Title: Text mining patents for biomedical knowledge

Author: Raul Rodriguez-Esteban Markus Bundschus

PII: S1359-6446(16)30152-0
DOI: http://dx.doi.org/doi:10.1016/j.drudis.2016.05.002
Reference: DRUDIS 1808

To appear in:

Received date: 12-1-2016
Revised date: 7-4-2016
Accepted date: 4-5-2016

Please cite this article as: Rodriguez-Esteban, R., Bundschus, M.,Text mining patents for biomedical knowledge, *Drug Discovery Today* (2016), http://dx.doi.org/10.1016/j.drudis.2016.05.002

Text mining patents for biomedical knowledge

Raul Rodriguez-Esteban[1] and Markus Bundschus[2]

[1]Roche Pharmaceutical Research and Early Development, pRED Informatics, Roche Innovation Center Basel, 4070 Basel, Switzerland
[2]Scientific & Business Information Services, Roche Diagnostics GmbH, 82377 Penzberg, Germany
Corresponding author: Rodriguez-Esteban, R. (raul.rodriguez-esteban@roche.com)

Keywords: text mining; patents; patent mining; natural language processing; named-entity recognition; information retrieval.

Teaser: We review research in the area of text mining of biomedical patents and highlight the associated technical challenges that emerge from the automatic extraction of patent information.

Biomedical text mining of scientific knowledge bases, such as Medline, has received much attention in recent years. Given that text mining is able to automatically extract biomedical facts that revolve around entities such as genes, proteins, and drugs, from unstructured text sources, it is seen as a major enabler to foster biomedical research and drug discovery. In contrast to the biomedical literature, research into the mining of biomedical patents has not reached the same level of maturity. Here, we review existing work and highlight the associated technical challenges that emerge from automatically extracting facts from patents. We conclude by outlining potential future directions in this domain that could help drive biomedical research and drug discovery.

Introduction

Data availability has traditionally had a key role in scientific progress. Although patents are, by their nature, public documents, large-scale free electronic access to patents has only been made possible recently. Notably, since 2010, the US Patent and Trademark Office (USPTO) has offered its patents online through a bulk download service (https://bulkdata.uspto.gov). This release of more than ten terabytes of data could be considered comparable to the launch of PubMed in 1996 [1], which made the majority of Medline content available online and stimulated the field of biomedical text mining more than increases in computational power or advances in natural language processing (NLP). The increased availability of patents has represented an opportunity for researchers to create and improve algorithms specialized in extracting the knowledge contained therein. In this review, we describe advances in the mining of biomedical patents, which are patents with application in areas such as medicinal chemistry, medical diagnostics, biological assays, large-molecule therapies, gene sequencing, bioinformatics, and other areas related to medical and pharmaceutical research and development. A rough estimate of the number of such patents is over 11.6 million, of which 8.7 million have been issued by the ten most prolific patent jurisdictions (as of January 11, 2016*) (Figure 1). Henceforth, we refer to this particular set of patents simply as 'patents'.

Here, we highlight the differences between patents and scientific articles that bear an influence on text mining strategies. We then review the state-of-the-art in the text mining of patents and conclude with potential future research directions.

Differences between scientific articles and patents

Differences in writing style and structure

A large portion of the research in biomedical text mining has focused on scientific abstracts, particularly on Medline abstracts. However, there is a substantial number of differences between scientific abstracts and patents that make mining patents a specialized challenge. Full-text articles are more similar to patents but there are still differences between the two that potentially influence text-mining strategies, such as: (i) Patents tend to be lengthier than full-text articles and have a different document structure. For example, patents are often divided into abstract, claims, description, and examples [2], whereas scientific articles are typically parceled into introduction, methods, results, and discussion (IMRaD) [3]. Differences in length and structure between scientific abstracts and full-text articles have been shown to be important for text-mining performance and, therefore, are potentially important when applying text-mining algorithms to patents [4]; (ii) the writing style and discourse in patents is different from that of scientific articles. For example, patent claims are often made of complex, long sentences with multiple clauses and dependencies to other claims [5]. The density of used terms also differs between patents and articles [6]. This, for example, can have an influence on sentence splitters and tokenizers; (iii) patent inventors might use nonstandard vocabulary that is not identified by text-mining approaches that work in the scientific literature; (iv) as discussed by [7], patent inventors try to increase the coverage of their claims by generalizing, hedging, or making 'prophetic' claims; (v) repetition of legal 'boilerplate' is common in patents. Furthermore, the content of patents can be redundant and patents might be grouped into families of equal or similar content; (vi) patents do not undergo scientific peer review and do not face the same level of scientific scrutiny. Thus, patents cite scientific articles but scientific articles rarely cite patents [8], reflecting the fact that academic scientists are not aware of what is being published in patents; (vii) an abundant number of patents are