Knowledge Neurons in Pretrained Transformers

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, Furu Wei

1. MOE Key Lab of Computational Linguistics, Peking University,

2. Microsoft Research

- Introduce the concept of **knowledge neurons** and propose a **knowledge attribution method** to identify the knowledge neurons that express specific factual knowledge in the fill in-the-blank cloze task.

- Conduct both qualitative and quantitative analysis to show that knowledge neurons are positively correlated to knowledge expression.

- Present preliminary studies of leveraging knowledge neurons to edit factual knowledge.

Related Work

- **Probing Knowledge in Pretrained Models**
- To retrieve knowledge in pretrained models (such as BERT) using cloze queries.
- Mining-based and paraphrasing-based methods.
- Closedbook question answering to measure how much knowledge a pretrained model has stored in its parameters.
- Measure and improve the consistency of pretrained models with respect to factual knowledge prediction.

- **Attribution Methods**
- Aim to attribute the model output to input features using different measures
- Integrated gradient method

Introduction

- Present preliminary studies on how factual knowledge is stored in pretrained Transformers by introducing the concept of knowledge neurons.
- Examine the fill-in-the-blank cloze task for BERT.
- Propose a knowledge attribution method, based on integrated gradients, to identify knowledge neurons.
- Have investigated that the activation of such knowledge neurons is positively correlated to the expression of their corresponding facts.
- Attempt to leverage knowledge neurons to edit (such as update, and erase) specific factual knowledge without fine-tuning.
- Results shed light on understanding the storage of knowledge within pretrained Transformers.

- View feed-forward network (i.e., two-layer perceptron) modules in Transformer as key-value memories.

- The hidden state is fed into the first linear layer and activates knowledge neurons; then, the second linear layer integrates the corresponding memory vectors.

Analysis shows that- activation of the identified knowledge neurons is positively correlated to the knowledge expression (3 ways)

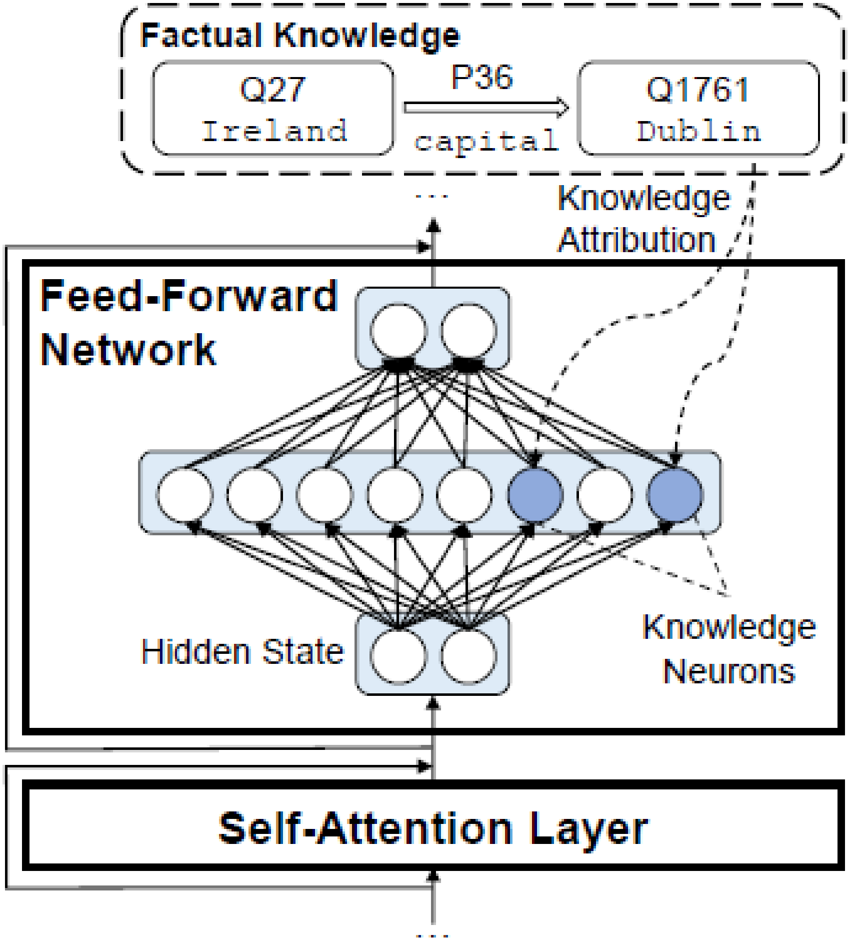


Figure 1: Through knowledge attribution, we identify knowledge neurons that express a relational fact.

