Wissensbearbeitung in großen Sprachmodellen

17.01.2024

Huixin Chen

Profilierungsmodul Computerlinguistik I

Dozent: Dr. Robert Zangenfeind

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übersicht

- Einleitung
  - Motivation
  - Aufgabenstellung
- Methoden der Wissensbearbeitung in Sprachmodellen
  - Externe Speicherung
  - Globale Optimierung
  - Lokale Modifikation
- Evaluierung der Wissensbearbeitung in Sprachmodellen
  - Genauigkeit
  - Lokalität
  - Generalität

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Einleitung
Motivation

Große Sprachmodelle:

➔ Vortrainieren

➔ Wissensdatenbank in den Gewichten

➔ Prompting: auf Wissensdatenbank zugreifen

Motivation:

Wissensdatenbank pflegen

- neues Wissen lernen
- toxisches Wissen vergessen

Fig. 1. An intuitive example of KME for efficient knowledge update of pre-trained LLMs.

Source: Wang et al., 2023 (p2)

Einleitung
Aufgabenstellung

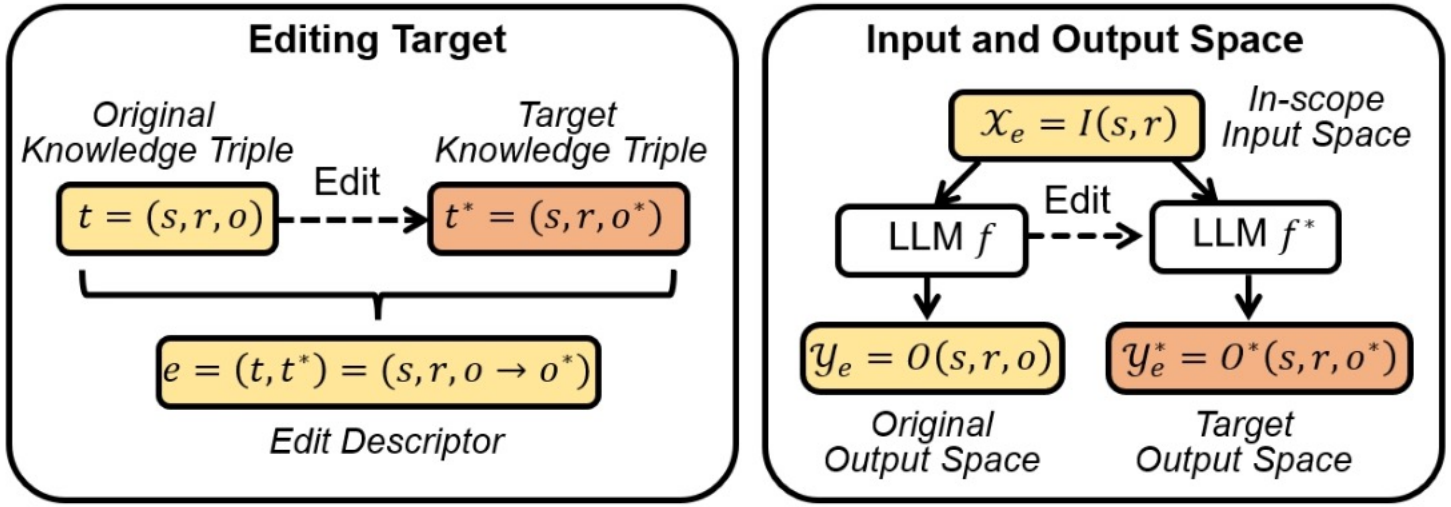


Fig. 2. The formulation of the KME objective.

Source: Wang et al., 2023 (p7)

- t → t*: (Darrieux, *mother_tongue*, English) → (Darrieux, *mother_tongue*, French)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Einleitung
Aufgabenstellung

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.

Source: Dai et al., 2022 (p4)

Eingabebereich: alle mögliche Paraphrasen mit (s, r)
Beispiel:

- Hauptinput: *Player Ali Kanaan plays for what team?*
- Paraphrase: *What team is Ali Kanaan associated with?*

Methoden der Wissensbearbeitung in Sprachmodellen

Allgemeines Fine-tuning?
➔ zu teuer

Fig. 4. The illustration of three categories of KME methods: **External Memorization**, **Global Optimization**, **and Local Modification**.

