Leveraging Taxonomic Information from Large Language Models for Hyponymy Prediction

Polina Chernomorchenko[1], Alexander Panchenko[2,3], and Irina Nikishina[4(✉)]

[1] HSE University, Moscow, Russia
pvchernomorchenko@edu.hse.ru
[2] Skolkovo Institute of Science and Technology, Moscow, Russia
[3] Artificial Intelligence Research Institute, Moscow, Russia
a.panchenko@skol.tech
[4] Universität Hamburg, Hamburg, Germany
irina.nikishina@uni-hamburg.de

Abstract. Pre-trained language models contain a vast amount of linguistic information as well as knowledge about the structure of the world. Both of these attributes are extremely beneficial for automatic enrichment of semantic graphs, such as knowledge bases and lexical-semantic databases. In this article, we employ generative language models to predict descendants of existing nodes in lexical data structures based on IS-A relations, such as WordNet. To accomplish this, we conduct experiments utilizing diverse formats of artificial text input containing information from lexical taxonomy for the English and Russian languages. Our findings demonstrate that the incorporation of data from the knowledge graph into a text input significantly affects the quality of hyponym prediction.

Keywords: taxonomy enrichment · IS-A relations · generative transformers · hyponym prediction

1 Introduction

Large pre-trained language models such as LLama-2 [24], Flan-T5 [4], InstructGPT [21] show impressive results in solving a wide range of tasks. However, the results of even such advanced models strongly depend on the input data [2,28]. In this paper, we assume that the prompting approach can be also extrapolated to lexical semantic tasks, such as IS-A relationship prediction. There are several studies exploring the ability of transformers to predict IS-A relationships through the use of natural language prompts [10,12]. However, they do not exploit sufficient information about taxonomy structure and the particular meaning of the lexeme which leads to the word sense disambiguation problem.

On the other hand, lexical taxonomies like WordNet [11] store a lot of important linguistic data. First of all, nodes (synsets) in such taxonomy graphs may accumulate several surface forms that represent the same meaning. Furthermore,

D. I. Ignatov et al. (Eds.): AIST 2023, LNCS 14486, pp. 49–63, 2024.
https://doi.org/10.1007/978-3-031-54534-4_4

they contain not only information about IS-A relations, but also words's synonyms, definitions and sense numbers specifying meaning of the particular word.

Taxonomy is a specific type of a knowledge graph that represents the relationships between the real-world entities and linguistic features. Taxonomies play a key role in a wide range of Natural Language Pricessing (NLP) tasks [9,15,27] and numerous studies are focused now on automatic enrichment of such structures [1,6]. Thus, this study aims to find out what information from the taxonomy can be useful for the accurate prediction of hyponyms using prompting. We focus on the hyponym prediction task, which aims at predicting new descendants for an existing node of the taxonomic graph. The formulation of the task is as follows: given taxonomy $G = (V, E)$; $G_{global} = (V_{global}, E_{global})$, $G \in G_{global}$. For given $v \in V$ find all $w \notin V : (v, w) \in E_{global}$ and $(v, w) \notin E$.

The main contributions of the paper are as follows: (i) we introduce new datasets for hyponym prediction for the Russian language; (ii) we explore the transformer-based generative architectures, specifically decoder and encoder-decoder models for hyponym prediction; (ii) we conduct experiments with various formats of artificial input data.

2 Related Work

Pre-trained language models demonstrate exceptional ability in encoding and understanding semantic information. For instance, Wiedemann et al. [26] demonstrate that homonyms can be differentiated using the k-NN search algorithm based on BERT embeddings [8]. Furthermore, BERT embeddings outperforms static embeddings in predicting lexical relationships [25].

In [10,12] authors exploring BERT's hypernymy knowledge. While BERT shows a decent level of acquisition of IS-A relations, there remain some limitations to prompting in natural language. Specifically, Ettinger notes in [10] that model predictions highly depend on the particular input, while in [12] is noted that universal prompts marking IS-A relations do not provide enough information for the model to distinguish homonyms.

