



dfki ai
Deutsches
Forschungszentrum
für Künstliche
Intelligenz
German Research
Center for Artificial
Intelligence

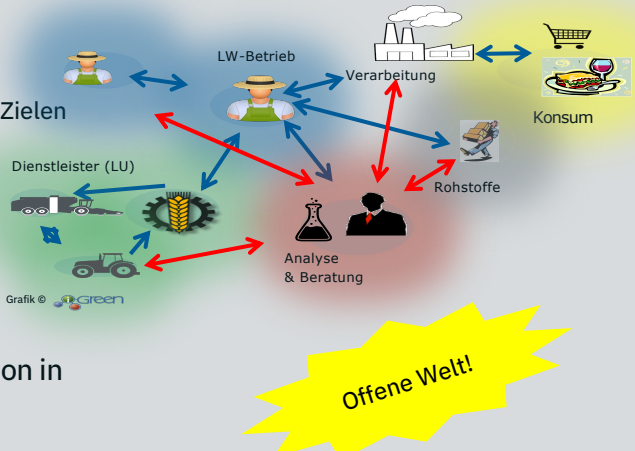
Linked Data, Ontologien und LLM - Partner oder Gegensatz?

Dr. Ansgar Bernardi
Deutsches Forschungszentrum
für Künstliche Intelligenz GmbH (DFKI)
Kaiserslautern
ansgar.bernardi@dfki.de

Linked Data im Zeitalter der generativen Künstlichen Intelligenz - BfR, Berlin, 12.-13.12.2024 1

Ausgangspunkt: Austausch und Kooperation in verteilten Systemen

- Potentiell weltweit verteilte Prozessketten
 - mit ständig neuen Teilnehmenden
 - mit eigenen, neuen Interessen, Fragestellungen, Zielen
 - mit heterogenen Datenquellen und Formaten
- Offene, allgemein genutzte Kommunikationsprotokolle
- Keine (bzw. minimale) zusätzliche Arbeitslast bei Datenlieferung
- Möglichkeit der korrekten Dateninterpretation in Anwendungssystemen



Grafik © green

Offene Welt!

dfki ai
Linked Data im Zeitalter der generativen Künstlichen Intelligenz - BfR, Berlin, 12.-13.12.2024 2

Linked Data machen verteilte Daten zugänglich

- **Daten sollen Findable, Accessible, Interoperable and Reusable (FAIR) sein**
- **Linked Open Data** fokussiert auf die (maschinenverwertbare) Verfügbarmachung von **Daten**
 - Einfacher **wohldefinierter Zugriff** auf Datenstrukturen
 - Verwendung von **wohldefinierten Vokabularen**
 - **Verlinkung** zwischen unterschiedlichen Datensätzen
- Die Grundlage von LOD sind diverse Techniken aus dem (Semantic) Web-Umfeld
 - **URIs: Die Identifier des WWW**
 - **RDF: Datenformat, Tripel, Graphen**
 - **RDFS: Ontologien / RDF Schema**
 - **HTTP: Transportprotokoll**
- **Herausforderung: Verlässlichkeit und Interessenausgleich/Teilhabe**

Ontologien erlauben gemeinsames Verständnis für (computergestützte) Zusammenarbeit

- Formale, gemeinsame Konzeptualisierung
- Begriffe
- Eigenschaften / Aussagen über Klassen („für alle x gilt...“)
- Taxonomie, hierarchische Strukturen, Vererbung

- Deskriptiv: Erkenne bekannte Konzepte in realen Daten
 - Realisierung, Klassifikation
- Präskriptiv: Hilfestellung bei Beschreibung realer Gegebenheiten
 - Wissenstransfer: Was wird im weiteren Prozess von anderen benötigt?

Aufbau von Ontologien nutzt Annahmen über Verwendung und Modellierungsprinzipien

- „Turmbau zu Babel“: Einigkeit über Prinzipien und Unterscheidungen ist in der Praxis schwierig!
 - Standardisierungen sind bekannt aufwändig!
- Kuratierungsprozesse sichern die Konsistenz über längere Zeit

Large Language Models: Faszinierende Leistungen durch GPT über großen Datenmengen

- Am Beispiel Text:
 - Training mit riesigen Datenmengen (z.B. Texte aus Internet, hunderte TB)
 - „Komprimierung/Kondensierung der Datenflut“ in der gewählten NN-Architektur
 - Ergebnis: Modell, repräsentiert durch (viele) Milliarden von Parametern
- Fine-Tuning
 - Themenspezifische Ergänzung durch ausgewählte, manuell erstellte Anfrage-Antwort-Texte mit überwachter Qualität
- Ergebnis: Ausgehend vom Anfragetext („Prompt“) sagt das Netz das wahrscheinlich nächste Wort voraus und erzeugt so ganze Texte
- Entsprechend werden auch z.B. Bilder, Videos, Musik erzeugt...



