Mutilingual Hierarchical Expected Answer Type Classification using the SMART 2021 Dataset

Aleksandr Perevalov and Andreas Both

Anhalt University of Applied Sciences, Köthen (Anhalt), Germany {aleksandr.perevalov,andreas.both}@hs-anhalt.de

Abstract. The Knowledge Graph Question Answering (KGQA) systems are required to understand natural language in order to transform textual questions into structured queries to a knowledge graph. One of the important tasks of natural-language understanding (NLU) in the context of KGQA systems is the classification of the expected answer type (EAT). In this paper, we present our approach on the EAT classification within the SeMantic Answer Type Prediction Task 2021. The approach is based on machine-translation-based data augmentation, it supports 104 input languages¹ and works over DBpedia and Wikidata. The obtained evaluation results demonstrate reasonable quality in comparison to both last year's and this year's solutions.

Keywords: Expected Answer Type Classification \cdot Target Type Identification \cdot Multilingual Question Answering \cdot Knowledge Graph Question Answering.

1 Introduction

The expected answer type (EAT, or answer type) prediction in the context of knowledge graph question answering (KGQA) is one of the tasks for natural-language understanding (NLU). The ability to know the EAT enables the system to significantly narrow down the search space for an answer to a given question. For example, given the following question: "In what city was Angela Merkel born?" it is possible to reduce the search space from 861 to 6 entities while querying all the 1-hop entities of the Angela Merkel resource in the DBpedia [1] knowledge base dbr:Angela_Merkel². The illustration for the corresponding example is in Figure 1.

In this work, we present our solution for the 2021 SeMantic Answer Type and Relation Prediction³ (SMART) Task [11] over the DBpedia and Wikidata datasets in the Answer Type Prediction track (SMART2021-AT). In the simplest

https://github.com/google-research/bert/blob/master/multilingual.md# list-of-languages

² dbr:Angela_Merkel is the prefix-based shortcut for the DBpedia resource http: //dbpedia.org/resource/Angela_Merkel

³ cf. https://smart-task.github.io/2021/

```
# without EAT prediction
SELECT (COUNT(DISTINCT ?obj) as ?count)
WHERE {
    dbr:Angela_Merkel ?p ?obj .
}
# ?count = 861
# with EAT prediction
SELECT (COUNT(DISTINCT ?obj) as ?count)
WHERE {
    dbr:Angela_Merkel ?p ?obj .
    ?obj rdf:type ?type .
FILTER(?type = dbo:City)
}
# ?count = 6
```

Fig. 1. Motivation for the EAT classification in KGQA. The example question: "In what city was Angela Merkel born?"

case, the EAT classification task can be treated as a multi-class text classification task. In the context of the SMART task, the data structure is more sophisticated. There are two class levels: answer category (resource, literal, boolean) and answer type. Hence, the class taxonomy is not flat and requires approaches for hierarchical classification.

The official description of the data states⁴: If the category is "resource", answer types are ontology classes from either the DBpedia ontology⁵ or the Wikidata ontology⁶. If the category is "literal", answer types are either "number", "date", or "string". For the category "boolean", no additional specialization is defined. The number of unique "resource" classes is high. Moreover, they are represented in the form of a list.

While following our long-term research agenda of increasing the accessibility of KGQA systems and their components through multilingualization, the presented solution is mainly based on the multilingual language models and datasets (as in our last year iteration [13]). We used open-source machine translation models [17] to translate the provided data into 10 languages (German, Spanish, Mandarin Chinese, Italian, Romanian, Vietnamese, Russian, French, Czech, Japanese). Thereafter, a multilingual BERT-based [7] classifier was fine-tuned on the original and translated data. We used a multi-level classification pipeline to make the final predictions. In the conclusion of the paper, we discuss the final results as well as our findings during the working process.

