

Assignment 4: Final Project – Creating Chatbot

Kathryn LiPetri

MSDS 453: DL Section 56, Fall 2022

December 4, 2022

Introduction and Problem Statement

The goal of this project is to develop a chatbot on a chosen topic. The business context of the chatbot created for this assignment is a request from a travel agency to automate answers to general questions that customers may have when visiting a specific place, in this case Italy. A question/answer chatbot is chosen as the best option for the given purpose. The data used for this chatbot is general information about Italy copied from the Wikipedia page (Italy 2022). The chatbot will be retrieval based, i.e. when asked a question (input from the user), it will read the text in the data file and output the text sequence that best matches the question asked, using cosine similarity. A couple of different types of models will be used to develop the chatbot including Tf-Idf and sentence transformers. The chatbots will be evaluated based on how accurate their responses are to the given questions.

Literature Review

Chatbot is a general name for a program that simulates conversations with a user by text message. It involves the user entering an input, and the chatbot replying with an output generated based on the user input and how the programming tells it to answer. There are two main types of chatbots, retrieval-based and generative-based, that can be programmed to serve numerous purposes (Zaboj 2022). Retrieval-based chatbots utilize a pre-defined dataset from which to retrieve the best response to the user input. These types of models are simpler to develop because they do not necessarily require large amounts of data to operate. However, the model is not intelligent, so it will not be able to learn or generate any new responses. Generative-based chatbots generate new responses based on training data input, essentially translating the user input to an output response. These types of chatbots are difficult to implement because they require large amounts of data to train and may not have accurate outputs (Sanjeevi 2018).

The concept of chatbots came about in 1950 when Alan Turing wrote a paper implying that a computer program could think and talk like a human. By 1966, a computer program called Eliza was developed at MIT by Joseph Weizenbaum and was considered to be the first chatbot. Eliza was able to communicate with users by answering user questions with pre-scripted responses from a database. From there, numerous chatbots were developed over the decades, and in 2010 the creation of Apple's Siri launched the first globally available virtual personal assistant. Shortly following were Google's Google Now, Microsoft's Cortana, and Amazon's Alexa virtual personal assistants (Zaboj 2022).

Currently, chatbots are developed for several different industries, such as e-commerce, travel, finance, healthcare, media, and education, to serve as customer service agents. Using chatbots for this purpose has several benefits including 24/7 support, automated resolutions to customer issues, lower operational costs, and a better customer experience. In addition, chatbots can be created to help employees in the workplace with training and onboarding, automating certain processes, and on-the-job question answering (Freshdesk Blogs 2022).

Pre-training language models has improved the accuracy of Natural Language Processing (NLP) tasks in recent years, including question/answer chatbots. One of the most successful models used for question/answer chatbots is BERT (Bidirectional Encoder Representation from Transformers) (Song et al 2020). Traditionally, RNN (Recurrent Neural Network) models were used to train language models because of the sequential structure of language. These types of models are slow to train because they process each token sequentially. It has been found that Transformers are better for this question/answer chatbot application because they process tokens in parallel and have a self-attention mechanism, which makes training larger models possible (Singh 2021).

However, not all models are created equal. An improvement on the BERT model has come about recently, and that is the MPNet model, which has been proposed in “MPNet: Masked and Permuted Pre-training for Language Understanding” (Song et al 2020). BERT uses Masked Language Modeling (MLM) for pre-training. While it is one of the most successful pre-training models for NLP tasks, it neglects dependency among predicted tokens. Another NLP model, XLNet, uses Permuted Language Modeling (PLM) to capture the dependency of the predicted tokens, but the position information of tokens in the full sentence is not known by the model. In the article, a new pre-training model is introduced to remedy the issues of both the BERT and XLNet model, the MPNet model. The MPNet model is able to take into consideration dependency among the tokens in addition to taking into consideration the position information of all tokens as input so that the model can see this information. The MPNet model has been shown to outperform models such as BERT, XLNet, and RoBERTa on the same developmental data sets (Song et al 2020).

