SemSketches-2021: experimenting with the machine processing of the pilot semantic sketches corpus

Maria Ponomareva[♠,♡] **Maria Petrova**[♠] **Julia Detkova**[♠] **Oleg Serikov**[♡,♢] **Maria Yarova**[♣]

[♠]ABBYY, Moscow, Russia

[♡]National Research University Higher School of Economics, Moscow, Russia

[♣]Moscow Institute of Physics and Technology, Moscow, Russia

[♢]Deeppavlov MIPT, Moscow, Russia

Abstract

The paper deals with elaborating different approaches to the machine processing of semantic sketches. It presents the pilot open corpus of semantic sketches. Different aspects of creating the sketches are discussed, as well as the tasks that the sketches can help to solve. Special attention is paid to the creation of the machine processing tools for the corpus. For this purpose, the SemSketches-2021 Shared Task was organized. The participants were given the anonymous sketches and a set of contexts containing the necessary predicates. During the Task, one had to assign the proper contexts to the corresponding sketches.

Key words: word sketches, semantic sketches, frame semantics, semantic role labeling, corpus lexicography

DOI: 10.28995/2075-7182-2021-20-560-570

SemSketches-2021: опыт автоматической обработки пилотного корпуса семантических скетчей

Мария Пономарева	**Мария Петрова**	**Юлия Деткова**	**Олег Сериков**	**Мария Ярова**
ABBYY, ВШЭ	ABBYY	ABBYY	ВШЭ, Deeppavlov	МФТИ
Москва	Москва	Москва	Москва	Москва

Аннотация

Статья посвящена различным подходам к автоматической обработке семантических скетчей. В статье представлен первый открытый корпус семантических скетчей для русского языка. На примере данного корпуса рассматриваются особенности семантических скетчей и проблемы, возникающие при их построении, обсуждаются задачи, которые могут решаться с привлечением скетчей, а также дальнейшие перспективы использования скетчей. Особое внимание уделяется возможности создания инструментов автоматической обработки корпуса. В качестве эксперимента по созданию подобных инструментов авторами проведено соревнование SemSketches-2021, в рамках которого участникам предлагалась задача по работе с корпусом скетчей, где требовалось соотнести анонимизированные скетчи с рядом контекстов для соответствующих предикатов.

Ключевые слова: скетчи слов, семантические скетчи, семантика фреймов, разметка семантических ролей, корпусная лексикография

1 Introduction

The current paper continues the work on the semantic sketches which were first presented at the Dialogue-2020 conference.

The idea of the semantic sketch was introduced in [7]. The semantic sketch is a special representation of a word's compatibility where all semantic links of the word are grouped according to their semantic relations with the core they depend on. All possible semantic dependencies are statistically ranged: first, the frequency of the collocation between the parent and the child is taken into account; second, the frequency of the semantic role for the given core (for instance, the frequency of the Agent, Locative, Object, or Time).

The most frequent collocations form the semantic sketch of the word. In [7], the authors focused on the creation of the semantic sketches and on testing the semantic mark-up used for the sketches. Namely, they measured the correctness of the predicate's choice in a set of sentences and the choice of the proper semantic roles for the predicates' dependencies.

In the present work, the focus has been made on building the pilot corpus of the semantic sketches themselves, **the SemSketches corpus**. The corpus is aimed at achieving several purposes:

1. to evaluate how representative the sketches are,
2. to elaborate some tools for processing the sketches,
3. to specify what kind of tasks the semantic sketches can help to solve, as our further plan is to integrate the sketches into the General Internet-Corpus of Russian (GICR, [4], [3]),
4. to analyze what kind of mistakes we happen to face while creating the sketches.

The idea to represent a word's meaning in the form of the semantic sketch is closely related to the main idea of distributional semantics according to which the meaning of the word can be represented through its lexical co-occurrence. The famous formula for the idea given in [10] says: "You shall know a word by the company it keeps".

Over the past few years, vector representations have become a traditional method of representing the word's semantics. The static embeddings such as word2vec [8] and FastText [9] as well as the dynamic embeddings that followed, such as ELMo [5], ULMFit [13], and BERT [2], have completely changed the NLP field. However, quality evaluations of the vector representations pose a challenge, as their serious drawback is that one can neither assess nor interpret them directly.

Whereas the vector is a numeric meaning representation, appropriate for computers, the semantic sketch can be considered its human-interpretable counterpart.

As an experiment on processing the sketches automatically, we have introduced **the SemSketches Shared Task**. One of its goals is to connect these two methods of semantic representation.

The Shared Task suggested the following problem. Participants were given the corpus of the semantic sketches with the core predicates unknown, that is, the semantic roles of the dependencies and the word-fillers of the roles were given, but not the predicates they were attached to. We have presented a set of such anonymous sketches and a list of contexts containing the predicates. The task was to create a tool that assigns the sketch to the corresponding contexts.

