Name-to-Structure Converters One Is Not Enough!

Plamen Petrov, R&D Information, AstraZeneca R&D, Mölndal, Sweden

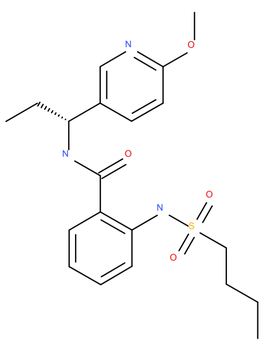
Introduction

Name-to-Structure (N2S) converters are commonly used for automatic extraction of chemical structures from text like patents, journal articles, documents in corporate repositories. While in most cases (for example patents) extracting only some of the structures is ok, there are applications where the user expects a high level of recovery and accuracy of the output. In this work we compare different N2S converters and the effect of more advanced text-mining tools.

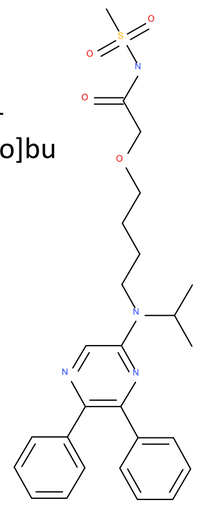
N2S converters comparison

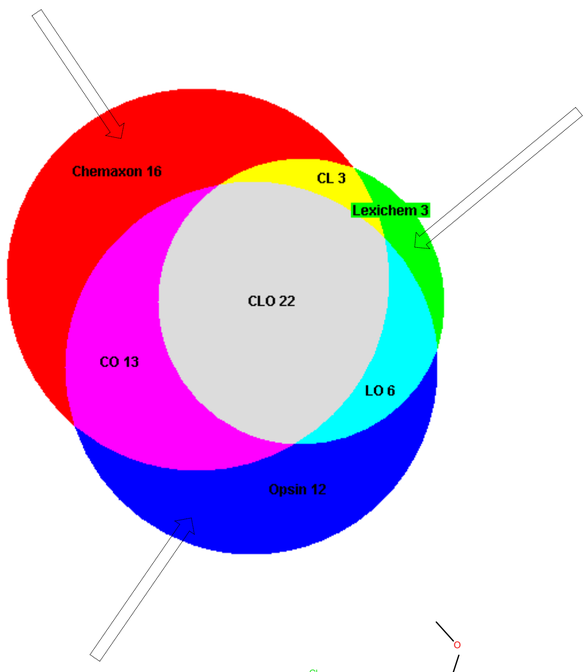
We processed the FDA orphan drug product database containing 2700 chemical names. The goal was to compare the ability of the N2S tools to parse IUPAC-like names, so we deleted all names which could be resolved by dictionary lookups. The names were passed to the Chemaxon's N2S, Lexichem by OpenEye and the open source OPSIN (the latest versions as of April 2013). Total of 240 chemical names were processed.

2-(Butane-1-sulfonyl-amino)-N-[1-(R)-(6-methoxypyridin-3-yl)-propyl]-benz-amide



2-{4-[(5,6-diphenylpyrazin-2-yl)(isopropyl)amino]butoxy}-N-(methylsulfonyl) acetamide





(R)-N-[2-(6-Chloromethoxy-1H-indol-3yl)propyl]acetamide

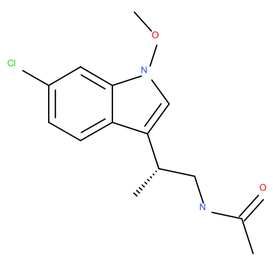
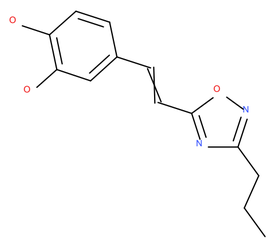


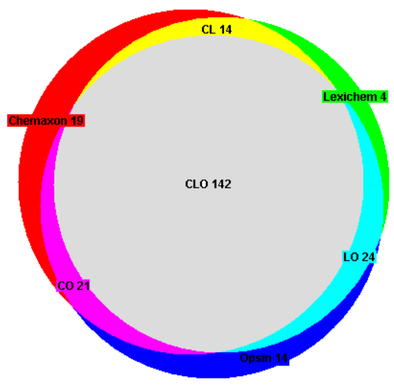
Fig.1 Venn diagram showing the number of structures recognized by each of the tools, the unique structures with respect to the other two and the structures-in-common. C is for Chemaxon, L – Lexichem, O – OPSIN. The example structures demonstrate names recognized by only one of the tools.

Effect of error correction tools

Chemical documents often contain typos or OCR errors in case of scanned PDFs. There are highly specialized tools, like NextMove's LeadMine or Chemaxon's Document-to-Structure which apply error correction techniques and generally improve the N2S success.

4-[2-(3-Propyl-[1,2,4]0xadiazol-5-yl)-vinyl]-benzene-1,2-diol





(2Z)-2-cyano-3-3hydroxy-N-[4-(trifluoromethly)phenyl]-2-hepten-6-ynamide

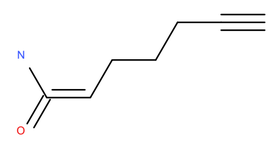


Fig.2 Effect of data correction on the N2S. About 50 more structures are recognized correctly, but 120 are false positives due to breaking the text and extracting a sub-structure.

The workflow

Using different tools from different vendors is not the most user-friendly approach, so we went on developing a workflow in Pipeline Pilot which can be published as a web application. All N2S tools we license in AstraZeneca were wrapped as restful web services in our ChemistryConnect data warehouse and subsequently exposed as PipelinePilot components.

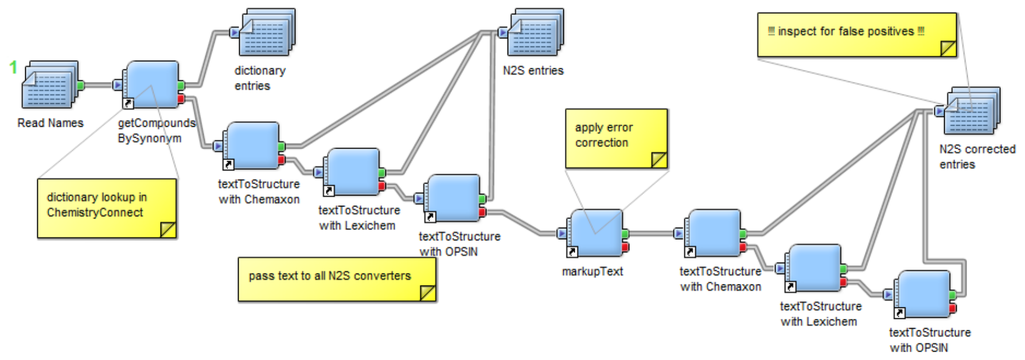


Fig.3 PipelinePilot protocol for processing chemical names in structured documents.

Conclusions

- **N2S tools are complimentary** to each other. Combining several increases the recognition ration
- **Error correction tools** help recover more structures, but introduce large number of false positives which mandates manual inspection of the output. Any warnings thrown by the tools should be taking into account when automatically processing large data sets.