FLORIDA ATLANTIC UNIVERSITY

DATA SCIENCE, ANALYTICS, AND ARTIFICIAL INTELLIGENCE CONFERENCE

Natural Language Processing and AI: Neural and Symbolic Approaches

Nelson Correa, Ph.D.
Andinum AI (Consultant, Bank of America CDSO)

@nelscorrea - /in/ncorrea

SATURDAY NOVEMBER 14, 2020 FAU.EDU/DATA

NLP and AI: Neural and Symbolic Approaches

State of the Art: Neural NLP

- Neural networks (deep learning) have replaced symbolic and statistical approaches to NLP and AI. NLP is the "*hottest area of AI*" (*stateof.ai 2020*).

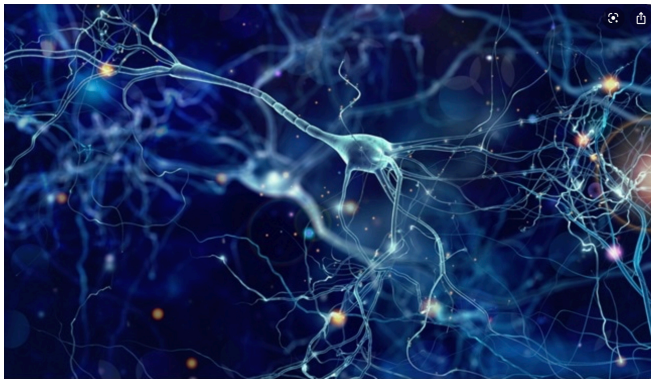
Pros: Human and super-human performance

- Unprecedented degree of accuracy on many NLP and AI tasks (end-to-end learning). Tasks that require high level cognitive ability can be successfully automated.

Cons: Large back box models

- Deep neural NLP models have grown to astronomical sizes (10^{11} parameters) raising issues of amount of training data required, computing and cost (vs. 0.1% parameters)
- Ethics of AI issues such as bias, privacy, fairness, risk, transparency, and accountability, all of which require model interpretability and prediction explainability.
- Data modeling, data governance and model risk management.

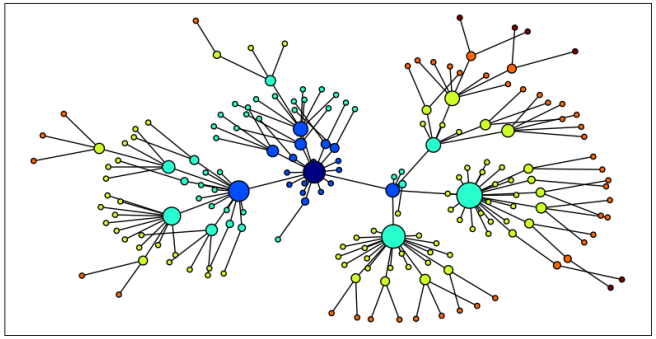
Hybrid Neuro-Symbolic AI / NLP



Outline

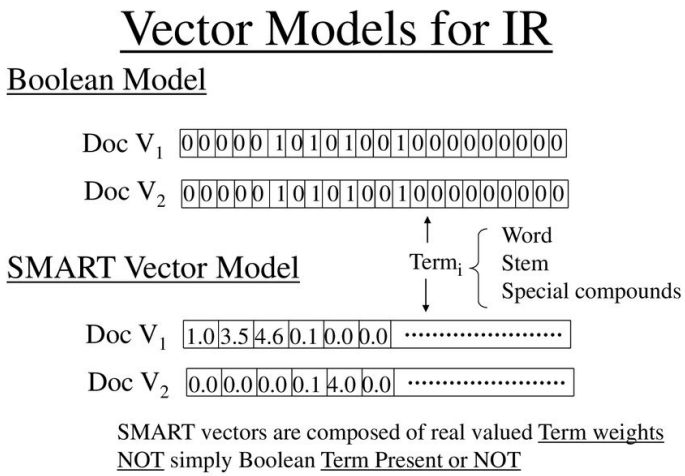
1. Words, symbols, sequences and vector spaces
2. The NLP pipeline
3. Language models
4. Neural NLP
5. Neuro-symbolic systems

