Sayle *et al. Journal of Cheminformatics* 2011, **3**(Suppl 1):O16
http://www.jcheminf.com/content/3/S1/O16

Journal of Cheminformatics

ORAL PRESENTATION **Open Access**

Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction

Roger A Sayle[1*], Plamen Petrov[2], Jon Winter[3], Sorel Muresan[2]

From 6th German Conference on Chemoinformatics, GCC 2010
Goslar, Germany. 7-9 November 2010

The text mining of patents and patent applications for chemical structures of interest to medicinal chemists poses a number of unique challenges not encountered in other fields of text analytics. Traditional text mining relies on the co-occurrence of common terms between documents to provide similarity measures that can be used to cluster and rank related documents. The more words shared between two documents, the more similar they are, and the greater the probability that they discuss the same topic. By contrast, in pharmaceutical "composition of matter" patents the novel and unique chemical entities are far more significant than those that can be found elsewhere. Although the text of a pharmaceutical patent may explicitly name thousands of individual compounds, and via generic Markush structures claim an infinite number, the role of these patents is to protect the intellectual property of only one or perhaps two drug candidates.

In this work, we present an analysis of the "quality not quantity" of structures extracted by automatic Chemical Named Entity Recognition (CNER) methods both on a small hand-curated benchmark set [1] and a large-scale analysis of a comprehensive database of 12 million patents [2,3]. Our results show the limited value of traditional lexicon/dictionary based approaches in extracting "key" compounds and that the major impediment is not the performance of the name-to-structure software used, but the high rate of OCR errors, typos and lexicographic problems found in patent-office data feeds. To address this problem, novel algorithms for automatic chemical spelling correction have been developed, that take advantage of the grammar used in IUPAC-like nomenclature. This forms a preprocessing pass, independent of the name-to-structure software used, and is shown to greatly improve results in our study.

Author details
[1]NextMove Software, Santa Fe, New Mexico, 87501, USA. [2]DECS, AstraZeneca, Molndal, Sweden. [3]CIRA, AstraZeneca, Alderley Park, Cheshire, UK.

Published: 19 April 2011

References

1. Hattori K, Wakabayashi H, Tamaki K: **Predicting Key Example Compounds in Competitor's Patent Applications using Structural Information Alone.** *J Chem Inf Model.* 2008, **48**:135-142.
2. Rhodes J, Boyer S, Kreulen J, Chen Y, Ordonez P: **Mining Patents using Molecular Similarity Search.** *Pac Symp Biocomput.* 2007, **12**:304-315.
3. Suriyawongkul I, Southan C, Muresan S: **The Cinderella of Biological Data Integration: Addressing the Challenges of Entity and Relationship Mining from Patent Sources.** In *Data Integration in the Life Sciences. Lect Notes Bioinf. Volume 6254.* Springer; 2010.
4. Sayle R: **Foreign Language Translation of Chemical Nomenclature by Computer.** *J Chem Inf Model.* 2009, **49**:519-530.

doi:10.1186/1758-2946-3-S1-O16
Cite this article as: Sayle *et al.*: Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction. *Journal of Cheminformatics* 2011 **3**(Suppl 1):O16.

* Correspondence: roger@nextmovesoftware.com
[1]NextMove Software, Santa Fe, New Mexico, 87501, USA
Full list of author information is available at the end of the article