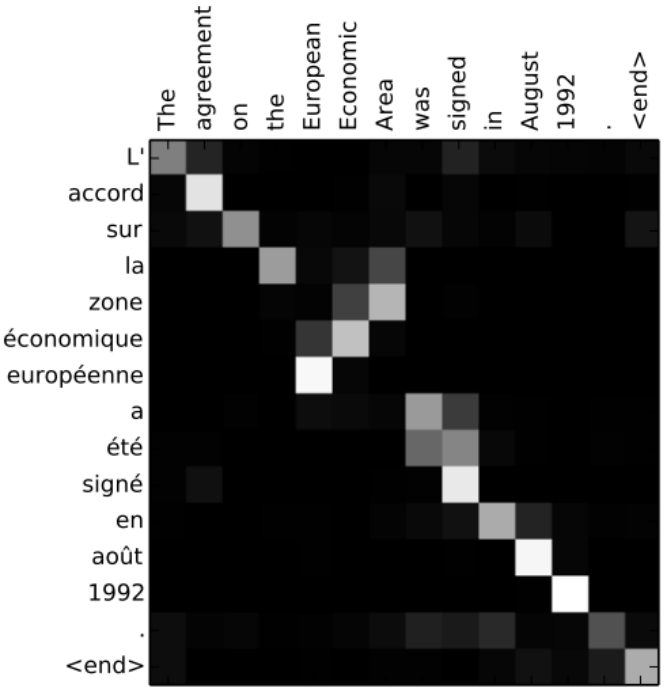

Figure 2.3: Sample is obtained by alignment model

state.

$$h_i = [\overrightarrow{\mathrm{h}}_i; \overleftarrow{\mathrm{h}}_i], i \in [1, ..., n] \tag{2.6}$$

The context vector is obtained by:

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i, t \in [1, ..., m] \tag{2.7}$$

$$\alpha_{t,i} = align(y_t, x_i) = \frac{exp(score(s_{t-1}, h_i))}{\sum_{i'=1}^{n} exp(score(s_{t-1}, h_{i'}))} \tag{2.8}$$

where *align* is alignment model calculate the compatibility between position of input i and the position of output t. And the *score* function:

$$score(s_t, h_i) = v_a^\top tanh(W_a[s_t; h_i]) \tag{2.9}$$

where v_a and W_a is learnt from alignment model and tanh is non-linear activation function.

Through Figure 2.3 [14], We can see the close connection between input and output when attention is applied for machine translation task.

2.2.3 BERT embeddings

BERT is stand for Bidirectional Encoder Representations from Transformer which is released in the end of 2018. From then we can use it as a method of pre-training language representations that was used to create models that natural language processing researchers can then download and use for free. We can use these models to extract high quality language features from our words or sentences data, the experiments could also use fine-tune models on a specific task like question answering, classification and entity recognition, etc.

I will focus on the Embedings. These are moderately low dimensional representations of a point in a higher dimensional vector space. In the same manner, word embeddings are dense vector representations of words in lower dimensional space. Since it was introduced, word embeddings are applied in almost every natural language processing model proposed these days. Its quite obvious the result of its effectiveness.

We can see an example of its in compare with a simple word2vec [15]: *'I like apples'* and *'I like Apple macbooks.'*. word2vec method could captured a static meaning of these sentences buy about the contextualized meaning, BERT word embedding could put them in the right positions. Since then, we could come up with two major benefits showing why we should use BERT embedding in our research:

1. These embddings are useful for keyword/search expansion, semantic search and information retrieval.

2. More importantly, these vectors created by this embedding could be used as high-quality feature inputs to downstream models. Some natural language processing model such as LSTMs or CNNs require inputs in the form of numerical vectors which translate the vocabulary and parts of speech which is vital for why-question answering.

Before feeding into BERT, we have to do some pre-processing step:

1. **Tokenization**: The very first step is to split the input sentence into words, remove noise from data such as "/, * , etc". from the dataset.

2. **Token embedding:** In there work, they use WordPiece [16] to convert token into vector. Moreover, they add token [CLS] stands for classification as the first token of the sentence, and [SEP] token as the ending token of the sentence.

3. **Segment Embeddings**: To distinguish which sentence that words belong to, they use the additional embedding to each word to identify whether A or B is the sentence which contains that word.

4. **Positional Embedding:** The last embedding is used to indicate word position in the sentence.

5. **Final representation:** After conducting three embeddings, the representation for a word from the sentence is a vector by summing three embedding vectors.

2.2.4 Transformer

The main parts of Transformer are the encoder stack and the decoder stack. The encoder stack contains layers of the encoder, and every single layer encoder has two layers inside. The first is a self-care layer, and the second is a feed-forward neural network (FFNN). Also, the decoder has the same structure. However, there is an additional layer between the self-attention and the FFNN decoder that is the encoder-decoder attention. This layer takes the information from the encoder and helps the decoder to memorize the relevant pieces of information from the input sentence.

Self-Attention

Self-Attention mechanism is a method to help the encoder to check all the other words of the input for figure out relevant parts in the sentence. This method can improve the current word has better embedding. For example, we have a sentence:

```
Mary can not go to work because she catches a cold.
```

Without information from other words, "she" here can not have any relation to Mary. This is the reason that self-attention is needed for embedding words form sentence.

To calculate self-attention for each word in the given sentence, we can refer to this formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.10}$$

where:

- Q: Query matrix.

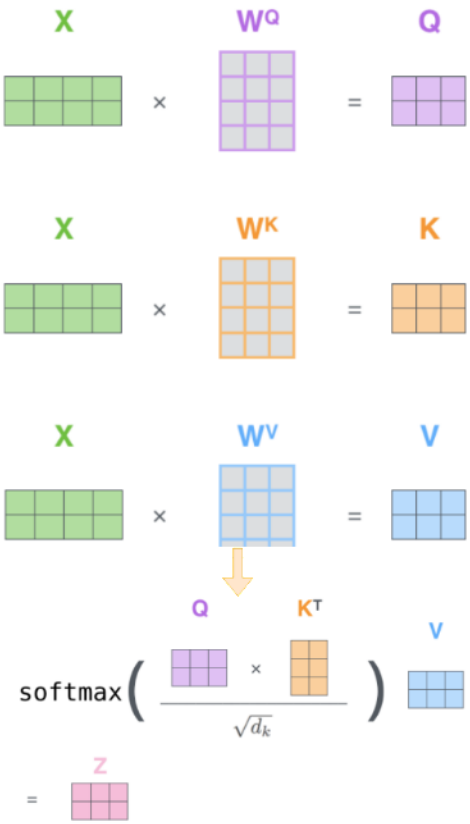

Figure 2.4: Process of calculating self-attention score

- K: Key matrix.
- V: Value matrix.
- softmax: generalized logistic function.
- d_k: the dimension of three vector Q, K.

