Natural Language Processing 1

Large Language Models (LLMs)

Katia Shutova

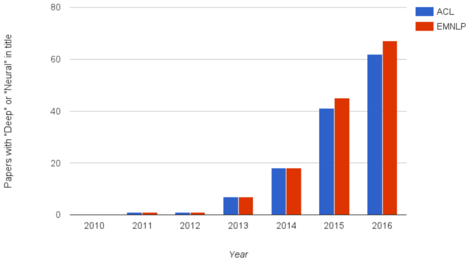
ILLC
University of Amsterdam

Outline.

The advent of representation learning

Deep learning

- dominates the field since ≈2014
- led to performance improvements in many tasks



The advent of representation learning

But why?

- Neural networks have been around for decades.
- What has changed in the way they are applied in NLP?
- Key conceptual innovation:

  *learning **intermediate meaning representations** in the process of end-to-end training for a particular task.*

The advent of representation learning

But why?

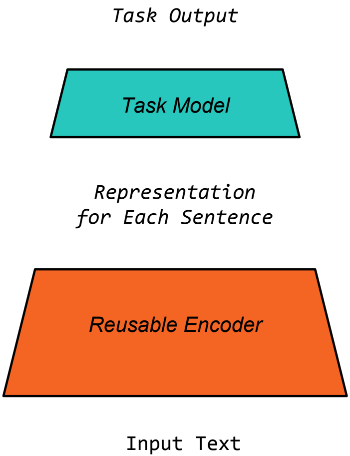
- Neural networks have been around for decades.
- What has changed in the way they are applied in NLP?
- Key conceptual innovation:

  learning ***intermediate meaning representations*** *in the process of end-to-end training for a particular task.*

General-purpose models

Paradigm shift:

- instead of training task-specific models
- train a **general-purpose** neural network **sentence encoder**
- which can be applied across diverse NLP tasks.



Why is this useful?

1. Improve performance
   - produce **rich semantic representations** for downstream NLP tasks
2. Improve data efficiency
   - provide a model of sentence representation for language understanding tasks which **lack training data**

What can we expect this model to capture?

- Lexical semantics and meaning disambiguation in context
- Word order
- Some syntactic structure
- Semantic composition
- Idiomatic/non-compositional phrase meanings
- Connotation and social meaning.

ELMo: Embeddings from Language Models

Peters et al. 2018. *Deep contextualized word representations*

- Pretrain a biLSTM model in the language modelling task
- Model context in both directions, produce contextualised word representations
- Use them as input to a task-specific model.

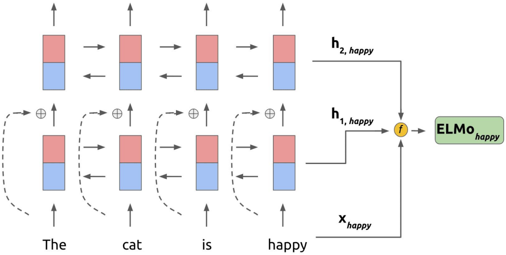


Image credit: Victor Zuanazzi

The ELMo model

Pretraining:

- The encoder is a 2 layer BiLSTM
- The model is trained with the language modelling objective
- jointly maximize log likelihood of the forward and backward directions.

Application:

- ELMo word representations: weighted sum of hidden representations at all layers
- Weights are learned in a given task.

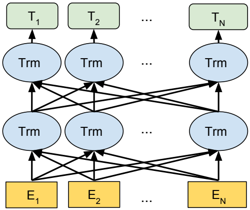
The contributions of ELMo

- **Contextualised word representations** provide a level of disambiguation
- **Deep** representations allow to capture linguistic information at various levels (syntax – lower layers; semantics – higher layers)
- (Large) performance improvements in many NLP tasks
- **Paradigm shift** towards sentence encoder pretraining
- Started the rich history of naming LMs based on Sesame Street characters.