It is worth noting that there are several modified versions of the BERT architecture that each have different pre-training objectives, such as RoBERTa, SpanBERT, DistilBERT, and ALBERT (Singh 2021). For the chatbot in this report, DistilBERT, which is a smaller and faster BERT architecture, was one of the better performing pre-training architectures. The experiments in this report utilize different pre-training models, including BERT and MPNet architectures, to reach the desired outcome of the question/answer chatbot.

Data

The dataset used to extract answers for the chatbot was copied from Wikipedia. The entire “Italy” page from Wikipedia was copied into a .txt file, and this was the corpus used for the chatbot answers to user questions (Italy 2022). The corpus consists of 25,207 words and has

information regarding numerous topics about the country of Italy including history, geography, politics, economy, demographics, and culture. Minimal pre-processing of the data was performed to convert everything to lowercase, remove stop words, and lemmatize.

Research Design and Modeling Methods

The first part of the assignment was to create an ontology with the chosen corpus in order to get a clearer understanding of the corpus material. The ontology was developed using Protégé software. The main class of the ontology is called “Italy Thing”, and there are several subclasses related to Italy including “Architecture”, “Art”, “City”, “Fashion”, “Festival”, “Food”, and “Weather”. The ontology shows specific entities for each subclass related to things that a tourist might want more information on when planning a trip or traveling in Italy.

The main part of the assignment was to create a chatbot that would help a travel agency answer general questions that customers may have about traveling to Italy. An extractive chatbot was created to read in a user’s question and output a response based on the corpus data. One part of the chatbot was rule-based, with greetings, greeting responses, and exit responses manually programmed into the chatbot. The question response portion of the chatbot was programmed to use either tf-idf vectorization of the data to output a response or different pre-trained sentence transformer models to output a response based on the corpus data. In all cases, cosine similarity was used to output a response that best matched to the user’s input question.

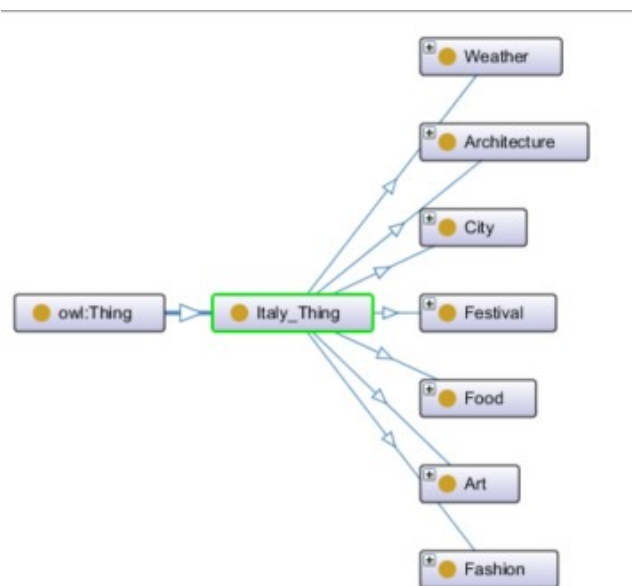
For each chatbot instance, the same 5 questions were asked of the chatbot: 1) What kind of food is there in Italy?; 2) What are popular cities to visit?; 3) What is the weather like in Italy?; 4) What is there for tourists to do in Italy?; 5) What kind of architecture does Italy have?. Chatbots were evaluated based on how accurate and relevant their responses were based on the user’s input question and the intended information the user was asking for.

Results, Analysis, and Interpretation

Ontology

The ontology was developed using Protégé software. The main class in the ontology is called “Italy Thing”, and there are several subclasses related to Italy including “Architecture”, “Art”, “City”, “Fashion”, “Festival”, “Food”, and “Weather”. The subclass “Weather” also has another subclass, “Season”. The ontology shows specific entities for each subclass related to things that a tourist might want more information on when planning a trip or traveling in Italy. Figure 1 shows the class and sub-class structure of the ontology, before revealing any entities.

