

The logo features the letters 'NLP' in a large, 3D, light blue font with a glowing effect. An orange speech bubble containing the word 'MEETUP' in white capital letters is positioned over the 'P'. To the right of the 'P', the word 'VIENNA' is written in a smaller, white, sans-serif font. The background is dark blue with a faint network of white lines and dots.

# NLP MEETUP VIENNA

---

NATURAL LANGUAGE PROCESSING

---

# NLP Meetup #4

---

18.2.2020 @ Scible

# NLP Meetup #4



19:00 - NLP News & Trends - **Liad Magen**

19:30 - Emotion Recognition in Textual Conversations - **Philipp Möhl**

20:00 - Networking & Refreshments



Stefan Gindl



Dr. Andreas Rath



Jason

Hoelscher-Obermaier



Liad Magen

# NLP - News & Trends

---

Liad Magen

# News Topics

- Language Models
- Chatbots
- Information Retrieval/Extraction
- Applied AI Frameworks

# Language Models

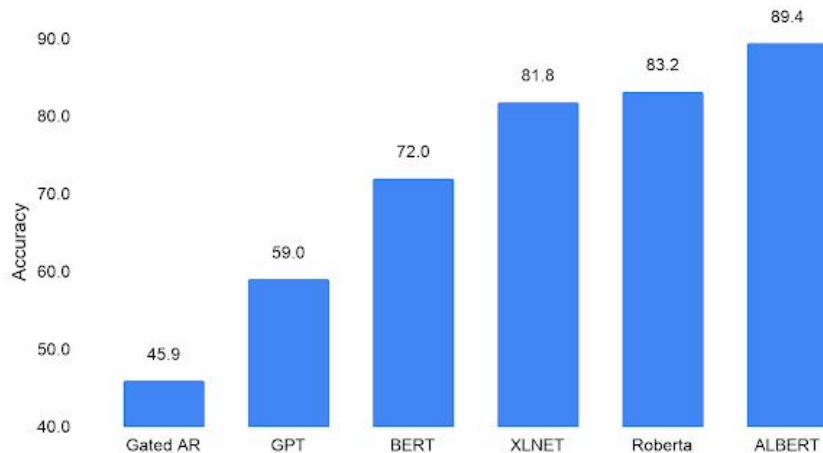
---

# BERTology



# BERT's children

- ALBERT
- XLNet
- RoBERTa
- DISTILBERT
- CamemBERT
- FlauBERT
- CTRL
- T5
- Reformer



\* Check out [TRAX](#)

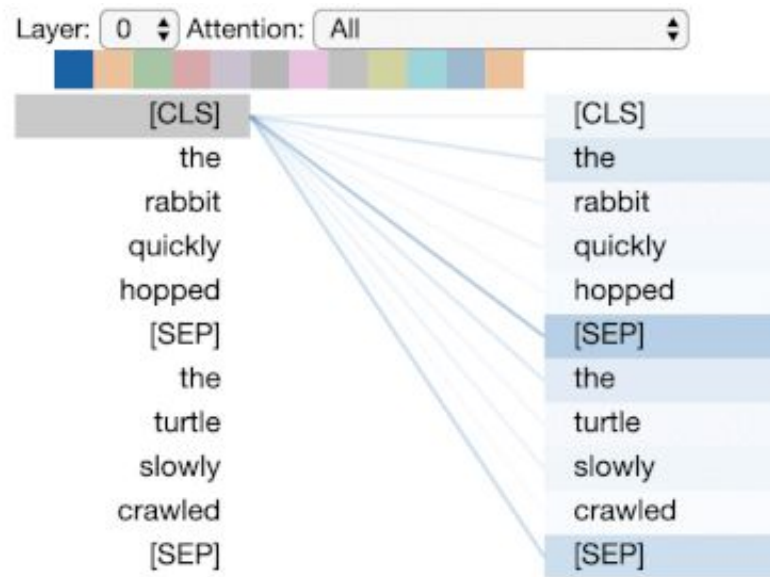


# BERT Analyzing tools

[jessevig/bertviz](#): Tool for visualizing attention in the Transformer model (BERT, GPT-2, Albert, XLNet, RoBERTa, CTRL, etc.)

