Chatbot

Based on Prepared Answer Set

J . A . R . V . I . S .

Xu Wenhao(20660739)

Li Yihang(20662036)

Zhang Yanran(20660090)

Lai Kunyao(20661991)

2020.05

1. Introduction

With the prosperity of online markets and the popularity of internet life, it is necessary for companies and other kinds of parties to build up a convenient and effective communication way to serve their customers or members online. In couple of years, more and more researchers pay attention to Dialogue System because of its potential and attractive commercial value. In particular, chatbots have risen to the spotlight due to the development of big data and Deep-Learning Technology. More and more innovative chatbots serve at different kinds of industries to solve several kinds of problems, like Duolingo, a study partner of the language learning APP, and Sunny, the digital assistant of the tourism association. According to different communication background and tasks, company can design their own chatbots for question answering or tasks oriented.

Nowadays, more and more banks try to build up or buy their own chatbots on their websites and mobile apps in order to reduce labor costs and offer better services, like the KAI of Master card. In this paper, we built up a retrievable chatbot based on given answer set for a bank website. By exploring different kinds of ways to write our own chatbot, we use the Bert Model and SVM classifier as the key point of our chatbot.

2. Related works

In August 1956, artificial intelligence (AI) was initially proposed by J. McCarthy, M.L. Minsky, N. Rochester and C. E. Shannon at Dartmouth Conference. Since scientists have always been struggling with the aim that machines will assist humans with more intellective work, one of the most significant developments for human-machine interaction is a variety of understanding and caring chatbots, such as Siri, Cortana, Google Now, etc. The core technique solving for natural and effective communication between humans and machines is natural language processing (NLP).

NLP consists of two procedures: natural language understanding (NLU) and natural language generation (NLG). To construct a chatbot on a certain topic, the standard answer usually can be prepared and saved as an answer set. In other words, the architecture of the chatbots brings higher demand on NLU and how to specify and provide the most suitable answer from the answer set.

2.1 NLU

Natural language understanding (NLU) is a challenging project, involving a series of subjects including not only computer science and math but also linguistic, logic and psychology. One of the most difficult topics in NLU is to construct or obtain a language database, which is a complicated work costing a lot of time and human sources. Therefore, various researches of NLU are proposed on public sharing language databases.

Pre-trained language representations are a vital component with a long history in NLU

models. (1)In 2001, a neural language model was initially proposed by Bengio etc., in order to solve the problem of "the curse of dimensionality" in statistical language modeling. And it promotes the appearance of word embedding. (2)Especially in 2013, Word2vec was proposed by Mikolov etc., which provides a more effective method for word embedding. Skip-gram model was applied to accurately learn the vector representations of the phrases. (3)After that, Sutskever etc. proposed the sequence to sequence learning (seq2seq) with DNNs, which is a framework to mapping a sequence to another sequence by Long Short-Term Memory (LSTM). Recently, several supervised learning are applied for training neural network, including the researches by (4)Alexis etc. (5)McCann etc. and (6)Subramanian etc. Nowadays, (7)ELMo proposed by Peters etc. and (8)OpenAI GPT proposed by Radford etc. are the mostly prevailing methods of pre-trained language representation. (9) BERT proposed by Devlin etc. has refreshed the performance score on several tasks, and is a welcomed pre-trained language representation method.

2.2 Classifier

There are mainly 4 types of text classification method, including rule-based, probability-based, geometry-based and statistical-based classification models.

Firstly, rule-based classification models are relatively simple and can be implemented. Its classification in specific fields can often achieve better results. Compared with other classification models, the advantages of rule-based classification models are low time complexity and fast operation speed. In rule-based classification models, many rules are used to express categories. Category rules can be defined by domain experts or obtained through computer learning.

Decision tree is a common classification model based on training and learning method to obtain classification rules. It establishes a mapping between object attributes and object values. Classify and discriminate unlabeled text by constructing a decision tree. Commonly used decision tree methods include CART algorithm, ID3, C4.5, CHAID and so on. Hierarchical forms generally exist in the field of Web text applications, and this hierarchical form can be described by a decision tree. Divide the data set into two or more partitions. The leaves of the decision tree are the data collection of the corresponding category.