These limitations can be addressed through the use of manually created prompts, although this process can be labor-intensive. Lexical taxonomies already provide structured information about lexemes and their relationships. Nevertheless, it is possible that incorporating information from graphs into language models can lead to more accurate hyponymy and hypernymy predictions.

Recent investigations have explored the incorporation of graph embeddings into language models [3,13,20]. In [3], for instance, vector representations of graphs were concatenated with text embeddings to provide model with knowledge of medical domain. In [20], the authors projected graph embeddings into the BERT space to enrich lexical taxonomies. Drawing inspiration from the significant achievements of language models in understanding and solving NLP tasks from bare text input, as exemplified by Brown et al. [2], we propose providing models with information about graph structure in textual format to improve their ability in comprehending IS-A relations.

3 Datasets

In this section, we present the datasets for both English and Russian languages. For each language, we perform our experiments on two different types of dataset: randomly and manually curated. We assume that the results in [20] for the English dataset collected automatically might be too low because the dataset comprises very uncommon and specific words from different domains (e.g. biological "protoctist family"), which significantly affects the results. Therefore, we also test on a smaller version of English dataset, where manually selected nodes are located at least 5 hops away from the root node and have 1–4 hop to descendants. We try to select similar words in both languages when creating similar datasets for Russian and present them within their features in Appendix A. We also collect two training datasets consisting of nonterminal nodes that are not included in any of the test datasets. Training datasets contain 15,000 and 10,000 synsets for English and Russian, respectively. The contents of the manually collected datasets as well as data on statistical parameters of the datasets can be found in Appendix A in Tables 6.

3.1 English Datasets

We utilize the CHSP dataset [20], consisting of 1000 preterminal nodes randomly selected from English WordNet [11], as an automatically generated dataset for English. One of the advantages of the CHSP dataset is that it closely resembles real data for the taxonomy enrichment task. However, the dataset contains highly specific and uncommon concepts, making it challenging to evaluate how well the models assimilate hyponymic relations. Based on the literature review on the acquisition of hyponymy with transformer-based models, we find out that past studies often use semantically simple datasets, as typified by the Battig dataset employed by Hanna and Mareček [12]. To provide an approximate understanding of the efficacy of proposed approach, we posit that a less intricate dataset is required. Therefore, to more accurately assess the models' hyponymy acquisition, we also test on a smaller dataset featuring 22 frequently used concepts from a common domain. While collecting the smaller dataset the formal criterion of a distance of at least five hops from the root while allowing nonterminal nodes is maintained. For the synset meanings, simple generic concepts that have at least 4 hyponyms including indirect ones are selected.

3.2 Russian Dataset

We generate same-sized random dataset for the Russian language based on formal criteria that match the CHSP dataset. When creating a manual dataset of common knowledge concepts, we try to find corresponding nodes in the Russian WordNet (RuWordNet)[1] for English ones. However, due to the different structures of the English and Russian taxonomies, some of the corresponding synsets

[1] https://ruwordnet.ru/en.

do not meet the formal criteria. For instance, the synsets with the meanings “room”, “furniture”, “monetary unit” and “board game” do not satisfy the condition for the distance from the root. Additionally, some synsets are replaced due to semantic inconsistency of concepts and specific hyponyms as a consequence. For example, the Russian synset with the meaning “color” has such hyponyms as *mimicry of organisms* and *animal’s color* on the same taxonomy level as usual color names like *red* or *green*.