For most sketches, the task did not seem difficult for a human, as some of the examples will demonstrate below, but it turned out to be rather complicated for the computer, as the results of the competition showed. The corpus and the Shared Task results are available at the SemSketches github[1].

2 What is a semantic sketch

There is no need to underline the importance of using text corpora for various purposes nowadays. The size of the corpora is growing quickly. On the one hand, it gives the users more opportunities and allows one to receive more representative data. On the other hand, with a bigger corpus, more sophisticated tools are demanded to process the results of the search queries.

One of the methods to describe the word's compatibility is to present it in the form of a syntactic sketch [22]. The syntactic sketch is a lexicographical profile of a word, where word dependencies are classified by their grammatical roles and ranged by the statistics of their compatibility with the core. The syntactic sketches were first introduced in the Sketch Engine project[2] and over the past years have become widely used in lexicography, language teaching, multilingual corpora creation, various translation resources, and in a number of other areas.

The evident advantage of the syntactic sketch is its vividness: it shows simultaneously all of the most frequent syntactic dependencies of a word and arranges them in a table according to the roles. At the same time, the syntactic sketches have one strong limitation: the grammatical information they are based on does not allow one to take lexical homonymy into account, which complicates the interpretation of the obtained results.

In order to solve this problem, an attempt was made to create the semantic sketches [7], where the representation of a word's compatibility is supplemented with semantic relations between words (each

[1] `https://github.com/dialogue-evaluation/SemSketches`

[2] `www.sketchengine.eu`

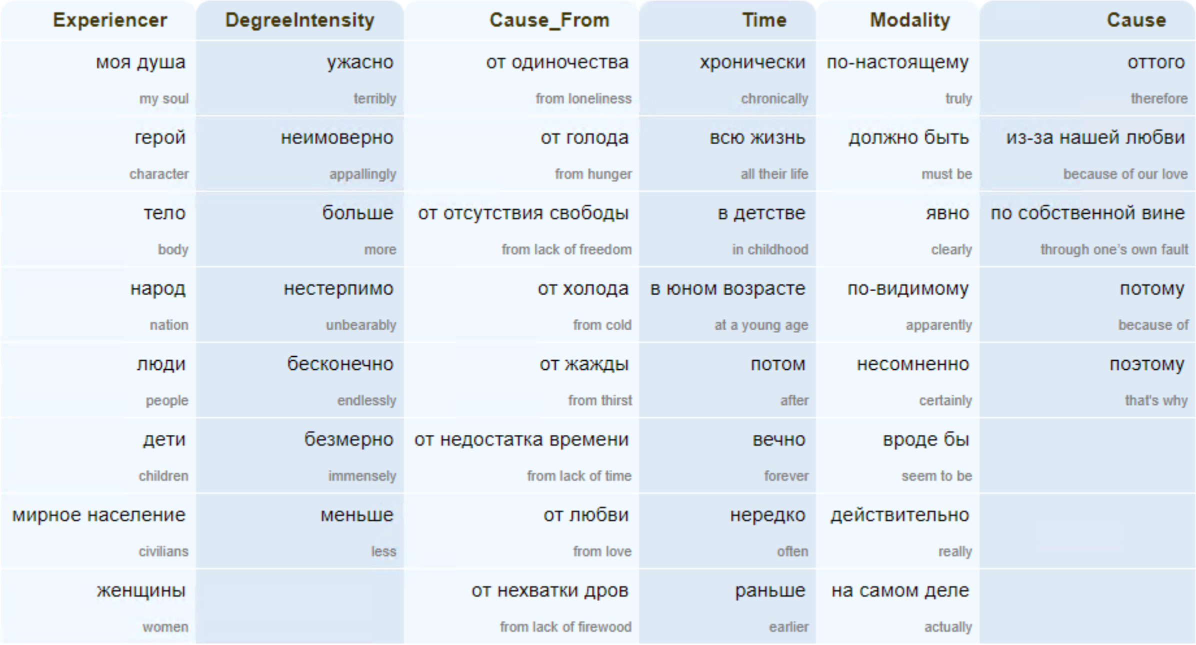
relation is marked not only with a syntactic, but with a semantic role as well) and semantic classes of words (which mark the specific semantic meaning of a word in a context).

Therefore, the semantic sketch is understood as a generalized lexicographic portrait of a word, which includes the most frequent semantic dependencies of the verb. In other words, it is a way of representing the compatibility of words, where the description of each word includes a set of its most frequent semantic dependencies classified according to their semantic roles. For each role a number of relevant "fillers" (words and phrases) are given, and the fillers are ranked according to the frequency of their compatibility with the core. Each sketch illustrates a word with a certain meaning.

The semantic sketches are built with the help of the Compreno parser [24]. Unlike other parsers, the Compreno suggests full semantic mark-up, namely, it deals not only with the actant semantic dependencies of the predicates, but with the adjuncts, modifiers, and other dependencies as well [18]. It makes the sketches an important tool for dealing with the semantic role labeling (SRL) problem which has attracted many researchers recently.