Transformer Block

- Self-Attention Layer
- Feed Forward Layer

Input matrix : $X \in \mathbb{R}^{n \times d}$

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V,$$
$$\text{Self-Att}_h(X) = \text{softmax}\left(Q_h K_h^T\right) V_h,$$
$$\text{FFN}(H) = \text{gelu}\left(HW_1\right) W_2,$$

Parameter matrices:

$W_h^Q, W_h^K, W_h^V, W_1, W_2$

Hidden state: H

- Connections Between Self-Attention and FFN
  - Self-Att(.) similar to FFN(.)
  - **query-key-value** vs **query vector**(**input to FFN**), **keys & values**(**two linear layers of FFN**)

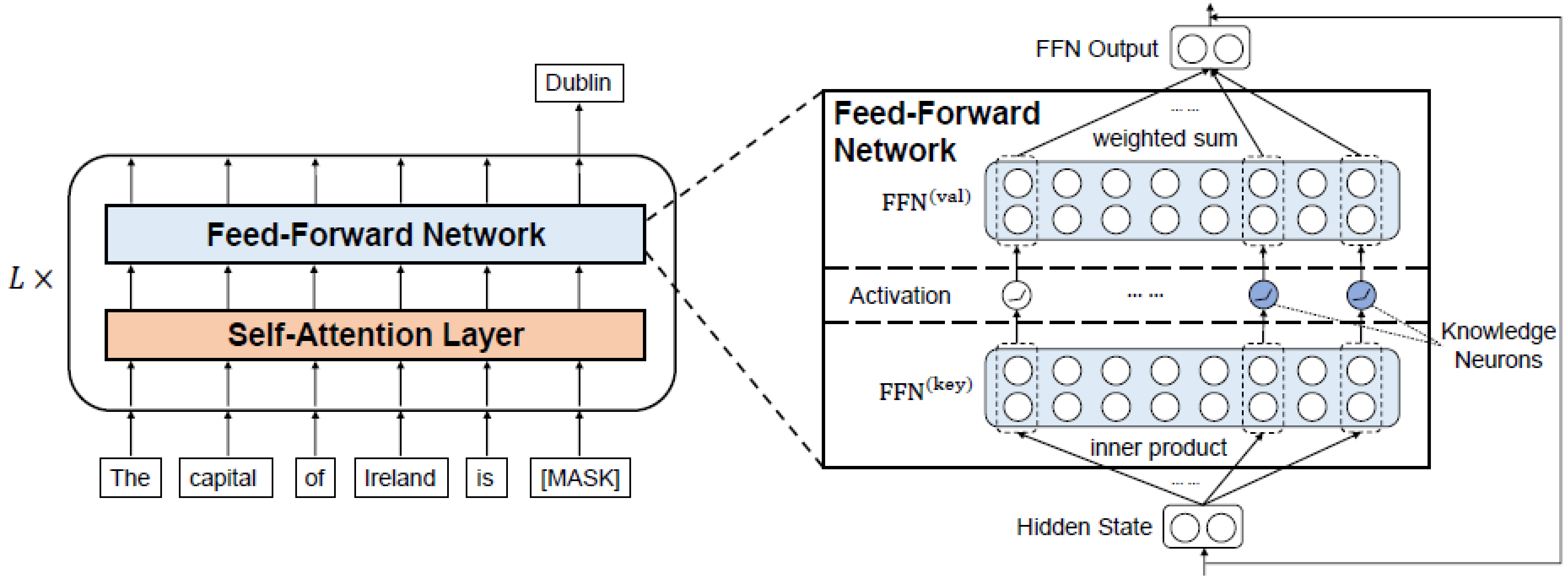


Figure 2: Illustration of how an FFN module in a Transformer block works as a key-value memory

- Hypothesize that factual knowledge is stored in FFN memories and expressed by knowledge neurons.