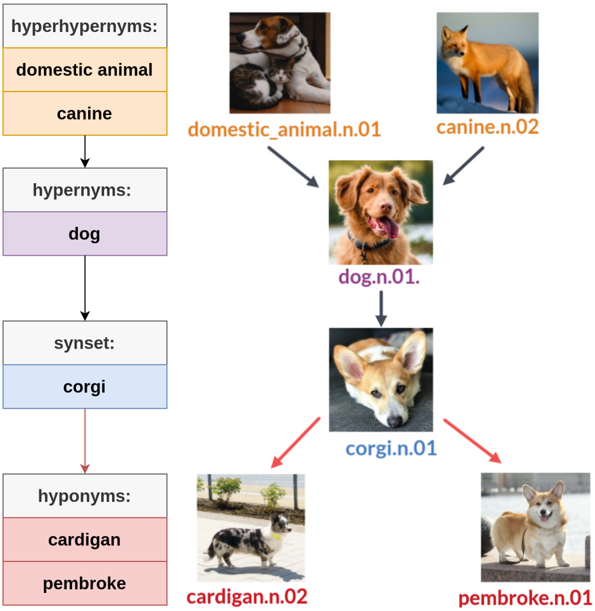

Fig. 1. Transformation of a graph data structure into a linear text representation.

4 Methodology

The current section presents a methodology for hyponymy prediction task. We compare artificial input formats to find out what information from the taxonomy the model needs for correct predictions and then use best input format to fine-tune models for both languages.

4.1 Artificial Prompt Selection for Fine-Tuning

There exist at least two studies exploring transformer-based models’ IS-A relations acquisition through prompting [10,12]. While this approach offers a way to get knowledge from transformers with minimal computational costs, it also has some drawbacks. One major issue is the need to disambiguate polysemous words when enriching taxonomies. Universal prompts in natural language, without additional context, do not allow for this kind of disambiguation. For example, it is possible to end sentence *Bat is ...* with both *an animal* and *a wooden club*. Another challenge is that certain text patterns may only work for a small subset of concepts. For example, the hyponymy-related text pattern “My favorite X is a Y” cannot be applied to negative concepts, as using the word “favorite” would be inappropriate (e.g. “My favorite retinopathy is diabetic retinopathy” would not make sense). In addition, many languages specify particular names for subtypes of different entities. Thus, a subclass of a dog is a *breed*, and a subclass of a plant is a *variety* and none of these words can be replaced by a *type* or *kind*. To overcome this obstacle, we narrow the attention to artificial text patterns providing model with information from taxonomy structure.

The basic format of the artificial text pattern for English includes information about hypernym and hyperhypernym of the target node and looks as following:

Example 1. hyperhypernyms: h_{1-n} | hypernyms: h_{1-m} | synset: s | hyponyms: l_{1-k}.

Here s denotes the target vertex, l_{1-k} is the list of correct hyponyms contained in the taxonomy, h_{1-m} and h_{1-n} are the parent and grandparent nodes relative to the target. Figure 1 shows the alignment between lexical taxonomy subgraph containing the target synset and proposed artificial input format.

In the English WordNet database [11], synsets contain not only the names of surface forms but also additional information that may be useful for predicting hyponyms. We take into account the following parameters: sense number, definitions and lemmas[2]. Based on these parameters, we create 8 artificial prefixes, the shortest of which contains only target synset and its hypernym, and the longest - all possible additional information. A description and examples of all types of artificial prefixes can be found in the Appendix A.

We fine-tune GPT-2 base [22] and T5-base [23] using constructed artificial patterns as input data. GPT-2 is fine-tuned for the language modeling task, while T5 is fine-tuned for seq2seq generation. Each experimental model is trained for three epochs with a batch size of 8 and evaluated with manually curated test data.

The structure of stored lexical information is the same in both the Russian and English WordNet, except for the absence of sense numbers in the former. Based on this fact we assume that outcomes of the experiments for English WordNet can be also extrapolated to the Russian data. Therefore, we restrict our comparison of artificial prefixes solely to the English WordNet.

4.2 Fine-Tuning Generative Transformers for Hyponymy Prediction Task

Based on the comparison of artificial prefixes, we select the best performing prompt for the decoder model. Then we conduct fine-tuning for three models

[2] In WordNet lemma represents a specific sense of a particular word. Each synset can contain multiple lemmas. (e.g., *color* and *colour* are two different lemmas, but have the same meaning).

in both languages: the decoder, the encoder-decoder and a third model which is instruction-based decoder of larger size. Thus, we use GPT2-large [22], T5-large [23] and Dolly-v1-6b [5] for English, and RuGPT2-large[3], RuT5-large[4] and Saiga-7b-LoRa[5] for Russian.