Despite high interest in the problem ([12], [11], [17], [15], [6], [16], [25]), until the current moment no research among the SRL papers has been presented (or, at least, we have not seen any), where all semantic roles are taken into account. Most works focus on the actant dependencies only, such as Agent, Object, or Experiencer. In the meantime, for many predicates, circumstantial dependencies are enough frequent and significant to get into the predicate's sketch together with its actants, and, moreover, in some cases, help to identify the core even better than the actants do.

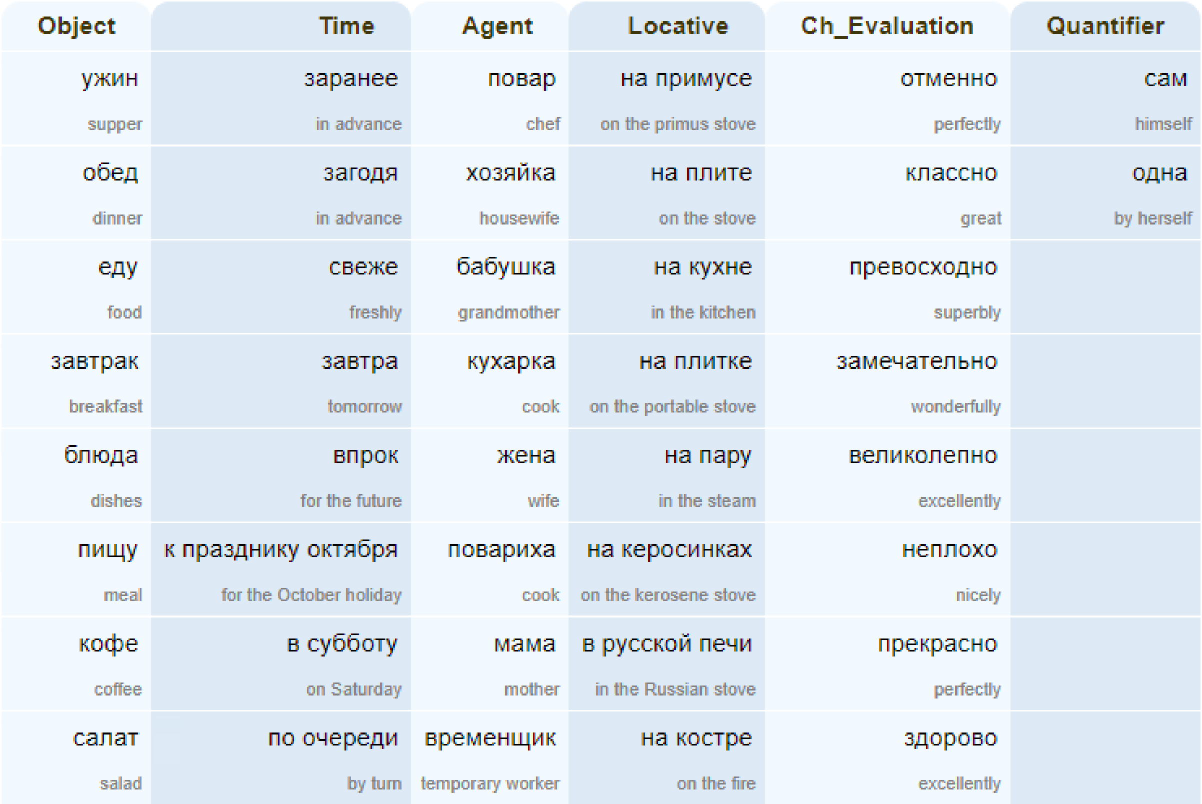
The sketches are illustrated in the two examples below, the first one — for the verb «страдать:SUFFERING_AND_TORMENT» 'to suffer' (Figure 1) and the second one — for the verb «готовить:TO_PREPARE_FOOD_SUBSTANCE» 'to prepare food, to cook' (Figure 2):



Experiencer	DegreeIntensity	Cause_From	Time	Modality	Cause
моя душа my soul	ужасно terribly	от одиночества from loneliness	хронически chronically	по-настоящему truly	оттого therefore
герой character	неимоверно appallingly	от голода from hunger	всю жизнь all their life	должно быть must be	из-за нашей любви because of our love
тело body	больше more	от отсутствия свободы from lack of freedom	в детстве in childhood	явно clearly	по собственной вине through one's own fault
народ nation	нестерпимо unbearably	от холода from cold	в юном возрасте at a young age	по-видимому apparently	потому because of
люди people	бесконечно endlessly	от жажды from thirst	потом after	несомненно certainly	поэтому that's why
дети children	безмерно immensely	от недостатка времени from lack of time	вечно forever	вроде бы seem to be	
мирное население civilians	меньше less	от любви from love	нередко often	действительно really	
женщины women		от нехватки дров from lack of firewood	раньше earlier	на самом деле actually	

Figure 1: the sketch for the verb «страдать:SUFFERING_AND_TORMENT» ('to suffer'). Here the elements of the sketch are given with their rough translations.

