Transfer learning

Vlad Shakhuro

MISIS
UNIVERSITY

21 November 2025

Outline

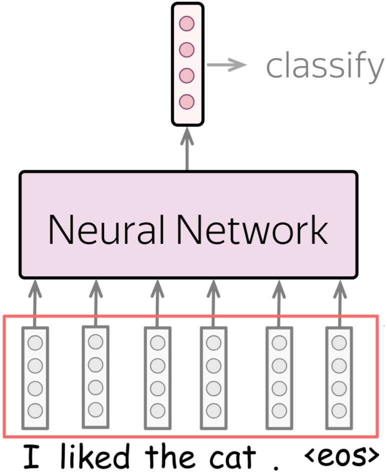
1. Recap: word embeddings

2. Idea 1: Words — Words in context

3. Idea 2: Task-specific — Unified

4. BERT analysis

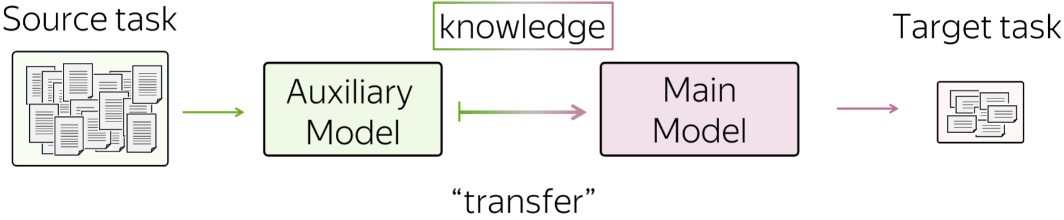
Word embeddings



Input word embeddings:

- train from scratch
- take pretrained (word2vec, GloVe)
- initialize with pretrained, then finetune

Transfer learning idea



Source task

knowledge

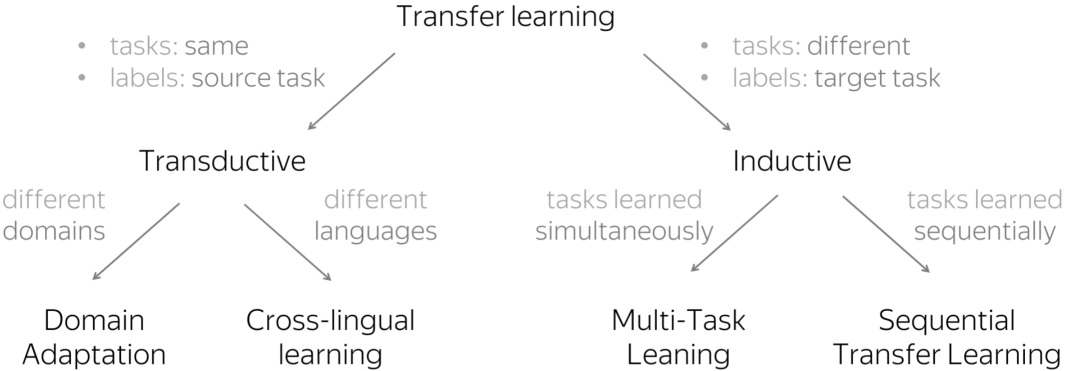
Target task

Auxiliary Model

Main Model

“transfer”