To obtain Q, K and V vector, we sequential multiply the embedding matrix (X) to matric W^Q ,W^K and W^V, which are pre-trained matrices. Since we have done with matrices, we can apply the formula (2.10) to calculate the attention score for the current position. The process of calculating attention score is showed as Figure 2.4 [1]

[1] http://jalammar.github.io/illustrated-transformer/

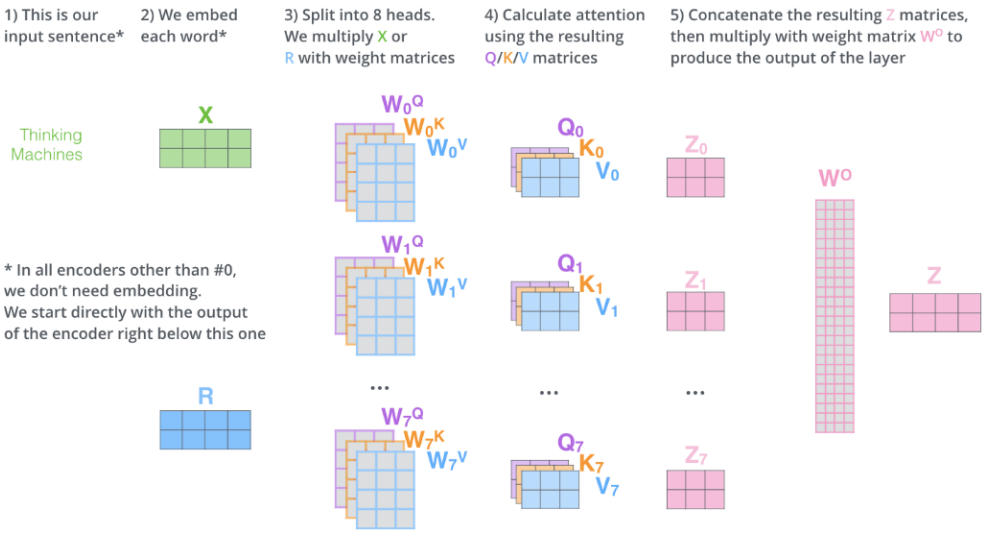

Figure 2.5: Multi-head attention illustration

Multi-Headed Attention

From Figure 2.1, we have a sample of single h head (h = 8 in Ashish Vaswani paper [17]) . It proposed that based on the input sentence, we can experiment eight times with eight different sets of W^Q ,W^K and W^V. To calculates Multi-head attention, we apply these formular:

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_h)W^O \\ head_i &= Attention(W_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.11}$$

where W^O is a weight matrix that was trained parallel with the model. We can have a high-level look of attention layer of Transformer through Figure 2.5 [2].

An encoder-decoder architecture model which used attention mechanisms to forward a more complete picture of the whole sequence to the decoder at once rather than sequentially as illustrated in the figures 2.6:

2.2.5 Sentence BERT

This is a modification of a pretrained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity [6]. A common method to address clustering and semantic search is to map each sentence

[2]http://jalammar.github.io/illustrated-transformer/

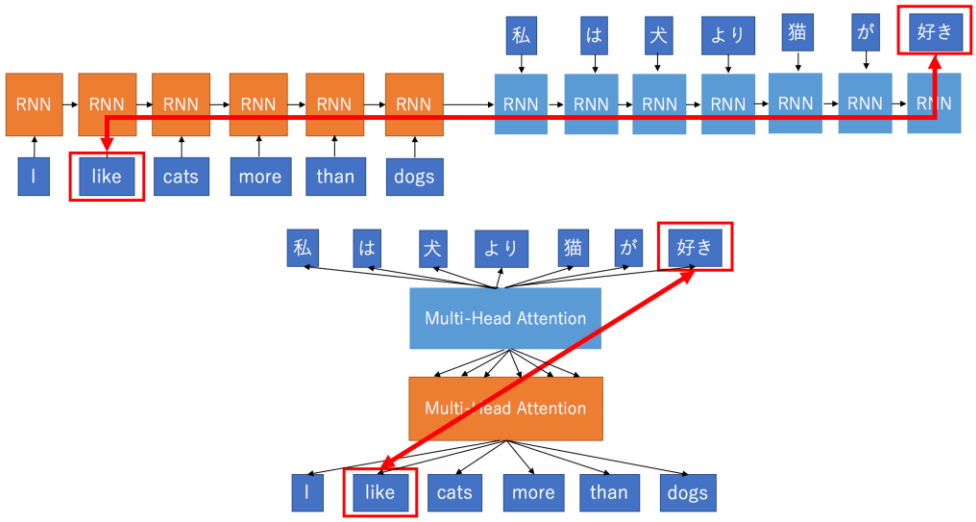

Figure 2.6: Source: http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/

to a vector space such that semantically similar sentences are close. Several approaches have started to bring individual sentences into BERT and to derive fixed size sentence embeddings. The most commonly used approach is to average the BERT output layer (known as BERT embeddings) or by using the output of the first token (the [CLS] token). This was shown that score would be often worse than average Grove embeddings. In order to overcome this matter, SBERT was developed. The siamese network architecture enables that fixed-sized vectors for input sentences can be derived. Using a similarity measure like cosine similarity or Manhatten/Euclidean distance, semantically similar sentences can be found. These similarity measures can be performed extremely efficient on modern hardware, allowing SBERT to be used for semantic similarity search as well as for clustering.

SBERT could be adapted to a specific task. In my research, I want to focus its capable in Argument similarity. The model will be used to show how much phrase will be the same as correct answer. To be considered similar, arguments must not only make similar claims, but also provide a similar reasoning. On the same hand, the lexical gap between sentences could be large, the same with my intentional data characteristic.

