LOC: Livestock Ontology Construction Approach From Domain based Text Documents

Nandhana Prakash, Amudhan A, Nithish R, Krithikha Sanju Saravanan
Sri Sivasubramaniya Nadar College of Engineering

Abstract

Livestock plays an irreplaceable role in rural and global economies and as a part of its progression livestock ontology can unlock its potential of cross-domain applications of Natural Language Processing (NLP). Domain data is essential for the retrieval of semantic and syntactic understanding of the input text data given to the model. The paper presents a Livestock based Ontology Construction (LOC) framework. The input data undergoes anaphora resolution employing semantic methods based on rules then the pre-trained BERT model with regular expression is utilized for retrieving terms (entities) from the data. Now the Graph Neural Network (GNN) is constructed with regular expressions for extricating relationships from the input documents for designing the livestock ontology. The effectiveness of the proposed LOC based on the BERT model with regular expressions and GNN method with regular expressions, demonstrates noteworthy results when compared to existing methods, showing a precision and recall of 97.56% and 95.24%.

1 Introduction

Livestock is an essential element for the survival and progress of the human population. Livestock plays a vital role in upholding the livelihoods around the globe, which encompasses various practices in the livestock industry. A country’s economic power is significantly influenced by livestock because the livestock and agriculture industry, resources, political policies and overall economic structure are the main contributors to the country’s economic fluctuations. Developments in supporting sectors in turn develop the livestock sector and it serves as a virtuous cycle. In the current era, information plays a vital role in everything and in this context, it can help educate people who manage livestock. Analyzed information based on trends and innovative techniques with mindful practices enables livestock owners to make informed decisions to increase their profit and hence improve the sector as a whole. Web resources play an important role in the maintenance and display of data specific in the domain via information sharing and data crowd sourcing, AI systems and innovations in research. All rounded data via social media, newspapers, articles, and websites in livestock is constantly increasing. The fact that the data from the livestock domain is different when they are in different web sources is a notable issue, and content extraction has to be taken with huge precision. The extraction of relevant data from text within the domain encompasses huge syntactic and semantic knowledge. Drawing an ontology is a great method for knowledge extraction (Saravanan and Bhagavathiappan 2024). The creation of domain-specific ontology requires the identification of all relevant terms and the meaningful comprehension of terms and sentences that are in the domain. But extracting these domain-specific terms from the textual input, despite all the forms of variations, is a challenging task. Implementing a powerful knowledge extracting strategy is vital for proper ontology construction and development. A good identification system must detect domain-specific compound words. Building an ontology for this domain has great potential and uses. Ontology studies the links between all entities and their relationships, which are then represented as a directed graph that connects all of them (Saravanan and Bhagavathiappan 2024). Establishing links between the terms in the text documents is essential for the development of response systems that use Generative AI that are specific to the domain and can answer domain-specific questions. While developing a thorough ontology construction methodology from a list of text sources, the following aspects must be considered:

1. Entities in the text must be recognized by applying semantic knowledge to the sentences.
2. Relationships between entities should be extricated using both semantic and syntactic knowledge.
3. In order to properly construct the ontology, relationships between entities must be properly acknowledged.

The following techniques are used by the LOC framework to extract domain-specific terms in the dataset

1. To guarantee that no information is lost during the entity extraction process, nominal anaphora resolution is applied to the dataset.

2. To create word embeddings, the third method combines NLP approaches with a pre-trained BERT model. An unsupervised Graph Neural Network (GNN) is then used to analyze these embeddings and identify associations in conjunction with positional information and regular expressions.

3. The GNN model's performance is optimized through hyperparameter adjustment.

2 Literature Survey

In this section, literature survey for the LOC work was carried out and is summarized below.

Gkoutos et al. (2012) in their recent studies in genetic and environmental factors have led us to understand its influence on its behaviour. The neurobehaviour ontology (NBO) depicts the systematic representation of behavioural processes and supports drawing commonalities across species and facilitate genotype comparisons that can be used to understand human diseases as well.

Kang et al. (2018) in their paper focus on studying diseases in calves by taking into account their birth history, housing conditions, immunity status. The findings throw light on the importance of reporting tagging livestock and environment management of breeding farms which are great measures to manage diseases in livestock.

