Tech Accelerator
A guide to artificial intelligence in the enterprise

DEFINITION

language modeling

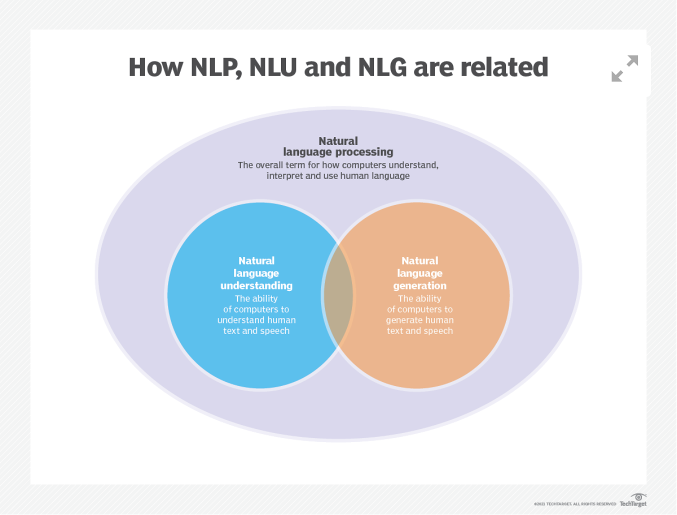
By Nick Barney, Technology Writer | Ben Lutkevich, Technical Features Writer

What is language modeling?

Language modeling, or LM, is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Language models analyze bodies of text data to provide a basis for their word predictions.

Language modeling is used in artificial intelligence (AI), natural language processing (NLP), natural language understanding and natural language generation systems, particularly ones that perform text generation, machine translation and question answering.

Large language models (LLMs) also use language modeling. These are advanced language models, such as OpenAI's GPT-3 and Google's Palm 2, that handle billions of training data parameters and generate text output.


How NLP, NLU and NLG are related
Natural language processing
The overall term for how computers understand, interpret and use human language
Natural language understanding
The ability of computers to understand human text and speech
Natural language generation
The ability of computers to generate human text and speech
©2021 TECHTARGET. ALL RIGHTS RESERVED TechTarget

Natural language processing incorporates natural language generation and natura language understanding.

How language modeling works

Language models determine word probability by analyzing text data. They interpret this data by feeding it through an algorithm that establishes rules for context in natural language. Then, the model applies these rules in language tasks to accurately predict or produce new sentences. The model essentially learns the features and characteristics of basic language and uses those features to understand new phrases.

THIS ARTICLE IS PART OF

A guide to artificial intelligence in the enterprise

Which also includes:

The future of AI: What to expect in the next 5 years

Types of AI algorithms and how they work

10 top artificial intelligence certifications and courses for 2023

There are several different probabilistic approaches to modeling language. They vary depending on the purpose of the language model. From a technical perspective, the various language model types differ in the amount of text data they analyze and the math they use to analyze it. For example, a language model designed to generate sentences for an automated social media bot might use different math and analyze text data in different ways than a language model designed for determining the likelihood of a search query.

Language modeling types

There are several approaches to building language models. Some common statistical language modeling types are the following:

- **N-gram.** This simple approach to a language model creates a probability distribution for a sequence of *n*. The *n* can be any number and defines the size of the *gram*, or sequence of words or random variables being assigned a probability. This allows the model to accurately predict the next word or variable in a sentence. For example, if *n* = 5, a gram might look like this: "can you please call me." The model then assigns probabilities using sequences of *n* size. Basically, *n* can be thought of as the amount of context the model is told to consider. Some types of n-grams are unigrams, bigrams, trigrams and so on. N-grams can also help detect malware by analyzing strings in a file.
- **Unigram.** This is the simplest type of language model. It doesn't look at any conditioning context in its calculations. It evaluates each word or term independently. Unigram models commonly handle language processing tasks such as information retrieval. The unigram is the foundation of a more specific model variant called the query likelihood model, which uses information retrieval to examine a pool of documents and match the most relevant one to a specific query.
- **Bidirectional.** Unlike n-gram models, which analyze text in one direction, backward, bidirectional models analyze text in both directions, backward and forward. These models can predict any word in a sentence or body of text by using every other word in the text. Examining text bidirectionally increases result accuracy. This type is often used in machine learning models and speech generation applications. For example, Google uses a bidirectional model to process search queries.
- **Exponential.** Also known as maximum entropy models, exponential models are more complex than n-grams. Simply put, it evaluates text using an equation that combines feature functions and n-grams. Basically, this type of model specifies features and parameters of the desired results and, unlike n-grams, leaves the analysis parameters more ambiguous --- it doesn't specify individual gram sizes, for example. The model is based on the principle of entropy, which states that the probability distribution with the most entropy is the best choice. In other words, the model with the most chaos, and least room for assumptions, is the most accurate. Exponential models are designed to maximize cross-entropy, which minimizes the amount of statistical assumptions that can be made. This lets users have more trust in the results they get from these models.
- **Neural language models.** Neural language models use deep learning techniques to overcome the limitations of n-gram models. These models use neural networks, such as recurrent neural networks (RNNs), and transformers to capture complex patterns and dependencies in text. RNN language models include long short-term memory and gated recurrent unit models. These models can consider all previous words in a sentence when predicting the next word. This allows them to capture long-range dependencies and generate more contextually relevant text. Transformers use self-attention mechanisms to weigh the importance of different words in a sentence, enabling them to capture global dependencies. Generative AI models, such as GPT-3 and Palm 2, are based on the transformer architecture.
- **Continuous space.** This is another type of neural language model that represents words as a nonlinear combination of weights in a neural network. The process of assigning a weight to a word is also known as word embedding. This type of model becomes especially useful as data sets get bigger, because larger data sets often include more unique words. The presence of a lot of unique or rarely used words can cause problems for linear models such as n-grams. This is because the amount of possible word sequences increases, and the patterns that inform results become weaker. By weighting words in a nonlinear, distributed way, this model can "learn" to approximate words and not be misled by any unknown values. Its "understanding" of a given word isn't as tightly tethered to the immediate surrounding words as it is in n-gram models.

The models listed above are more general statistical approaches from which more specific variant language models are derived. For example, as mentioned in the n-gram description, the query likelihood model is a more specific or specialized model that uses the n-gram approach. Model types can be used in conjunction with one another.

The models listed also vary in complexity. Broadly speaking, more complex language models are better at NLP tasks because language itself is extremely complex and always evolving. Therefore, an exponential model or continuous space model might be better than an n-gram for NLP tasks because they're designed to account for ambiguity and variation in language.

A good language model should also be able to process long-term dependencies, handling words that might derive their meaning from other words that occur in far-away, disparate parts of the text. A language model should be able to understand when a word is referencing another word from a long distance, as opposed to always relying on proximal words within a certain fixed history. This requires a more complex model.

Importance of language modeling

Language modeling is crucial in modern NLP applications. It's the reason that machines can understand qualitative information. Each language model type, in one way or another, turns qualitative information into quantitative information. This allows people to communicate with machines as they do with each other, to a limited extent.

Language modeling is used in a variety of industries including information technology, finance, healthcare, transportation, legal, military and government. In addition, it's likely that most people have interacted with a language model in some way at some point in the day, whether through Google search, an autocomplete text function or engaging with a voice assistant.

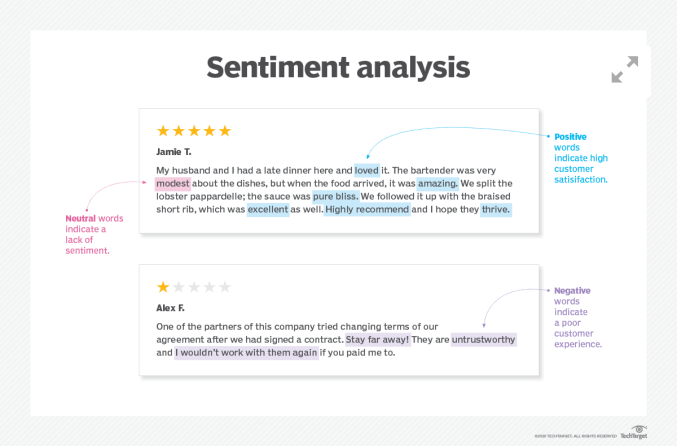
The roots of language modeling can be traced back to 1948. That year, Claude Shannon published a paper titled "A Mathematical Theory of Communication." In it, he detailed the use of a stochastic model called the Markov chain to create a statistical model for the sequences of letters in English text. This paper had a large impact on the telecommunications industry and laid the groundwork for information theory and language modeling. The Markov model is still used today, and n-grams are tied closely to the concept.

