TOWARDS OPEN DOMAIN LITERATURE BASED DISCOVERY

O. M. Bensch*, Maastricht University, [6200] Maastricht, The Netherlands
T. Hecking †, German Aerospace Center (DLR), [51147] Cologne, Germany

Abstract

Literature based discovery (LBD) is concerned with the extraction of implicit knowledge from large corpora of scientific publications inferring previously unseen links between terms and concepts, which can potentially lead to new hypotheses and findings. Most LBD systems are used in biomedical and related domains, where the existence of well elaborated domain taxonomies and ontologies support the automatic extraction of relevant information from texts. Due to a lack of such resources and different nature of publications LBD has been rarely used in other scientific areas. This work explores and evaluates text and graph methods for open domain concept and relation discovery in scientific literature. First results indicate that several different approaches have to be combined to detect a sufficient amount of concepts and meaningful relationships in an open domain corpus. The work can contribute to broaden the scope of LBD systems and potentially lead to new applications.

INTRODUCTION

Search applications let users express their information needs as specific search queries that are matched against a search index for documents or translated into database queries. In contrast, discovery systems focus on exploration of information items in a less targeted manner. Here the goal is to exploit rich datasets to discover something new, unexpected, and possibly inspiring. This paradigm has been applied in literature-based discovery (LBD) systems [1] that aim at fostering new scientific developments and hypotheses in an automated manner. The ratio behind this is that large publication databases contain a lot of implicit knowledge that is not manifested in one publication alone but becomes salient when insights from several publications are combined. In this sense, Swanson established the "ABC-Model" for LBD to automatically generate and evaluate new hypotheses [2]. In a first step relations between concepts (or meaningful terms) are extracted from scientific text corpora, e.g. based on co-occurrence in documents, which eventually results in a concept network. In the discovery part, one aims at predicting previously unseen relationships in such networks from transitive relations. A prominent example from Swansons seminal work [3] is: The relation that fish oil (A) lowers the blood viscosity (B) was found in one set of publications. Another set of publications reports that a high blood viscosity (B) causes Raynaud's disease (C). With that explicit knowledge a new hypothesis can be stated that there is an implicit relations between A and C. This hypothesis was later proven correct [3].

* o.bensch@student.maastrichtuniversity.nl
† tobias.hecking@dlr.de

Based on the ABC model two search approaches can be used to generate new hypotheses from literature corpora (illustrated in 1. In open discovery, the search term "A" is given. This concept is used to identify "B" concepts that are related to "A", as well as "C" concepts that are related to the "B" but not "A". In closed discovery two concepts "A" and "C" are given and the aim is to find bridging concepts "B" that connect "A" and "C".

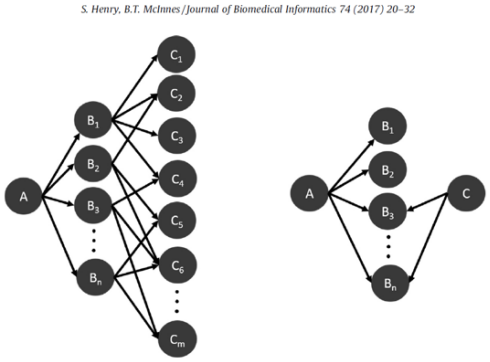

Figure 1: LBD approaches from [4]

The typical workflow comprises of (1) identification of important concepts in documents, (2) relation extraction and creation of a concept network, and (3) ranking of unconnected concept pairs. While the first literature-based discoveries mainly came from manual inspection of documents, with the advancing methods in natural language processing and graph mining the process becomes increasingly automatised [4].

However, the vast majority of LBD systems focus on the biomedical domain, which might be attributed to highly specific and descriptive content in research papers in this domain, as well as the existence of well elaborated taxonomies such as the Unified Medical Language System (UMLS), which alleviates the extraction of meaningful information [5].

