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

Maria Ponomareva, Maria Petrova, Julia Detkova, Oleg Serikov, Maria Yarova

18.06.2021

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,
- the most frequent collocations form the semantic sketch of the word.

Work on the semantic sketches

Last year

- creation of the semantic sketches
- testing the semantic mark-up used for the sketches

This year

- creation of the first pilot open corpus of the semantic sketches
- experiment on creating the machine processing tools for the corpus

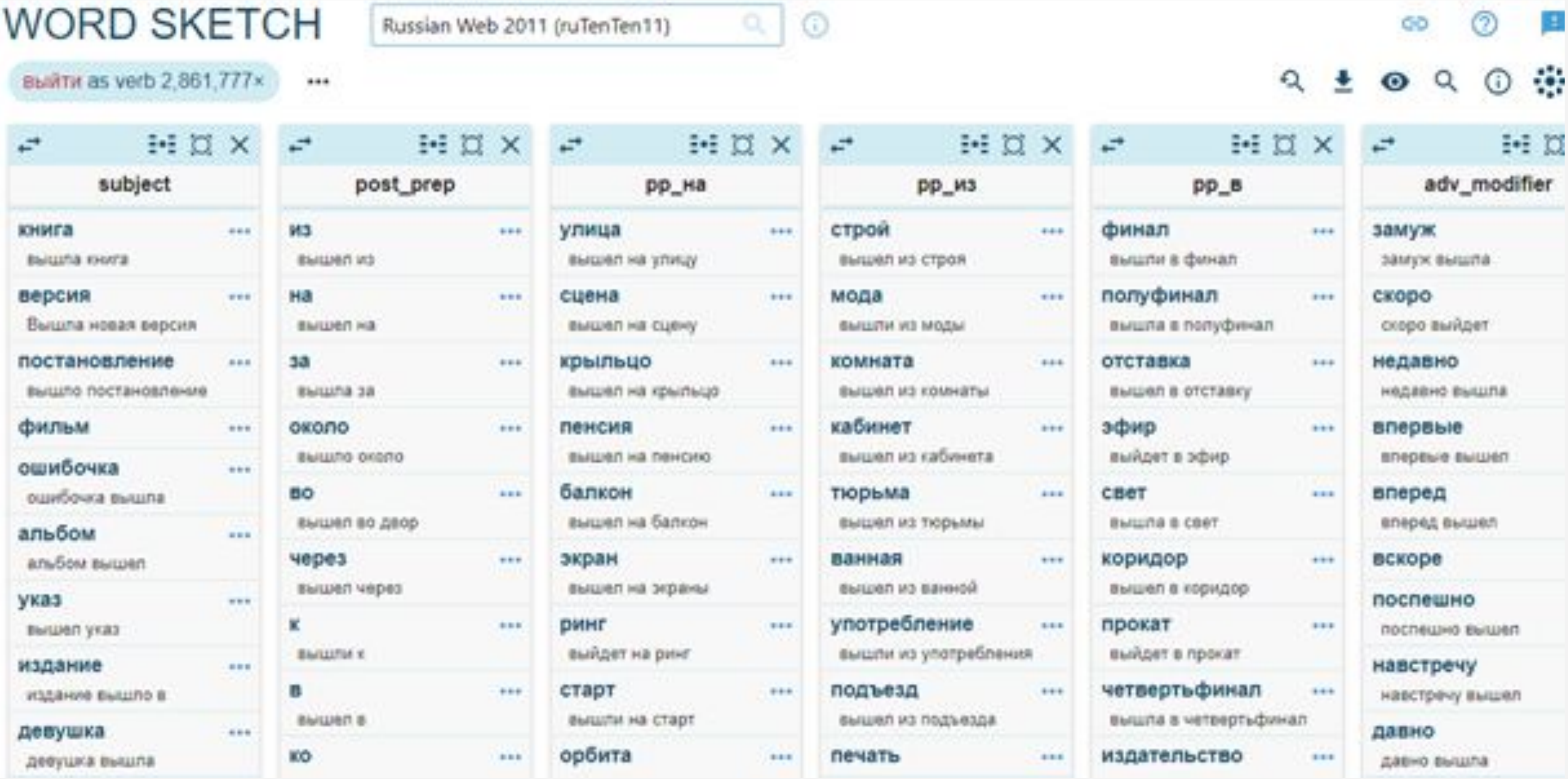
Purposes of the corpus

- to evaluate how representative the sketches are,
- to elaborate some tools for processing the sketches,
- 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,
- to analyze what kind of mistakes we happen to face while creating the sketches.