The rise of the Transformer

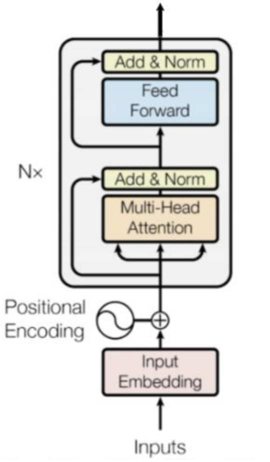
Devlin et al. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

- Transformer architecture
- Bidirectional context representation
- Two pretraining tasks: masked language modelling (MLM) and next sentence prediction (NSP)
- Pretrain the encoder and then fine-tune it for a specific task.



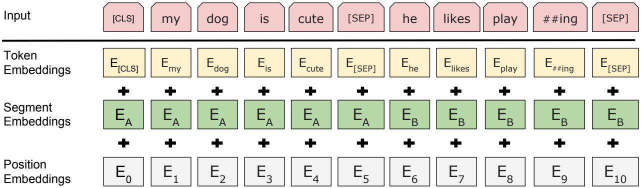
BERT: Architecture

- Stacked Transformer blocks (multi-head attention followed by feed-forward neural network)
- BASE model: 12 Transformer layers, 12 attention heads (110M params)
- LARGE model: 24 Transformer layers, 16 attention heads (340M parameters)



BERT: Input representations

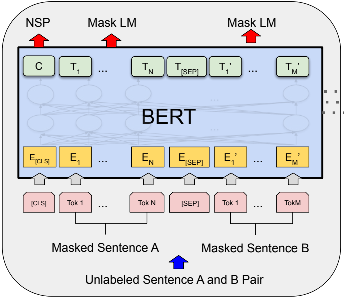
- Introduce special [CLS] and [SEP] tokens
- The [CLS] token represents the whole input sequence
- The [SEP] token indicates a boundary between two segments
- **Input representations** are a sum of token embeddings + position embeddings + segment embeddings.



BERT: Pretraining tasks

Masked language modelling

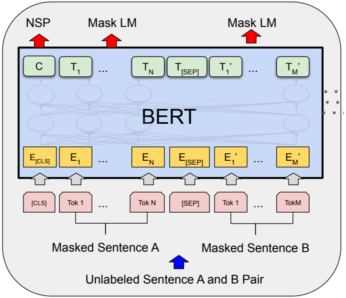
- standard conditional language models only model context in one direction at a time
- BERT performs bidirectional encoding by masking 15% of the input tokens
- Inspired by the cloze task



BERT: Pretraining tasks

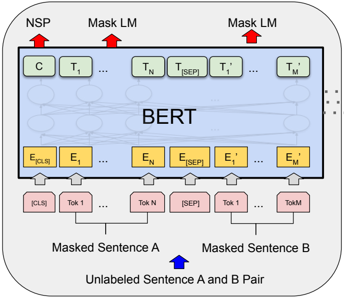
Next sentence prediction

- Randomly sample sentence pairs, such that 50% of the time the sentences follow each other.
- Predict whether the second sentence follows the first or not.
- This models the relations between sentences (useful for many tasks, e.g. QA)

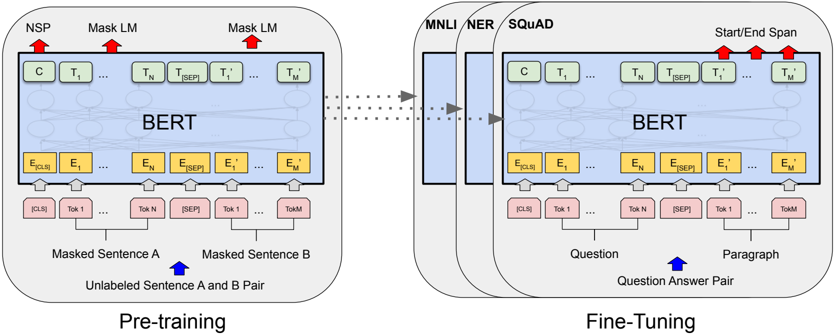


BERT: pretraining

- **Pre-training loss**: the sum of the mean MLM likelihood and the mean NSP likelihood
- **Data**: BooksCorpus (800M words) and English Wikipedia (2500M words)