[\[1906.04341\] What Does BERT Look At? An Analysis of BERT's Attention](#)

Partial explainability tool.



# BERT Research: Can BERT capture Linguistic structure?

[\[1905.05950\] BERT Rediscovered the Classical NLP Pipeline](#)

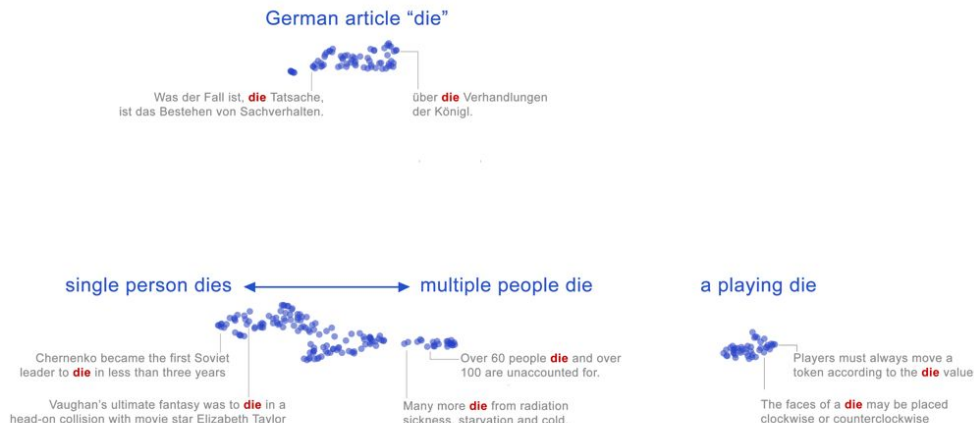
POS → parsing → NER → semantic roles → coref.

[\[1906.02715\] Visualizing and Measuring the Geometry of BERT](#)

“We find evidence of a fine-grained geometric representation of word senses”

[Revealing the Dark Secrets of BERT](#)



BERT is severely overparameterized;  
Not all attention heads are required

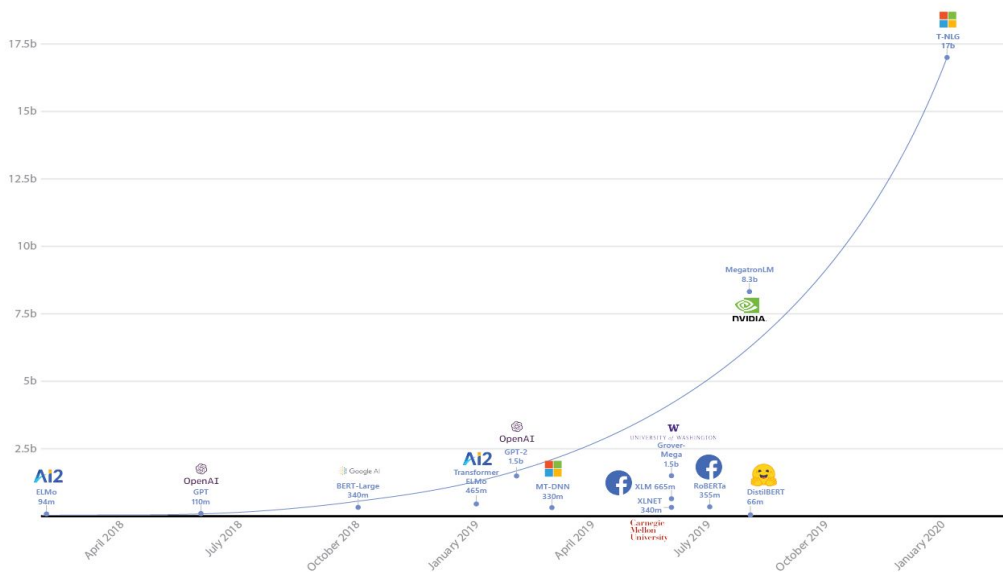
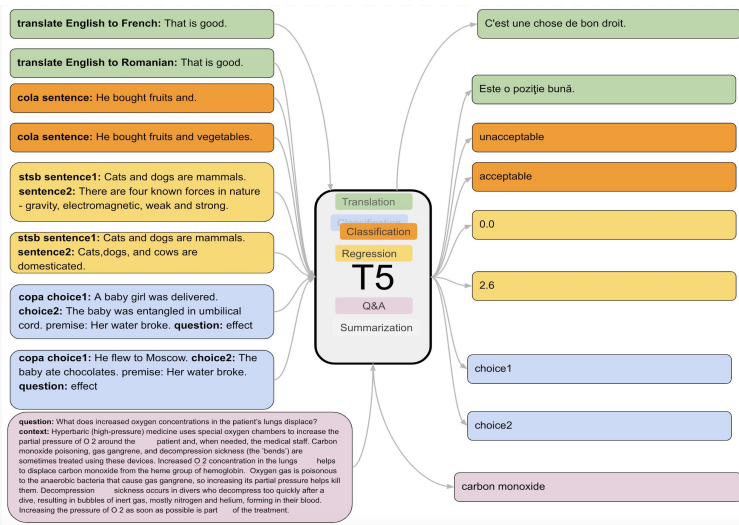


