Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

RUSSE’2020: FINDINGS OF THE FIRST TAXONOMY ENRICHMENT TASK FOR THE RUSSIAN LANGUAGE

Nikishina I. (Irina.Nikishina@skoltech.ru),
Logacheva V. (V.Logacheva@skoltech.ru),
Panchenko A. (A.Panchenko@skoltech.ru)

Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia

Loukachevitch N. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

This paper describes the results of the first shared task on taxonomy enrichment for the Russian language. The participants were asked to extend an existing taxonomy with previously unseen words: for each new word their systems should provide a ranked list of possible (candidate) hypernyms. In comparison to the previous tasks for other languages, our competition has a more realistic task setting: new words were provided without definitions. Instead, we provided a textual corpus where these new terms occurred. For this evaluation campaign, we developed a new evaluation dataset based on unpublished RuWordNet data. The shared task features two tracks: “nouns” and “verbs”. 16 teams participated in the task demonstrating high results with more than a half of them outperforming the provided baseline.

Keywords: shared task, taxonomy, language resources, taxonomy enrichment, hypernymy, RuWordNet

DOI: 10.28995/2075-7182-2020-19-579-595

RUSSE'2020: ТЕСТИРОВАНИЕ МЕТОДОВ ПОПОЛНЕНИЯ ТАКСОНОМИИ ДЛЯ РУССКОГО ЯЗЫКА

Никишина И. (Irina.Nikishina@skoltech.ru),
Логачева В. (V.logacheva@skoltech.ru),
Панченко А. (A.Panchenko@skoltech.ru)
Сколковский институт науки и технологий (Сколтех), Москва, Россия

Лукашевич Н. (louk_nat@mail.ru)
Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В данной работе описываются результаты первой дорожки RUSSE'2020 по пополнению таксономии терминов русского языка. Задача участников состоит в расширении существующей таксономии (RuWordNet): для новых слов необходимо предсказать их возможные гиперонимы. В отличие от соревнований, проводившихся для других языков, RUSSE'2020 имеет более реалистичную постановку: участникам не предоставляются определения для новых слов, при этом они могут использовать корпус текстов, в которых встречаются новые термины. Для оценки качества методов был подготовлен «золотой стандарт»: новые слова и их гиперонимы из неопубликованной версии RuWordNet, размеченные вручную. RUSSE'2020 включает в себя два трека по частям речи: «существительные» и «глаголы». Всего в соревновании участвовали 16 групп, более чем половине из них удалось превзойти предоставленное организаторами базовое решение.

Ключевые слова: соревнование, таксономия, обогащение таксономии, лингвистические ресурсы, гиперонимия, RuWordNet

1. Introduction

Taxonomies are tree structures that organize terms into a semantic hierarchy. Taxonomic relations (or hypernyms) are "is-a" relations: cat *is-an* animal, banana *is-a* fruit, Microsoft *is-a* company, etc. This type of relations is useful in a wide range of Natural Language Processing (NLP) tasks for performing semantic analysis.

While substantial interest is drawn to the extraction of hypernyms and taxonomic structures from text [6], [7], [9], the fully automatic taxonomy induction methods are still not widely used for routine construction of lexical resources, such as taxonomies. Nevertheless, the automatic hypernym candidate generation can facilitate and accelerate the manual taxonomy extension. Therefore, it is extremely useful to develop support tools for creation, enrichment, and maintenance of the existing semantic resources as well as their tuning to specific tasks and/or text collections.

Multiple evaluation campaigns tackling taxonomy problems have been organized for English and other Western European languages. Among them are SemEval-2018 task 9 on hypernym extraction [9], SemEval-2016 task 13 [7] and SemEval-2015 task 17 [6] on taxonomy induction, and SemEval-2016 task 14 [13] on taxonomy enrichment.

The main contribution of this paper is to report about RUSSE'2020—the first shared task on Taxonomy Enrichment for Russian, as well as for any other Slavic language. The goal of this semantic task is to extend an existing taxonomy with previously unseen words. For each new word—an *orphan*—the participants should provide a ranked list of possible hypernyms. RUSSE'2020 is similar to the SemEval-2016 task 14 [13], but has a more realistic setting. The participants are not given the definitions of the words to be added, but only a list of these words. However, the participants are allowed to use any additional resources.

We create a gold standard dataset for evaluating the participating systems. We consider the unreleased data from RuWordNet [19] as our gold standard and split it into two subsets: "nouns" and "verbs". Moreover, we develop and release a baseline taxonomy enrichment model that uses an unsupervised approach based on word embeddings.

This paper is organized as follows. **Section 2** reviews the previous shared tasks on taxonomy creation, extension, and maintenance as well as hypernym extraction. **Section 3** introduces the task, the data, and the baseline model. The participating systems are described in **Section 4**, the overall results are provided in **Section 5**.