Conclusion



Words, symbols, sequences & vector spaces

- Vocabulary: (finite) set of words (tokens, symbols, graphemes, “forms”)
  - Vocabulary V: 10,000s to millions of word forms (inflection, compound)
  - Discrete symbol representation (label, integer, one-hot binary vector)
- Language: set of sentences or documents (s*equences* of words); *n-grams*
  - Complexity: for vocabulary V of size $|V|$, there are $|V|^n$ n-grams
  - for n=3, $|V| = 100{,}000$, there are 10^{15} trigrams
- Vector spaces: technique for word, document and meaning representation
  - binary/discrete/continuous vectors; sparse/dense
- Examples:
  - Indexing/search: Salton SMART system, 1972
  - Document representation: *Bag of Words* (BoW): Binary, Count, TF-IDF
  - Word, sentence and document representation as fixed-length continuous dense vectors (word2vec, GloVe, doc2vec, etc.)



Dense semantic vector representations

- Vocabulary and documents
- Real-valued vector of dimension *d*
- d << |V|, independent of |V|
- word2vec, GloVe (d = 50, 100, 300)
- Word/document similarity measures

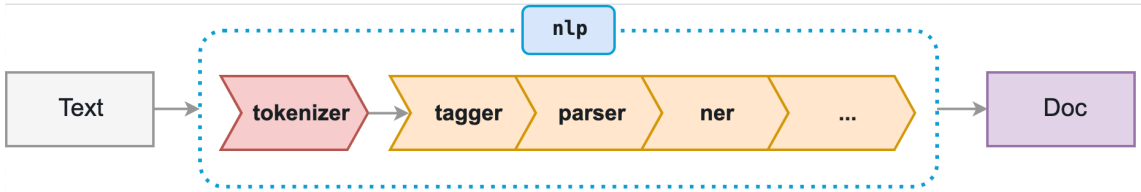
The NLP pipeline

Base NLP Tasks
- Tokenization, Tagging, Chunking, Parsing

Applications
- Search, Classification, Information extraction (KG), Text generation, Transcription, Translation, Question answering, Dialog, Chatbots, Reasoning

NLP Pipeline
- Text-feature vectorizers
  - Scikit-Learn
  - Core NLP, NLTK, spaCy, HuggingFace
- Symbolic NLP: Higher-level analysis
  - Parsing, logical form, semantics/SLR, discourse
- Neural NLP: Fixed-length vector representations
  - Output classifier or decoder (seq2seq)
  - e.g., Text classification: Logistic classifier
  - End of parsing with neural NLP?

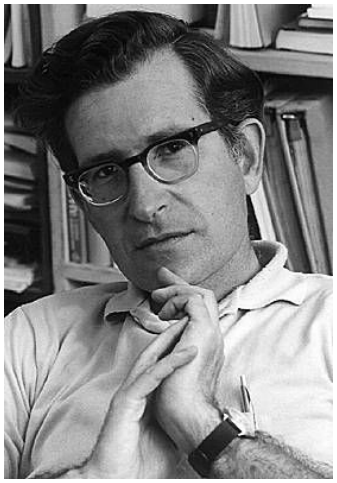
```
# spaCy NLP Pipeline
# "pipeline": ["tagger", "parser", "ner"]

import spacy
nlp = spacy.load("en_core_web_lg")
tokens = nlp("My input text")

https://spacy.io/usage/
```



Symbolic NLP: Sentence Parsing & Logical Form

Formal languages & linguistic theory (Chomsky, 1955-2000)

- Generative grammar (human language faculty)
- Compositionality of representations
- Acceptability of a sentence (binary)



