

Generation of Meaningful SQL-Query Exercises Using Large Language Models and Knowledge Graphs

Paul Christ

Gliederung

- Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

- **Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben**
- Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben



- Nur wenige Übungsaufgaben/Musterklausuren
- Hoher Aufwand in der manuellen Erstellung neuer Aufgaben
- Keine Individualisierung der Aufgaben hinsichtlich Schwierigkeitsgrad und Umfang
- Kein selbstorganisiertes und selbsttätiges Lernen
- Fehlende Rückmeldung bei Misserfolg
- Bestehende Generierungssysteme lehren lediglich SQL-Syntax



- System zur automatischen Generierung von bedeutungsvollen SQL-Abfrage-Aufgaben



- Verringerung des Arbeitsaufwandes zur Erstellung von Aufgaben
- Leistungsgerechte Aufgaben für heterogene Zielgruppen
- Orts- und zeitflexibles Lernen
- Einsehbare Musterlösung

Gliederung

- Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- **Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben**
- Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgabe

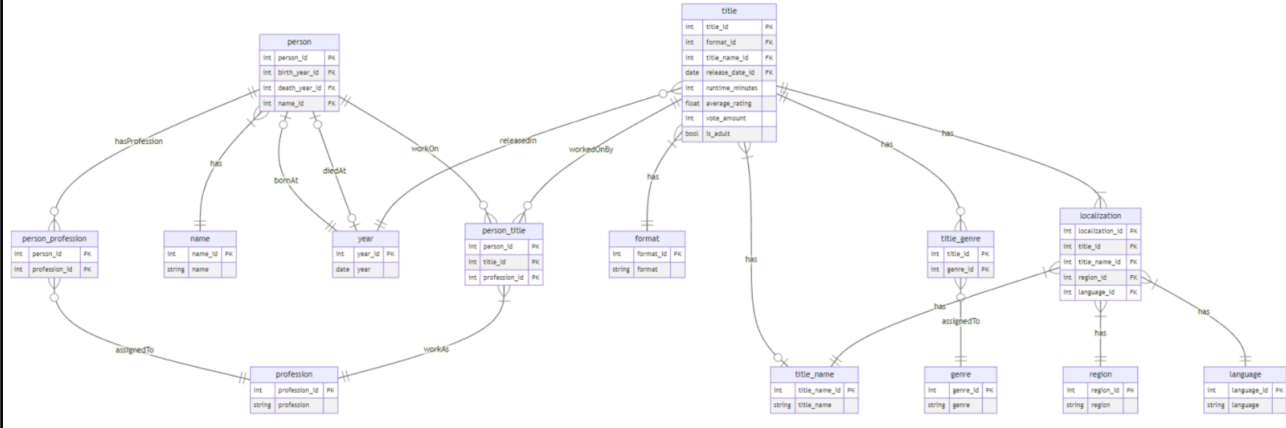
Aufgabenstellung:

Schreibe eine SQL-Abfrage, welche die im Text beschriebenen Informationen aus der darunter abgebildeten Datenbank extrahiert.


Textbeschreibung:

Find the count of title-ids, the runtime, the number of votes and the language-ids for all persons and the titles they were involved with, if the region contains 'ge' and the count of title_id is not '4883174'. Sort the result by the person-id in ascending order.

Datenbankschema:



```
SELECT COUNT(t.title_id), t.runtime_minutes, t.vote_amount, lo.
       language_id
FROM imdb2.person as p
INNER JOIN imdb2.year as y
ON p.birthyear_id = y.year_id
INNER JOIN imdb2.title as t
ON y.year_id = t.release_date_id
INNER JOIN imdb2.localization as lo
ON t.title_id = lo.title_id
INNER JOIN imdb2.region as r
ON lo.region_id = r.region_id
WHERE r.region LIKE 'ge%'
GROUP BY t.runtime_minutes, t.vote_amount, lo.language_id
HAVING COUNT(t.title_id) <> '4883174'
ORDER BY p.person_id ASC;
```



Zeig mir die Lösung

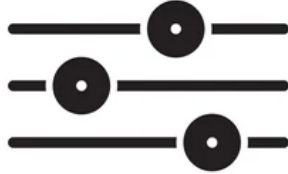
Kopieren!

Ausführen!

```
1 SELECT t.region_id, t.territory_id
2 FROM northwind.employee_territories as et
3 CROSS JOIN northwind.territories as t
4 WHERE t.region_id = '1'
5 GROUP BY t.region_id, t.territory_id;
```

Anforderungen an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

Parametrisierbarkeit
der SQL-Abfrage

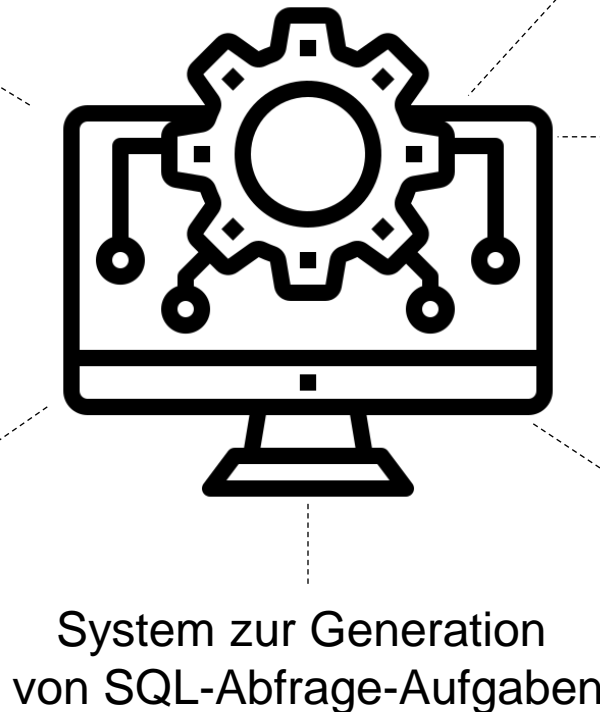


- SQL-Bestandteile
(bspw. WHERE-Clause)
- Constraint-Art
(bspw. BETWEEN numerisch)
- Anzahl Bestandteil
(bspw. 3-5 Constraints)