2. Related Work

Various methods were proposed for hypernym extraction, including pattern-based methods [12], [23], unsupervised and supervised methods based on word embeddings [27], [32], and hybrid approaches integrating several types of features [4], [29], [30].

In the majority of settings, hypernym extraction is cast as a binary classification task. Thus, the hypernym extraction algorithms are usually evaluated on purpose-built datasets containing positive and negative examples. One of such datasets is BLESS created by Baroni and Lenci [3] to test distributional models that predict several types of relationships between words.

In the semantic taxonomy enrichment task at SemEval 2016 [13], the organizers studied the possibilities of automatic addition of concepts from online glossaries and lexicographic resources into existing taxonomies such as WordNet [22]. Each new word was provided with a definition (gloss) from Wiktionary. The baseline model attached a new term to the first word from its gloss with the matching part of speech. Despite its simplicity, this approach turned out to be difficult to beat. It was outperformed by only one participating system. All participants used only Wiktionary glosses and did not try to employ any additional features from Wiktionary or text collections.

Bordea et al. [6], [7] evaluated taxonomy construction models based on the extracted hypernym relations. The evaluation was performed for several domains. Gold standard datasets were collected from WordNet and EUROVOC thesaurus[1]. The authors suggested several metrics tailored for taxonomy evaluation.

[1] Eurovoc: http://eurovoc.europa.eu/drupal

Levy et al. [17] suggested that the results achieved in classification settings of hypernym extraction are mainly explained by the so-called "lexical memorization phenomenon"—a situation when models learn that in a relation "*x* is-a *y*" a word *y* is a prototypical hypernym. For example, if a classifier obtains many positive examples with the word *y*=animal, it may learn that anything that appears with *y*=animal should generate a positive answer. Camacho-Collados [8] argues that hypernym classification is not a realistic scenario. Instead, hypernym-oriented evaluation should be organized as a hypernym discovery task, i.e. given a word *dog*, the system should be able to discover its hypernyms *mammal* or *animal* among a large number of other possible candidates. He suggests evaluating models' performance in this task with information-retrieval evaluation measures such as mean reciprocal rank (MRR) or mean average precision (MAP).

In the hypernym discovery task at SemEval 2018 [9], the organizers attempted to improve the quality of evaluation and formulated the hypernym extraction task as a ranking task. They created a list of hypernym candidates—these were all unigrams, bigrams, and trigrams that occurred more than *N* (for example, 5 times in the corpus). For each of the new words and phrases, the participants were asked to rank the hypernym candidates by their relevance. Moreover, the participants had to find as many hypernyms as possible. The gold standard list of answers contained hypernyms of all hierarchy levels excluding only the most abstract concepts such as "entity".

Panchenko et al. [24] describe the shared task on semantic similarity for Russian. One of the subtasks was to predict the similarity between words (synonym or hypernym relations). Each target word had the same number of related and unrelated source words. Reference answers were taken from the RuThes thesaurus [18].

Compared to the above mentioned competitions, RUSSE'2020 is closely related to the SemEval-2016 Taxonomy Enrichment Task [13] and SemEval-2018 Hypernym Discovery Task [9]. As in the mentioned SemEval tasks, in our competition the participants are asked to attach new words to the existing synsets, to create a ranked list of hypernym candidates, and the performance is evaluated using MAP and MRR metrics.

3. Shared Task Description

The goal of the task can be formulated as follows: given words that are not yet included in the taxonomy, we need to associate each word with the appropriate hypernym synset(s) from the existing taxonomy RuWordNet. For example, given an input word "утка" (duck) the participants are asked to provide a ranked list of its most probable 10 candidate hypernym synsets, e.g. "животное" (animal), "птица" (bird), and so on. We assume that an *orphan* may be a "child" of one, two, or more "ancestors" (hypernym synsets) at the same time.

The task featured two tracks: detection of hypernym synsets for nouns and verbs. We provided to participants the following resources: (i) training set based on the RuWordNet taxonomy, (ii) a collection of news texts from the year 2017 (2.2 billion tokens), (iii) a parsed Wikipedia corpus[2], and (iv) a hypernym database from the Rus-

[2] https://doi.org/10.5281/zenodo.3827903

sian Distributional Thesaurus[3] [26], [28], which contains a set of hypernyms and a set of distributionally related terms both extracted from a huge text corpus. The participants were allowed to use any additional data and were asked to indicate the additional resources in their model descriptions.

The competition was hosted on the Codalab platform[4]. To allow the participants to evaluate their models on real data, we split the gold standard data into public and private test sets (denoted as "PRACTICE" and "EVALUATION" phases in Codalab). Thus, the participants could test their models before the deadline on the public test set by submitting the results to the "PRACTICE" leaderboard. During the "EVALUATION" phase the leaderboard was hidden, so the participants were not able to overfit the test data.