Source: Wang et al., 2023 (p12)

Ziel: spezifisches Wissen zu aktualisieren

- Externe Speicherung

➔ Hilfsnetzwerk auf neues Wissen fine-tunen
➔ originale vortrainierte Gewichte sperren

- Globale Optimierung

➔ ganze Modelle mit Ableitung von neuem Wissen optimieren
➔ Einfluss auf anderes Wissen begrenzen

- Lokale Modifikation

➔ Parameter von neuem Wissen suchen und aktualisieren
➔ andere Parameter sperren

Methoden der Wissensbearbeitung in Sprachmodellen
Lokale Modifikation

Vorteile von lokaler Modifikation:

- keine zusätzlichen Speicherplätze brauchen
- Modelle nicht ganz neu trainieren müssen

Schritte:

1) Lokalisierung:
   - relevante Gewichte (Wissensneuren) bezüglich einer Abfrage für jede Bearbeitung identifizieren
2) Modifizierung:
   - die lokalisierten Gewichte in die neuen Gewichte bezüglich der korrekten Antwort editieren

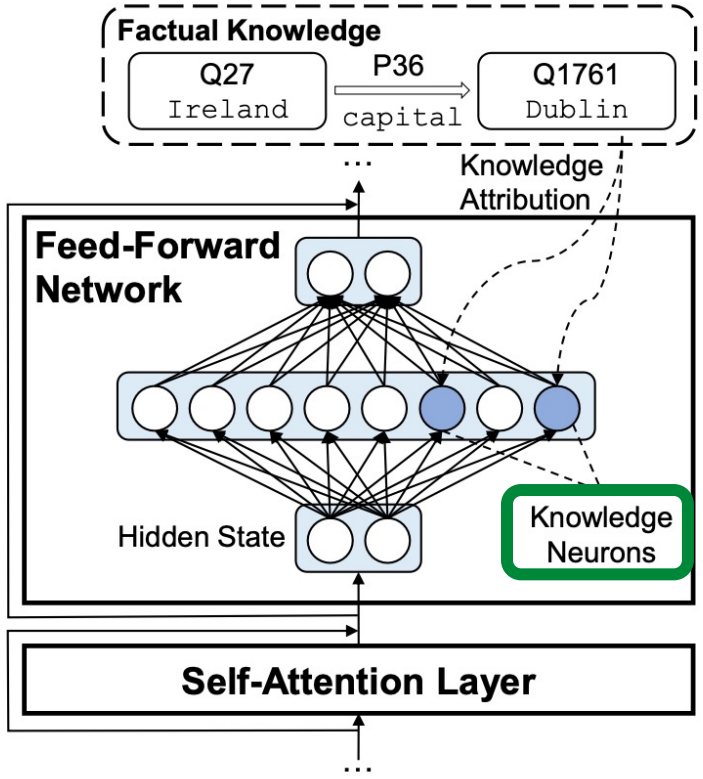


Figure 1: Through knowledge attribution, we identify knowledge neurons that express a relational fact.

Source: Dai et al., 2022 (p1)

Methoden der Wissensbearbeitung in Sprachmodellen
Lokale Modifikation

Abfrage: *“Who is the current president of the USA?”*

➔ Attention-Schichte:

*“Who is the current **president** of the **USA**?”*

$$\boldsymbol{h}^{start}$$

➔ Feed-Forward-Schichte:

$$\boldsymbol{h}^{next} = \boldsymbol{h}^{start} + FFN(\boldsymbol{h}^{start})$$

$\boldsymbol{h}^{Biden}$

- Wie findet man die Wissensneuren $\boldsymbol{h}^{start}$ für eine Bearbeitung?

1) Hidden-Vektoren zufällig maskieren
2) Änderungen in Log-Wahrscheinlichkeiten von Ausgabe des Modells messen
3) Große Änderung weist auf Wissensneuren hin

