Improving Literature-Based Discovery with Advanced Text Mining

Anna Korhonen[1], Yufan Guo[1,2], Simon Baker[1], Meliha Yetisgen-Yildiz[2], Ulla Stenius[3], Masashi Narita[4], and Pietro Lio'[1]

[1] Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
alk23,yg244,sb895,pl219@cam.ac.uk
[2] Biomedical and Health Informatics, School of Medicine, University of Washington
Box 358047 Seattle, WA 98109, USA
melihay@uw.edu
[3] Institute of Environmental Medicine, Karolinska Institute
SE-171 77 Stockholm, Sweden
Ulla.Stenius@ki.se
[4] Cancer Research UK Cambridge Institute, University of Cambridge
Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
masashi.narita@cruk.cam.ac.uk

Abstract. Automated Literature Based Discovery (LBD) generates new knowledge by combining what is already known in literature. Facilitating large-scale hypothesis testing and generation from huge collections of literature, LBD could significantly support research in biomedical sciences. However, the uptake of LBD by the scientific community has been limited. One of the key reasons for this is the limited nature of existing LBD methodology. Based on fairly shallow methods, current LBD captures only some of the information available in literature. We discuss how advanced Text Mining based on Information retrieval, Natural Language Processing and data mining could open the doors to much deeper, wider coverage and dynamic LBD better capable of evolving with science, in particular when combined with sophisticated, state-of-the-art knowledge discovery techniques.

1 Scientific Background

The volume of scientific literature has grown dramatically over the past decades, particularly in rapidly developing areas such as biomedicine. PubMed (the US National Library of Medicine's literature service) provides now access to more than 24M citations, adding thousands of records daily[5]. It is now impossible for scientists working in biomedical fields to read all the literature relevant to their field, let alone relevant adjacent fields. Critical hypothesis generating evidence is often discovered long after it was first published, leading to wasted research time and resources [20]. This hinders the progress on solving fundamental problems such as understanding the mechanisms

[5] PubMed: http://www.ncbi.nlm.nih.gov/pubmed

underlying diseases and developing the means for their effective treatment and prevention.

Automated Literature Based Discovery (LBD) aims to address this problem. It generates new knowledge by combining what is already known in literature. It has been used to identify new connections between e.g. genes, drugs and diseases and it has resulted in new scientific discoveries, e.g. identification of candidate genes and treatments for illnesses [6, 21].

Facilitating large-scale hypothesis testing and generation from huge collections of literature, LBD could significantly support scientific research [15]. However, based on fairly shallow techniques (e.g. dictionary matching) current LBD captures only some of the information available in literature. Enabling automatic analysis and understanding of biomedical texts via techniques such as Natural Language Processing (NLP), advanced Text Mining (TM) could open the doors to much deeper, wider coverage and dynamic LBD better capable of evolving with science. The last decade has seen massive application of such methodology to biomedicine and has produced tools for supporting important tasks such as literature curation and the development of semantic data-bases [20, 19]. Although advanced TM could similarly support LBD, little work exists in this area, e.g. [25].

In this paper we discuss the state of the art of LBD and the benefits of an approach based on advanced TM. We describe how such an approach could greatly improve the capacity of LBD, in particular when combined with sophisticated knowledge discovery techniques. We illustrate our discussion by highlighting the potential benefit in the literature-intensive area of cancer biology. Since LBD is of wide interest and its potential applications are numerous, improved LBD could, in the future, support scientific discovery in a manner similar to widely employed retrieval and sequencing tools.

2 Materials and Methods

2.1 Literature-Based Discovery: The State of the Art

Literature-based discovery was pioneered by Swanson [22] who hypothesised that the combination of two separately published premises “A causes B“ and “B causes C“ indicates a relationship between A and C. He discovered fish oil as treatment for Raynaud’s syndrome based on their shared connections to blood viscosity in literature. Since then, considerable follow-up research has been conducted on LBD (see [6] for a recent review).