Table 1: Number of RuWordNet synsets in datasets used in the shared task

	Nouns	Verbs
Total in RuWordNet	29,297	7,636
Train set	12,393	2,109
Private test set	1,525	350
Public test set	763	175

3.1. Datasets and Additional Resources

We provided the gold standard dataset which contains words with manually defined hypernyms. These words were included in the extended version of RuWordNet which has not been published yet. We split this data into two parts: public (763 nouns and 175 verbs) and private (1,525 nouns and 350 verbs).

The words included in the gold standard test dataset (*orphans* in Table 1) were collected in the following way. First, we extracted words (nouns and verbs) which are present in the extended RuWordNet, but absent in the published RuWordNet. We selected only single words (not phrases) with at least 50 occurrences in the corpus of news texts from 2017. Then we filtered the obtained list excluding the following words:

- all three-symbol words and the majority of four-symbol words;
- diminutive word forms and feminine gender-specific job titles;
- words which are derived from words which are included in the published RuWordNet;
- words denoting inhabitants of cities and countries;
- geographic and personal names;
- compound words that contain their hypernym as a substring.

The gold hypernyms of the *orphan* words were assigned manually by linguists. However, it should be noted that these gold hypernyms are not necessarily the closest hypernyms. The extended RuWordNet can contain whole chains of hypernyms none

[3] https://doi.org/10.5281/zenodo.3827834

[4] https://competitions.codalab.org/competitions/22168

of which is included in the published version. If one of the synsets selected for the test set belongs to this chain and its immediate hypernym is not presented in the published version, we set its closest published "ancestor" as a gold hypernym.

The training dataset (words paired with hypernyms) was generated from the current version of the RuWordNet taxonomy and annotated analogously to the test data. To create the training set we sampled all leaves (synsets with no hyponyms) of depth equal or more than 5. Overall, it comprises 12,393 nouns and 2,102 verbs.

The news text collection, which was provided to the participants, consists of 8 million news articles written in 2017 collected from more than 1,000 news sources. It contains a total of 2.2 billion tokens. This corpus was initially collected so that it contains at least 50 occurrences of the majority of words from the test data. However, it was further discovered that 17 words in the public test (1.8%) and 42 words in the private set (2.2%) have fewer occurrences in the corpus, due to the use of different lemmatization tools and morphological ambiguity.

Table 2: Various ground truth representations for the term "cruise". The connectivity component representation allows us to take into account the fact that all three direct hypernyms are related to the same word sense, as depicted in Figure 1(d), and do not wrongly penalise a system that predicted only one of them

Set of direct hypernyms	{**entertaining journey**, **journey**, **tour**}
Sets of direct hypernyms and their parents	{**entertaining journey**, travel, entertainment, active leisure}, {**journey**, travel, move}, {**tour**, travel, journey, active leisure}
Connectivity component	{**entertaining journey**, **journey**, **tour**, travel, entertainment, active leisure, move}

3.2. Evaluation Metrics

The participants were asked to generate a ranked list of 10 most probable hypernym candidates for each word in the test set. The results were evaluated using the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) scores. MAP score evaluates the whole range of produced hypernym candidates, whereas MRR looks at how close the first correct prediction is to the top of the list. We consider MAP as the official metric of our competition.

Both metrics are widely used in the Hypernym Discovery shared tasks, where systems also need to output ranked lists of candidate hypernyms [9]. In contrast to [9], we limited the number of possible answers to $k = 10$, because the correct answers from lower positions will have small weights and will not contribute much to the final score.

To be less restrictive during the evaluation, we consider as correct answers not only the immediate hypernyms of new words but also the hypernyms of these hypernyms. Therefore, if a system predicts a hypernym of a correct hypernym, this will also be considered a match.

One hypernym may be a "parent" of another hypernym (synset "plane" has two parents—"aircraft" and "aviation technology", whereas "aviation technology" itself is the hypernym for "aircraft"). While computing the MAP score, it may not be clear

which hypernym gains the score: "aviation technology" synset as the immediate hypernym or "aviation technology" as the second-order hypernym. Hypernyms may also have common parents: "string instrument" and "folk instrument" both have a hypernym "musical instrument". In this case, if "musical instrument" appears in the candidate list, the MAP score will also be confused.