Performant

Plattformunabhängig

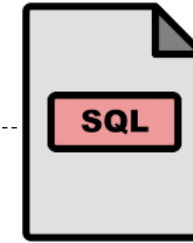
Konstant zugänglich



Eindeutig

Vollständig

Natürlichsprachige Beschreibung



SQL-Abfrage

Syntaktisch korrekt

Semantisch plausibel

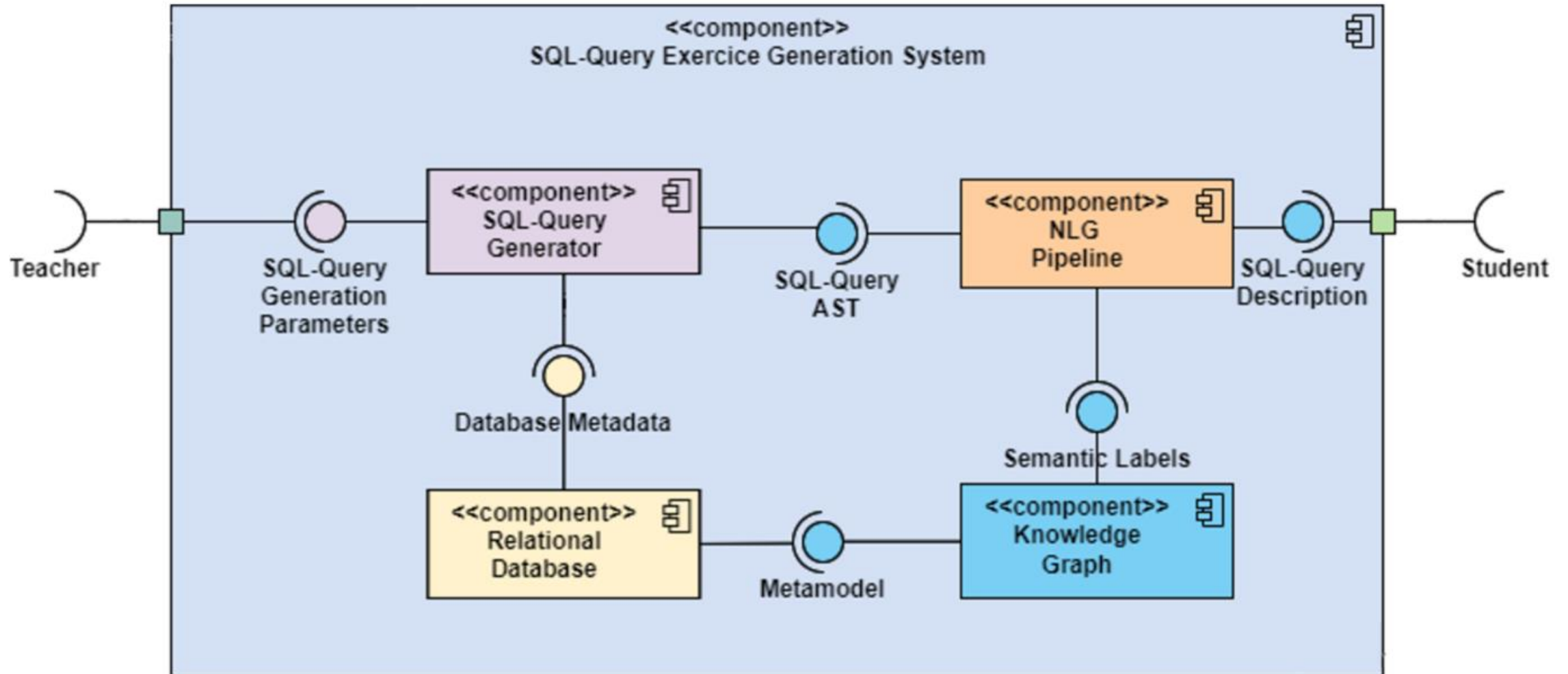


Bewertung von Lösungsversuchen

Gliederung

- Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- **Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben**
- Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

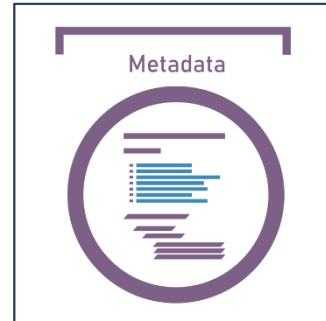
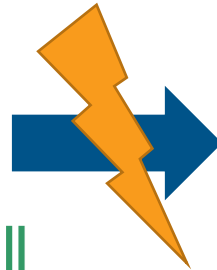
Übersicht über ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben



Experiment: Automatisches Ableiten domänenspezifischer Datenbanken aus bestehenden Wissensgraphen

