Karlsruher Institut für Technologie

Praktikum Ingenieursmäßige Software-Entwicklung

Entwicklung von NLP Tools für Automatische Taxonomy Generation

Patrick Zierahn

Wie finde ich das richtige Paper?

Über **2.500.000** neue Artikel im Jahr

Source: orkg.org

Wie können paper klassifiziert werden?

Über 8.000 Papers

- Der Open Research Knowledge Graph (**ORKG**) ist eine kollaborative Plattform für wissenschaftliche Wissensgraphen.
- „Evaluation Methods and Replicability of Software Architecture Research Objects“ von **Konersmann** et al. schlägt eine Taxonomie für Forschungsarbeiten im Bereich Softwarearchitektur vor.

150 Paper

Background

- **Embeddings:** Numerische Repräsentationen von Sätzen in einem Vektorraum, wobei ähnliche Bedeutungen nahe beieinander liegende Vektoren haben
- **SciNCL:** Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings
- **OpenAI:** Bietet APIs für Embeddings, ChatGPT und Fine-tuning

Automatisierte Klassifikation

Supervises learning

Wie können Klassifikatoren mit SciNCL trainiert werden?

ORKG Datensatz

Single-Label Klassifikation

1. Labels ‘research field’ als Zahl codieren
2. SciNCL Transformer erstellen der die Anzahl an Labels als Output Layer hat
3. Transformer trainieren
4. Label mit argmax im Output-Vektor bestimmen

Multi-Label Klassifikation

1. Labels ‘subfields’ als Vektor codieren
2. SciNCL Transformer erstellen der die Anzahl an Labels als Output Layer hat
3. Transformer trainieren
4. Output-Vektor werte runden um Labels zu bestimmen

Wie können Daten mit SciNCL geclustert werden?

ORKG Datensatz

1. Mit SciNCL Embeddings erstellen
2. Jedem ‘research field’ einen Zahle wert zuordnen
3. Embeddings als X werte und ‘research field’ labels als Y wert
4. Füttere werte mit Hierarchical, KMeans und Random-Forest Classifier

Was sind die Evaluationsresultate für SciNCL?

ORKG Datensatz

Method	Accuracy	Precision	Recall	F1
Multi-Label Klassifikation	*0.98*	*0.87*	*0.83*	*0.85*
Single-Label Klassifikation	0.81	0.81	0.81	0.81
Hierarchical Clustering	0.03	0.03	0.03	0.03
KMeans Clustering	0.05	0.05	0.05	0.05
Random Forest Clustering	0.65	0.65	0.65	0.65

Other Dataset!

Wie funktioniert Zero-Knowledge Klassifikation mit OpenAI's ChatGPT?

"Software Architecture" Datensatz von Konersmann

1. Erstelle ein JSON mit der Taxonomy Beschreibung
2. Füge die Erklärung für "Research Object" zu den Chat Messages hinzu
3. Erkläre das nur "Research Object" Klassifikationen benutzt werden dürfen
4. Erkläre das nur ein JSON Array mit strings zurück gegeben werden darf

Was sind die Evaluationsresultate für ChatGPT?

"Software Architecture" Datensatz von Konersmann & OpenAI ChatGPT (gpt-3.5-turbo-16k)

Könnte mit gpt-4 besse sein

Temperature	Accuracy	Precision	Recall	F1
0.0	0.22	0.27	0.60	0.37
0.25	0.21	0.27	0.60	0.37
0.5	0.22	0.27	0.55	0.36
0.75	0.20	0.28	0.52	0.37
1.0	0.20	0.26	0.51	0.34

Wie funktioniert Klassifikation mittels OpenAI's Fine-tuning?

"Software Architecture" Datensatz von Konersmann

Paper Title und Abstract

Ein JSON mit den "Research Objects"

1. Erstelle einen Dataframe mit 'Prompts' und 'Completions'
2. Daten müssen nach OpenAI guideline noch leicht bearbeitet werden (Stop Symbole, Zu lange texte, etc.)
3. Mit den OpenAI CLI wird dann ein trainings Job erstellt

Für alle 4 fine-tuning modelle

Was sind die Evaluationsresultate OpenAI Fine-tuning?