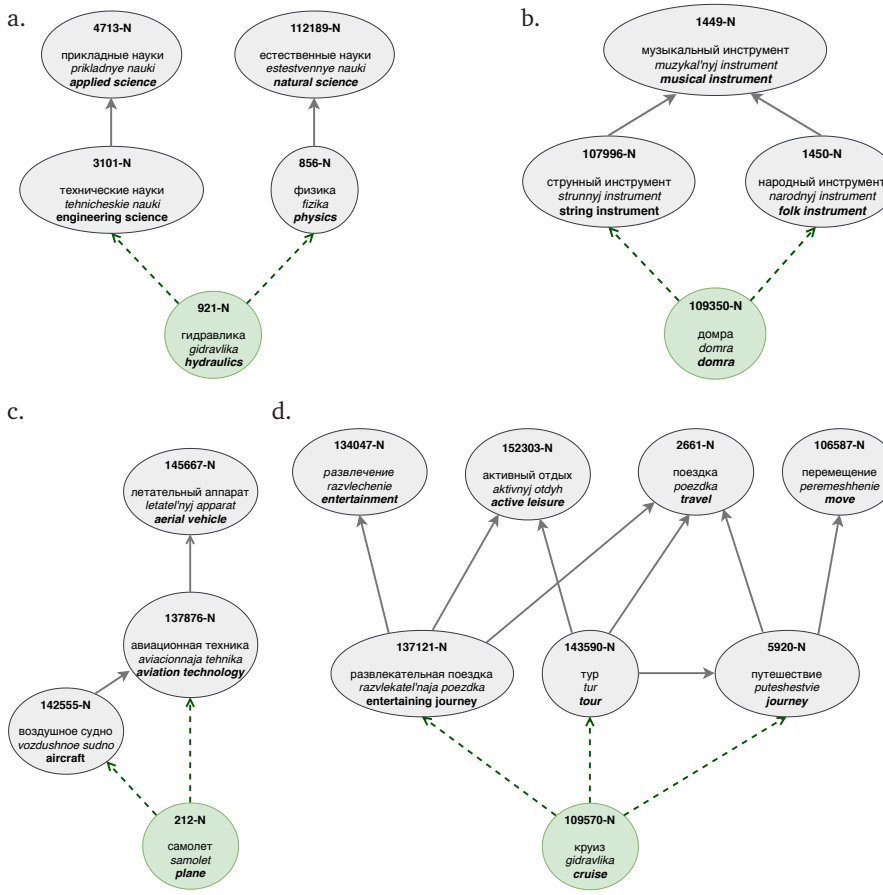

Figure 1: Examples of hypernym subgraphs from RuWordNet ground truth: direct and second-order hypernyms may be related in various ways motivating the evaluation metric based on connectivity components. While in (a) two parents lead to different senses, in (b, c, d) two parents lead to the same sense. Dashed lines indicate ground truth hypernyms.

To avoid this hypernym ambiguity, we split all hypernyms of a word (both immediate and second-order) into groups. Each group corresponds to the connectivity component in the subgraph reconstructed from all hypernyms. The process is shown in Figure 1. We see that the first and the second subgraphs consist of only one connectivity component, whereas in the third graph the immediate hypernyms form different hypernym groups. Therefore, the list of possible candidates of a given word should contain at least one hypernym from each hypernym group. Thus, connectivity

components allow us to distinguish between cases depicted in Figure 1(a) where a system must predict hypernyms for both word senses from two independent branches and (b)/(c)/(d) where only one word sense is to be predicted.

Overall, to compute the score, we extend the standard MAP reference and group hypernyms into connectivity components (see evaluation examples in Table 2 for the word "cruise"). **The answer is given a full score if there is at least one hypernym from each connectivity component in the list of possible candidates**. To get the highest score for the example from Table 2, it is enough to predict one of the synsets. Moreover, all hypernyms of all connectivity components are considered equally relevant: predictions starting with "applied science" and "physics" or with "natural science", and "engineering science" will get the same score.

3.3. Baseline

We implemented a simple baseline that makes use of non-contextualized (standard) word embeddings. We chose fastText embeddings[5] [5] to solve this task for two reasons: pre-trained fastText models are easy to deploy and they do not require any additional data or training for the out-of-vocabulary words, because they incorporate subword tokens.

Our baseline comprises the following steps:

1. Compute embeddings of all synsets in RuWordNet by averaging embeddings of all words from senses belonging to a synset.
2. Get embeddings for *orphans*. For multi-word *orphans* the embeddings are computed by averaging vectors for all words comprising an *orphan*.
3. For each *orphan* compute the top $k = 10$ closest synsets of the same part of speech as the orphan using the cosine similarity measure.
4. Extract hypernyms for each of these closest synsets from the previous step. Take the first $n = 10$ results (as each synset may have several hypernyms).

Our method is unsupervised and does not require any additional data. Nevertheless, it turned out to be a strong baseline as shown below.

4. Participating Systems

RUSSE'2020 shared task attracted 16 participants in the "nouns" track and 14 in the "verbs" track (excluding the baseline). We provide descriptions of the top 7 solutions which outperformed the baseline at any track. We denote each team either with its team name (if any) or with their CodaLab user names. In cases of multiple submissions from one team, we report only the best result. The scores of the teams are shown in Tables 3 and 4.

