Pan *et al. Energy Informatics* 2023, **6**(Suppl 1):22
https://doi.org/10.1186/s42162-023-00277-0

Energy Informatics

RESEARCH **Open Access**

Schema matching based on energy domain pre-trained language model

Zhiyu Pan[1*], Muchen Yang[1] and Antonello Monti[1,2]

From The 12th DACH+ Conference on Energy Informatics 2023
Vienna, Austria. 4-6 October 2023. https://www.energy-informatics2023.org/

*Correspondence:
zhiyu.pan@eonerc.rwth-aachen.de

[1] Institute for Automation of Complex Power Systems, RWTH Aachen University, Mathieustraße 10, 52074 Aachen, Germany
[2] Fraunhofer FIT, Schloss Birlinghoven, 53757 Sankt Augustin, Germany

Abstract

Data integration in the energy sector, which refers to the process of combining and harmonizing data from multiple heterogeneous sources, is becoming increasingly difficult due to the growing volume of heterogeneous data. Schema matching plays a crucial role in this process by giving each representation a unique identity by matching raw energy data to a generic data model. This study uses an energy domain language model to automate schema matching, reducing manual effort in integrating heterogeneous data. We developed two energy domain language models, Energy BERT and Energy Sentence Bert, and trained them using an open-source scientific corpus. The comparison of the developed models with the baseline model using real-life energy domain data shows that Energy BERT and Energy Sentence Bert models significantly improve the accuracy of schema matching.

Keywords: Pre-trained language model, Schema matching, Energy domain

Introduction

Data sharing between different energy sectors is essential to improve the efficiency of modern energy systems. However, the energy sector generates large amounts of data from various sources, including sensors, devices and systems that can have different formats, structures, and representations (Malarvizhi and Kalyani 2013; Zhang et al. 2023). As proposed in the EU Data Act, industrial data should be shared, stored and processed in a fair and secure way. It is an essential step to share the data among the whole energy sector by harmonizing the data schema into a common data representation (Sayah et al. 2021), which requires the involvement of schema matching to map the data from the different sources to a generic data representation. The matching of data between different companies is a challenging task, which costs both time and human resources. Manually matching, as the current solution, while generally providing high accuracy, usually needs collaboration between organizations (Sutanta et al. 2016), which can be quite time-consuming when dealing with huge varieties of data. A common data representation for data sharing is crucial for decision making and analysis, as well as for the development

of new energy technologies and solutions. By automating the schema matching process, the accuracy and efficiency of data integration can be improved, leading to better insights and outcomes in the energy sector.

Schema matching aims to find semantic correspondence between elements of two schemas. To illustrate the process of schema matching, two datasets are matched with data models in Fig. 1. The state-of-the-art automatic schema matching technology (Fernandez et al. 2018; Pan et al. 2022) typically involves the use of machine learning algorithms, such as deep learning neural networks, to identify and match similar structures in energy-related datasets. These technologies often rely on graph-based representations of schemas and can leverage additional information such as ontologies and contextual data to enhance their accuracy. There are several ways of implementing automatic schema matching. A taxonomy covering a number of existing approaches is presented in Rahm and Bernstein (2001). Afterward, a new semantic-based schema matching approach using WordNet was proposed in Giunchiglia et al. (2007). SEMPROP (Fernandez et al. 2018), which is based on syntactic and semantic similarities using word embeddings and proposed by Raul Castro Fernandez et al, was proved to work better than other state-of-the-art automatic schema matching methods.

Unfortunately, word embedding cannot distinguish the word with multiple meanings. However, with the recent breakthrough of natural language processing (NLP), the BERT (Bidirectional Encoder Representations from Transformers) pre-training model (Kenton and Toutanova 2019) in 2018 changed the traditional standard one-directional training model to use a deep bidirectional transformer. It is designed to understand and generate human-like text by capturing contextual relationships between words in a given sentence. BERT learns information from the left and right contexts of the target word. In contrast to the previous word embedding model, BERT captures the different meaning for the same word in different contexts.