Example – Knowledge Accessing Task

- Fact : <Ireland, capital, Dublin>
  - Triplet <h,r,t>
  - h - head entity, t – tail entity, r – relation between head & tail

- Possible query : “The capital of Ireland is _____”

(knowledge-expressing prompt)

Knowledge Attribution

- Propose a **knowledge attribution method** based on integrated gradients.
- Can evaluate the contribution of each neuron to knowledge predictions.

- Complete process to identify knowledge neurons:

(1) produce n diverse prompts

(2) for each prompt, calculate the knowledge attribution scores of neurons;

(3) for each prompt, retain the neurons with attribution scores greater than the attribution threshold t, obtaining the coarse set of knowledge neurons;

(4) considering all the coarse sets together, retain the knowledge neurons shared by more than prompts p%

Process of Knowledge Attribution

- 1) Given an input prompt *x*, Calculate probability of the correct answer predicted by a pretrained model

$$\mathrm{P}_x(\hat{w}_i^{(l)}) = p(y^*|x, w_i^{(l)} = \hat{w}_i^{(l)}), \qquad (4)$$

*y** - correct answer

$w_i^{(l)}$ - *i*-th intermediate neuron in the *l*-th FFN

$\hat{w}_i^{(l)}$ - a constant

- 2) Calculate the attribution score
  - Change $w_i^{(l)}$ from 0 to its original value $\overline{w}_i^{(l)}$ calculated by the pretrained model
  - Integrate the gradients

$$\mathrm{Attr}(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\alpha=0}^{1} \boxed{\frac{\partial\, \mathrm{P}_x(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}}} d\alpha, \qquad (5)$$

Calculates gradient of the model output wrt $w_i^{(l)}$

- If the neuron has a great influence on the expression of a fact, the gradient will be salient, which in turn has large integration values.

problem

- Riemann approximation

$$\tilde{\mathrm{Attr}}(w_i^{(l)}) = \frac{\overline{w}_i^{(l)}}{m} \sum\nolimits_{k=1}^{m} \frac{\partial \, \mathrm{P}_x(\frac{k}{m}\overline{w}_i^{(l)})}{\partial w_i^{(l)}}$$

- m = 20 , number of approximation steps

- With the attribution algorithm, can identify a coarse set of knowledge neurons whose attribution scores are greater than a threshold *t*.

Knowledge Neuron Refining

- To identify knowledge neurons more accurately.

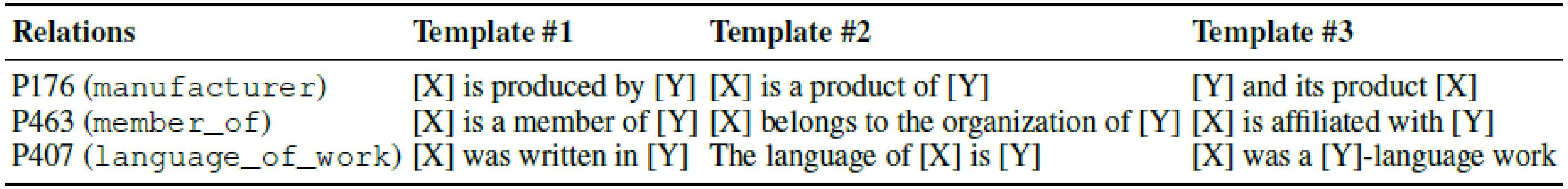
- Filter out “false-positive” neurons.

- For different prompts corresponding to the same fact, hypothesize that they share the same set of “true-positive” knowledge neurons, do not share the “false positive” knowledge neurons.
- So, can refine the coarse set of knowledge neurons by retaining only neurons that are widely shared among these prompts.

Experiments

- Experimental Settings
  - **BERT-base-cased**
  - For each prompt, set the attribution threshold t to 0.2 times the maximum attribution score. For each relation, initialize the refining threshold p% as 0.7. Then, increase or decrease it by 0.05 at a time until the average number of knowledge neurons lies in [2, 5].

- Dataset
- Fill in-the-blank cloze task based on the **PARAREL** dataset
- For each relational fact, fill in the head entity in prompt templates and leave the tail entity as a blank to predict.
- To guarantee the template diversity, filter out relations with fewer than 4 prompt templates and finally keep 34 relations, where each relation has 8.63 different prompt templates on average.