Golik et al. (2012) introduce the multi species Animal Trait Ontology for Livestock (ATOL) which indexes phenotype databases that covers key areas like growth, nutrition, milk, production with welfare and draws species-specific reasoning which was developed using existing ontologies and analyzing scientific terminology. It features a taxonomy of 1,654 traits and solves challenges in phenotype data integration.

Mullen et al. (2024) use the Vertebrate Breed Ontology (VBO) which was synthesized to standardize breed names in veterinary science improves data portability within diagnostics, treatments and precision medicine in veterinary science. VBO compiles breed names and other information and presents each breed as an entity with metadata, including over 19000 breeds across 41 species which are classified using description logic for advanced data analytics, these terms are linked to the NCBI Taxonomy which enhance contextual understanding.

Zhao et al. (2024) show that past methodologies used in ontology requirements engineering (ORE) have been predominantly relying on manual techniques like conducting interviews and discussions to state their requirements obtained from domain experts, especially in large-scale projects. The OntoChat framework introduces a novel approach to ORE by utilizing large language models (LLMs) in order to narrow the process via 4 major functions which are user story generation, competency question (CQ) extraction, CQ filtering and analysis and support for ontology testing.

Peroni (2017) presents SAMOD (Simplified Methodology for Ontology Development) which is a modern agile approach to construct ontologies via an iterative job flow which consists of small manageable incremental steps. SAMOD emphasizes on the designing of well-constructed and properly documented models beginning with precise domain descriptions.

Fathallah et al. (2024) in their paper introduce ontology learning by using the structured NeOn methodology utilizing large language models

(LLMs) to transform natural language domain descriptions into Turtle syntax ontologies. The major contributor is a prompt pipeline constructed for domain agnostic modelling, which is depicted in a case study on the ontology of wine. The constructed workflow, NeOn-GPT automates ontology and aids in its development within the platform. Result evaluation via the Stanford wine ontology shows that LLMs struggle with procedural tasks and do not have much reasoning skills but can majorly impact the time and expertise required by a great deal.

Martin (2024) presents System of systems (SoS) using the advantages of independent systems to complete complex missions. This paper tackles the progressing complexity of SoS by utilizing artifacts to improve understanding and engineering procedures. They include a core ontology for missions and capabilities, a decision support framework, and a premature construction model of the ontology. The artifacts depicted through wildfire management and road paving use cases, strive to improve conceptual clarity and feasibility in SoS development which also addresses complexity at a higher level.

Poveda-Villalón et al. (2024) show that vocabularies and ontologies are vital for standardizing and integrating data from diverse resources into Knowledge Graphs. As more models have been constructed, many other engineering methodologies have been produced across various semantic artifacts in various domains. Ensuring that ontologies adhere to the Findable, Accessible, Interoperable and Reusable (FAIR) principles from the outset poses a challenge. It reviews the existing guidelines and rules available for constructing ontologies FAIR and maps them to the development cycle, showing current gaps where no guidelines support FAIR-by-design practices.

Fuentes et al. (2021) depict that precision farming is used to improve agricultural procedures using data analysis techniques along with dairy farming which increases its potential due to advanced computational techniques. This work explores the usage of data analytics to optimize dairy production utilizing a data center from the Valacta center in Eastern Canada that records dairy cow and farm performance, unifying and centralizing heterogenous dairy data and provides a common vocabulary for stakeholders and automated tools.

Song et al. (2024) demonstrate that Invariant based Contrastive Learning (ICL) methods have been used to augment and make it more robust in downstream tasks. It is used in various domains but doesn’t have latent space representation for augmentation information. Incorporating equivariance in ICL can make it more effective. This paper uses augmentation strategies and shows the role of equivariance in improving ICL’s performance. It proposes CLeVer (Contrastive Learning Via Equivariant Representation), a framework integrating equivariance in contrastive learning.

2.1 Challenges

The following are the challenges in the construction of ontology.

1. Anaphora resolution for domain related task is not an easy job as there is a significant lack of domain information on all sides. If these can be resolved, then NLP processing can easily be done.
2. Getting livestock information is the most important step to build the skeleton of the system.
3. If there is a lack of proficiency in a particular domain then it is very difficult to process NLP in the domain. Domain strength is the main pillar NLP relies on. Without any prior knowledge, it might be very difficult to process information, assess it or recognize any of the entities in that particular area.
4. Getting to know about all the entities is a very difficult task as on also it is very important to ensure that inter relationships between all the entities in a domain are known without which a skeletal system in the particular domain cannot be created or put to use in a practical manner.