Noam Chomsky

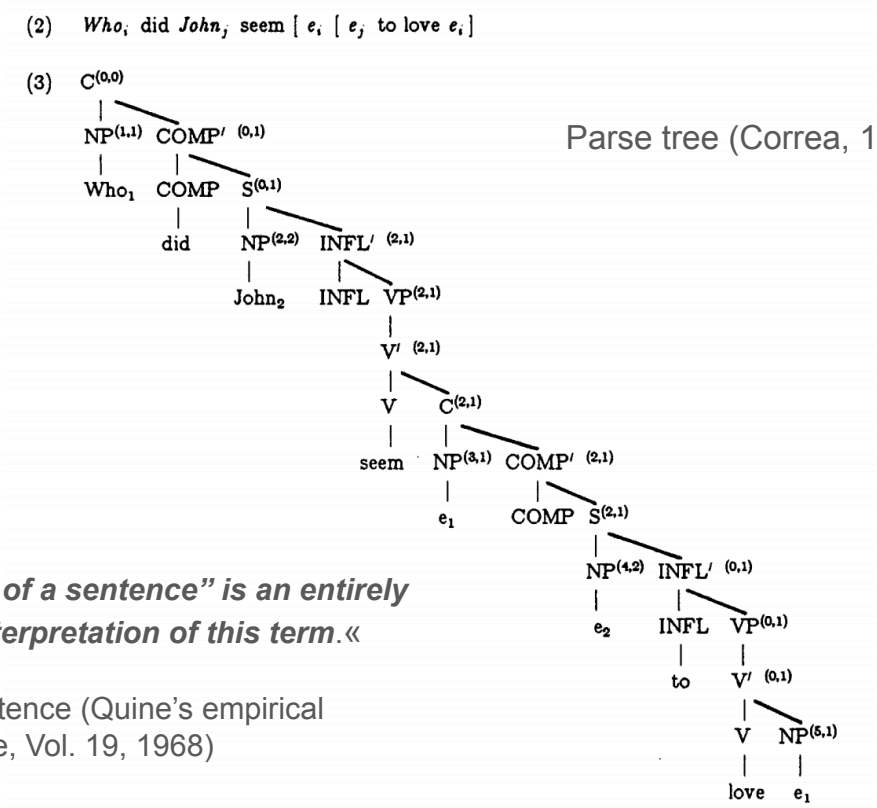
Who did John seem to love?

Bag of Words (BOW):

- `[john, seem, love]`

»… ***the notion of the "probability of a sentence" is an entirely useless one, under any known interpretation of this term***.«

Chomsky on the probability of a sentence (Quine's empirical assumptions, N. Chomsky, Synthese, Vol. 19, 1968)



Parse tree (Correa, 1988)

Language models

- Probabilistic view of language given a sample (corpus)
  - Probability or a word or a sentence in a corpus
  - Words: single (Zipf law), co-occurrence (Fitch)

- Language models: N-Grams and HMMs
  - Given a word history, predict the next word
  - Statistical grammars (discrete vocabularies)
  - Markov assumption for word sequences
  - Metrics of a LM on a corpus: Entropy, perplexity
  - Probabilistic model estimation

- Baker, Jelinek, 1974, 1976, 1980, 1992
- Automatic speech recognition (ASR/STT)
- Machine translation (MT)

- HMMs are symbolic ML models

Prediction and Entropy of Printed English

By C. E. SHANNON

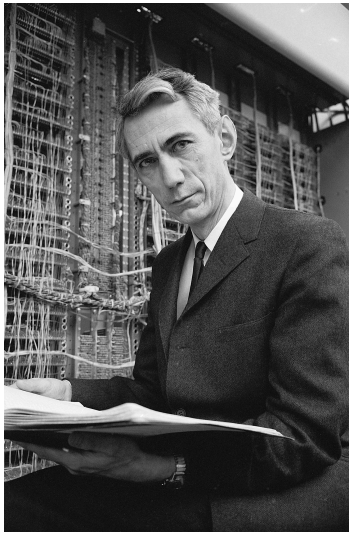
(*Manuscript Received Sept. 15, 1950*)

Entropy of English (Shannon, 1951)

- Corpus: History book
- Character language model: 26 characters
- Character entropy: 2.3 bits/character
- Character perplexity: 4.9 ($= 2^{2.3}$)

Entropy of English (WSJ)

- Corpus: 40M words (test set 1.5M words)
- Word language model: 20K words
- Entropy (3-gram): 6.8 bits/word
- Perplexity (3-gram): 109
- (Jurafsky and Martin, 2009)