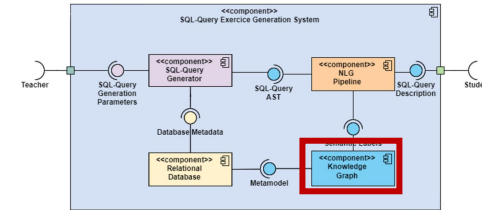
Wissensgraph

- Entitäten
- Instanzen
- Hyperonymiestruktur
- Meronymiestruktur



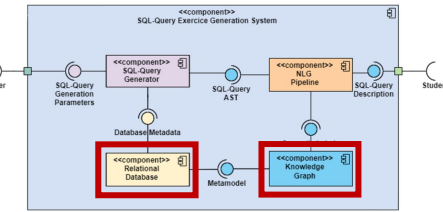
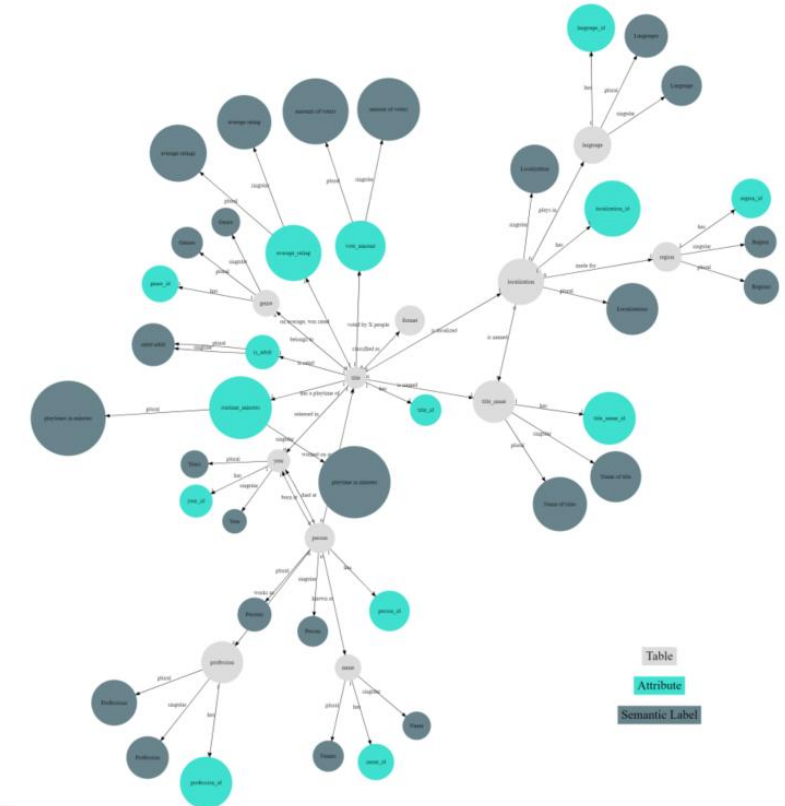
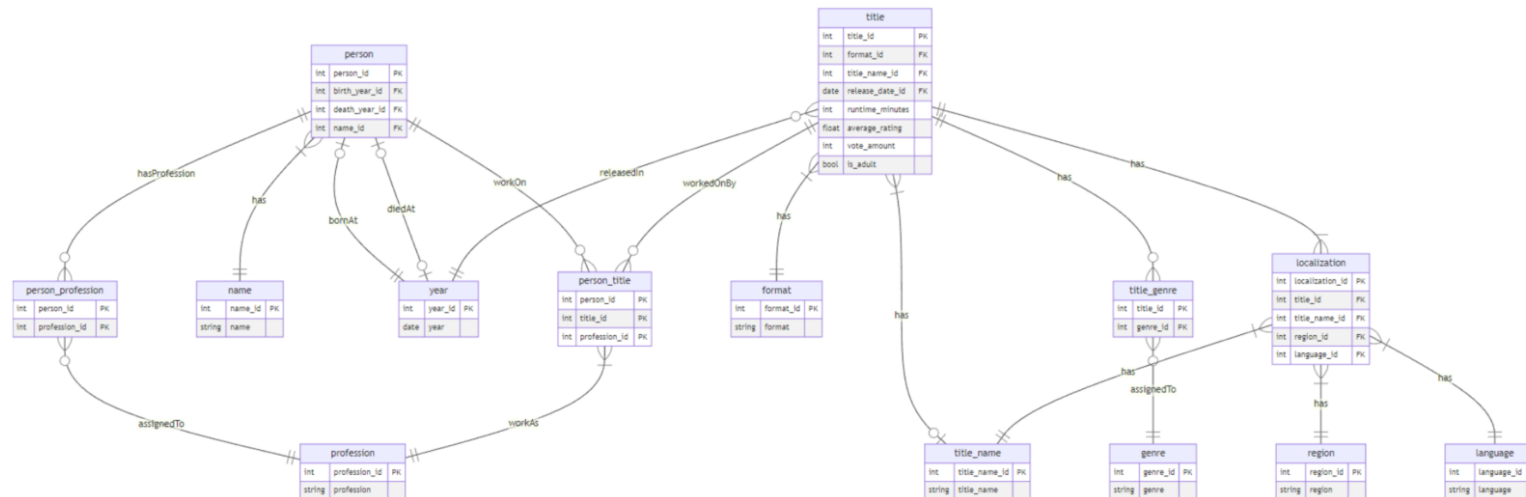
Limitationen:

- Automatisches Ableiten von Tabellenstrukturen entweder zu spezifisch oder zu generisch
- Attributzuordnung unklar, da auf Instanz-, anstelle Klassenebene
- Wann ist Entität eine Klasse/Instanz? Kontext wird nicht in Wikidata modelliert
- ...

