Critical Essay

Kaicheng wu

Abstract — **I study the problem of machine learning, whose goal is to train a Its goal is to train a chatbot, call a corpus with the theme of constellations supplemented by chat (e.g., dataset called AIML with some new corpus written by myself), and conduct the most simulated human-machine dialogue. While the chatting part, the anthropomorphic of a robot, i.e. how plausible a conversation is, is a hot topic. Domain-specific corpus dialogue is less explored in the literature, large amount of datasets are also necessary to train the bot into smarter. Without enriching datasets, directly repeated training applied to chatbot often leads to too one-sided dialogue direction. In this paper, I used the existing AIML corpus as a basis to write new xml files as a corpus to enrich the machine chat under the constellation theme.**

Keywords — ***Chatbot, Chatbot Systems, AIML, Conversation Systems, Natural Language Processing, Machine Learning, Chatbot Knowledge, Chatbot Domain, Signal Processing, Python*** **.**

1. Introduction:

A comprehensive corpus is necessary for an intelligent chatbot to perform machine learning. Apart from the simple dialogue function, the function of calling datasets is also widely used in artificial intelligence, machine learning, text processing, optimization experience and personalized analysis.[1] For example, a classification project of scraping a corpus of posts from a social network relating to certain groups of online recipe sharing communities in different countries, which would be cumbersome but necessary when tracing datasets[2]. And also related to cultural heritage, e-learning, e-government, web base model, dialog model, semantic analysis framework, interaction framework, humorist expert, network management, adaptive modular architecture as well.[3] There are many problems in the corpus, including the generation of a large amount of data, there will be more and more data redundancy, and it is not convenient to update because the version is fixed.

- The main purpose of this project is to train existed corpus - AIML, with specific topic conversations, to serve specific area.

2. Literature review:

The usual reason of any technology development is to lead lives into a more convenient life. This is also what machine learning are researched and why the conversational systems, which also called chatbot get popularity in decades. “Chatbot is a program that talks in voice or text. It is equipped with artificial intelligence (AI) technology and is developed in a way to combine with messenger.”[4]

At the same time, with the advent of smartphones, the core of the Internet market has shifted from existing web browsers to mobile platforms and various application ecosystems. However, chatbot-based messaging apps are poised to become a strong alternative as major mobile platform companies, such as KakaoTalk of South Korea, Wechat of china, Line in Japan, and Whatsapp in America establish messaging apps as medium to long-term growth engines The current role of the web browser in the future.

As Daniel J. and James M. mentioned in 1999, Natural Language Processing(NLP) is one of the most applications in artificial intelligence. Furthermore, conversation mode is also one of the most important applications in NLP.[5] Apart from widely used, the reason why the chatbot system is popular is that it is approachable, improves customer experience, can manage a large number of customers, and is cost-effective. It has also been found to help reduce overall operating costs. Because from the customer's perspective, robots provide a novel experience, it has also been noticed that consumers are more inclined to interact with robots rather than human-to-human interactions.

Many users claimed in interviews that "chatting (with people offline) is a burden", so many merchants have to hire more online chat staff (customer service) in addition to store staff and provide 24-hour service. So an automatic question-answering robot with simple question-and-answer function, common phrases and FAQ is considered to be used, but at the same time a new problem arises: although it can reduce the problem to a certain extent, it is not very effective.

Chatbots try to answer users' questions by being endowed with three types of knowledge, including structured database, knowledge bases and unstructured dataset.

[6]

Many variable methods of machine learning or deep learning can be explored to make the chatbot system accurate. And basic processing of NLP field could be analysed and discover deeper, so it could be end by there is still a lot of scope for research in this area of machine learning for NLP.

- Related work:

TURING ROBOT: An artificial intelligence company with semantic and dialogue technology as its core has hundreds of core patents in the field of artificial intelligence robots, [Online] "http://www.turingapi.com/"

Mitsuku/Kuki: An embodied artificial intelligence robot designed to befriend humans in the Metaverse. Kuki, formerly known as Mitsuku, is a chatbot created by Steve Worswick using Pandorabots AIML technology, [Online] "https://www.kuki.ai/"

3. Methodology and findings

I conduct extensive experiments on several datasets of chatterbot, AIML and directly define an empty chatbot, and create conversations on the spot to use as the bot's corpus. In the end, because the python version is 3.9, chatterbot can only be installed in an environment lower than python3.8. The function of creating data sets on the spot is too simple, otherwise it is necessary to create a large amount of dialogue composition corpus. Also chatbots based on AIML (Artificial Intelligence Markup Language) used to be popular because they were lightweight and easy to configure. So I used the AIML dataset as my chatbot operating environment.

AIML is a derivative of Extensible Markup Language (XML). It has a data object class called an AIML object, which describes the behavior of a computer program. It contains units called topics and categories. A category is the basic unit of knowledge in AIML. Each category contains patterns containing the input and templates containing the chatbot's answers. Additionally, some optional context is included, called "that" and "topic". ‘that’ contains the last sentence of the chatbot and ‘topic’ contains a set of categories. Its prototype was a highly extended Eliza robot named "A.L.I.C.E." (Artificial Linguistic Internet Computer Entity). Since ALICE's AIML setup is released under the GNU GPL license, there have been many clones based on the program and the AIML library. As a result, AIML currently has versions in Java, Ruby, Python, C, C#, Pascal and other languages.

