Data sets for training AI-based prior art search

Thank you chair! Good day, dear colleagues!

Next slide, Please!

Slide 2

Rospatent is step by step expanding the use of AI. I'm going to tell you about our work on creating datasets, which are necessary for training AI searching for state of the art during the examination of inventions.

Our approach to creating datasets is based on two ideas; you can see them on the slide:

First thing. The most important task of examining the application for an invention is to search for many documents characterizing the state of the art for a given application. Let's call this set of documents the Semantic Cluster of Patent Documents. The semantic cluster of patent documents is the set of all documents defining the state of the art for a given application, including all documents of the patent families.

AND Second thing. The unique property of information about inventions is that each publication of information about the grant of a patent contains a list of documents that, in the opinion of the examination, characterize the state of the art for the given application (that is, it contains a Semantic Cluster of Patent Documents). It is in accordance with the WIPO standard. So, if in total at least 2 hundred million patent documents have been published, including applications, patents, and other publications, then, assuming that from a quarter to a third of them are patents, we can expect that approximately 40 to 70 million of them are, technically speaking, **marked** by examination. We can calculate more precisely, but in any case we are dealing with tens of millions of documents. This is a plenty amount of data for AI training!

Next slide, please!

Slide 3

Here you see more formal definition of the semantic cluster, on the left.

Well-known idea is The good and voluminous data are crucial for success in AI training. We have made a dataset generator based on the new concept of semantic clusters of patent documents. We formed new datasets with this generator. With

these datasets we have reached fairly good results of AI training for the prior art search.

The generator features a special API that allows to prepare and receive datasets for a wide range of conditions.

Next slide, please!

Slide 4

In conclusion, here are tools available today:

- Dataset of semantic clusters of Russian-language patent documents (1.5 million patent clusters);

- Generator of datasets of semantic clusters of English-language patent documents (8.5 million patent clusters).

More details are provided in the authors' publications, here on the slide.

That all for today, next slide, please!

Thank you for attention!