Syntactic sketches

Adam Kilgarriff   Sketch Engine Project *www.sketchengine.eu*

Syntactic sketch - a lexicographic profile of a word, where word dependencies are classified by their **grammatical** roles and ranged by the statistics of their compatibility with the core.



WORD SKETCH

Russian Web 2011 (ruTenTen11)

выйти as verb 2,861,777×

subject	post_prep	pp_на	pp_из	pp_в	adv_modifier
книга вышла книга	из вышел из	улица вышел на улицу	строй вышел из строя	финал вышли в финал	замуж замуж вышла
версия Вышла новая версия	на вышел на	сцена вышел на сцену	мода вышли из моды	полуфинал вышла в полуфинал	скоро скоро выйдет
постановление вышло постановление	за вышла за	крыльцо вышел на крыльцо	комната вышел из комнаты	отставка вышел в отставку	недавно недавно вышла
фильм	около вышло около	пенсия вышел на пенсию	кабинет вышел из кабинета	эфир выйдет в эфир	впервые впервые вышел
ошибочка ошибочка вышла	во вышел во двор	балкон вышел на балкон	тюрьма вышел из тюрьмы	свет вышла в свет	вперед вперед вышел
альбом альбом вышел	через вышел через	экран вышел на экраны	ванная вышел из ванной	коридор вышел в коридор	вскоре
указ вышел указ	к вышли к	ринг выйдет на ринг	употребление вышли из употребления	прокат выйдет в прокат	поспешно поспешно вышел
издание издание вышло в	в вышел в	старт вышли на старт	подъезд вышел из подъезда	четвертьфинал вышла в четвертьфинал	навстречу навстречу вышел
девушка девушка вышла	ко	орбита	печать	издательство	давно давно вышла

Syntactic sketches

Advantage - vividness:

- shows simultaneously all of the most frequent dependencies
- arranges them in a table according to the roles

