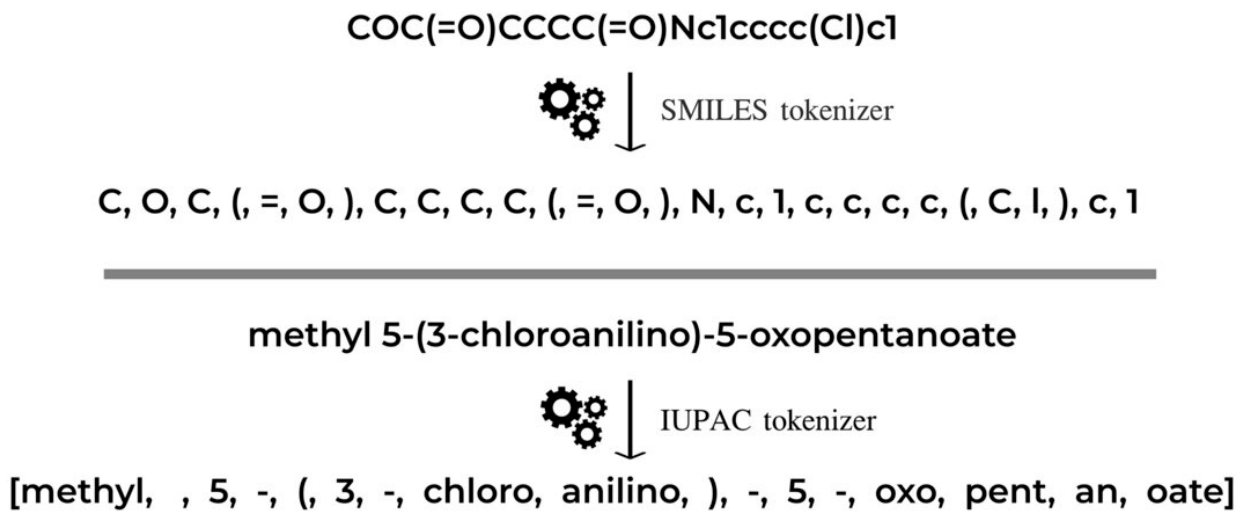
Neural network trained to properly name organic molecules

July 28 2021



Demonstration of SMILES tokenization (top) and IUPAC names tokenization (bottom). Credit: 10.1038/s41598-021-94082-y

Skoltech researchers and their colleagues from Lomonosov Moscow State University and the Syntelly start-up have developed and trained a neural network to generate names for organic compounds in accordance with the IUPAC nomenclature system. Their research published in the *Scientific Reports* shows that modern neural networks are able to efficiently deal with exact algorithmic problems.

Chemistry uses the nomenclature system of IUPAC, the International

Union of Pure and Applied Chemistry, as a generally accepted language for giving names to organic compounds. For example, in the IUPAC terms, sucrose is called (2R,3R,4S,5S,6R)-2-[(2S,3S,4S, 5R)-3,4-dihydroxy-2,5-bis(hydroxymethyl)oxolan-2-yl]oxy-6-(hydroxymethyl)oxane-3,4,5-triol, and paracetamol, the active ingredient of antipyretic drugs like Tylenol, is N-(4-hydroxyphenyl)acetamide.

Since the IUPAC name is a full representation of a compound's structure, complex molecules tend to have long and tedious names. Omitting even a single digit or symbol is unacceptable, so chemists have to pay attention to what they write down and have deep knowledge of IUPAC's numerous rules. Off-the-shelf software tools that generate IUPAC names are widely available on the market but open-source software is not.

"Initially, we wanted to create an IUPAC name generator for Syntelly, our AI chemistry platform. Soon we realized that it would take us more than a year to create an algorithm by digitizing the IUPAC rules, so we decided instead to leverage our experience in neural network solutions," says Skoltech research scientist Sergey Sosnin, lead author of the study and co-founder of the Syntelly startup.

The team used Transformer architecture, one of the most powerful machine translation neural networks initially designed by Google, as the basis for their research and trained it to convert a molecule's structural representation to a IUPAC name and vice versa.

The new network was trained and tested using PubChem, the world's largest open chemical database of over 100 million compounds. Designed in a matter of six weeks, the network learned to do the conversion with nearly the same accuracy (about 99%) as rule-based algorithmic solutions.

In addition, the study showed that neural networks can solve algorithmic problems fairly accurately. "Telling a cat from a dog in a picture is an equally easy task for humans and neural networks, while there is no way to make an efficient purely algorithmic solution. At the same time, multiplying multi-digit numbers is hard for humans but easy for a primitive calculator that instantly produces an absolutely accurate result. Both this task and IUPAC name generation are examples of purely algorithmic problems," Sosnin explains.

"We have shown that neural networks can cope with exact problems, disproving the formerly prevalent notion that they should not be used for this kind of problem. Replacing a word with a synonym is quite possible in machine translation, whereas in our task, a single wrong symbol results in an incorrect molecule. Yet, Transformer successfully copes with this task," Sosnin adds.

The new solution has been implemented in the Syntelly platform and is available online. The researchers hope that their method can be used for conversion between chemical notations and for other technical notation-related tasks, such as generation of mathematical formulas or translation of software programs.

More information: Lev Krasnov et al, Transformer-based artificial neural networks for the conversion between chemical notations, *Scientific Reports* (2021). DOI: 10.1038/s41598-021-94082-y

Provided by Skolkovo Institute of Science and Technology