LBD has been used for both closed and open discovery. Closed discovery (i.e. hypothesis testing) assumes a potential relationship between concepts A and C and searches for intermediate concepts B that can bridge the gap between A and C and support the hypothesis. It can help and find an explanation for a relationship between two concepts. Open discovery (i.e. hypothesis generation), in contrast, takes as input concept A and aims to identify a set of concepts C that are likely to be linked to A via an intermediate concept B. It can, for example, be used to find new treatments for a given disease or new applications for an existing drug.

The first step of LBD is to identify the concepts of interest (e.g. genes, diseases, drugs) in literature. Most current systems use dictionary-based matching for this. The

MetaMap tool (http://metamap.nlm.nih.gov/) which identifies biomedical concepts by mapping text to the Unified Medical Language System (UMLS) Metathesaurus [7] is a popular choice. Unfortunately, the dictionary-based method suffers from poor coverage because it cannot find linguistically complex concepts (e.g. event-like concepts describing biomedical processes), concepts indicated by anaphoric expressions (e.g. pronouns or anaphoric expressions spanning sentences) or newly introduced concepts still missing in dictionaries.

The second step of LBD is to discover relations between concepts. This is typically done using co-occurrence statistics. However, since most co-occurring concepts are unrelated, this simple approach is error-prone and also fails to explain *how* two concepts might be related (e.g. that there is an *interaction* or *activation* relationship between them, or possibly a negative association). Semantic filtering based on relations in a thesaurus such as UMLS can help [6] but suffers from the limitations of dictionary-based approaches. While use of advanced text mining could enable the discovery of novel concepts and relations in context, it remains relatively unexplored in LBD [25].

For knowledge discovery, most systems use Swanson's ABC model or its extensions, e.g. concept chains [8], network analysis [16], and logical reasoning [24] (see [21] for a survey of such extensions). For a concept pair A and C, these models identify the most obvious B and return a ranking of pairs using measures such as average minimum weight, linking term count and and literature cohesiveness [26]. Based on partial B evidence only, these models are not optimally accurate and also do not produce data suitable for statistical hypothesis testing. The latter would be valuable for users of LBD as it could guide them towards highly confident hypotheses.

Evaluation of LBD is challenging as successful techniques discover knowledge that is not proven valuable at the time of discovery. Metrics for direct system comparisons are now available [26] and some existing techniques have been integrated in practical LBD tools which have been made freely available to scientists. Examples of such tools include Arrowsmith [23], BITOLA [5], Semantic MEDLINE [1], and FACTA+ [25], among others. These tools have been used to generate new scientific discoveries (e.g. candidate genes for Parkinson's disease, a link between hypogonadism and diminished sleep quality); see [6] and [21] for recent reviews. However, confirmation of such discoveries via actual laboratory experiments remains rare.

Due to combination of these factors, LBD is not in wide use yet, despite its recognised potential for scientific research [15]. Although closer engagement with end-users, better consideration of end-users needs, and increased validation of findings in the context of laboratory experiments is needed, the fundamental bottleneck lies in the current LBD methodology which suffers from poor coverage as it is capable of identifying only some of the relevant information in literature.

2.2 Advanced Text Mining

LBD could be greatly improved via use of advanced TM. Combining methodology from Information Retrieval (IR), NLP and data mining, TM aims to automatically identify, extract and discover new information in written texts [20, 19]. It can be used to organise vast amounts of unstructured textual data now generated through economic, academic and social activities into structured forms that are easily accessible and intuitive for users [15].

