Using BERT Models to Automatically Classify Domain Concepts into DOLCE Top-Level Concepts: A Study of the OAEI Ontologies

Guilherme Sousa[1], Rinaldo Lima[2], Renata Vieira[3] and Cassia Trojahn[1]

[1]IRIT: Institut de Recherche en Informatique de Toulouse, France
[2]Universidade Rural de Pernambuco, Recife, Brazil
[3]CIDEHUS, Universidade de Évora, Portugal

Abstract

Top-level ontologies provide a set of foundational concepts that have a well-founded philosophical meaning, being a useful tool in ontology engineering. However, in practice, few domain ontologies integrate top-level concepts. One of the difficulties refers to the selection of appropriate top-level concepts. This paper presents an analysis of top-level categories of a set of well-known domain ontologies from ontology matching benchmarks. Our main hypothesis is that training classification models using only concept comments (i.e., `rdfs:comment`) from top-level concepts can improve reported results in the literature. We then consider the best classifiers to estimate the distribution of concepts from Ontology Alignment Evaluation Initiative (OAEI) ontologies aligned to DOLCE top-level concepts.

Keywords

Foundational Ontologies, Top Level Prediction, Ontology Matching

1. Introduction

Top-level ontologies provide a set of foundational concepts that have a well-founded philosophical meaning, being a useful tool in ontology engineering [1]. They play an essential role in different tasks, such as ontology matching, providing a bridge for different ontologies [2, 3]. However, not all ontologies were built using top-level ontologies as a foundation and some existing ones are too large to be manually annotated. In this sense, the use of automatic top-level classifiers can help to establish a link between domain and foundational ontologies. In order to train these classifiers, a large amount of labeled data aligned with top-level concepts is required. One relevant source of such data is OntoWordNet [4] which aligns WordNet [5] synsets to DOLCE [6] concepts.

A recent effort in such direction has been done in [7], where a training dataset was constructed using labels and comments associated with the entities in OntoWordNet, which are aligned with

FOUST VII: 7th Workshop on Foundational Ontology, 9th Joint Ontology Workshops (JOWO 2023), co-located with FOIS 2023, 19-20 July, 2023, Sherbrooke, Québec, Canada.
guilherme.santos-sousa@irit.fr (G. Sousa); rinaldo.jose@ufrpe.br (R. Lima); renatav@uevora.pt (R. Vieira); cassia.trojahn@irit.fr (C. Trojahn)
0000-0002-2896-2362 (G. Sousa); 0000-0002-1388-4824 (R. Lima); 0000-0003-2449-5477 (R. Vieira); 0000-0003-2840-005X (C. Trojahn)

top-level DOLCE concepts. Using this data, several classifiers were evaluated for predicting top-level concepts of entities. Based on that previous work, in this paper, we evaluate the performance of a set of classification models and the impact of using comments as features in the classification task. We address the cases of multi-inheritance, which may lead to different top-level concepts in DOLCE, in a different manner from this previous work by disambiguating cases that lead to a unique top-level concept and filtering those that lead to multiple concepts. We then select the best classifier to study the distribution of top-level concepts in well-known domain ontologies from benchmarks used for evaluating matching systems. Our study analyses the distribution of the concepts of the ontologies from each track from the Ontology Alignment Evaluation Initiative (OAEI) [8]. This is the first effort in such direction and our intuition is that concepts in ontology correspondences (a correspondence is a triple involving a source concept in the source ontology and a target concept from the target ontology, together with a relation between them).

The remainder of this paper is structured as follows: in Section 2 related work is presented. In Section 3 the multi-inheritance is discussed along with the approaches adopted. Section 4 presents an evaluation of the performance of the models. Section 5 presents the analysis of OAEI ontologies and, finally, in Section 6 the conclusion and future work are discussed.

2. Related Work

Previous work has shown the importance of associating concepts from top-level to domain ontologies. In [9], correspondences between DBPedia ontology and DOLCE-Zero [10] have been used to identify inconsistent statements in DBPedia. In [11] an alignment between a foundational ontology (BFO) and a biomedical ontology (GO) is used for filtering out correspondences at the domain level that relate two different kinds of ontology entities.

Analyzing the impact of using top ontologies as semantic bridges in ontology matching have been done in [12], where a set of algorithms exploiting such bridges are applied and the circumstances under which foundational ontologies improve matching approaches are studied. They propose algorithms that use structural and mixed information and show that different combinations have different impacts on precision and recall. In [13], where OAEI ontologies were manually aligned to UFO, adopting a set of patterns grounded by UFO ontology. In [14], a domain ontology describing web services (OWL-S) has been manually aligned to DOLCE-Lite-Plus, in order to overcome conceptual ambiguity, poor axiomatization, loose design, and narrow scope of the domain ontology. The difficulties of such a manual alignment have also been addressed in [15], where the authors evaluate the performance of manual classification of entities in top-level concepts. The experiment was conducted by asking experts to manually classify a set of entities into top-level concepts. They showed a high level of disagreement between experts and that a methodological framework for this integration is needed.

In order to automate this process, in [16], word sense disambiguation and word embedding models have been used to automatically align top and domain concepts. The evaluation has been conducted with the task of associating DOLCE and SUMO top-level concepts to ontologies from three different domains. Automatisation has been also addressed in [7]. The authors organize two datasets based on OntoWordNet with the goal of training top-level concept classifiers of

ontology entities. The first dataset contains OntoWordNet entities with their respective DOLCE concepts. The second dataset contains the same entities but is classified into 5 top-level concepts (Endurant, Perdurant, Quality, Situation, and Abstract). Along with the datasets, the authors evaluate several models that predict the top-level concept based on entity labels and comments. In their following work [17], they propose a method to extract two datasets from OntoWordNet that have target concepts from DOLCE Lite and DOLCE Lite Plus. Different language models are evaluated in the task of predicting the top-level concept from the textual comments, the BERT base achieves the best results in predicting the concept from comments. Their models were not available at the time we conducted our study.

3. Materials and Methods

3.1. Training Datasets

This section describes the characteristics of the datasets from [7] and how they have been rebuilt and extended in order to deal with multi-inheritance cases[1].