The participants of the Shared Task got the same representations, but did not get the titles of the sketches. However, as the pictures above demonstrate, it does not seem difficult for a human to guess the proper predicates for the sketches, which allows us to regard the sketches as representative illustrations for the verb's compatibility.



Object	Time	Agent	Locative	Ch_Evaluation	Quantifier
ужин supper	заранее in advance	повар chef	на примусе on the primus stove	отменно perfectly	сам himself
обед dinner	загодя in advance	хозяйка housewife	на плите on the stove	классно great	одна by herself
еду food	свеже freshly	бабушка grandmother	на кухне in the kitchen	превосходно superbly	
завтрак breakfast	завтра tomorrow	кухарка cook	на плитке on the portable stove	замечательно wonderfully	
блюда dishes	впрок for the future	жена wife	на пару in the steam	великолепно excellently	
пицу meal	к празднику октября for the October holiday	повариха cook	на керосинках on the kerosene stove	неплохо nicely	
кофе coffee	в субботу on Saturday	мама mother	в русской печи in the Russian stove	прекрасно perfectly	
салат salad	по очереди by turn	временщик temporary worker	на костре on the fire	здорово excellently	

Figure 2: the sketch for the verb «готовить:TO_PREPARE_FOOD_SUBSTANCE» ('to prepare food, to cook'). Here the elements of the sketch are given with their rough translations.

3 The SemSketches Shared Task

To explore the semantic sketches as far as their quality and representativeness are concerned, we have created the pilot corpus of Russian semantic sketches and made it the basis for the SemSketches Shared Task. The problem was formulated as follows: *given a set of anonymized sketches and a set of contexts for different predicates, one should match each predicate in its context to a relevant sketch.*

The second part of the competition data is the set of the contexts given for different predicates. In the case of ambiguous predicates, the WSD problem can be stated.

3.1 Data preparation

Sketches

The sketches were built on the texts from the Magazine Hall of the GICR.

Although the parser gives us the full semantic mark-up, we have implemented some restrictions for the present research. As in [7], the authors have taken only verbal cores and their subtrees: all verbs are marked with semantic classes (denoting their meanings) and the semantic roles for their direct dependencies.

We did not mark the dependencies of the non-vebal cores, the dependencies of the ellipted verbs and the ellipted groups themselves, as well as the syntactically moved groups. In addition, we have introduced some additional restrictions for the purpose of the current competition, namely, we have excluded pronouns and personal nouns, as they complicate the work with the anonymized sketches.

For the current corpus, we have chosen only verbs which have at least two meanings, as it makes the task of defining proper sketches more interesting, on the one hand, and, on the other hand, contributes to solving the WSD problem. It means that each verb chosen entered at least two semantic classes.

The number of such verbs for the Russian language turned out to be more than ten thousand. Then we chose a subset of the list through selecting the verbs by the following principles.

At the beginning, we have ranged the sample so that the verbs with the most frequent meanings came first: for instance, the verb *рубить* meaning TO_HACK (***рубить*** *дерево* — 'to hack a tree') is sufficiently frequent, while the same verb meaning TO_KNOW_ABOUT (***рубить*** *в математике* — 'to understand mathematics well') is rather marginal and has thus been positioned at the end of our list. The frequency of different meanings has been obtained with the help of the Compreno parser.

Next, we have collected the verbs' sketches taking into account the number of the relations the verb has in the corpus. Namely, we have collected all the semantic dependencies for each meaning of each verb in our marked-up corpus, and if the number of the dependent nodes exceeded the threshold of 2000, the predicate in the certain value was selected for inclusion in the final set. During this procedure, all dependencies were taken into account — both different and repeated, in order not to lose any frequent predicates with limited lexical compatibility. At the same time, the threshold was rather high to preserve the quality of the sketches.

At last, the final number of sketches in the pilot corpus became 915. Due to the exclusion of rare meanings, some verbs kept only one meaning in the sample, that is, the terminal verb list contained both polysemantic verbs with several meanings in the sample and polysemantic verbs which entered in our sample only in one (the most frequent) meaning.

The next step was to analyze the correctness of the sketches, namely, to check whether the semantic dependencies and the fillers of the dependencies that got into the sketch really refer to the verb in the given meaning. The errors check was performed for a subsample of the corpus which formed the manual Dev data (see below).

Most errors refer to situations where the more frequent homonym influences the less frequent one. For instance, the verb *писать* meaning 'to paint' (***писать*** *портрет с кого-л.* — 'to paint smb.'s picture') is less frequent than *писать* meaning 'to write' (***писать*** *письмо* — 'to write a letter'), so the sketch for the *писать* — 'to paint' contains some incorrect examples in the Object dependency — such as 'to write letters'.

The reason is that when building the semantic structures for the sentences the sketch is based on, the structure with the incorrect but more frequent homonym gets a higher evaluation due to the high statistics of the more frequent verb.

