Seminar CS715

Data Integration using Large Language Models
Hallo

- **Prof. Dr. Christian Bizer**
- **Professor for Information Systems V**
- **Research Interests:**
  - Web Data Integration
  - Data and Web Mining
  - Deployment of Data Web Technologies
- **Room:** B6 - B1.15
- **eMail:** christian.bizer@uni-mannheim.de
- **Consultation:** Wednesday, 13:30-14:30
Hallo

- M. Sc. Wi-Inf. Ralph Peeters
- Graduate Research Associate
- Research Interests:
  - Entity Matching using Deep Learning
  - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de
Hallo

- M. Sc. Wi-Inf. Alexander Brinkmann
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Product Data Categorization
- Room: B6, 26, C 1.04
- eMail: alexander.brinkmann@uni-mannheim.de
Hallo

- **M. Sc. Wi-Inf. Keti Korini**
- Graduate Research Associate
- Research Interests:
  - Table Annotation using Deep Learning
  - Schema Matching
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de
You and Your Experience

- A Short Round of Introductions
  - What are you studying?
  - Which DWS courses did you attend?
  - What kind of experience do you have with data science/data integration projects?

- Participants
  1. Baumann, Nick
  2. Chen, Chun-Yi
  3. Vogli, Aleksandro
  4. Golchha, Pujit
  5. Joseph, Abhay
  6. Chauhan, Vishal
  7. Steiner, Aaron
  8. Shyamsundar, Sharan
  9. Joo, Eun
  10. Lichwa, Mateusz
Agenda of Today’s Kickoff Meeting

1. Seminar organization
2. Topic Assignment
3. How to structure your seminar paper / presentation?
4. Your Questions
1. Seminar Organization
Learning Goals

- Writing a seminar thesis as an exercise for your master thesis
- Understanding and presenting state-of-the-art scientific work
- Designing experiments and present experimental results
- Searching and citing scientific papers / journal articles
- How to structure your thesis and presentation
- How to write a scientific paper using LaTeX
## Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday, <strong>02.03.2023</strong> (10:00-11:30)</td>
<td>Kick-off meeting and topic/mentor assignment</td>
</tr>
<tr>
<td></td>
<td>Read papers about your topic and search for additional literature</td>
</tr>
<tr>
<td></td>
<td>Design experimental setup (if applicable)</td>
</tr>
<tr>
<td></td>
<td>Prepare outline and argumentation line for the presentation</td>
</tr>
<tr>
<td><strong>Until 24.03.2023</strong></td>
<td>Meet with your mentor to discuss outline and argumentation</td>
</tr>
<tr>
<td></td>
<td>Prepare draft of your presentation</td>
</tr>
<tr>
<td><strong>Until 21.04.2023</strong></td>
<td>Send draft presentation to your mentor</td>
</tr>
<tr>
<td></td>
<td>Finalize your presentation</td>
</tr>
<tr>
<td><strong>Friday, 12.05.2023</strong> (10:00-12:30)</td>
<td>Presentation and discussion of your topic (30 % of your final grade)</td>
</tr>
<tr>
<td></td>
<td>Write seminar thesis</td>
</tr>
<tr>
<td><strong>Sunday, 09.07.2023</strong></td>
<td>Submission of your seminar thesis (70 % of your final grade)</td>
</tr>
</tbody>
</table>
Formal Requirements

- Presentation
  - 12 minutes + 8 minutes discussion
  - should be 100% understandable for all participants

- Written report (paper)
  - 12-15 pages single column
    - including abstract and appendixes
    - not including bibliography
    - every additional page reduces your grade by 0.3
  - written in English
  - use latex template of Springer Computer Science Proceedings

- Final grade
  - 70% written report
  - 30% presentation
Which template to use?

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
2. Seminar Topics and Topic Assignment
Data Integration using Large Language Models (LLM)

Symbolic Representations $\rightarrow$ Distributed Representations

Traditional Machine Learning $\rightarrow$ Deep Neural Networks

Large Language Models
Topics

1. Deep Learning for Table Search
   - Student: Vogli, Aleksandro
   - Mentor: Ralph Peeters

2. Experimental Topic: Evaluating Large Language Models on the Task of Entity Matching

- Student: Steiner, Aaron
- Mentor: Ralph Peeters


3. Representation Learning for Missing Value Imputation

- Student: Joseph, Abhay

- Mentor: Alexander Brinkmann


4. Experimental Topic: Evaluating Large Language Models on the Task of Missing Value Imputation for Knowledge Graph Completion