This work is structured as follows, in Section 2 we review the related research, Section 3 describes the exploratory data analysis of the provided datasets, we describe our approach in Section 4 and present the evaluation results in Section 5. Section 6 concludes the paper.

2 Related Work

The entity- and type-centric models were introduced in [2] to identify the answer type of a question. These models are used to rank the queries given the entity- or type-related content [9]. The idea of incorporating an additional context to improve answer type predictions was proposed in work [18]. One of the

⁴ https://smart-task.github.io/

⁵ http://mappings.dbpedia.org/server/ontology/classes/

⁶ https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology

ISWC 2020's Semantic Web challenges was addressing the answer type classification (SeMantic Answer Type prediction task, SMART) [10]. It has shown that transformer-based models demonstrate the highest results in this task [16,12]. The approach based on using external data (e.g., KGQA datasets) was introduced in [14]. Recently, the authors of [6] proposed a system for EAT prediction in a "distantly supervised fashion" (i.e., no manual data annotation is required).

3 Exploratory Data Analysis

There are two class levels in the datasets: answer type category (resource, literal, boolean) and answer type. As the class taxonomy is not flat it requires us to use approaches for hierarchical classification. The official description of the data states⁷: If the category is "resource", answer types are ontology classes from either the DBpedia ontology⁸ or the Wikidata ontology⁹. If the category is "literal", answer types are either "number", "date", or "string". For the category "boolean" no additional specialization is defined. Such data has to be analyzed on different levels for class distribution, noise, missing values, etc. In this section, we demonstrate the analysis of both DBpedia and Wikidata answer type classification datasets.

3.1 DBpedia Dataset

The DBpedia dataset contains Train (40,621 examples) and Test (10,093 examples) subsets. After removing the null values from the Train subset, the number of examples decreased to 37,061 (-3,560 examples). The example of the DBpedia dataset for the answer type prediction is shown in Figure 2. Here, if the category is resource, the answer type field is represented as an ordered list of DBpedia ontology types, where the first type is the most specific and the last is the most general one (see the very first question in the Figure). While manually analyzing this year's SMART dataset we were capable to find several noisy examples (see Figure 3). In this example, the type field contains duplicated DBpedia ontology types.

We also analyzed the distributions of values for category, literal, type fields. The results are demonstrated in Figure 4. There is a huge imbalance towards the resource value while the literal values are more or less balanced. The distribution of the resource field values is demonstrated in Figure 5. Surprisingly, the most frequent values of the answer type field contain noise. For example, the second most frequent value has incorrect order of DBpedia types. The third most frequent value has DBpedia types of different hierarchies (dbo:Place and dbo:Agent).

⁷ https://smart-task.github.io/

⁸ http://mappings.dbpedia.org/server/ontology/classes/

⁹ https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology

A. Perevalov, A. Both

4

Fig. 2. An example of DBpedia dataset

```
{
    "id":26178,
    "question":"What kind of music is the album farewell aldebaran",
    "category":"resource",
    "type":[
         "dbo:Genre", "dbo:TopicalConcept", "dbo:MusicGenre",
         "dbo:Genre", "dbo:TopicalConcept", "dbo:MusicGenre"
]
}
```

Fig. 3. A noisy example in the DBpedia dataset

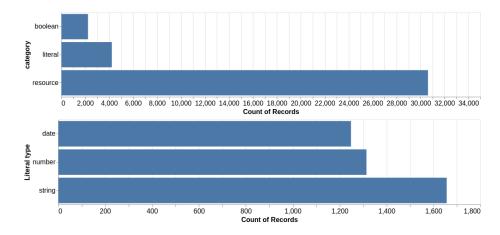


Fig. 4. Data distribution for the category and literal literal values for DBpedia training subset

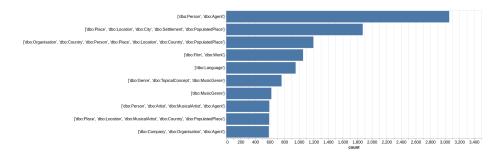


Fig. 5. Data distribution of the type values for the Wikidata training subset

Fig. 6. An example of Wikidata dataset. Only "resource" category is demonstrated as the other ones are the same as in DBpedia dataset in Figure 2.