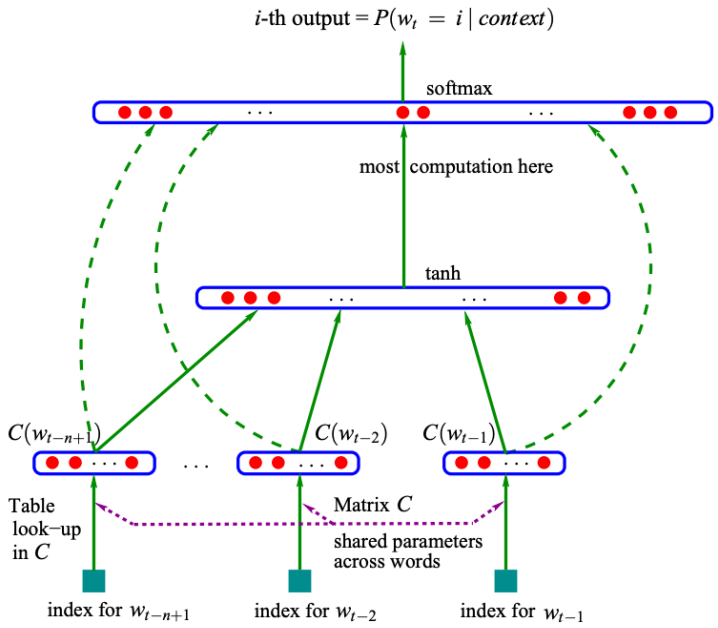
Claude E. Shannon

Neural language models

Bengio et al., 2003, A Neural Probabilistic Language Model, JMLR.
Multi-layer perceptron with one hidden layer, softmax output, residual connections

- N-gram model (N = 3) that jointly computes dense word embeddings
- No recurrence or attention mechanism. Train on Brown corpus, AP News.

> 1. associate with each word in the vocabulary a distributed *word feature vector* (a real-valued vector in $\mathbb{R}^m$),
> 2. express the joint *probability function* of word sequences in terms of the feature vectors of these words in the sequence, and
> 3. learn simultaneously the *word feature vectors* and the parameters of that *probability function*.



Neural network language model architectures: MLP, RNN, LSTM
Recurrence allows (in theory) to capture full sequence context
Mikolov (2010), Zaremba (2017) and Merity (2018)

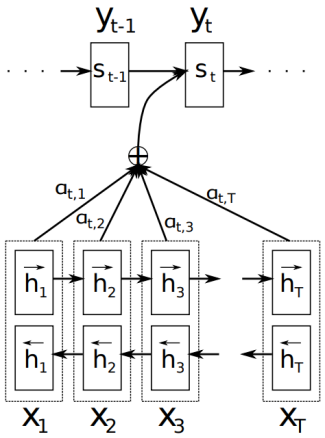
Newer models: Attention and transformers, since 2015 ("*Muppet*" models)

Neural language models: Recurrence vs. Attention

Mikolov *et al*., 2010
- Simple RNN LM

Zaremba, 2017; Merity, 2018
- RNN / LSTM LM

Bahdanau *et al*., 2015
- Source X1-T, targ y1-M
- RNN states h1-T
- Context vector C



Vaswani *et al*., 2017
- Encoder-Decoder
- Architecture: d-model, N, h, …
- Max input length
- Parameters: 213 million (large model)
- Training cost: 2.3×10^{19} FLOPS
- Training on eight V100: 3.5 days

- Task: SMT EN-DE, EN-FR
- Data: 4.5M / 36M sentence pairs
- Vocabulary: 37K / 32K tokens