Secondly, Naive Bayes classifier is the most widely used probability classification model in the probability-based model. The basic idea of naive Bayes classification is to use the joint probability of phrase and category to estimate the category probability of a given document. d, the set of categories is C = {c1, c2, ..., cm}, the probability model classification is to find the conditional probability model P (ci | d) for $1 \leqslant i \leqslant n$, and the category with the highest conditional probability of the document d is used as the output category of the document.

Thirdly, the geometry-based model is to use the vector space model to represent text, and the text is expressed as a multi-dimensional vector, then it is a point in the multi-dimensional space. Construct a hyperplane through the principle of geometry to distinguish texts that do not belong to the same category. The most typical geometry-based classifier is the support vector machine (SVM). The simplest SVM application is binary classification, which is a common positive and negative example. The goal of SVM is to construct an N-dimensional space decision hyperplane that can distinguish positive examples from negative

examples.

Finally, statistical-based machine learning methods have become the mainstream research methods in the field of natural language research. In fact, both the naive Bayes classification model and the support vector machine classification model also use statistical methods. One of the most typical statistical classification models in text classification algorithms is k-Nearest Neighbor (kNN) model, which is one of the better text classification algorithms.

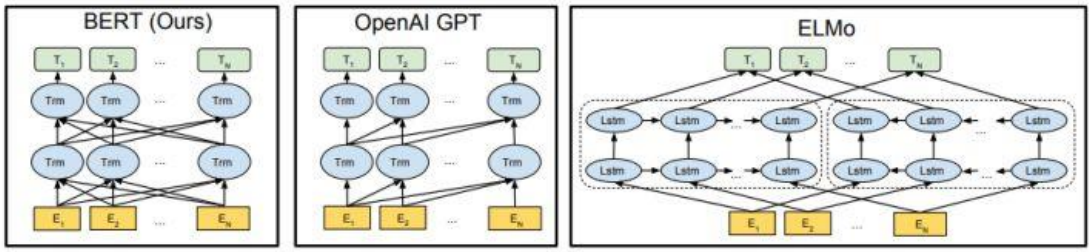
The main idea of the kNN classification model: by giving an unlabeled document d, the classification system finds the k nearest (similar or the same) labeled documents that are closest to it in the training set, and then labels the k adjacent documents according to the classification To determine the category of document d.

3. Model introduction

3.1 BERT

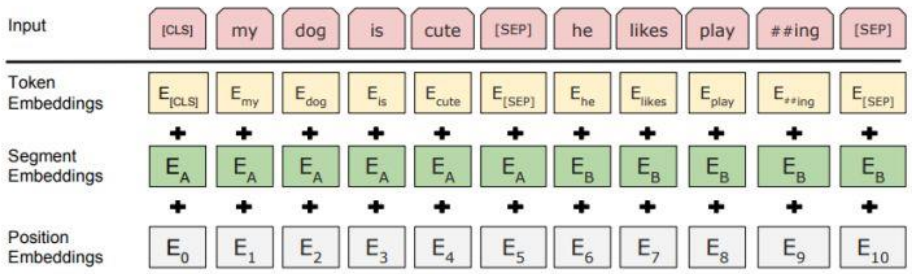
3.1.1 Network Architecture

The full name of BERT is Bidirectional Encoder Representation from Transformers, that is, the encoder of the bidirectional Transformer, because the decoder cannot obtain the information to be predicted. The main innovation of the model is the pre-train method, which uses Masked Language Model and Next Sentence Prediction to capture word and sentence-level representation respectively.



3.1.2 Embedding

The Embedding here is made by summing three Embeddings:



Token Embeddings are word vectors, the first word is the CLS logo, which can be used for subsequent classification tasks.

Segment Embeddings are used to distinguish between two kinds of sentences, because pre-training not only does Language Model but also has to perform classification tasks with two sentences as input.

