Digging for data with Chemlist and ChemSpider

March 22 2010

Just like the rest of us, scientists today are swamped with information. As more chemical resources become freely available, text mining applications - previously focused on correctly identifying gene and protein names - are now shifting towards also correctly identifying chemical names. Now database experts have compared two chemical name dictionaries head to head, and report on the payoffs of manual versus automatic data curation in the open access publication, *Journal of Cheminformatics*.

Chemlist's creators wanted to investigate the effect extensive manual curation of a multi-source chemical dictionary would have on chemical term identification in text. Kristina Hettne and her team based in the Netherlands, together with US-based colleagues, compared Chemlist, a dictionary for identifying small molecules and drugs in text automatically generated from a number of publicly available databases, with a second dictionary extracted from the ChemSpider database which has been curated manually to establish valid chemical name to structure relationships. To compare automatic curation with manual curation, the authors used only the ChemSpider component containing manually curated names and synonyms in their research.

The researchers tested the dictionary from ChemSpider on an annotated corpus and compared the results with those for the Chemlist dictionary. The ChemSpider dictionary of some 80,000 names was less than a third of the size of Chemlist at around 300,000. The ChemSpider dictionary had a precision of 0.43 and recall of 0.19 before filtering and

disambiguation, with results of 0.87 and 0.19 after filtering and disambiguation. Meanwhile the Chemlist dictionary scored 0.20 for precision and 0.47 for recall before filtering and disambiguation, and 0.67 and 0.40 for these two measures afterwards.

This means that although ChemSpider achieved the best precision, the Chemlist dictionary had a higher recall and the best F-score, a function of a test's accuracy incorporating both precision and recall. "Rule-based filtering and disambiguation is necessary to achieve high precision for both automatically generated and the manually curated dictionaries," Hettne concludes. Antony Williams, project lead for ChemSpider comments "Such validated name-structure dictionaries studied in this work provide a strong foundation for semantic markup technologies, interlinking and various online resources." Both ChemSpider and the chemical databases included in Chemlist continue to grow at high speed, and further investigation is needed to see how this growth affects the performance of the dictionaries.

More information:
ChemSpider is available at www.chemspider.com/ and the Chemlist dictionary is freely available as an XML file in Simple Knowledge Organization System format on the web at www.biosemantics.org/chemlist

Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining, Kristina M Hettne, Antony J Williams, Erik M van Mulligen, Jos Kleinjans, Valery Tkachenko and Jan A Kors, Journal of Cheminformatics (in press), www.jcheminf.com/

Provided by BioMed Central