3.2 Wikidata Dataset

The Wikidata dataset contains Train (43,604 examples) and Test (10,864 examples) subsets. We also cleaned the training data by removing the null values. Consequently, the number of examples decreased to 43,554 (-50 examples). In this aspect, the Wikidata data has significantly fewer null values rather than the DBpedia data. An example of the Wikidata dataset for the answer type prediction is shown in Figure 6.

Here, if the category is resource, the answer type field is represented as a list of Wikidata classes that are retrieved according to the following SPARQL query:

```
PREFIX wd: <a href="http://www.wikidata.org/entity/">http://www.wikidata.org/entity/>
PREFIX wdt: <a href="http://www.wikidata.org/prop/direct/">http://www.wikidata.org/prop/direct/>
SELECT ?subClasses WHERE {
    wd:Q123456789 wdt:P31 ?x . # subject to be replaced with actual answer entity
    ?x wdt:P279 ?subClasses .
}
```

While manually analyzing the training subset of Wikidata we were capable to find several noisy examples of different nature in comparison to DBpedia (see Figure 7). In this example, the type field contains duplicated DBpedia ontology types.

We also analyzed the distributions of values for category, literal, type fields. The results are demonstrated in Figure 8. For Wikidata we observed the same data distribution patterns: there is a huge imbalance towards the resource

```
6 A. Perevalov, A. Both
```

Fig. 7. A noisy example in the Wikidata dataset

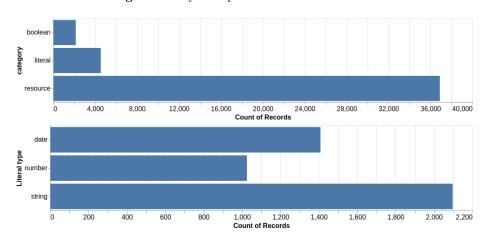


Fig. 8. Data distribution for the category and literal literal values for Wikidata training subset

value while the literal values are more or less balanced. The distribution of the resource field values is demonstrated in Figure 9. Obviously, the answer type values are extremely imbalanced towards the ones related to the person class.

3.3 Summary

We assume that the observed noise in the data was unintended and there was a risk of having the data quality in the test dataset. Hence, we decided not to use this year's SMART Task data. In our local training and evaluation process, the data from the previous year was used [10]. The corresponding data analysis for the previous year's data is available in our paper [13].

4 Approach

While following our long-term research agenda on enhancing multilingual accessibility of KGQA systems [15], we base our approach on multilingual data augmentation. We used only the data from SMART 2020 (the previous year's challenge). For both DBpedia and Wikidata, all the textual questions in the corresponding datasets were machine-translated from English into German, Spanish,

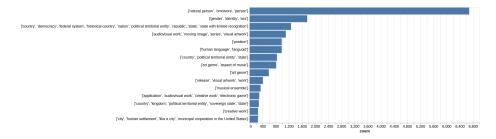


Fig. 9. Data distribution for the answer type values for the Wikidata training subset

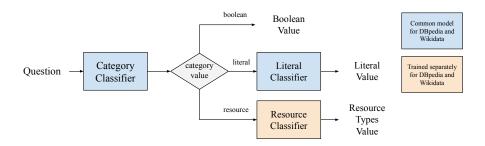


Fig. 10. Architecture of the classification pipeline

Chinese, Italian, Romanian, Vietnamese, Russian, French, Czech, and Japanese using Helsinki NLP tool [17]. For the DBpedia dataset, we fetched additional data from LC-QuAD 1.0 [19] dataset. The same was done w.r.t. the Wikidata dataset and LC-QuAD 2.0 [8].