Given the rapid growth of scientific literature in biomedicine, biomedical TM has become increasingly popular over the past decade. Basic resources (e.g. lexicons, databases, annotated corpora, datasets) and NLP techniques such as part-of-speech (POS) tagging (i.e. classifying words) and parsing (i.e. analysing the syntactic structure of sentences) have been developed for biomedicine. IR (i.e. identification of relevant documents) and Information Extraction (IE) (i.e. identification of specific information in documents) is now developed, and relatively accurate techniques are now available for identification of named entities (e.g. concept name such as protein names, e.g. AntP), relations (e.g. specific interactions between AntP and BicD), and events (i.e. identifying facts about named entities, e.g. that the AntP protein represses BicD, repress(AntP,BicD)) in texts. Progress has also been made on increasingly complex tasks such as biological pathway or network extraction [10]. Not only direct evaluations against gold standard datasets but also evaluations in the context of practical tasks such as literature curation, literature review and semantic enrichment of networks have produced promising results, highlighting the great potential of deep TM in supporting biomedicine [20, 19, 9].

Much of recent TM research has focussed on enhancing TM further for demanding real-life tasks. In terms of accuracy, TM is challenged by the linguistic nature of biomedical texts. The biomedical language is characterized by heavy use of terminology and long sentences that have high informational and structural complexity (e.g. complex co-referential links and nested and/or inter-related relations). In addition, the mapping from the surface syntactic forms to basic semantic distinctions is not straightforward. For example, the same relation of interest may be expressed by nominalizations (e.g. phosphorylation of GAP by the PDGF receptor) and verbal predications (e.g. X inhibits/phosphorylates Y) which may not be easy to recognize and relate together.

NLP techniques such as statistical parsing and anaphora resolution which yields richer representations (e.g. internal structure of nominalisations, co-referential links in texts such as it, the protein, the AntP protein) are not challenged to the same extent as shallow techniques are [20, 19]. Integration of lexical, semantic, and discourse analysis could help and improve accuracy further [4, 13].

In terms of portability, TM has traditionally relied on expensive, manually developed resources (e.g. corpora consisting of thousands of sentences annotated for events by linguists) which are expensive to develop and therefore available for a handful of areas only (e.g. molecular biology, chemistry). Due to strong sub-domain variation resources developed for one area are not directly applicable to others [12]. Researchers are now improving the adaptability of TM by reducing the need for manual annotations via minimally supervised machine learning [3] and use of declarative expert (e.g. task, domain) knowledge in guiding learning [4]. Because text mining components typically build on each other, traditional systems have a pipeline architecture where errors tend to propagate from one level to another. Leveraging mutual disambiguation among related tasks and avoiding error propagation, joint learning and inference of various TM tasks is also gaining popularity and has been shown to further improve accuracy [17].

2.3 Towards LBD based on Advanced Text Mining

Based on much deeper analysis and understanding of texts, advanced TM could enable considerably more accurate, broader and dynamic LBD than current, largely dictionary-based methods. While this potential has been recognized, e.g. [15, 6], very little work has been done on TM-based LBD, e.g. [25]. For long, application of TM to LBD has been challenged by the interdisciplinary nature of biomedical research - the fact that research in one area draws increasingly on that in many others, while TM has been typically optimised to perform well in a clearly defined area. However, given recent developments in the field aimed at optimising both accuracy and portability of TM (see the developments discussed in the section above), the approach is now ripe for application in real-life LBD. Whilst TM is challenging by nature and will not produce fully accurate output, errors can be reduced e.g. via statistical filtering to produce maximally accurate input to LBD. Filtering has proved effective in previous works which have demonstrated the usefulness of adaptive TM for practical tasks in biomedicine, e.g. [19, 3]

The use of such enhanced, adaptive TM will enable targeting not only basic concepts (i.e. terms or named entities) like most previous LBD, but also complex concepts describing biomedical processes (i.e. events), and relations between concepts in diverse biomedical literature. The latter can be used to restrict search space by permitting direct connections only between concepts which are involved in specific relations [6]. All this information can be learned dynamically from relevant biomedical literature as science evolves, and LBD can be performed on the resulting complex network of concepts.