The justification for fine-tuning the third model is two-fold. Firstly, we assume that higher capacity of large language models to store information about the world and language will allow them to predict a wider variety of candidates from the texts it had seen during pre-training. Secondly, given the observed success of the instruction-scaled models [4], we expect that the model that has seen numerous instructions would finetune better and faster for yet another linguistic task of hyponym prediction.

For the first two models, we perform the full fine-tuning procedure and for the third - parameter efficient one using LoRa [14] and 8-bit version [7] of the model. For Saiga-7b fine-tuning we merge trained LoRa on Russian data adapter with the base LlaMA model [16] and than train new LoRa adapter via our data.

Models for English are trained during 3 epochs with batch size equal to 16. RuGPT2 and RuT5 are trained during two epochs with the same batch size, while Saiga is trained for one epoch with batch size equal to 4. All computations are performed on hardware of type NVIDIA RTX A6000-48GB.

5 Evaluation Setup

To compute evaluation metrics, we generate 50 sequences up to 15 tokens long (excluding prefix) for each input sample in the test dataset using top-k sampling ($k = 20$). We believe that this experimental formulation provides a more accurate evaluation of the models' hyponymy acquisition than using greedy search.

Next, we split the output by comma (since the models learn the expected output format correctly) and sort the final list of n-grams by frequency of occurrence.

5.1 Evaluation Metrics

Predicted n-grams are compared with the actual hyponyms in the taxonomy using a set of metrics. We use the Precision@k (P@k) metric to evaluate precision, which calculates the proportion of correct results achieved at a predetermined rank k. This helps to determine the number of accurate answers among the top-k results. Additionally, we use the Mean Reciprocal Rank (MRR) metric, which evaluates the multiplicative inverse of the rank of the first correct answer. To assess the sum of correct answers and their rank in the candidate list, we also use the Mean Average Precision (MAP). We also use Recall to consider the coverage.

[3] https://huggingface.co/ai-forever/rugpt3large_based_on_gpt2.

[4] https://huggingface.co/ai-forever/ruT5-large.

[5] https://huggingface.co/IlyaGusev/saiga_7b_lora.

Table 1. Results of fine-tuning with different formats of artificial prefix

Prefix format	MAP	MRR	P@1	P@2	P@5	P@10	R@a
GPT2-base							
Sense num	0.05	**0.64**	**0.50**	0.43	0.30	0.27	0.17
Default	0.05	0.63	0.46	**0.50**	**0.40**	**0.31**	0.17
Lemmas	0.05	0.58	0.41	0.36	0.38	0.30	0.20
Definition	0.04	0.59	0.41	0.39	0.29	0.26	0.20
Add. lemmas	0.04	0.55	0.36	0.39	0.28	0.28	0.21
Hyperdefinition	0.04	0.59	0.46	0.36	0.31	0.26	0.21
Hypernym only	0.04	0.55	0.36	0.39	0.30	0.31	0.15
All additions	0.03	0.44	0.32	0.25	0.18	0.17	0.21
T5-base							
Default	0.07	**0.67**	0.50	**0.57**	**0.46**	0.32	0.13
Lemmas	0.07	0.65	**0.55**	0.53	0.41	0.32	0.12
Sense num	0.07	0.61	0.50	0.46	0.38	0.33	0.14
Hyperdefinition	0.07	0.59	0.46	0.48	0.41	**0.36**	0.14
Hypernym only	0.07	0.50	0.36	0.43	0.38	0.31	0.14
Add. lemmas	0.06	0.63	0.55	0.43	0.39	0.32	0.13
Definition	0.06	0.60	0.50	0.46	0.37	0.31	0.13
All additions	0.05	0.50	0.32	0.39	0.37	0.30	0.14

6 Results

In this section we discuss the results of comparing artificial inputs as well as the results of fine tuning.