We used a multi-level hierarchical classification pipeline for both DBpedia and Wikidata. The pipeline consists of the following models: (1) category classifier, (2) literal classifier – are common models for both knowledge graphs, and (3) resource classifier – trained separately for DBpedia and Wikidata. The architecture of the classification pipeline is shown in Figure 10. The resource classifier for DBpedia was trained in the multi-class classification setting to predict the most specific type of the hierarchy. When the prediction is executed, the rest of the hierarchy is fetched from DBpedia via the following SPARQL query based on the predicted type:

The resource classifier for Wikidata was trained in the multi-label classification setting. It works without any additional steps in the prediction phase.

DBpedia	Accuracy	NDCG@5	NDCG@10
	0.991	0.643	0.577
Wikidata	Accuracy	\mathbf{MRR}	
	0.980	0.430	

Table 1. Final results obtained on the private test set by organizers

5 Evaluation Results

For the classification model, BERT multilingual model [7] was used. We utilized "transformers" ¹⁰ Python library for implementing the classification pipeline. For the multi-class classifiers (category, literal, resource for DBpedia), one fully-connected layer of size n was added to the BERT model, where n – is the number of classes. The input for this layer was the last hidden state of the BERT model (i.e., [CLS] token). We used categorical cross-entropy [5] as a loss function for the multi-class models. The multi-label classifier – resource for Wikidata – was also provided with a fully connected layer of the same size, however, binary cross-entropy loss [5] was used as each of the output neurons represents the probability of predicted label. The training of the models was done using early stopping criteria targeted at minimizing the loss with the patience of one epoch.

The prediction results were evaluated with the following metrics. For both DBpedia and Wikidata, the quality of the category predictions was measured using accuracy. The quality of the answer type predictions was measured using Mean Reciprocal Rank (MRR) for Wikidata and lenient Normalized Discounted Cumulative Gain @k with a Linear decay (NDCG@k) for DBpedia, where $k \in 5, 10$ [3].

The evaluation was done by organizers on the private dataset using an internal process. The final results are shown in Table 1. The obtained results are demonstrating reasonable quality in comparison to the other participants as well as to the results of the last year. The Accuracy score of the solution demonstrates one of the best results on both DBpedia and Wikidata, while metrics related to the ranking of the answer type hierarchies (NDCG@k, MRR) achieved relatively poor results. The source code of our solution is available online¹¹. In addition, we also have deployed a demo-interface for EAT classification over DBpedia online¹².

6 Discussion and Conclusion

We would like to raise the following questions concerning the evaluation process. First, as we observed a significant imbalance in the data w.r.t. the category values (see Section 3), we think that the usage of accuracy score to measure

¹⁰ https://huggingface.co/bert-base-multilingual-cased

¹¹ https://github.com/Perevalov/smart-2021

 $^{^{12}\ \}mathtt{https://webengineering.ins.hs-anhalt.de:41009/eat-classification}$

the results is not the best option since that it is not robust to the imbalanced data. Instead, we propose to use precision and recall scores computed in the classification setting. Secondly, as the answer types of the resource category questions in Wikidata are not ordered and do not form a hierarchy, we think that the usage of mean reciprocal rank is not acceptable as this metric is used to evaluate ordered result sets. Hence, the measure for unordered lists is naturally applicable to this task, such as precision and recall calculated in an information-retrieval setting.

In this paper, we demonstrated our approach for hierarchical EAT prediction based on a multi-level classification pipeline. As we used multilingual data and models for training, the classification pipeline supports input in 104 languages. The evaluation process demonstrated reasonable results w.r.t. the quality metrics. For future work, we are targeting on improving the quality of the resource answer types classification, and we are planning to create a study on the impact of EAT classification on QA quality for multiple QA systems using a component-oriented QA framework (e.g., [4]).