My methods are add topic-specific corpora to avoid random or repetitive data due to extensive training. To this end, I used the off-the-shelf chatbot corpus as the basis of the project, and wrote some new questions and answers as the corpus of a specific topic (constellation) that I needed, to some extent enriching some of the dialog areas that the data set lacked, given paired input and output , With carefully-designed generators and discriminators to activate the chatbot function, and a new corpus, my methods can make a chatbot that discusses constellation. Furthermore, I can scrape data from the Internet and write the corpus about constellations I need.

#Code: Due to my computer is installed in the python3 code environment, I need to use pip *install python-AIML* to install the AIML module. Then use the *import* function to call the *AIML, os* and *sys* libraries.

#Get the installation directory of the ALICE library by:

```
def get_module_dir(name):
    path = getattr(sys.modules[name], '__file__', None)
     if not path:
        raise AttributeError('module %s has not attribute __file__' % name)
     return os.path.dirname(os.path.abspath(path))
```

#Switch to the directory where the corpus is located by:

```
alice_path = get_module_dir('AIML') + '/botdata/alice'
os.chdir(alice_path)
print(alice_path)
```

#Load the corpus file by:

```
alice = AIML.Kernel()
alice.learn("startup.xml")
alice.respond('LOAD ALICE')
```

And finally by:

```
while True:
    print(alice.respond(input("My name is K. Feel free to chat with me >> ")))
```

It can run the round robin function.

The most important step is to write a corpus composed of my established dialogues, and use the *<catergory></catergory>* element to include the questions in *<pattern></pattern>*, which are the user's question sentences. It is mainly used to match user input and supports fuzzy matching by using "_" and "*". The answer of the robot in *<template></template>* is that when the user input matches the Pattern under the same category, the template under this category element will be output. By the way, in the "template" element, the <random></random> element can be used to make the chatbot randomly answer the answers in <li></li>.

e.g.

```
<category>
    <pattern> hi* </pattern>
    <template>
        <random>
            <li> hi </li>
            <li> hello </li>
            ...
        </random>
    </template>
</category>
```

Finally to create a standard startup file called std-startup.xml as the main entry point for loading AIML files. Here we'll write a basic file that matches a pattern and performs an action. We want to match the pattern *load AIML b*, and have it load into our AIML brain. Through this file, we can call the corpus file with xml suffix written in the datasets.

4. Conclusion and prospect

The basic function as a chatbot has been realized, that is, it can talk and reply to most of the daily conversations and established conversations on horoscope topics. The daily conversations in it are not described much, but the constellation conversations set by me are roughly as follows:”what are your constellations”,”talk about your constellations”,”talk about pisces” and the other 11 constellations, “do you like pisces” and also the rest with some others conversations.

But as a robot that calls the corpus to answer, the answer of the robot is only the answer set by the program, so the answer of the machine cannot always cover all the questions, and it seems very blunt. It is far from reaching the level of passing the Turing test, so this robot needs to add a learning function in the future. It is a function that I have tried during the coding process but has problems and cannot be successfully run. Through a large number of conversations with a large number of users and using the following Scikit-learn-based learning library, which expands new dialogues by the computer itself.

The machine learning process will frequently obtain user and network data, and many ethical issues will inevitably arise. User privacy is always an inevitable topic of discussion in the information age, and chatbots certainly cannot escape the steps of obtaining user data. But I think that under the restriction of privacy use that crosses the border, that is, only the questions entered by the user and the permissions of some device storage are obtained, and the user's right to know is reserved, so that the user knows that the robot will save the conversation that occurred and save it to the database (for machine learning)). Under this limitation, the data obtained by the robot is open and transparent, and its importance would basically not have adverse effects and harms.

5. References

[1] Martin Breuss, “ChatterBot: Build a Chatbot With Python”, [Online] Available at: <https://realpython.com/build-a-chatbot-python-chatterbot/#step-2-begin-training-your-chatbot> [Access 2 November 2022]

[2] Cynthia V, H & Gilles J, “Automatic detection of cyberbullying in social media text”, [Online] Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203794> [Access 2 November 2022]

[3] M. S. Satu, M. H. Parvez and Shamim-Al-Mamun, "Review of integrated applications with AIML based chatbot," 2015 International Conference on Computer and Information Engineering (ICCIE), 2015, pp. 87-90

[4] Miri H, Kyoung J.L, “Chatbot as a New Business Communication Tool: The Case of Naver TalkTalk”, Bus. Commun. Res. Pract. 2018;1(1):41-45.

[5] Daniel J, James H.M, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition”, Prentice Hall Inc Publications, 2nd Edition, 1999.

[6] T. P. Nagarhalli, V. Vaze, N. K. Rana, "A Review of Current Trends in the Development of Chatbot Systems," 2020, pp. 706-710