enRichMyData

SINTEF *

^ INSID&S Lab

Department of Informatics, Systems and Communication University of Milan - Bicocca

Semantic Data Enrichment: from Interactive Exploration to Scalable Deployment

Roberto Avogadro *, Flavio De Paoli ^, Dumitru Roman *, Matteo Palmonari ^

Part V – Conclusions

*This work presented in this presentation has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No 732590 - **EW-Shopp** - and No 732003 – **euBusinessGraph** - and from the European Union's Horizon Europe research and innovation program under grant agreements No 101070284 - **enRichMyData**.*

Conclusions #1

- Semantic data enrichment
  - Knowledge graphs and semantic techniques to support data preparation for AI
  - Demand in the industry
  - The *Link & Extend* paradigm
    - Developer POV: service-based interoperability and architecture
    - Publisher POV: publising data not enough: reconciliation + data extension services
  - Interactive exploration to understand, design, consfigure and refine pipelines
  - Scalable deployment solutions
  - Ongoing work in the enRichMyData project
    - Comprehensive toolkit, covering also enrichment of textual data
    - Updates at https://enrichmydata.eu/

Conclusions #2

- Key techniques for link & extend
  - Tabular data annotation
    - Specific: heuristic-based and feature-based ML approaches, e.g., Alligator
    - Generalistic: LLM-based approaches, e.g., TURL, TableLlama
    - Challenges vs LLM-based approaches
      - Performancei
      - Cross-dataset generalization (?)
      - Scalability and costs
      - Interpretability
- Ongoing work towards better solutions presented in the tutorial
  - SemTUI: interactive data enrichment tool
  - Scalable deployment of pipelines designed with Argo Workflows and TAO

Discussion and Open Challenges

- How to use LLMs for table annotation and enrichment tasks with better scalability and sustainability?
  - Unclear if they beat heuristic and fine-tuned feature-based approaches on every dataset
  - Fine-tuning on specific tables with limited data
  - AI-in-the-loop? When to use large LLMs wisely?
- LLMs for prompt-based data enrichment
  - Several ongoing work with text-to-code approaches
  - Still limited application to enrichment with third-party sources
- Can LLMs bring semantic services back again?
  - Agile interoperable solutions with LLMs vs. ontology-based annotations

References

Data Enrichment

- [Bucher et al. 2021] Bucher, T. C., Jiang, X., Meyer, O., Waitz, S., Hertling, S., & Paulheim, H. (2021). scikit-learn Pipelines Meet Knowledge Graphs: The Python kgextension Package. ESWC Satellite Events: Revised Selected Papers 18 (pp. 9-14).
- [Hameed & Naumann 2020] Hameed, M., & Naumann, F. (2020). Data preparation: A survey of commercial tools. ACM SIGMOD Record, 49(3), 18-29.
- [Harari & Katz 2022a] Harari, A., & Katz, G. (2022). Automatic features generation and selection from external sources: a DBpedia use case. Information Sciences, 582, 398-414.
- [Harari & Katz 2022b] Harari, A., & Katz, G. (2022, May). Few-shot tabular data enrichment using fine-tuned transformer architectures. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1577-1591).
- [Lehmberg et al 2016] Lehmberg, O., Ritze, D., Meusel, R., & Bizer, C. (2016, April). A large public corpus of web tables containing time and context metadata. In Proceedings of the 25th international conference companion on world wide web (pp. 75-76).
- [Nawaz et al. 2022] Nawaz, M. S., Khan, S. U. R., Hussain, S., & Iqbal, J. (2022). A study on application programming interface recommendation: state-of-the-art techniques, challenges and future directions. Library Hi Tech, 41(2), 355-385.
- [Nozza et al. 2017] Nozza, D., Ristagno, F., Palmonari, M., Fersini, E., Manchanda, P., & Messina, E. (2017, April). TWINE: A real-time system for TWeet analysis via INformation Extraction. In EACL Demos (pp. 25-28).

Solutions for Scalable Data Transformations

[Dessalk et al. 2020] Dessalk, Y. D., Nikolov, N., Matskin, M., Soylu, A., & Roman, D. (2020, November). Scalable execution of big data workflows using software containers. In Proceedings of the 12th International Conference on Management of Digital EcoSystems (pp. 76-83).

[Liu et al. 2011] Liu, X., Thomsen, C., & Pedersen, T. B. (2011). ETLMR: a highly scalable dimensional ETL framework based on MapReduce. In *Data Warehousing and Knowledge Discovery: 13th International Conference, DaWaK 2011, Toulouse, France, August 29-September 2, 2011. Proceedings 13* (pp. 96-111). Springer Berlin Heidelberg.

[Sakr & Sherif 2011] Sakr, Sherif, et al. "A survey of large scale data management approaches in cloud environments." IEEE communications surveys & tutorials 13.3 (2011): 311-336.

Tabular Data Annotation

- [Avogadro et al. 2022] Avogadro, R., Cremaschi, M., D’adda, F., De Paoli, F., Palmonari, M.: Lamapi: a comprehensive tool for string-based entity retrieval with type-base filters. (2022) In OM 2022.
- [Avogadro et al. 2023] Avogadro, R., Ciavotta, M., De Paoli, F., Palmonari, M., Roman, D.: Estimating link confidence for human-in-the-loop table annotation. (2023) In WI-IAT. pp. 142–149
- [R. Avogadro 2024] Semantic Enrichment of Tabular Data with Machine Learning Techniques. PhD Thesis. Available at https://boa.unimib.it/handle/10281/465138
- [Cremaschi et al. 2022] Cremaschi, M., Avogadro, R., Chieregato, D.: s-elbat: a semantic interpretation approach for messy table-s. (2022). In SemTab 2022
- [Sarthou-Camy et al. 2022] Sarthou-Camy, C., Jourdain, G., Chabot, Y., Monnin, P., Deuzé, F., Huynh, V. P., ... & Troncy, R. (2022, May). DAGOBAH UI: a new hope for semantic table interpretation. In ESWC Demo (pp. 107-111).
- [Dasoulas et al. 2023] Dasoulas, I., Yang, D., Duan, X., Dimou, A.: Torchictab: Semantic table annotation with wikidata and language models. In: CEUR Workshop Proceedings. pp. 21–37. CEUR Workshop Proceedings (2023)
- [Korini & Bizer 2024] Korini, K., Bizer, C.: Column type annotation using chatgpt. ESWC 2024. arXiv preprint arXiv:2306.00745 (2023)
- [Deng et al. 2024] Deng, X., Sun, H., Lees, A., Wu, Y., Yu, C.: Turl: Table understanding through representation learning. ACM SIGMOD Record 51(1), 33–40 (2022)
- [Suhara et al. 2022] Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, c., Chen, C., Tan, W.C.: Annotating columns with pre-trained language models. In: Proceedings of the 2022 International Conference on Management of Data. p. 1493–1503. SIGMOD ’22, Association for Computing Machinery, New York, NY, USA (2022)
- [Huynh et al. 2022] Huynh, V. P., Chabot, Y., Labbé, T., Liu, J., & Troncy, R. (2022). From heuristics to language models: A journey through the universe of semantic table interpretation with dagobah. In SemTab@ISWC 2022: 45-58.

Related Tasks

- [Peeters & Bizer 2023] Peeters, R., & Bizer, C. (2023, August). Using chatgpt for entity matching. In *European Conference on Advances in Databases and Information Systems* (pp. 221-230). Cham: Springer Nature Switzerland.