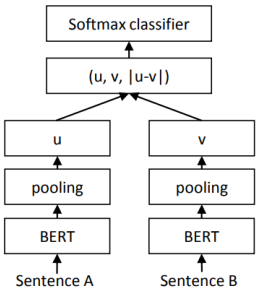

Figure 2.7: SBERT architecture with classification objective function

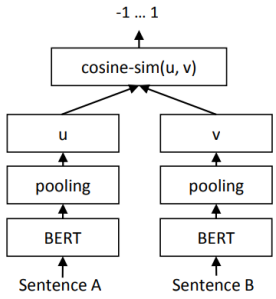

Figure 2.8: SBERT architecture at inference, to compute similary scores or used with the regression objective function

2.2.6 BERT-base-uncased

This is an auto-training transformer model based on a wide set of English results **??**. This means that the basic texts have been pre-trained only and have been classified in no way by humans (which is why much data accessible to the public) for an artificial mechanism to create inputs and labels out of these texts. More importantly, two goals were pretrained:

- Masked language modeling (MLM): The random model masks 15% of the terms within the input for each phrase and then continues through the whole masked phrase and must estimate the masked words. This is separate from conventional recurrent neural networks (RNNs), which typically interpret words one after another, or self-defense models such as GPT, which mask potential tokens internally. The paradigm enables the expression to be interpreted bidirectionally.

- Next estimation of phrase (NSP): models integrate two masked sentences as pre-workout inputs. Sometimes they refer to the sentences of the original text, sometimes not next to each other. The model would then predict whether or not the two phrases obey one another.

Thus model therefore generates an inner representation of the English language, which can then be used to extract the functions that are useful for downstream task: for example, you can use the features generated by the BERT model as input to train a regular classificator when you have a data set of labeled sentences.

I want to use the raw model for either masked language modeling or next sentence, but mostly for the downstream task. See the model hub for fine-

tuned models in a fascinating job. This model is mainly intended to refine functions, such as the grouping of sequences, the token classification, or the interrogation, and use the whole sentence (which may have been masked). You should look at templates like GPT2 for tasks like text generation.

2.2.7 RoBERTa

Thi model [10] builds on BERT and amend existing key hyperparameters, replacing the following pre-workout objective and practicing for even larger miniatures. Pre-training of the language model has led to substantial efficiency improvements but it is difficult to compare different approaches carefully. Training on private datasets of different sizes is computationally costly, and, as we can illustrate, option on hyperparameters has a big influence on the ultimate performance. We are presenting a BERT pretraining replication study which closely measures the influence of several major hyperparameters. We find that the performance of each model reported afterwards is substantially under-trained, and can equal or surpass. Our best model provides cutting-edge GLUE, RACE and SQuAD results. In this researh, I focused on several types of Roberta which are:

- **Roberta For Sequence Tagging Classification**: Pre-training of the language model has led to substantial efficiency improvements but it is difficult to compare different approaches carefully. Training on private datasets of different sizes is computationally costly, and, as we can illustrate, option on hyperparameters has a big influence on the ultimate performance. We are presenting a BERT pretraining replication study which closely measures the influence of several major hyperparameters. We find that the performance of each model reported afterwards is substantially under-trained, and can equal or surpass. This model could provides cutting-edge several famous GLUE, RACE and SQuAD results.

- **Roberta for token classification**: For Name Entity Recognition (NER) activities, for example RoBERTa Model is headed up with a token classification (linear layer on top of hidden state output).The TFPreTrainedModel is heritary. Check the superclass documents for the standardized methods for the whole library model (such as downloading or saving, resizing the input embeddings, pruning heads etc.) This model is also a subset of tf.keras.Model. Using it for general use and actions as a regular model of the TF 2.0 Keras, refer to the TF 2.0 documentation.

2.2.8 ALBERT

This model was proposed to presents two parameter reduction strategies available to minimize memory usage and improve BERT's training speed:

- Divide in two smaller matrices the embedding matrix.
- Repeated layers separated between classes are used.

Growing model size also contributes to better success at downstream activities when planning for natural language representations. At one point though, the GPU/TPU memory constraints, longer training cycles and unexplained model deterioration make more model increases more difficult. We propose two parameter reducing strategies to minimize memory use and improve the BERT training pace in order to resolve these issues. Complete empirical data reveals that our techniques contribute to far higher scale models than the initial BERT. We also utilize a self-controlled loss which focuses on modeling coherence between sentences and consistently helps m downstream tasks. It is generally recommended to embellish inputs on the right instead of on the left. ALBERT is a model with absolute embedding of position. ALBERT uses repeated layers that provide a limited amount of memory space, but the cost of processing is equal to a BERT-like architecture, with the same number of hidden layers as the iterated layers.

2.2.9 BM25

The problem that BM25 (Best Match 25) is trying to solve is similar to that of TF-IDF (Term Frequency, Inverse Document Frequency), which represents our text in vector space (it can be applied to field outside of text, but text is where it has the greatest presence) so that we can search/find similar documents for a given document or query. The essence of TF-IDF is to decide if a document is equivalent to our question by two key factors:

- Term Frequency aka tf: how often does the term appear in the document? 3 times, 5 times?
- Inverse Document Frequency aka idf: measures the number of documents in which the term appears. Inverse document frequency $(1/df)$ then measures the specificity of the term. Is the term a very rare (only one doc) word? Or is it a relatively common one (occurs in almost all the documents)?

Using these two factors, TF-IDF measures the relative concentration of the term in a given document. If the term is common in this article but is

relatively rare elsewhere, the TFIDF score will be high and documents with a higher TF-IDF score would be considered to be very relevant to the search term. BM25 function scores each document in a corpus according to the document's relevance to a particular text query. For a query Q, with terms $q_1, \ldots, q_n$, the BM25 score for document D is:

$$\text{BM25}(D, Q) = \sum_{i=1}^{n} IDF(q_i, D) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i) + k_1 \cdot (1 - b + b \cdot |D|/d_{avg}))} \quad (2.12)$$

where:

- $f(q_i, D)$ is the number of times term q_i occurs in document D
- $|D|$ is the number of words in document D
- d_{avg} is the average number of words per document
- b and k_1 are hyper parameters for BM25

Like all hype parameters in general, defaults are usually a good starting point, and we should probably focus on tweaking other items before jumping into the rabbit hole of the hyper parameter tuning. In the context of the search, it may well be necessary to ensure that our ranking scores older documents lower in the application, such as news ranking. But if we were to start tuning, remember to always measure the performance of the different settings, and the following questions are general points of departure to something that we can refer.[18]