In the past three years, BERT has developed rapidly with variants such as BioBert (Lee et al. 2020), SciBert (Beltagy et al. 2019) and RoBERTa (Liu et al. 2019), which have made great improvements in terms of size and structures according to domain-specific application. Unfortunately, most of the research focuses on biomedical, computer science,

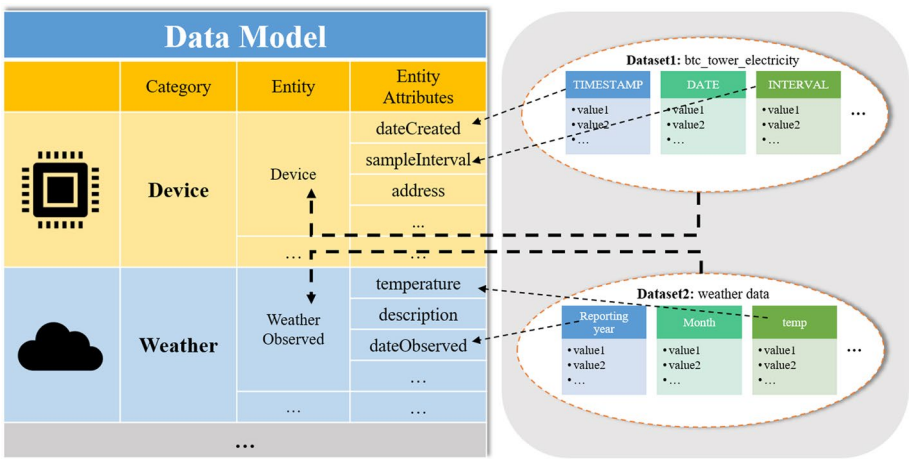

Fig. 1 Schema matching example

and social media, neglecting the exploration of energy domain. Additionally, the original BERT language model is focused on text-mining tasks: Named Entity Recognition (NER), Relation Extraction (RE), Question Answering (QA) and etc. The language model for the schema matching task requires further development and testing. In Hättasch et al. (2022) and Pan et al. (2022), both used the existing pre-trained language model for the schema matching tasks and achieved a better result than the state-of-the-art method. However, they have not applied the domain-specific language model to the schema matching task. Both works use the existing Bert model without any additional training. As mentioned in Hättasch et al. (2022) and Gururangan et al. (2020), domain special training is possible to improve the accuracy of schema matching in the special domain. In this paper, we extend the existing schema matching process with a domain-specific pre-trained model.

In summary, the contributions of the paper are:

- A new automatic schema matching process using a domain-specific pre-trained language model;
- Development of Energy Bert and Energy Sentence Bert for schema matching task;
- Evaluation of the new schema matching process and the developed language model with energy domain datasets.

Related work

Schema matching method

There are two main automatic schema-matching methods. Firstly, the COMA is a rule-based system that uses a path-based approach to compare schema and ontology elements and is a highly evolved and customizable system that allows for the combination of different matching algorithms and supports RDF (Resource Description Framework), OWL (Web Ontology Language), Linguistic based and structure based input (Do and Rahm 2002). It has been developed for three different versions COMA++ and COMA3.0 (Aumueller et al. 2005; Massmann et al. 2011). COMA++ is an extension of COMA that introduces machine learning techniques to improve matching accuracy. It also includes advanced matching strategies such as fragment matching and context-based filtering. COMA 3.0 is a further development of COMA++ that includes support for enriched mappings and ontology merging. It introduces new components such as an Enrichment Engine, a Merge Engine, and a Transformation Engine. But COMA is developed for general domain schema matching and does not take into account the specific characteristics and abbreviations of the energy domain. It relies on the availability of a large set of training data to generate accurate schema matching. In the energy domain, such labeled training data is limited, making it more challenging to obtain accurate schema mappings.

Secondly, in Pan et al. (2022) a schema matching process based on semantic similarity was developed, which used the active learning method to combine the different matcher with a small amount of labeled data and achieved a better result. It uses a two-step matching process: dataset level and attribute level, to reduce the computation time. Moreover, this methodology incorporates the existing Bert language model for schema

matching; regrettably, it lacks domain-specific training. This process which is a more general approach based on different semantic similarity calculation methods (e.g., edit distance, WordNet, and other corpus-based methods), is used as a starting point for this paper. We modified and further developed it to a domain-specific and language model-based schema matching process in this paper.

Pre-trained language model

The two representative models trained on the corpus in their respective domains without changing the basic Bert structure, BioBert and SciBert, were proven to significantly improve their performances of tasks in their fields. BioBert shifts the word distribution from general domain corpus to biomedical corpora, which performs significantly better in three biomedical text mining tasks (NER, RE and QA) (Lee et al. 2020). Similarly, the training data in Scibert consists of 82% biomedical and 12% computer science. Scibert even outperforms BioBert results on BC5CDR and ChemProt in the biomedical domain and achieves new SOTA results on ACLARC, and the NER part of SciERC (Beltagy et al. 2019). The advantage of these well-trained models is that they can more accurately identify words in their corresponding domain.