BERT: fine-tuning



The contributions of BERT

- **Advanced the state-of-the-art** in a range of NLP tasks
- Demonstrated the importance of bidirectional pretraining
- Reduced the need for task-specific architectures
- Most widely-used NLP model (118K+ citations)
- Traditional linguistic hierarchy emerges within layers of BERT (Tenney et al. 2019)
- lower layers – syntax; higher layers – semantics and discourse.

Tenney et al. 2019. *BERT Rediscovers the Classical NLP Pipeline*

Outline.

Generative language models: The GPT family

Radford et al. 2019. *Language Models are Unsupervised Multitask Learners*

GPT, GPT2, GPT3, GPT4 etc.

- Left-to-right language model
- Generative model, i.e. able to generate text (unlike BERT)
- Transformer architecture (GPT comparable in size to BERT BASE)
- Interesting intuition: multitask learning from natural language instructions.

More than a language model?

- Many tasks are already described in the data in some way
- Can language models learn to perform tasks from natural language instructions found in web text?

> If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".
>
> If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
>
> "**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

InstructGPT and ChatGPT

InstructGPT

- trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT

- optimized for dialogue
- make GPT generations more "conversational": can provide more natural answers, ask follow-up questions etc.

An example from ChatGPT

You

What is the Dutch city famous for its canals?

ChatGPT

The Dutch city famous for its canals is Amsterdam. Amsterdam is known for its extensive canal system, which includes numerous picturesque waterways that crisscross the city. These canals are a defining feature of Amsterdam's landscape and have earned it the nickname "Venice of the North." The canals are not only a beautiful sightseeing attraction but also serve as an important part of the city's transportation infrastructure. They are lined with historic buildings, houseboats, and bridges, making Amsterdam a popular destination for tourists from around the world.

Reinforcement learning from human feedback

What makes a ”good” text?

This is task and context dependent:

- if writing a story we may want creativity,
- when providing information it has to be truthful.
- we want the model outputs to be safe, unbiased and polite
- this goes beyond the LM objective of predicting next words

How to evaluate and balance these factors?

RLHF: Humans provide feedback on model outputs, and the model is then adjusted to better reflect human preferences.

Reinforcement learning from human feedback

Key components:

1. Pretrained language model
2. gathering data and training a reward model
3. fine-tuning the LM with reinforcement learning.

Training a reward model

- Collect a set of prompts
- Generate continuations for these prompts using multiple LMs
- Humans rank these continuations, obtaining a scalar score for each
- This score numerically represents human preferences
- Train a reward model to predict this score.

Training a reward model

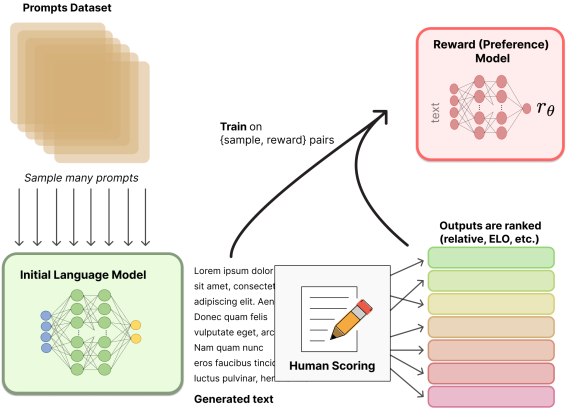


Image credit: Huggingface

Fine-tuning with reinforcement learning

Fine-tune the LM to better **match human preferences**

At each iteration:

- Given the prompt x, the LM generates continuation y
- Concatenate x and y, and pass as input to the reward model
- Reward model outputs a reward score r_θ
- Fine-tune the LM to maximize the reward score for the current batch of data
- Regularisation to ensure the per token probability distributions don't change too much (from original LM)

Outline.