[5] https://fasttext.cc/docs/en/crawl-vectors.html

4.1. Yuriy

This participant-generated candidate hypernyms and calculated features for them. Then candidates were ranked by a linear model with handcrafted weights. The list of features is provided below:

1. candidate is in top 10 similar words from RuWordNet;
2. candidate is in hypernyms of top 10 similar words from RuWordNet;
3. candidate is in hypernyms of hypernyms of top 10 similar words from RuWordNet;
4. candidate is in hypernyms on Wiktionary[6] page about the word;
5. candidate is in hypernyms of hypernyms on Wiktionary page about the word;
6. candidate is in "en-ru" translation of WordNet[21] hypernyms of "ru-en" translation of the word (extracted with Yandex Machine Translation model[7]);
7. candidate is in the word definition in the Wiktionary page;
8. candidate is in the Yandex search result page;
9. candidate is in the Google search result page.

The candidates were collected using features 1–6. Features 1–3 are based on the fastText model[8]. This approach was applied for both "nouns" and "verbs" tracks.

4.2. xeno

This participant merged candidates extracted by several methods. Those methods included: Russian Wiktionary semantic graph (taxonomic relations, synonymy, antonymy); rule-based plain text definition parsing; rule-based plain text parsing with Hearst patterns on Russian Wikipedia from [26] and Russian language corpus; graph-based analysis of the nearest neighbor list obtained from word2vec. The definitions were taken from Russian Wiktionary, Russian Wikipedia, Big English-Russian polytechnic dictionary, Efremova dictionary [11]. The above-mentioned methods were used for nouns. For verbs, the team used only the Russian Wiktionary semantic graph and rule-based plain text definition parsing.

4.3. KuKuPl

[14] This team trained a classifier on the official train data provided by the organizers. They considered synsets (occurring more than *n* times in the training data) as classes, representing words with the embeddings (standard CBOW from word2vec) pretrained a concatenation of four corpora: Araneum Russicum Maximum, Russian Wikipedia, Russian National Corpus, and a corpus of Russian news (9.5 billion word tokens overall). The corpus was specially tailored for this task: all multi-word entities which also occurred in the RuWordNet were merged into single tokens, thus making sure that the majority of the RuWordNet entries received their respective vector representations.

---

6 https://ru.wiktionary.org

7 https://translate.yandex.ru

8 https://fasttext.cc/docs/en/crawl-vectors.html

A neural network classifier with one hidden layer (size 386), dropout of 0.1, ReLu activation, and a softmax output layer was trained on all the training data until convergence, using hypernym synset ids as class labels. At test time, the trained model obtains the vector representation of a query word and predicts possible classes (hypernym synsets) for this vector. 10 synsets with the highest probability are considered predictions. This approach is applied for both "nouns" and "verbs" tracks.

4.4. RefalMachine, Parkat13

[31] This team implemented the algorithm consisting of three stages. Firstly, they created a list of similar words using a combination of vector representations of words obtained with PPMI (positive pointwise mutual information) weighting and SVD factorization (window = 1). Secondly, they selected candidates from those similar words (depending on pattern matching), their hypernyms, and second-order hypernyms. These candidates were ranked based on the following features:

- cosine similarity;
- patterns matching co-hyponyms;
- patterns matching hypernyms (Hearst patterns). The patterns were extracted from the news corpus provided by the organizers;
- the number of synset occurrences in the candidate list;
- probabilities based on ruBERT predictions [15].

The final rank for each candidate was computed using the weighted feature combination; the weights are hand-picked during the experiments. This approach was applied for both "nouns" and "verbs" tracks.

4.5. MorphoBabushka (alvadia, maxfed, joystick)

[1] This team used the following pipeline. First, they retrieved nearest neighbors for the target word from word2vec "SkipGram with Negative Sampling" model trained on Librusec book collection [2] and search for their direct and indirect hypernyms in RuWordNet. Then they counted direct and indirect hypernyms of the nearest neighbors, combining their counts, converting (or excluding if not possible) inappropriate ones with wrong part-of-speech. They took 10 most frequent hypernyms of nearest neighbors' synsets. Finally, they combined those hypernyms with the hypernyms extracted from Wiktionary by matching definition N-grams with the synsets. This method was applied for both "nouns" and "verbs" tracks.

4.6. cointegrated

[10] The participant used similarity scores between word embeddings to predict hypernym relations. For each RuWordNet synset, the team computed the embedding of its title, all senses, and the mean embedding of the title and all senses. Each type of the above-mentioned embeddings was computed as an L2-normalized weighted mean of its word embeddings from RusVectores[16] (weight is of 1.0 nouns, 0.1 for prepositions, and 0.5 for all other POS). For OOV words, the embedding was

computed as a mean embedding of all words in the vocabulary with the longest prefix matching the target word.

For each query word (orphan), the participant found its 100 nearest neighbors from RuWordNet and all the first and second-order hypernyms of the corresponding synsets, considering them as answer candidates. The resulting list of hypernyms comprises 10 candidates with the highest scores. The score for each candidate is a sum of "neighbor scores" overall nearest neighbors from RuWordNet; if the candidate is a second-order hypernym, its "neighbor score" is multiplied by 0.5. The "neighbor score" is calculated as $\exp(-3 \cdot d) \cdot s^5$, where d is the distance between the queries and neighbor embeddings; s is their cosine similarity. The described approach was applied for both "nouns" and "verbs" tracks.

