Matching semantic sketches to predicates in context using the BERT model

Aleksandrova Polina
NRU HSE
Moscow, Russia
paleksandrova37@gmail.com

Mokhova Anna
NRU HSE
Moscow, Russia
ann.mokhova@gmail.com

Nikolaenkova Maria
NRU HSE
Moscow, Russia
hublbublgog@gmail.com

Abstract

Modern language models have extensive information about the compatibility and meanings of various words. One of the ways to represent such lexical information, which is presented in the present study, is the construction of semantic sketches.

This paper presents a solution to the task of predicting a predicate from its most frequent actants and sirconstants using the application of the BERT neural network, which showed the best quality metrics in the Dialogue Evaluation SemSketches competition. The study analyzed several solutions approaching this task and ways to improve them based on the peculiarities of the architecture and the nature of data in terms of linguistics.

The results of testing the selected methods showed that the most successful tool for determining the semantic sketch of a predicate is the Conversational RuBERT model combined with the search for synonyms of the verbs sought in the training data.

Other promising ways to improve the quality of mapping the predicate to its semantic sketch include the use of contextualized embeddings to be able to take context into account, as well as fine-tuning of the models used.

Keywords: semantic sketches, lexical compatibility, language modeling.
DOI: 10.28995/2075-7182-2021-20-1-7

Соотнесение скетча с предикатом в контексте с использованием модели BERT

Александрова Полина
НИУ ВШЭ
Москва, Россия
paleksandrova37@gmail.com

Мохова Анна
НИУ ВШЭ
Москва, Россия
ann.mokhova@gmail.com

Николаенкова Мария
НИУ ВШЭ
Москва, Россия
hublbublgog@gmail.com

Аннотация

Современные языковые модели обладают широкой информацией о сочетаемости и значениях различных слов. Один из способов представлений таких лексических сведений, который представлен в настоящем исследовании, — конструирование семантических скетчей. В данной работе представлено решение задачи предсказания предиката по его наиболее частотным актантам и сирконстантам с помощью применения нейронной сети BERT, которое показало наилучшие метрики качества в рамках соревнования Dialogue Evaluation SemSketches. В ходе исследования было проанализировано несколько подходов, приближающих к решению этой задачи, а также способы их улучшения, основанные на особенностях архитектуры и природы данных с точки зрения лингвистики. Результаты тестирования выбранных методов показали, что наиболее успешным инструментом для определения семантического скетча предиката является модель Conversational RuBERT в сочетании с поиском синонимов искомых глаголов в тренировочных данных. К другим перспективным способам улучшения качества сопоставления предиката с его семантическим скетчем можно отнести использование контекстуализированных эмбеддингов для возможности учитывать контекст, а также дообучение (fine-tuning) используемых моделей. гвистики.

Ключевые слова: семантические скетчи, лексическая сочетаемость, языковое моделирование.

1 Introduction

The concept of semantic sketch, which will be used in this paper, can be defined as follows: a semantic sketch is a representation of all the actants and sirconstants of a predicate, which are distributed into classes according to their semantic roles. Another definition, more often used when working with corpus methods, is the idea of representation "in the form of statistics of combinability of the analyzed word with syntactically related lexical units [2].

An example of a semantic sketch of the verb ‘играть’ is shown in Table 1:

	Sphere Special	Time	Agent	Locative	ContrAgent
0	в карты	в детстве	дети	на бирже	с детьми
1	в шахматы	на большой перемене	мальчишки	во дворе	с мальчишками
2	в футбол	по вечерам	пацаны	на бильярде	с читателем
3	в азартные игры	каждый день	игроки внизу	на компьютере	с собакой
4	в игры	допоздна	ребята	на площадке	с сыном
5	в прятки	в молодости	команда	на чужом поле	с ребятами

Таблица 1: Example of semantic scetch for verb ’играть’

The separation of semantic word sketches is widely applicable for lexical analysis of linguistic units and corpus representation of linguistic data. This approach was first proposed and implemented by Adam Kilgarriff within the SketchEngine project [8].