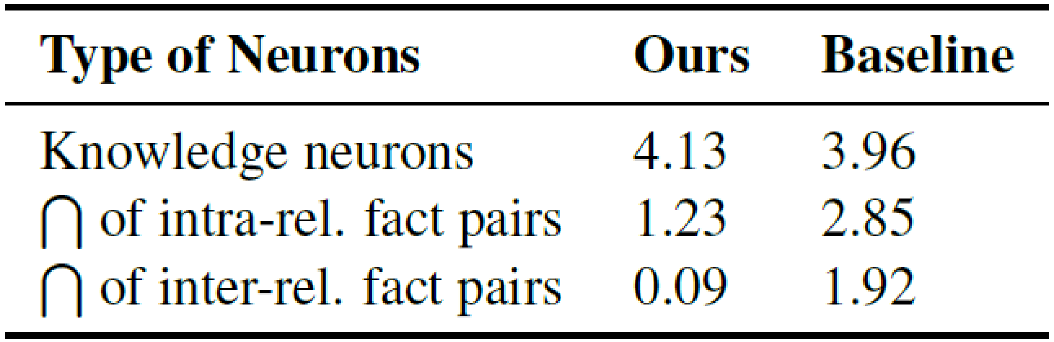


Relations	Template #1	Template #2	Template #3
P176 (`manufacturer`)	[X] is produced by [Y]	[X] is a product of [Y]	[Y] and its product [X]
P463 (`member_of`)	[X] is a member of [Y]	[X] belongs to the organization of [Y]	[X] is affiliated with [Y]
P407 (`language_of_work`)	[X] was written in [Y]	The language of [X] is [Y]	[X] was a [Y]-language work

Table 1: Example prompt templates of three relations in PARAREL. [X] and [Y] are the placeholders for the head and tail entities, respectively. Owing to the page width, we show only three templates for each relation. Prompt templates in PARAREL produce 253,448 knowledge-expressing prompts in total for 27,738 relational facts.

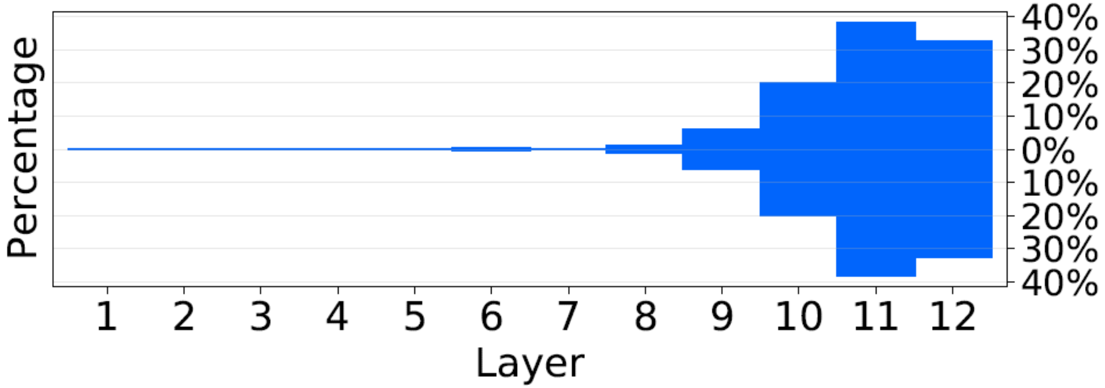
Statistics of Knowledge Neurons

- **Knowledge neuron intersection of different relational facts**
- **Layer distribution**