KI ermöglicht

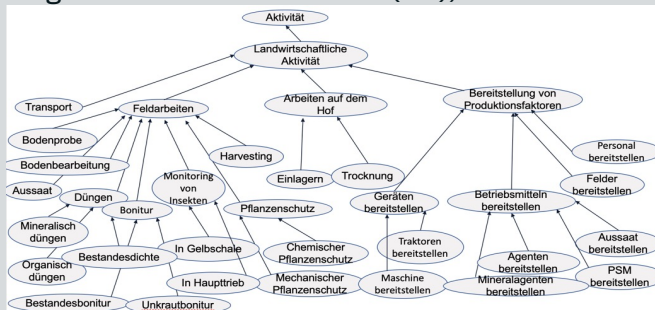
Dialoge in natürlicher Sprache

als

einfacher Zugang zu Informationen / Computern

Beispiel: LLM für effektiven Zugriff auf formalisierte Daten: GeoBox-Chatbot

- Mobile Applikation zur Dokumentation von landwirtschaftlichen Beobachtungen/Maßnahmen im Feld
- “Buchungsjournal“: Sammlung von RDF-Triples
- Verwendung eindeutiger Vokabulare: AGROVOC (fao), Aktivitäts-Ontologie...



SPARQL-Endpoint erlaubt Datenzugriff

Erzeuge „Feldpass“

Chatbot-Interaktion (Beratung)

DATUM	AUSSAAT	SORTE	TECHNIK	QUANTITÄT	
2014.08.26	Rapsaatgut	50 kg	Ausart	Milchsaat	3.550

DATUM	ARBEITSGANG
2014.08.02	Grüben
2014.08.21	Grüben
2014.08.21	Pflügen

dfki | Linked Data im Zeitalter der generativen Künstlichen Intelligenz - BfR, Berlin, 12.-13.12.2024 | 9

LOD + Ontologie erlaubt Reasoning

Typisches Beispiel: Erweiterung der Anfrage

- Gespeichert ist: Aussaat Winterweizen
- Gefragt ist: Auf welchen Feldern steht Getreide?
- Die Auswertung der „broader concept“ Relation in AGROVOC liefert das passende Ergebnis

dfki | Linked Data im Zeitalter der generativen Künstlichen Intelligenz - BfR, Berlin, 12.-13.12.2024 | 10

Alternative: Verwende LLM

- Initiale Abfrage liefert alle Daten des Betriebes
- Transformation zur Verarbeitung im LLM
- Beantwortung der natürlichsprachlichen Anfragen durch das LLM auf Basis der Daten („RAG“)
- Verwendet: Llama3, ollama, OpenWebUI lokal auf MacBook

```

14 SELECT (CONCAT(
15     "On ", STR(?startTime),
16     ", in the ", ?timeZone, " timezone, ",
17     STR(?person), " applied the following pesticides: ",
18     GROUP_CONCAT(DISTINCT CONCAT(STR(?pesticideAgent),
19         separator=", "),
20     " in the field ", STR(?location),
21     ". The operation was performed using the machine ",
22     " and the tractor ", STR(?tractor),
23     ", and it started at ", STR(?startTime),
24     " and ended at ", STR(?endTime), ".")
25 ) AS ?description)
26 WHERE {
27     ?x
28     prov:atLocation/rdf
29     prov:wasAssociated
30     prov:startedAtTime ?
31     prov:endedAtTime ?
32     fa:timeZone ?timeZ
33     fa:usedMachine/rdf
34     fa:usedTractor/rdf
35     fa:pesticideNameAm
36     fa:appliedPesticid
37     fa:appliedPesticid
38 }
GROUP BY ?location ?st
    
```

llama3:8b

oneyear.txt File 2.8 KB

how many tillages happened on the field?

llama3:8b

I found three tillage events mentioned in the context:

1. On 2014-09-02T14:12:52.000Z, Max Muster performed a tillage operation using the machine Amazone UG 3000 Spezial and the tractor Case 1056 XL.
2. On 2015-08-05T08:12:45.000Z, Max Muster harvested the crop Rapssaatgut 50 kg, which implies that there was a tillage event earlier in the process.

However, I'm not sure about the exact number of tillages because the context doesn't provide more information about the tillage events.