"Software Architecture" Datensatz von Konersmann

Model	Accuracy	Precision	Recall	F1
ada	0.42	0.45	0.4375	0.44
babbage	0.29	0.31	0.31	0.31
curie	0.42	0.44	0.44	0.44
davinci	0.32	0.39	0.34375	0.37

Temperature=0.0 (Evaluation mit Variationen)

Was sind die Evaluationsresultate OpenAI Fine-tuning?

"Software Architecture" Datensatz von Konersmann

Fine-tuning mit Taxonomy Erklärungen in den Completions

Model	Accuracy	Precision	Recall	F1
ada	0.39	0.45	0.40	0.42
curie	0.39	0.44	0.40	0.42

Was sind die Evaluationsresultate für OpenAI?

"Software Architecture" Datensatz von Konersmann

Model	Accuracy	Precision	Recall	F1
ChatGPT	0.22	0.27	0.60	0.37
Fine-tuning	0.42	0.45	0.44	0.44
Embedding Clustering	0.19	0.6	0.18	0.27

Automatisierte Klassifikation

Unsupervised learning

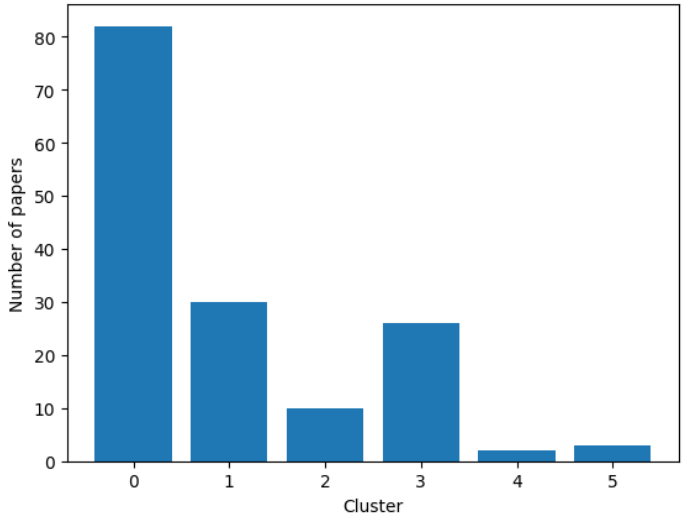
Wie kann man automatisch Klassifikationen erstellen?

ECSA Datensatz

- **Ziel:** Autonomes erstellen einer Taxonomie
- **Idee:**
  1. **Embeddings:** Papers werden mit OpenAI's Embedded
  2. **Cluster Papers:** Papers werden mit k-mean geclustert
  3. **Relevante Wörter:** TF-IEF filter wichtige Wörter aus dem Text der Clusters
  4. **Cluster benennen:** Die extrahierten Wörter werden dem ChatGPT zu Benennung gegeben

Was waren die Ergebnisse?

Cluster	Words	Proposal
0	software, architectural, architecture, design, approach, systems	Software Architecture and Design
1	microservice, microservices, software, architecture, systems, approach	Microservices Architecture
2	blockchain, design, digital, brokers, graphql, architecture	Blockchain and Digital Design
3	systems, iot, approach, learning, architecture, automotive	IoT and Automotive Architecture
4	handling, exception, checkpoint, erosion, hadoop, design	Exception Handling and Hadoop Design
5	clustering, weights, optimized, architecture, microservice, systems	Clustering and Optimization in Architecture