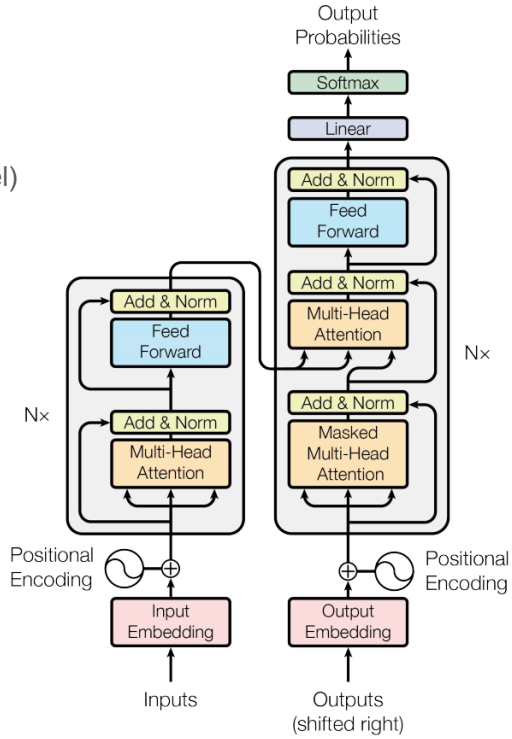
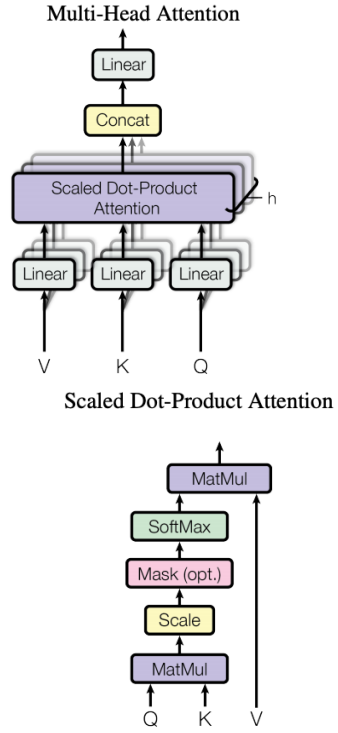


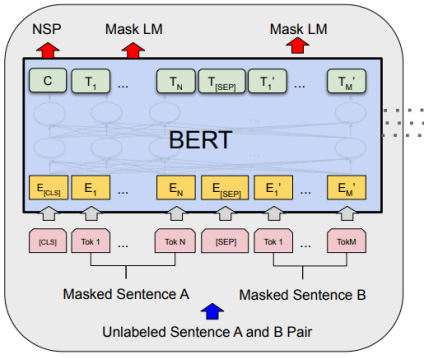
Figure 1: The Transformer - model architecture.



Attention and transformers: BERT and GPT

BERT: Devlin et al., 2018

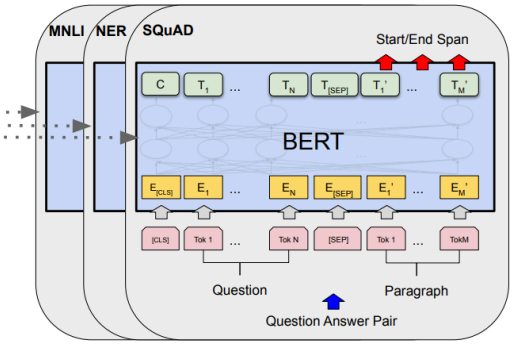
- Encoder of Transformer
- Universal encoder
- Masked LM; Next sentence
- Pre-train; Fine-tune
- BERT-large: 340M parameters



Pre-training

GPT - GPT-3: Radford, 2018 - 2020

- Decoder of Transformer
- Task: Predict next word
- Language prompt
- Zero-shot, Few-shot training
- GPT-2 large: 1.5B parameters
- GPT-3: 175B parameters

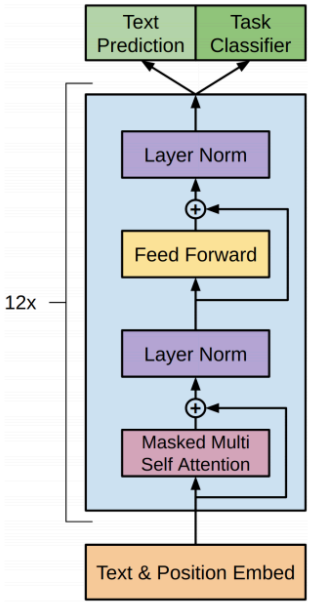


Fine-Tuning

GPT-2 (2019)

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

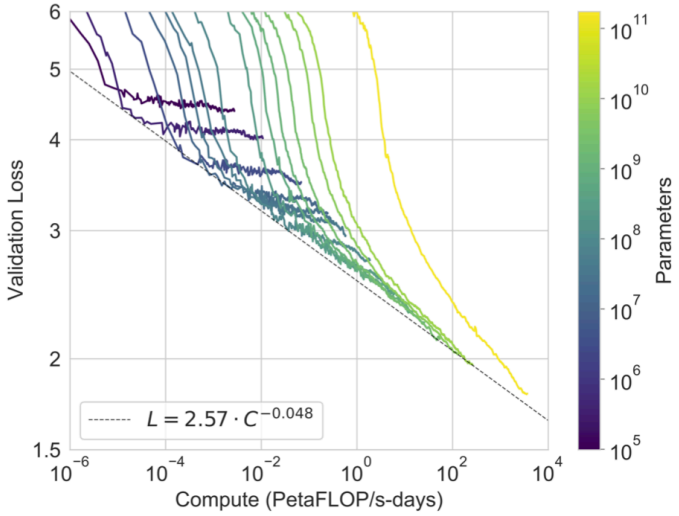
Table 2. Architecture hyperparameters for the 4 model sizes.