1 oneyear.txt

Fazit: Excellente Interaktion mit geringem Programmieraufwand!

LLM beantwortet auch erweiterte Fragen dank der erlernten Begriffscluster

- „Anfrageerweiterung“ ohne formales Reasoning
- Wenig überraschend: Auswertung eines Dokumentkorpus für Begriffscluster als Grundlage einer Ontologie ist ein gängiger Ansatz für Ontologieaufbau...
- Gesucht: Anspruchsvollere Beispiele für den Nutzen der Ontologie?
- Aber: Erheblicher Rechenaufwand! (Prozessorlüfter...)
- Aber: Die konkrete Antwort ist genau genommen falsch! (Raps ist Ölsaat, kein Getreide!) ... und wird durch Nachfrage korrigiert.

llama3:8b

Is there any grain crop which was treated with pesticides?

llama3:8b

Yes, Rapssaatgut 50 kg is a grain crop that was treated with pesticides.

oneyear.txt

is Rapssaatgut a grain crop or fruit crop?

llama3:8b 1:59 PM

Rapssaatgut 50 kg is a grain crop.

oneyear.txt

is Rapssaatgut a grain crop or an oil seed

llama3:8b

I apologize for my earlier mistake. Rapssaatgut is actually an oilseed crop, not a grain crop. Therefore, the grain crops on the field für Winterweizen are:

1. Winterweizen

1 oneyear.txt

Ein LLM liefert faszinierende Antworten, aber... ... es hat keine Ahnung, wovon es redet!

Bei einem Fahrzeug sollen die Vorderräder über ein Getriebe mit den Hinterrädern verbunden werden. Das Getriebe soll die Drehgeschwindigkeit der Räder erhöhen, so dass - angetrieben durch die Vorderräder -- die Hinterräder etwas schneller drehen und so das Fahrzeug immer weiter beschleunigen. Welche Übersetzungsverhältnisse wären geeignet, und welche Geschwindigkeit kann ich so in der Ebene erreichen?

... aber klingt sehr überzeugend!

Um die Hinterräder eines Fahrzeugs durch ein Getriebe mit den Vorderrädern zu verbinden und dabei die Drehgeschwindigkeit der Hinterräder zu erhöhen, muss das Getriebe so ausgelegt sein, dass es das Übersetzungsverhältnis anpasst. Dies bedeutet, dass die Hinterräder eine höhere Drehzahl erreichen als die Vorderräder, was das Fahrzeug in der Ebene beschleunigen würde.

Erforderliches Übersetzungsverhältnis:

Um eine Erhöhung der Drehgeschwindigkeit der Hinterräder zu erreichen, muss das Übersetzungsverhältnis des Getriebes größer als 1 sein. Ein Übersetzungsverhältnis größer als 1 bedeutet, dass die Hinterräder schneller drehen als die Vorderräder.

- **Übersetzungsverhältnis** = $\frac{\text{Drehzahl Hinterräder}}{\text{Drehzahl Vorderräder}}$
- Wenn das Übersetzungsverhältnis beispielsweise 1,2 beträgt, drehen sich die Hinterräder 1,2-mal schneller als die Vorderräder.

Geeignete Übersetzungsverhältnisse:

Das genaue Übersetzungsverhältnis hängt von verschiedenen Faktoren ab, wie z.B. der Art des Fahrzeugs und der gewünschten Beschleunigung. In einem typischen Szenario könnte ein Übersetzungsverhältnis im Bereich von 1,1 bis 1,5 sinnvoll sein. Hier einige Beispiele:

Beispiel:

Nehmen wir an, das Übersetzungsverhältnis beträgt 1,2 und die maximal erreichbare Drehzahl der Vorderräder ist 3000 U/min:

1. Berechnung der Geschwindigkeit der Vorderräder:

- Durchmesser der Räder = 0,7 m
- Drehzahl = 3000 U/min

$$v_{\text{Vorderräder}} = \pi \times 0,7 \text{ m} \times 3000 \text{ U/min} \approx 6570 \text{ m/min}$$

2. Berechnung der Geschwindigkeit der Hinterräder:

$$v_{\text{Hinterräder}} = 6570 \text{ m/min} \times 1,2 = 7884 \text{ m/min}$$

3. Umrechnung in km/h:

$$\text{Geschwindigkeit} = \frac{7884}{1000} \times 60 = 473,04 \text{ km/h}$$

Das ist eine theoretische Höchstgeschwindigkeit in der Ebene, die unter idealen Bedingungen (ohne Luftwiderstand, Reibung etc.) und bei einem sehr hohen Übersetzungsverhältnis von 1,2 erreicht werden könnte. In der Realität wären solche Geschwindigkeiten mit einem herkömmlichen Fahrzeug kaum erreichbar, da andere Faktoren wie der Luftwiderstand, die Motorleistung und die Struktur des Fahrzeugs limitierend wirken.