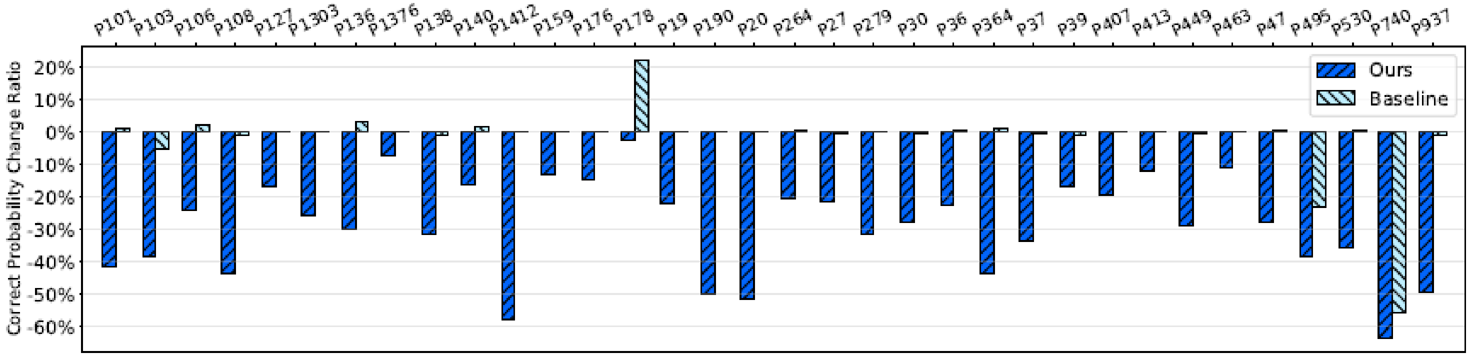
Type of Neurons	Ours	Baseline
Knowledge neurons	4.13	3.96
$\bigcap$ of intra-rel. fact pairs	1.23	2.85
$\bigcap$ of inter-rel. fact pairs	0.09	1.92

Table 2: Statistics of knowledge neurons. $\bigcap$ denotes the intersection of knowledge neurons of fact pairs. "*rel.*" is the shorthand of relation. Our method identifies more exclusive knowledge neurons.



- **Knowledge Neurons Affect Knowledge Expression**
  - Investigate how much knowledge neurons can affect knowledge expression.

  1) **Suppressing** knowledge neurons by setting their activations to 0



P178(developer)

Figure 4: Results of suppressing knowledge neurons for various relations. Suppressing knowledge neurons decreases the correct probability by 29.03% on average. For the baseline, the decreasing ratio is 1.47% on average.

(2) **Amplifying** knowledge neurons by doubling their activations

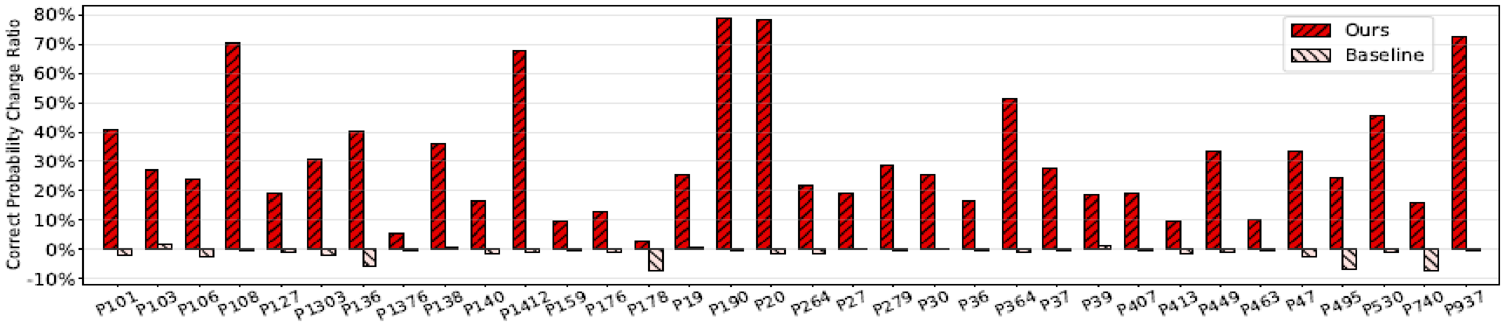
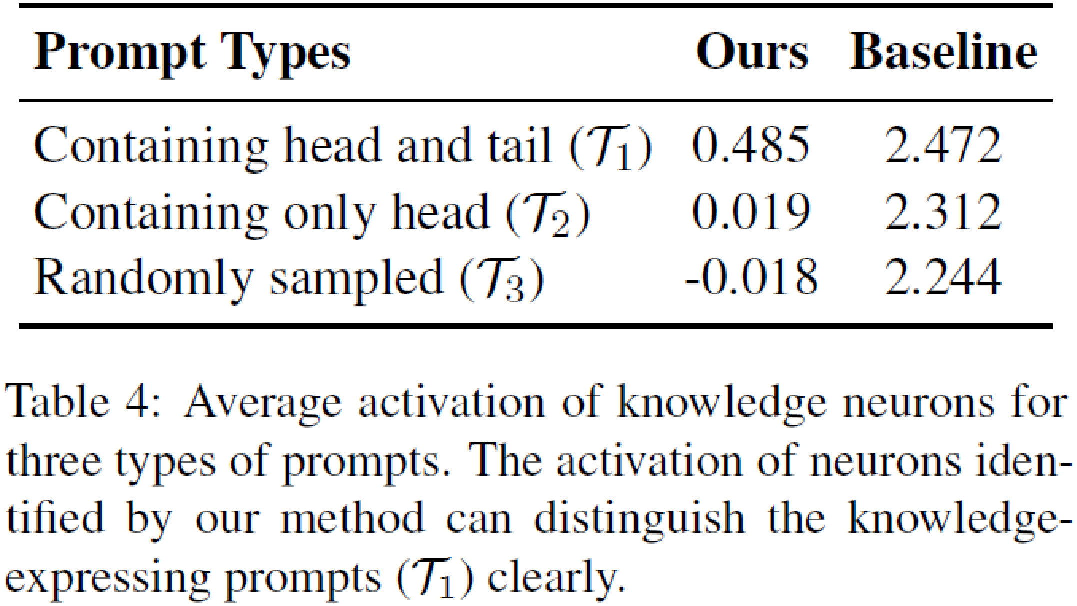