Uses and examples of language modeling

Language models are the backbone of NLP. Below are some NLP use cases and tasks that employ language modeling:

- **Speech recognition.** This involves a machine being able to process speech audio. Voice assistants such as Siri and Alexa commonly use speech recognition.
- **Text generation.** This application uses prediction to generate coherent and contextually relevant text. It has applications in creative writing, content generation, and summarization of structured data and other text.
- **Chatbots.** These bots engage in humanlike conversations with users as well as generate accurate responses to questions. Chatbots are used in virtual assistants, customer support applications and information retrieval systems.
- **Machine translation.** This involves the translation of one language to another by a machine. Google Translate and Microsoft Translator are two programs that do this. Another is SDL Government, which is used to translate foreign social media feeds in real time for the U.S. government.
- **Parts-of-speech tagging.** This use involves the markup and categorization of words by certain grammatical characteristics. This model is used in the study of linguistics. It was first and perhaps most famously used in the study of the Brown Corpus, a body of random English prose that was designed to be studied by computers. This corpus has been used to train several important language models, including one used by Google to improve search quality.
- **Parsing.** This use involves analysis of any string of data or sentence that conforms to formal grammar and syntax rules. In language modeling, this can take the form of sentence diagrams that depict each word's relationship to the others. Spell-checking applications use language modeling and parsing.
- **Optical character recognition.** This application involves the use of a machine to convert images of text into machine-encoded text. The image can be a scanned document or document photo, or a photo with text somewhere in it -- on a sign, for example. Optical character recognition is often used in data entry when processing old paper records that need to be digitized. It can also be used to analyze and identify handwriting samples.

- **Information retrieval.** This approach involves searching in a document for information, searching for documents in general and searching for metadata that corresponds to a document. Web browsers are the most common information retrieval applications.
- **Observed data analysis.** These language models analyze observed data such as sensor data, telemetric data and data from experiments.
- **Sentiment analysis.** This application involves determining the sentiment behind a given phrase. Specifically, sentiment analysis is used to understand opinions and attitudes expressed in a text. Businesses use it to analyze unstructured data, such as product reviews and general posts about their product, as well as analyze internal data such as employee surveys and customer support chats. Some services that provide sentiment analysis tools are Repustate and HubSpot's Service Hub. Google's NLP tool Bert is also used for sentiment analysis.



Sentiment analysis

★★★★★
Jamie T.
My husband and I had a late dinner here and loved it. The bartender was very modest about the dishes, but when the food arrived, it was amazing. We split the lobster pappardelle; the sauce was pure bliss. We followed it up with the braised short rib, which was excellent as well. Highly recommend and I hope they thrive.

★☆☆☆☆
Alex F.
One of the partners of this company tried changing terms of our agreement after we had signed a contract. Stay far away! They are untrustworthy and I wouldn't work with them again if you paid me to.

Positive words indicate high customer satisifaction.
Neutral words indicate a lack of sentiment.
Negative words indicate a poor customer experience.
©2020 TECHTARGET. ALL RIGHTS RESERVED
TechTarget

Sentiment analysis uses language modeling technology to detect and analyze key customer reviews and posts.

The future of language modeling

State-of-the-art LLMs have demonstrated impressive capabilities in generating human language and humanlike text and understanding complex language patterns. Leading models such as those that power ChatGPT and Bard have billions of parameters and are trained on massive amounts of data. Their success has led them to being implemented into Bing and Google search engines, promising to change the search experience.

New data science techniques, such as fine-tuning and transfer learning, have become essential in language modeling. Rather than training a model from scratch, fine-tuning lets developers take a pre-trained language model and adapt it to a task or domain. This approach has reduced the amount of labeled data required for training and improved overall model performance.

As language models and their techniques become more powerful and capable, ethical considerations become increasingly important. Issues such as bias in generated text, misinformation and the potential misuse of AI-driven language models have led many AI experts and developers such as Elon Musk to warn against their unregulated development.

Language modeling is one of the leading techniques in generative AI. Learn the top eight biggest ethical concerns for generative AI.

This was last updated in October 2023

Continue Reading About language modeling

- Q&A: How to start learning natural language processing
- Intersection of generative AI, cybersecurity and digital trust
- What does NLP mean for augmented analytics?
- Experts predict NLP to be biggest BI trend this year
- NLP and AI boost the automated data warehouse

Related Terms