Open and closed LBD from such rich, TM-based data could also benefit from improved methodology for knowledge discovery. This methodology could be based on recent data mining techiques which enable considering all the intermediate concepts between target concepts. Just one example method is link prediction in complex networks [14] which has been applied successfully to to related problems in social network analysis [11] and web mining [2]. In comparison with most current LBD which is based on extensions of Swanson's ABC model [21] and considers only the most obvious intermediate concepts, such enhanced techniques could provide improved estimate of links between concepts. They could also generate data needed for calculating the likelihood of different concept pairs using statistical tests. This can be highly useful for scientists as it enables them to focus on highly confident hypotheses.

3 A Case Study in Cancer Biology

To illustrate the benefit of TM-based LBD we will describe how such an approach could be used to support the rapidly growing, literature-intensive area of cancer biology. Cancer biology is one of the "interdisciplinary" areas of biomedicine where knowledge discovery draws from advances made in a variety of sub-domains (rather than one well-defined sub-domain) of the field. This makes it particularly difficult for scientists to keep on top of all the relevant literature and highlights the need for automated LBD. From the perspective of TM-based LBD, an area such as cancer biology offers the research challenges needed for the development of adaptive TM technology as many sub-domains involved do not have annotated datasets that could be used for full supervision of systems.

The starting point is to gather relevant literature via PubMed – for example, all the MEDLINE abstracts and freely available full text articles from journals in relevant sub-areas of biomedicine (e.g. cell biology, toxicology, pharmacology, and medicine, among others). The resulting texts will be cleaned and processed using sophisticated NLP techniques such as part-of-speech tagging, parsing, semantic and discourse processing. Concepts of relevance to cancer research (e.g. cancer types, genes, proteins, drugs, physiological entities, symptoms, hallmarks of cancer) will then be extracted from NLP-processed texts, along with relations of interest (e.g. physical, spatial, functional, temporal) between the concepts.

While LBD that uses dictionary-based techniques can find mentions of simple concepts (e.g. gene names) and their known synonyms, TM can also find mentions of concepts "hidden" in anaphoric expressions, those appearing in complex linguistic constructions and those missing in resources such as UMLS, yielding more complete information for LBD. The concepts and relations would be extracted from rich NLP-annotated data using minimally supervised, adaptive TM-techniques. In the absence of relevant in-domain training data, TM can be guided by use of expert knowledge (e.g. constraints that capture task knowledge [4]) and joint inference of related tasks [17].

The network of concepts emerging from TM will be richer than that created by traditional methods. While it will also be noisier due to the challenging nature of advanced NLP and TM, previous work has demonstrated that the impact of noise on practical tasks, in particular after applying statistical noise filtering, will be minimal and unlikely to affect the usefulness of TM. Finally, sophisticated knowledge discovery techniques, e.g. [14], will be applied to the resulting complex network of concepts to conduct maximally accurate closed and/or open discovery.

Figure 1 illustrates how such a TM-based LBD tool could be used to support cancer biology. It shows an example that focuses on anti-carcinogenic effects of statins. Statins are known to have anti-carcinogenic properties but the underlying mechanism by which these drugs prevent cancer is not fully understood [18]. This problem can be studied by investigating whether specific proteins and hallmarks of cancer act as intermediate concepts between statins and different cancer cell types and if yes, whether such concepts could help to explain the mechanism. In the case study illustrated in Figure 1, cancer biologists use a TM-based LBD tool for closed discovery to investigate the question *In which way do statins prevent prostate cancer?* The given concepts are

Concept A: Drug: Statin
Concept C: Cancer type: Prostate cancer

The tool will

1. gather literature: PubMed articles on "statin" and "prostate cancer",
2. identify Concepts B (Hallmarks, Proteins) in the resulting literature using TM,
3. build a concept map for Concepts A, B and C,
4. return B that link to both A and C.