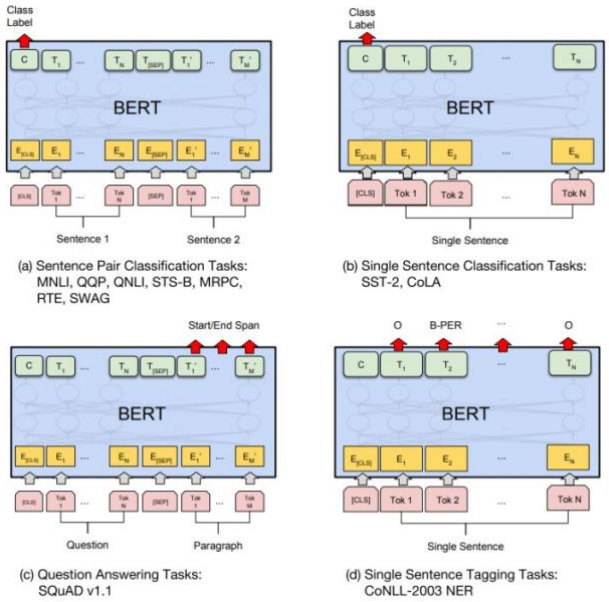
Position Embeddings are different from the Transformer in the previous article, it is not a trigonometric function but a learned one.

3.1.3 Fine-tuning

Classification: For sequence-level classification tasks, BERT directly takes the final hidden state of the first $C \in \Re^H$ [CLS] token, adds a layer of weight $W \in \Re^{K \times H}$, and softmax predicts the label probability:

$$P = softmax(CW^T)$$

Other prediction tasks require some adjustments, as shown in the figure:



3.2 SVM Classifier

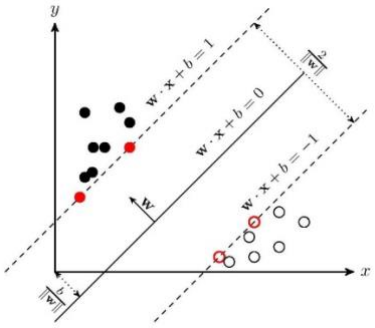
3.2.1 Introduction

Support vector machines (SVM) is a binary classification model. Its basic model is a linear classifier with the largest interval defined in the feature space. The maximum interval makes it different from the perceptron; SVM also includes Kernel technique, which makes it a substantially non-linear classifier. The learning strategy of SVM is to maximize the interval, which can be formalized as a problem of solving convex quadratic programming, and is also equivalent to the problem of minimizing the regularized hinge loss function. The learning algorithm of SVM is the optimal algorithm for solving convex quadratic programming.

3.2.2 Algorithm Principle

The basic idea of SVM learning is to solve the separation hyperplane that can correctly

divide the training data set and has the largest geometric interval. As shown in the following figure, $\omega x + b = 0$ is a separation hyperplane. For a linearly separable data set, there are an infinite number of such hyperplanes, that is, perceptron, but the separation hyperplane with the largest geometric interval is unique.



4. Chatbot construction

4.1 Preparation and modeling

We should set the BERT-service on using the CMD, and the BERT-service uses TF1 and python3.

4.1.1 Say hello and choose the mode

```
PRESS Q to QUIT
TYPE "DEBUG" to Display Debugging statements.
TYPE "STOP" to Stop Debugging statements.
TYPE "TOP5" to Display 5 most relevent results
TYPE "CONF" to Display the most confident result


Bot: Hi, Welcome to our bank!
```

In the DEBUG mode we can debug the chatbot, and we can stop the debug mode by typing 'STOP'. And if we choose "TOP5" it will give us answers to the 5 most relevant questions. If we choose "CONF" it only gives us the most confident answer.

4.1.2 Modeling and answer questions

Step 1. Embed the questions and classify the question

We first embed the questions into vectors using the BERT model, then we can train a classifier which can tell us which type of the questions it is. Here we use a SVM classifier, the accurate score is 0.89.

```
SVC: 0.8934240362811792
```

Step 2. Get a set of questions that might be the same type of question.
If we choose the DEBUG mode, we can see which type the question is predicted to be.

Step3. Find the most 'similar' question in the set of the questions we got in step 1.
We use the cosine between the question vector and the question vectors from the set of the questions to measure the similarity. Then we got the question that is most similar with the question given by the customer. We can give him the answer to the questions.