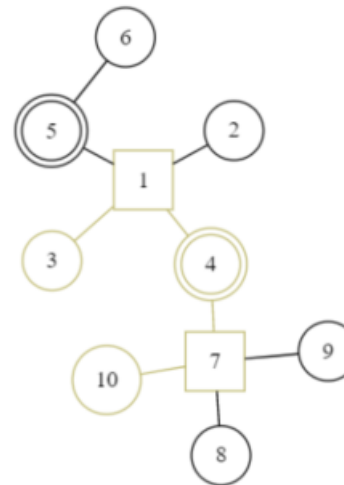
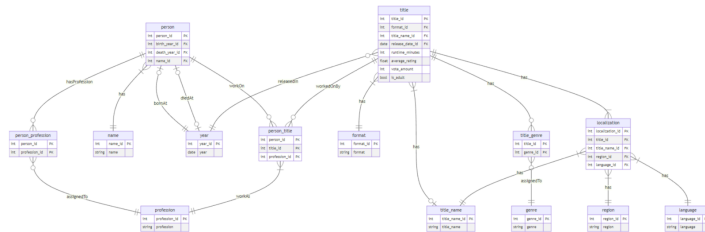
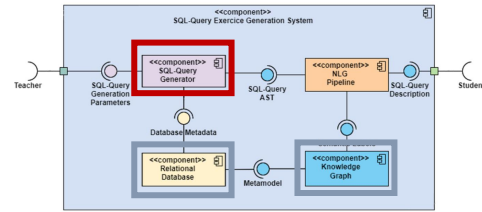


Ausweichlösung: Manuelle Erzeugung eines, zu einer existierenden Datenbank korrespondierenden, Wissensgraphen

Film- und Mediendatenbank



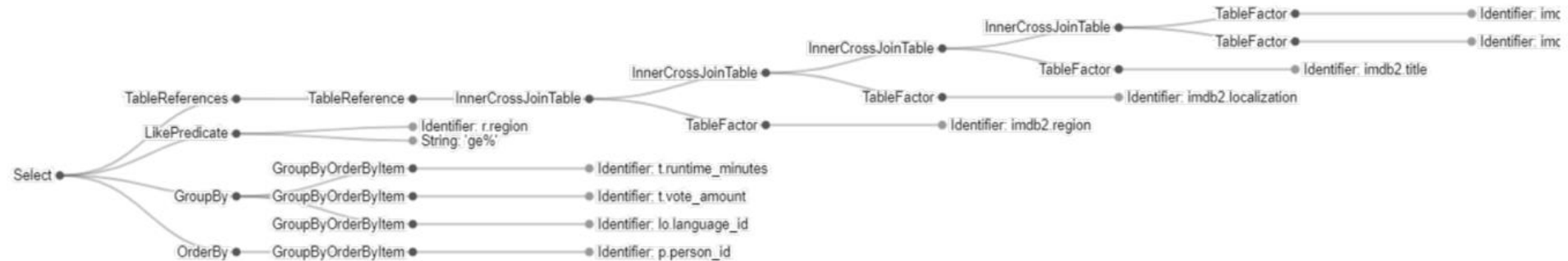
SQL-Abfrage-Generator



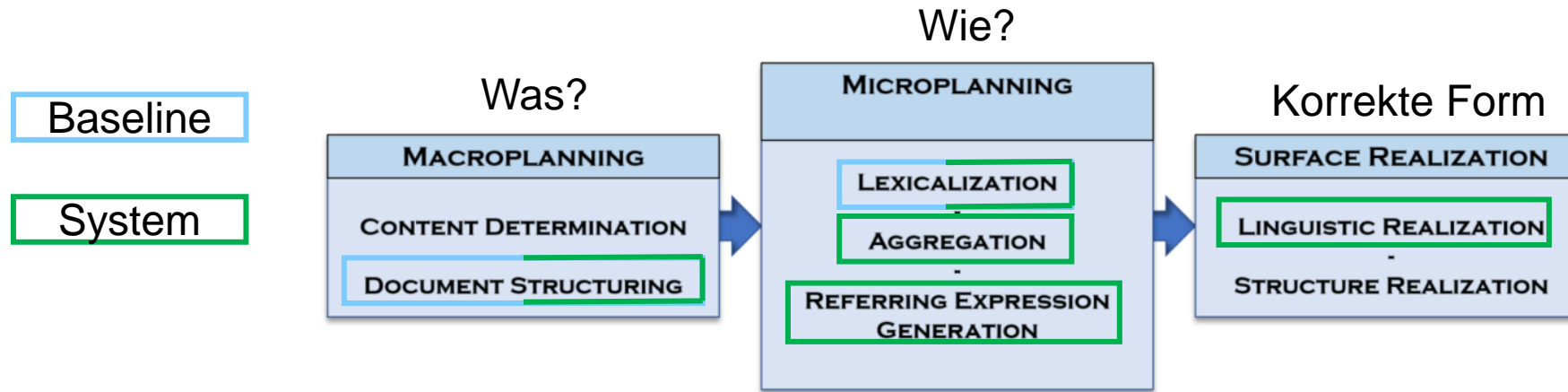
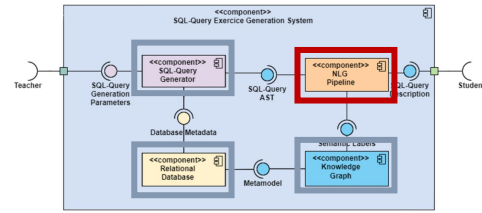
1. FROM