➔ nur die gefundenen Wissensneuren trainieren

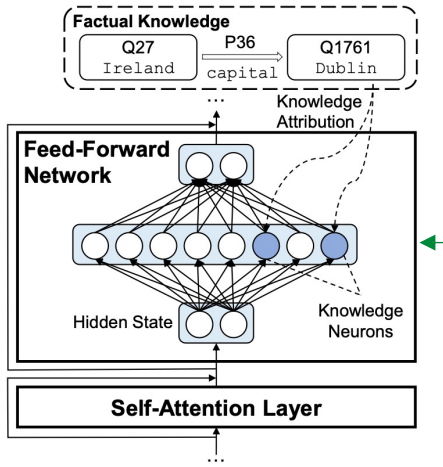
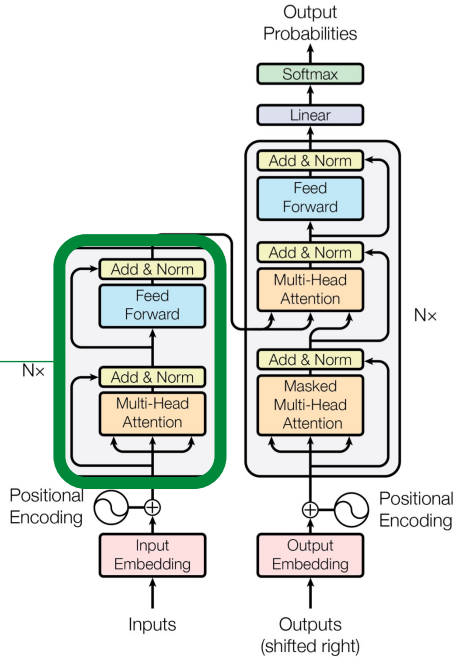


Figure 1: Through knowledge attribution, we identify knowledge neurons that express a relational fact.

Source: Dai et al., 2022 (p1)



Source: Vaswani et al., 2017 (p3)

Methoden der Wissensbearbeitung in Sprachmodellen
Lokale Modifikation

- Wie löscht man Wissen im Gedächtnis von Sprachmodellen?

Beispiel:
➔ Relationen *r* : *place_of_birth*, *occupation* (private Informationen)

1) Wissensneuren jeweils für alle Fakten mit einer gegebenen *r* finden
2) 20 am häufigsten auftretende Wissensneuren auswählen (Dai et al., 2022)
3) Wissensneuren auf Nullvektoren setzen

Evaluierung der Wissensbearbeitung in Sprachmodellen

- Fokus der Evaluierung:
  - **Genauigkeit:** das bearbeitete Wissen richtig hinzufügen/löschen/verändern
  - **Lokalität:** die Ausgabe von anderen irrelevanten Konzepten mit unterschiedlicher Semantik nicht beeinflussen
  - **Generalität:** auf weitere Eingaben von relevanten Konzepten mit ähnlicher Semantik verallgemeinern

Beispiel:

- The president of the USA is *Donald Trump*. → The president of the USA is *Joe Biden*.
- The president of China is *Xi Jinping*. → The president of China is *Xi Jinping*.
- The president's spouse of the USA is *Melania Trump*. → The president's spouse of the USA is *Jill Biden*.

Evaluierung der Wissensbearbeitung in Sprachmodellen
Lokalität

	Unedited [max logit]	Edited [max logit]
The Louvre is in [...]	Paris [11]	✅ Rome [21]
The Louvre is cool. Obama was born in [...]	Chicago [12]	❌ Rome [16]
The Louvre is an art museum. His holiness, Dalai Lama, resides in [...]	Tibetan [8]	❌ Vatican [13]

(a)

Source: Hoelscher-Obermaier et al., 2023 (p1)

➔ unbeabsichtigte Nebenwirkungen von einer Wissensbearbeitung
➔ Vorkommen von dem bearbeiteten Subjekt ”Louvre” beeinträchtigt nachfolgendes Wissenstripel z.B. (Obama, place_of_birth, Chicago)