# BERTrend

1. BERT leads a major field breakthrough
  - a. **Transfer learning**
  - b. Task independency
  - c. (officially powering google search)
2. Active development
  - a. Academic research
  - b. Industry collaboration
3. Language dependent
  - a. Performs better on single language models
  - b. Independant releasing of LMs: French, Chinese, Finish...
4. Size doesn't always matters
  - a. DistillBERT / ALBERT

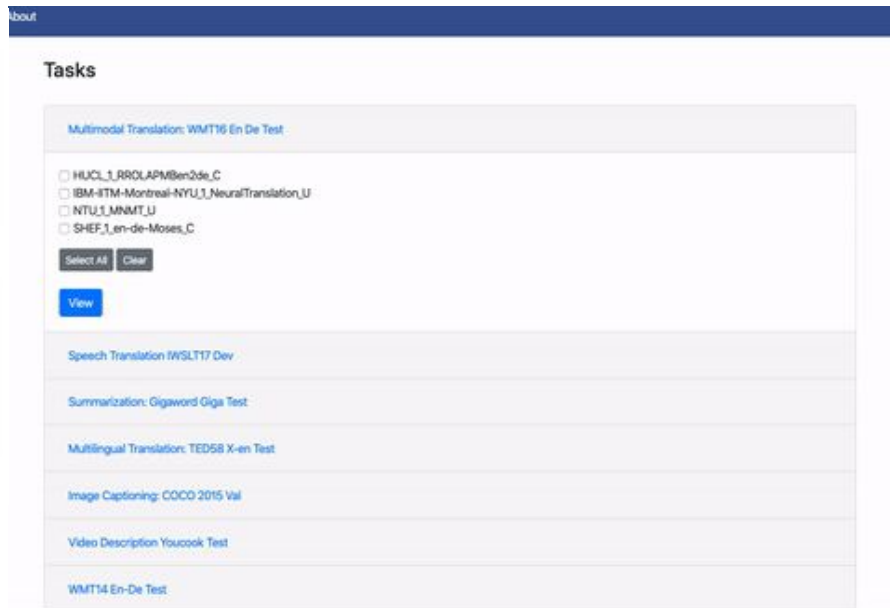
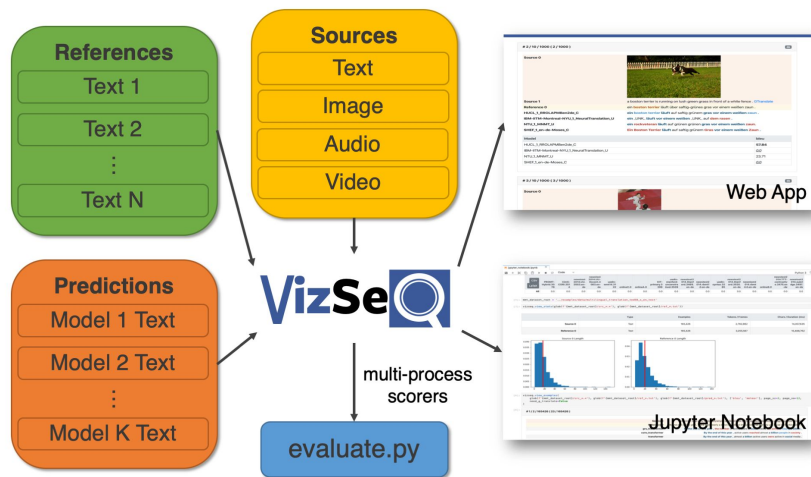
# "My LM is bigger than yours"