Currently, various methods of applying semantic sketches as a particular type of corpus rendition are becoming a subject of research in computational linguistics. This paper will address the problem of correlating a semantic sketch with a predicate in context, which has been put forward in the SemSketches competition.

2 Task Description

The task of semantic sketches prediction was introduced in the Dialogue Evaluation SemSketches competition. Baseline solution is presented in the work [7]. The contexts of use (sentences) for the most frequent Russian predicates, as well as a set of anonymized semantic sketches, were chosen as the initial data. Several such predicates are: выйти (go out), сидеть (sit),действовать (act), подумать (think), написать (write), воспринимать (perceive), завершать (complete), принять (accept), встречаться (meet), продать (sell), говорить (talk). Anonymized sketches are such sketches, for each of which information about the essential roles and their fillers is provided, but the predicate itself is hidden. The task comes down to matching each context with one of the anonymized sketches. In other words, a set of contexts for different predicates is given, and the selected predicate in the context must be mapped to a sketch. And several different contexts can correspond to the same sketch. The data for the training sample consisted of 2000 sentences and 20 sketches. The results of the models were tested on a benchmark sample of 44750 sentences and 895 sketches, from which, in turn, 4347 sentences and 100 sketches, dev.gold and manual dev.gold, were manually selected, respectively. The Bidirectional Encoder Representations from Transformers (BERT) model [1] was used to solve the problems of this study. BERT is a neural network from Google, created in 2018, which showed by a large margin state-of-the-art results on a number of tasks. BERT can be used to create artificial intelligence programs to solve problems from various fields, in particular for natural language processing. RuBERT is a BERT model trained on the Russian-language part of Wikipedia and news data. Methods for adapting multilingual masked language models are presented in [5]. The use of this model has significantly improved the handling of Russian language data, and it will be used in some of the approaches described in the paper.

2.1 Baseline

This section will briefly present the solution of the organizers of the SemSketches competition, the results of which were used as a baseline. Determining the correspondences between a predicate and its semantic sketch was divided into several subtasks. First, syntactic parsing of sentences was applied to existing contexts, searching for the predicate and masking its direct dependents. In the next step, masked words were predicted using the RuBERT model. The resulting predicted sketches were compared with the reference variants, and the final accuracy quality metric was 0.1535.

3 Methods

3.1 Sketches

The solution [1] to the problem at hand is based on the idea of predicting a predicate for each sketch by generating templates from its data. The concept of masked language modeling, implemented in the BERT model used, is essential for template creation. During the neural network training, individual tokens were randomly masked in the input data, and the main task was to predict the token in place of the mask. This training procedure has a clear advantage in the described task over those models that learn to predict each next word based on the previous context because there is a possibility to predict a specific and any possible position in the sequence.

So, as a result of processing the sketch, the output is templates of two types:

1. MASK + role
   For each of the fillings of each role of the sketch the masked context was mapped to both left and right. The need to generate both [MASK] + role and role + [MASK] templates simultaneously is due to the fact that BERT, configured to process whole sentences, is very likely to predict punctuation marks in place of the mask in the right-hand position. In the final analysis, the predicted fillers on such patterns were treated separately.
   There is also a problem with predictions for templates with agentive roles and masks since in this case there are almost no verbs among the results: the model tends to construct sentences with nominal predicates. Therefore, it was decided to consider agentive roles separately.

2. Agent + [MASK] + role
   Each of the filler sets for roles was divided into two groups: fillers for agentic roles (these include ’Agent’, ’Agent Metaphoric’, ’Agent Route’) and the rest. In the absence of agentic roles, we limited ourselves to the pronouns ’he’ and ’they’.
   Thus, each of the role fillers other than agentic roles was matched with templates of the form Agent + [MASK] + role, where Agent is all the possible agentic fillers described above. The number of templates is not fixed - it depends on how many of the most frequent role fills are present in a particular sketch.