Compared to the literature survey the proposed LOC system is unique in the following ways.

1. Compared to the above system the LOC system is better than other systems as it uses all the literary entity terms as input. With this input system all details about the entity are given and the system starts understanding how the entire domain works. A complete understanding of the

domain and how it works isn't necessary. With the input given, the system takes care of creating complete character of how the entire domain works
2. Semantactic and syntactic positions features are used to resolve anaphora.
3. The complex livestock entities are extracted effectively using trained N gram model.
4. Using semantic features and syntactic rules, the model can find out how the entire domain systems can be extracted feature by feature.

With this and the entire ontology construction program can be completed easily in the LOC Framework system. Different entities in the domain can be brought together and also can be clearly be connected between each so that a complete skeletal system can be clearly formed.

Here the LOC system tries to create an ontology of the livestock program with the given documents as well as the architectural diagram provided which shows the features of the methodology involved. For creating a livestock ontology, all the entities in the livestock domain need to be extracted properly and also the relationship between all the entries in place need to be identified and then need to be interconnected perfectly. While handling text documents, it is very important to identify the anaphora in the system so that there is no loss of data. So the text document undergoes anaphora resolution using semantic approach so that there is no loss of data.

2.2 Materials and methods

The LOC system is properly elaborated here. The methodologies depicted in the entire construction diagram is shown in Fig 1. The construction of an ontology from the input text is the main aim of the proposed LOC system. The entities and relationships present in the text should be properly extricated to design a precise ontology (Saravanan and Bhagavathiappan).

2.2.1 Anaphora Resolution:

The presence of anaphors are common and should be removed before processing the input text documents, so as to avoid data loss. The anaphora resolution phase which applies the semantic approach helps with this.

During the anaphora resolution phase, the anaphors present in the livestock domain based text are mapped to their respective antecedents. This is a crucial step in resolving uncertainties and ambiguities in the document (Saravanan and Velammal 2024). By removing these we get precise and accurate information from the input. All compound sentences are resolved into simple sentences that contain at least two entities and one relationship mapping the two entities.

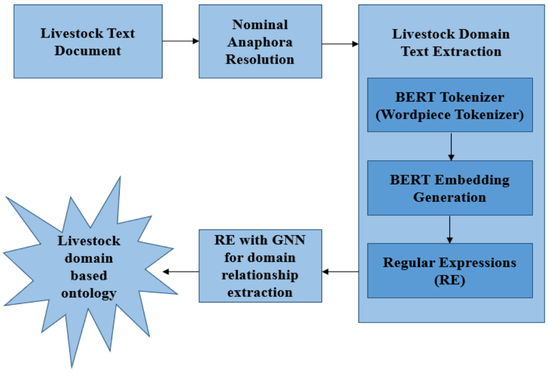

Livestock Text Document
Nominal Anaphora Resolution
Livestock Domain Text Extraction
BERT Tokenizer (Wordpiece Tokenizer)
BERT Embedding Generation
Regular Expressions (RE)
RE with GNN for domain relationship extraction
Livestock domain based ontology

Fig 1. Block diagram of proposed LOC approach

2.2.2 Domain term extraction:

The BERT model has been pre-trained with Masked Language Modelling (MLM) to be able to predict sentences. Thus, livestock related terms can be extracted from the anaphora resolved documents. For pre training, deep bidirectional depiction of words in the unlabeled input data with joint conditioning on all facets of the context of the BERT model is utilized. The model contains an embedding layer at the input end which is used to convert the data in the input text into a mathematical vector form which is required for future processing. At the output end, a Softmax layer used to transform the output into a probability distribution over the possible classes.

The BERT model generates three types of embeddings: token embeddings, used to convert words to its pretrained vector representation; segment embeddings, which form a numerical value via encrypting the data number; and position embeddings, which encrypt the index of all tokens in the sequence ranging from 0 to (512-1). BERT lacks structure for sequences like Recurrent Neural Network, so positional embeddings are used for capturing the context by encoding the order of the text. [CLS] and [SEP] are two special tokens in the BERT model for proper understanding of text. The [CLS] token is a classification token utilized in the final encrypted units of the model. The [SEP] token is inserted at the end of every input sentence to determine the end of one sentence and the start of another. Fig.1 shows the architecture of the BERT model. RE enhance BERT's entity extraction by identifying domain-specific text patterns. Thus, enhances the domain term extraction for the livestock domain.