- “Text-to-Text Transfer Transformer” - T5 (11B) 
- T-NLG (17B) 



# Measuring LM for text generation

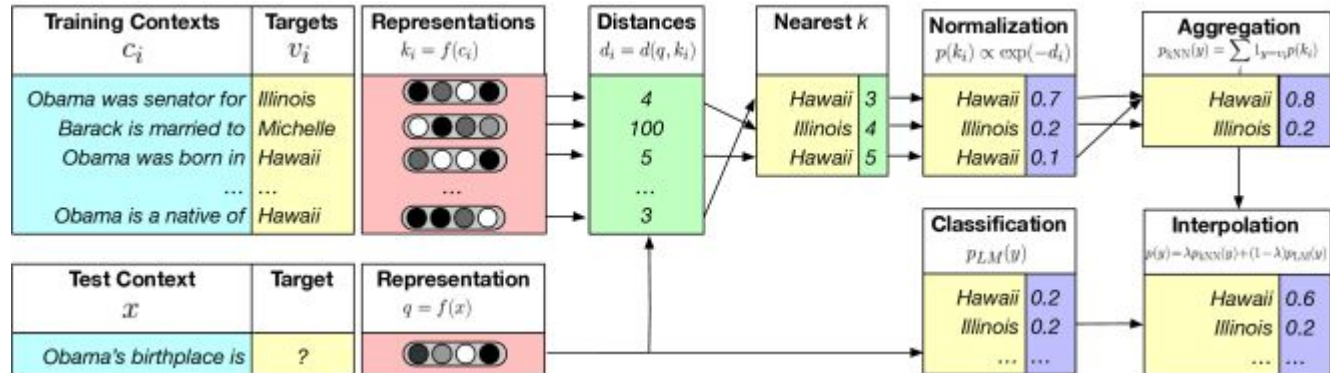
[facebookresearch/vizseq: A Research Toolkit for Natural Language Generation \(Translation, Captioning, Summarization, etc.\)](https://facebookresearch.github.io/vizseq/)



# New LM Approaches

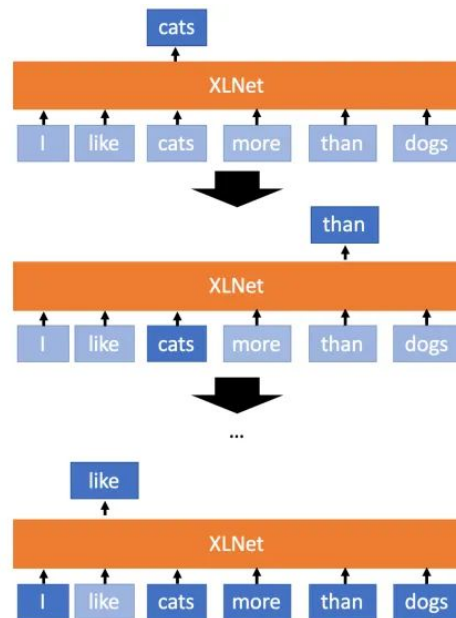
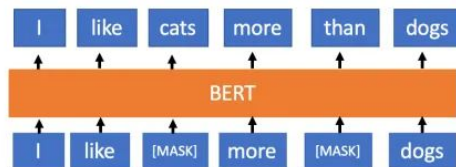
[1906.08237] **XLNet**: Generalized Autoregressive Pretraining for Language Understanding