Dataset Lopes22-5c This is the original dataset from [7], which is used to train models for top-level concept prediction. The dataset is built from OntoWordNet containing 116838 entities. It provides links for each concept of OntoWordNet to one of the 5 top concepts of DOLCE (Endurant, Perdurant, Quality, Situation, and Abstract). This dataset is composed of 3 columns (Table 1): Concept (DOLCE top-level concept), Label (OntoWordNet entity label, `rdfs:label`, and Comment (OntoWordNet entity comment, `rdfs:comment`).

Table 1: Example of the Lopes22-5c dataset.

Concepts	Label	Comment
endurant	order	established customary state esp. of society. 'order ruled in the streets'. 'law and order'
endurant	ritual	the prescribed procedure for conducting religious ceremonies
endurant	celebration	the public performance of a sacrament or solemn ceremony with all appropriate ritual. 'the celebration of marriage'
endurant	ritual	stereotyped behavior

Dataset Sousa23-5c The Sousa23-5c rebuilds the Lopes22-5c dataset while taking into account the problem of multi-inheritance, which is further detailed in Section 3.1.1. Hence, the resulting dataset Sousa23-5c differs from Lopes22-5c since the strategy of dealing with multi-inheritance filters ambiguous entities, while in Lopes23-5c, the entity is inserted multiple times with each possible top-level concept.

[1]The source code used in dataset generation and experiments can be found at https://gitlab.irit.fr/melodi/ontology-matching/top-level. For the rebuild of the dataset, the version of OntoWordNet used was downloaded from http://www.loa.istc.cnr.it/ontologies/OWN/OWN.owl(on 01/04/23) having 66065 entities

Dataset Sousa23-6c Another characteristic of Lopes22-5c dataset is its highly imbalance concept distribution (Endurant 76%, Perdurant 10%, Quality 4%, Situation 6%, Abstract 3%). Our proposal to deal with this problem is to break Endurant into two groups of concepts, generating a more balanced dataset with 6 concepts. This dataset was built from Sousa23-5c by following the hierarchy of entities until reaching one of the 5 top-level concepts (Endurant, Perdurant, Quality, Situation, Abstract), and in the case that the type Endurant is found, it is replaced by the immediate child in the path (Physical-endurant or Non-physical-endurant).

The concept distribution of the 3 datasets is presented in Table 2.

Table 2: Concept distribution of the datasets. Sousa-5c deals with multi-inheritance and Sousa23-6c does not have the Endurant concept. The number of perdurants can vary since Spatio-temporal-particular is the super concept of both Endurant and Perdurant and due to the path selection method used, different types can be reached by varying the number of perdurants in the two datasets.

concept	Lopes22-5c	Sousa23-5c	Sousa23-6c
Endurant	88410 (75.7%)	27900 (53.0%)	-
Physical-endurant	-	-	44553 (41.7%)
Non-physical-endurant	-	-	35853 (33.5%)
Perdurant	11683 (10.0%)	9045 (17.2%)	10847 (10.1%)
Quality	4948 (4.2%)	4245 (8.1%)	4245 (4.0%)
Situation	7763 (6.6%)	7157 (13.6%)	7157 (6.7%)
Abstract	4035 (3.5%)	4268 (8.1%)	4268 (4.0%)

3.1.1. Dealing with Multi-Inheritance in DOLCE and OntoWordNet

In order to deal with the case of multiple paths to the top-level concepts, we consider two distinct scenarios: one when the multi-inheritance occurs in the Wordnet part of OntoWordNet, and the other when it is present in the DOLCE hierarchy. If an entity in WordNet has multi-inheritance, many paths are traversed, and if they all lead to the same type, the entity is added to the dataset. If the paths diverge, the entity is ignored.

In DOLCE, one example of a concept without a direct path to the proposed top-level concepts is *Physical-realization* that is a sub-concept of *Spatio-temporal-particular* and which in turn is the super concept of Endurant, Perdurant, and Quality causing ambiguity. To deal with these cases, when multi-inheritance occurs in the DOLCE hierarchy, a breadth-first search is performed until one of the defined top-level concepts is found. If the concept found is *Spatio-temporal-particular*, then the WordNet entity is not added to the dataset.

From the total of 66065 entities present in OntoWordNet, 889 entities in the WordNet hierarchy have multi-inheritance. However, 5023 entities in the DOLCE hierarchy remain ambiguous even after applying the strategy mentioned above. To solve this problem, for DOLCE concepts that do not have a direct superclass, a breadth-first search is performed by traversing the predicates *RDFS.subClassOf*, *OWL.equivalentClass*, *OWL.intersectionOf*, *OWL.unionOf*, *RDF.first*, *RDF.rest* in decreasing order priority, adding the resulting objects to the priority queue used for the search. Using this approach, the distribution of concepts remained deterministic over several runs.

After this disambiguation process, entities are post-processed, in which labels are selected from the entity name in the WordNet part and `rdfs:comment` as comments. comments are then converted to lowercase. Both newline characters and quotes are removed, and semicolons are replaced with periods. Labels having synonyms separated by two underscores ('__') are split and generate new rows in the dataset. For example, *SOFTHEARTEDNESS__TENDERNESS* is split and generates two entries in the dataset *SOFTHEARTEDNESS* and *TENDERNESS*.

3.1.2. Test Dataset

For the purpose of evaluating the performance of the classification models, 2 testing datasets based on OAEI Conference track ontologies[2] were created. The top-level concepts are assigned using an existing reference alignment provided in [18] that aligns the highest concepts in Conference to DOLCE. The Conference dataset contains 70 correspondences between the concepts in Conference with concepts in DOLCE-Lite-Plus (DLP). The sub-concepts of top-concepts in Conference are aligned, by transitivity, to the top-level concepts in DOLCE. From the 70 alignments present in the initial reference alignment, 1 has multiple paths leading to the same top-level concept, whereas 34 were ambiguous and were manually assigned with the concept Endurant. The resulting datasets have 5 concepts (Conference-5c) and 6 concepts (Conference-6c) and their respective distributions are shown in Table 3.