4.2 Example

Let's see an example in the DUBUG MODE.

```
You: >? How can I get an IVR Password from the bank
...:
...:
...:
Question classified under category: ['security']
57 Questions belong to this class
Assuming you asked: How can I obtain an IVR Password
Bot: By Sending SMS request: Send an SMS 'PWD<space>1234' to 9717465555 or to 5676712 from your registered (with Bank) mobile number.(Note: 1234 are the
```

If I ask "How can I get an IVR Password from the bank?"
The bot will first consider the question belongs to the 'security' type of question.
Then it finds the most similar question: How can I obtain an IVR Password?
Finally it gives the answer.

Then if you can't get your satisfy answer, the bot will give you the five most relevant to your question, and you can choose from them.

```
Was this answer helpful? Yes/No: >? NO
Bot: Do you want me to suggest you questions ? Yes/No: >? yes
1 Question: How can I obtain an IVR Password
--------------------------------------------------
2 Question: How do I register my Mobile number for IVR Password
--------------------------------------------------
3 Question: How should I get the IVR Password  if I hold an add-on card
--------------------------------------------------
4 Question: How will the online store know that I have Verified by Visa/ MasterCard SecureCode
--------------------------------------------------
5 Question: In how much time will the IVR Password be delivered to my mobile phone/email ID
--------------------------------------------------
Please enter the question number you find most relevant:
```

Then you can choose the question you find most relevant. And you can judge the answer.

```
Was this answer helpful? Yes/No:
You: >? yes
Bot: Yes!
```

Finally, you can quit the system by typing "Q", and the bot will say goodbye to you.

```
You: >? Q
Bot: It was good to be of help.
```

5. Existing shortcomings & Future improvements

1. At present, it is just a question and answer that lacks chat-like interaction with customers. For example, when customers say hi, they will find the question closest to hi, instead of saying hello to customers like real customer service. In the future, we can specifically deal with this hello problem, we may use the deep learning model.
2. All answers are still matched based on the existing question and answer library. Choose the answer to the question that is closest to the known question. In the future, we will need to continue to enrich the question and answer library to match more questions and we can try more deep learning methods for question answering, rather than just embedding and matching questions.

References

1. *A Neural Probabilistic Language Model.* **Yoshua Bengio, Rejean Ducharme, Pascal Vincent.** Vancouver, British Columbia, Canada : s.n. NIPS 2001.

2. *Distributed Representations of Words and Phrases and their Compositionality.* **Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean.** Advances in Neural Information Processing Systems 26 : s.n. NIPS 2013.

3. *Sequence to Sequence Learning with Neural Networks.* **Ilya Sutskever, Oriol Vinyals, Quoc V.Le.** NIPS 2014.

4. **Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, Antoine Bordes.** Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv preprint arXiv:1705.02364, 2017.*

5. *Learned in Translation: Contextualized Word Vectors.* **Bryan McCann, James Bradbury, Caiming Xiong, Richard Socher.** NIPS 2017.

6. **Sandeep Subramanian, Adam Trischler, Yoshua Bengio, Christopher J Pal.** Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. *arXiv preprint arXiv:1804.00079, 2018.*

7. **Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power.** Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108, 2017.*

8. **Radford A, Narasimhan K, Salimans T, et al.** Improving language understanding with unsupervised learning. *Technical report, OpenAI.* 2018.

9. **Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805, 2018.*

Contributions of each member:

We all studied together to explore useful tools related to retrievable chatbot, including articles related to NLP and popular classifiers. Besides, we also study typical cases of chatbot and published coding on websites. Specifically,

Xu Wenhao(20660739) pay more attention on different types of classifiers and gathered several cases related to the chatbot used in bank industry.

Li Yihang(20662036) pay more attention on NLP, and read and shared many related articles and main ideas with us.

Zhang Yanran(20660090) pay more attention on SVM classifier and Bert model, in order to use them in our chatbot.

Lai Kunyao(20661991) make most efforts to refine our chatbot and prepared the data set which is used as the answer set of our chatbot.

Link of coding and video:

https://pan.baidu.com/s/1nVk3CDF49BhB5Ic9h574ag 提取码:5311