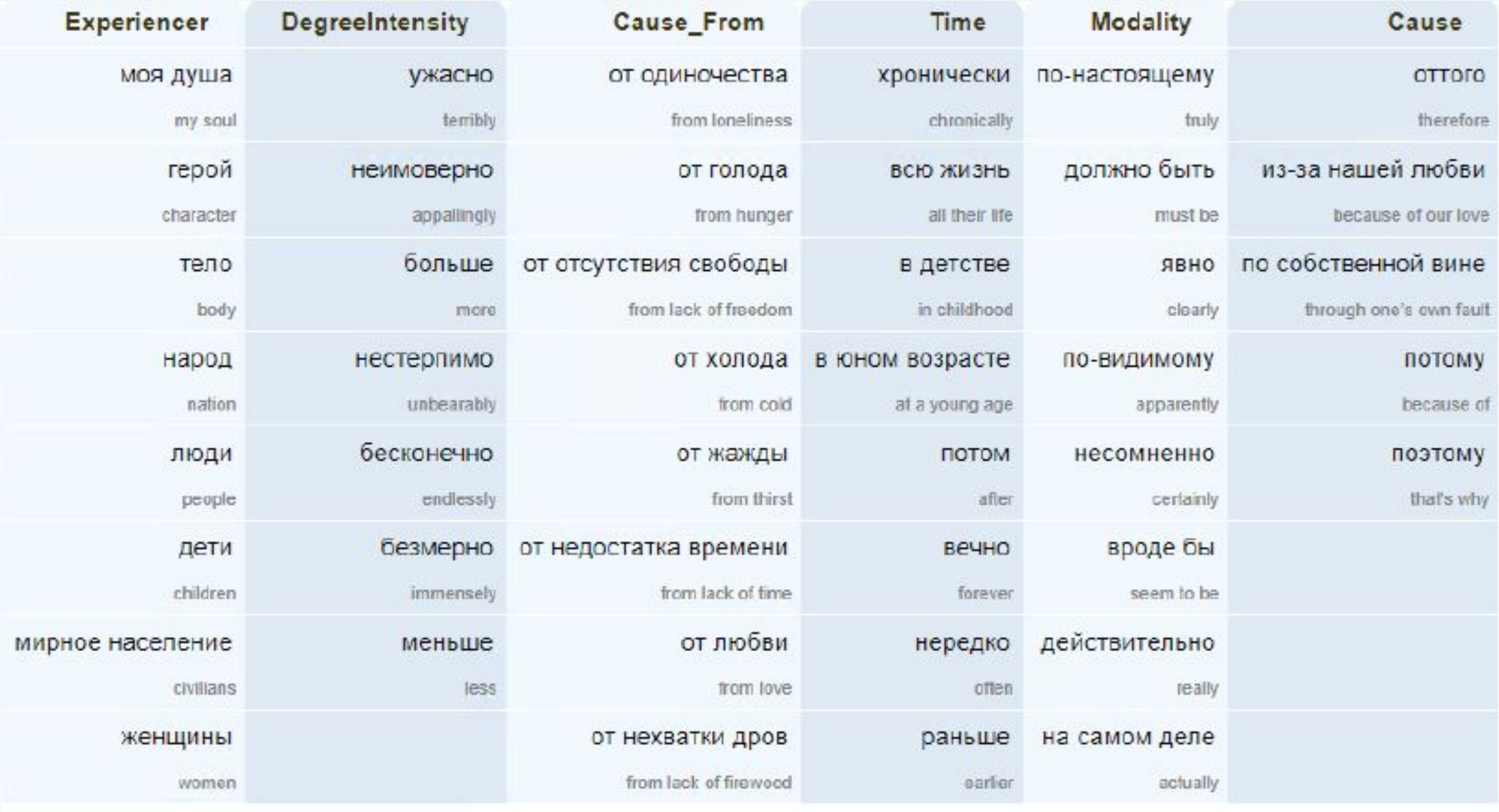
Disadvantage:

no opportunity to take lexical homonymy into account

Semantic sketches

Semantic sketch - a generalized lexicographic portrait of a word, where word dependencies are classified by their **semantic** roles and ranged by the statistics of their compatibility with the core

SemSketch for <<страдать:SUFFERING_AND_TORMENT>> ‘to suffer’



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	

Semantic sketches

- are built on the Compreno parser with full semantic mark-up
- include both actants and adjuncts/modifiers
- one sketch = one meaning
- each “filler” of a semantic role enters a sketch in one meaning
- include the frequency of the collocation between the parent and the child
- include the frequency of the semantic role for the given core

Semantic sketches can contribute to the tasks of:

- semantic role labeling (SRL)
- word sense disambiguation (WSD)
- all tasks bound with word compatibility

The SemSketches Pilot Corpus

- texts from the Magazine Hall of the GICR

- all verbs are marked with

  - semantic classes (denoting their meanings)

  - the semantic roles for their direct dependencies

1. Restrictions on the mark-up:

- only verbal cores and their subtrees
- we did not mark:
  - the dependencies of the non-verbal cores,
  - the dependencies of the ellipted verbs and the ellipted groups themselves,
  - the syntactically moved groups
- no pronouns and personal nouns (as they complicate the work with the anonymized sketches)

2. Choice of verbs for the corpus:

Stage 1: only verbs with at least two meanings => more than 10 000 verbs

Stage 2: ranging the sample by frequency of meanings (by the Compreno parser)

рубить `to hack a tree' (frequent => top of the list) vs

рубить `to understand well' (marginal => end of the list)

Stage 3: collecting all semantic dependencies for each meaning of each verb in our marked-up corpus

Stage 4: if the number of the dependent nodes (both different and repeated)

> 2000, the predicate (in this meaning) enters the final set

Final corpus

Final number of sketches in the pilot corpus - 915.