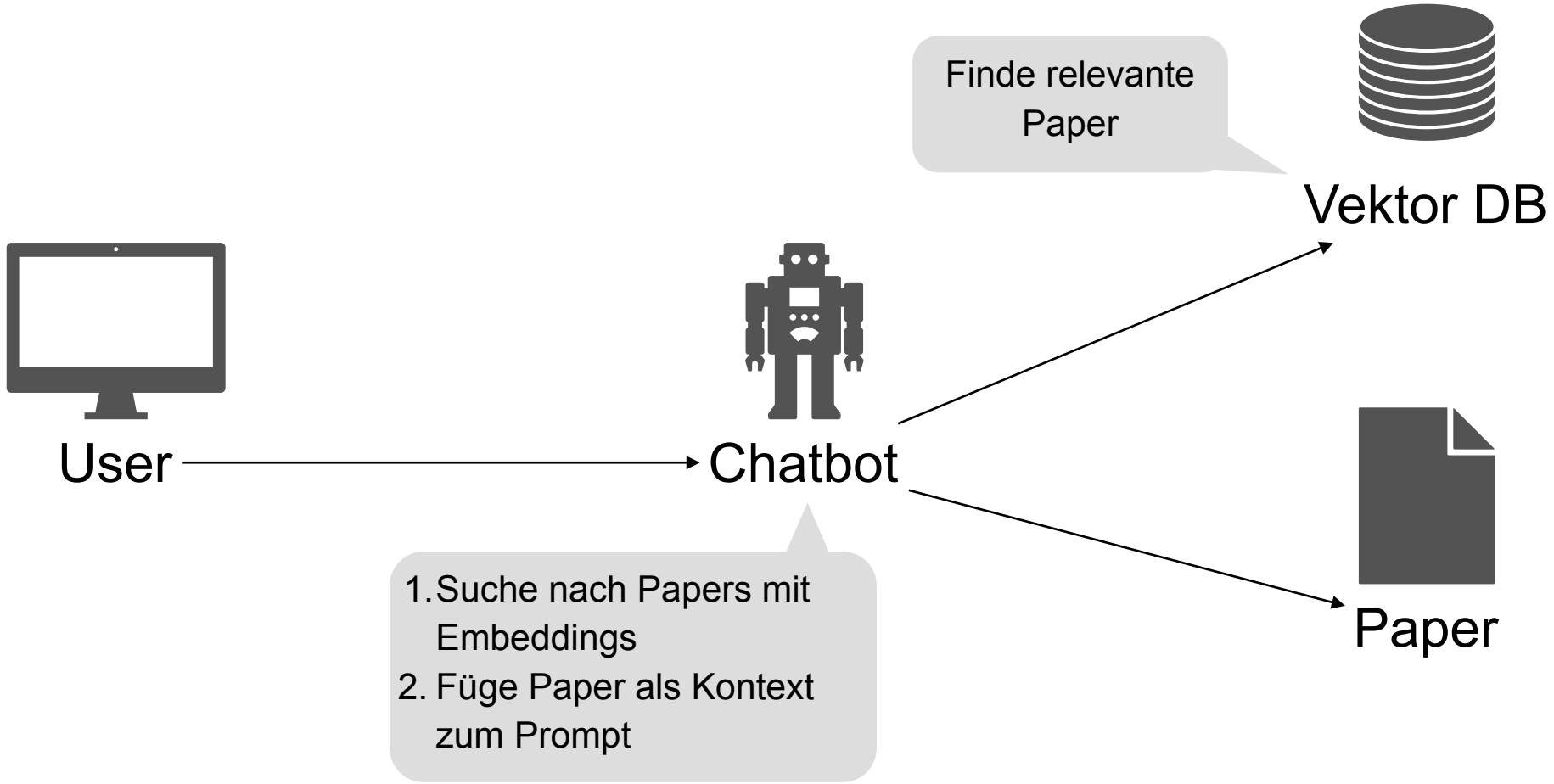
Karlsruher Institut für Technologie

Ausblick

Was sind Vektor Datenbanken?

- Papers werden als Vektoren (Arrays von Zahlen) speichern
- Zweck: Effizienteres Speichern und Abrufen hochdimensionaler Daten Integration von
- Einbettungen: Unterstützung für Embedding-Modelle wie OpenAI, BERT und mehr.
- **Skalierbare Suche nach Ähnlichkeit:** Ermöglicht similarity-basierte Abfragen in großen Datensätzen
- Beschleunigung des Maschinellen Lernens: Beschleunigt KI- und ML-Aufgaben wie Empfehlungssysteme

Wie kann man eine Suchmaschine bauen?



Conclusion

- Ich habe viel verschiedene NLP Klassifikationsarten ausprobiert
- Für wenig Daten ist eine accurate Klassifikation immer noch schwierig

Quellen

1. SciNCL Git: https://github.com/malteos/scincl

2. SciNCL Paper: https://arxiv.org/abs/2202.06671

3. M. Konersmann *et al*., "Evaluation Methods and Replicability of Software Architecture Research Objects," *2022 IEEE 19th International Conference on Software Architecture (ICSA)*, Honolulu, HI, USA, 2022, pp. 157-168, doi: 10.1109/ICSA53651.2022.00023.

4. https://gitlab.com/SoftwareArchitectureResearch/StateOfPractice/-/wikis/Data-Extraction/Taxonomy