As a step towards broader adaptation in scientific discovery and monitoring systems, the main goal of our research is to explore techniques for LBD in open domain text corpora. This paper reports on our first results as well as technical issues along examples from a literature corpus of 25.161 English abstracts retrieved from the publication server elib [6] of the German Aerospace Center (DLR). We, furthermore, reflect on possible future directions for open scientific discovery systems including semantic augmentation and the usage of existing scientific knowledge graphs.

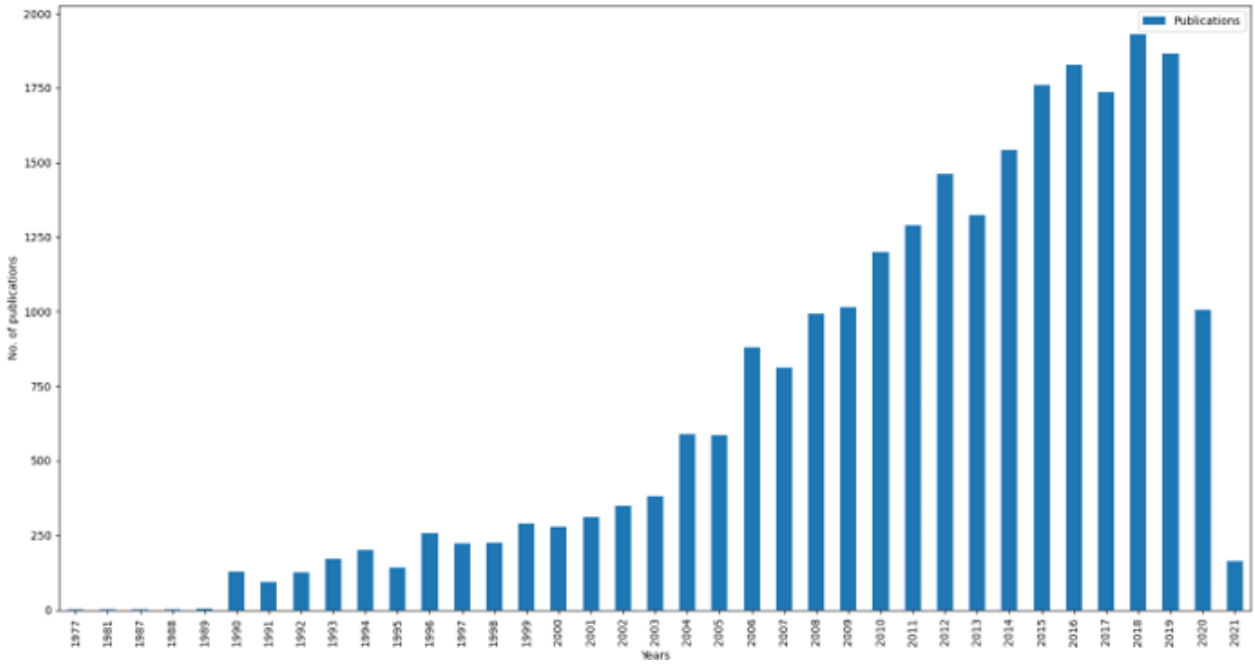

Figure 2: Abstracts per year.

RELATED WORK

Although literature-based discovery has a long research history originating in the 80's [1], it is still mainly focused on variations of the classical ABC model with applications in biomedicine and chemistry.

In these contexts, LBD systems often use the Unified Medical Language System (UMLS) [7] to identify important concepts mentioned in medical and biological literature. It contains more than 2 million terms, 900,000 concepts, and over 12 million relations. These are used in conjunction with natural language processing methods, like named entity recognition for information extraction from text corpora [8]. One state of the art example for an existing LBD system in the medical domain is LION LBD [9] focusing on the molecular biology of cancer. It uses the PubTator [10] to annotate important medical concepts in publications. Based on co-occurrences of concepts LION LBD uses machine learning to infer promising new relationships for open as well as closed discovery. Due to the lack of comparable ontologies in other domains, semantic augmentation for LBD is rarely used in other domains than biology, chemistry, medicine, and biomedicine [5].