Table 3: Distribution of concepts in the alignment between Conference and DOLCE with 5 concepts.

Endurant		Perdurant	Situation	Abstract	Quality
473 (77.8%)		95 (15.6%)	34 (5.6%)	6 (1.0%)	0 (0.0%)
Physical-endurant	Non-physical-endurant				
0	473 (77.8%)				

3.2. Learning Models

In [7], the prediction model relies on the use of labels and comments. The system is composed of two parts as can be seen in Figure 1 a). The first Feed-Forward Neural Network (FNN) part has as input the average of the embeddings of the words contained in the labels, whereas the second part consists of a BiLSTM [19] neural architecture that contextualizes learned embeddings for each word in the dataset comment. After passing through the BiLSTM, an average pooling is applied to generate the embedding representation of the whole comment. The BiLSTM part of this architecture has the same setting as ELMO [20] one.

However, using more robust architectures like BERT [21] may achieve improved results in this task as also reported in [17]. One of the reasons is that BERT can generate better natural language text representations due to its capacity of managing context. Another point is that some entities have the same label while being assigned to different top concepts in the dataset Lopes22-5c. This can hamper the model's ability to distinguish among the different concepts

[2] https://oaei.ontologymatching.org/2022/conference/index.html (on 01/07/23)

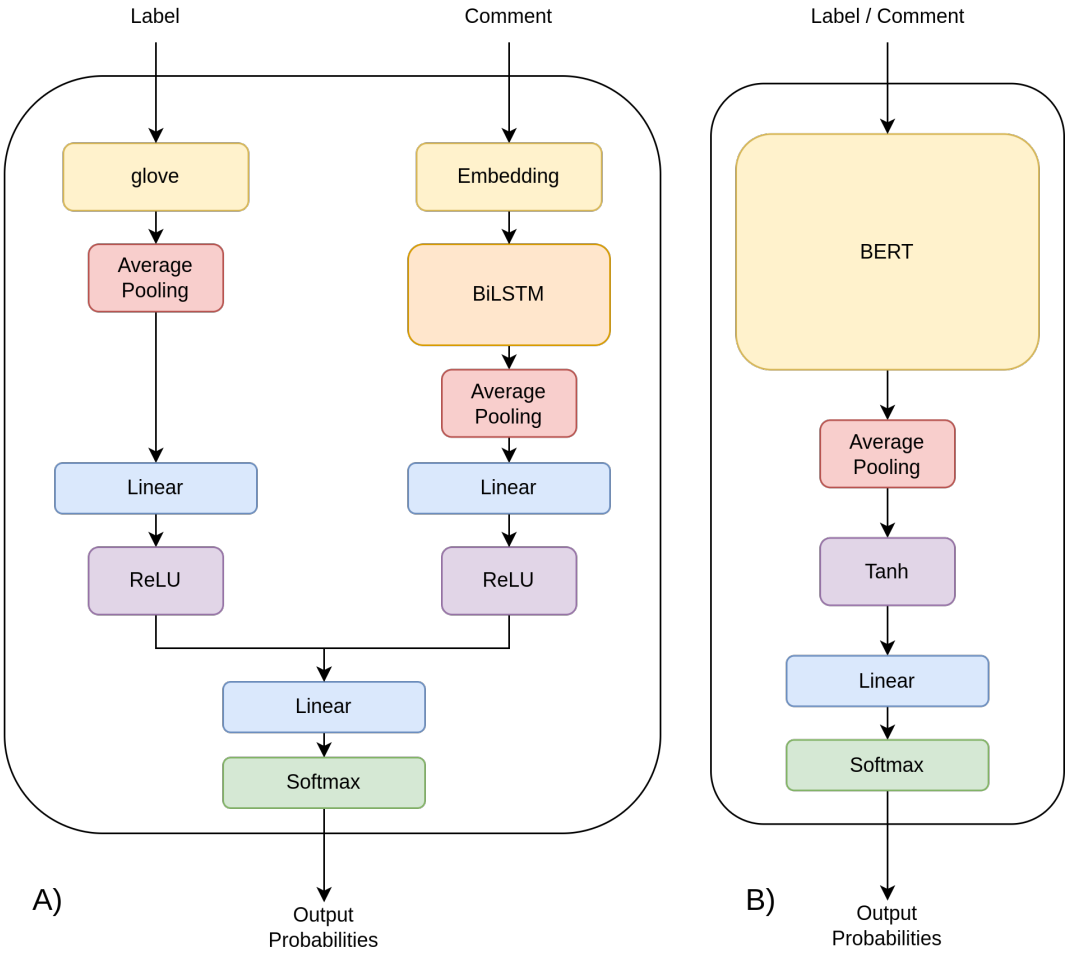

Figure 1: a) The proposed model in [7]. b) The BERT model.

while giving less importance to the label part of the input. Another issue is that, while comments can impact the training step, ontologies often contain a low amount of comments. Since the model makes a distinction between labels and comments, its capacity for generalization in the test phase can be reduced.

In this way, unifying the model's input between labels and comments can improve the model's performance since it will be able to take advantage of the information from comments during training while being able to work only with labels when comments are not present. Based on those assumptions, for better generalization, we used BERT with a classification head to predict the top-level concept that accepts a single text input that can be both labels or comments. The architecture using BERT is present in Figure1 b).

4. Experimental Evaluation

4.1. Performance of the Classification Models

In order to verify the hypothesis that the use of comment improves the baseline model results in BERT and in the model proposed in [7], the 3 datasets (Lopes22-5c, Sousa23-5c, and Sousa23-6c) were split using 10-fold cross-validation. Before training, the majority concept instances are reduced to match the number of instances in the minority concept using downsampling [22]. The exceeding entities are added to the test folds.