In 2019, SBERT (Sentence BERT) (Reimers and Gurevych 2019) based on transformer networks like BERT was proposed to improve efficiency on sentence-pair regression tasks like semantic textual similarity (STS). Since BERT and ROBERTA need to feed both sentences into the network, they modified the pre-trained BERT network to reduce computational overhead. However, sentence pairs used for training are expensive and difficult to produce, Sequential Denoising Auto-Encoder (TSDAE) method (Wang et al. 2021) trains the sentence embedding in an unsupervised manner by adding a certain type of noise into the encoder layer and then reconstructing the vectors into the original input. This solves the problem of limited labeled sentence pairs in the energy domain. Therefore, we reused TSDAE to train our domain-specific language model for schema matching.

Method

In this section, the new schema matching process is described in detail. As the input is raw data with a heterogeneous schema and a general data model, the relevant domains are identified based on the input data, and the domain paper in the corpus is selected. This method is developed to achieve a domain specific schema matching. Therefore, we first identify the domain of the schema matching process. In our use case, the energy domain is the relevant domain for this paper. There are two possibilities to determine the domain if the domain is not pre-defined. First, the keywords of the application domain are provided manually, which is usually the case if the domain expert works in one specific domain and knows exactly what is the data. Second, the domain keywords are extracted from the input data based on the keyword extraction method (e.g., Sharma and Li 2019 and Beliga et al. 2015). The corpus data containing domain specific keywords are extracted with help of the pre-defined S2ORC metadata. After preprocessing the domain paper, the generated training data are used in the domain language model. After the training, the language model is used in the dataset level matching and attribute level matching as a matcher. Although there have been many benchmark datasets with

homogeneous schema in the generic domain (Wang et al. 2021), heterogeneous schema data in the energy domain is still missing. The preprocessing step removes all special symbols, separates the connected phrases, and ensures the uniformity of word case (e.g., 'idStation' to 'id station'). The overview of the process is illustrated in Fig. 2.

Domain corpus

After identifying the domain of the raw data and data model, corpus data used for training data is selected and generated based on the keyword. A large corpus S2ORC (Lo et al. 2020) contains 81.1 million English-language academic papers spanning many academic disciplines. This general-purpose corpus for NLP and text-mining research over scientific papers are supported by the Semantic Scholar search engine. S2ORC introduces the sources, which these papers are derived from. In addition to trusted article sources, S2ORC also has advantages over the traditional IEEE and DBLP in construction. They break ties by minimizing the total number of sources from which they select metadata. With the help of metadata, 20 fields of study can be easily located and extracted. Instead of directly extracting the original papers, all the articles only kept the text part and remove all figures, tables, and references.

Based on the collected data, we fixed the training paper domain to energy. Since the energy domain is not included in the 20 fields in S2ORC, we extracted all abstracts and papers containing the word 'energy' separately. Finally, we processed these articles into a text file with a size of 1.40 GB, comprising 636,132 bodies of text and 554,240 abstracts. It is important to note that the numbers are inconsistent since we did not extract articles along with abstracts if the word "energy" wasn't present in both.

Pre-trained language model

Currently there is no Bert model in the energy domain, although the advantage of such a domain-specific model has been proved in Lee et al. (2020) and Beltagy et al. (2019). Tai et al. (2020) provides a method to extend the existing model by adding the domain-specific vocabulary in the tokenizer. To collect the existing energy domain vocabulary, we identified two categories of resources: data model (e.g. FIWARE (https://www.fiware.org/smart-data-models/) and EPC4EU Serna-González et al. 2021) and ontology (e.g. Brick Balaji et al. 2018 and SAREF Daniele et al. 2015). Those vocabularies first need to go through the preprocessing step to be divided into word pieces (e.g. in FIWARE