In the task that applies BM25 as a ranking function for retrieval, the values for the internal parameters b and k_1 must be chosen, and also instantiate RSJ. With reference to the weight of RSJ, all the previous observations shall apply. With more details, it could be applied with or without relevance information. If the information is missing, it returns to basic form of idf. In this case, BM25 weight will looks pretty the same as the traditional weight of $tf - idf$. Though the tf component involves the saturation function discussed and is therefore somewhat unlike most other tf functions seen in the equations. where common choices are tf itself and $(1+logtf)$. The latter has a somewhat similar shape curve, but does not have an asymptotic maximum — it goes to infinity, even if somewhat slower than tf itself. In terms of internal parameters, the model offers no instruction on how to adjust them. This might be considered a model limitation. However, given a set of evaluated queries and relevance judgments in the standard retrieval experiment form, it allows for optimization. A significant number of such experiments

have been done, and suggest that in general values such as $0.5 < b < 0.8$ and $1.2 < k1 < 2$ are reasonably good in many circumstances. HThere is also evidence, however, that optimum values rely on other aspects (such as the type of documents or queries).

Published versions of BM25 can vary somewhat. Hence I indicate some differences that might be encountered in different versions of the function in published sources:

- The original had a component for within-query term frequency $q_t f$, for longer queries where a term might occur multiple times. In its full generality, this had a similar saturation function to that used for tf , but with its own k_3 constant. However, experiments suggested that the saturation effect for $q_t f$ was unimportant, leading to a formula which was linear in $q_t f$. In other words, one could simply treat multiple occurrences of a term in the query as different terms.

- The based line algorithm had higher accuracy on correction for document length, to the total document score. This correction was tent to be unimportant.

- A frequently used variant is to include a $(k_1 + 1)$ component to the saturation function. This is the same in all terms, and hence does not impact on the final ranking. The explain for adding it was to help the final formula more compatible with the RSJ weight used on its own. Then a single occurrence of a term would have the same weight in both schemes.

- There are several researchers use specific values assigned to b and k1. A well-know combination is b = 0.5 and $k_1 = 2$. (In another hand, many studies show that it might have a lower value of k1 stand with higher value of b.)

A variety of approaches to information retrieval from basic to complex models have been developed. BM25 is based on a probabilistic set of knowledge. Model that includes paper properties, including term frequencies.[5] The frequencies of the message and the length of the document. Recently, the BM25 may also have a simplified reverse text frequency model. BM25, a traditional basis in the knowledge recovery culture, is one of the most generally used recovery techniques.[19]

2.3 Data-set

Starting with the idea for my research, I want to conduct my research on a highly flexible data-set with variety kinds of relations. Therefor I want to conduct research on BBC datasets [20]. This is a kind of article datasets, originating from BBC news, provided for use as benchmarks for machine learning research.

Considering the benefits from data-set for why-question answering, normally we are expected several work we could do to data as follow:

- *Causality extraction:* The methods of extraction of causality can be classified with respect to what constitutes cause and effect: noun phrases, verb phrases or clauses. The noun and verb phrase type has mostly been addressed by RE methods [21], The clause type has also been studied [22]. The last type is causal embeddings which can be used for causal question answering [23]. we focused here on phrase embeddings which can be combined of verb and noun phraes and hence, this work can provide the most contribution for question-anwering system in general.

- *Relation extraction:* The target relations included "Cause-Effect", Relationship cases whose component entities are not present together in a phrase should be extracted. More recent researches have addressed inter-sentential relation extraction for specialized domains and constructed a large-scale dataset for this task [24]. This is obviously out of range that our experiments can be done so I am not doing this extraction.

- *Knowledge extraction:* The most known research in this kind of extraction is from Wikipedia, there are several studies on extracting class concepts [25], taxonomies [26], infobox contents[27], trivia [28], and various semantic relations[29]. This is also not the kind of relation that we want to extract and insert to out model.

- *Temporal relation extraction:* there is notice that causality extraction and temporal relation extraction share some properties and can complement each other [21]. Possibility of exploiting temporal relation extraction method is still needed further conduct to ensure their effective on question-answering problem.

This data-set is contain five class labels: business, entertainment, politics, sport, tech. This set of articles also consist of 2225 documents from the BBC news website corresponding to the stories from 2004 to 2005. There

cause part	**effect part**
Ad sales	Time Warner profit
Sales of high-speed internet connections and higher advert sales	one of the biggest investors in Google, benefited
offering the online service free to TimeWarner internet customers	increase subscribers
if they join the offshoring drive	they fear that they will damage their brand
because increased responsibility is not going hand-in-hand with more training	problems are occurring
when closing domestic call centre operations	saving money is the main consideration

Table 2.1: cause-effect relation examples

are several standout publications experimented on this data such as topic modeling [30], spam filtering [31], etc.

In this data, we have multiple articles which could use for extracting and conduct research of cause-effect relations. These relations would be difficult to extract cause of multiple situations and connection words between them. For examples, from a business articles, *war trade* will lead to *economy collapse* as well as *new trading potential*. In addition, we have multiple connections words for cause and effect parts in this data-set which could be excellent for highlight the effectiveness of words or sentences embedding methods.

The length of each article is from 10 to 20 sentences average. Articles mention different problems individually and therefore the subjective has different circumstances and motivations respectively. From there, we can have multiple pairs of cause-effect relation which is not duplicate in semantic meaning.

As we can observe from the 2.1, there are different types of cause-effect recognition ways among the data. Normally, they are just chains of sequence events happened. Therefor we could identify them based on their time appearance. Underneath them is the second type of cause-effect which have cause-effect connective words. Based on the manual English Oxford grammar, we can manage to know them as several particular words like 'if', 'because', 'because of', etc.