- 2. [WHERE CONSTRAINTS]
- 3. SELECT COLUMN
- 4. [GROUP BY]
- 5. [HAVING CONSTRAINTS]
- 6. [ORDER BY]

```
SELECT COUNT(t.title_id), t.runtime_minutes, t.vote_amount, lo.
    language_id
FROM imdb2.person as p
INNER JOIN imdb2.year as y
ON p.birthyear_id = y.year_id
INNER JOIN imdb2.title as t
ON y.year_id = t.release_date_id
INNER JOIN imdb2.localization as lo
ON t.title_id = lo.title_i
INNER JOIN imdb2.region as r
ON lo.region_id = r.region_id
WHERE r.region LIKE 'ge%'
GROUP BY t.runtime_minutes, t.vote_amount, lo.language_id
HAVING COUNT(t.title_id) <> '4883174'
ORDER BY p.person_id ASC;
```



Natural Language Generation Pipeline



It is [MASK] to
[MASK] that

Mask 1 Predictions:	Mask 2 Predictions:
11.5% important	22.9% say
11.1% difficult	15.5% note
8.8% easy	5.7% see
5.0% possible	3.3% suggest
3.5% hard	2.1% conclude

Form the intersection that contains the corresponding entries of the tables person and year and the intersection that contains the corresponding entries of the tables year and title and the intersection that contains the corresponding entries of the tables title and localization and the intersection that contains the corresponding entries of the tables localization and region. Return the columns the amount of title_id, runtime_minutes, vote_amount and language_id. Only return the data for which region contains 'ge'. A further constraint is the amount of title_id doesn't equal '4883174'. Group the result by runtime_minutes, vote_amount, title_id, language_id and person_id. Sort the result ascending by person_id.

Find the count of title-ids, the runtime, the number of votes and the language-ids for all persons and the titles they were involved with, if the region contains 'ge' and the count of title_id is not '4883174'. Sort the result by the person-id in ascending order.

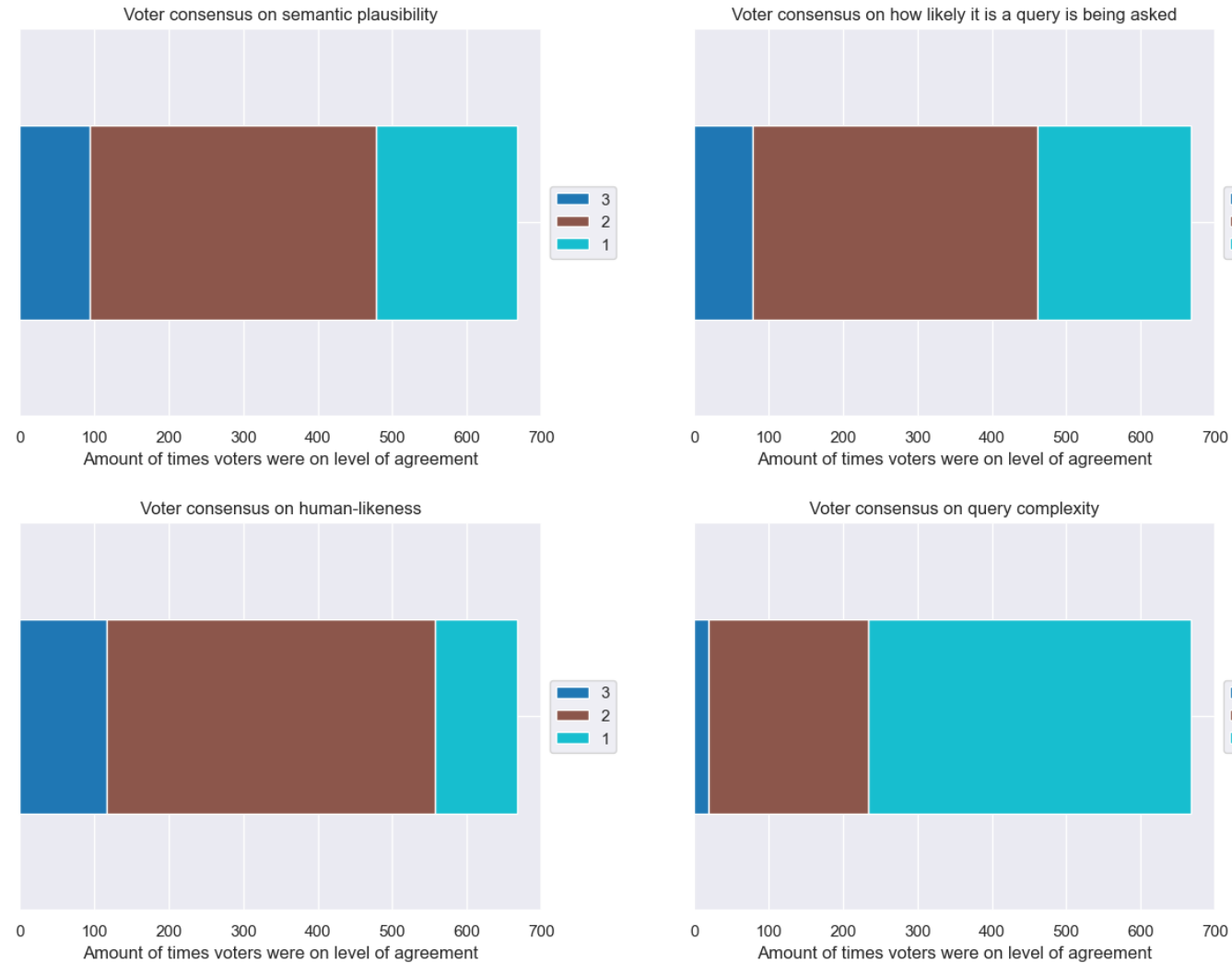
Gliederung

- Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- **Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben**
- Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