References

- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web. pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
- 2. Balog, K., Neumayer, R.: Hierarchical target type identification for entity-oriented queries. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2391–2394. CIKM '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2396761.2398648
- Balog, K., Neumayer, R.: Hierarchical target type identification for entity-oriented queries. pp. 2391–2394 (10 2012). https://doi.org/10.1145/2396761.2398648
- 4. Both, A., Diefenbach, D., Singh, K., Shekarpour, S., Cherix, D., Lange, C.: Qanary a methodology for vocabulary-driven open question answering systems. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) The Semantic Web. Latest Advances and New Domains. pp. 625–641. Springer International Publishing, Cham (2016)
- 5. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958)
- Dash, S., Mihindukulasooriya, N., Gliozzo, A., Canim, M.: Type prediction systems. CoRR (2021), https://arxiv.org/abs/2104.01207
- 7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv e-prints (2018)
- 8. Dubey, M., Banerjee, D., Abdelkawi, A., Lehmann, J.: LC-QuAD 2.0: A large dataset for complex question answering over Wikidata and DBpedia. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web ISWC 2019. pp. 69–78. Springer International Publishing, Cham (2019)

- Garigliotti, D., Hasibi, F., Balog, K.: Target type identification for entity-bearing queries. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 845–848. SIGIR '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3080659
- Mihindukulasooriya, N., Dubey, M., Gliozzo, A., Lehmann, J., Ngomo, A.C.N., Usbeck, R.: SeMantic Answer Type prediction task (SMART) at ISWC 2020 Semantic Web Challenge. CoRR/arXiv abs/2012.00555 (2020), https://arxiv.org/abs/2012.00555
- 11. Mihindukulasooriya, N., Dubey, M., Gliozzo, A., Lehmann, J., Ngonga Ngomo, A.C., Usbeck, R., Rossiello, G., Kumar, U.: Semantic answer type and relation prediction task (smart 2021). arXiv (2022)
- 12. Nikas, C., Fafalios, P., Tzitzikas, Y.: Two-stage semantic answer type prediction for question answering using BERT and class-specificity rewarding. In: Proceedings of the SeMantic AnsweR Type prediction task (SMART) at ISWC 2020. CEUR Workshop Proceedings, vol. 2774, pp. 19–28. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2774/paper-03.pdf
- Perevalov, A., Both, A.: Augmentation-based answer type classification of the SMART dataset. In: Proceedings of the SeMantic Answer Type prediction task (SMART) at ISWC 2020. CEUR Workshop Proceedings, vol. 2774, pp. 1–9. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2774/paper-01.pdf
- Perevalov, A., Both, A.: Improving answer type classification quality through combined question answering datasets. In: Knowledge Science, Engineering and Management. pp. 191–204. Springer International Publishing, Cham (2021)
- 15. Perevalov, A., Diefenbach, D., Usbeck, R., Both, A.: QALD-9-plus: A multilingual dataset for question answering over DBpedia and Wikidata translated by native speakers. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). IEEE (2022)
- 16. Setty, V., Balog, K.: Semantic answer type prediction using BERT IAI at the ISWC SMART task 2020. In: Proceedings of the SeMantic Answer Type prediction task (SMART) at ISWC 2020. CEUR Workshop Proceedings, vol. 2774, pp. 10–18. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2774/paper-02.pdf
- 17. Tiedemann, J., Thottingal, S.: OPUS-MT Building open translation services for the World. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)
- Tonon, A., Catasta, M., Prokofyev, R., Demartini, G., Aberer, K., Cudré-Mauroux, P.: Contextualized ranking of entity types based on knowledge graphs. Journal of Web Semantics 37-38, 170–183 (2016). https://doi.org/10.1016/j.websem.2015.12.005
- Trivedi, P., Maheshwari, G., Dubey, M., Lehmann, J.: LC-QuAD: A corpus for complex question answering over knowledge graphs. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) The Semantic Web – ISWC 2017. pp. 210–218. Springer International Publishing, Cham (2017)