An example of several templates for the verb to ‘собираться’:
[MASK] в стаи
в стаи [MASK]

[1] The code is available on https://github.com/psaleksandrova/Matching-semantic-sketches-to-predicates-in-context-using-the-BERT-model

[MASK] по выходным
регулярно [MASK]
друзья [MASK] в последний раз
друзья [MASK] у кого-нибудь дома
публика [MASK] в 9 часов
военный совет [MASK] впервые
вся семья [MASK] в дорогу

3.2 RuBERT

After the templates were generated for the sketch, for each of them, a list of placeholders was predicted using RuBERT, which was also used in the organizers' solution.

After predicting the fillers in place of the mask, only verbs were selected from the resulting lists. The results obtained, as well as the predicates from the sentences, were lemmatized to find matches. Morphological analysis and lemmatization were implemented using the pymorphy2 library [4].

On the one hand, this is a necessary step to find identical verbs in predicates and sentences; on the other hand, an inaccurate definition of the initial form could disrupt the matching process (the initial form шило for the past tense verb masculine шить (sting)). Thus, each sketch was matched with a list of predicates predicted for each of the patterns. Next, the resulting list was ranked by frequency of predictions, and the most frequent predicate was selected as the final single predicate for the sketch, which was then matched to the sentence with the corresponding predicate. If there were several such sentences, the very first of them was chosen without any analysis, which entailed an unresolved polysemy, since the choice of the sketch sentence was determined by the verb alone and not by its context.

An example of a cross-section of the frequency-ranked list of predictions of the verb sketch 'собираться' in the Table 2.

быть	0.42058823529411765
прийти	0.21764705882352942
собрать	0.18529411764705883
ходить	0.17352941176470588
собираться	0.1470588235294117
ждать	0.1323529411764706
собраться	0.0941176470588235
играть	0.07941176470588235
войти	0.06764705882352941
ехать	0.06764705882352941

Таблица 2: Example of the frequency-ranked list of predictions of the verb sketch 'собирать'

Accordingly, if there is a context for the predicate 'быть' in the set of 2000 sentences, then that is what will be predicted for the given sketch.

3.3 Conversational RuBERT

As an improvement to the method proposed above, it was decided to use Conversational RuBERT. It is assumed that the model, which was trained on the texts of subtitles, blogs, social networks, etc. should better process various stable word combinations and, in general, better summarize the features of Internet data, which were just used to highlight the semantic sketches proposed in the competition. It is possible that the Wikipedia and news texts on which the standard RuBERT was trained do not in principle contain, or contain in small amounts insufficient for the model to "remember"some of the roles from the sketches or predicates in mind.

This approach also solves the problem of cases where the predicate predicted for the sketch was not found in any of the sentences. In the first iteration, the randomized selection was used in such cases as the answer, but the new method solved the problem by searching for verbs synonymous with the predicted predicate. A cosine distance comparison of vector word2vec representations was used to search by synonyms [3]. A static model from the RusVectores resource [6], trained on the texts of the National Corpus of the Russian language (NRU)[2], was used. Synonyms are those words of a similar part of speech whose embeddings are the least different (cosine distance is minimal) from the vector of the original verb.

4 Analysis

The work resulted in a best-effort error analysis, namely the reference predicate - predicted predicate pairs for each sketch were examined. In addition to explicit differences between verbs, which are challenging to explain linguistically, several explicit groups were identified for well-interpreted mismatches, which, with proper correction, can result in a correct prediction.

Among these discrepancies, there are cases where the verbs in the reference data and the prediction differ only in the feature of species, the presence of the category of return. Quite a few pairs with the same root morpheme and similar meaning but different prefixes. And also exciting cases are pairs of synonyms, both of which obviously fit the roles of a particular sketch, as well as antonyms. In fact, some roles may contain information about the sign of meaning itself (for example, the fact of winning in the example lose-win) but do not reflect its presence or absence.

