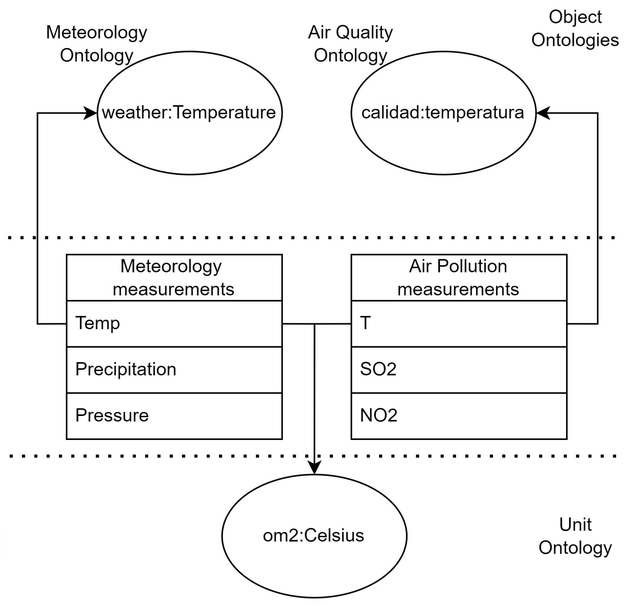
Generating Semantic Context for Data Interoperability in Relational Databases using BGE M3-Embeddings

Bugra Altug, Martin Weise, Andreas Rauber
bugraaaltug@gmail.com, martin.weise@tuwien.ac.at, andreas.rauber@tuwien.ac.at

Institute of Information Systems Engineering

Semantic Context

- Machine-understandable context provides findability, accessibility, interoperability and reusability in data repositories (DBRepo) [1]
- Data within relational databases can:
  - Use custom object ontologies from scientific domains
  - Interoperate with one another through a shared unit ontology [2]



Semi-Automatic Mapping

- Mapping attributes to ontological concepts capturing their semantics and measurement units
- Encoding is done by BGE M3-Embeddings[3]:
  - "unit" keyword is added to the unit entity labels
  - Entity labels and column names are converted into embedding vectors
- Similarity score:
  - Entity score: Cosine similarity between an entity's and a column's embedding
  - Ontology score: Average of the highest entity score for each column
- Highest ontology score and its highest entity scores are suggested to the user

Object Mapping Results

- Two evaluation rounds with validation and test datasets. **219** columns have a target object entity
- **1,932** entities from 6 object ontologies across different scientific domains
- **42%** automatically mapped, **45.4%** required a correction within the top-96 suggestions

1st round

	1st (%)	1st - 5% (%)	5% - 10%	MRR
Edit distance n = 1	26.9	40.7	8.3	0.3441
Edit distance n = 2	27.8	45.4	0.9	0.3776
Jaro–winkler	26.9	34.3	6.5	0.3363
Jaccard index	13.9	**47.2**	**10.2**	0.1857
Longest common subsequence	21.3	42.6	6.5	0.2624
Cosine Similarity with BGE	**43.5**	45.4	0.9	**0.5259**
Cosine Similarity Nasa-SMD[4]	40.7	45.4	5.6	0.5093

Micro-average for measuring the system's overall performance

	1st (%)	1st - 5% (%)	5% - 10%	MRR
Edit distance n = 1	32.3 +/- 8.3	35.2 +/- 3	7.3 +/- 0.5	0.384 +/- 0.077
Edit distance n = 2	33 +/- 8.2	39.5 +/- 5.3	1.3+/- 0.1	0.407 +/- 0.075
Jaro–winkler	32.4 +/- 9.2	31 +/- 5.5	4.6 +/- 0.4	0.377 +/- 0.089
Jaccard index	17.4 +/- 5.7	**43.4 +/- 5.3**	**10.4 +/- 1.2**	0.223 +/- 0.071
Longest common subsequence	26.2 +/- 7.8	37.1 +/- 6.3	5.8 +/- 0.4	0.304 +/- 0.074
Cosine Similarity with BGE	**51.4 +/- 9.8**	37.9 +/- 7.4	1.3 +/-0.1	**0.593 +/- 0.077**
Cosine Similarity Nasa-SMD	43.9 +/- 4.4	42.7 +/- 1.7	6.1 +/- 0.9	0.538 +/- 0.044

Macro-average for measuring the user experience

	Without removing the columns that are not grounded			With removing		
	(Micro) Average Clicks	(Macro) Average Clicks	Correct Ontology	(Micro) Average Clicks	(Macro) Average Clicks	Correct Ontology
Edit distance n = 1	1.98	1.998 +/- 0.639	6/10	1.98	2.044 +/- 1.024	6/10
Edit distance n = 2	2.06	2.097 +/- 1.15	6/10	2.07	2.168 +/- 1.814	6/10
Jaro–winkler	2.38	2.197 +/- 1.003	6/10	2.38	2.3 +/- 1.589	6/10
Jaccard index	2.69	2.598 +/- 0.712	**5/10**	2.88	2.856 +/- 1.29	4/10
Longest common subsequence	2.35	2.328 +/- 0.781	5/10	2.44	2.49 +/- 1.341	5/10
Cosine Similarity with BGE	**1.51**	**1.379 +/- 0.393**	7/10	**1.36**	**1.17 +/- 0.676**	9/10
Cosine Similarity Nasa-SMD	1.54	1.537 +/- 0.356	**8/10**	1.4	1.408 +/- 0.545	**10/10**

Average character length of the correct object entities is **13.4**

2nd round

- **1st**: 41.1% (41.0 +/- 11.4)
- **1st - 5%**: 45.5% (47.0 +/- 8.6)
- **5% - 10%**: 1.8% (1.7 +/- 0.1)
- **MRR**: 0.486 (0.49 +/- 0.09)
- Without removal: 7/10 **cases** with 1.779 (1.603 +/- 0.987) **clicks**
- With removal 8/10 **cases** with 1.723 (1.6 +/- 0.980) **clicks**

Average character length of the correct object entities is **14.402**

Unit Mapping Results

- Two evaluation rounds. **173** columns have a target unit entity
- **3, 811** entities from 3 unit ontologies
- **10.9%** automatically mapped, **51.4%** required a correction within the top-381 suggestions. Object feedback provided additional **19.9**% to top-381.