In several previous approach, answers for why-question answering must be learned as solid facts extracted from knowledge based data like Wikipedia 2.10. The facts that we choose this kind of data to conduct research is to

Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (Â£600m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

Figure 2.9: One article from business section

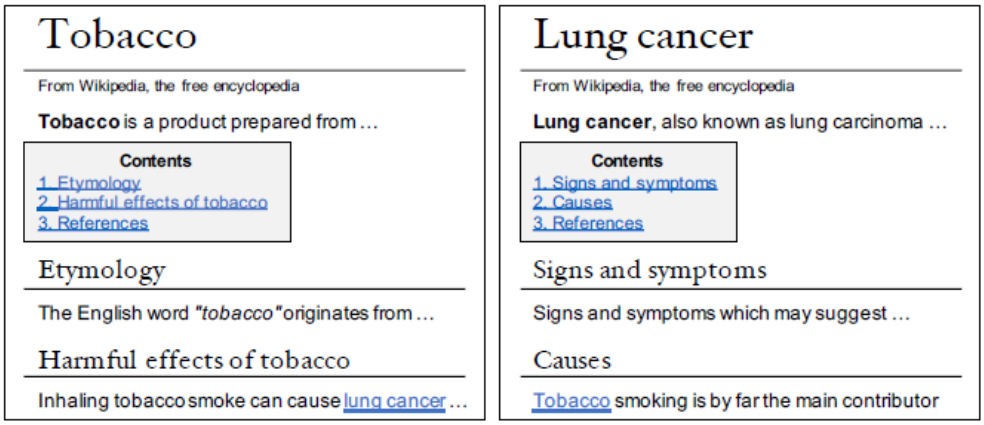

Figure 2.10: cause-effect relation pair example

improve the performance and adaptation of model among variety kinds of research.

Chapter 3

Methodology

In this chapter, we introduce the causality extraction from own built annotation data as the first phase. The next phase is comparing two kinds of indexing to find most similar effect for why-question answer based on BM25. Following this ideal, I divined the thesis into two major parts **??** and the details of each part are displayed in underneath figure 3.1, 3.2.

In figure 3.1, we could observe that from the data-set we mentioned above, we annotated the articles into multiple pairs which include cause phrases and effect phrases, as well as their connection as cause-effect relationships. Into the next step, several BERT-based model will take charge and create a cause-effect relation extraction model. In the continuously figure, we divine the phase into the proposed method and baseline method experiments. The base-line method is using directly the why-question as effect to extract relevant cause (also know as answer for why-question) from data base by BM25 and sentence BERT to calculate the most similar answer as possible for the question. In the other hand, the propose method of ours is using the cause-effect relation extraction method to extract relation between question and articles in data-base and applying BM25 and sentence BERT to find the most similar cause for answer the why-question.

3.1 Extracting Cause-Effect pairs from news using causality model

In this section, I will annotate data and use that into transformer model to bring out cause-effect model that could based on words embedding to extract cause-effect pairs from sentences.

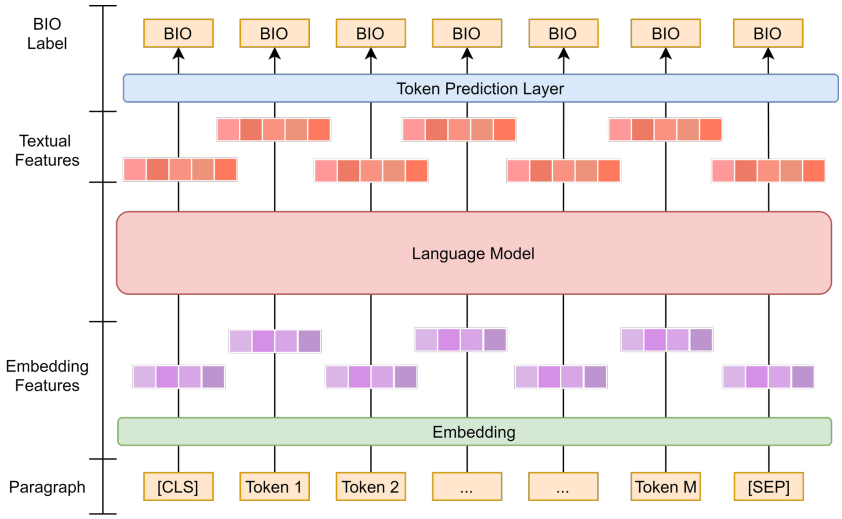

Figure 3.1: language modeling sample

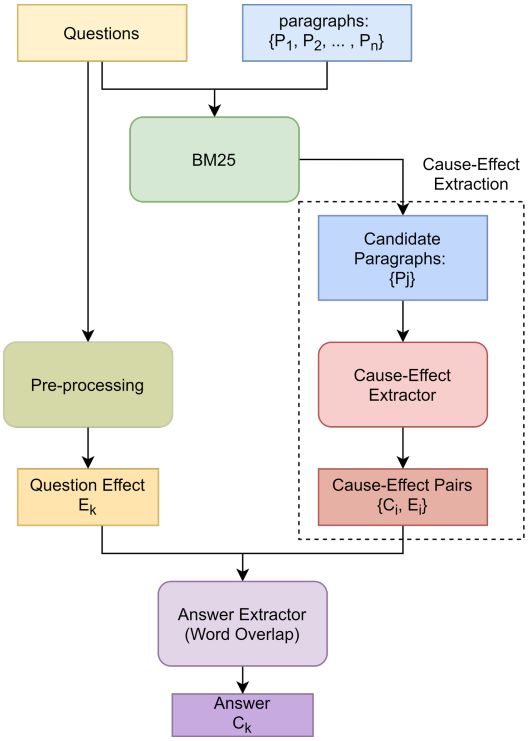

Figure 3.2: Phase two model

3.1.1 Data annotation

First of all, I would use brat which is a web-based tool for text annotation. I focus only on illustrating two basic categories of annotation:

- *text span*: marking words with which classes they are in like person, company, etc.
- *relation*: link text spans or word phrases and note their relation with each other.

An original type of text span category is useful for providing annotations for named entity identification, and binary relationships for uncomplicated relational knowledge extraction tasks, among other.

The annotation of n-ary group which could be combined together is also supported here with brat. It will work with any number of other annotations that take part in a particular position. This annotation category can be used for event annotations. But in my experiment, I only use single target relation between two phrases.

The detail explanation and characteristics of other annotations can be more particularly advanced thanks to annotation attributes, such as making an occurrence as factual or unique objects, as relation to a community or person or by making an entity. This function is not as well as utilize in my research because relation between my phrases is constructed as cause-effect relationship.

Finally, while not the tool's primary target, Brat also allows an annotation to be supplied with free form text 'notes'. Annotation categories, their forms and limitation on their use-age are all entirely configurable. For instance, the fact that a 'Student' relationship must constantly connect 'person' type annotations. There for brat is applicable in almost any text annotation job. Inside the original design of Brat, they also uses natural language processing methods to support attempts at human annotation by integrating a variety of features.

As combinations those advanced features, I have the annotation example as follow in 3.3 and data representation is in 3.4 and explanations for the data marking is as in 3.1

The baseline conditions for me to marking these annotations are cause-effect recognition thanks to English grammar and semantic meaning identification based on English dictionary.