In the table below, for the highlighted groups of inconsistencies, examples are given from the results of predictions compared to the reference predicate.

Characteristics of a mismatch	Example (benchmark - prediction
aspectual pairs	выходить - выйти останавливаться - остановиться вздохнуть - вздыхать бросить - бросать
reflexivity	менять-меняться катить - катиться завершить - окончиться схватить - схватиться
prefixal verbs	пожать-нажать просидеть-сидеть подействовать - действовать исполнить - выполнить
synonyms	вложить - вкладывать расстреливать-убивать отметить - обнаружить вскакивать - встать падать - упасть
antonyms	проиграть - выиграть прощаться - встретиться

Таблица 3: Example of semantic scetch for verb 'играть'

5 Discussion

As an improvement of the two main methods, some ideas were proposed which theoretically should have resulted in a quality gain, but unfortunately their implementation was not possible.

[2] https://ruscorpora.ru/new/

As noted earlier, when matching the predicted predicate for a sketch with the corresponding sentences, the context was not taken into account in any way, which gave rise to false matches given the polysemy of the predicate. As a solution to this problem, the idea of using embeddings that use information about context begs to be solved. Contextualized embeddings of each token or the whole sentence can be extracted from the BERT model: they contain information about the entire sequence. Usually, the latent states from the last layers of the neural network are used as embeddings. Accordingly, it is possible to present as a contextualized vector both verbs, potentially suitable for the mask places in the templates, and predicates in all sentences. And select the most appropriate predicate for the sketch by ranking by cosine closeness between vector representations. The reason for the inability to implement the method is the limited memory in the Google Colab service, which allowed to predict only 500 predicates out of 44750 available.

Fine-tuning of the model on specific data gives good results on a number of NLP problems. The volume of available sentences should be sufficient for the necessary tuning of weights. The assumption is that if we mask in contexts only the particular predicate in question, and BERT predicts it in the process of retraining, then in pattern prediction, more attention will be paid to verbs and from the required finite set. It would then be possible to implement one of the previously proposed methods on the pre-trained BERT model. The limited amount of computational resources is the factor that prevented the implementation of this concept.

6 Conclusion

As a result of the present work we have studied the nature of semantic sketches and possible approaches to predicate prediction based on its possible actant and syrconstant fillers. Table 4 presents the results of the evaluation experiments with both methods that were implemented in our study.

Baseline	0.154
RuBERT	0.212
Conversational RuBERT	0.309

Таблица 4: Accuracy scores

We have carried out appropriate experiments on predicate prediction based on its semantic sketch and analyzed the results in terms of both the approach itself and the semantic nature of the data.

In the future, we plan to improve the implemented approach with an adjustment based on the semantic analysis of the current results; and, provided the computational resource problem is solved, to try to implement other methods we have proposed.

References

[1] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[2] Detkova J. Novitskiy V. Petrova M. Selegey V. Differential semantic skethes for Russian Internet-corpora // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2020. — Vol. 17. — P. 20.

[3] Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // arXiv preprint arXiv:1310.4546. — 2013.

[4] Korobov Mikhail. Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2015. — P. 320–332.

[5] Kuratov Yuri, Arkhipov Mikhail. Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213. — 2019.

[6] Kutuzov Andrey, Kuzmenko Elizaveta. WebVectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2016. — P. 155–161.

[7] SemSketches-2021: experimenting with the machine processing of the pilot semantic sketches corpus / Maria Ponomareva, Maria Petrova, Julia Detkova et al. // Proc Dialogue, Russian International Conference on Computational Linguistics. — Moscow, 2021.

[8] The Sketch Engine: ten years on / Adam Kilgarriff, Vít Baisa, Jan Bušta et al. // Lexicography. — 2014. — Vol. 1, no. 1. — P. 7–36.