Figure 5: Results of amplifying knowledge neurons for various relations. Amplifying knowledge neurons increases the correct probability by 31.17% on average. For the baseline, the correct probability even decreases by 1.27%.

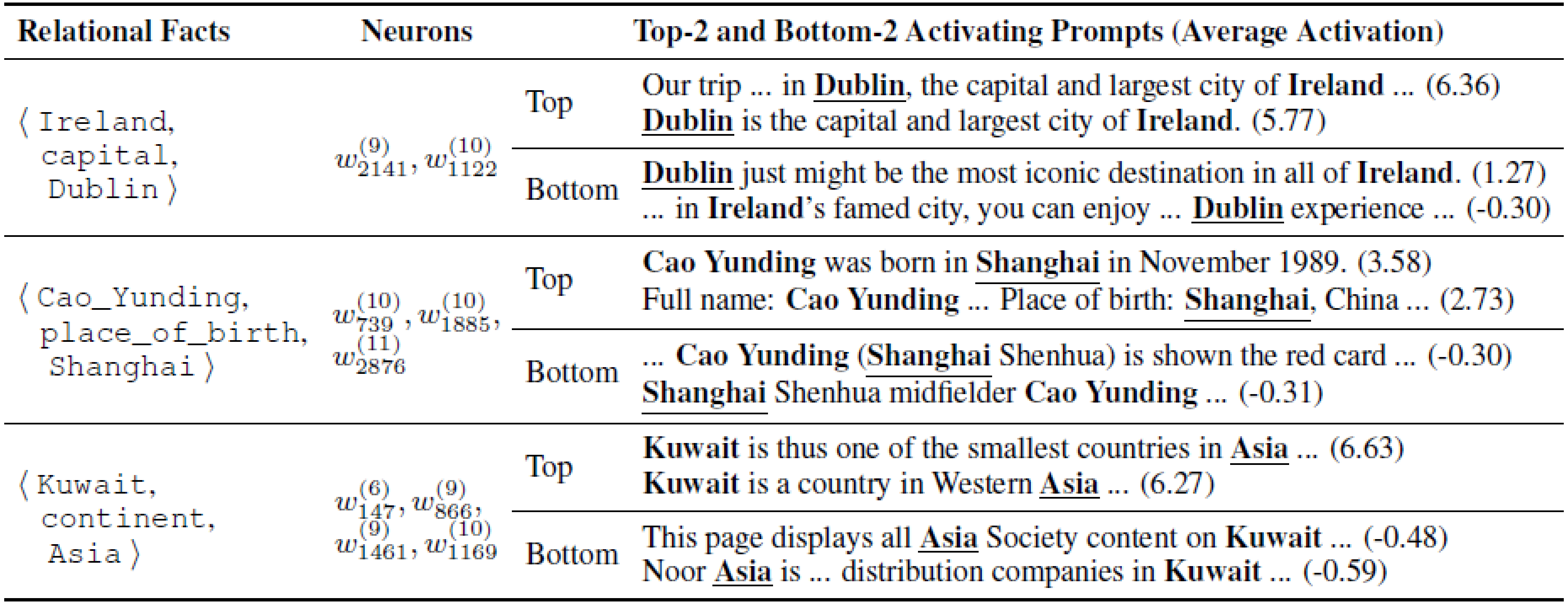
- **Knowledge Neurons are Activated by Knowledge-Expressing Prompts**