- **grammar recognition**: I use Oxford dictionary of English Grammar (1 rev. ed) book. This book is published online by Oxford University press and written by Sylvia Chalker and Edmund Weiner in 2003. Since

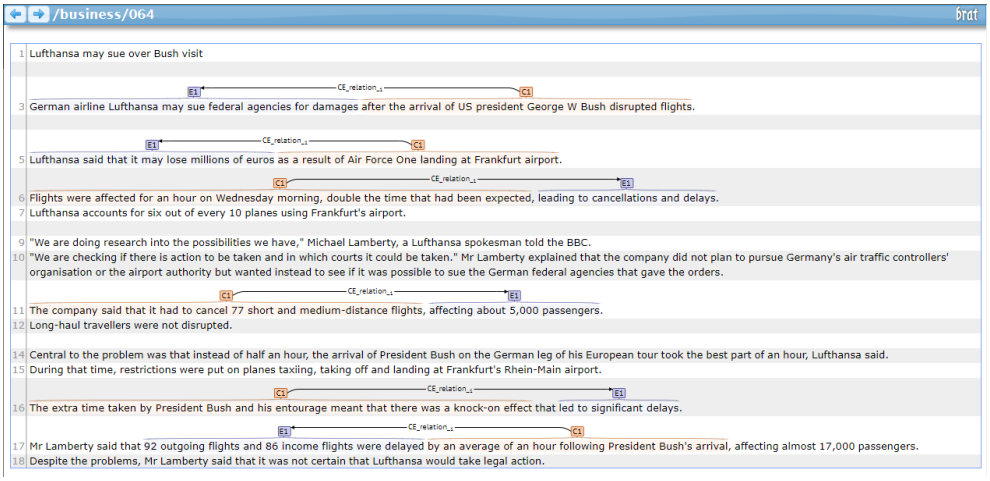

Figure 3.3: Article annotation example

```
T3  Effect1 35 96   German airline Lufthansa may sue federal agencies for damages
T6  Effect1 165 214 Lufthansa said that it may lose millions of euros
T9  Cause1 274 368  Flights were affected for an hour on Wednesday morning, double the time that had been expected
T10 Cause1 920 995  The company said that it had to cancel 77 short and medium-distance flights
T11 Effect1 997 1029    affecting about 5,000 passengers
R7  CE_relation_1 Arg1:T10 Arg2:T11
T14 Cause1 1363 1458    The extra time taken by President Bush and his entourage meant that there was a knock-on effect
T17 Effect1 1513 1567   92 outgoing flights and 86 income flights were delayed
T1  Cause1 97 162   after the arrival of US president George W Bush disrupted flights
R1  CE_relation_1 Arg1:T1 Arg2:T3
T2  Cause1 215 272  as a result of Air Force One landing at Frankfurt airport
R2  CE_relation_1 Arg1:T2 Arg2:T6
T4  Effect1 370 405 leading to cancellations and delays
R3  CE_relation_1 Arg1:T9 Arg2:T4
T5  Effect1 1464 1489   led to significant delays
R4  CE_relation_1 Arg1:T14 Arg2:T5
T7  Cause1 1568 1627    by an average of an hour following President Bush's arrival
R5  CE_relation_1 Arg1:T7 Arg2:T17
```

Figure 3.4: Annotated data representation

Annotation name tagging	Phrase identify	Phrase location	Phrase example
T3	Effect1	35 96	German airline Lufthansa may sue federal agencies for damages
T1	Cause1	97 162	after the arrival of US president George W Bush disrupted flights

Table 3.1: Tagging explanation

the beginning of the 20th century the grammar of the English language has changed greatly and is a topic which could present a complicated minefield of language uncertainties. The reader and the student and instructor will use this open and detailed dictionary to provide easy and immediate access to 1.000 grammar concepts and their definitions. The existing grammar concepts as well as old, common names and controversial new colloquial expressions as well as other language research products are included. Examples of the language in use and regular quotations in current grammar works are presented with succinct descriptions of the broader field of linguistic, including phonetics and transformative grammar. From the knowledge there, I could hypothesize that to recognize cause-effect relation I can based on two subjects which are coordinating conjunction and subordinating conjunction

coordinating conjunction: A coordinating conjunction is a term that combines two components that have the same grammar and syntax. Two verbs, two nouns, two adjectives, two sentences or two separate words may be added to them. For and nor, however, or still and so are the several example coordinating conjunctions. We have here conjunction that connecting two verb, two phrases, two clauses, or a sentence starting with one and conjunction adverbs. there example for my research could be seen in table 3.2

conjunction type	**example**
two words connection	company has bring out *sale for commercial* purpose.
two phrases connection	British Airways has blamed *high fuel prices for a 40% drop in profits.*
two clauses connection	he said that *he was in favour of floating exchange rates because they help countries cope with economic shocks.*
sentence starting with connection	*If you have good management and the right processes in place*, you can make call centres perform anywhere.
adverbs conjunction	we expect the deficit to continue to widen in 2005 *more over* the dollar gets back to its downward trend.

Table 3.2: Coordinating conjunction example

subordinating conjunction: The most straightforward way to describe cause-effect relation is subordinate conjunction. Since it has only a single goal which demonstrate a relationship between a subordinate clause and a main clause, a correlation of cause and effect. A clause itself begins with the justification that it is incomplete. We have several example of what subordinating conjunctions are *for, as, since, therefore, hence, as a result, consequently, though, due to, provided that, because of, unless, as a result of, etc.*

- **Semantic meaning recognition**:

  Sentence meaning: As this part, there are little examples which is not included any of above recognition signatures. Hence I have to carefully checking all the sentences as well as paragraph carefully for semantic meaning that could lead to confirm cause-effect relation exist. There are also number of hints which could combine to serve that purpose like *similar subject identification* and *comma placement and subordinating conjunctions* 3.3

Sample methods	**Example**
Similar subject identification	There are a strong performance of *Nestle* in the Americas and China. Revenue dipped 1.4% to 86.7bn Swiss francs in 2004. Still *Nestle*'s profits margins were helped.
Comma placement and subordinating conjunctions	During the year WPP bought US rival Grey Global, creating a giant big enough to rival sector leader Omnicom.