5. Results

Tables 3 and 4 present respectively the results "nouns" and "verbs" tracks. As one can observe, the absolute difference in scores of the two tracks is quite large. Apparently, the "verbs" track is more difficult, because word embeddings for verbs are not as accurate and exhaustive as for nouns: verbs are more abstract and can be seen in a context with a wider range of words than nouns [25].

All the methods applied by the participants can be divided into two classes. The first class applies supervised learning (binary or multi-class classification). The second one performs ranking based on a range of features (similarity measures, hypernyms of different orders, etc.). Surprisingly, the majority of approaches are not stable across the tasks: they can demonstrate promising results on the "nouns" track, but lag behind on "verbs" (e.g. *KuKuPl*, *RefalMachine*) or vice versa (e.g. *cointegrated*).

Another interesting point is the type of embeddings that was used by the top-7 participants. Apart from *RefalMachine*'s, no methods used contextualized embeddings. The most popular vector model is word2vec [20], pre-trained (*Yuriy*, *cointegrated*) or trained on the provided datasets (*KuKuPl*, *MorphoBabushka*).

Interestingly, all the top-7 participants resort to additional data. The most popular additional source are text corpora: *KuKuPl*, *MorphoBabushka* use corpora to train custom word embeddings, *cointegrated* and *Yuriy* apply pre-trained embeddings. The 2017 news corpus with contexts for word occurrences is used by three teams (out of the top 7 teams described in this paper): *KuKuPl*, *Parkat13* and *RefalMachine*. Another promising source of information are dictionaries: *MorphoBabushka* and *Yuriy* give their preference to Wiktionary, whereas *xeno* uses Big English-Russian polytechnic dictionary, Efremova dictionary. The most outstanding range of additional resources (from *Yuriy*) includes Wiktionary, Yandex Translate, Google, and Yandex search pages results. However, we cannot draw any conclusions about the efficiency of the use of additional data, as these sources are not the only factors that influenced the final results.

Table 3: Evaluation results for "nouns" track on hte private test dataset

Rank	User	Entries	MAP	MRR
1	Yuriy	5	**0.5522**	**0.5940**
2	xeno	5	0.5054	0.5433
3	KuKuPl	2	0.4976	0.5332
4	RefalMachine	6	0.4930	0.5314
5	MorphoBabushka	5	0.4497	0.4835
6	baseline	1	0.4210	0.4518
7	cointegrated	5	0.4178	0.4503
8	adhaesitadimo	1	0.3759	0.4043
9	vvyadrincev	2	0.3095	0.3342
10	vimary	4	0.2951	0.3187

Table 4: Evaluation results for "verbs" track on the private test dataset

Rank	User	Entries	MAP	MRR
1	cointegrated	3	**0.4483**	**0.5049**
2	Yuriy	2	0.4355	0.5135
3	MorphoBabushka	5	0.3890	0.4419
4	baseline	1	0.3335	0.3817
5	xeno	2	0.3075	0.3547
6	RefalMachine	5	0.2542	0.2969
7	KuKuPl	3	0.2470	0.2897
8	vimary	2	0.1783	0.2115
9	vvyadrincev	3	0.1474	0.1786
10	Arshehremen	2	0.0000	0.0000

In order to analyse the results obtained by the participants, we provide several examples for both verbs and nouns (Tables 5 and 6)[9]. We took 3 nouns from *Yuriy*'s answer and 3 verbs from *cointegrated*'s to compare with the gold standard hypernym synset subgraphs ("ground truth" part of Tables Tables 5 and 6). For the nouns "сахарин" (saccharin), "селфи" (selfie) and the verb "тусить" (to party) candidate lists contain either all hypernyms or at least one hypernym from all subgraphs. These examples also demonstrate that the systems are capable of accurate and correct predictions. Moreover, even for verbs "прохлаждаться" (to be hanging around) and "фотошопить" (to photoshop) and for the noun "кэшбэк" (cashback) the systems predicted synsets which are very close to the correct meaning, but they either cannot predict the whole variety of synsets or predict hypernyms in the proximity to the correct ones. The task of automatic taxonomy enrichment is technically feasible, but it still requires more sophisticated approaches.