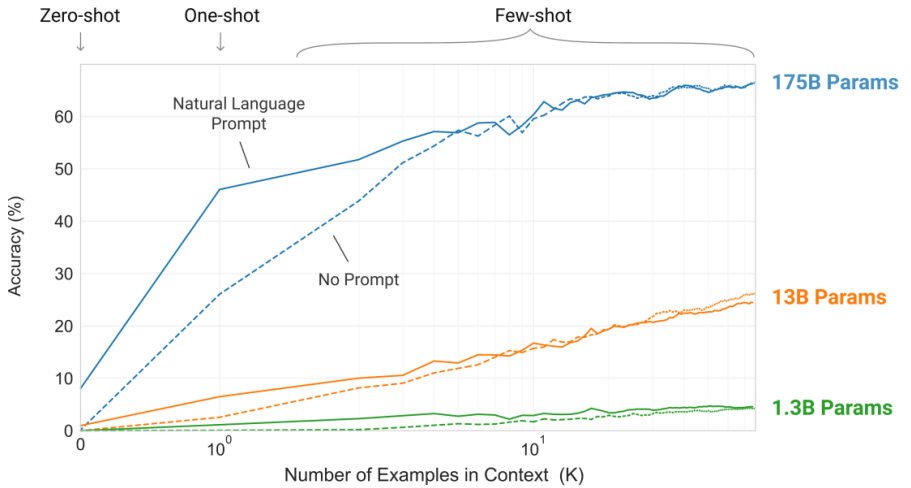
GPT (2018)

Attention and transformers: GPT-3

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%



Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128



Size of neural language models: From ELMo to GPT-3

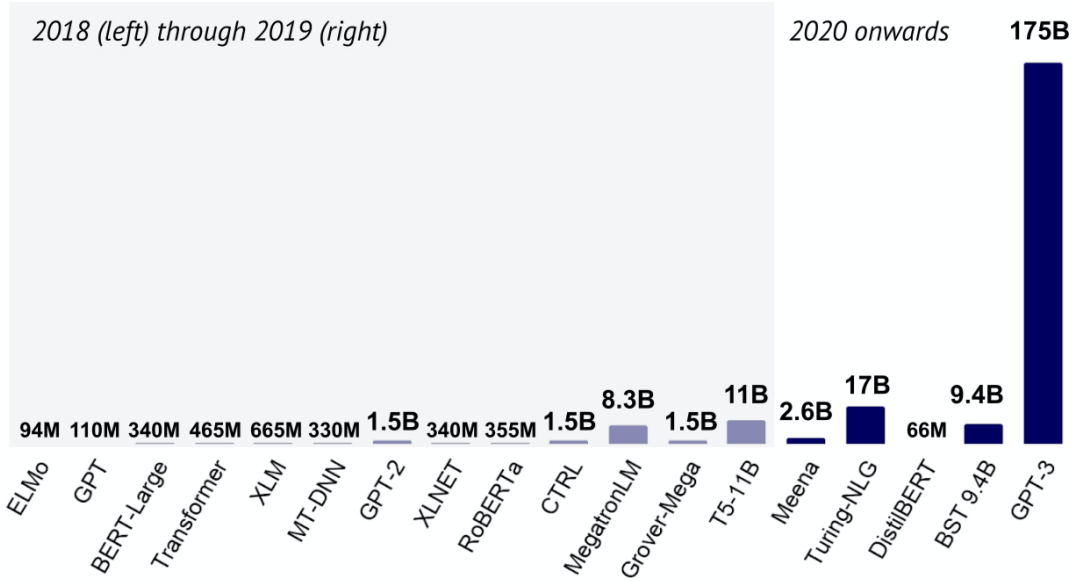
GPT-3 Model size
- 175B parameters, 1 TB memory
- Enough to store training data (300B tokens)
- Enough to store Wikipedia 300 times over