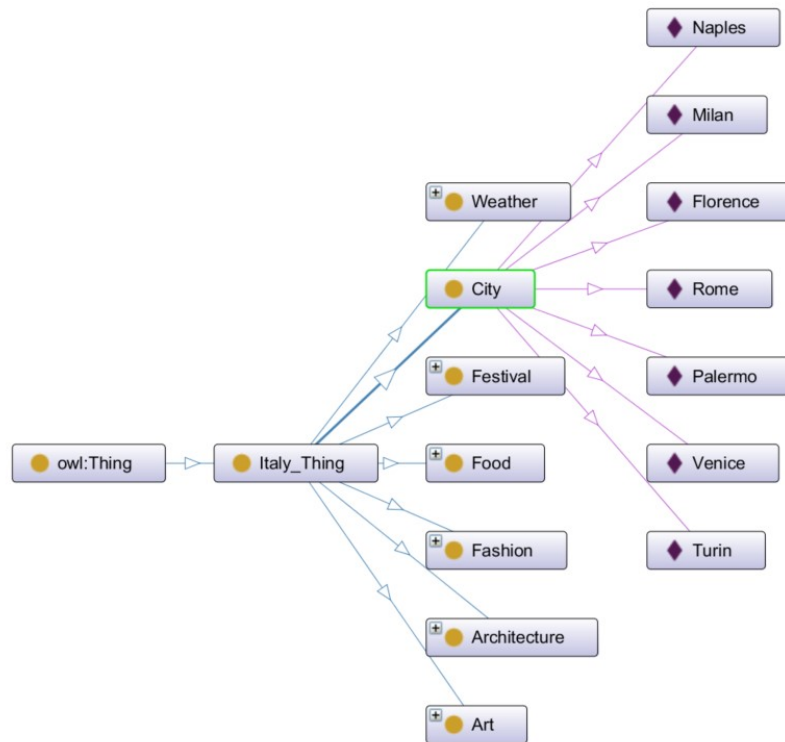
Figure 1. Ontology Class Structure



Subclasses were then linked to their specific entities listed in the corpus used for this project. The entities represent things that are anticipated to be output by the chatbot when a user asks a question about one of the subclasses listed. In addition, creating the ontology also helps to determine where additional data could be added to the knowledge base to improve the chatbot’s ability to answer questions. For instance, if the “City” subclass is expanded to show its entities,

these are some of the cities that are anticipated that a user would ask about. This expanded version of the ontology is shown in Figure 2.

Figure 2. Expanded Ontology by “City”



From this expanded ontology view, it is clear that the entity cities shown are all mentioned in our current corpus or knowledge base. The fully expanded ontology can be found in the appendices.

Chatbot

An extractive chatbot was created to read in a user’s question and output a response based on the corpus data. One part of the chatbot was rule-based, with greetings, greeting responses, and exit responses manually programmed into the chatbot. The question response portion of the chatbot was programmed to use either tf-idf vectorization of the data to output a response or different pre-trained sentence transformer models to output a response based on the corpus data. In all cases, cosine similarity was used to output a response that best matched to the

user's input question. Table 1 summarizes the main characteristics of the different chatbots created with either tf-idf or pre-trained sentence transformers. The example code for the tf-idf chatbot originated from a medium.com article (Dass 2018). Information regarding the pre-trained models can be found at sbert.net (Reimers 2022).