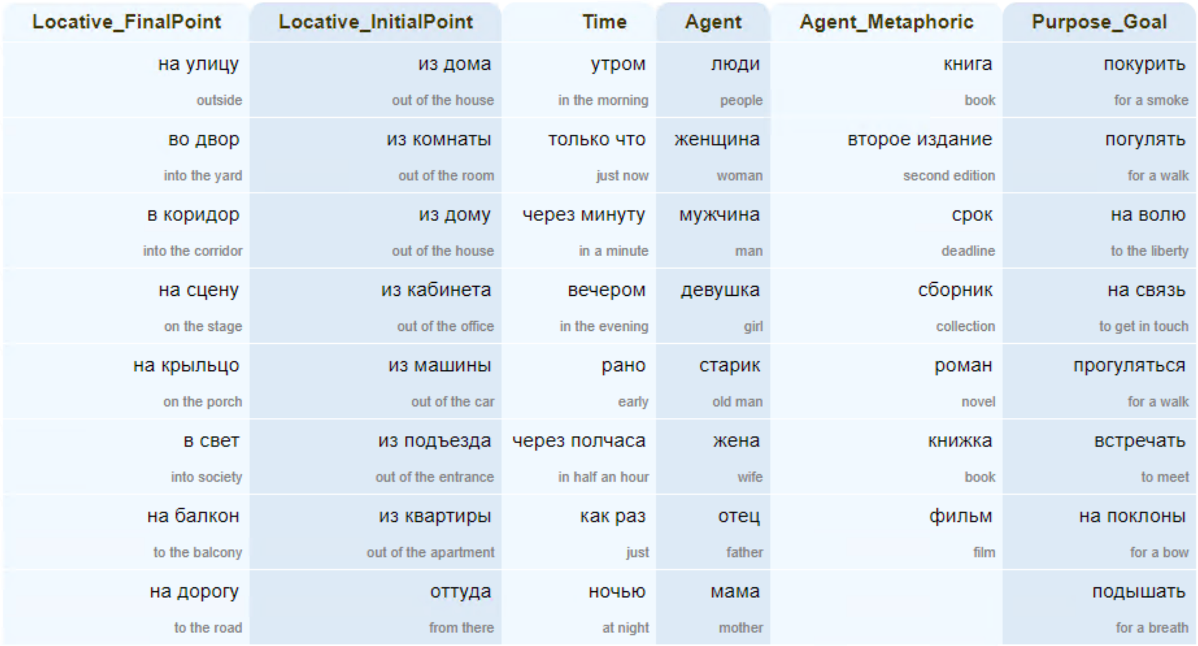
Another error can be illustrated with the sketch «готовить:TO_PREPARE_MEDICINE _OR_FOOD» 'to cook'. It contains combinations like ***готовить*** *резервную копию* — 'to prepare a reserve copy'. Here the problem is that the compatibility of 'copy' with the verbs depends not on the 'copy' itself but on the semantics of the noun following it, that is, 'the copy of the cake' is also possible.

As an instance of the sketch with the incorrect semantic dependency, let us take the sketch «выходить:идти:TO_WALK» 'go out' on the Figure 3. The sketch contains the Agent_Metaphoric slot which must be definitely referred to another meaning, and the Purpose_Goal slot contains the incorrect filler *на связь* (*выйти* ***на связь*** means 'to get in touch', and here another homonym of the verb *выйти* is supposed to be):

The main reasons for the mistakes in the sketches are the incorrect influence of the statistics, certain inaccuracies of the semantic models in the parser, and the impossibility of distinguishing between the homonyms due to the closeness of their meanings or lack of distinguishing context in the sentences.

Contexts

Every meaning from the chosen set is illustrated with contexts. A context is a sentence with one target predicate highlighted. No additional mark-up is presented. Each meaning corresponds to several dozens of contexts with the target words having this meaning. The contexts were collected from news, fiction, publicistic texts, being close by genre to those presented in Magazine Hall. It is important that the contexts do not overlap with the corpus which the sketches were built on. The excerpt from the contexts is given in Table 1.



Locative_FinalPoint	Locative_InitialPoint	Time	Agent	Agent_Metaphoric	Purpose_Goal
на улицу outside	из дома out of the house	утром in the morning	люди people	книга book	покурить for a smoke
во двор into the yard	из комнаты out of the room	только что just now	женщина woman	второе издание second edition	погулять for a walk
в коридор into the corridor	из дому out of the house	через минуту in a minute	мужчина man	срок deadline	на волю to the liberty
на сцену on the stage	из кабинета out of the office	вечером in the evening	девушка girl	сборник collection	на связь to get in touch
на крыльцо on the porch	из машины out of the car	рано early	старик old man	роман novel	прогуляться for a walk
в свет into society	из подъезда out of the entrance	через полчаса in half an hour	жена wife	книжка book	встречать to meet
на балкон to the balcony	из квартиры out of the apartment	как раз just	отец father	фильм film	на поклоны for a bow
на дорогу to the road	оттуда from there	ночью at night	мама mother		подышать for a breath

Figure 3: the semantic sketch for the verb «выходить:идти:TO_WALK» ‘to go out’). Here the elements of the sketch are given with their rough translations.

