Localizing AI: The Path to Philadelphia's Specialized Chatbot

Will Schenk

19 December 2023

The AI revolution is here in the form of chatbots. However, many of these chatbots merely redirect their queries to models from OpenAI and Google. High-powered chatbots tailored and trained for specific purposes, rivaling OpenAI's ChatGPT or Google Bard's abilities, still need to be added to the market. A great potential still exists for tailored AI chatbots in the hope that the technical barriers to entry in training their neural networks are diminished. Specifically, developing a robust, highly specialized chatbot could significantly diminish barriers faced by Philadelphia businesses and residents, providing accurate, personalized, and timely responses to inquiries related to city regulations, business operations, community programs, and daily living.

There is opportunity in the massive wealth of documents stored on the internet. Over two decades ago, Fabrizio Sebastiani of the National Research Council of Italy, in his paper, Machine Learning in Automated Text Categorization, defined the backbone of working AI knowledge. He stated, "In the last 10 years content-based document management tasks (collectively known as information retrieval—IR) have gained a prominent status in the information systems field due to the increased availability of documents in digital form and the ensuing need to access them in flexible ways" (Sebastiani 1). Similar to the wide-scale adoption of the internet at the turn of the millennium, 24 years later, the public's ability to access digital documentation will rapidly change again with the advancement of AI chatbots. But now, chatbots have begun to bridge direct communication between users and their computers.

Before training AI chatbots, various methods, and machine learning techniques are employed to decompose textual data into a numerical vector format the computer can understand. Vectorization methods such as Bag of Words (BoW) counts word occurrences, and Term Frequency-Inverse Document Frequency (TF-IDF) accesses word importance. Additionally, cosine similarity measures the degree to which these vectors relate. In an article defining these methods, Prasoon Singh of Analytics Vidhya explains, "In language processing, the vectors are derived from textual data to reflect various linguistic properties of the text" (Singh, 2019). These vectors are then fed into neural networks with multiple layers of nodes that mimic the human brain's functioning, enabling chatbots to process language in a human-like way.

Neural networks consist of layers of nodes, or artificial neurons, that simulate the functioning of the human brain. Building these networks on a mass scale is a complicated endeavor, but the fundamental requirements for industry-level AI can be gleaned from leading bot models such as OpenAI's GPT-4, or Google's PaLM 2. Two state-of-the-art and

open-source deep learning systems appear in both models: BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory).

BERT revolutionizes language understanding by pretraining deep bidirectional representations reading the content forward and backward. Researchers at Google's AI laboratory released BERT's architecture to the public in a 2018 paper, stating, "the pre-trained BERT model can be fine tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications" (Devlin et al.), and later saying, BERT is "conceptually simple and empirically powerful."

In conjunction with BERT, LSTM enhances the chatbot's capabilities. Van Houdt et al. describe LSTM as a model that "has transformed both machine learning and neurocomputing fields... One reason for this recurrent network's success is its ability to handle the exploding/vanishing gradient problem" (Van Houdt, Mosquera, and Nápoles 1). The LSTM model's ability to decidedly remember or forget data based on relevance mimics the human brain and is vital for a chatbot's conversational fluency and context awareness.

Building a specialized AI chatbot with capabilities akin to ChatGPT is significantly expensive and complex. Based on a co-published paper by Nvidia and Microsoft Research, Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM, even reconstructing models like GPT-3 would be daunting. Their paper notes the following reasons, "training such models... is challenging for two reasons: (a) it is no longer possible to fit the parameters of these models in the main memory of even the largest GPU... and (b) the high number of compute operations required can result in unrealistically long training times" (Narayanan et al. 1). The resource intensity and technical demands of creating advanced AI chatbots are only feasible for some developers. Still, it is helpful to define the fundamentals and prepare data and plans for when the technological opportunity for smaller groups arises to create a highly curated chatbot.

Developing a chatbot for Philadelphia begins with accessing and compiling extensive data from various city-related documents. The City of Philadelphia's official website, phila.gov, is a rich source, hosting over 1,500 publications from its 83 departments covering various topics from systems, resources, and programs. The 161 programs consist of operations from Discarded Tire Clean Up to Tuberculosis Control. Additionally, nearly 400 extensive datasets are publicly available at OpenDataPhilly. The city also shares software repositories for all of its public websites on GitHub, including tools like the Recycling and Donation Finder

Map and School Let Out Times. Furthermore, since 2018, the City Council has enacted 610 regulations and laws, detailed at phila.gov/departments/department-of-records/proposed-regulations. The comprehensive book of codes is managed through American Legal Publishing and is accessible at codelibrary.amlegal.com. It's important to mention that this is only documentation from the City of Philadelphia. Even more voluminous and comprehensive resources, regulations, and laws come from the State of Pennsylvania and the United States in which Philadelphia resides.

The output of an AI chatbot could guide anyone in tasks such as creating businesses, buildings, and schools, reconstructing current regulations and resources to fill in gaps, and creating a simplified system. Managing a city, in addition to navigating its bureaucratic processes, is challenging. In the study "Organizational Effectiveness and Bureaucratic Red Tape," authors Pandey, Coursey, and Moynihan analyze bureaucratic systems and their impact on organizational effectiveness. They highlight the significant challenge posed by red tape in public management, stating, "Efforts to improve the effectiveness of government agencies inevitably target red tape... the scant consideration of the effect of red tape on organizational effectiveness is akin to ignoring the proverbial 'elephant in the room'" (Pandey et al., 2007, p. 399). Cumbersome processes from the government, while possibly being unavoidable for governance, can be better managed or navigated by AI.