6.1 Artificial Prefixes Comparing

The results, presented in Table 1, demonstrate that both GPT2 and T5 models achieve the best results by using prefixes with either no or very few number of additions, such as “default”, “sense numbers”, and “lemmas”. This suggests that we can specify the meaning of the target synset by pointing to higher levels of the taxonomic structure. Surprisingly, the most informative input data format in both cases result in the lowest scores. “All additions” prefix variation contains sense numbers and also provides additional lemmas and definitions for each synset specified in the input. On the other hand, a minimalist input data format, which only indicates the parent vertex of the target synset, produced higher scores. We assume that such results may be related to the fact that long definitions combined with lemmas violate the formal data structure and make it harder for the model to understand the information. Additionally, prefix lengthening causes reduction of training examples due to input size constraints,

Table 2. Fine-tuning results for hyponymy prediction on automatically generated datasets for English and Russian

Method	MAP	MRR	P@1	P@2	P@5	P@10	R@a
English							
GPT-2	0.039	0.172	0.110	0.104	0.091	0.074	0.243
T5	0.048	0.189	0.123	0.115	0.096	0.076	0.127
Dolly-6b 8-bit LoRa	**0.111**	**0.324**	**0.226**	**0.202**	**0.164**	**0.122**	0.318
Russian							
RuGPT-large	**0.082**	**0.244**	0.142	**0.142**	**0.117**	**0.092**	**0.275**
RuT5-large	0.072	0.240	**0.162**	**0.142**	0.104	0.079	0.197
Saiga-7b	0.069	0.224	0.157	0.140	0.104	0.076	0.202

Table 3. Fine-tuning results for hyponymy prediction on manually collected datasets for English and Russian

Method	MAP	MRR	P@1	P@2	P@5	P@10	R@a
English							
GPT2-large	0.13	0.65	0.50	0.50	0.47	0.45	**0.34**
T5-large	0.15	0.85	0.77	0.66	0.64	0.54	0.24
Dolly-6b 8bit LoRa	**0.18**	**0.93**	**0.86**	**0.84**	**0.70**	**0.57**	0.29
Russian							
RuGPT-large	**0.177**	**0**.734	0.591	0.591	**0.600**	**0.523**	0.317
RuT5-large	0.156	0.664	0.500	0.523	0.500	0.486	0.256
Saiga-7b 8bit LoRa	0.129	**0.851**	**0.818**	**0.682**	0.555	0.450	**0.269**

while a minimalistic data format, on the contrary, allows more hyponyms in the input sequence (however, it's worth noting that this is only true for decoders, since in the seq2seq task formulation input and output are separated). Despite this fact, the prefix format under consideration exhibits the highest Recall scores for both models. This valuable characteristic makes the format particularly useful in the context of the final task of taxonomy enrichment. It can also be observed that GPT2 yields a greater extent of correct hyponym coverage in comparison to T5, as it generates a more diverse list of candidates. Higher MAP scores of T5 can therefore be attributed to this fact.

6.2 Hyponym Prediction for English and Russian

To fine-tune final models, we opt for the prefix containing sense numbers for English, as it demonstrated superior performance on the decoder model (since 2 out of 3 models for each language are decoders). On the other hand, for Russian, the default prefix is employed due to the absence of sense numbers in the Russian

WordNet that would indicate the ordinal number of given token's value in the current synset.

Below are examples of the prefixes selected for English and Russian for the synset *coat*:

Example 2. hyperhypernyms: garment.n.01 | hypernyms: overgarment.n.01 | synset: coat.n.01 | hyponyms:

Example 3. гипергиперонимы: одежда | гиперонимы: верхняя одежда | синсет: куртка | гипонимы:

We present the results for manual and automatic datasets in Tables 3 and 2, respectively. The results indicate that, in general, small datasets tend to yield significantly higher scores, suggesting that generative models are adept at assimilating hyponymic relations of frequently used words. However, automatic datasets, as previously mentioned, often comprise of rare and narrow concepts that lead to lower performance.