ID	dev.sent.rus.116
target	наполнились
start	46
end	57
context	Когда доктор вошел, она вспыхнула, и глаза ее наполнились слезами

Table 1: The example of the context. The position of the target word *наполнились* ‘filled’ in the context ‘When the doctor came in, she flushed, and her eyes filled with tears’ is defined by the offsets.

Datasets

The task was meant to be solved in a few-shot or unsupervised learning manner. During the Shared Task, we provided the participants with two sets of data. In the first phase, the Trial data was published. It comprises three parts: a set of sketches, a set of contexts, and mapping between these two sets. The participants could use the data to get familiar with the formats, to test their hypotheses and to fine-tune their systems. During the second phase, we provided the participants with the main set of the sketches and corresponding contexts, which will be referred to as Dev data.

In contrast to trial data, where the mapping had been given, for Dev data the participants were asked to find the relations between the sketches and the contexts themselves.

For the third phase, we manually selected 100 sketches and evaluated the corresponding contexts. This data formed the gold standard set for the task, which we will refer to as Manual Dev data. Table 2 shows the size of the obtained datasets.

During the second phase, the participants were able to commit their answers to the CodaLab[3] to know the results on the Dev data and to choose the best decision. During the third phase, the performance of the best variants was finally evaluated on the Manual Dev data.

After the announcement of the results, we published the answers (the mapping between the sketches and the contexts) on the SemSketches github.

[3] https://competitions.codalab.org/competitions/29992

split	number of sketches	number of contexts
Trial	20	2000
Dev	895	44750
Manual Dev	100	4347

Table 2: The size of the SemSketches datasets, *Manual Dev* data forms a subset of *Dev* data

3.2 Evaluation metric

The submitted systems were evaluated using the **accuracy** metric. For the Shared Task, accuracy was calculated as the number of matched pairs between the participants' answers and test markup divided by the total number of contexts.

The evaluation script is publicly available on the SemSketches github.

3.3 Baseline

The participants were provided with a weak baseline solution. The solution was based on the masked language modeling (MLM) mechanism of the RuBERT [14] model.

For a given context *cont*, *sketch* was chosen according to the computed sketch scores based on MLM candidates. MLM candidates (MLM^N_{cont}) were calculated as follows:

1. syntactic analysis using the UDPipe ([23]) has been performed to find the direct dependents of the target predicate;
2. for each of the direct dependents, top-*N* mask replacements Rep^N_{dep} were stored;
3. stored replacements were intersected, i.e. $MLM^N_{cont} = \bigcap\{Rep^N_{dep}\ \forall dep \in cont\}$;
4. sketch *Score* was computed as the number of tokens present in the intersection of the sketch representation and the stored MLM candidates.

$$Score\,(sketch, cont) = \left|MLM^{1000}_{cont} \cap Tokens_{sketch}\right|$$

The intersection was performed over lemmas thus treating *на заре* and *заря* as intersecting entries.

The weak baseline system has shown 0,0094 accuracy on the Dev data set thus overperforming the random baseline.

3.4 Submitted systems

Three teams participated in the Shared Task: `paleksandrova`, `good501`, `smpl`. All teams suggested the solutions based on different approaches, and each solution managed to overcome the baseline. However, the final scores of each team turned out to be rather modest. To compare the results achieved, see Table 3 where the score of each team and the baseline score are presented.

Team	Dev Score	Manual Dev Score
`paleksandrova`	0.309	0.277
`good501`	0.104	0.127
`smpl`	0.182	0.121
`baseline`	0.0094	0.0035

Table 3: SemSketches Shared Task: the results of the submitted and baseline systems.

Let us now shortly characterize each decision and analyze what core problems they faced.

The team `smpl` used the brute-force approach. LM score has been used to rank sketches and choose the best one for each context. To estimate how well the predicate *pred* fits into the given sketch *sketch*, the *LM score* was used. *LM score* is the average probability of *pred* to replace *[MASK]* token in template sentence «*[MASK] cell*». Template sentences were generated for each *cell* present in *sketch*.

The team `Good501` used the approach based on the sentences' similarity objective, which is a popular objective when training language models. Target predicate was highlighted in the sentence using special tags. Sketch tables were flattened into pseudo-sentences. For the given sentence, the most similar sketch was chosen by using the Sentence-BERT[21] siamese similarity mechanism.

The team `paleksandrova` [1] used the MLM approach, which consisted of first restoring the covered predicate for each of the given sketches, and then picking the relevant sketch for the target sentence.

