

Deep Learning Improves Natural Language Processing

Making conversational agents real.

Abstract

Modern natural language processing (NLP) and its subfield natural language understanding (NLU) combine sophisticated computational linguistics, probabilistic approaches, machine learning, and deep learning.

Natural language processing (NLP) and natural language understanding (NLU) use advanced computing techniques to analyze, contextualize, and understand human speech. NLP and NLU are improving at such a rapid rate that they can enable meaningful and normal conversations between machines and people. They make conversational agents a reality.

Modern natural language processing (NLP) and its subfield natural language understanding (NLU) combine sophisticated computational linguistics, probabilistic approaches, machine learning, and deep learning. These technologies enable computers to organize and structure the knowledge required for conversational agents to understand speech. They give computers the capability to summarize, translate, recognize named entities, extract relationships, analyze sentiments, and detect and segment topics.

Deep Learning Enables Lifelike Conversations

Recently, deep learning techniques have made NLP far more effective. Conversational agents powered by NLP and enabled by deep learning can have natural, normal conversations that are indistinguishable from conversations between two humans.

NLP Model #1: Retrieval


NLP approaches use one of two models, retrieval or generative. The retrieval model is easier to design and implement. It typically uses canned responses combined with heuristics to select probable responses based on an input text and its derived context. The model outputs many probable responses with their respective scores. To select an appropriate response, the designer of the agent would calculate the scores for these responses and generally select the response with the highest score.

Retrieval models don't make grammatical mistakes, but they are unable to handle conversations for which there are no predefined responses. A retrieval model is fine for dealing with queries requiring simple answers—for example, "What is my bank balance?" But they are terrible for dealing with complex, highly variable queries and responses. Regardless of the tone of voice used by the machine, these models will seem robotic, canned, and scripted—in other words, not conversational.

NLP Model #2: Generative

Generative models, on the other hand, are more flexible and human-like than retrieval models—in part because they don't need a repository of canned responses. As their name suggests, generative models can generate completely new responses and can handle unforeseen cases dynamically. They can refer back to entities used in context throughout the dialog, for example, and engage in more human-like conversations. They are, in short, smarter and more complex.

A generative model, for example, can be trained to handle a natural, unprompted request for a flight reservation. A consumer could say, for example, "I need to fly to Mumbai in about a week but first I need to get to Chennai with at least two days there." This would be essentially impossible for even a very good retrieval model, but can be done with a generative one.



Generative models, on the other hand, are more flexible and human-like than retrieval models—in part because they don't need a repository of canned responses.

Real conversations—free-form, unstructured—are often referred to as “open domain.” Instead of limiting the speaker to a finite set of responses, the speaker can ask anything and the agent should respond appropriately.

But the intelligence and complexity of the generative model come at a price. First, training generative models is very hard; second, they require large amounts of data. And even with good training and good data, generative models are currently prone to making grammatical errors. Generative models use techniques like machine translation to translate an input and match it to the response output (see Figure 1).

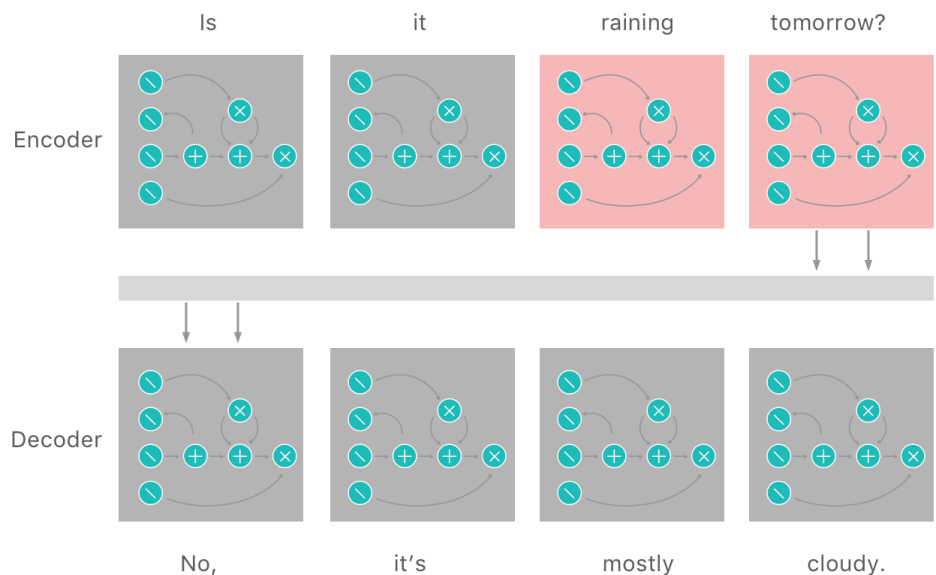


Figure 1: Shown are Long Short-Term Memory cells running the same algorithm to encode and decode a query and response. The question is worded strangely (most people would just say, "Will it rain tomorrow?"), but a generative model-based NLP can easily understand the intent (the query is about the weather) and the entity, in this case the timeframe of tomorrow. Note that the response is conversational; not just "No" or even "No, it's not going to rain tomorrow" but a fuller more contextually appropriate response. The system isn't just answering the explicit question but the implied question.