In [7], different word embeddings are tested, however, here we selected Glove 6B [23] because it provides a good balance between performance and model size. It also provides a more straightforward implementation compared to fastText [24] which is trained using character n-grams and needs a further tokenization procedure. Other baseline models were tested including Bernoulli Naive Bayes (BNB), Feed Forward Neural Network (FNN), Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Feed-forward Neural Network (FNN), and Support Vector Machine (SVM). The proposed model is [7] Model-Lopes and the FNN was trained using the Adam [25] optimizer with a learning rate of 0.001 for 10 epochs with a batch size of 64. The BERT model was also trained with Adam optimizer, with a learning rate of 0.00003, employing 1 epoch with a batch size of 64. The models are evaluated using the micro-F1 metric. For each model, we tested the following alternatives of input: using the comment only, the label only, and the label+comment. The results of the evaluation on the datasets Lopes22-5c, Sousa23-5c, and Sousa23-6c are presented in Table 4.

Table 4: Results of all models in terms of micro F1 score evaluated in the datasets Lopes22-5c, Sousa23-5c, Sousa23-6c.

	Lopes22-5c			Sousa23-5c			Sousa23-6c		
model	c	l	l+c	c	l	l+c	c	l	l+c
BERT	0.87	0.71	0.88	0.88	0.71	0.88	0.87	0.69	0.87
Model-Lopes	0.75	0.36	0.73	0.76	0.37	0.75	0.74	0.35	0.74
SVM	0.68	0.36	0.43	0.68	0.38	0.45	0.64	0.37	0.43
RF	0.64	0.38	0.37	0.65	0.38	0.38	0.58	0.36	0.36
FNN	0.59	0.33	0.39	0.58	0.34	0.39	0.57	0.32	0.38
LR	0.52	0.26	0.37	0.53	0.25	0.37	0.49	0.24	0.35
BNB	0.49	0.26	0.35	0.5	0.25	0.34	0.40	0.23	0.31
DT	0.40	0.27	0.25	0.41	0.27	0.25	0.40	0.25	0.23
GNB	0.31	0.43	0.43	0.30	0.44	0.44	0.46	0.33	0.37

One can notice that all classifiers achieved higher results using only comment as input, except the Gaussian Naive Bayes in the datasets Lopes22-5c and Sousa23-5c. The BERT model achieved the highest performance in all categories. And, in some cases, the BERT model obtained the same results even using label+comment input. One possible reason for that result is that the model can make better use of label information when appropriate due to the attention mechanism.

The confusion matrix for the BERT model in the 3 datasets can be seen in Tables 5, 6, and 7. One can see that, in the dataset Lopes22-5c, the model tends to misclassify a considerable

amount of Perdurants into Situation (16%), and Situations into Perdurant (17%). In the dataset Sousa23-5c, this misclassification is between Perdurant and Quality 21%. In the Sousa23-6c dataset, the model also misclassifies 13% of Perdurants into Situation and 23% of situations into Perdurant. Similar misclassification in Situation concept is found in [7], and these results may have a relation to the fact that Situation is nondisjoint with the other classes. In that sense, it is not an appropriate class for a top-level classification model if a single class is required for the task.

Table 5: BERT confusion matrix evaluated in (Lopes22-5c).

concept	Endurant	Perdurant	Quality	Abstract	Situation
Endurant	**753515 (88.9%)**	22208 (2.6%)	19023 (2.2%)	21977 (2.6%)	31062 (3.7%)
Perdurant	1584 (2.0%)	**62592 (77.7%)**	2220 (2.8%)	895 (1.1%)	13224 (16.4%)
Quality	116 (0.9%)	381 (2.9%)	**11666 (88.7%)**	397 (3.0%)	595 (4.5%)
Abstract	142 (3.5%)	31 (0.8%)	44 (1.1%)	**3795 (94.1%)**	23 (0.6%)
Situation	718 (1.7%)	7017 (17.0%)	1449 (3.5%)	230 (0.6%)	**31901 (77.2%)**

Table 6: BERT confusion matrix evaluated in (Sousa23-5c)

concept	Endurant	Perdurant	Quality	Abstract	Situation
Endurant	**682031 (89.1%)**	22033 (2.9%)	20942 (2.7%)	24965 (3.3%)	15884 (2.1%)
Perdurant	1410 (2.0%)	**54861 (78.1%)**	10339 (14.7%)	995 (1.4%)	2660 (3.8%)
Quality	571 (1.7%)	7052 (21.1%)	**24064 (72.1%)**	326 (1.0%)	1352 (4.1%)
Abstract	94 (2.1%)	59 (1.3%)	20 (0.4%)	**4047 (90.4%)**	255 (5.7%)
Situation	49 (1.2%)	130 (3.1%)	112 (2.6%)	99 (2.3%)	**3855 (90.8%)**

4.2. Evaluation on the OAEI Conference Datasets

The models trained on Sousa23-5c and Sousa23-6c were tested on Conference-5c and Conference-6c, respectively. Since the downsampling technique used for balancing the training datasets generates different dataset partitions, the results are evaluated in 150 steps to account for possible variations. In this evaluation phase, all the models receive as input only the labels of the entities, while the BERT model was trained only with comments. Model-Lopes is trained

Table 7: BERT confusion matrix evaluated in (Sousa23-6c)

concept	Physical-endurant	Non-physical-endurant	Perdurant	Quality	Abstract	Situation
Physical-endurant	**377547 (92.7%)**	12629 (3.1%)	2422 (0.6%)	2062 (0.5%)	10657 (2.6%)	2008 (0.5%)
Non-physical-endurant	9329 (2.9%)	**264356 (82.5%)**	13729 (4.3%)	8926 (2.8%)	11090 (3.5%)	12895 (4.0%)
Perdurant	437 (0.6%)	1145 (1.6%)	**55421 (78.9%)**	2430 (3.5%)	1412 (2.0%)	9420 (13.4%)
Quality	21 (0.5%)	27 (0.6%)	122 (2.9%)	**3688 (86.9%)**	218 (5.1%)	169 (4.0%)
Abstract	90 (2.0%)	26 (0.6%)	29 (0.6%)	144 (3.2%)	**4168 (93.1%)**	18 (0.4%)
Situation	203 (0.6%)	1001 (3.0%)	7693 (23.1%)	1308 (3.9%)	290 (0.9%)	**22870 (68.5%)**

and tested in different cases. Case 1: the model was trained only with labels and the test input is fed into the label input of the model. Case 2: the model was trained only with comments and the test input is fed into the comment input of the model. Case 3: the model was trained with both labels and comments and the test input is fed into the label input of the model. And Case 4: the model was trained only with both labels and comments and the test input is fed into the comment input of the model. The results are present in Table 8.