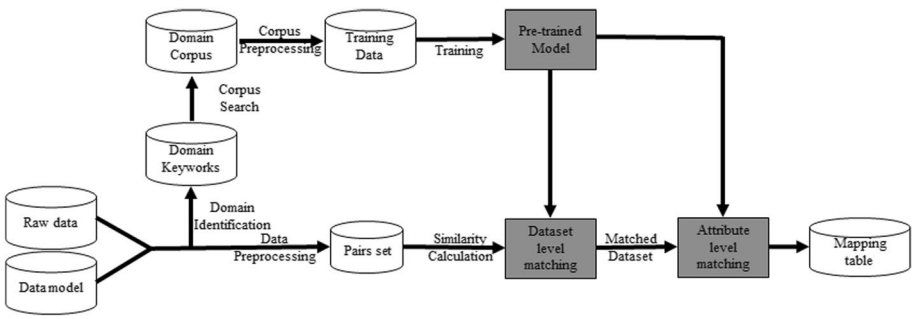

Fig. 2 Schema matching process overview

energy data model "phaseType" to "phase type") and the overlapped word needs to be removed (e.g. "building" in Brick and "building" in FIWARE building data model). Those are summarized and integrated into the BERT model vocabulary. In total, 460 additional tokens are added into the original vocabulary. Additionally, two energy domain pre-train language models are developed based on energy domain vocabulary, which are Energy Bert and Energy Sentence Bert.

Energy Bert

The Energy Bert starts with the basic structure of Bert like BioBert and SciBert and adjusts the training method as follows to improve the ability in the energy field.

- *Dynamic masking and larger batch size* In Zhuang et al. (2021), a robustly optimized BERT pre-training approach (RoBERTa) uses a new masking method called dynamic masking to make full use of semantic information contained in the corpus. This method performs several random masking on one sentence and uses the masked sentences to train in different epochs. This approach is used in our training process to utilize the domain corpus.
- *No NSP* Next Sentence Prediction (NSP) is a binary classification loss for predicting whether two segments follow each other in the original text. However, in RoBERTa, the authors questioned the necessity of the NSP loss. Instead of using the NSP loss, they trained the Bert model only with Masked Language Model (MLM) task with fine-tuning of parameters. The performance is even better than using NSP. The same conclusion is also mentioned in the article of Spanbert (Joshi et al. 2020). In addition, since NSP is more applied to QA tasks, which is not so helpful with semantic matching, in our experiment only MLM task is used for training.

Energy Sentence Bert

Because the goal of pre-trained language model is to help the schema matching process. The model is used to calculate the semantic similarity between pairs. The Sentence Bert is developed specifically for the text similarity and also achieves a good performance in the previous study (Pan et al. 2022). Therefore, Sentence Bert is used as starting point for the developed Energy Sentence Bert.

However, sentence pairs used for training are expensive and difficult to produce, so in our model, we adopted TSDAE method (Wang et al. 2021), which is also based on a sentence transformer to train our data. Sequential Denoising Auto-Encoder (TSDAE) trained unsupervised sentence embeddings by adding a certain type of noise into the encoder layer and then reconstructing the vectors into the original input.

Dataset level matching

All the steps above generate the domain-specific language model. The actual matching process starts with the generation of pairs sets: dataset pairs, which consist of the dataset in the raw data with the entity in the data model and attribute pairs, which consist of attributes in the raw dataset with attributes in the entity. We then apply these models to the dataset pairs and calculate the semantic similarity. The method of calculating

similarity is to create embeddings of the pairs, which are then passed through an attenuation mechanism and mean-pooling method to obtain their features by creating an embedding vector. The semantic similarity is calculated by computing the cosine similarity between the embedding vectors. In this part, the models and tokenizers are obtained from the pre-trained models mentioned above. The dataset-level matching calculates the similarity for the dataset pairs and matches the most similar dataset. After the dataset-level matching, the attributes between the most similar dataset pairs are compared and the most similar attribute pairs are found. With this two-step matching, the computational time is significantly reduced, as the attribute-level matching only needs to compare the most probable dataset pairs rather than all the datasets.

Attribute level matching

The attribute level matching utilizes the pre-trained language model and the results from dataset level matching to match the attribute pairs. The pre-trained language model calculated the semantic similarity of the attribute pairs. A threshold value is set to 0.3 according to Pan et al. (2022), if the similarity is smaller that 0.3, the similarity score will be 0. This method is used to avoid data noise. Finally, a mapping table is generated based on the attribute level and dataset level results, which contain the matched entities between raw data and data model.

Results

In this section, the datasets for schema matching, the pre-training setups, and the experimental results are discussed in detail. The experiment evaluates the performance of the developed Energy Bert model in schema matching problems. It contains the following two experiments. The dataset level matching experiment compares the developed model with the baseline model. The second experiment focuses on the vocabulary overlap between the developed model with the raw data and the baseline model, which provides a more insightful explanation of the model performance.