Compute 3.3×10^{23} FLOPS
- Training time on one Tesla V100: 355 years
- Cost: $4.5 to $10M USD to train

Are the model sizes justified?
- Yes, by GPT performance charts
- But, size can be reduced to 0.1% to 3% of parameters (alternate models; model distillation)
- Schick and Schütze, 2020; Adhikari et al., 2020

Data modeling and governance
- conflate knowledge of language, world knowledge and data facts into a “black box” representation
- provenance, consistency, completeness, bias, priority

Better-than-GPT3 performance with 0.1% the number of parameters
- PET: 223 Million parameters, 74.0 average SuperGLUE score.
- GPT3: 175 Billion parameters, 71.8 average SuperGLUE score.



source: stateof.ai

Language models: Welcome to the Billion Parameter club

Neuro-Symbolic Systems

- Architecture and data
  - data representation (words, categories, relations, productions/rules, time, facts)
  - distributed continuous vector representations
  - trainable neural-like substrates for all components (learning)
  - modular architecture, inductive biases: perception, memory, cognitive faculties, levels of representation, interfaces
  - (Extended/continuous) physical symbol systems, Newell and Simon, 1976, 10th ACM Turing Award

- Probabilistic logic and databases (PDBs). Graph Networks. Neural Graph Networks. Compositionality.
  - Anima Anandkumar, 2020, How to Create Generalizable AI, ACM TechTalks 08/11/2020.
  - Guy van den Broeck, 2019, IJCAI Computers and Thought award.
  - Peter Battaglia et al., 2018, Relational inductive biases, deep learning, and graph networks.

- Data modeling, data governance and model risk management
  - It is desirable for AI/ML/NLP models to separate (i) knowledge of language, (ii) world/data knowledge and (iii) data facts (language specification, data schema, data elements)
  - e.g., SQL BNF/semantics specification vs. data schema/architecture vs. data elements vs. data queries
  - Model risk: Interpretability and explainability are business requirements, especially in evolving regulated industries