In the results, the BERT model trained with the adopted hyperparameter settings is unstable and has the highest standard deviation. This model collapses in some cases, giving the same output for every input, causing it to have values ranging between 0 and 0.78 F-measure. The model achieves 0.00 F-1 when it outputs Quality for every input, as the test dataset does not have any element labeled as Quality. On the other hand, it achieves the highest F-1 (0.78) when the model predicts Endurant for every input as, in the test dataset, 0.78% of the entities are Endurant. The BERT model had the highest scores in Conference-5c when considering the 75% percentile results, excluding instances of collapse. In Conference-6c, the BERT model and the model trained and tested with comments had the best and nearly equal results. Since the BERT model achieves the highest scores on Conference test datasets, it was selected to analyze the OAEI datasets described in Section 5.

5. Applying the Best Classifier: An Evaluation on the OAEI Tracks

This section presents an analysis of the top-level concepts in different OAEI tracks (Conference, Anatomy, Complex, Food, BioML, BioDiv, MSE, and KG[3] along with the characteristics of

[3]These tracks are described on https://oaei.ontologymatching.org/2022/ (on 01/07/23)

Table 8: Results of the models in the Conference test dataset.

model	mean	std	min	25%	50%	75%	max
BERT (Sousa23-5c)	0.45	0.25	0.0	0.17	0.54	**0.66**	**0.78**
BERT (Sousa23-6c)	0.41	0.23	0.0	0.29	0.48	0.6	**0.78**
Model-Lopes (l) (Sousa23-5c)	0.37	0.03	0.29	0.35	0.37	0.4	0.46
Model-Lopes (l) (Sousa23-6c)	0.19	0.03	0.12	0.17	0.19	0.21	0.25
Model-Lopes (c) (Sousa23-5c)	0.5	0.08	0.27	0.45	0.49	0.54	0.72
Model-Lopes (c) (Sousa23-6c)	0.55	0.07	**0.38**	0.51	0.56	0.61	0.71
Model-Lopes (l+c test l) (Sousa23-5c)	0.13	0.04	0.05	0.11	0.13	0.15	0.27
Model-Lopes (l+c test l) (Sousa23-6c)	0.2	0.08	0.07	0.15	0.2	0.25	0.52
Model-Lopes (l+c test c) (Sousa23-5c)	**0.56**	0.07	**0.38**	**0.52**	**0.57**	0.61	0.72
Model-Lopes (l+c test c) (Sousa23-6c)	0.38	0.07	0.22	0.34	0.38	0.43	0.58

comments present in ontology entities. In the first subsection, an analysis of the distribution of top concepts in the ontologies is provided. The second subsection evaluates the consistency of the reference alignments in terms of their top-level concepts. The third subsection evaluates the distribution of label length compared to the comments present in the training datasets.

5.1. Distribution of Top Concepts

Using the best model (BERT), the distribution of the top-level types of the concepts in the ontologies from schema alignment tracks is estimated. The distribution concerns the entities of each ontology present in the tracks, excluding blank nodes, and properties. For each entity, one label is collected searching for label predicates *rdfs.label*, *skos.prefLabel*, *skos.altLabel*, or if no label is found the label is retrieved from the resource identifier. As can be seen in Figure 2, the distribution of concepts for each track is distinct. The estimation by the model trained in Sousa23-5c shows that Complex, Food, and BioDiv have a high concentration of Endurants while the others are more distributed.

The distribution by the model trained on Sousa23-6c is presented in Figure 3 as well as its estimation. The Anatomy, Food, BioML, and KG ontologies have a high concentration of entities in one concept. For BioDiv, the majority of entities concentrate on Physical-endurant and Non-physical-endurant. The two models disagree with the distribution of Quality entities between tracks. The model trained on Sousa23-5c tends to classify some Endurant as Quality compared to the model trained on Sousa23-6c. The two models achieve similar distributions of Perdurant and Abstract types.

5.2. Alignment Consistency across Correspondence Entities

We analyzed the number of correspondences having entities associated with the same top-level types (a correspondence is composed of a source and a target ontology entities). The proportion can be seen in Table 9 for the estimations given by the model trained on Sousa23-5c and Sousa23-6c. The distribution of correspondences that have the same type is similar for the two models in Conference, MSE, CommonKG, BioML, and KG. The difference in scores for the

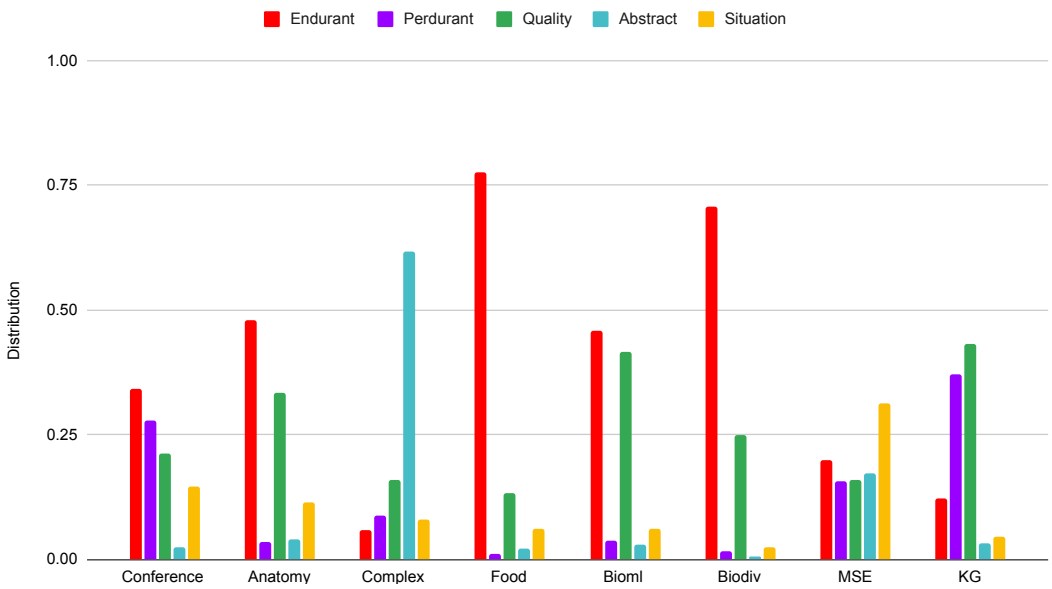

Figure 2: OAEI tracks distribution with 5 concepts.