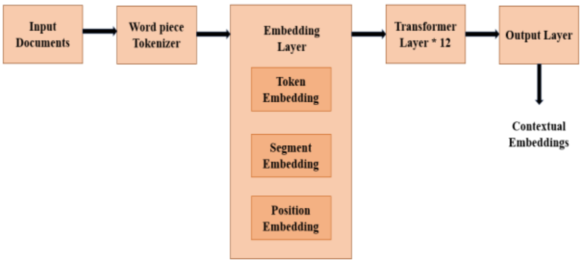


Fig 2. Architecture of BERT Model

2.2.3 Domain Relationship Extraction and Ontology Construction:

The Graph Neural Network, GNN, is a special neural network designed to process graph-structured data from which graphs can be constructed. A graph has nodes (entities) and edges (relationships) that connect these nodes, which has attributes or features. These may or may not have attributes or features. The GNN learns a function that can map the features of the nodes and edges to outputs like labels, embeddings, graph embeddings, or graph labels. Hyperparameter tuning of the GNN involves experimenting with learning rates, optimizers, and activation functions. Table 1 represents the mean squared error losses obtained during the hyperparameter tuning of the GNN model at the learning rate of 0.01 with Relu activation function and RMSProp optimization. In the proposed model, twelve specific relationships are considered (type of, has a, produces, consumes, located in, managed by, treated with, has offspring, requires, sold in, monitored by, vaccinated against). The first method employs POS-based Hearst patterns with positional features to draw relationships between terms. The second method uses regular expression in combination with Hearst patterns and features for retrieving relationships. The third method uses a straightforward GNN employing regular expression to identify relationships among these entities. The extracted terms and relationships are depicted on an ontology graph using Networkx, and Matplotlib libraries in Python.

The GNN contains two primary components, a message passing and a readout layer. The message passing layer updates features of the node by combining features from neighboring nodes and edges and iterating many times to allow the node features to contain local and global data that is necessary for construction of the directed graph. The readout layer uses these node features to synthesize the final graph. To implement, the GNN model uses PyTorch and PyTorch geometric libraries and applies regular expression to extract the relationships from the documents. The input dimension corresponds to the total number of entities in the graph, and the output dimension corresponds to the hidden dimension of 50, and the GNN model comprises two linear layers to completely transform node features from hidden to output dimension and from input to hidden dimension. The forward method uses node features and edge indices and returns modified features of the node after using the linear layers and activation function. The GNN model is trained using the mean squared error loss function and an optimizer with required relationships which is extracted through regular expression. Its performance loss is quantified as MSE. In this, hyperparameter tuning has also been conducted to minimize MSE and is shown in Table 1. So, RE identifies initial relationships between entities, forming structured graph input for GNN. The GNN then refines and generalizes these relationships to build the ontology.

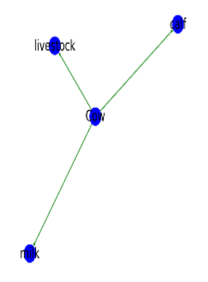
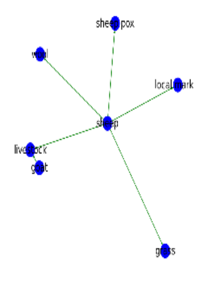
Table 1. Depicts the results obtained from applying optimizer function

	RELU	GELU
Adam	$3.8838e^{-06}$.	$1.9429e^{-06}$
AdaGrad	0.0020	$5.7975e^{-06}$
SGD	0.0799	0.1276
RMSProp	$9.2518e^{-16}$	$3.1880e^{-13}$
AdaDelta	0.2392	0.3882
SGD with Momentum	$2.1281e^{-05}$	$6.596e^{-06}$

Experimental Outcomes

The text document dataset in the livestock domain has been obtained from various government websites and blogs for 100 pages. The step by step results of the proposed LOC is proposed is shown Table 2 for the sample data and the sample ontology is shown in Fig 3.