Neuro-Symbolic Architecture

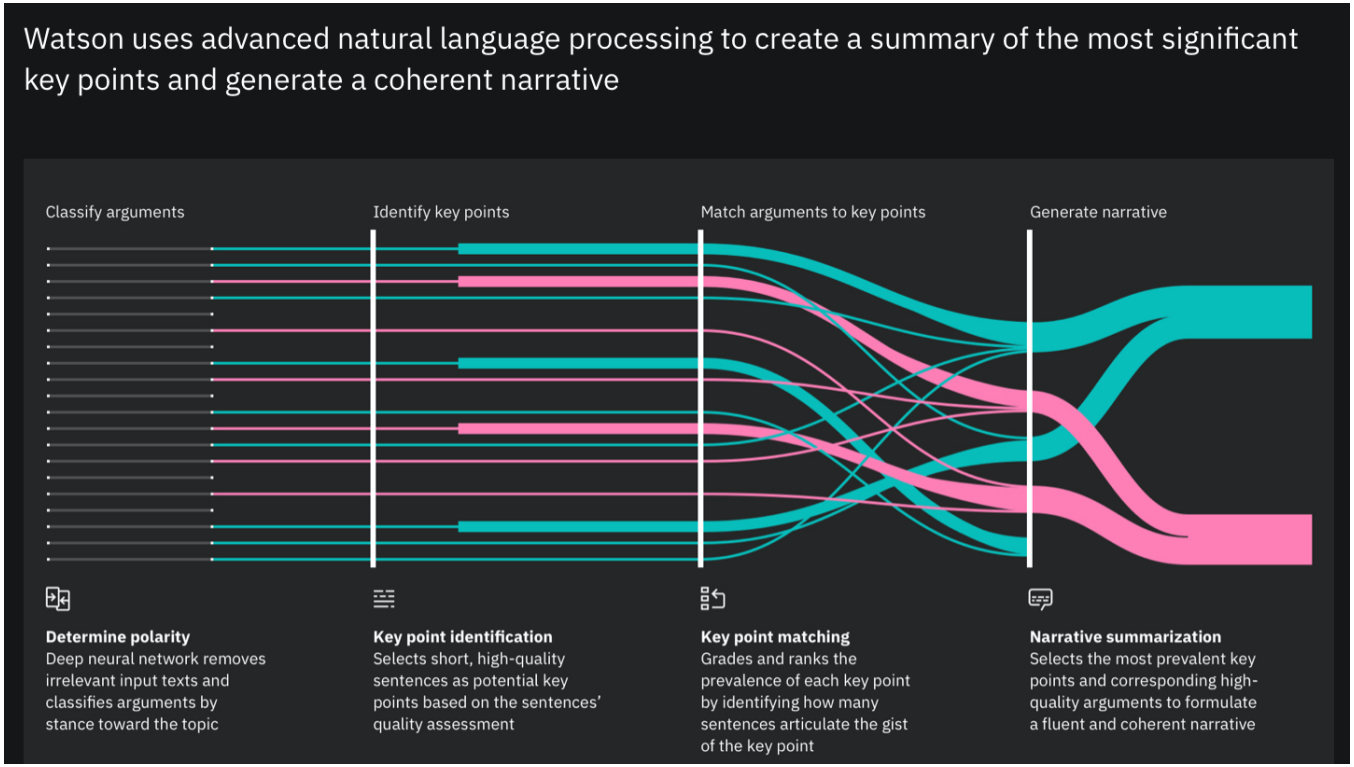
Hybrid AI systems

- IBM Watson Debater, 2020
- IBM Watson Jeopardy, 2011, was symbolic/statistical
- Defined system interface levels

Neuro-symbolic hybrid systems

- less training data
- track inference steps to draw conclusions
- interpretability, explainability

- [MIT IBM Watson Lab]



Conclusion

- Neural and symbolic approaches to AI and NLP
  - complementary strengths/weaknesses
  - hybrid models to refine the notion of a symbol system

- Model risk
  - From data: Incomplete data, inconsistent data, irrelevant data; bias, malicious data, etc.
  - From model: Interpretability/Explainability involve model and data.

- Societal implications: Future of work, model misuse, safety, ethics

- Ethics of AI: Stanford University HAI, MIT, IBM AI, Google AI Ethics as a Service, …
  - e.g., discussions at Standord HAI by O. Etzioni, 08/2020; C. Potts, 10/2020

- Opportunity: Model understanding, distillation, performance, accuracy, modularity

Thank you

Nelson Correa, Ph.D.
@nelscorrea

Slides:
- https://nelscorrea.github.io
- https://nelscorrea.github.io/fau2020/neurosymbolic

For useful discussions, thanks to:
- A. Correa, K.P. Unnikrishnan, R. Wesslen, W. Zadrozny

References (short name, web ref/arxiv.)

- Adhikari et al., 2019, Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT, https://www.aclweb.org/anthology/2020.repl4nlp-1.pdf
- Anima Anandkumar, How to Create Generalizable AI, ACM TechTalks 08/11/2020, https://learning.acm.org/techtalks/generalizable.
- Badhanau 2015
- Peter Battaglia et al., 2018, Relational inductive biases, deep learning, and graph networks. https://arxiv.org/abs/1806.01261
- Bender Kroll, ACL 2020
- Bengio 2003
- Bhattamishra et al., Limitations of Transformers to Recognize Formal Languages (https://arxiv.org/abs/2009.11264)Danqi CHen, ACL 2020
- Guy van den Broeck, 2019, IJCAI Computers and Thought award
- Chomsky
- Collobert, Weston, 2008
- Correa
- CMU KBMT'89
- Dehaene
- Devlin
- Ellman
- HAI
- Hinton, 1986
- HuggingFace, 2019
- Jelinek
- Kanerva, 2009
- Mikolov 2012, 2013
- Minervini, Scholar
- Newell and Simon, 1976, Computer science as empirical inquiry: symbols and search, Communications of the ACM, https://dl.acm.org/doi/10.1145/360018.360022
- Pantel, Turner, Vector semantics
- Pennington, GloVe
- Radford
- Salton 1972
- Shannon
- T. Schick, H. Schütze, 2020, It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners, https://arxiv.org/abs/2009.07118
- The Guardian, 2020,
- van de Broeck, Guy, IJCAI 2019
- Vasbani 2017
- WSJ 2020
- Zaremba 2017