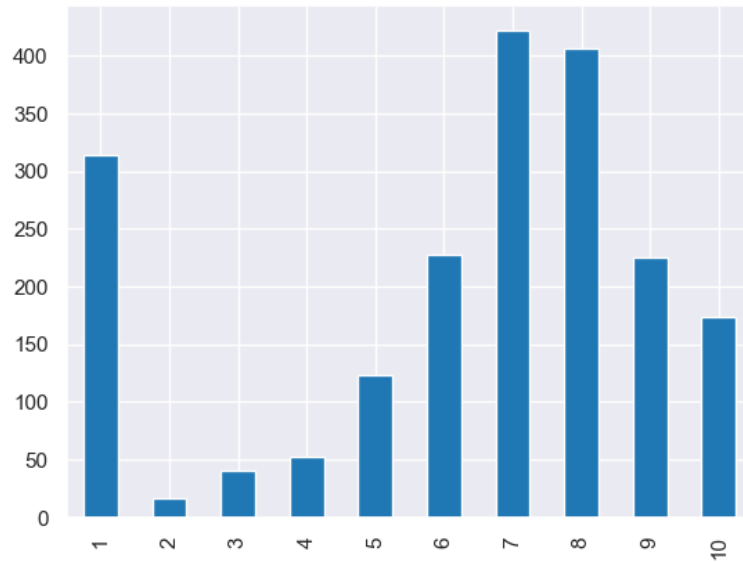
- Datensatz von 667 SQL-Abfragen
- Bewertung mittels Amazon Mechanical Turk (Crowdworking)
- Bewertung einer SQL-Aufgabe durch je 3 Crowdworker
 - Selbstangabe: SQL-Vorkenntnisse
- Survey-Input:
 - Datenbankschema
 - SQL-Abfrage
 - Natürlichsprachige Beschreibung
 - Baseline
 - System
- Survey-Fragen:
 - SQL-Abfrage
 - Komplexität
 - Semantische Plausibilität
 - Natürlichsprachige Beschreibung
 - Fehler (Stil, Fehlende Information, Verständlichkeit)
 - „Menschenähnlich“