Generative is the Future

Although research is progressing in the area of generative models, most current NLP systems are based on retrieval-based models. However, given their rather elementary nature, retrieval-based models will always be limited in application and scope to what are called “closed domain” situations. A closed domain will rely on a fairly limited vocabulary and a set of responses that can be predicted, scripted, and programmed. But even some closed domain applications can be very complex to model for retrieval-based systems—for example, banking or airline customer support.

Real conversations—free-form, unstructured—are often referred to as “open domain.” Instead of limiting the speaker to a finite set of responses, the speaker can ask anything and the agent should respond appropriately. For a natural language processing system, an open domain system is vastly more complex than a traditional retrieval-based model can possibly handle.

In fact, it is currently impossible for retrieval-based NLP models to converse in open domains. The topics are infinite, and human programmers won't be able to create rules and responses for every situation. However, newer deep learning techniques like sequence-to-sequence are rapidly taking us closer to the goal of highly efficient and accurate generative models.

One characteristic of generative models is the ability to handle longer, more free-form conversations. Longer conversations can be composed of multiple intents, contexts, and entities. A good conversational agent must derive and maintain both linguistic and physical contexts to generate understandable and sensible responses in a conversation.

One experimental approach is the generative hierarchical neural network model, which tries to address this by plotting conversational vectors. Several derived data attributes like location, event, date, time, and user information assist in maintaining, creating, and deriving context.¹

For example, in each of the following similar queries, the conversational agent should generate similar, consistent responses:

- How are my accounts looking?
- What are the balances in my account?
- Are my balances looking ok?
- What's the status of my account balances?

A human customer service representative could easily understand any of those queries and respond with an appropriate answer. But the challenge is enormous for a machine-based conversational agent.

A consistent output can be achieved by training that uses a lot of data across several users. Persona-based conversational neural models² are enabling the modeling of such consistent responses.

Generative models tend to respond with generic responses until they achieve a certain level of diversity in their models, or through some of the latest and most cutting-edge techniques such as intentional objective functional modeling.

One characteristic of generative models is the ability to handle longer, more free-form conversations.

Deep Learning Enables “Real” Conversation

Conversational agents are proving to be a highly effective way to assist and engage with customers across several industries. Today, it’s possible to build, production-quality conversational agents using retrieval-based models and generative models in closed domains (see Figure 2). Even in these more limited spheres, a conversational agent requires state-of-the-art deep learning capabilities. Rapid progress in deep learning is leading to iterative advances in the capabilities and applications of conversational agents powered by NLP and NLU.

Open Domain	Currently Not Possible	Deep Learning (Hardest)
Closed Domain	Rules-Based (Easy)	Machine Learning (Hard)
	Retrieval Based	Generative Based

Figure 2: This grid plots the two NLP models against the two types of domains; it will require deep learning techniques to create generative-based, open domain systems—key to the creation of truly conversational agents for all kinds of customer-facing applications.

Deep learning offers the most powerful way to build truly great generative models that can work in open domains—and, accordingly, the most powerful way to build truly great conversational agents.

Here at TCS, we are working with several clients to incorporate deep learning and natural language processing into real-world applications. Two examples: We are creating state-of-the-art conversational agents for some major international airlines and global financial institutions that will make responding to customer queries not just more efficient but also more effective.

References

- [1] For more information on this approach, see “Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models” by Iulian V. Serban et al. at <https://arxiv.org/abs/1507.04808>
- [2] For more information on persona-based conversational neural models, visit <https://arxiv.org/abs/1603.06155>

About The Author

Sunil Karkera

Head of Digital Reimagination™ Studio

Sunil is an engineer by profession, and over the past 20 years, he founded three successful startups in Silicon Valley (wireless products, SaaS solutions, and mobile applications), led information systems groups at News Corporation, and worked and consulted for companies such as Nest, Westinghouse, Dell-Sonicwall, Lockheed, and Siebel. He is also a trained typographer.

About Artificial Intelligence at TCS and our Digital Reimagination™ Studio

At TCS, we operate with the belief that the future of business will be driven by five powerful digital forces: mobility and pervasive computing, the cloud, Big Data, social media, and artificial intelligence (AI). We are applying AI, notably deep learning, to all kinds of applications from autonomous vehicles to the analysis of sensor data from the Internet of Things, from fraud detection to natural language processing and conversational agents.

The TCS Digital Reimagination™ Studio is dedicated to helping businesses create fundamentally new experiences by reimagining industries through creative thinking. The Studio brings a start-up culture to large enterprise clients by leveraging the best of world-class creative, design, engineering, and business domain experts. The result is business transformation through rapid product prototyping and extremely agile collaboration.

Contact

To learn more, contact the TCS Digital Reimagination™ Studio at analytics.insights@tcs.com

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w

Feedburner: <http://feeds2.feedburner.com/tcswhitepapers>

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting, and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering, and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com