A taxonomy of transfer learning in NLP



Transfer learning

- tasks: same
- labels: source task

- tasks: different
- labels: target task

Transductive

Inductive

different domains

different languages

tasks learned simultaneously

tasks learned sequentially

Domain Adaptation

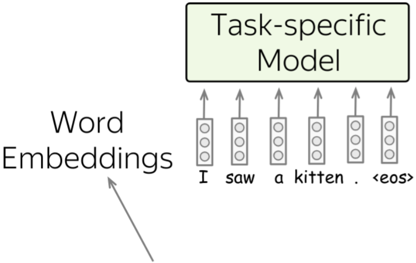
Cross-lingual learning

Multi-Task Leaning

Sequential Transfer Learning

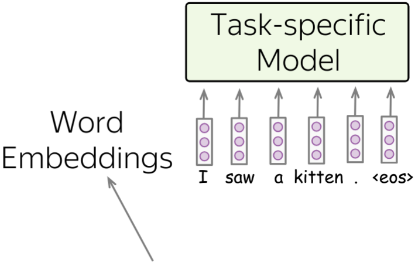
Transfer through word embeddings

Before



Know nothing, trained from scratch with the model

After



Word2Vec, GloVe, ... - saw a lot of data, know a lot

Outline

1. Recap: word embeddings

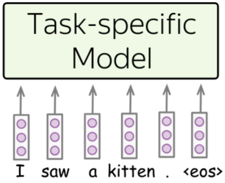
2. Idea 1: Words — Words in context

3. Idea 2: Task-specific — Unified

4. BERT analysis

From words to words in context

Before

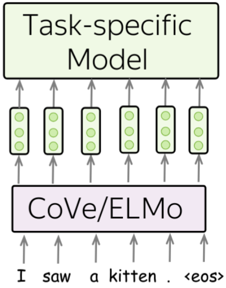


Word Embeddings

Word2Vec, GloVe, ... - saw a lot of data, know a lot about words

Model is (almost) the same

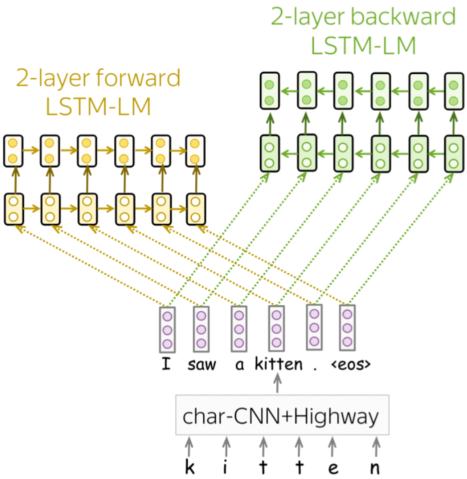
After



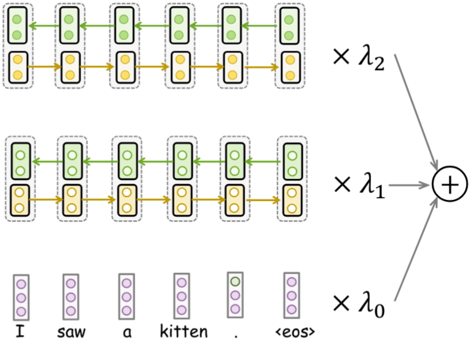
Representations of words **in context**

Know not only about word in general, but how it is used in each particular sentence

ELMo: From words to words in context



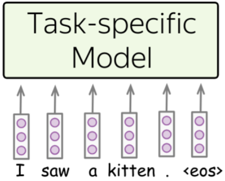
Learn specific $\lambda_0, \lambda_1, \lambda_2$ for each task



ELMo usage

Before

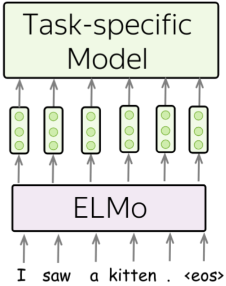
Word Embeddings



Word2Vec, GloVe, ... - saw a lot of data, know a lot

After

Representations of words **in context**



We used huge unlabeled dataset to learn not only words, but **how to process entire texts**