Table 1. Summary of Chatbots

Chatbot	Technique/Model Used	Max Sequence Length	Dimensions	Training Data	Size	Average Performance	Encoding Speed (sentences/s)
1	Tf-Idf Vectorizer	N/A	N/A	N/A	N/A	N/A	N/A
2	Multi-qa-MiniLM-L6-cos-v1	512	384	215M (question, answer) pairs from diverse sources	80 MB	58.08	14200
3	All-MiniLM-L6-v2	256	384	1B+ training pairs	80 MB	58.80	14200
4	All-mpnet-base-v2	384	768	1B+ training pairs	420 MB	63.30	2800
5	All-distilroberta-v1	512	768	1B+ training pairs	290 MB	59.84	4000
6	Multi-qa-mpnet-base-dot-v1	512	768	215M (question, answer) pairs from diverse sources	420 MB	62.18	2800
7	Multi-qa-distilbert-cos-v1	512	768	215M (question, answer) pairs from diverse sources	250 MB	59.14	4000

For each chatbot, the same 5 questions were asked of the chatbot: 1) What kind of food is there in Italy?; 2) What are popular cities to visit?; 3) What is the weather like in Italy?; 4) What is there for tourists to do in Italy?; 5) What kind of architecture does Italy have?. Chatbots were evaluated based on how accurate and relevant their responses were based on the user's input question and the intended information the user was asking for. Table 2 summarizes the answers from each model for the 5 questions.

Table 2. Chatbot Response Summary

Question	What kind of food is there in Italy?	What are popular cities to visit?	What is the weather like in Italy?	What is there for tourists to do in Italy?	What kind of architecture does Italy have?
Tf-Idf Vectorizer	the country is well known for its creative and innovative business,a large and competitive agricultural sector (with the world's largest wine production),and for its influential and high-quality automobile, machinery, food, design and fashion industry	the most popular sport in italy is football.	demographics main article: demographics of italy see also: italians, italian diaspora, genetic history of italy, list of cities in italy, and racism in italy map of italy's population density at the 2011 census	tourism main article: tourism in italy the amalfi coast is one of italy's major tourist destinations.	roman painting does have its own unique characteristics.

Question	What kind of food is there in Italy?	What are popular cities to visit?	What is the weather like in Italy?	What is there for tourists to do in Italy?	What kind of architecture does Italy have?
Multi-qa-MiniLM-L6-cos-v1	italian cuisine in itself takes heavy influences, including etruscan, ancient greek, ancient roman, byzantine, and jewish.	rome is the 3rd most visited city in europe and the 12th in the world, with 9.4 million arrivals in 2017 while milan is the 27th worldwide with 6.8 million tourists.	because of the length of the peninsula and the mostly mountainous hinterland, the climate of italy is highly diverse.	factors of tourist interest in italy are mainly culture, cuisine, history, fashion, architecture, art, religious sites and routes, naturalistic beauties, nightlife, underwater sites and spas.	italy is also prominent in the field of design, notably interior design, architectural design, industrial design and urban design.
All-MiniLM-L6-v2	the mediterranean diet forms the basis of italian cuisine, rich in pasta, fish, fruits and vegetables and characterised by its extreme simplicity and variety, with many dishes having only four to eight ingredients.	in addition, venice and florence are also among the world's top 100 destinations.	because of the length of the peninsula and the mostly mountainous hinterland, the climate of italy is highly diverse.	factors of tourist interest in italy are mainly culture, cuisine, history, fashion, architecture, art, religious sites and routes, naturalistic beauties, nightlife, underwater sites and spas.	italy is also prominent in the field of design, notably interior design, architectural design, industrial design and urban design.
All-mpnet-base-v2	italy is the world's largest wine producer, and one of the leading in olive oil, fruits (apples, olives, grapes, oranges, lemons, pears, apricots, hazelnuts, peaches, cherries, plums, strawberries and kiwifruits), and vegetables (especially artichokes and tomatoes).	rome is the 3rd most visited city in europe and the 12th in the world, with 9.4 million arrivals in 2017 while milan is the 27th worldwide with 6.8 million tourists.	average winter temperatures vary from around 0 °c (32 °f) in the alps to 12 °c (54 °f) in sicily, so average summer temperatures range from 20 °c (68 °f) to over 25 °c (77 °f).	factors of tourist interest in italy are mainly culture, cuisine, history, fashion, architecture, art, religious sites and routes, naturalistic beauties, nightlife, underwater sites and spas.	the city of venice, built on 117 islands the duomo and the leaning tower of pisa italy is known for its considerable architectural achievements, such as the construction of arches, domes and similar structures during ancient rome, the founding of the renaissance architectural movement in the late-14th to 16th centuries, and being the homeland of palladianism, a style of construction which inspired movements such as that of neoclassical architecture, and influenced the designs which noblemen built their country houses all over the world, notably in the uk, australia and the us during the late 17th to early 20th centuries.
All-distilroberta-v1	[589][590] italian cuisine is noted for its regional diversity,[591][592][593] abundance of difference in taste, and is known to be one of the most popular in the world,[594] wielding strong influence abroad.	in addition, venice and florence are also among the world's top 100 destinations.	these seas constitute a reservoir of heat and humidity for italy.	there are many festivals and festivities in italy.	the city of venice, built on 117 islands the duomo and the leaning tower of pisa italy is known for its considerable architectural achievements, such as the construction of arches, domes and similar structures