However, most recently it has been shown that automatic discovery of implicit knowledge in publication databases is also possible in other domains. For example, in the field of material science Tshitoyan et al. [26] could predict scientific discoveries years before corresponding experiments were actually conducted by applying machine learning techniques on word embeddings created from older literature.

METHODS AND EXAMPLES

We created a test data set of 25.161 English abstracts retrieved from the publication server of the German Aerospace Center (DLR) elib [6], to evaluate open domain LBD approaches. The abstracts in this dataset were composed on average of 188,74 words in 8,3 sentences. The distribution of abstracts per year of this dataset can be seen in figure **??**. It can be seen that the amount of papers published per year increases from year to year. Abstracts until the end of may 2021 were included in this dataset.

In the following different approaches for the three main steps of the LBD process (concept detection, relationship identification, and concept pair ranking) are described pointing out their strength and weaknesses for the task at hand.

Concept detection

There are several ways to detect terms in sentences that refer to concepts of interest. Several approaches that solely work on the syntactic level can only detect single words, which will miss out multi-word expressions such as 'machine learning', while other techniques involving grammatical analysis can also detect coherent terms as one concept. Another important aspect of concept detection is matching expressions to ontologies as external knowledge bases, which are, however, often not available or not sufficiently elaborated.

Named Entity Recognition One example of a technique that can detect multi-word expressions as one concept, as well as a corresponding ontology is named entity recognition (NER).

The most common open domain model for NER is the Stanford NER [18] which was trained on the OntoNotes 5.0

dataset. This NER model was built to detect 18 different types of entities such as persons, organizations, products, dates, or monetary terms. Consequently, this model works well for information extraction in economic domains. However, our experiments have shown that the Standford NER could not be used to extract meaning full concepts from scientific literature since this model was not trained to detect terms such as technical components, methods, etc., which are important for LBD. Overall this model detected 2-3 terms per abstract, which was not sufficient to extract meaning full concept relations.

An example of this approach can be seen in figure 3.

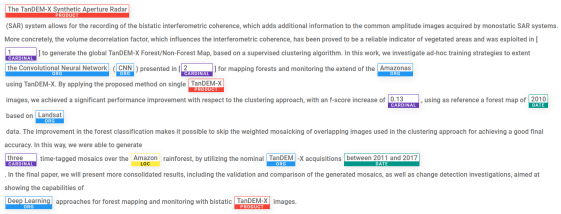

Figure 3: Stanza example.

Wikification Another recent approach to detect open domain terms is wikification [15]. Wikification makes use of Wikipedia as external knowledge base to identify meaningful concepts in texts by matching terms to titles of Wikipedia articles (see [15] for an example). Apart from a good precision also for multi-word expression, another advantage is that one can make use of additional information associated with an article especially the category of the page for concept classification or redirects to other articles for synonym resolution. To evaluate the wikification approach a list of all occurring words in the test dataset was created. In the next step, we used the following SPARQL query to query DBPedia (an ontology based on Wikipedia) to detect English terms with matching articles.

```
SELECT DISTINCT * WHERE {
 ?url rdfs:label "' + searchText + '"@en .
}
```

The "searchText" variable was replaced by a single word for each query. In this example, the query was only performed for single terms. However, it can easily be modified to also match multi-word expressions by checking expressions with variable length.

This approach performs better for abstracts of older publication, as Wikipedia entries are created over time. Very novel concepts, for which no wiki article exists will be missed which is definitely a disadvantage for LBD. On average 6.3 concepts could be identified in the abstracts in our dataset, and most of them appear to be useful for scientific information extraction.

An example of this approach can be seen in Figure 4.