As has been noted above, the most similar competition to ours is the SemEval-2018 hypernym discovery task (task 9). However, the setting used at SemEval is still quite

[9] English: https://competitions.codalab.org/competitions/22168#learn_the_details-results

different from ours—in particular, there, the participants of the task had to construct a taxonomy from scratch, whereas we ask our participants to extend an existing taxonomy. If we compare the scores of SemEval participants and models submitted to our task, we can see that models participating in our task yielded significantly higher MRR scores—almost 0.6 for the best-performing models compared to 0.3 for the winners of SemEval. This suggests that our task turns out to be easier than the full taxonomy construction. Obviously, the settings are quite diverse and cannot be compared rigorously—we asked participants to output $K = 10$ hypernym candidates, while at SemEval K was set to 15, the lexis were different, so we have no information about whether one test set was easier than the other. Finally, the tasks were for different languages. However, we can still speculate that such a large difference in scores is mainly because in our task the participants were using the existing taxonomy for their predictions. If they were not using it, as in SemEval, this task would not be any easier.

Table 5: Predicted hypernym synsets from RuWordNet for nouns from *Yuriy*'s answer (top-1 for nouns). Green color denotes predictions of the model from the ground truth.

rank	сахарин	селфи	кэшбэк
1	подсластитель	изображение (результат)	скидка
2	заменитель	фотографическое изображение	сфера деятельности
3	пищевые добавки	фотосъемка	предоставление услуги
4	добавление (то, что добавлено)	кинофотосъемка	учетная операция
5	вещество	портрет (изображение)	вексельная операция
6	сахарозаменитель	ателье бытовых услуг	учетная ставка
7	материал для изготовления	фотоателье	понизить величину
8	сахара	движение, перемещение	льгота
9	сахар	автопортрет	действие, целенаправленное действие
10	продукты питания	постоянная сущность	банковская операция
ground truth	заменитель подсластитель сахарозаменитель пищевые добавки	автопортрет портрет (изображение)	вернуть взятое возврат имущества, средств
		фотографическое изображение фотопортрет	премия бонус (вознаграждение)

Table 6: Predicted hypernyms synsets from RuWordNet for verbs from *cointegrated*'s answer (top-1 for verbs). Green color denotes predictions of the model from the ground truth.

rank	тусить	прохлаждаться	фотошопить
1	собраться в одном месте	бездельничать	воспроизвести (воссоздать, повторить в копии)
2	общение, связь	недостойное поведение	исправить недостатки, ошибки
3	веселиться	бродить туда-сюда	копирование, снятие копии
4	занятие, деятельность	находиться, пребывать	изобразить (воспроизвести)
5	отношения между людьми	лежать (находиться всем телом на поверхности)	проверить, удостовериться в правильности
6	пробыть, провести время	пробыть, провести время	обеспечить, снабдить
7	развлечься, приятно провести время	отдых	создать (сделать существующим)
8	добраться до места	идти ногами	устранить (уничтожить)
9	идти ногами	веселиться	исправить, улучшить
10	отдых	медлить	находиться, пребывать
ground truth	пробыть, провести время развлечься, приятно провести время занятие, деятельность	недостойное поведение бездельничать	преувеличить представить в виде приукрасить, выгодно представить
		медлить действие, целенаправленное действие	
	тусоваться добраться до места собраться в одном месте	освежить, восстановить силы восстановить прежнее состояние	изменить, сделать иным видоизменить

6. Conclusion

We present the results of the first shared task on Taxonomy Enrichment for Russian. For this shared task, we created a new dataset from the unpublished data of RuWordNet. 16 teams participated in the task, and almost half of them outperformed the baseline model.

Undoubtedly, the provided gold standard may not be perfect and exhaustive. Such manual evaluation of system answers would provide a more objective result, but we did not perform it because of the time constraints. Manual inspection of system outputs by an expert could reveal valid hypernyms identified by systems but absent in the gold standard data.

Moreover, the best-performing methods presented by participants might not be optimal for some words. These methods are based on fastText and similar distributional models, such as word2vec. However, it is known that these low-variance and high-bias models tend to identify the dominant meaning of a word and populate nearest neighbor lists with words related to this dominant meaning. Therefore, some rare senses of hypernyms can be underrepresented based on such methods. Identifying them correctly requires using alternative approaches.

According to the provided results, we see that the automatic hypernym candidate generation from an existing taxonomy is a feasible task, so it can be used to assist manual taxonomy enrichment. We hope that the evaluation datasets will foster further development of taxonomy induction and enrichment methods. Besides, the obtained levels of quality will allow direct use of some of the best-performing methods in the further development of lexical resources, such as thesauri, taxonomies, and ontologies.

7. Acknowledgements

The work of Natalia Loukachevitch in the current study (preparation of RuWordNet data for the shared task) is supported by the RFBR foundation (project N 18-00-01226 (18-00-01240)). We thank Dmitry Ustalov for updating the RUSSE web site with the information about the current shared task. Finally we are grateful to RUSSIR, AIST, and AINL conference organizers, Moscow NLP Seminar organizers, and Vladislav Lialin for sharing the information about this shared task in their media resources.