[1911.00172] Generalization through Memorization: **Nearest Neighbor Language Models**



# XLNet vs BERT

XLNet randomly selects the next word to predicts:



# Information Extraction

---

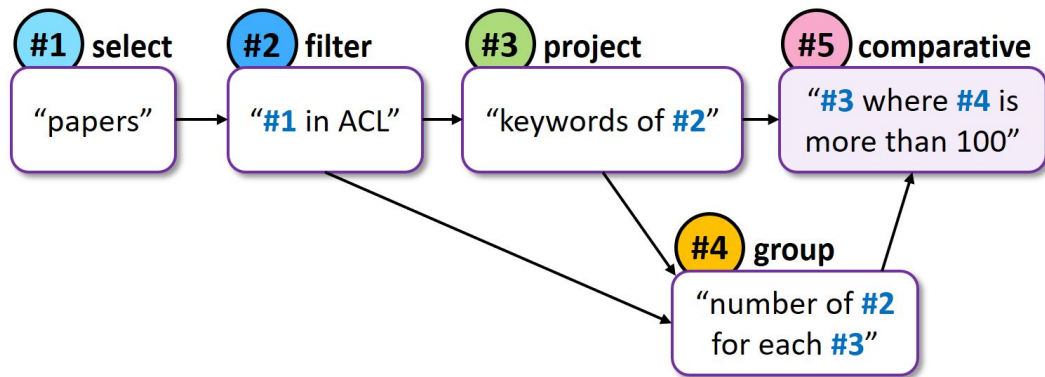


# BREAK Dataset

[A Question Understanding Benchmark | BREAK](#)

83K questions for training models to reason over complex questions

**Q: Which keywords have been contained by more than 100 ACL papers?**



[1910.02915] Commonsense Knowledge Base Completion  
with Structural and Semantic Context

## REALM: Retrieval-Augmented Language Model

### Pre-Training

# Ai2

Allen Institute for AI

## Commonsense Knowledge Base Completion with Structural and Semantic Context

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, Yejin Choi

UNIVERSITY OF WASHINGTON

Code: [github.com/allenai/commonsense-kg-completion](https://github.com/allenai/commonsense-kg-completion)

## Introduction

- KB Completion:** Induce missing edges between existing nodes in graph.
- Challenges** for KB completion of commonsense knowledge graphs (ConceptNet, ATOMIC): **Large** (# nodes is high = ~18x FB15K) and **highly sparse** (~0.01x graph density) due to semantic diversity.

## Approach

- Learning from Local Graph Structure:** Using weighted Graph Convolution Networks (GCNs).
  - Sparse graph structure: Enrich connectivity with **similarity-induced edges** ( $s^{sm}$ ) (i).
  - GCN Operation (ii):  $h_i^k = \tanh\left(\sum_{j \in N(i)} \alpha_{ij} g_j^k W_h^k + W_b^k b_i\right)$
- Learning from Text:** Using expressive representations from pre-trained language models (BERT).
  - Fine-tune BERT to node text and use as feature extractor (iii).
- Fusion & Decoding:**
  - Progressive masking of BERT representations (iv).
  - Convolutional Decoder to compute scores for candidate tuples (v).

## Experiments & Results

### 1. Effect of sparsity on KB completion performance

Note In-Degrees      Graph Density vs Scores

| In-Degrees Range | FB15K  | CN-100K | ATOMIC |
|------------------|--------|---------|--------|
| 10-100           | ~0.237 | ~0.237  | ~0.237 |
| 100-1000         | ~0.237 | ~0.237  | ~0.237 |
| 1000-10000       | ~0.237 | ~0.237  | ~0.237 |
| 10000-100000     | ~0.237 | ~0.237  | ~0.237 |
| 100000-1000000   | ~0.237 | ~0.237  | ~0.237 |
| 1000000-10000000 | ~0.237 | ~0.237  | ~0.237 |