Numerous institutions and companies have already worked to move tailored AI chatbots forward. The study "Building Emotional Support Chatbots in the Era of LLMs" by Zheng, Liao, Deng, and Nie exemplifies an approach where, rather than focusing on retraining a new model, the researchers recognized the enduring value of input-output datasets as a means of encoding information for a future AI system. The team meticulously compiled datasets of dialogues, drawing from scientific literature used by psychiatrists and therapists in patient interactions. They repeatedly had ChatGPT modify the dialogues, asking the chatbot to change scenarios, generating discussions with increased variance. Each conversation was validated to ensure its suitability for professional emotional support. They assembled 11,000 conversations with 36 different emotional support scenarios. This research demonstrates the use of a valuable tool for knowledge-building. Instead of spending time vectorizing data, use ML and careful work to curate and expand upon the text.

Alternatively, researchers working on the EduChat application decided to focus on constructing an AI model using existing, publicly available technologies. Developed by Ph.D. student Yuhao Dan of East China Normal University and Ze Zhou of ZhuQingTing Data Technology Ltd. EduChat, a highly curated chatbot past initial development, can offer in-

sights. EduChat is not built on ChatGPT or another existing AI. It is trained from scratch on educational materials, including textbooks, exams, and assignments from Chinese secondary schools. The platform leverages well-known developer tools such as Twilio Conversations API for managing its chat service and Google Firebase for data storage. The team used open-source tools and datasets, including four million cleaned, diverse instructions. They focused on essay assessment, Socratic teaching, and emotional support. This endeavor showcases that with existing technology and a focused approach, even small teams can develop a tool showing initial potential as an advanced domain-specific chatbot.

Notice how emotional support dialogue was a central component of the data of the two real-world applications of EduChat and the emotional support bot. The AI needs its knowledge co-mingled with guidance, precedence, and examples on how to communicate information. In addition, it is vital to maintain order between topics and categories of information. A chatbot geared toward the people of Philadelphia should be trained on real conversations amongst its factual data. Realistic sources include a high focus on Q&A sections of Philadelphia-related websites, online forums discussing Philadelphia, or potential inquiry data provided by the city's government.

An approach for a developer or small organization to start building a chatbot tailored to Philadelphia will require the effective storage and organization of large amounts of relevant data. For the simple storage and archiving of information, he Hadoop Distributed File System is an open-source industry go-to. Developers can send massive amounts (thousands of terabytes) of "documents" into storage. The system prioritizes rapid retrieval, which could effectively train a specialized chatbot when the opportunity arises. Conversely, NoSQL is an open-source database with more controls, such as data manipulation, except it does not have the more intensive relational requirements of SQL.

Although only some industry-leading examples exist, a solid framework of requirements can be established. A massive wealth of data about Philadelphia exists, which could be expanded through machine learning and have its meaning encapsulated through data vectorization techniques. By evaluating case studies via comparative analysis, we can discover lessons and strategies for the future development of this type of bot. In the context of Philadelphia, the advent of a customized, high-performance chatbot harnessing the power of advanced AI technologies like GPT-4, BERT, and LSTM, promises a transformative impact. This would be a pivotal tool in breaking down information barriers by offering assistance tailored to the unique needs of the city's businesses and residents.

Works Cited

Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Google AI Language, Oct. 2018, research.google/pubs/pub47751/.

Dan, Yuhao, et al. "EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education." School of Computer Science and Technology, East China Normal University; Institute of AI for Education, ECNU; School of Computer Science, Fudan University; ZhuQingTing Data Technology (Zhejiang) Co., Ltd., 5 Aug. 2023, arxiv.org/pdf/2308.02773.pdf.

Narayanan, Deepak, et al. "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM." arXiv, NVIDIA, Stanford University, Microsoft Research, 2021, arxiv.org/pdf/2104.04473.pdf.

Pandey, Sanjay K., David H. Coursey, and Donald P. Moynihan. "Organizational Effectiveness and Bureaucratic Red Tape: A Multimethod Study." Public Performance & Management Review, vol. 30, no. 3, Mar. 2007, jstor.org/stable/20447639.

Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization." ACM Computing Surveys, vol. 34, no. 1, Mar. 2002, pp. 1–47, doi:10.1145/505282.505283.

Singh, Prasoon. "Fundamentals of Bag Of Words and TF-IDF." Analytics Vidhya, 4 Sept. 2019, medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22.

Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles. "A Review on the Long Short-Term Memory Model." Artificial Intelligence Review, vol. 53, no. 1, Dec. 2020, doi:10.1007/s10462-020-09838-1, researchgate.net/publication/340493274_A_Review_on_the_Long_Short-Term_Memory_Model.

Zheng, Zhonghua, et al. "Building Emotional Support Chatbots in the Era of LLMs." Harbin Institute of Technology, Shenzhen, Singapore Management University, National University of Singapore, 17 Aug. 2023, arxiv.org/pdf/2308.11584.pdf.