NB:

Due to the exclusion of rare meanings, the terminal verb list contained both verbs with several meanings in the sample and verbs with one (the most frequent) meaning.

Correctness of the sketches

The check was performed on a subsample of the corpus - manual Dev data:

- 100 sketches.

Types or errors

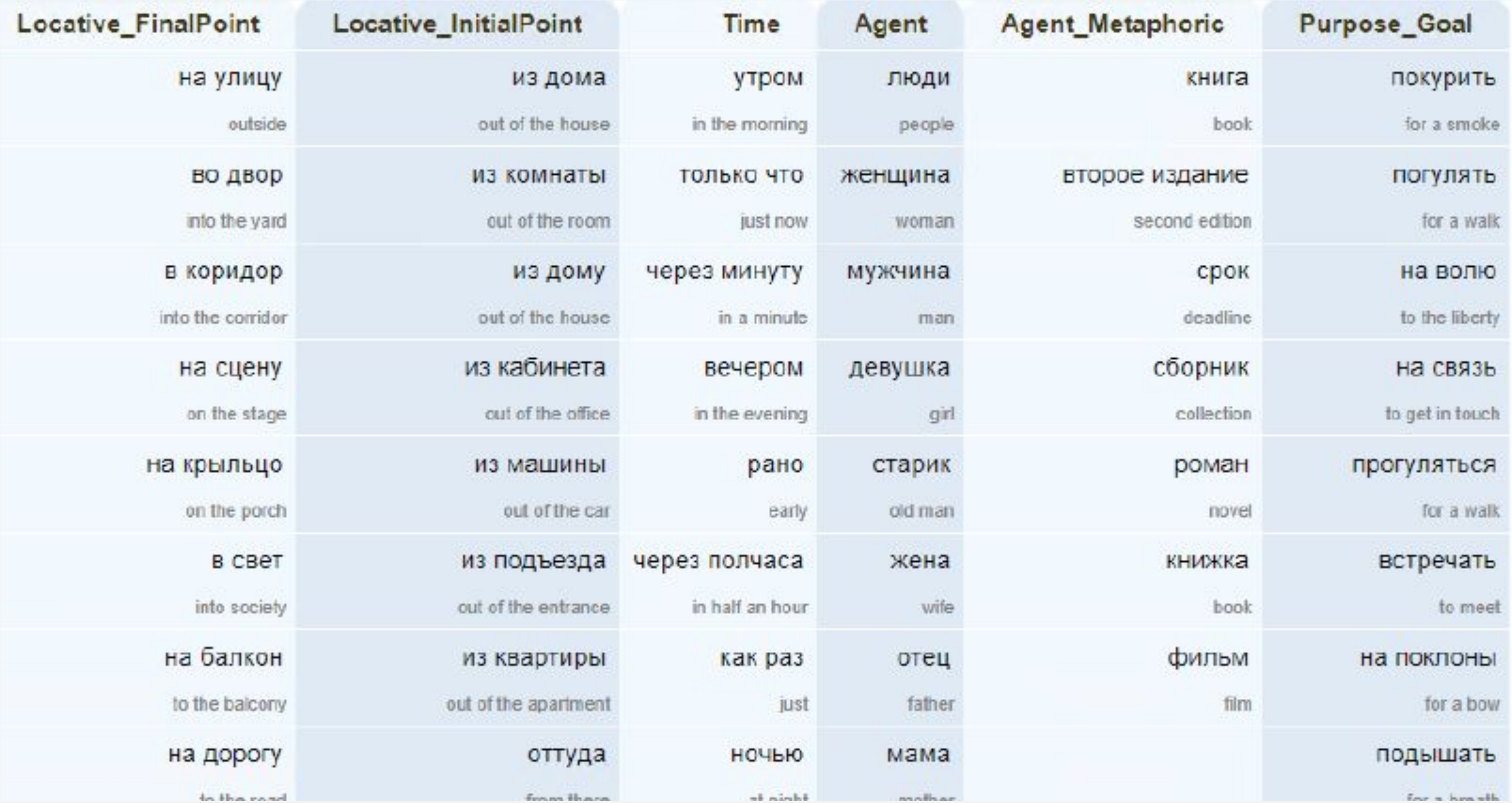
(1) More frequent homonym influences the less frequent one:
писать портрет с кого-либо ‘to paint smb.’s picture’ vs *писать* ‘to write’