Übereinstimmung der Crowdworker

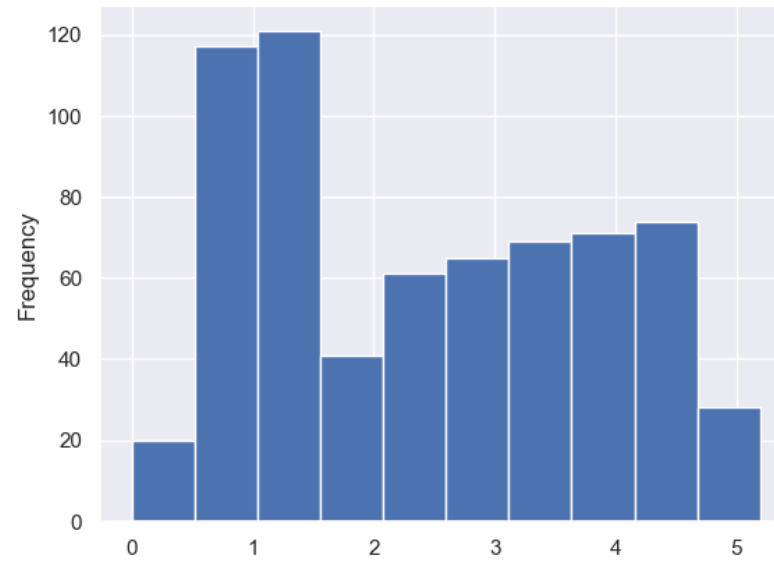


Komplexitätsbewertung der SQL-Abfragen

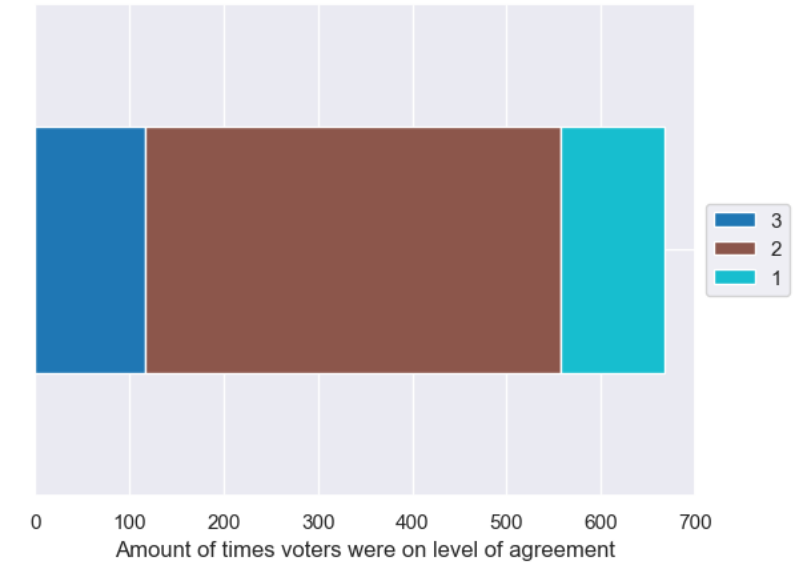
Global complexity scores



Standard deviation in complexity per query

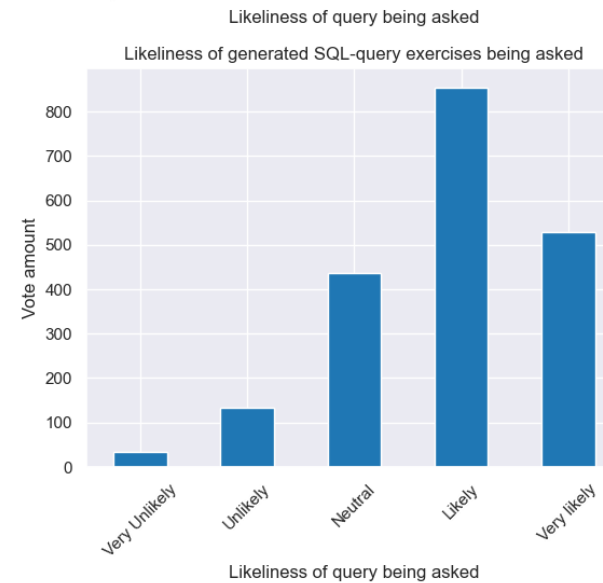
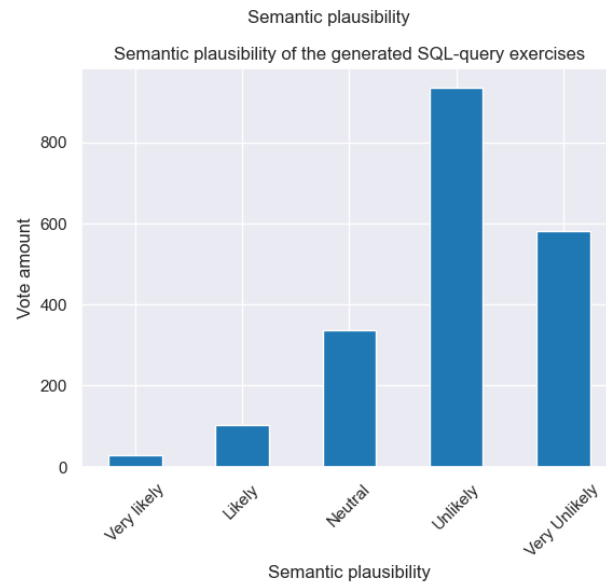
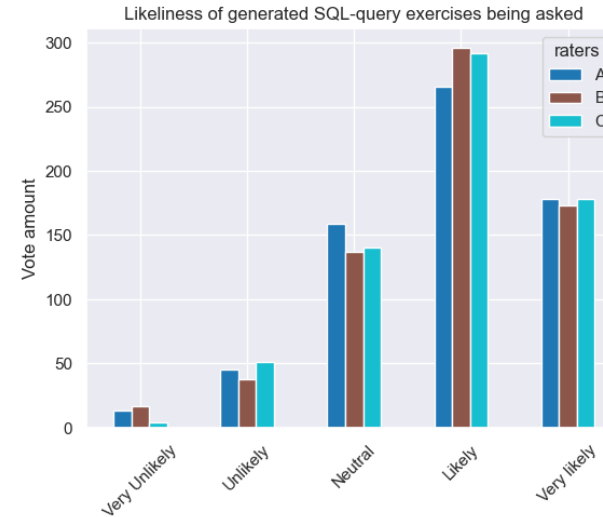
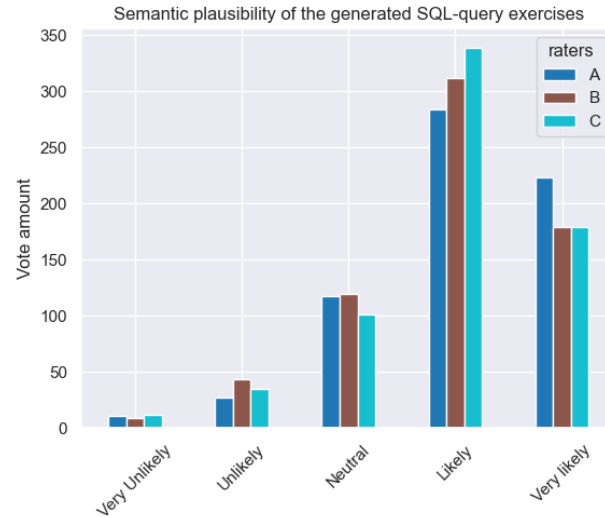


