UDK 004.932.72

Parsing system for chemical reactions

Luzanova A.M. (ITMO University, Infochemistry Scientific Center),
Tonkii I.S. (ITMO University, Infochemistry Scientific Center)
Golovinsky R.P. (ITMO University, Infochemistry Scientific Center)
Scientific supervisor - Skorb E.V. PhD (ITMO University, Infochemistry Scientific Center)

Up until recently, most scientific and patent papers on chemistry described molecular structures by using either systematic names or graphical depictions of Kekule structures. Despite the huge number of developments in the recognition of molecules and chemical reactions, the problem of recording and deciphering chemical data in a computer-friendly format continues to be unresolved. The thing is - graphical images can't be directly interpreted by a computer. Currently, there are two types of programs by which molecules are recognized. The first type of mechanisms is rule-based. It consists of process of translating molecules graphical images into a machine-readable format called Optical Chemical Structure Recognition (OCSR) [1]. The method is based on vectorizing images and interpreting vectors and nodes as bonds and atoms. Molvec [2], OSRA [3] are some of these. It is relatively less flexible, but has the significant advantage - reliability. The second type, for instance Img2mol [4], is driven by neural network training using datasets of molecules. This field is the most interesting and promising at present. The development of neural networks is the subject of a broad scientific discourse, so its implementation in new types of systems is more relevant than ever.

Molvec is the program which, at the current stage of development, has proven to be the most suitable and easy to implement. It was improved and integrated into the parsing system. The article that must be processed is loaded into the parser. Reaction detection runs and the program removes unnecessary text, tables and pictures. The parsing of the reaction proceeds in several steps: at first, the scanner reads information about the structure of the reagents and products. Next, it determines the reaction conditions: catalyst, temperature, concentrations, solvent, etc. After refactoring, the data are distributed to the database columns and thus filled in automatically.

An open-source MVP of the parser, which is already able to recognize the structure of molecules and determine what is a reagent and what is a product, was developed. At the stage of development, problems with recognition of some radicals and identification of images with poor resolution appeared. These problems will be solved by writing tests and improving the readout algorithms. In the future the possibility of implementing neural networks for recognition is being considered. Training of existing mechanisms on large and complex datasets is required in order to achieve maximum accuracy and high parser efficiency.

Literature:

1. Kohulan Rajan, Henning Otto Brinkhaus, Achim Zielesny, Christoph Steinbeck A review of optical chemical structure recognition tools // J Cheminform 2020; 12(60);

2. Peryea T, Katzel D, Zhao T, Southall N, Nguyen D-T Merge pull request / Peryea T, Katzel D, Zhao T, Southall N, Nguyen D-T [Electronic resource] // GitHub: [website]. - URL: https://github.com/ncats/molvec?ysclid=le7d2l5m9l253888001;

3. Igor V. Filippov, Marc C. Nicklaus Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution // J. Chem. Inf. Model. 2009, 49(3): 740–743;
4. Djork-Arne Clevert, Tuan Le, Robin Winter, Floriane Montanari Img2Mol – accurate SMILES recognition from molecular graphical depictions // Chem. Sci. 2021, 12:14174-14181;