Anatomy track is related to the distribution given by each model. For the model Sousa23-5c, the majority of entities are distributed as Endurant and Quality. In contrast, as the model Sousa23-6c yielded a high concentration of entities in the same concept, the reference alignments also follow the same tendency. In BioDiv, the model trained on Sousa23-6c achieves only 9.65% of the alignments with the same type for the alignment between Agrovoc and Nat ontologies as the majority of the difference between them is that the model gives physical-endurant for one entity, and non-physical-endurant for the other. This problem does not happen with the model trained on Sousa23-5c since both will be classified as Endurant and so, the alignments will have the same type.

Table 9: Number of reference alignments with the same type in source and target ontologies using both models.

	BERT Sousa23-5c			BERT Sousa23-6c		
Track	Same	Total	%	Same	Total	%
Conference	213	258	82.56	197	258	76.36
Anatomy	908	1516	59.89	1248	1516	82.32
MSE	159	388	40.98	162	388	41.75
CommonKG	268	304	88.16	264	304	86.84
BioDiv	83427	96483	86.47	42807	96483	44.37
BioML	18007	25270	71.26	18158	25270	71.86
KG	13627	15359	88.72	13452	15359	87.58

Both models trained in 5 and 6 classes yielded a few correspondences with the same types

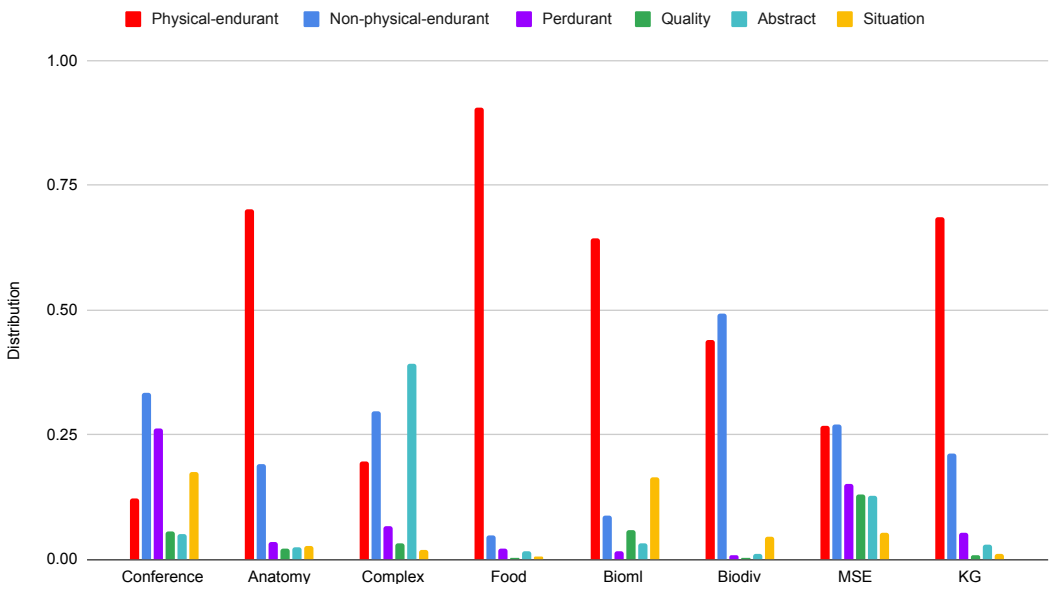

Figure 3: OAEI tracks distribution with 6 concepts.

in the MSE track. The main reason resides in the lack of information in the entities, such as acronyms for chemical elements, in which the model has enough information to give the correct classification. A similar problem appears in the Conference datasets, for instance, between the alignments of PaperAbstract and Abstract or Country and State. Without further information, the label State is ambiguous, and the model can yield an incorrect classification based on it. As can be seen in these examples, a dataset containing entities along with its contained subgraphs could lead to better results for this task, as the model should have more information to deal with the ambiguous labels.

Table 10: Some examples where the model predicts different types for the reference alignments in Conference and MSE test data.

Track	Source	Target
Conference	PaperAbstract (Endurant)	Abstract (Situation)
	Country (Situation)	State (Quality)
	Attendee (Quality)	Listener (Situation)
MSE	Mt (Endurant)	MeitneriumAtom (Perdurant)
	Dy (Situation)	DysprosiumAtom (Endurant)
	Ag (Endurant)	Silver (Situation)

5.3. Discussion of Terminological Distribution

As entity comments are equivalent to ontology comments, the models are expected to have the best results when evaluated on them. However, comments are rare in the ontologies present in

all tracks, and in this case, the labels need to be used to predict the top level. The distribution of labels and comments in all ontologies for schema matching is analyzed to verify the relation to the distribution of the proposed datasets. The number of labels and comments in all tracks are present in Table 11. Among all tracks, Anatomy, Food, and BioDiv have no comments. BioML (MONDO), BioML (UMLS) and BioDiv have less than 1% of comments. Conference and Complex have less than 5% and KGh while MSE have respectively 13.49% and 36.17%. Since the number of comments is low, the labels need to be used to predict the top-level types. However, as most machine learning models suffer from the well-known Out-of-Distribution Generalization (OOD) [26] problem, they are hampered by both the labels that do not have a similar syntactic structure of their comments and their length distribution differs.