Table 3.3: Semantic meaning recognition

After annotation, I divine the data into three packets for study and have the statistic as in table 3.4

3.1.2 Cause-Effect extraction

In this section we apply words embedding on the annotated data then bring it up to pre-trained-BERT model to extract highly accuracy experiment results then feed to the next phase. This is an normal step, the state-of-the-art models are overwhelming human performance with large scale dataset. The effects of them on small hand-annotate data might be difficult to reach that great but still showing the sign of acceptable success.

Statistic	Train	**Test**
Number of articles	80	40
Average cause length (words)	8	7
Average effect length (words)	9	10
Total cause-effect relations (pairs)	385	160

Table 3.4: Data annotation statistics

3.2 Detecting cause through most similar effect

3.2.1 Why-question preparation

Based on the grammar about coordinating conjunction and subordinating conjunction I mentioned above. I could use the general knowledge to create Why-question from cause-effect phrase pairs that appeared in the articles and use them to further research. The sample question is created is in table 3.5 and statistic for the data of the second phase is in table 3.6

cause-effect phrase	why-question
Ad sales boosts Time Warner profit.	Why does Time Warner profit boost?
With Indian assets now seen as lees of a gamble, hence more cash is expected to flow into its market.	Why is more cash expected to flow into its market?

Table 3.5: why-question generated sample

3.2.2 Answer for why-question

From the why-question and answer previously, we mark the phrase in the question as effect then use the cause-effect extraction model from the first part to extract the cause-effect relations in the data based by applying BM25 and sentence BERT to find most similar cause for the question why

Statistic	Test
Number of article	40
Average answer length (words)	9
Average question length (words)	10
Total question-answer (pairs)	188

Table 3.6: Phase 2 data statistic

We would like to print our BM25 and sentence BERT significance score for highest results corpus along with their original text, note that this has not been sorted by decreasing order of the relevant score yet. This involves identifying, processing and submitting the most relevant paper to the user. Here, also, we calculate the scores for each cause-effect pairs we found, which reduce the compromise of TF-IDF fasle alarms. The search engine thus uses the inverted index to accelerate items. An inverted index is composed of a list of all the unique terms in each text, which helps us to easily identify the cause-effect similar that have the term in our demand, and only after that measure the significance score for the smaller recall collection.

Chapter 4

Experiments and Results

4.0.1 Experiments

Experiments setting

I applied several pre-trained models for the first part which is in table 4.1 from the annotated data. These are to find out which model have the highest accuracy for such particular task after comparing their performance.

bert-base-uncased	12-layer, 768-hidden, 12-heads, 110M parameters.
xlnet-base-cased	12-layer, 768-hidden, 12-heads, 110M parameters.
xlm-mlm-en-2048	12-layer, 2048-hidden, 16-heads
roberta-base	12-layer, 768-hidden, 12-heads, 125M parameters
albert-base-v1	12 repeating layers, 128 embedding, 768-hidden, 12-heads, 11M parameters
albert-base-v2	12 repeating layers, 128 embedding, 768-hidden, 12-heads, 11M parameters

Table 4.1: The detail setting of the pre-trained model

Evaluation method

We create a baseline model which is from the why-question, extracting directly the answer from the raw text data without using cause-effect extrac-

tion model. The experimental results of this baseline model will be compared with the most similarity answers which is finale chosen by BM25 method.

To detail information for the result table:

- **True Positive** which is the positive value that is accurately estimated, indicating that the goal value and predicted value are the same true.
- **True Negative** These are the properly predicted negative values which implies that the real
- **False negatives** when the goal category is yes but the experiment predicted no.
- **Precision** is the proportion of positive observations correctly predicted to the overall positive observations predicted.

$$Precision = \frac{TruePositives}{TruePositive + TrueNegative} \tag{4.1}$$

- **Recall** is based on optimistic observations the accurately estimated to all the observations in real class yes.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.2}$$

- F1 score is the score balance both of Precision and Recall. Hence, False Positives and False negative are used in the calculation. F1 score is typically more helpful than Accuracy score in the case of not equal class distribution.

$$F1score = 2.\frac{Recall.Precision}{Recall + Precision} \tag{4.3}$$

4.1 Experimental Results

4.1.1 Phase one

The results for the experiments of pretrained model is presented in table 4.2

The sample cause effect pairs could be automatically extracted from the articles are display in table 4.3 The best result is only 46.2% F1 score from Albert-based-v1. The parameter changing is fluctuate slowly between the other models. There are hardly any improvement on the performance of the models if I switch on several different parameters.

Pretrained model	**Precision**	**Recall**	**F1 score**
BERT-base-uncased	20.69%	**39.13**%	**27.07**%
ALBERT-based-v1	20.78%	34.78%	26.02%
xlnet-based-case	**20.97%**	28.26%	24.07%

Table 4.2: Cause-effect relation extraction model accuracy

phrase	**test result**	**text prediction**
cause	the automatic route	the automatic route
effect	Construction companies need only one set of official approvals and do not need to gain clearance from Foreign Investment Promotion	construction companies need only one set of official approvals
cause	Improving India infrastructure	Improving India infrastructure
effect	boost foreign investment in other sector too	boost foreign investment in other sectors too

Table 4.3: Sample extracted cause-effect pair

The cause and effect pairs extraction working properly on this data. There are a lot of cause-effect pairs extracted but most of them do not include the hole phrase of each pairs, leading to low score as exact match evaluation.

4.1.2 Phase two

I conclude the experiments for my method at the table 4.4

Experiment methods	**Precision**	**Recall**	**F1 score**
BM25 only	31.76%	**92.79%**	45.43%
albert-base-v1	59.54%	84.49%	**67.06**%
bert-base-uncased	60.24%	79.31%	65.73%
xlnet	**61.64%**	79.26%	66.58%

Table 4.4: why-question answering experimental results

In this phase, the F1 scores between different methods are almost similar between each other in the same page. On the other hand, albert-base-v1 with 37.1% on F1 score. The base line score of sentence BERT is the lowest among the results. The sample result phrases comparing of sentence bert (baseline) and Albert-base-v1 output are showed in table 4.5

4.2 Result Analysis

4.2.1 The first Phase

For the first phase, xlnet-based-case, ALBERT-based-v2 and Roberta have the same low experiment result beneath expectation. Under other else models, ALBERT-based-v1 and BERT-based-uncased. The different appeared because of two major factors: the model technical and the annotation data. The data that used for the research are hand annotated in a small corpus, the large data base must provide more suitable experiment environment for the training. With the model, i can define several characteristics make these highly gap for this research:

- The grammar detail make the sentence embedding impact with struggling in phrase details. The predicted cause-effect pairs are normally annotated without the hold sentences which make low accuracy when applied to extra accurate pre-trained models.