We can also observe the already mentioned trend that the Recall scores for GPT2 are significantly higher than those of T5 for both languages.

Regarding the architecture of the models, it is challenging to determine whether decoders or encoder-decoders perform better since the results are inconsistent between English and Russian. Among the smaller English models, the best scores in terms of MAP are achieved by T5, while for Russian, GPT2 fares better. Comparison of the results for Russian and English is not straightforward due to the usage of different prefixes and a varying number of training vertices - 15,000 for English and 10,000 for Russian.

Upon comparing the language-specific data, it becomes evident that the performance of smaller models is slightly better for Russian than to English according to the MAP scores. We connect this finding to the presence of a larger number of noun synsets for English, signifying a higher degree of fine-grained concepts in the taxonomy.

Regarding the large-scale instructional models, Dolly exhibits a significant lead over smaller models in relation to its performance on English-language data. This finding highlights the superior ability of larger models to assimilate hyponymic relations of both frequent and rare concepts. However, the same substantial increase in performance is not readily apparent for Saiga-7b. This outcome can be elucidated by the fact that the LlaMA model upon which Saiga-7b is based was primarily trained on English-language data. Moreover, the additional LoRa fine-tuning was confined to a relatively modest corpus of artificially generated Russian dialogue data, resulting in fewer instances of Russian lexical diversity being encountered during the training of the model.

7 Conclusion

In the presented study, we introduce a novel approach for constructing input data by incorporating information on the structure of the lexical taxonomy into large pre-trained language models. Our method shows that generative models provided with information about the graph in a well-perceived textual form can significantly improve the quality of hyponymy prediction. In this paper, we also provide a manually assembled dataset of general concepts for English, as well as the first datasets for evaluating the quality of predicting hyponyms in the context of enriching taxonomies for the Russian language.

Despite the fact that the prediction of hypernyms has a wider practical application than the prediction of hyponyms, the findings of the presented study are valuable for creating comprehensive common domain lexical taxonomies from scratch, benefiting low-resource languages.

In our research, we fine-tune relatively small-sized models, which demonstrate decent performance on both English and Russian data. However, the results from Dolly-v1-5b demonstrate that larger models can yield substantial improvements. As observed by Logan IV et al. [19] when comparing his findings to other studies [17,18], full fine-tuning is more effective for smaller models than prompt-tuning. Authors assumes that as model size increases, prompt-tuning yields better results. Thus, we can observe a promising path for future work to prompt-tune sizable generative models. We also anticipate that our approach can be extended both for other languages and other taxonomy enrichment tasks such as inserting the new node in the middle of the graph.

Acknowledgements. This work was supported by the DFG through the project "ACQuA: Answering Comparative Questions with Arguments" (grants BI 1544/7- 1 and HA 5851/2- 1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999).

A Appendix

Formats of Artificial Prefixes:

Default: part of speech (POS) tags and sense number are excluded from synset names.

Example 4. hyperhypernyms: undertaking | hypernyms: assignment | synset: school assignment | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Sense number: the full name of the synset is used, including POS tag and sense number.

Example 5. hyperhypernyms: undertaking.n.01 | hypernyms: assignment.n.05 | synset: school assignment.n.01 | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Lemmas: a list of lemmas is used instead of the name of the synset.

Example 6. hyperhypernyms: undertaking, project, task, labor | hypernyms: assignment | synset: school assignment, schoolwork | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Additional lemmas: the lemmas included in the synset are listed after its full name.

Example 7. hyperhypernyms: undertaking.n.01 (undertaking, project, task, labor) | hypernyms: assignment.n.05 (assignment) | synset: school assignment.n.01 (school assignment, schoolwork) | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Definitions: definitions for target synsets are given in parentheses.