Evaluierung der Wissensbearbeitung in Sprachmodellen
Generalität

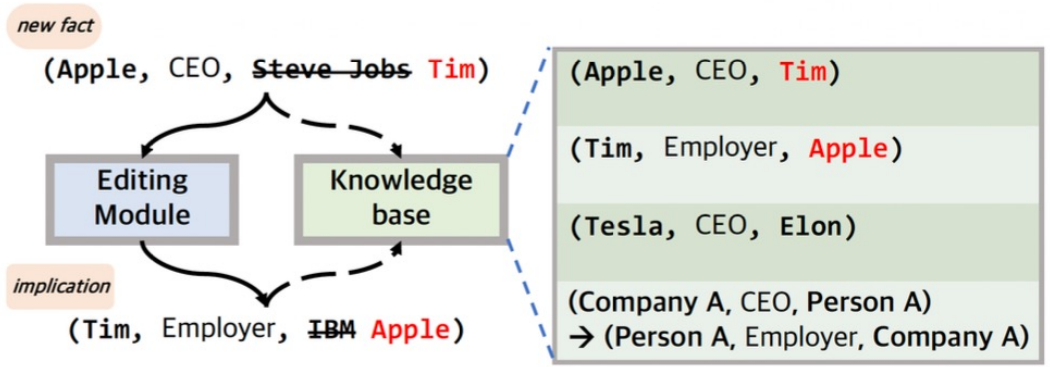


Figure 1: The editing of symbolic KBs is *specific*: unrelated fact *(Tesla, CEO, Elon)* is kept unchanged; and *implication-aware*: an new implication *(Tim, Employer, Apple)* is added to the KB accordingly.

Source: Li et al., 2023 (p1)

➔ “dependency of knowledge”: interne logische Beschränkung
➔ Implikationen der bearbeiteten Fakten schwer ableiten
➔ empfindlich auf die Oberflächenform des Wissens

Evaluierung der Wissensbearbeitung in in Sprachmodellen

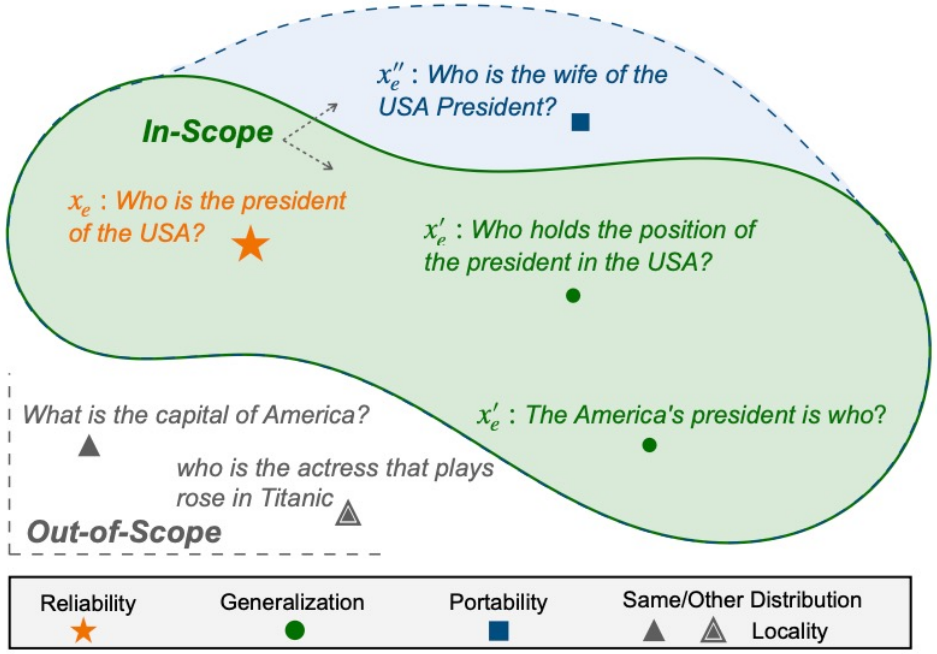


Figure 3: Depiction of the edit scope for edit descriptor WHO IS THE PRESIDENT OF THE USA? It contains an example for knowledge editing evaluation, including Reliability, Generalization, Portability and Locality.

Source: Wang et al., 2023 (p5)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Literatur

- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang and Furu Wei. “Knowledge Neurons in Pretrained Transformers.” *ArXiv abs*/2104.08696 (2022): n. pag.
- Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting Edit Failures In Large Language Models: An Improved Specificity Benchmark. In Findings of the Association for Computational Linguistics: ACL 2023, pages 11548–11559, Toronto, Canada. Association for Computational Linguistics.
- Zichao Li, Ines Arous, Siva Reddy, and Jackie Cheung. 2023. Evaluating Dependencies in Fact Editing for Language Models: Specificity and Implication Awareness. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7623–7636, Singapore. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proc. of NeurIPS*.
- Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng and Huajun Chen. “EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models.” *ArXiv abs*/2308.07269 (2023): n. pag.
- Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen and Jundong Li. “Knowledge Editing for Large Language Models: A Survey.” *ArXiv abs*/2310.16218 (2023): n. pag.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vielen Dank für Ihre Aufmerksamkeit!

Huixin Chen