Question	What kind of food is there in Italy?	What are popular cities to visit?	What is the weather like in Italy?	What is there for tourists to do in Italy?	What kind of architecture does Italy have?
					during ancient rome, the founding of the renaissance architectural movement in the late-14th to 16th centuries, and being the homeland of palladianism, a style of construction which inspired movements such as that of neoclassical architecture, and influenced the designs which noblemen built their country houses all over the world, notably in the uk, australia and the us during the late 17th to early 20th centuries.
Multi-qa-mpnet-base-dot-v1	italy is the world's largest wine producer, and one of the leading in olive oil, fruits (apples, olives, grapes, oranges, lemons, pears, apricots, hazelnuts, peaches, cherries, plums, strawberries and kiwifruits), and vegetables (especially artichokes and tomatoes).	in addition, venice and florence are also among the world's top 100 destinations.	these seas constitute a reservoir of heat and humidity for italy.	there are many festivals and festivities in italy.	byzantine architecture was also widely diffused in italy.
Multi-qa-distilbert-cos-v1	cheese, cold cuts and wine are central to italian cuisine, with many regional declinations and protected designation of origin or protected geographical indication labels, and along with pizza and coffee (especially espresso) form part of italian gastronomic culture	in addition, venice and florence are also among the world's top 100 destinations.	climate main article: climate of italy köppen-geiger climate classification map of italy the climate of italy is influenced by the large body of water of the mediterranean sea that surrounds italy on every side except the north.	there are many festivals and festivities in italy.	the city of venice, built on 117 islands the duomo and the leaning tower of pisa italy is known for its considerable architectural achievements, such as the construction of arches, domes and similar structures during ancient rome, the founding of the renaissance architectural movement in the late-14th to 16th centuries, and being the homeland of palladianism, a style of construction which inspired movements such as that of neoclassical architecture, and influenced the designs which noblemen built their country houses all over the world, notably in the uk, australia and the us during the late 17th to early 20th centuries.

Red = wrong answer; Green = best answer; No Color = answer makes sense but is not the best

Based on the responses from the different chatbots, the tf-idf chatbot performed the worst with 4 of the 5 answers not related to the question at all. One of the answers for the tf-idf chatbot made sense, but was not the best possible answer when compared to the other chatbot outputs. The chatbot that performed the best answered 4 out of 5 questions with the best answer, and 1 out of 5 questions with an answer that made sense but was not the best. This was the chatbot using pre-trained sentence transformer model 'all-mpnet-base-v2'. All other sentence-transformer models responded to up to 2 of the 5 questions with the best answer.