Table 11: Count of labels and comments in the analyzed tracks.

track	entities	rdfs.label	skos.prefLabel	skos.altLabel	rdfs.comment
Conference	1020	0 (0.00%)	0 (0.00%)	0 (0.00%)	32 (3.14%)
Anatomy	12498	12017 (96.15%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Complex	32620	9565 (29.32%)	0 (0.00%)	0 (0.00%)	434 (1.33%)
Food NC	11737	0 (0.00%)	9396 (80.05%)	0 (0.00%)	0 (0.00%)
BioML (MONDO)	34872	33755 (96.80%)	0 (0.00%)	0 (0.00%)	60 (0.17%)
BioML (UMLS)	145486	145093 (99.73%)	55312 (38.02%)	23948 (16.46%)	43 (0.03%)
BioDiv (ncbitaxon)	233776	233746 (99.99%)	0 (0.00%)	5861 (2.51%)	0 (0.00%)
BioDiv	3599317	2582980 (71.76%)	359314 (9.98%)	155987 (4.33%)	882 (0.02%)
MSE	2411	953 (39.53%)	451 (18.71%)	47 (1.95%)	872 (36.17%)
KG	2050682	808385 (39.42%)	321695 (15.69%)	708059 (34.53%)	276631 (13.49%)

The frequency of the lengths of labels and comments for each entity in all ontologies of all tracks is compared to the distribution of the comment lengths in the Sousa23-5c dataset. As can be seen in Figure 4, the average length of the comments in the training datasets is 50 characters, however, it was noticed that the majority of the labels are shorter. Furthermore, comments, which are relatively rare, have a high standard deviation. Also, the differences in the distribution of text length between labels and comments hinder the capacity for generalization of the models.

6. Conclusion and Future Work

In this work, we investigated the task of top-level concept prediction. We generated one dataset with 5 top-level concepts (Sousa23-5c) and one with 6 concepts (Sousa23-6c) based on OntoWordNet. In this generation, we discussed the multi-inheritance problem and proposed procedures to obtain top-level concepts for entities in unambiguous cases. In addition, classifier models were then trained and tested varying between the use of only labels, comments, and labels+comments. The yielded results show that the use of `rdfs:comment` improves the prediction performance of classification models. These results show the importance of `rdfs:comment` for automated system understanding of concepts. We selected the best-generated model to estimate the distribution of concepts in ontologies from well-known ontology matching benchmarks (OAEI).

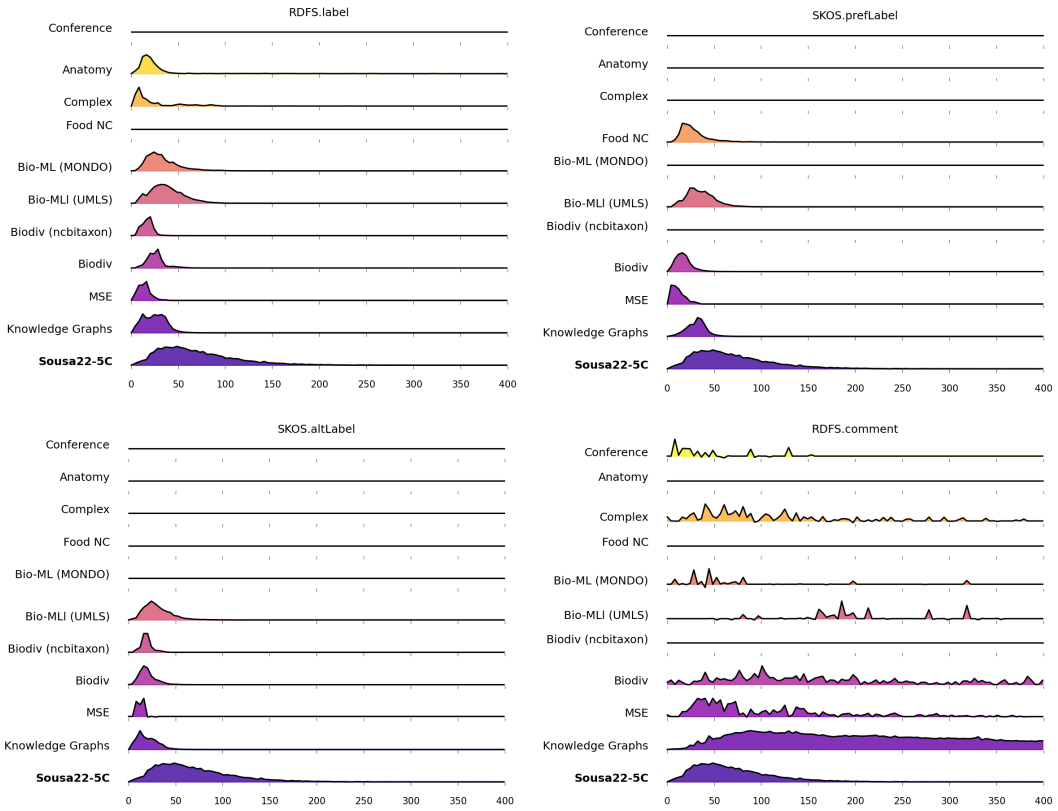

Figure 4: Distribution between labels and comments in the analyzed tracks compared with the Sousa23-5c comments. The x-axis represents the length of the text and the y-axis is the number of labels or comments in each length. The visualization is limited to texts ranging from 0 to 400 characters long.

The results show that tracks have different distributions among top-level types. The performed analysis of the reference alignments showed that a high number of correspondences, as expected, are of the same type. We consider that this gives us an estimation of the accuracy of the trained classifiers.

In future work, we intend to conduct experiments with new deep-learning architectures that should improve the results reported in this paper. Dynamically predicting the top-level types for each concept of an ontology should help in downstream tasks such as ontology matching. Therefore, as different ontologies have distinct top-level distributions, we expect that our present analysis could be used for generating better classification models in the near future. Since the labels of entities are ambiguous in some cases, including the ontology structure as contextual information for the classification models may improve the prediction of top-level concept types. Also, the high number of correspondences that have the same type in some OAEI tracks shows that using these tools could help increase matching systems performance by increasing the similarity of entities with the same top-level type.

References

[1] M. McDaniel, V. C. Storey, Evaluating domain ontologies: Clarification, classification, and challenges, ACM Comput. Surv. 52 (2019) 70:1–70:44. URL: https://doi.org/10.1145/3329124. doi:10.1145/3329124.