- Student: Chen, Chun-Yi
- Mentor: Alexander Brinkmann

- https://paperswithcode.com/task/knowledge-graph-completion
5. Schema Matching using Deep Learning

- Student: Golchha, Pujit
- Mentor: Keti Korini

6. Cell Entity Annotation in Tabular Data

- Student: Lichwa, Mateusz
- Mentor: Keti Korini

- More references and benchmarks: Papers with Code: Cell Entity Annotation
Topics

7. Deep Tabular Learning for Domain-Specific Prediction Tasks

- Student: Shyamsundar, Sharan
- Mentor: Keti Korini

8. Information Extraction for E-Commerce Product Data

- Student: Baumann, Nick

- Mentor: Alexander Brinkmann

• Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022: OA-Mine: Open-World Attribute Mining for E-Commerce Products with Weak Supervision. WWW 2022, 3153–3161.

• Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title. ACL 2019, 5214–5223.

Topics


- Student: Chauhan, Vishal
- Mentor: Alexander Brinkmann

10. Experimental Topic: Combining WebAPIs and Large Language Models for Question Answering via In-Context Learning

- Student: Joo, Eun
- Mentor: Ralph Peeters

- Example Task: Ask ChatGPT or GPT3 questions about restaurants or hotels in Mannheim using TripAdvisor data and in-context learning.
3. How to Structure Your Paper / Presentation
Goals of Literature and Experimental Papers

- **Goals of Literature Papers**
  1. describe the problem / task
  2. describe several existing methods/systems for handling the task,
  3. compare the methods/systems and their evaluation using a systematic set of comparison criteria

- **Goals of Experimental Papers**
  1. describe the problem / task
  2. design an experimental setup for evaluating LLM on the task
  3. present and discuss the results of your experiments
How to Structure Your Literature Paper?

1. Introduction and Problem Statement
   • Which problem/task is addressed? Why is the problem important?
   • Structure of your paper

2. Description of Existing Approaches
   • Overview of existing methods and features used by the methods
   • Detailed description of selected methods (likely two)
   • Comparison of the selected methods using a set of comparison criteria

3. Evaluation
   • Comparison and discussion of the evaluation tasks, metrics
   • Comparison of the evaluation results using a set of comparison criteria

4. Conclusion
   • What did the comparison of the methods and evaluation results show?
   • Can something be concluded for future work?

5. Bibliography
How to Structure Your Experimental Paper?

1. Introduction and Problem Statement
   • Which problem/task is addressed? Why is the problem important?
   • Overview of existing methods and benchmarks used for evaluation
   • Structure of your paper

2. Description of Experimental Design
   • How do you select examples for which challenges from which benchmark?
   • Which prompt designs do you test?
   • Which Large Language Models do you benchmark?

3. Presentation of Experimental Results
   • Present the results of your experiments (tables containing values and deltas).
   • Present the results of your error analysis (types of errors, frequency of these types)

4. Conclusion
   • What did the experiments and the error analysis show?
   • What can be concluded for future work?

5. Bibliography
Learn from Examples

- Read survey articles and previous experimental papers and identify the structure from the previous slides
  • Why can this paragraph be found at that position?
  • What is the purpose of some section / subsection?

- Important
  • Read survey articles!
  • Read conference or journal papers

- Textbook on how to write a thesis

- University Library: Academic Writing Consultancy
  • https://www.bib.uni-mannheim.de/en/writing-consultancy/
Citing different Types of Publications

- **Journal article**
  - Good to cite, current research results
  - Survey articles (very good for an overview)

- **Conference and workshop paper**
  - Good to cite, current research results

- **Books (sometimes cited)**
  - Textbooks
  - Collections of articles/papers => Cite specific paper in book

- **Websites**
  - better not cited, exceptions are, e.g., documents like W3C Specifications
  - *Do not cite Wikipedia, ever!*
  - Use footnotes to refer to project pages, download pages, or technical documentation

- **Slide sets (especially from our lectures)**
  - Never cite!
How to Find Relevant Publications?

- Use Standard Search Engines
- **Use Google Scholar**
  - we use it a lot ourselves
- Search Engines of the University’s library
  - see slides from the library course
- **Exploit references:** Given a relevant document $x$
  - Follow references in the past: papers $y$ that $x$ has cited
  - Follow references in the future: papers $y$ that cited $x$
    ("cited by" functionality in Google scholar)
4. Questions?