TF-IDF Another approach is to detect terms with TF-IDF. In contrast to the previous approaches, TF-IDF solely

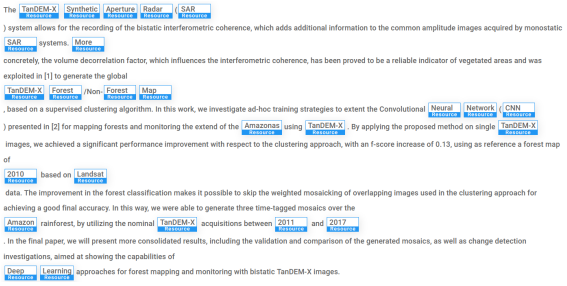

Figure 4: Wikification example.

relies on term occurrence statistics and no ontologies or grammatical analysis are needed. On the downside, this approach can only detect single terms as a concept resulting. Furthermore, one has to specify a threshold for TF-IDF scores of words to be included as a concept. Depending on this threshold common words like "within" can also be classified as a concept. This can be circumvented by using proper stopword lists to remove such common language words.

With well adjusted threshold (in our example case 0.7) and in combination with stopword filtering and lemmatization for pre-processing and possibly further post processing steps, this approach can detect single-word concepts quite reliably (see Figure 5). While TF-IDF detects more concepts than wikification, there are also terms that are not useful e.g. 'single' or 'exploited'. This indicates that TF-IDF may have a higher recall in concept detection but lower precision compared to wikification.

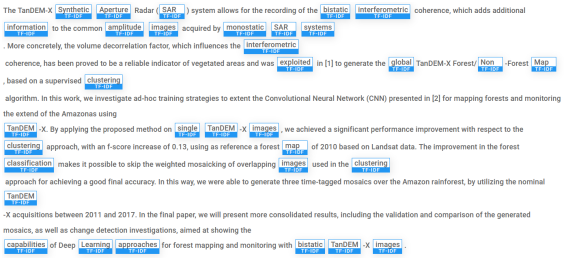

Figure 5: TF-IDF example.

Relation identification

Once important concepts are identified, the next step is to discover their connections. This is often done based on co-occurrences. The most simple approach would be to connect all concepts from the same abstract pairwise. However, co-occurrence can also be defined on sentence or paragraph level. Although co-occurence based concept linking discards semantic information, i.e. the nature of their relationship, it is known that co-occurrence models often inherently captures the semantic structure of a text to a sufficient extent [19]. The significance of a relation between two concepts can be determined from the number of such co-occurrences, which can be used to filter sporadically and noisy relations.

Based on the selected technique for concept detection, this method has to tackle several issues. For methods that only

detect single words, like TF-IDF, a co-occurrence-based concept relation detection might also detect coherent concepts instead of relations between independent concepts.

Our experiments have shown that relation detection on sentence-based co-occurrence with concepts extracted with TF-IDF and a threshold of 0.7 can not only be used to detect concept relations but also coherent concepts. In this case, the highest concept relation rank was achieved by the two tokens "open" and "source" as these often co-occur in sentences. As a result, a concept relation with a high co-occurrence rank and a low word distance between those concepts could indicate a coherent concept. Based on a threshold this approach could be used to detect coherent concepts with methods that would normally only detect single word concepts. Depending on the selected threshold words like "within" might be considered as a concept by TF-IDF. Stopword lists could be used as a filter for those words.

In our experiment we used a threshold of 0.1 for TF-IDF. The top 50 co-occurring words can be seen in figure 6.

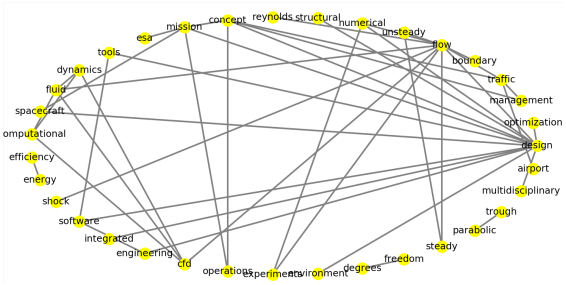

Figure 6: TF-IDF top 50 co-occurring concepts and relations.

By increasing the amount of concept relations detected by this method more and more sub-graphs were combined to larger graphs. However, less important concept relations were included this way.

In a zoomed version of this figure example relations between the extracted concepts can be seen. This zoomed version can be seen in figure 7.