Dataset for schema matching

In this paper, the datasets are collected from 11 different energy network operators, energy suppliers and building energy management companies. The total number of raw data is 25 tabular datasets, which contain 140 attributes. 28 entities are in the data model and the total number of attributes in all the entities is 755. As a preparation for the experiment, one entity attribute has been found for each attribute in raw data manually. However, not every raw data could be matched with a suitable attribute in reality and one raw data could also correspond to multiple attributes. In this case, we only take the semantically closest group and vice versa.

Pre-training setups

For the pre-training of Energy Bert, the epoch of 2, the learning rate of 2e-5, batch size of 32 and drop out of 0.1 is selected according to the default setting. For the pre-training of Energy Sentence Bert, it trained with TSDAE method and applied STSB (The Semantic Textual Similarity Benchmark) to validate the cosine similarity scores. For the

pre-training of Energy Sentence Bert, the epoch of 1, batch size of 8 and max sequence length of 75 are selected.

The hardware used for training is one NVIDIA V100 (40GB) GPU. It takes nearly seven days to pre-train Energy Bert and nearly one day for Sentence Energy Bert. All models with their corresponding tokenizers used in training are loaded from Huggingface.

Dataset level matching experiment

The dataset level matching experiment compares the developed model with the baseline model. COMA is selected as a baseline model for the schema matching task. COMA provides a straightforward algorithm and user friendly GUI. BERT is selected as the baseline model for the pre-trained language model. Because it is the most widely used baseline model and compared in the most of literature in recent years. Another powerful language model is GPT (Generative Pre-trained Transformer) model. GPT is more suitable for task like text generation or dialogue applications, which is not considered as a baseline model for schema matching. In this part, the two pre-trained Energy Bert models and the baseline model are compared in the dataset level matching process. The results show in Table 1. The precision of COMA is higher than the recall of COMA. Because the dataset is collected from different European countries for example it contains Latvian (e.g. paterinš means consumption), which is not supported in COMA. The Energy Bert compared with the original Bert model further improve the F1 score from 0.571 to 0.714 (14.3 percent higher). The Energy Sentence Bert, which used the same corpus as Energy Bert, but with TSDAE for training, achieved the best result. This shows the Sentence Bert model is more suitable for the schema matching task.

Vocabulary overlap

In this part, we compare the overlap rate of the vocabulary between the original BERT base model and raw data. The vocabulary file of BERT model works as a dictionary, which indicates how much content a model has. In addition to single words, there are also word pieces denoted by the character "##". If a part of the token already exists in the vocab, the token will continue to be divided. After the division, except for the first part, other parts will be added with "##". In the following calculation of overlap rate the impact of "##" is ignored and treated as normal words. We split the vocab into a list and the overlap rate is calculated by the length of the same part in two compared lists, which is divided by the total length of the BERT model. By calculating the overlap rate between vocab from models and our raw data, the denominator is the length of the model. By BERT base model and other kinds of models, the denominator is the BERT model. The results are shown as followed. The overlap rate of Bert and raw data

Table 1 Dataset level matching results

Model	Recall	Precision	F1 score
COMA	0.164	0.5	0.247
BERT	0.516	0.640	0.571
Energy Bert	0.645	0.8	0.714
Energy Sentence Bert	0.677	0.84	0.75

is 21.759%. The overlap rate of energy domain papers and raw data is 42.361%, which means the energy domain papers provide much more meaningful training data compared with BERT model. This provides an explanation of why Energy BERT achieved a better performance in the dataset level matching experiment. The energy domain paper contains more domain-specific vocabulary compared with the general language model. This proves that our approach can select the meaningful domain-specific corpus to train the language model.

Conclusion and future work

This paper developed the new automatic schema matching process with the energy domain pre-trained language models. Energy Bert and Energy Sentence Bert are pre-trained with the energy domain paper in S2ORC corpus. After the integration of the energy domain language model, the schema matching process achieved a better result in terms of accuracy compared with the baseline model. This shows the importance of further research on domain special schema matching with pre-trained language models. The advantage of language models is their universal applicability. The developed Energy Bert is applied for the specific task schema matching. In the future, other semantic data related applications with Energy Bert will be further evaluated.

Acknowledgements
The authors would like to thank all the Enershare consortium partners.

About this supplement
This article has been published as part of Energy Informatics Volume 6 Supplement 1, 2023: Proceedings of the 12th DACH+ Conference on Energy Informatics 2023. The full contents of the supplement are available online at https://energyinformatics.springeropen.com/articles/supplements/volume-6-supplement-1.

Author contributions
Conceptualization, ZP; methodology, ZP writing—original draft preparation, ZP, MY; software, MY; writing—review and editing, ZP; supervision, AM.