The covered predicates were restored by generating templates (e.g. «*[MASK] в школу*» — '[MASK] to school') using the sketch content cells. The most frequent predicate of all the MLM hypotheses for the sketch's templates was treated as the re-covered predicate. The first sketch whose predicate matched the sentence predicate was used as the system answer. When no sketch was found by exact matching, the sketch whose restored predicate was word2vec-closest [8] to the sentence predicate was used as the answer.

3.5 The analysis of the submitted systems

During the Shared Task, we formulated the experimental problem leaving enough room for different approaches. Although the performance of the submitted systems may be improved significantly, the proposed ideas were encouragingly diverse and thought provoking. The common feature of all three systems is using the pretrained Language Models.

The team `501good` which adopted the approach from Sentence Transformers introduced the only system that included training. The model was trained on the *Trial* data (20 sketches).

The systems of `smpl` and `paleksandrova` defined their unsupervised strategies for mapping the sketches and the contexts. While the `smpl` team estimated how well each target predicate fits to each sketch using the score from the Masked Language Model, the `paleksandrova` team suggested an original approach imitating the way humans guess the core of the anonymous sketch.

It is worth mentioning that the approaches of `paleksandrova` and `smpl` by design cannot disambiguate the polysemous predicate, as they take only the target verb into account but not its context.

The team `smpl` approach could be thought of as scoring how well the sketch could account as the sentence predicate core. LM is trained on sentence-level objective, therefore, the successful application of the similarity approach demands more sophisticated preprocessing of the input sequence, for example, taking the predicate contexts into account. Such modification could improve the results.

The team `paleksandrova` approach seems to be the most promising one. But the accuracy turned out to be rather low for the following reason. The sketch accumulates several verb forms, namely, it includes all tense, aspect and voice forms. For instance, the verbs *строить* 'build' <Imperfective, NonReflexive>, *построить* 'build' <Perfective, NonReflexive>, *строиться* 'build' <Imperfective, Reflexive>, *построиться* 'build' <Perfective, Reflexive> refer to one sketch. As far as `paleksandrova` approach is concerned, the team regarded such verbs as different candidates for a sketch. At the same time, they chose only one top candidate for each sketch. Therefore, only one grammatical form of the necessary set could be referred to the right sketch.

4 Discussion

In the current paper, we demonstrated the pilot corpus of the semantic sketches, gave a brief analysis of the problems we faced during the corpus creation, and described the results of the SemSketches Shared Task aimed at applying the machine processing tools to the corpus.

The sketches are based on the parser with full semantic mark-up, which defines their value and uniqueness: first, the sketches allow one to analyze not only the actant dependencies, but a full semantic model of a word; second, they differentiate between the various meanings of the verbs.

As far as the opportunities for theoretical investigations are concerned, the sketches can help in dealing with all problems bound with the semantic compatibility of words. Especially, the SRL and the WSD problems must be mentioned here.

As noted above, most researchers focus mainly only on the actant roles, while other dependencies do not usually get much attention. The semantic sketches suggest interesting data in this respect. The sketches include most frequent collocations, that is, the most natural, most typical contexts of a word. Among the dependencies the sketches include, modifiers and adjuncts are quite frequent. For some verbs, they seem to be even more specific than the actants and give more help in identifying the predicate.

For instance, the Locative is a typical circumstantial adjunct, but it is an obligatory slot for the verbs with the position meaning such as *быть* 'be', *находиться* 'be situated'. The Locative slot helps to differentiate between the 'be' with the position meaning and other be-homonyms, while the semantic role corresponding syntactically to the Subject of 'be' does not really contribute to differentiating between the be-homonyms.

It seems that the meaning of the adjuncts and the modifiers is sometimes underevaluated, therefore, an interesting task is to evaluate the correlation between the actant and circumstantial dependencies in the sketches.

As for the applied tasks, one of the promising directions in using the semantic sketches is their implementation for probing tasks for the pretrained language models. The interpretation of the linguistic knowledge encoded in the pretrained models has attracted much attention recently ([26], [19], [20]). We believe that the semantic sketches can serve as a basis for both probing tasks and linguistically-motivated fine-tune tasks for such models.

To summarize, the ideas from the proposed approaches can be used to embed effectively semantic sketches, making them not only a tool for manual lexicographical work but a semantic representation valid for automatic methods of Natural Language Processing.

5 Further plans

Our next plan is to add the sketches into the GICR, which brings two problems to consider.

The first one deals with the errors evaluation: in the current work, we did not check all the sketches in the pilot corpus manually — only the manual Dev data. Therefore, we did not evaluate the total number of the mistakes in the whole corpus. This task is still to be done, including work on both, that is, sketches that seem to be unsuitable (checking the manual Dev data shows that such cases are rare) and sketches containing single mistakes in either the semantic dependencies or their fillers.

The second question is about the processing tools the sketches should be provided with. The SemSketches Shared Task demonstrated that machine tools can be successfully applied to the sketches processing (in spite of the fact that the precision of the solutions suggested by the applicants was not really high). What the tools should look like, depends significantly on the tasks the sketches will be used to solve.