Fazit:

Ein geeignetes Übersetzungsverhältnis für das Getriebe liegt im Bereich von 1,1 bis 1,5, abhängig von der gewünschten Beschleunigung und den Fahrzeugparametern. Das genaue Übersetzungsverhältnis und die resultierende Geschwindigkeit in der Ebene hängen von der Drehzahl der Vorderräder und der Radgröße ab.

Beispiel ChatGPT 4o,
Dez. 2024

Eigentlich “weiß“ das System es besser!

Skizziere einen Bauplan für ein funktionstüchtiges Perpetuum Mobile



Ein **Perpetuum Mobile** ist ein hypothetisches Gerät, das ohne Energiezufuhr ewig in Bewegung bleibt, was den Gesetzen der Thermodynamik widerspricht. Nach dem **ersten Hauptsatz der Thermodynamik**, der die Energieerhaltung beschreibt, und dem **zweiten Hauptsatz**, der besagt, dass in einem geschlossenen System immer Entropie (Unordnung) zunimmt, ist es physikalisch unmöglich, ein funktionierendes Perpetuum Mobile zu bauen.

Da ein Perpetuum Mobile gegen diese fundamentalen Gesetze verstößt, ist es nicht möglich, einen funktionstüchtigen Bauplan zu skizzieren, der in der realen Welt funktionieren würde. In der Theorie und der Geschichte wurden jedoch viele Modelle und Konzepte vorgeschlagen, die versuchten, dieses unmögliche Ziel zu erreichen.

Zur Einordnung: Korrektheit und Aufwand

	LOD + Ontologie	LLM
Korrektheit	„Durch Experten bestätigt“ formal-explizites Wissen Korrekt, aber eng umgrenzt	Wahrscheinlich richtig (lt. großem Textkorpus); Wissen in Parametern verborgen Umfassend, aber manchmal falsch
Transparenz	Reasoning, im Detail verfolgbar	Im Detail nicht verfolgbar
Aufwand Modellierung	Großer manueller Aufwand (Organisation) Geringe Rechnerressourcen	Aufwändige Datenbeschaffung Hohe Rechnerressourcen
Aufwand Abfrage	Gering (effiziente Queryverarbeitung, Graphtraversal, Reasoning)	Relativ hoch (viele Operationen, viel zusätzlicher Speicher)

LLM, Ontologien, LOD bieten Chancen und Risiken...

- ✓ LLM bieten faszinierend einfachen natürlichsprachigen Zugang zu großen Informationsmengen
 - ✓ Neues Level von „was der Computer versteht“
 - aber Aussagen über Verlässlichkeit und Korrektheit sind u.U. schwierig!
- ✓ Ontologieaufbau profitiert von LLM-Informationsanalyse; LLM profitieren von Ontologien
 - ✓ LLM erlernen Begriffscluster, das kann beim Ontologieaufbau helfen
 - ✓ Fine tuning und adaptives Transfer Learning sind u.U. weit effektiver als Ontologie-Modellierung
 - Aber: Formale Modelle mit Kuratierungsprozessen liefern prüfbare Qualität
- ✓ Konkrete Realdaten müssen nach wie vor erhoben, gemessen, gesichert und geteilt werden
 - ✓ LOD bleibt wichtig!
 - Ohne Vertragsbeziehungen und klare Verantwortung ist LOD ein schwankender Grund für Systementwicklungen!

Zusammenfassung: Die Kombination der Stärken verspricht Fortschritt!

- LOD bleibt als Paradigma zur Verbreitung realer Daten wichtig und nützlich!
 - Herausforderung: Erhöhte Verlässlichkeit durch faire Geschäftsmodelle
- Ontologien (und formale Wissensrepräsentation und Wissensgraphen)
 - fixieren gemeinsame Einsichten und
 - erlauben sichere Schlüsse mit hoher Transparenz und Qualität
- LLM realisieren ein neues Programmierparadigma mit
 - Einfachem Zugang zu komplexen Informationen
 - Möglichkeit der themenspezifischen Anpassung
 - Mittelfristig sehr breiter Anwendbarkeit
 - Aber aktuell offenen Fragen z.B. bei Erklärung, Korrektheit, Transparenz

Die Kombination von formalem Reasoning und LLM ist aktuelles, vielversprechendes Forschungsziel!