Example 8. hyperhypernyms: undertaking.n.01 | hypernyms: assignment.n.05 | synset: school assignment.n.01 (a school task performed by a student to satisfy the teacher) | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Hyperdefinition: definitions for hypernyms and hyperhypernyms are given in parentheses.

Example 9. hyperhypernyms: undertaking.n.01 (any piece of work that is undertaken or attempted) | hypernyms: assignment.n.05 (an undertaking that you have been assigned to do (as by an instructor)) | synset: school assignment.n.01 | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

Hypernym only: only the hypernym is given.

Example 10. hypernyms: assignment.n.05 | synset: school assignment.n.01 | hyponyms: classroom project, classwork, homework, prep, preparation, lesson

All additions: all possible information is used: sense number, definitions and lemmas.

Example 11. hypernyms: assignment.n.05 (assignment) (an undertaking that you have been assigned to do (as by an instructor)) | synset: school assignment.n.01 (school assignment, schoolwork) (a school task performed by a student to satisfy the teacher) | hyponyms: classroom project, classwork, homework, prep, preparation, lesson (Tables 4 and 5).

Table 4. Content and statistic of manually curated dataset for Russian. Here d_l denotes to leaf distance, d_r to root distance and n_h to the total number of hyponyms including indirect ones.

Id	Title	d_l	d_r	n_h
6892-N	ПАЛЬТО	2	7	4
108048-N	КУРТКА	1	7	7
108194-N	ШТАНЫ, БРЮКИ	2	6	10
109093-N	МАКАРОННЫЕ ИЗДЕЛИЯ	3	6	6
3921-N	СЫР	4	8	16
1225-N	МЯСО	1	5	29
8367-N	ВИНО	1	7	25
5239-N	КОНФЕТА	2	6	8
107283-N	ПИРОГ	1	6	9
549-N	НАПИТОК	2	5	82
107842-N	ЯГОДА	1	9	28
109620-N	ДЕТСКАЯ ИГРУШКА	1	6	20
7992-N	УДАРНЫЙ МУЗЫКАЛЬНЫЙ ИНСТРУМЕНТ	1	6	10
107996-N	СТРУННЫЙ МУЗЫКАЛЬНЫЙ ИНСТРУМЕНТ	2	6	20
1045-N	ВРАЧ	1	5	85
354-N	ФРУКТ	1	9	37
348-N	ОВОЩ	1	8	27
107795-N	ХИЩНОЕ МЛЕКОПИТАЮЩЕЕ	1	7	56
4454-N	СОБАКА	2	7	53
109170-N	КОСМЕТИЧЕСКОЕ СРЕДСТВО	1	6	21
4318-N	КРУПА	1	6	12
965-N	ЦВЕТКОВОЕ РАСТЕНИЕ	2	5	51

Table 5. Content and statistic of manually curated dataset for English. Here d_l denotes to leaf distance, d_r to root distance and n_h to the total number of hyponyms including indirect ones.

Id	Title	d_l	d_r	n_h
3057021	coat.n.01	2	9	53
3045337	cloak.n.02	2	9	29
4489008	trouser.n.01	2	8	26
7698915	pasta.n.02	1	5	26
7850329	cheese.n.01	3	5	37
7649854	meat.n.01	1	5	197
7891726	wine.n.01	3	7	68
7597365	candy.n.01	1	8	62
7625493	pie.n.01	1	7	25
7881800	beverage.n.01	3	5	339
7742704	berry.n.01	4	7	21
3219135	doll.n.01	1	6	8
3249569	drum.n.01	1	9	8
3467517	guitar.n.01	1	9	6
502415	board_game.n.01	1	8	18
4105893	room.n.01	1	7	195
3405725	furniture.n.01	1	7	196
13388245	coin.n.01	1	8	41
3597469	jewelry.n.01	1	7	39
3714235	makeup.n.01	1	8	11
4959672	chromatic_color.n.01	1	6	91
1699831	dinosaur.n.01	1	12	50

Table 6. Number of synsets and hyponyms in test datasets.

