Title: Language model

URL: https://en.wikipedia.org/wiki/Language_model

PageID: 1911810

Categories: Category:Language modeling, Category:Markov models, Category:Statistical natural language processing

---

A language model is a model of the human brain's ability to produce natural language . [1] [2] Language models are useful for a variety of tasks, including speech recognition , [3] machine translation , [4] natural language generation (generating more human-like text), optical character recognition , route optimization , [5] handwriting recognition , [6] grammar induction , [7] and information retrieval . [8] [9]

Large language models (LLMs), currently their most advanced form, are predominantly based on transformers trained on larger datasets (frequently using texts scraped from the public internet). They have superseded recurrent neural network -based models, which had previously superseded the purely statistical models, such as the word n -gram language model .

History

Noam Chomsky did pioneering work on language models in the 1950s by developing a theory of formal grammars . [10]

In 1980, statistical approaches were explored and found to be more useful for many purposes than rule-based formal grammars. Discrete representations like word n -gram language models , with probabilities for discrete combinations of words, made significant advances.

In the 2000s, continuous representations for words, such as word embeddings , began to replace discrete representations. [11] Typically, the representation is a real-valued vector that encodes the meaning of the word in such a way that the words that are closer in the vector space are expected to be similar in meaning, and common relationships between pairs of words like plurality or gender.

Pure statistical models

In 1980, the first significant statistical language model was proposed, and during the decade IBM performed ‘ Shannon -style’ experiments, in which potential sources for language modeling improvement were identified by observing and analyzing the performance of human subjects in predicting or correcting text. [12]

Models based on word n -grams

A word n -gram language model is a purely statistical model of language. It has been superseded by recurrent neural network –based models, which have been superseded by large language models . [13] It is based on an assumption that the probability of the next word in a sequence depends only on a fixed size window of previous words. If only one previous word is considered, it is called a bigram model; if two words, a trigram model; if n – 1 words, an n -gram model. [14] Special tokens are introduced to denote the start and end of a sentence ■ s ■ {\displaystyle \langle s\rangle } and ■ / s ■ {\displaystyle \langle /s\rangle } .

Exponential

Maximum entropy language models encode the relationship between a word and the n -gram history using feature functions. The equation is

P (w m ■ w 1 , … , w m – 1) = 1 Z (w 1 , … , w m – 1) exp ■ (a T f (w 1 , … , w m)) {\displaystyle P(w_{m}\mid w_{1},\ldots ,w_{m-1})={\frac {1}{Z(w_{1},\ldots ,w_{m-1})}}\exp(a^{T}f(w_{1},\ldots ,w_{m}))}

where Z (w 1 , … , w m – 1) {\displaystyle Z(w_{1},\ldots ,w_{m-1})} is the partition function , a {\displaystyle a} is the parameter vector, and f (w 1 , … , w m) {\displaystyle f(w_{1},\ldots ,w_{m})} is the feature function. In the simplest case, the feature function is just an indicator of the presence of a certain n -gram. It is helpful to use a prior on a {\displaystyle a} or some form of regularization .

The log-bilinear model is another example of an exponential language model.

Skip-gram model

Skip-gram language model is an attempt at overcoming the data sparsity problem that the preceding model (i.e. word n -gram language model) faced. Words represented in an embedding vector were not necessarily consecutive anymore, but could leave gaps that are skipped over (thus the name "skip-gram"). [15]

Formally, a k -skip- n -gram is a length- n subsequence where the components occur at distance at most k from each other.

For example, in the input text:

the set of 1-skip-2-grams includes all the bigrams (2-grams), and in addition the subsequences

In skip-gram model, semantic relations between words are represented by linear combinations , capturing a form of compositionality . For example, in some such models, if v is the function that maps a word w to its n -d vector representation, then

v (k i n g) – v (m a l e) + v (f e m a l e) ≈ v (q u e e n) {\displaystyle v(\mathrm {king})-v(\mathrm {male})+v(\mathrm {female})\approx v(\mathrm {queen})}

Neural models

Recurrent neural network

Continuous representations or embeddings of words are produced in recurrent neural network -based language models (known also as continuous space language models). [18] Such continuous space embeddings help to alleviate the curse of dimensionality , which is the consequence of the number of possible sequences of words increasing exponentially with the size of the vocabulary, further causing a data sparsity problem. Neural networks avoid this problem by representing words as non-linear combinations of weights in a neural net. [19]

Large language models

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks
Transformer Vision
Vision
Mamba
Spiking neural network
Memtransistor
Electrochemical RAM (ECRAM)
Q-learning
Policy gradient
SARSA
Temporal difference (TD)
Multi-agent Self-play
Self-play
Active learning
Crowdsourcing
Human-in-the-loop
Mechanistic interpretability
RLHF
Coefficient of determination
Confusion matrix
Learning curve
ROC curve
Kernel machines
Bias–variance tradeoff
Computational learning theory
Empirical risk minimization
Occam learning
PAC learning
Statistical learning
VC theory
Topological deep learning
AAAI
ECML PKDD
NeurIPS
ICML
ICLR
IJCAI
ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Although sometimes matching human performance, it is not clear whether they are plausible cognitive models . At least for recurrent neural networks, it has been shown that they sometimes learn patterns that humans do not, but fail to learn patterns that humans typically do. [23]