Outline

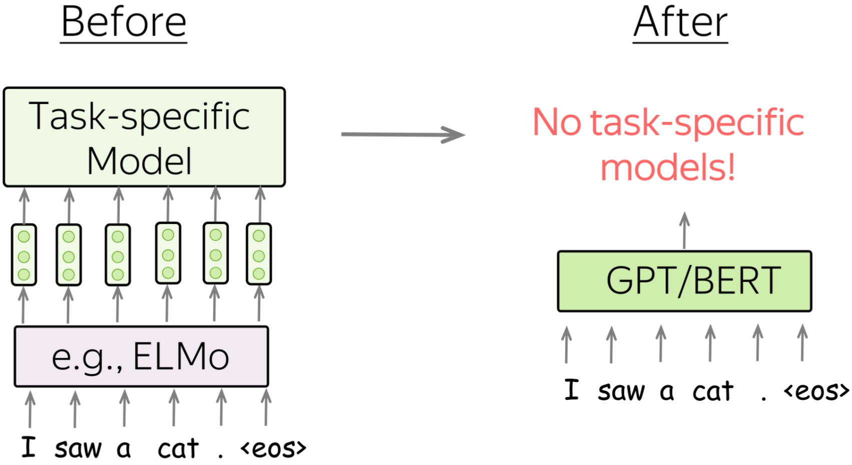
1. Recap: word embeddings

2. Idea 1: Words — Words in context

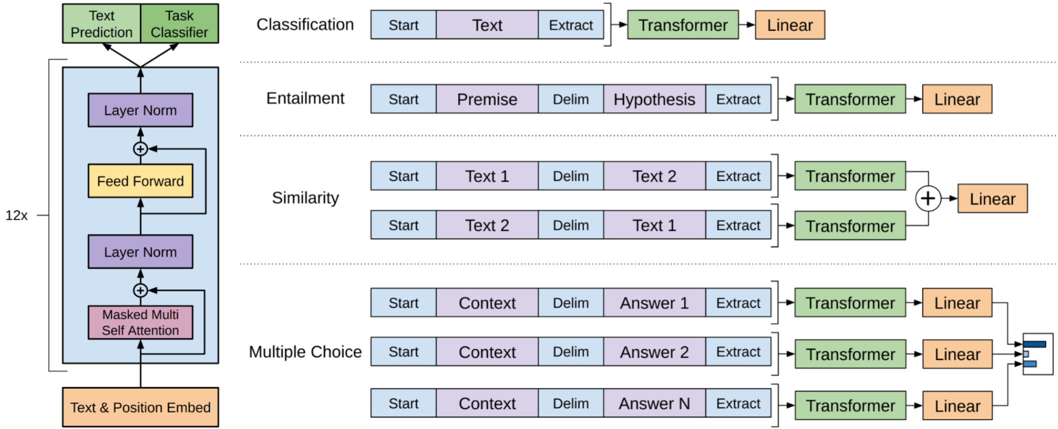
3. Idea 2: Task-specific — Unified

4. BERT analysis

From task-specific models to unified

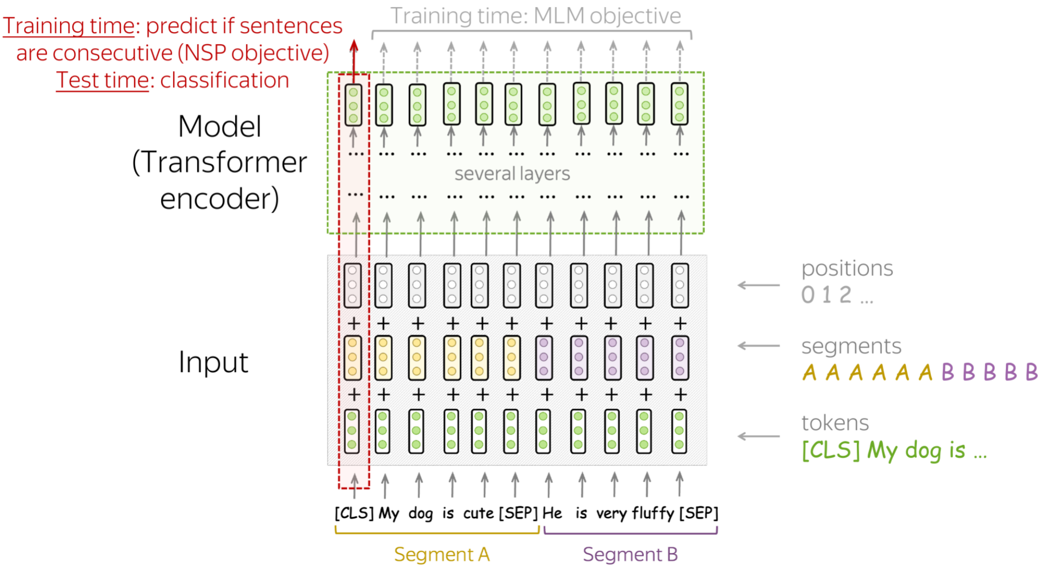


GPT-1 arch and training



Fine-tuning loss: $L = L_{xent} + \lambda L_{task}$

BERT: Transformer encoder with novel training

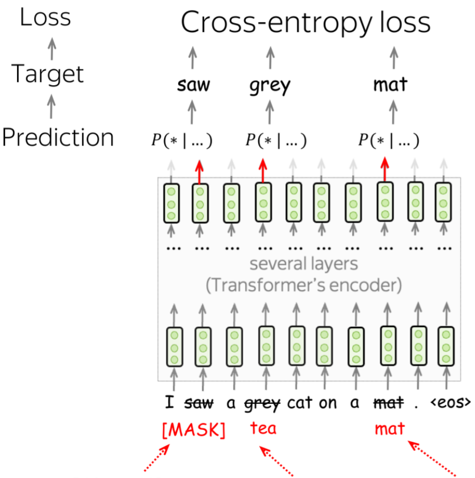


Training time: predict if sentences are consecutive (NSP objective)
Test time: classification

Model (Transformer encoder)

Input

BERT: Masked language modelling



- [MASK], with p = 80%
- Random token, with p = 10%
- Original token, with p = 10%

At each training step:

- pick randomly 15% of tokens
- replace each of the chosen tokens with something
- predict original chosen tokens

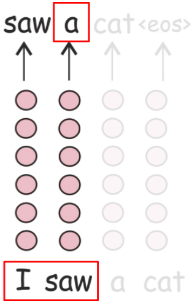
Training objectives: LM vs MLM

Language Modeling

Target: next token

Prediction: $P(* \mid \text{I saw})$

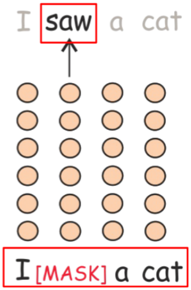
left-to-right, does not see future



Masked Language Modeling

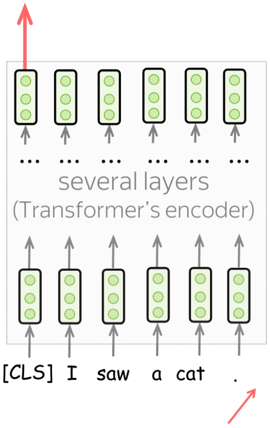
Target: current token (the true one)

Prediction: $P(* \mid \text{I [MASK] a cat})$



sees the whole text, but something is corrupted

Finetuning BERT: classification



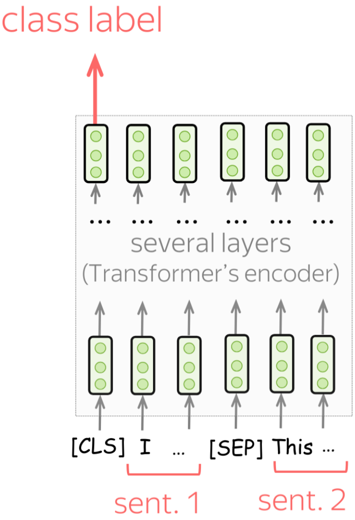
class label

several layers (Transformer's encoder)

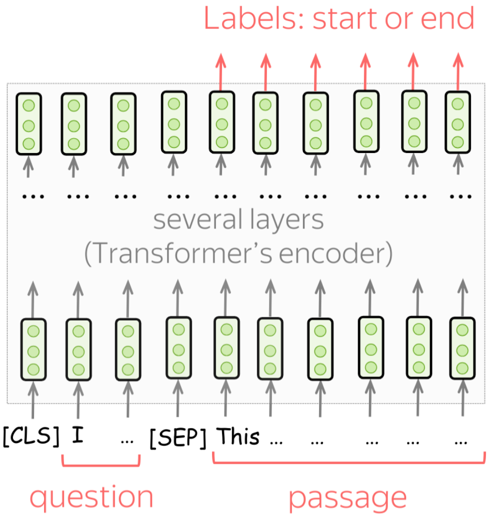
[CLS] I saw a cat .