Voter consensus on query complexity with increased bin-size *

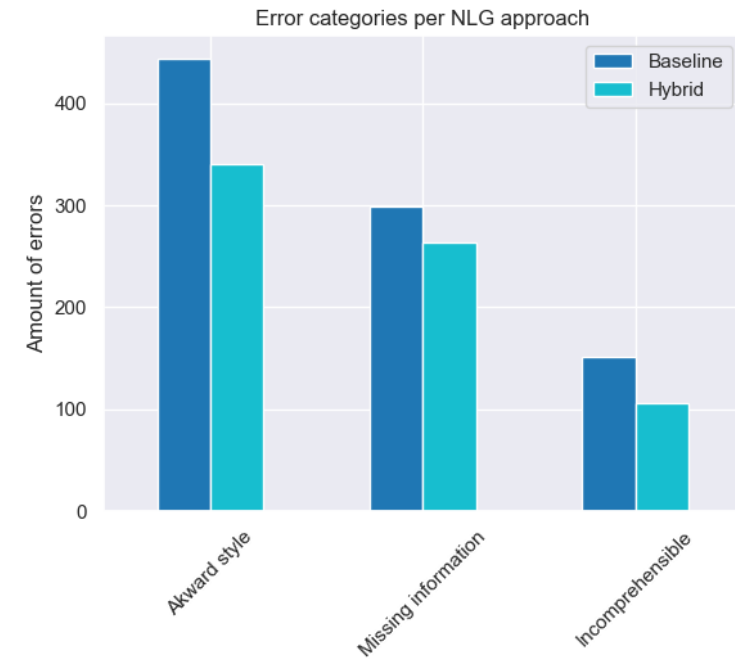
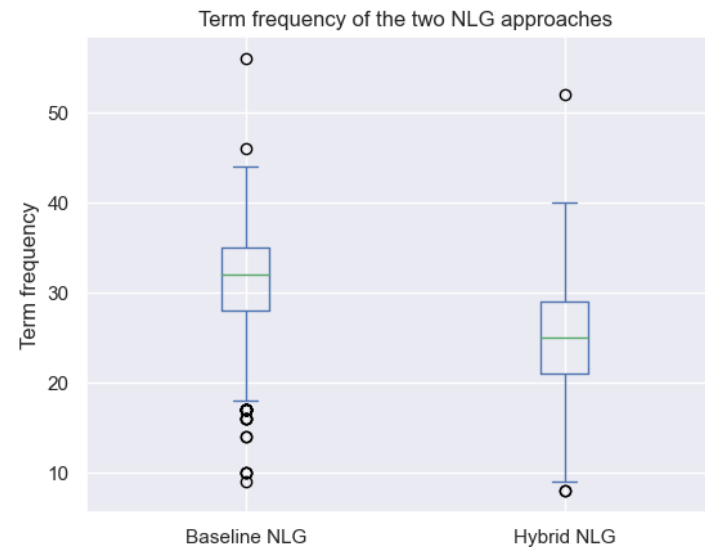
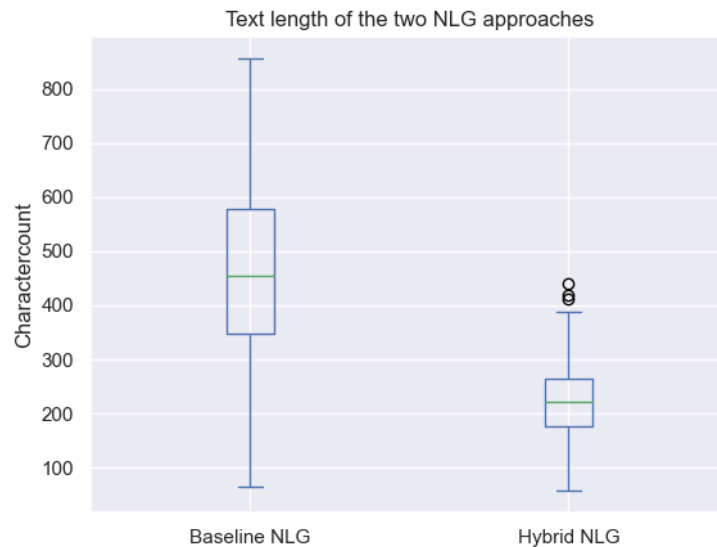


*Bins: [1,3], [4,6], [7,10]

Semantische Plausibilität der SQL-Abfragen



Vergleich der Ansätze zur Generierung natürlichsprachiger Beschreibungen von SQL-Abfragen



Limitation der Evaluation

- Unklarheit über tatsächliche Qualifikation der Crowdworker
- Fehlender Expertenabgleich
 - Inner-Rater-Agreement
- Keine Evaluation hinsichtlich pädagogischer Relevanz

Gliederung

- Motivation zur Entwicklung eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Anforderungen an eine bedeutungsvolle SQL-Abfrage-Aufgaben und an ein System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Entwurf und Implementation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- Evaluation eines Systems zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben
- **Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben**

Fazit zum entwickelten System zur Generierung bedeutungsvoller SQL-Abfrage-Aufgaben

- Anforderungen an das System grundlegend erfüllt, jedoch:
 - Natürliche Sprachgenerierung noch unzureichend
 - Automatische Bewertung von Lösungsversuchen nur mit binärer Entscheidung
- System übertrifft zwar Baseline und vorherige Systeme; dennoch hohe Anzahl an Fehlern in den Bewertungen der textuellen Beschreibung der SQL-Abfrage
- Zukünftige Arbeiten
 - Spiegeln der Semantik von SQL-Operationen in die Sprachgenerierung
 - Feingranulare Bewertung von Lösungsversuchen durch Normalisierung der Abfragen und Editier-Distanz zwischen den ASTs
 - Erweitern der unterstützten SQL-Syntax (DDL, etc.)
 - Evaluation in geeignetem Setting (Datenbankübung an Hochschulkursen)