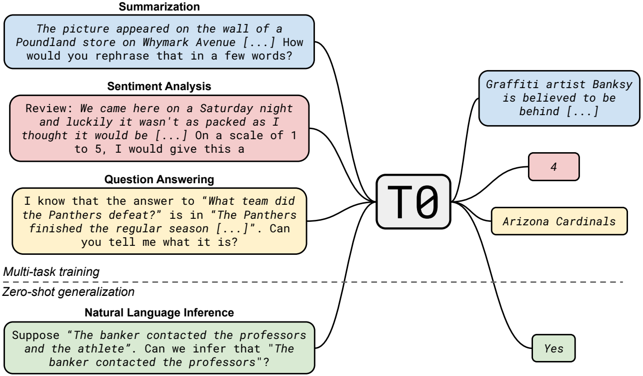
From task-specific to general-purpose models

Generative language models

LLMs in research and practice

Instruction-tuned LLMs and multi-task learning

Sanh et al., 2022. *Multitask Prompted Training Enables Zero-Shot Task Generalization*. ICLR 2022.



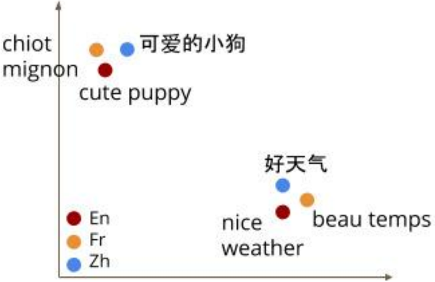
Multilingual LLMs

Goal: a single model that captures universal language structures and can reason across all known languages

- E.g. nearly all languages make a distinction between nouns and verbs and distinguish function words from content words.
- find such commonalities between languages
  - on the lexical, syntactic and semantic level
- and exploit them
- balance *language-agnostic* and *language-specific* information

Multilingual LLMs: Intuition

Phrases with similar meaning should obtain similar representations, irrespective of the language



Multilingual LLMs: Models

Up to 110 languages encoded within one model

Architectures:

- 5 layer LSTM: LASER model
- Transformer (mid-size): Multilingual BERT, XLM, XLM-R
- Transformer (large): BLOOM, XGLM, mT5 etc.

Pretraining tasks:

- Monolingual: Masked language modelling or generative LM
- Cross-lingual: Sentence translation or translation language modelling (TLM)

Multilingual LLMs: Application

1. Pretrain a multilingual LLM
2. Fine-tune the LLM on one or more high-resource languages to obtain a task model
3. Perform zero-shot or few-shot transfer to other (low-resource) languages

How successful is such transfer?

High performance for typologically-close languages, much lower performance for typologically-different and low-resource languages.

Can LLMs solve NLP?

They are an exciting step forward, offering many opportunities

- They generate fluent text
- Practically useful in many contexts
- They provide a unified framework for solving many tasks
- They can learn in-context: few-shot learning from a task instruction and a small number of demonstration examples
- They can explain their own decisions (sometimes)

But! *See next slide...*

Can LLMs solve NLP?

But many challenges remain (and some new ones emerged)

- Can LLMs generalise systematically?

  robin is a bird; robins are flowers

- Factual errors and inference errors
- Hallucinations: making up content
- Memorisation vs. generalisation?
- Data contamination: performance evaluations in many tasks are misleading
- Many issues related to safety and bias

Outstanding challenges and future directions

- Interpretability
- Better learning algorithms, e.g. continual learning
- Low-resource languages
- Few-shot learning and generalisation
- Common sense reasoning
- Ethics and alignment

We will discuss these topics in advanced NLP courses, such as *NLP2* and *Advanced Topics on Computational Semantics* (block 5)