Question	goal answer	albert-base-v1 (BM25) answer	sentence BERT base-line answer
Why had India's rupee hit a five year high?	after Standard and Poor's rasied the country's foreign currency rating.	after Standard and Poor's rasied the country's foreign currency rating.	With Indian assets now seen as less of a gamble.
Why is today's economic crisis in the West Bank?	Closures are a key factor	Closures are a key factor	US Secretary of State was visiting the West Bank to revive its reform programme and maintain financial discipline after Closures are a key factor.
Why did China lent Russia $6bn?	to help the Russian government renationalise the key Yuganskneftegas unit of oil group Yukos	to help the Russian government renationalise the key Yuganskneftegas unit of oil group Yukos	*undetected*

Table 4.5: answer extracted sample

- The corpus is quite limited as well as the phrase have large scale of words make the predicted sample is commonly missing several words but in most of the cases, they haven't effect the context meaning that much.

- The outputs still reduce the accuracy score and challenging the semantic meaning of cause-effect pairs extracted.

Therefor the grammar detail make the sentence embedding impact with struggling in phrase details. The predicted cause-effect pairs are normally annotated without the hold sentences which make low accuracy when applied to extra accurate pre-trained models. The characteristic of albert-based in the other hand, bring up a lot of compact information for the predicted token. On the small data-base could benefit its effectiveness which have all the context relevant or query stream of annotated words thanks to all articles are focus on each specific problem of themselves.

Talking about the positive score of albert-base-v1 and BERT-based model. the Albert went into separated paths compared to almost all another bert-based models which reduce the parameters in the model in order to faster training and lower computational capability demands. Hence maybe thanks to that specific distinctive, ALBERT-based-v1 and BERT-based-uncased bring up higher hope into the next phrase of research.

The sample results of the experiments also bring up the same problem with the annotation data. The corpus is too small as well as the phrase have large scale of words make the predicted sample is commonly missing several words but in most of the cases, they haven't effect the context meaning that much. But it still reduce the accuracy score and challenging the semantic meaning of cause-effect pairs extracted.

4.2.2 The second phase

The reason for choosing BM25 and sentence BERT as the last method of comparing cause-effect extraction model trained in the first phase was mentioned above. Thanks to the results from the first phase, it is obviously that the F1 score for the model trained based on Albert-based-v1 is used for the next phase. The best score in this section is belonged to Albert-based-v1(BM25) with 37.1% accuracy. The effect of the pre-trained model on such small scale of data could impact the arguably one of the most efficiency and widely used information retrieval function is highly promising. While commonly used, few studies have investigated its efficacy on a single-field and multiple-field combinations document definition [32]

- The effect of the pre-trained model on such small scale of data could impact the arguably one of the most efficiency and widely used information retrieval function is highly promising.
- We could not overcome just yet is the difficulty in optimizing the function parameters for a given information retrieval measure.
- BM25 show slightly different impact to the improvement which help it be suitable to be used with common similarity measure which is cosine-similarity.

Chapter 5

Conclusion

Our main contributions in this work:

- Proposing a framework for why-question answering task which based on detected cause-effect relations to archive the most accurate answer.
- Creating datasets for cause-effect relations and why-question answering from collecting and annotating,
- Integrating dominant language models into why-question answering task but in small scale data.

Bibliography

[1] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, "Baseball: an automatic question-answerer," pp. 219–224, 1961.

[2] W. A. Woods and W. WA, "Lunar rocks in natural english: Explorations in natural language question answering." 1977.

[3] R. C. Staudemeyer, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks." 2019.

[4] R. Yamashita, "Convolutional neural networks: an overview and application in radiology," 2018.

[5] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," 2004.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[8] J.-H. Oh, K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, Y. Wang *et al.*, "Why question answering using sentiment analysis and word classes," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 368–378.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[16] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] J. Gao, Q. Wu, C. Burges, K. Svore, Y. Su, N. Khan, S. Shah, and H. Zhou, "Model adaptation via model interpolation and boosting for Web search ranking," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Aug. 2009, pp. 505–513. [Online]. Available: https://www.aclweb.org/anthology/D09-1053

[19] D. Metzler, "Generalized inverse document frequency," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 399–408. [Online]. Available: https://doi.org/10.1145/1458082.1458137

[20] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," *Proc. ICML*, 2006.

[21] H. W. Qiang Ning, Zhili Feng and D. Roth, "Joint reasoning for temporal and causal relations," in *In Proceedings of the 56th Annual Meeting of the Association for Computational (ACL)*, 2018.

[22] L. L. Jesse Dunietz and J. Carbonell, "Automatically tagging constructions of causation and their slot-fillers," in *Transactions of the Association for Computational Linguistics (TACL)*, 2017.

[23] C. K. R. I. Jong-Hoon Oh, Kentaro Torisawa and J. Kloetzer, "Multi-column convolutional neural networks with causality-attention for why-question answering," in *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

[24] S. S. T. Z. W. X. W. Enrique Noriega-Atala, Paul Douglas Hein and C. T. Morrison, "Inter-sentence relation extraction for associating biological context with events in biomedical texts," in *IEEE International Conference on Data Mining Workshops(ICDM)*, 2018.

[25] M. pasca, "Finding needles in an encyclopedic haystack: Detecting classes among wikipedia articles," in *Proceedings of the 2018 World-WideWeb Conference (WWW)*, 2018.

[26] T. P. Tiziano Flati, Daniele Vannella and R. Navigli, "Two is bigger (and better) than one: the wikipedia bitaxonomy project," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.

[27] J. L. J. T. ZhigangWang, Zhixing Li and J. Z. Pan, "Transfer learning based cross-lingual knowledge extraction for wikipedia," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.

[28] I. G. David Tsurel, Dan Pelleg and D. Shahaf, "Fun facts: Automatic trivia fact extraction from wikipedia," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

[29] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[30] J. C. P. C. D OCallaghan, D Greene, "An analysis of the coherence of descriptors in topic modeling," *ESWA*, 2015.

[31] D. G. S J Delany, M Buckley, "Sms spam filtering: Methods and data," *Expert Systems with Applications, pp. 9899-9908*, 2012.

[32] K. Svore and C. Burges, "A machine learning approach for improved bm25 retrieval," in *CIKM*, 2009.