(2) The filler of the dependency is a ‘lexical core’:

<<готовить:TO_PREPARE_MEDICINE_OR_FOOD>> `to cook': > *готовить резервную копию* `to cook a reserve copy'

(3) Certain inaccuracies of the semantic models in the parser (see next slide):

Mistakes in SemSketch *выходить* ‘go out’



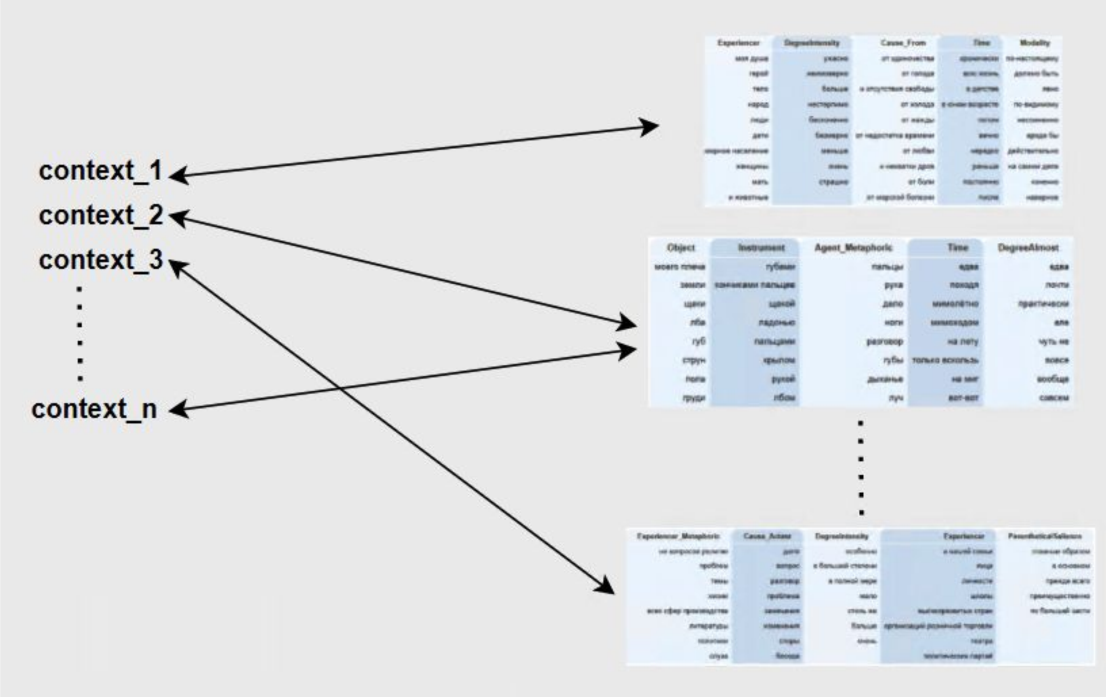
Locative_FinalPoint	Locative_InitialPoint	Time	Agent	Agent_Metaphoric	Purpose_Goal
на улицу outside	из дома out of the house	утром in the morning	люди people	книга book	покурить for a smoke
во двор into the yard	из комнаты out of the room	только что just now	женщина woman	второе издание second edition	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

SemSketches Shared Task

- Formalizing the task
- Data
- Baseline
- Overview of participating systems
- Results and Discussion

SemSketches Shared Task

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.