Table 2. Step by step results of proposed LOC

Sentences	**Sample Data 1**	**Sample Data 2**
Input text	Cow is a type of livestock. Its offspring is calf. It produces milk	Goat and sheep are types of livestock. Sheep produces wool and consumes grass. Sheep is sold in markets. Sheep is vaccinated against sheep pox.
Anaphora resolved text	Cow is a type of livestock. Cow's offspring is calf. Cow produces milk.	Goat is a type of livestock. Sheep is a type of livestock. Sheep produces wool. Sheep consumes grass. Sheep is sold in local markets. Sheep is vaccinated against sheep pox
Entities	Cow, livestock, calf, milk	Goat, sheep, livestock, wool, grass, local markets, sheep pox.
Relationships	(Cow, livestock) (Cow, calf) (Cow, milk)	(Goat, livestock) (Sheep, livestock) (Sheep, wool) (Sheep, grass) (Sheep, local market) (Sheep, sheep pox)
Ontology		

The term extraction method of the proposed LOC framework is evaluated using performance metrics and is compared against existing methods as shown in Table 3. From the Table 3 it is clearly seen that the proposed model of BERT with regular expressions of LOC framework performs slightly better than other methods which makes a big impact in the construction of domain ontologies. The sample ontology constructed using LOC framework is depicted in Fig 3.

Table 3. The proposed LOC framework is compared with existing models

Method	Precision	Recall	F1 score	Accuracy
NER in Spacy	0.8571	0.7500	0.8000	0.7500
Basic NLP methods	0.9524	0.9091	0.9302	0.9032
Proposed method (BERT with regular expressions)	0.9756	0.9756	0.9756	0.9690

The proposed LOC approach is evaluated using the standard metrics and is shown in Table 4. Also, it is compared with existing methods which is shown in Table 4.

Table 4. Evaluation of proposed model

Metric	Value Obtained
Sensitivity	0.9524
Specificity	0.9474
Precision	0.9746
Negative Predictive Value	0.9000
False Positive Rate	0.0526
False Discovery Rate	0.0244
False Negative Rate	0.0476
Accuracy	0.9508
F1 score	0.9639
Matthews correlation coefficient	0.8876

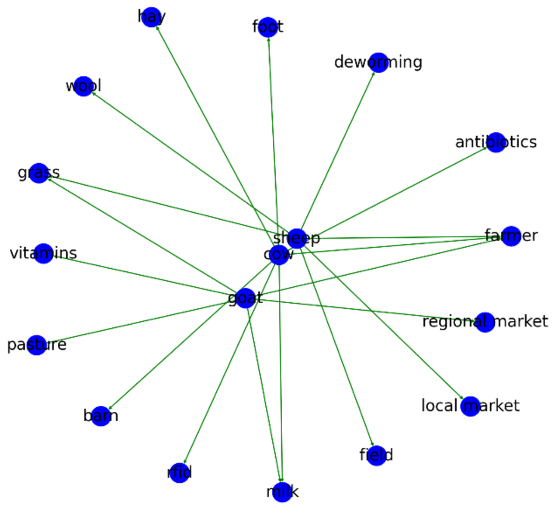


Fig 3. Model ontology construction using LOC framework

Table 5. Comparison of relationship extraction for ontology construction

Method	Precision	Recall	F1 score	Accuracy
Parts of Speech (PoS)	0.5263	0.5714	0.5479	0.5147
PoS based Hearst Pattern	0.7353	0.6849	0.7092	0.6239
Regular expression based Hearst Pattern	0.7826	0.8182	0.8000	0.7188
Proposed method (GNN with regular expression)	0.9756	0.9524	0.9639	0.9508

The comparison of proposed LOC method with existing methods is shown in Table 5. From the Table 5 it can be clearly inferred, the proposed LOC method shows outstanding results for performance metrics when compared with the existing methods.

Also, the proposed model utilizing the domain experts is assessed on the basis of clarity which is satisfied by easy and clear understandability of the model with the cumulative score of 97.4 %. On the basis of coherence, showing how the content flows logically and cohesively, the score is 97.9%. On the basis of minimal encoding bias, ensuring the model does not promote stereotypes or unfairness, the cumulative score is 97.5 %. On the grounds of conciseness measuring how efficient the expressions are and avoiding irrelevant details, the cumulative score is 97.2%. On the basis of completeness, showcasing how comprehensive the inferences provided in outputs are, it can be measured by comparing the content against reference data. This is satisfied in the framework and is given a cumulative score of 98.1%.