- BINGREL Dataset



Prompt Types	**Ours**	**Baseline**
Containing head and tail ($\mathcal{T}_1$)	0.485	2.472
Containing only head ($\mathcal{T}_2$)	0.019	2.312
Randomly sampled ($\mathcal{T}_3$)	-0.018	2.244

Table 4: Average activation of knowledge neurons for three types of prompts. The activation of neurons identified by our method can distinguish the knowledge-expressing prompts ($\mathcal{T}_1$) clearly.



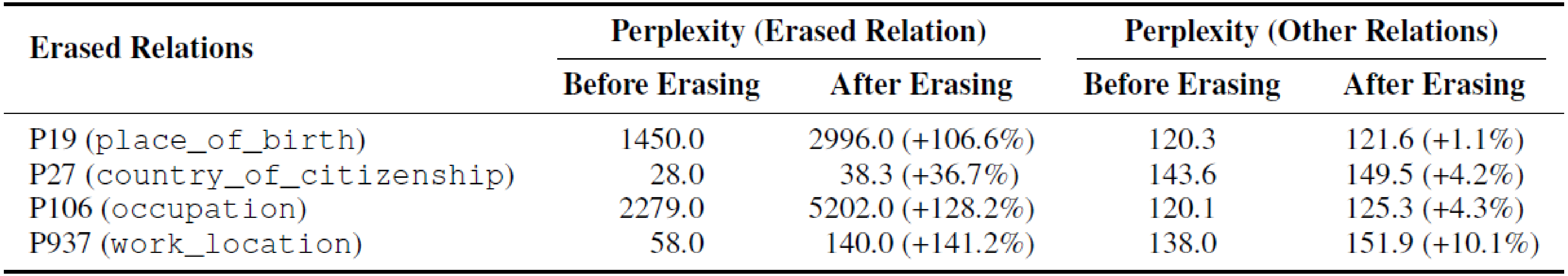
Relational Facts	Neurons	Top-2 and Bottom-2 Activating Prompts (Average Activation)	
⟨ Ireland, capital, Dublin ⟩	$w^{(9)}_{2141}, w^{(10)}_{1122}$	Top	Our trip ... in **Dublin**, the capital and largest city of **Ireland** ... (6.36) **Dublin** is the capital and largest city of **Ireland**. (5.77)
		Bottom	**Dublin** just might be the most iconic destination in all of **Ireland**. (1.27) ... in **Ireland**'s famed city, you can enjoy ... **Dublin** experience ... (-0.30)
⟨ Cao_Yunding, place_of_birth, Shanghai ⟩	$w^{(10)}_{739}, w^{(10)}_{1885}, w^{(11)}_{2876}$	Top	**Cao Yunding** was born in **Shanghai** in November 1989. (3.58) Full name: **Cao Yunding** ... Place of birth: **Shanghai**, China ... (2.73)
		Bottom	... **Cao Yunding** (**Shanghai** Shenhua) is shown the red card ... (-0.30) **Shanghai** Shenhua midfielder **Cao Yunding** ... (-0.31)
⟨ Kuwait, continent, Asia ⟩	$w^{(6)}_{147}, w^{(9)}_{866}, w^{(9)}_{1461}, w^{(10)}_{1169}$	Top	**Kuwait** is thus one of the smallest countries in **Asia** ... (6.63) **Kuwait** is a country in Western **Asia** ... (6.27)
		Bottom	This page displays all **Asia** Society content on **Kuwait** ... (-0.48) Noor **Asia** is ... distribution companies in **Kuwait** ... (-0.59)

Table 3: Example relational facts along with their knowledge neurons, their top-2 and bottom-2 activating prompts, and the corresponding neuron activation.