At the same time, we have recently started work on the English sketches, so our further plans include adding other languages to the sketch model, starting with the English sketches.

References

[1] Aleksandrova Polina, Mokhova Anna, Nikolaenkova Maria. Matching semantic sketches to predicates in context using the BERT model // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2021.

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[3] Big and diverse is beautiful: A large corpus of Russian to study linguistic variation / Alexander Piperski, Vladimir Belikov, Nikolay Kopylov et al. // Proc 8th Web as Corpus Workshop (WAC-8). — 2013.

[4] Corpus as language: from scalability to register variation / Vladimir Belikov, Nikolay Kopylov, Alexander Piperski et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Bekasovo, 2013.

[5] Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 2018. — Jun. — P. 2227–2237. — Access mode: `https://www.aclweb.org/anthology/N18-1202`.

[6] Deep Semantic Role Labeling With Self-Attention / Zhixing Tan, Mingxuan Wang, Jun Xie et al. // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018 / Ed. by Sheila A. McIlraith, Kilian Q. Weinberger. — AAAI Press, 2018. — P. 4929–4936. — Access mode: `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16725`.

[7] Differential Semantic Sketches For Russian Internet-Corpora / Julia Detkova, Valeriy Novitskiy, Maria Petrova, Vladimir Selegey // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2020.

[8] Distributed Representations of Words and Phrases and their Compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Neural and Information Processing System (NIPS). — 2013. — Access mode: `https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf`.

[9] Enriching Word Vectors with Subword Information / Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — P. 135–146.

[10] Firth J. A Synopsis of Linguistic Theory 1930-1955 // Studies in Linguistic Analysis. — Philological Society, Oxford, 1957. — reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

[11] Generalized Inference with Multiple Semantic Role Labeling Systems / Peter Koomen, Vasin Punyakanok, Dan Roth, Wen-tau Yih // Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). — Ann Arbor, Michigan : Association for Computational Linguistics, 2005. — Jun. — P. 181–184. — Access mode: `https://www.aclweb.org/anthology/W05-0625`.

[12] Gildea Daniel, Jurafsky Daniel. Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.

[13] Howard Jeremy, Ruder Sebastian. Fine-tuned Language Models for Text Classification // CoRR. — 2018. — Vol. abs/1801.06146. — 1801.06146.

[14] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — 1905.07213.

[15] Lang Joel, Lapata Mirella. Unsupervised Semantic Role Induction with Graph Partitioning // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. — Edinburgh, Scotland, UK. : Association for Computational Linguistics, 2011. — Jul. — P. 1320–1331. — Access mode: `https://www.aclweb.org/anthology/D11-1122`.

[16] Learning Structured Natural Language Representations for Semantic Parsing / Jianpeng Cheng, Siva Reddy, Vijay Saraswat, Mirella Lapata // Proceedings of the 55th Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada : Association for Computational Linguistics, 2017. — Jul. — P. 44–55. — Access mode: https://www.aclweb.org/anthology/P17-1005.

[17] Palmer Martha Stone. Semantic role labeling. Synthesis lectures on human language technologies ; #6. — San Rafael, Calif.] : Morgan & Claypool Publishers, 2010. — ISBN: 9781598298314.

[18] Petrova M.A. The Compreno Semantic Model: The Universality Problem // International Journal of Lexicography. — 2013. — 12. — Vol. 27, no. 2. — P. 105–129. — https://academic.oup.com/ijl/article-pdf/27/2/105/2731792/ect038.pdf.

[19] Probing Pretrained Language Models for Lexical Semantics / Ivan Vulić, E. Ponti, Robert Litschko et al. // ArXiv. — 2020. — Vol. abs/2010.05731.

[20] Ravichander Abhilasha, Belinkov Yonatan, Hovy Eduard. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? — 2021. — 2005.00719.

[21] Reimers Nils, Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 3982–3992. — Access mode: https://www.aclweb.org/anthology/D19-1410.

[22] The Sketch Engine: ten years on / Adam Kilgarriff, Vít Baisa, Jan Bušta et al. // Lexicography. — 2014. — P. 7–36.

[23] Straka Milan, Straková Jana. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe // Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Vancouver, Canada : Association for Computational Linguistics, 2017. — August. — P. 88–99. — Access mode: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

[24] Syntactic and semantic parser based on ABBYY Compreno linguistic technologies / K. V. Anisimovich, K. Ju. Druzhkin, F. R. Minlos et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Bekasovo, 2012.

[25] Syntax for Semantic Role Labeling, To Be, Or Not To Be / Shexia He, Zuchao Li, Hai Zhao, Hongxiao Bai // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 2018. — Jul. — P. 2061–2071. — Access mode: https://www.aclweb.org/anthology/P18-1192.

[26] What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties / Alexis Conneau, German Kruszewski, Guillaume Lample et al. // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 2018. — Jul. — P. 2126–2136. — Access mode: https://www.aclweb.org/anthology/P18-1198.