The tool will also identify relevant relations between concepts:

Interacts with (Statin, Akt kinase)
Causes (Akt kinase, Prostate cancer)
Prevents (Statin, Sustaining proliferative signaling)
Causes (Sustaining proliferative signaling, Prostate cancer)
Exhibits (Akt kinase, Sustaining proliferative signaling)

The answer emerging from the tool is that *statins prevent prostate cancer by inhibiting cell proliferation via Akt kinase*.

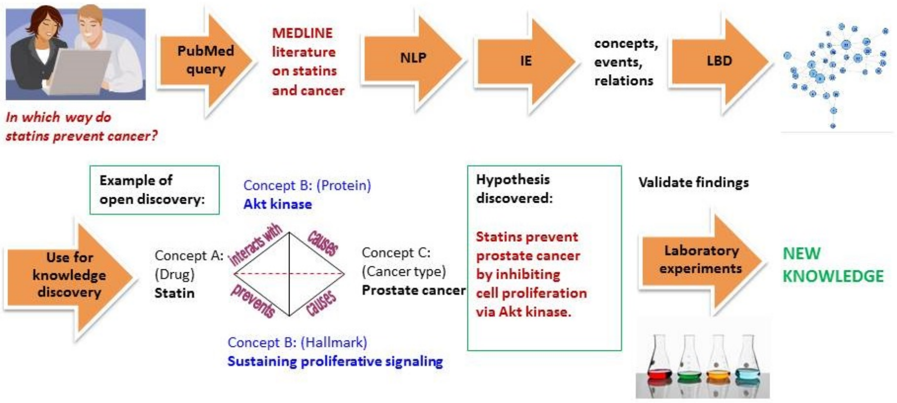
In which way do statins prevent cancer?
PubMed query
MEDLINE literature on statins and cancer
NLP
IE
concepts, events, relations
LBD
Example of open discovery:
Concept B: (Protein)
Akt kinase
Hypothesis discovered:
Statins prevent prostate cancer by inhibiting cell proliferation via Akt kinase.
Validate findings
Use for knowledge discovery
Concept A: (Drug) Statin
interacts with
causes
prevents
causes
Concept C: (Cancer type) Prostate cancer
Laboratory experiments
NEW KNOWLEDGE
Concept B: (Hallmark)
Sustaining proliferative signaling

Fig. 1. TM-based LBD for cancer biology. The figure illustrates how LBD can discover the mechanism by which statins prevent prostate cancer.

To be useful, such TM-based technology should be integrated in a practical tool aimed at supporting cancer researchers in LBD. The tool should allow uploading articles of interest e.g. via PubMed, performing open and close discovery using a set of queries to define the scope of interest in terms of concepts and relations, visualising the results and the statistical trends in the data, and navigating through individual articles, highlighting the scientific evidence in its actual context. Such a tool should be developed in close collaboration with scientists to ensure optimal integration with existing research practices.

Finally, any new hypotheses or discoveries resulting from LBD should ideally be confirmed experimentally by scientists. In the case of cancer biology, one might validate the promising findings from LBD experimentally in vitro according to their nature. Such experimentation and subsequent publication in relevant journals can encourage the uptake of LBD by the research community, leading to further benefits.

4 Conclusion

In biomedicine a number of LBD tools have been developed to support the testing and discovery of research hypotheses in scientific literature. Although such tools could, in principle, greatly support scientists in their work, their uptake has remained limited. We have highlighted a number of issues that act as barriers to wider exploitation of LBD in scientific research, and have focused, in particular, on limitations related to the current LBD methodology. We have explained how use of advanced TM could enable the discovery of much richer information in scientific texts than is possible using current largely dictionary-based methods. This potential has been previously recognised, but TM has only recently reached the point where it can be realistically applied to diverse literature without costly creation of manually annotated in-domain training data. While the development of a fully optimal LBD approach based on TM will require considerable research effort, it is now realistic – and looking into the future, the approach could open the doors to much wider coverage LBD capable of better evolving with the development of biomedical science.

Acknowledgments

We would like to thank the Royal Society (UK) and the Swedish Research Council (Sweden) for financial support.