Conclusion

This research introduces LOC framework for creating a livestock ontology. Domain terms are extracted through nominal anaphora-resolved text and relationships are identified using 3 distinct techniques. Among these the BERT model with regular expression and GNN with regular expression provides an accuracy of 95.08%. Involving domain experts is crucial for the system's accuracy and applicability. However this puts it at risk for biases from domain experts as well as the absence of a benchmark dataset. Future research can focus on refining the ontology construction process by integrating additional relationships and adapting to different livestock domains and applications. The developed ontology has different applications and can benefit livestock agents and assist policymakers in optimizing resource allocation. It can also support the education sector and help us better understand market trends and environmental impacts

References

Fathallah, N., Das, A., De Giorgis, S., Poltronieri, A., Haase, P. and Kovriguina, L., 2024. NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning. In *The Extended Semantic Web Conference*.

Fuentes, V., Martin, T., Valtchev, P., Diallo, A.B., Lacroix, R. and Leduc, M., 2021. Toward A Dairy Ontology to Support Precision Farming. In *ICBO* (pp. 24-35).

Gkoutos, G.V., Schofield, P.N. and Hoehndorf, R., 2012. The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. *International review of neurobiology*, *103*, pp.69-87.

Golik, W., Dameron, O., Bugeon, J., Fatet, A., Hue, I., Hurtaud, C., Reichstadt, M., Salaün, M.C., Vernet, J., Joret, L. and Papazian, F.,

2012. ATOL: the multi-species livestock trait ontology. In *Metadata and Semantics Research: 6th Research Conference, MTSR 2012, Cádiz, Spain, November 28-30, 2012. Proceedings 6* (pp. 289-300). Springer Berlin Heidelberg.

Kang, Y.J., Choi, D.O. and Yin, L., 2018, October. Prediction of livestock diseases using ontology. In *2018 International Conference on Sensor Networks and Signal Processing (SNSP)* (pp. 29-34). IEEE.

Martin, J., 2024. *Towards Mission and Capability Modelling for Systems of Systems* (Doctoral dissertation, Mälardalens universitet).

Mullen, K.R., Tammen, I., Matentzoglu, N.A., Mather, M., Mungall, C.J., Haendel, M.A., Nicholas, F.W. and Toro, S., 2024. The Vertebrate Breed Ontology: Towards Effective Breed Data Standardization. *arXiv preprint arXiv:240602623.*.

Peroni, S., 2017. A simplified agile methodology for ontology development. In *OWL: Experiences and Directions–Reasoner Evaluation: 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers 13* (pp. 55-69). Springer International Publishing.

Poveda-Villalón, M., Garijo, D., Gonzalez-Beltran, A.N., Jonquet, C. and Le Franc, Y., 2024. Ontology Engineering and the FAIR principles: A Gap Analysis.

Saravanan, K.S. and Bhagavathiappan, V., A new framework for building agricultural domain-based ontologies from text documents using natural language processing and artificial intelligence techniques. Journal of Intelligent & Fuzzy Systems, (Preprint), pp.1-19.

Saravanan, K.S. and Bhagavathiappan, V., 2024. AOQAS: Ontology Based Question Answering System for Agricultural Domain. International Journal of Computer Information Systems and Industrial Management Applications, 16(2), pp.16-16.

Saravanan, K.S. and Velammal, B.L., 2024, May. A New Semantic Relationship Extraction-based Ontology Construction for Agricultural Domain. In 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-10). IEEE.

Saravanan, K.S. and Bhagavathiappan, V., 2024. Innovative agricultural ontology construction using NLP methodologies and graph neural network. Engineering Science and Technology, an International Journal, 52, p.101675.

Song, S., Wang, J., Zhao, Q., Li, X., Wu, D., Stefanidis, A., Su, J., Zhou, S.K. and Li, Q., (2024) Representation. *arXiv preprint arXiv:2406.00262*.2024. Contrastive Learning Via Equivariant

Zhao, Y., Zhang, B., Hu, X., Ouyang, S., Kim, J., Jain, N., de Berardinis, J., Meroño-Peñuela, A. and Simperl, E., 2024. Improving Ontology Requirements Engineering with OntoChat and Participatory Prompting. *arXiv preprint arXiv:2408.15256*.