Some of the models were specifically trained with question/answer pairs and others were trained more generally. The 'all-mpnet-base v2' model was trained for general use, but performed better than any of the models trained specifically as a question/answer model. This is likely because of the very large volume of training data compared to the question/answer models. The 'all-mpnet-base-v2' model was trained on 1B+ training pairs while all of the question/answer models were trained with 215M question/answer pairs.

There were other models trained on the 1B+ training pairs, but these did not perform as well as the mpnet general model. As discussed in the literature review, the MPNET model was created to improve upon issues with other models and is able to take into consideration dependency among the tokens in addition to taking into consideration the position information of all tokens as input so that the model can see this information. These improvements in the MPNet model allowed it to perform better than the other models, despite being trained on the same 1B+ training pairs dataset. The 'all-mpnet-base-v2' model was also the model with the highest average performance of sentence embedding and semantic search (Reimers 2022).

Conclusions

An extractive question/answer chatbot was created to automate answering general questions for travel agency customers traveling to Italy. Data from the Wikipedia page for Italy was used as the corpus for this model. Several different techniques and models were used for the chatbot response feature including tf-idf vectorization and 6 different pre-trained sentence transformer models. Chatbots were evaluated based on accuracy and relevance of their outputs in response to the same 5 customer questions. The chatbot created with sentence-transformer model ‘all-mpnet-base-v2’ performed the best, answering 4 of the 5 questions with the best responses. The tf-idf chatbot performed the worst with 4 of the 5 questions answered incorrectly.

The ‘all-mpnet-base-v2’ model was hypothesized to have performed the best because of the very large training data (1B+ training pairs) and the improvements that the MPNet architecture has over other models, such as BERT, in the areas of token dependency and token position information.

Directions for Future Work

Based on the responses of the chatbots, it is proposed to move forward with future work using the ‘all-mpnet-base-v2’ sentence-transformer model. Based on the ontology, the corpus used for the chatbot could be expanded upon and fine tuned based on anticipated questions that customers would have. More tourist/travel information could be added to the corpus, and information not related to customer questions could be removed from the corpus. Another option to obtain a better performing model would be to train a model with only tourist and travel related corpora to specialize the model to this area of expertise. From there, the knowledge base could be expanded to allow the chatbot to answer customer questions about other travel destinations, further improving customer service and decreasing costs for the travel agency.

References

1. “20 Fascinating Chatbot Applications from Six Key Industries.” Freshdesk Blogs, November 17, 2022. <https://freshdesk.com/customer-service/chatbot-applications-blog/#2>.
2. Dass, Riti. “Create Your Chatbot Using Python NLTK.” Medium. Medium, September 24, 2018. <https://medium.com/@ritidass29/create-your-chatbot-using-python-nltk-88809fa621d1>.
3. “Italy.” Wikipedia. Wikimedia Foundation, November 17, 2022. <https://en.wikipedia.org/wiki/Italy>.
4. Sanjeevi, Madhu. “Chapter 11: Chatbots to Question & Answer Systems.” Medium. Deep Math Machine learning.ai, April 19, 2018. <https://medium.com/deep-math-machine-learning-ai/chapter-11-chatbots-to-question-answer-systems-e06c648ac22a>.
5. Singh, Rohit Kumar. “How to Train a Question-Answering Machine Learning Model.” Paperspace Blog. Paperspace Blog, April 9, 2021. <https://blog.paperspace.com/how-to-train-question-answering-machine-learning-models/>.
6. Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. “MPNet: Masked and Permuted Pre-Training for Language Understanding.” *ArXiv*, November 2, 2020.
7. Reimers, Nils. “Pretrained Models¶.” Pretrained Models - Sentence-Transformers documentation, 2022. https://www.sbert.net/docs/pretrained_models.html.
8. Zabo, Daria. “The Complete Chatbot Guide 2022 - from Basics to Advanced.” ChatBot Blog, August 31, 2022. <https://www.chatbot.com/blog/chatbot-guide/>.

Appendices

Figure A1. Fully Expanded Ontology