References

1. *Arefyev, N. et al.:* Word2vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions. In: Computational linguistics and intellectual technologies: Papers from the annual conference "Dialogue". (2020).
2. *Arefyev, N. et al.:* Evaluating three corpus-based semantic similarity systems for russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2015". 28, (2015).
3. *Baroni, M., Lenci, A.:* How we blessed distributional semantic evaluation. In: Proceedings of the gems 2011 workshop on geometrical models of natural language semantics. pp. 1–10 Association for Computational Linguistics (2011).
4. *Bernier-Colborne, G., Barrière, C.:* CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation. pp. 725–731 Association for Computational Linguistics, New Orleans, Louisiana (2018).

5. *Bojanowski, P. et al.:* Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
6. *Bordea, G. et al.:* Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1081–1091 (2016).
7. *Bordea, G. et al.:* Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1081–1091 (2016).
8. *Camacho-Collados, J.:* Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. arXiv preprint arXiv:1703.04178. (2017).
9. *Camacho-Collados, J. et al.:* SemEval-2018 task 9: Hypernym discovery. In: Proceedings of the 12th international workshop on semantic evaluation (semeval-2018); 2018 jun 5–6; new orleans, la. Stroudsburg (pa): ACL; 2018. P. 712–24. ACL (Association for Computational Linguistics) (2018).
10. *Dale, D.:* A simple solution for the Taxonomy enrichment task: Discovering hypernyms using nearest neighbor search. In: Computational linguistics and intellectual technologies: Papers from the annual conference "dialogue". (2020).
11. *Efremova, T.:* New dictionary of the russian language. Explanatory-derivational. Moscow: Russky yazyk. (2000).
12. *Hearst, M. A.:* Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics-volume 2. pp. 539–545 Association for Computational Linguistics (1992).
13. *Jurgens, D., Pilehvar, M. T.:* SemEval-2016 task 14: Semantic taxonomy enrichment. In: Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). pp. 1092–1102 (2016).
14. *Kunilovskaya, M. et al.:* Taxonomy Enrichment: Linear Hyponym-Hypernym Projection vs Synset ID Classification. In: Computational linguistics and intellectual technologies: Papers from the annual conference "dialogue". (2020).
15. *Kuratov, Y., Arkhipov, M.:* Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
16. *Kutuzov, A., Kuzmenko, E.:* WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
17. *Levy, O. et al.:* Do supervised distributional methods really learn lexical inference relations? In: Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 970–976 (2015).
18. *Loukachevitch, N., Dobrov, B. V.:* RuThes linguistic ontology vs. Russian wordnets. In: Proceedings of the seventh global wordnet conference. pp. 154–162 (2014).
19. *Loukachevitch, N. V. et al.:* Creating russian wordnet by conversion. In: Computational linguistics and intellectual technologies: Papers from the annual conference" dialogue. p. 22 (2016).

20. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
21. *Miller, G. A.:* Nouns in wordnet. WordNet: An electronic lexical database. 23–46 (1998).
22. *Miller, G. A.:* WordNet: A Lexical Database for English. Communications of the ACM. 38, 11, 39–41 (1995).
23. *Nakashole, N. et al.:* PATTY: A taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 1135–1145 Association for Computational Linguistics (2012).
24. *Panchenko, A. et al.:* RUSSE: The first workshop on russian semantic similarity. In: Computational linguistics and intellectual technologies: Papers from the annual international conference "Dialogue-2015". pp. 89–106 (2015).
25. *Panchenko, A. et al.:* Noun sense induction and disambiguation using graph-based distributional semantics. In: Proceedings of the 13th conference on natural language processing (konvens 2016). pp. 192–202 (2016).
26. *Panchenko, A. et al.:* Human and machine judgements for russian semantic relatedness. In: International conference on analysis of images, social networks and texts. pp. 221–235 Springer (2016).
27. *Roller, S. et al.:* Inclusive yet selective: Supervised distributional hypernymy detection. In: Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers. pp. 1025–1036 (2014).
28. *Sabirova, K., Lukanin, A.:* Automatic extraction of hypernyms and hyponyms from russian texts. In: Ignatov, D. I. et al. (eds.) Supplementary proceedings of the 3rd international conference on analysis of images, social networks and texts (aist'2014), yekaterinburg, russia, april 10–12, 2014. pp. 35–40 CEUR-WS.org (2014).
29. *Shwartz, V. et al.:* Improving hypernymy detection with an integrated path-based and distributional method. In: Proceedings of the 54th annual meeting of the association for computational linguistics. pp. 2389–2398 (2016).
30. *Snow, R. et al.:* Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics. pp. 801–808 Association for Computational Linguistics (2006).
31. *Tikhomirov, M. et al.:* Combined Approach to Hypernym Detection for Thesaurus Enrichment. In: Computational linguistics and intellectual technologies: Papers from the annual conference "Dialogue". (2020).
32. *Weeds, J. et al.:* Learning to distinguish hypernyms and co-hyponyms. In: Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers. pp. 2249–2259 Dublin City University; Association for Computational Linguistics (2014).