[2] I. G. Husein, S. Akbar, B. Sitohang, F. N. Azizah, Review of ontology matching with background knowledge, in: 2016 International Conference on Data and Software Engineering (ICoDSE), IEEE, 2016, pp. 1–6.

[3] C. Trojahn, R. Vieira, D. Schmidt, A. Pease, G. Guizzardi, Foundational ontologies meet ontology matching: A survey, Semantic Web 13 (2022) 685–704. URL: https://doi.org/10.3233/SW-210447. doi:10.3233/SW-210447.

[4] A. Gangemi, R. Navigli, P. Velardi, The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet, in: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003, volume 2888 of *Lecture Notes in Computer Science*, Springer, 2003, pp. 820–838. URL: https://doi.org/10.1007/978-3-540-39964-3_52. doi:10.1007/978-3-540-39964-3_52.

[5] G. A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (1995) 39–41. URL: http://doi.acm.org/10.1145/219717.219748. doi:10.1145/219717.219748.

[6] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, DOLCE: A descriptive ontology for linguistic and cognitive engineering, Appl. Ontology 17 (2022) 45–69. URL: https://doi.org/10.3233/AO-210259. doi:10.3233/AO-210259.

[7] A. G. L. Junior, J. L. Carbonera, D. Schimdt, M. Abel, Predicting the top-level ontological concepts of domain entities using word embeddings, informal definitions, and deep learning, Expert Syst. Appl. 203 (2022) 117291. URL: https://doi.org/10.1016/j.eswa.2022.117291. doi:10.1016/j.eswa.2022.117291.

[8] M. A. N. Pour, A. Algergawy, P. Buche, L. J. Castro, J. Chen, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, M. Huschka, L. Ibanescu, E. Jiménez-Ruiz, N. Karam, A. Laadhar, P. Lambrix, H. Li, Y. Li, F. Michel, E. Nasr, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, C. Trojahn, C. Verhey, M. Wu, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2022, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Hangzhou, China, held as a virtual conference, October 23, 2022, volume 3324 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 84–128.

[9] H. Paulheim, A. Gangemi, Serving DBpedia with DOLCE - More than Just Adding a Cherry on Top, in: The Semantic Web, 2015, pp. 180–196.

[10] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, Sweetening WORDNET with DOLCE, AI Magazine 24 (2003) 13–24.

[11] V. Silva, M. Campos, J. Silva, M. Cavalcanti, An Approach for the Alignment of Biomedical Ontologies based on Foundational Ontologies, Information and Data Management 2 (2011) 557–572.

[12] V. Mascardi, A. Locoro, P. Rosso, Automatic ontology matching via upper ontologies:

A systematic evaluation, IEEE Trans. Knowl. Data Eng. 22 (2010) 609–623. URL: https://doi.org/10.1109/TKDE.2009.154. doi:10.1109/TKDE.2009.154.
[13] N. F. Padilha, F. Baião, K. Revoredo, Alignment Patterns based on Unified Foundational Ontology, in: Proc. of the Brazilian Ontology Research Seminar, 2012, pp. 48–59.
[14] P. Mika, D. Oberle, A. Gangemi, M. Sabou, Foundations for Service Ontologies: Aligning OWL-S to Dolce, in: Proc. of the 13th Conf. on World Wide Web, 2004, pp. 563–572.
[15] R. Stevens, P. Lord, J. Malone, N. Matentzoglu, Measuring expert performance at manually classifying domain entities under upper ontology classes, J. Web Semant. 57 (2019). URL: https://doi.org/10.1016/j.websem.2018.08.004. doi:10.1016/j.websem.2018.08.004.
[16] D. Schmidt, R. Basso, C. Trojahn, R. Vieira, Matching domain and top-level ontologies exploring word sense disambiguation and word embedding, in: E. Demidova, A. Zaveri, E. Simperl (Eds.), Emerging Topics in Semantic Technologies - ISWC 2018 Satellite Events [best papers from 13 of the workshops co-located with the ISWC 2018 conference], volume 36 of *Studies on the Semantic Web*, IOS Press, 2018, pp. 27–38. URL: https://doi.org/10.3233/978-1-61499-894-5-27. doi:10.3233/978-1-61499-894-5-27.
[17] A. Lopes, J. L. Carbonera, D. Schmidt, L. F. Garcia, F. H. Rodrigues, M. Abel, Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models, Knowl. Based Syst. 265 (2023) 110385. URL: https://doi.org/10.1016/j.knosys.2023.110385. doi:10.1016/j.knosys.2023.110385.
[18] D. Schmidt, C. Trojahn, R. Vieira, M. Kamel, Validating top-level and domain ontology alignments using wordnet, in: Proceedings of the IX ONTOBRAS Brazilian Ontology Research Seminar, Curitiba, Brazil, October 3rd, 2016, volume 1862 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 119–130.
[19] A. Graves, N. Jaitly, A. Mohamed, Hybrid speech recognition with deep bidirectional LSTM, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013, IEEE, 2013, pp. 273–278. URL: https://doi.org/10.1109/ASRU.2013.6707742. doi:10.1109/ASRU.2013.6707742.
[20] M. E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 1756–1765. URL: https://doi.org/10.18653/v1/P17-1161. doi:10.18653/v1/P17-1161.
[21] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.
[22] Y. Sun, A. K. Wong, M. S. Kamel, Classification of imbalanced data: A review, International journal of pattern recognition and artificial intelligence 23 (2009) 687–719.
[23] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[24] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword

information, Trans. Assoc. Comput. Linguistics 5 (2017) 135–146. URL: https://doi.org/10.1162/tacl_a_00051. doi:10.1162/tacl_a_00051.

[25] Z. Zhang, Improved adam optimizer for deep neural networks, in: 26th IEEE/ACM International Symposium on Quality of Service, IWQoS 2018, Banff, AB, Canada, June 4-6, 2018, IEEE, 2018, pp. 1–2. URL: https://doi.org/10.1109/IWQoS.2018.8624183. doi:10.1109/IWQoS.2018.8624183.

[26] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: A survey, CoRR abs/2108.13624 (2021). URL: https://arxiv.org/abs/2108.13624. arXiv:2108.13624.