| Density    | FB15K  | CN-100K | ATOMIC |
|------------|--------|---------|--------|
| 0.001      | ~0.237 | ~0.237  | ~0.237 |
| 0.01       | ~0.237 | ~0.237  | ~0.237 |
| 0.1        | ~0.237 | ~0.237  | ~0.237 |
| 1.0        | ~0.237 | ~0.237  | ~0.237 |
| 10.0       | ~0.237 | ~0.237  | ~0.237 |
| 100.0      | ~0.237 | ~0.237  | ~0.237 |
| 1000.0     | ~0.237 | ~0.237  | ~0.237 |
| 10000.0    | ~0.237 | ~0.237  | ~0.237 |
| 100000.0   | ~0.237 | ~0.237  | ~0.237 |
| 1000000.0  | ~0.237 | ~0.237  | ~0.237 |
| 10000000.0 | ~0.237 | ~0.237  | ~0.237 |

- Performance drops observed when knowledge graphs (result: FB15K-237) become sparser.

### 2. Results (MRR) on CN-100K and ATOMIC

| Dataset         | DistMult | BERT + ConvTransE | Sim + GCN + ConvTransE | Sim + GCN + BERT + ConvTransE | ConvTransE | GCN + ConvTransE | GCN + BERT + ConvTransE |
|-----------------|----------|-------------------|------------------------|-------------------------------|------------|------------------|-------------------------|
| ConceptNet-100K | 8.97     | 18.68             | 49.50                  | 29.8                          | 29.01      | 29.01            | 29.01                   |
| ATOMIC          | 12.39    | 12.04             | 50.38                  | 51.31                         | 12.39      | 12.04            | 12.04                   |

- Subgraph sampling:** Crucial for scaling GCN operation with large graphs.
- Human Evaluation** (% of valid tuples among top 10 target entities) shows similar trends as automatic metrics.

## Main Findings

- BERT** is proficient at capturing taxonomic relations and hence provides **significant boosts for ConceptNet** (~9–35 points) but is not as effective for ATOMIC.
- Graph embeddings provide rich local context**, boosting performance with subgraph training.
- Similarity-induced edges boost learning of graph embeddings** resulting in improved performance.