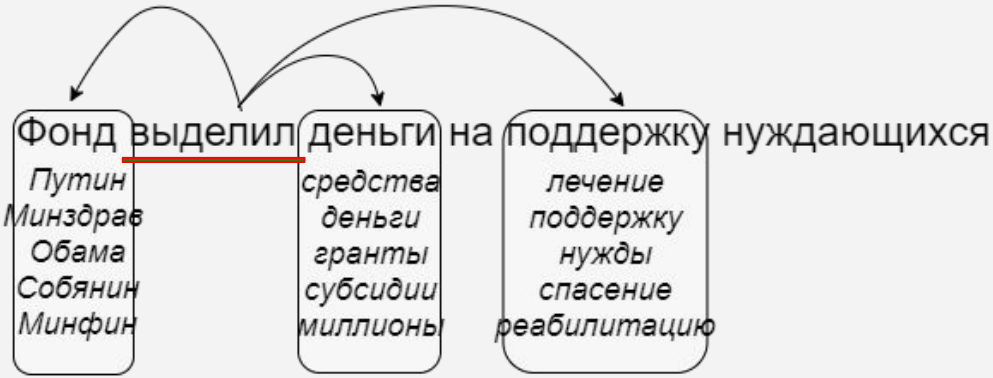


Data

Split	Number of sketches	Number of contexts
Trial	20	2000
Dev	895	44750
Manual Dev	100	4347

```
"Dev.sent.rus.1": {
        "instance": "пожал",
        "start": 44,
        "end": 49,
 "sentence": "Он не спеша подошел к полковнику Эмсуорту и пожал ему руку"}
```

Baseline



For each context:

- find the direct dependents of the target predicate (UDpipe);
- select top-N mask replacements for each of the direct dependents using MLM (RuBERT);
- unite the replacements to obtain MLM candidates;
- for each sketch compute the Score as the number of tokens present in the intersection of the sketch representation and the stored MLM candidates;
- map the context to the sketch with the max Score.

Submitted systems

- 3 participating systems
- 3 different approaches
- modest results, but much better than the baseline

Submitted system #1 (the **smpl** team)

Going from the context to the sketch

For each context:

- normalise the predicate norm(pred)

  “поиграл” ➔ “поиграть” ‘played’ ‘play’

- for every sketch generate 6 templates (for each semantic role): *norm(pred)* + *cell filler*

  "поиграть в карты", "поиграть с друзьями"... ‘play cards’ ‘play with friends’

- the number of templates may grow during the replacement of each subtoken of norm(pred) one by one with [MASK]

  [MASK, ‘##игр’, ‘##ать’, ‘в’, ‘карты’], [‘по’, MASK, ‘##ать’, ‘в’, ‘карты’] ...

- estimate the average probability of the *subtokens* to replace [MASK] token in the templates

  mean(lm_score("играть в карты"), lm_score('играть в детстве'), ...)

Submitted system #2 (the **501good** team)

Learning the similarity between the sketch and the context

- sketch tables were flattened into pseudo-sentences;
- The model was trained using the Sentence-BERT siamese similarity mechanism;
- two training pairs for each context in the dataset: one with matching sketch (label 1), second with random sketch (label 0);

Submitted system #3 (the **paleksandrova** team)

Going from sketch to context

For each sketch:

- Generate templates using all sketch content cells;

  “[MASK] нестерпимо”, “[MASK] от жажды” … ‘[MASK] unbearably’ ‘[MASK] from thirst’

- Obtain MLM hypotheses for each template;
- The most frequent candidate of all the MLM hypotheses is treated as the re-covered predicate;
- Map the sketch to the contexts with the matching target predicate.

For the sentences with no sketch found, the sketch with word2vec-closest predicate was used as an answer.

Results

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

The submitted systems were evaluated using the **accuracy** metric.

Results and discussion

- The task turns out to be rather difficult, unsupervised approaches leave enough room for different improvements.
- Two of three systems could improve its performance taking into account WSD problem.

Results and discussion

Possible directions of future investigation:

- evaluate the importance of circumstantial dependencies in the sketches;
- use semantic sketches as a basis for probing tasks for the pretrained language models;
- use semantic sketches as a basis for linguistically-motivated fine-tune tasks for the pretrained language models.

Further plans

- Quantitative and qualitative analysis of the sketches
- Integrate SemSketches into the GICR
- Work on parallel English-Russian sketches (some data can be already found in our github)

Competition:
https://competitions.codalab.org/competitions/29992

Github:
https://github.com/dialogue-evaluation/SemSketches

Thank you for your attention!