Dataset	Synsets	Hyponyms
CHSP	1000	13617
RuCHSP	1000	5673
EnManual	22	1546
RuManual	22	616

References

1. Aly, R., Acharya, S., Ossa, A., Köhn, A., Biemann, C., Panchenko, A.: Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4811–4817. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1474. https://aclanthology.org/P19-1474
2. Brown, T.B., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020)
3. Chang, D., Lin, E., Brandt, C., Taylor, R.: Incorporating domain knowledge into language models using graph convolutional networks for clinical semantic textual similarity (preprint). JMIR Med. Inform. (2020). https://doi.org/10.2196/23101
4. Chung, H.W., et al.: Scaling instruction-finetuned language models (2022)
5. Conover, M., et al.: Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM (2023). https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm
6. Dale, D.: A simple solution for the taxonomy enrichment task: discovering hypernyms using nearest neighbor search (2020)
7. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Llm.int8(): 8-bit matrix multiplication for transformers at scale (2022)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Minneapolis, Minnesota (Volume 1: Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
9. Dietz, L., Kotov, A., Meij, E.: Utilizing knowledge graphs in text-centric information retrieval, pp. 815–816 (2017). https://doi.org/10.1145/3018661.3022756
10. Ettinger, A.: What BERT is not: lessons from a new suite of psycholinguistic diagnostics for language models. Trans. Assoc. Comput. Linguist. **8**, 34–48 (2020). https://doi.org/10.1162/tacl_a_00298. https://aclanthology.org/2020.tacl-1.3
11. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication. MIT Press, Cambridge (1998)
12. Hanna, M., Mareček, D.: Analyzing BERT's knowledge of hypernymy via prompting. In: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Punta Cana, Dominican Republic, pp. 275–282. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.blackboxnlp-1.20. https://aclanthology.org/2021.blackboxnlp-1.20
13. He, B., et al.: BERT-MK: integrating graph contextualized knowledge into pre-trained language models. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2281–2290. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.207. https://aclanthology.org/2020.findings-emnlp.207
14. Hu, E.J., et al.: Lora: low-rank adaptation of large language models (2021)
15. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: Proceedings of the Twelfth ACM International Conference on Web

Search and Data Mining, WSDM 2019, pp. 105–113. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3289600.3290956

16. Izacard, G., Grave, E., Pilehvar, M.T., Alzantot, M., Baroni, M.: LLaMA: open and efficient foundation language models (2023)
17. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.243. https://aclanthology.org/2021.emnlp-main.243
18. Li, X.L., Liang, P.: Prefix-tuning: optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4582–4597. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.353. https://aclanthology.org/2021.acl-long.353
19. Logan, R.L., IV., Balažević, I., Wallace, E., Petroni, F., Singh, S., Riedel, S.: Cutting down on prompts and parameters: simple few-shot learning with language models. CoRR abs/2106.13353 (2021). https://arxiv.org/abs/2106.13353
20. Nikishina, I., Vakhitova, A., Tutubalina, E., Panchenko, A.: Cross-modal contextualized hidden state projection method for expanding of taxonomic graphs. In: Proceedings of TextGraphs-16: Graph-Based Methods for Natural Language Processing, Gyeongju, Republic of Korea, pp. 11–24. Association for Computational Linguistics (2022). https://aclanthology.org/2022.textgraphs-1.2
21. Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2018). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
23. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html
24. Touvron, H., et al.: LLaMA 2: open foundation and fine-tuned chat models (2023)
25. Vulić, I., Ponti, E.M., Litschko, R., Glavaš, G., Korhonen, A.: Probing pretrained language models for lexical semantics. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7222–7240. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.586. https://aclanthology.org/2020.emnlp-main.586
26. Wiedemann, G., Remus, S., Chawla, A., Biemann, C.: Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings (2019)
27. Zhou, R., et al.: WCL-BBCD: a contrastive learning and knowledge graph approach to named entity recognition (2022)
28. Zhou, Y., et al.: Large language models are human-level prompt engineers. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=92gvk82DE-