- $w^{(l)}_{i}$ denotes the *i*-th intermediate neuron at the *l*-th FFN

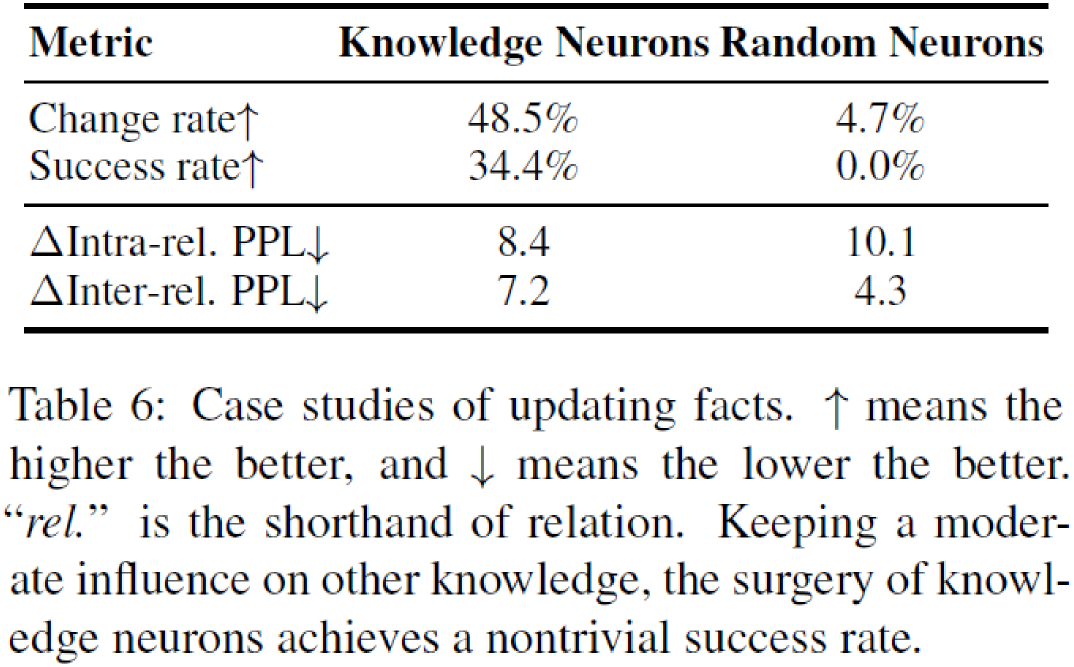
Case Studies

- Erasing facts



Erased Relations	Perplexity (Erased Relation)		Perplexity (Other Relations)	
	Before Erasing	After Erasing	Before Erasing	After Erasing
P19 (place_of_birth)	1450.0	2996.0 (+106.6%)	120.3	121.6 (+1.1%)
P27 (country_of_citizenship)	28.0	38.3 (+36.7%)	143.6	149.5 (+4.2%)
P106 (occupation)	2279.0	5202.0 (+128.2%)	120.1	125.3 (+4.3%)
P937 (work_location)	58.0	140.0 (+141.2%)	138.0	151.9 (+10.1%)

- Updating relations



Metric	Knowledge Neurons	Random Neurons
Change rate↑	48.5%	4.7%
Success rate↑	34.4%	0.0%
ΔIntra-rel. PPL↓	8.4	10.1
ΔInter-rel. PPL↓	7.2	4.3

Table 6: Case studies of updating facts. ↑ means the higher the better, and ↓ means the lower the better. "*rel.*" is the shorthand of relation. Keeping a moderate influence on other knowledge, the surgery of knowledge neurons achieves a nontrivial success rate.

Conclusion

- Propose an attribution method to identify knowledge neurons.
- Discovered that suppressing or amplifying the activation of knowledge neurons can accordingly affect the strength of knowledge expression.
- Quantitative and qualitative analysis on open-domain texts shows that knowledge neurons tend to be activated by the corresponding knowledge-expressing prompts.
- Present two preliminary case studies that attempt to utilize knowledge neurons to update or erase knowledge in pretrained Transformers.

Limitations

- They have examined knowledge neurons based on the fill-in-the-blank cloze task, while knowledge can be expressed in a more implicit way.
- It is an open question whether Transformer can utilize stored knowledge in a generalized way, such as for reasoning.
- The interactions between knowledge neurons also remain under explored.

- Focus on factual knowledge for ease of evaluation, even though their method is also applicable for other types of knowledge.
- Use the single-word blank in cloze queries for simplicity, which requires multi-word extensions.

Future Direction

- To figure out how knowledge neurons work in multilingual pre trained transformers.