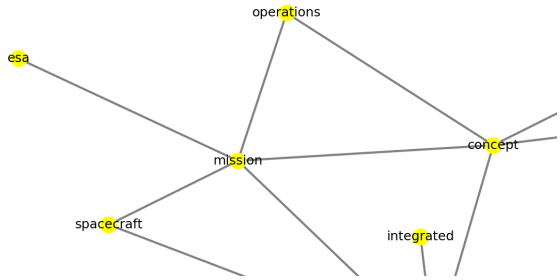

Figure 7: TF-IDF top 50 co-occurring concepts and relations, zoomed.

Concept pair ranking

The result from concept and relation identification can be represented as a network. For discovering new previously not observed relationships that can potentially lead to new hypotheses and ideas, one can borrow from link prediction methods established in social network analysis [20]. The most simple approach is to rank unconnected node pairs based on the number of their common neighbours, which corresponds to the classical ABC model of LBD [1] described in the beginning of this paper. The idea is that for two concepts that share a large number of common neighbours but are not connected by an edge themselves, there could exist an implicit connection. A drawback, however, is that two concepts that are well connected in the network have a higher chance to share neigbours than concepts sparsely connected to the rest of the network. To account for this, one can use the Jaccard similarity of the neighbourhoods of two concepts instead of the total number of shared neighbours. While the aforementioned methods only take two-step connections into account, the Katz coupling measure [20] counts all paths between two concept nodes and includes a weighting parameter that downweights longer paths. Thus, this measure can be considered as a generalisation of the common neighbours method. However, it suffers from the same problem that it can be a biased towards well connected nodes.

To illustrate the differences between different link prediction methods for concept pair ranking, we build a graph of concepts identified by the TF-IDF method with more than 10 co-occurrences on the sentence level. The result is a network of 6109 nodes with 275513 edges. Excerpts of the top ranked concept pairs are given in table 1

It can be seen that the Katz measure and the common neighbours measure share most of the new detected relations, as well as the weights between these. The term 'sar' (search and rescue) is dominant. As mentioned above, one reason can be that terms that have many more connections than the average (also called hubs), which is a well-known property of many networks [21], are connected to many others via short paths. In contrast, the relations discovered with the Jaccard measure differs completely from the other two and appears to be less biased toward such hub concepts.

CONCLUSION

We have explored different approaches for extracting meaningful concepts and their relationships from publication abstract for open-domain literature based discovery. Initial findings suggest that several approaches should be combined along with pre- and post-processing to improve the detection rate. Apart from that, open domain LBD remains challenging since every discipline differs in the nature of findings and how to communicate them. Consequently, there will be no one-fits-all solution so that open domain LBD systems should rather provide a set of configurable tools from which specific applications can be built.

We believe that the potential of using computational tools to link existing pieces of information to support scientific discovery has not yet fully been exploited, especially in the light of emerging open web indices and scientific knowledge

Table 1: Top concept pairs extracted by different link prediction methods

Common Neighbours	Katz	Jaccard
sar – aircraft	sar – aircraft	files – streams
images – design	images – design	exchangers – pumps
sar – temperature	sar – temperature	housekeeping – streams
presented – presents	sar – preasure	linearity – uniformity

graphs, e.g. Open Academic Graph (OAG) [22]. Advances in this direction can be a key element to make better use of the rapidly growing amount of scientific information available not only in traditional publications but also on the web.

FUTURE WORK

On the technical level, further approaches for entity extraction could be evaluated. Similar to the wikification approach, a dictionary model like wordnet [23] or, when available, curated domain taxonomy could be used. This would also alleviate the problem of synonyms and multi-word expressions.

Further processing steps like dependency parsing or machine learning models could be used to detect coherent concepts and relations between them. These could also be combined with pattern-based approaches similar to Hearst patterns for discovery of hyponyms [**?**], for example, based on keywords like "influences" that indicate specific connections.

To combine the mentioned approaches, we plan to use weak supervision frameworks such as Snorkel AI [24]. These, take the results of various heuristics expressed as (imperfect) data labelling functions and combines them in a probabilistic framework to create consistently labelled training data for building machine learning models for information extraction.