No second sentence!

Finetuning BERT: sentence pair classification



Finetuning BERT: question answering



Outline

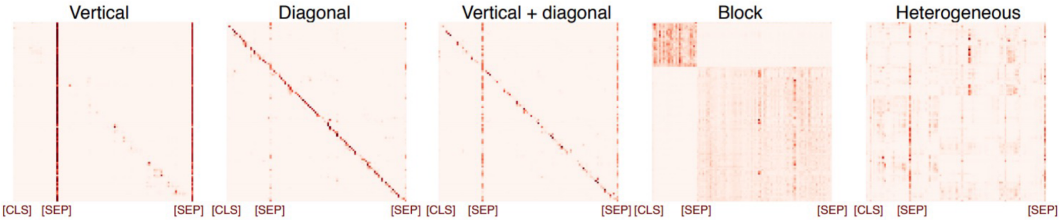
1. Recap: word embeddings

2. Idea 1: Words — Words in context

3. Idea 2: Task-specific — Unified

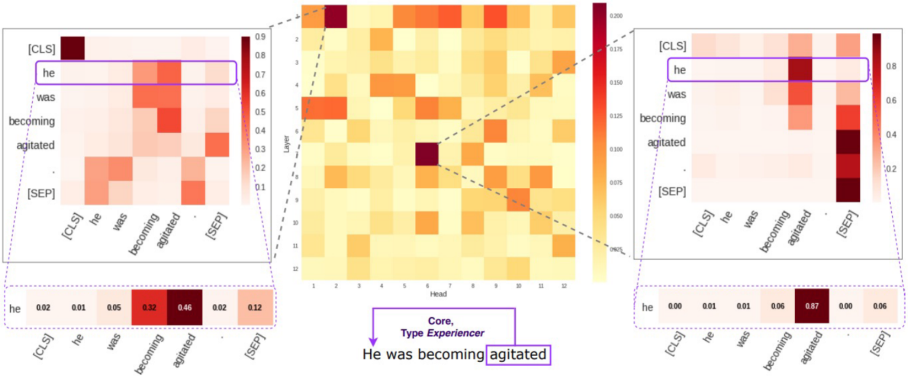
4. BERT analysis

Self-attention heads



Typical self-attention patterns

Self-attention heads

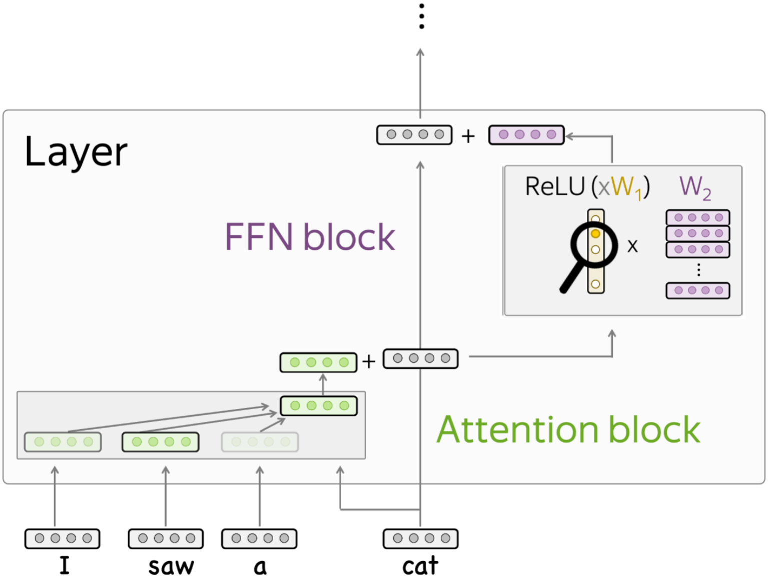


Core, Type *Experiencer*

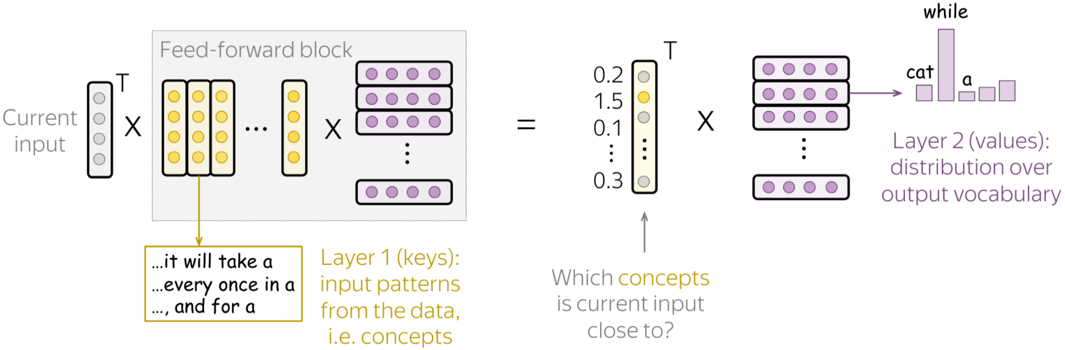
He was becoming agitated

Heads that encode information correlated to semantic links in the input text

FFNs as key-value memories



FFNs as key-value memories



...it will take a
...every once in a
..., and for a

Layer 1 (keys): input patterns from the data, i.e. concepts

Which concepts is current input close to?