Funding
This research was funded by Enershare, which is a European project funded by the European Union's Horizon 2020 research and innovation program under Grant Agreement No.101069831.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Published: 19 October 2023

References

Aumueller D, Do H-H, Massmann S, Rahm E (2005) Schema and ontology matching with coma++. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 906–908

Balaji B, Bhattacharya A, Fierro G, Gao J, Gluck J, Hong D, Johansen A, Koh J, Ploennigs J, Agarwal Y et al (2018) Brick: Metadata schema for portable smart building applications. Appl Energy 226:1273–1292

Beliga S, Meštrović A, Martinčić-Ipšić S (2015) An overview of graph-based keyword extraction methods and approaches. J Inf Org Sci 39(1):1–20

Beltagy I, Lo K, Cohan A (2019) Scibert: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620

Daniele L, Hartog Fd, Roes J (2015) Created in close interaction with the industry: the smart appliances reference (saref) ontology. In: International Workshop Formal Ontologies Meet Industries, pp. 100–112. Springer

Do H-H, Rahm E (2002) Coma-a system for flexible combination of schema matching approaches. In: VLDB'02: Proceedings of the 28th International Conference on Very Large Databases, pp. 610–621. Elsevier
Fernandez RC, Mansour E, Qahtan AA, Elmagarmid A, Ilyas I, Madden S, Ouzzani M, Stonebraker M, Tang N (2018) Seeping semantics: Linking datasets using word embeddings for data discovery. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 989–1000. IEEE
Fiware smart-data-models. https://www.fiware.org/smart-data-models/
Giunchiglia F, Yatskevich M, Shvaiko P (2007) Semantic matching: algorithms and implementation. In: Journal on Data Semantics IX, pp. 1–38 . Springer
Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360
Hättasch B, Truong-Ngoc M, Schmidt A, Binnig C (2022) It's ai match: A two-step approach for schema matching using embeddings. In: 2nd International Workshop on Applied AI for Database Systems and Ap- Plications
Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) Spanbert: Improving pre-training by representing and predicting spans. Trans Assoc Comput Linguistics 8:64–77
Kenton JDM-WC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
Lo K, Wang LL, Neumann M, Kinney R, Weld DS (2020) S2orc: The semantic scholar open research corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4969–4983
Malarvizhi R, Kalyani S (2013) Soa based open data model for information integration in smart grid. In: 2013 Fifth International Conference on Advanced Computing (ICoAC), pp. 143–148. IEEE
Massmann S, Raunich S, Aumüller D, Arnold P, Rahm E et al (2011) Evolution of the coma match system. Ontol Matching 49:49–60
Pan Z, Pan G, Monti A (2022) Semantic-similarity-based schema matching for management of building energy data. Energies 15(23):8894
Rahm E, Bernstein PA (2001) A survey of approaches to automatic schema matching. VLDB J 10(4):334–350
Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992
Sayah Z, Kazar O, Lejdel B, Laouid A, Ghenabzia A (2021) An intelligent system for energy management in smart cities based on big data and ontology. Smart Sustain Built Environ 10(2):169–192
Serna-González V, Hernández Moral G, Miguel-Herrero FJ, Valmaseda GC, Martirano Pignatelli F, Vinci F (2021) ELISE Energy & Location Applications: Use Case Harmonisation of Energy Performance Certificates of buildings datasets across EU- Final Report. Publications Office of the European Union, Luxembourg. https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.01.01_60/gs_cim009v010101p.pdf
Sharma P, Li Y (2019) Self-supervised contextual keyword and keyphrase retrieval with self-labelling
Sutanta E, Wardoyo R, Mustofa K, Winarko E (2016) Survey: models and prototypes of schema matching. Int J Electric Comput Eng 6(3):2088–8708
Tai W, Kung H, Dong XL, Comiter M, Kuo C-F (2020) exbert: extending pre-trained models with domain-specific vocabulary under constrained training resources. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1433–1439
Wang K, Reimers N, Gurevych I (2021a) Tsdae: using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 671–688
Wang J, Li Y, Hirota W (2021b) Machamp: a generalized entity matching benchmark. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4633–4642
Zhang Z, Chen Z, Zhao Q, Wang Y, Tian J (2023) Situation awareness and sensitivity analysis for absorption of grid-connected renewable energy power generation integrating robust optimization and radial basis function neural network. J Modern Power Syst Clean Energy
Zhuang L, Wayne L, Ya S, Jun Z (2021) A robustly optimized bert pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics, pp. 1218–1227

Publisher's Note