Apart from technical advancements of components in LBD pipelines, in the future one can also go beyond publication data for knowledge discovery. Since scientific output is increasingly available on the web and manifested also in form of software publications in public repositories or open datasets, one can also include these diverse sources into the discovery process (c.f. [25]).

REFERENCES

[1] M. Thilakaratne, K. Falkner, T. Atapattu "A Systematic Review on Literature-based Discovery: General Overview, Methodology, Statistical Analysis" in Association for Computing Machinery New York, USA, December 2019, Article No.: 129.

[2] D.R. Swanson "Migraine and magnesium: Eleven neglected connections." in Perspectives in Biology and Medicine 31 Summer 1988, pp. 526–557.

[3] R. A. DiGiacomo, J. M. Kremer, D. M. Shah "Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study" in The American Journal of Medicine Volume 86 Issue 2, New York, 1989, pp. 158–164.

[4] S. Henry, McInnes, T. Bridget "Literature based discovery: models, methods, and trends" in Journal of biomedical informatics 74, 2017 pp. 20–32.

[5] M. Thilakaratne, K.Falkner, T. Atapattu "A Systematic Review on Literature-Based Discovery: General Overview, Methodology, amp; Statistical Analysis" in Association for Computing Machinery 52, 2020.

[6] elib `https://elib.dlr.de`

[7] O. Bodenreider "The Unified Medical Language System (UMLS): integrating biomedical terminology." in Nucleic Acids Research, Volume 32, January 2004, pp. D267—D270.

[8] D. Hristovski, C. Friedman, T. Rindflesch, B. Peterlin "Exploiting Semantic Relations for Literature-Based Discovery" in AMIA Annual Symposium proceedings. AMIA Symposium, 2006, pp. 349–353.

[9] S. Pyysalo et al. "LION LBD: a literature-based discovery system for cancer biology", Bioinformatics, Volume 35, Issue 9, 1 May 2019, pp. 1553—1561.

[10] C. Wei, A. Allot, R. Leaman, Z. Lu "PubTator central: automated concept annotation for biomedical full text articles", Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, pp. W587—W593.

[11] T. C. Rindflesch, M. Fiszman "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text" in Journal of Biomedical Informatics 36 (6), pp. 462–477.

[12] C. Chantrapornchai, A. Tunsakul "Information Extraction on Tourism Domain using SpaCy and BERT" in ECTI Transactions on Computer and Information Technology (ECTI-CIT). 15, pp. 108-122.

[13] E. Sang, F. De Meulder "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition" in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142—147.

[14] R. Weischedel et al. "OntoNotes Release 5.0" in Linguistic Data Consortium Philadelphia, October 2013 `https://catalog.ldc.upenn.edu/LDC2013T19`

[15] Y. Taskin, T. Hecking, H. Hopppe "ESA-T2N: A Novel Approach to Network-Text Analysis" in Complex Networks and Their Applications VIII, pp.129–139.

[16] SPARQL `https://www.w3.org/TR/rdf-sparql-query`

[17] S. Auer et al. "DBpedia: A Nucleus for a Web of Open Data" in 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference Busan, Korea, 2007, pp. 722—735.

[18] J. Finkel, T. Grenager, C. Manning "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." in Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.

[19] , Carley, Kathleen, Palmquist, Michael (1992). Extracting, representing, and analyzing mental models. Social forces. 70(3), 601–636.

[20] Liben-Nowell, D., Kleinberg, J. (2007). The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7), 1019-1031.

[21] Barabási, A. L., Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.

[22] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[23] Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

[24] A. Ratner et al."Snorkel: rapid training data creation with weak supervision." in Proc. VLDB Endow. 11, 3 November 2017, pp. 269-–282.

[25] R. el Baff, S. Santhanam, T. Hecking "Quantifying Synergy between Software Projects using README Files Only" in Proceedings of the International Conference on Software Engineering Knowledge Engineering, 2021, Pittsburg, USA, pp. 265–270.

[26] Tshitoyan, V., Dagdelen, J., Weston, L. et al. (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 .