1st round

	1st (%)	1st - 5% (%)	5% - 10%	Average	MRR
Edit distance n = 2	**6.9**	16.1	8	5.77	0.0839
Nasa-SMD	1.1	**50.6**	5.7	6.23	0.0615
BGE	5.7	41.4	**14.9**	5.781	0.1049
BGE + unit keyword	**6.9**	47.1	10.3	**5.745**	**0.1164**

Average character length of the correct unit entities is **14.26**

MRR increases up to **0.139** with user provided object feedback

	1st (%)	1st - 5% (%)	5% - 10%	Average	MRR
Edit distance n = 2	7.3 +/- 0.9	16.3 +/- 3.1	10.0 +/- 1.6	**5.667 +/- 2.454**	0.092 +/- 0.008
Nasa-SMD	2.0 +/- 0.4	**51.4 +/- 7.5**	5.2 +/- 0.5	6.272 +/- 8.705	0.077 +/- 0.006
BGE	8.6 +/- 0.9	36.8 +/- 4.5	**14.9 +/- 1.4**	5.984 +/- 14.75	0.118 +/- 0.01
BGE + unit keyword	**9.2 +/- 0.9**	46.0 +/- 2.7	9.7 +/- 1.7	5.791 +/- 12.236	**0.13 +/- 0.012**

- Embedding models tends to favor composite units: "gram per day" is favored instead of "gram".
- Without the composite units the BGE model has **4.793** user clicks while the edit distance method has **5.819**.
- Object feedback: When a user selects an object entity, the embedding vectors of the column name and the entity are averaged. This new embedding is then used to calculate similarity.
- Column names can contain relevant information such as hyponym synonym, and unit abbreviations.

2nd round

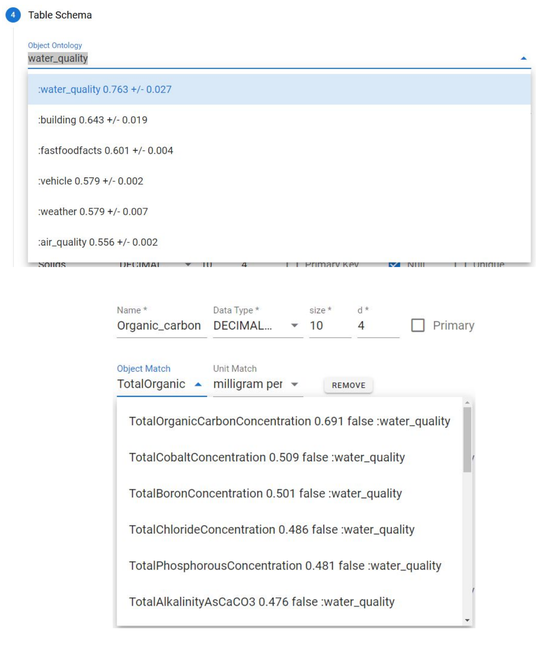
- **1st**: 14.9% (14.0 +/- 6.6)
- **1st - 5%**: 31% (33.7 +/- 6.7)
- **5% - 10%**: 14.9% (14.4 +/- 2.6)
- **MRR**: 0.1861 (0.177 +/- 0.075)
- **Average user clicks**: 6.162 (6.029 +/- 17.612)

Average character length of the correct object entities is **14.552**

MRR increases up to **0.203** with user provided object feedback

Conclusion

- Semantic context and interoperability can be achieved by mapping data to object and unit entities
- New UI [1] can generate:
  - Object mappings **8.36** times faster in terms of number of clicks compared to manual
  - Unit mappings **2.36** times faster in terms of number of clicks compared to manual
- Unified usage of embedding model provides user feedback to improve unit suggestions



References

[1] Weise, M., Staudinger, M., Michlits, C., Gergely, E., Stytsenko, K., Ganguly, R., & Rauber, A. (2022). DBRepo: a Semantic Digital Repository for Relational Databases. International Journal of Digital Curation, 17(1), 11. DOI: 10.2218/ijdc.v17i1.825

[2] Magagna, B., Rosati, I., Stoica, M., Schindler, S., Moncoiffe, G., Devaraju, A., Peterseil, J., & Huber, R. (2021). The i-adopt interoperability framework for fairer data descriptions of biodiversity. arXiv, DOI: 10.48550/arXiv.2107.06547.

[3] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. & Liu, Z. (2024). BGE M3-Embedding: Multi-lingual, multi-functionality, multigranularity text embeddings through self-knowledge distillation. arXiv, DOI: 10.48550/arXiv.2402.03216.

[4] NASA-IMPACT. (2024). nasa-smd-ibm-st-v2 (revision d249d84) DOI: 10.57967/hf/1800.