Journal of Cheminformatics

POSTER PRESENTATION Open Access

Comparing manual and automated extraction of chemical entities from documents

Christian Tyrchan*, Sorel Muresan

From 5th German Conference on Cheminformatics: 23. CIC-Workshop
Goslar, Germany. 8-10 November 2009

The chemical information landscape is changing rapidly with a yearly increase of over 1 million new compounds and more than 700,000 publications related to chemistry [1]. Exploring the chemical space covered by relevant journals and patents is a crucial step in early stage medicinal chemistry projects. Extracting chemical entities from unstructured text is a complex task and different approaches are currently used including manual extraction by expert curators, text mining supported by chemical NER or combinations thereof [2]. The chemical information and corresponding annotations are subsequently stored in relational databases allowing for complex chemical and text queries.

To assess the capability of chemical NER in documents and to understand the coverage and accuracy of the underlying data we compared the chemistry extracted by manual curation (GVKBIO) and text mining (SureChem) from a small patent corpus.

- GVKBIO databases are populated with explicit relationships between compounds, assays and sequence identifiers that have been manually extracted from journals and patents on a large scale [3].
- SureChem Portal [4] is a gateway for chemical patent search on full text collections for USPTO, EPO and WO. SureChem users can perform structure and keyword searches on more than 9 million unique compounds.

We have selected a set of 250 patents covering various target classes and for which a minimum of 25 records per patents were retrieved from GVKBIO Patent database. The analysis was done using PipelinePilot protocols [5].

These initial results demonstrate the benefits and challenges of text mining for chemical information extraction from unstructured text.

Published: 4 May 2010

References

1. Engel T: *J Chem Inf Model* 2006, **46**:2267-2277.
2. Banville DL, (Ed.): *Chemical Information Mining: Facilitating Literature-Based Discovery* CRC Press: Boca Raton, London New York 2009.
3. [http://www.gvkbio.com/informatics.html].
4. [http://www.surechem.org].
5. [http://accelrys.com/products/scitegic].

doi:10.1186/1758-2946-2-S1-P7
Cite this article as: Tyrchan and Muresan: **Comparing manual and automated extraction of chemical entities from documents.** *Journal of Cheminformatics* 2010 **2**(Suppl 1):P7.

AstraZeneca R&D, LG CVGI, Pepparedsleden 1, S-43183 Mölndal, Sweden