References

1. Semantic MEDLINE. Available at: http://skr3.nlm.nih.gov/Sem-MedDemo/
2. Chakrabarti, S.: Mining the web. Morgan Kaufmann (2002)
3. Guo, Y., Korhonen, A., Silins, I., Stenius, U.: Weakly-supervised learning of information structure of scientific abstracts - is it accurate enough to benefit real-world tasks in biomedicine. Bioinformatics 27(22) (2011)
4. Guo, Y., Reichart, R., Korhonen, A.: Improved information structure analysis of scientific documents through discourse and lexical constraints. In: NAACL (2013)
5. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature-based discovery to identify disease candidate genes 2-4(74), 289–298 (2005)
6. Hristovski, D., Rindflesch, T., Peterlin, B.: Using literature-based discovery to identify novel therapeutic approaches. In: Cardiovasc Hematol Agents Med Chem, vol. 11, pp. 14–24 (2013)
7. Humphreys, B.L., Lindberg, D.A.B.: The umls project. In: Bull. Med. Lib. Assoc., vol. 81, pp. 179–177 (1993)
8. Jin, W., Srihari, R.K., Ho, H.H., Wu, X.: Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In: ICDM. IEEE Computer Society, Washington, DC, USA (2007)
9. Kadekar, S., Silins, I., Korhonen, A., Dreij, K., Al-anati, L., Hogberg, J., Stenius, U.: Exocrine pancreatic carcinogenesis and autotaxin expression. PloS One 7(8) (2012)
10. Li, C., Liakata, M., Rebholz-Schuhmann, D.: Biological network extraction from scientific literature: state of the art and challenges. Brief Bioinform (2013)
11. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management. pp. 556–559 (2003)
12. Lippincott, T., D. O'Seaghdha, A.K.: Exploring subdomain variation in biomedical language. BMC Bioinformatics 12(212) (2011)

13. Lippincott, T., Rimell, L., Verspoor, K., Korhonen, A.: Approaches to verb subcategorization for biomedicine. Journal of Biomedical Informatics 46(2) (2013)
14. Lu, L., Zhou., T.: Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 6(1150–1170) (2011)
15. McDonald, D., Kelly, U.: Value and benefits of text mining. JISC Publications (2012)
16. Ozgur, A., Xiang, Z., Radev, D.R., He, Y.: Link prediction in complex networks: A survey. J Biomed Biotechnol pp. 426–479 (2010)
17. Poon, H., Vanderwende, L.: Joint inference for knowledge extraction from biomedical literature. In: HLT-NAACL (2010)
18. Roudier, E., Mistafa, O., Stenius, S.: Statins induce mammalian target of rapamycin (mtor)-mediated inhibition of akt signaling and sensitize p53-deficient cells to cytostatic drugs. Molecular Cancer Therapeutics 5(2706–2715) (2006)
19. Shatkay, H., Craven, M.: Mining the Biomedical Literature. MIT Press (2012)
20. Simpson, M.S., Demner-Fushman, D.: Biomedical Text Mining. Springer US (2012)
21. Smalheiser, N.R.: Literature-based discovery: Beyond the abcs. In: JASIST (2012)
22. Swanson, D.R.: Fish oil, raynauds syndrome, and undiscovered public knowledge. Perspect. Bio. Med 30, 7–18 (1986)
23. Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91(2), 183–203 (1997)
24. Tari, L., Anwar, S., Liang, S., Cai, J., Baral, C.: Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 26, 426–479 (2010)
25. Tsuruoka, Y., Makoto, M., Hamamoto, K., Tsujii, J., Ananiadou, S.: Discovering and visualizing indirect associations between biomedical concepts. In: Bioinformatics, vol. 27, pp. 111–9 (2011)
26. Yetisgen-Yildiz, M., Pratt, W.: A new evaluation methodology for literature-based discovery systems. J Biomed Inform 4(42), 633–43 (2009)