Layer 2 (values): distribution over output vocabulary

BERT rediscovers the classical NLP pipeline

Part of speech: I want to find more , [something] bigger or deeper . → NN (Noun)

Constituents: I want to find more , [something bigger or deeper] . → NP (Noun Phrase)

Dependencies: [I]$_1$ am not [sure]$_2$ how reliable that is , though . → nsubj (nominal subject)

Entities: The most fascinating is the maze known as [Wind Cave] . → LOC

Semantic Role Labeling: I want to [find]$_1$ [more , something bigger or deeper]$_2$. → Agr1 (Agent)

Coreference: So [the followers]$_1$ waited to say anything about what [they]$_2$ saw . → True

BERT rediscovers the classical NLP pipeline

Part of speech:

Constituents:

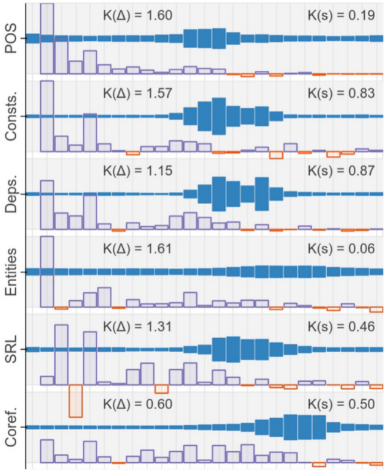
Dependencies:

Entities:

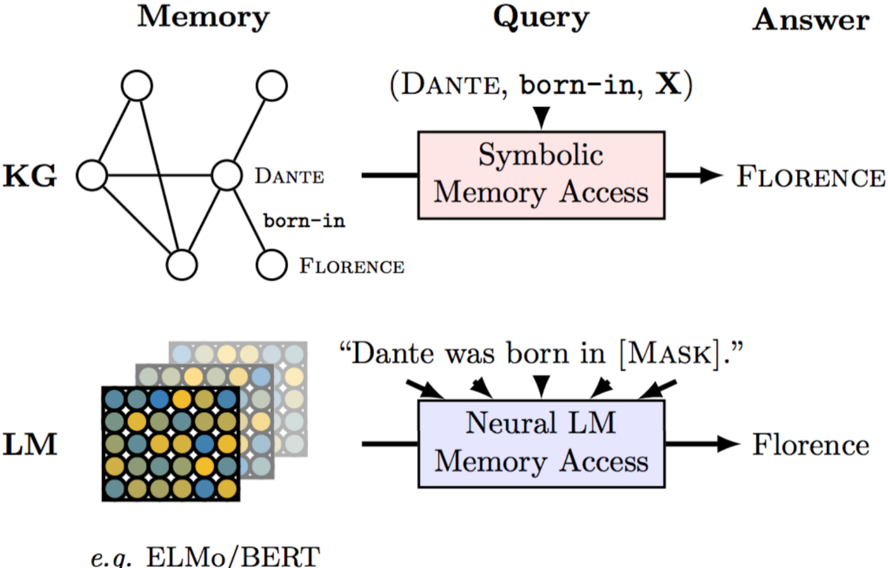
Semantic Role Labeling:

Coreference:

In classical NLP, to solve a subsequent task is was required to solve the previous one



Language models as knowledge bases



Memory

Query

Answer

KG

(Dante, born-in, X)

Symbolic Memory Access

Florence

LM

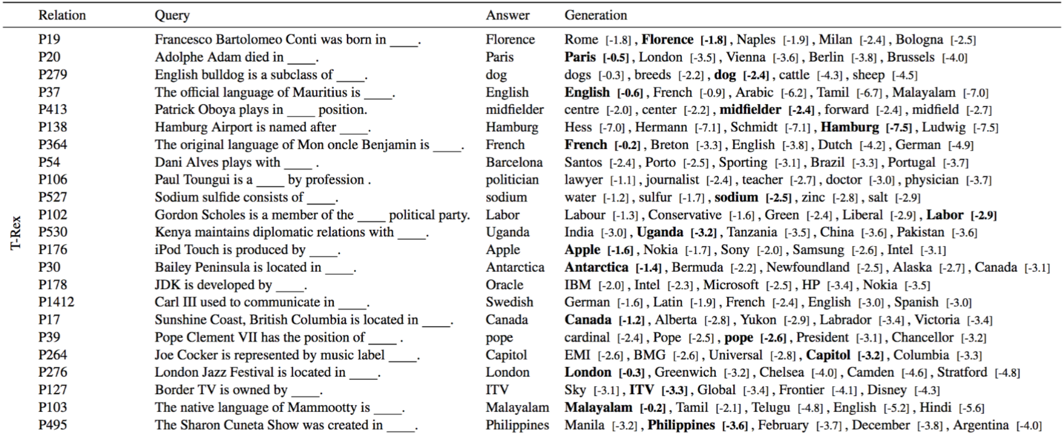
"Dante was born in [Mask]."

Neural LM Memory Access

Florence

e.g. ELMo/BERT

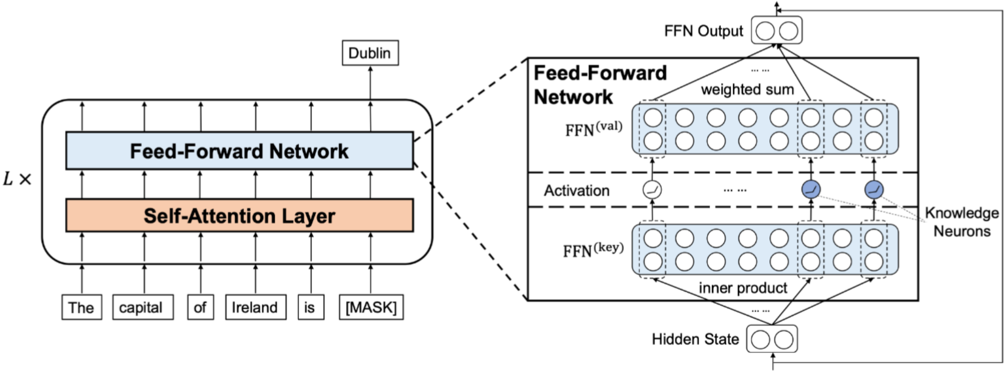
Language models as knowledge bases



	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0]

Knowledge neurons in FFNs

Data: (subject, relation) -> object triplets



Conclusion

We reviewed following topics:

- main idea of transfer learning
- shift from word embeddings to words-in-context embeddings (ELMo)
- from from task-specific to unified models (GPT-1, BERT)
- analysis of BERT model