# Chatbots

---

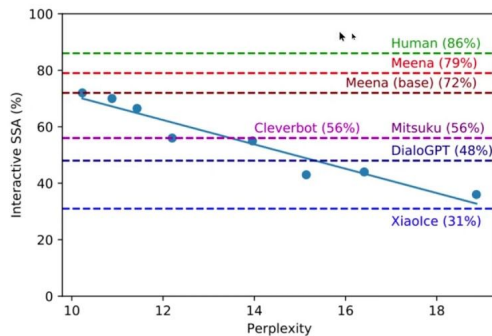
# Meet Meena



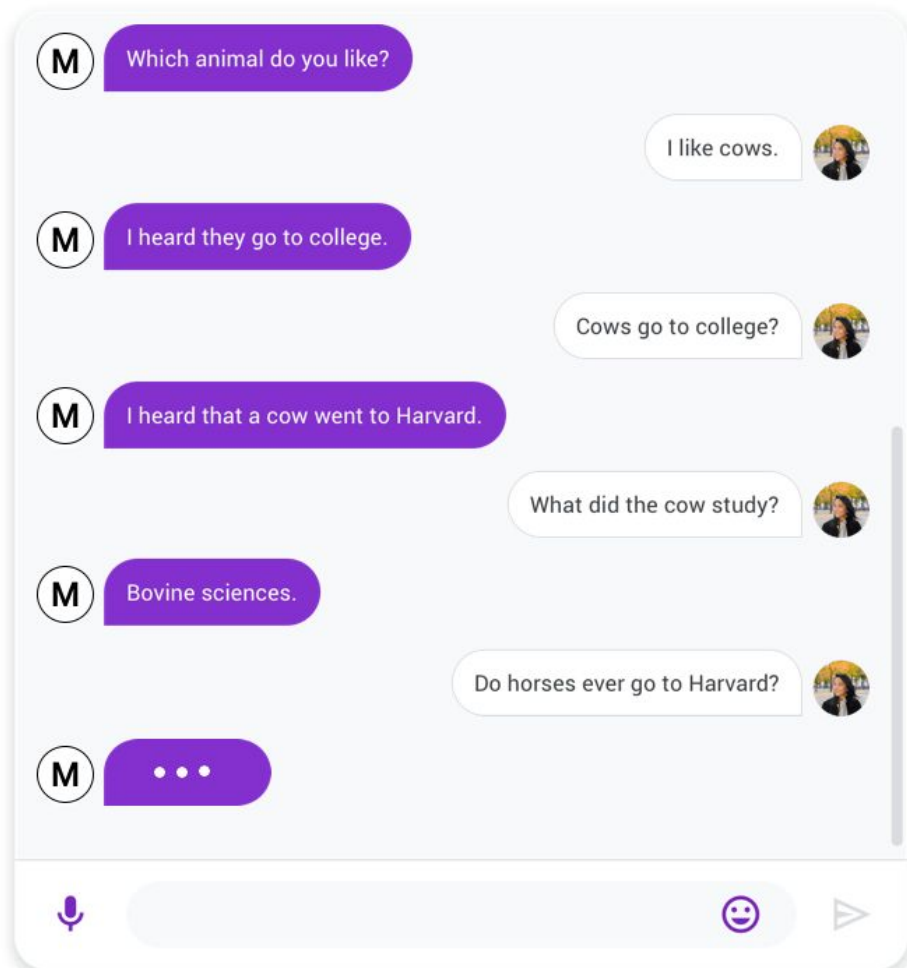
[2001.09977] Towards a Human-like Open-Domain Chatbot

- 2.6B params - GPT-2 x 2
- 40B words from social media conversations
- SSA  $\Rightarrow$  Sensibleness and Specificity Average  
Does the response make sense?  
Is it specific to the content?

## SSA vs Perplexity



4



# Applied AI

---

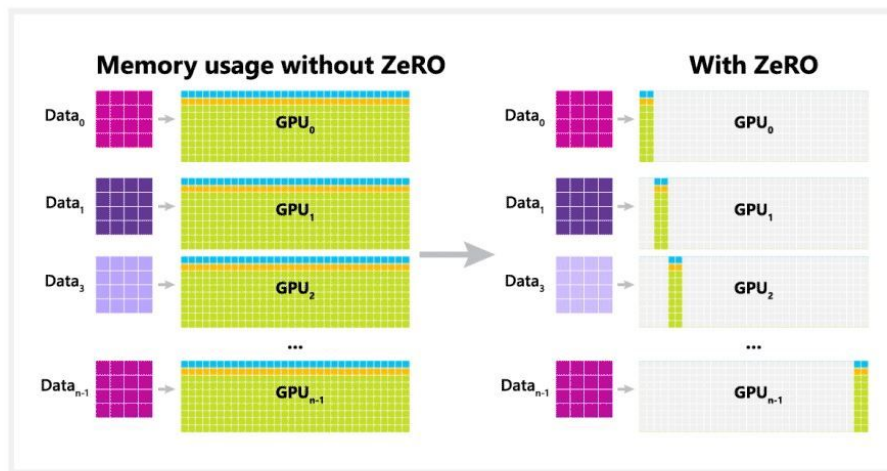
# Deep Speed

DeepSpeed 

x4 - x8 memory  
reduction

32GB GPUs

## DeepSpeed + ZeRO



### Scale

- 100B parameter
- 10X bigger

### Speed

- Up to 5X faster

### Cost

- Up to 5X cheaper

### Usability

- Minimal code change

# THiNC

- Framework agnostic (PyTorch, TF...)
- Configuration system
- High order functions
- Typed variables



Thank you!

---



# Bibliography

<https://www.scihive.org/paper/1911.00172>

<https://thegradient.pub/gpt2-and-the-nature-of-intelligence/>

<https://allenai.github.io/Break/>

<https://medium.com/dair-ai/nlp-newsletter-reformer-deepmath-electra-tinybert-for-search-vizseq-open-sourcing-ml-68d5b6eed057https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html>