Evaluation and benchmarks

Evaluation of the quality of language models is mostly done by comparison to human created sample benchmarks created from typical language-oriented tasks. Other, less established, quality tests examine the intrinsic character of a language model or compare two such models. Since language models are typically intended to be dynamic and to learn from data they see, some proposed models investigate the rate of learning, e.g., through inspection of learning curves. [24]

Various data sets have been developed for use in evaluating language processing systems. [25] These include:

Massive Multitask Language Understanding (MMLU) [26]

Corpus of Linguistic Acceptability [27]

GLUE benchmark [28]

Microsoft Research Paraphrase Corpus [29]

Multi-Genre Natural Language Inference

Question Natural Language Inference

Quora Question Pairs [30]

Recognizing Textual Entailment [31]

Semantic Textual Similarity Benchmark

SQuAD question answering Test [32]

Stanford Sentiment Treebank [33]

Winograd NLI

BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC, OpenBookQA, NaturalQuestions, TriviaQA, RACE, BIG-bench hard, GSM8k, RealToxicityPrompts, WinoGender, CrowS-Pairs [34]

See also

Linguistics portal

Mathematics portal

Technology portal

Artificial intelligence and elections – Use and impact of AI on political elections

Cache language model

Deep linguistic processing

Ethics of artificial intelligence

Factored language model

Generative pre-trained transformer

Katz's back-off model

Language technology

Semantic similarity network

Statistical model

References

Further reading

Jay M. Ponte; W. Bruce Croft (1998). "A Language Modeling Approach to Information Retrieval". Research and Development in Information Retrieval . pp. 275– 281. CiteSeerX 10.1.1.117.4237 . doi : 10.1145/290941.291008 .

Fei Song; W. Bruce Croft (1999). "A General Language Model for Information Retrieval". Research and Development in Information Retrieval . pp. 279– 280. CiteSeerX 10.1.1.21.6467 . doi : 10.1145/319950.320022 .

Chen, Stanley F.; Joshua Goodman (1998). An Empirical Study of Smoothing Techniques for Language Modeling (Technical report). Harvard University. CiteSeerX 10.1.1.131.5458 .

v

t

e

AI-complete

Bag-of-words

n -gram Bigram Trigram

Bigram

Trigram

Computational linguistics

Natural language understanding

Stop words

Text processing

Argument mining

Collocation extraction

Concept mining

Coreference resolution

Deep linguistic processing

Distant reading

Information extraction

Named-entity recognition

Ontology learning

Parsing Semantic parsing Syntactic parsing

Semantic parsing

Syntactic parsing
Part-of-speech tagging
Semantic analysis
Semantic role labeling
Semantic decomposition
Semantic similarity
Sentiment analysis
Terminology extraction
Text mining
Textual entailment
Truecasing
Word-sense disambiguation
Word-sense induction
Compound-term processing
Lemmatisation
Lexical analysis
Text chunking
Stemming
Sentence segmentation
Word segmentation
Multi-document summarization
Sentence extraction
Text simplification
Computer-assisted
Example-based
Rule-based
Statistical
Transfer-based
Neural
BERT
Document-term matrix
Explicit semantic analysis
fastText
GloVe
Language model (large)
Latent semantic analysis
Seq2seq
Word embedding

Word2vec
Corpus linguistics
Lexical resource
Linguistic Linked Open Data
Machine-readable dictionary
Parallel text
PropBank
Semantic network
Simple Knowledge Organization System
Speech corpus
Text corpus
Thesaurus (information retrieval)
Treebank
Universal Dependencies
BabelNet
Bank of English
DBpedia
FrameNet
Google Ngram Viewer
UBY
WordNet
Wikidata
Speech recognition
Speech segmentation
Speech synthesis
Natural language generation
Optical character recognition
Document classification
Latent Dirichlet allocation
Pachinko allocation
Automated essay scoring
Concordancer
Grammar checker
Predictive text
Pronunciation assessment
Spell checker
Chatbot
Interactive fiction

Question answering
Virtual assistant
Voice user interface
Formal semantics
Hallucination
Natural Language Toolkit
spaCy
v
t
e
History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions
Regression Bias–variance tradeoff Double descent Overfitting
Bias–variance tradeoff
Double descent
Overfitting
Clustering
Gradient descent SGD Quasi-Newton method Conjugate gradient method
SGD
Quasi-Newton method
Conjugate gradient method
Backpropagation
Attention
Convolution
Normalization Batchnorm
Batchnorm
Activation Softmax Sigmoid Rectifier
Softmax
Sigmoid
Rectifier
Gating
Weight initialization
Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing
Warren Sturgis McCulloch
Walter Pitts
John von Neumann
Claude Shannon
Shun'ichi Amari
Kunihiko Fukushima
Takeo Kanade
Marvin Minsky
John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category