Using Discourse Structure of Scientific Literature to Differentiate Focus from Background Entities in Pathogen Characterisation

Antonio Jimeno Yepes[1,2], Ameer Albahem[1], and Karin Verspoor[2,1]

[1]School of Computing and Information Systems, University of Melbourne
[2]School of Computing Technologies, RMIT University
Melbourne, Victoria, Australia

antonio.jimeno@gmail.com, ameer.albahem@gmail.com, karin.verspoor@rmit.edu.au

Abstract

In the task of *pathogen characterisation*, we aim to discriminate mentions of biological pathogens that are *actively studied* in the research presented in scientific publications. These are the pathogens that are the focus of direct experimentation in the research, rather than those that are referred to for context or as playing secondary roles. This task is an instance of the more general problem of identifying *focus entities* in scientific literature, in which key entities of interest must be discriminated from other potentially relevant entities of the same type mentioned in the articles.

In this paper, we explore the hypothesis that focus pathogens can be differentiated from other, non-actively studied, pathogens mentioned in articles through analysis of the patterns of mentions across different segments of a scientific paper, that is, using the discourse structure of the paper. We provide an indicative case study with the help of a small data set of PubMed abstracts that have been annotated with actively mentioned pathogens.

1 Introduction

Global monitoring of repositories of potentially harmful biological materials is an important component of ensuring the health and safety of our populations. In this context, we are building an information extraction system to identify information related to experimentation with potentially dangerous biological pathogens – e.g. viruses, bacteria, and biological toxins – as well as to detect facilities that may serve as repositories of harmful pathogens. This system will systematically scan open access data sets for evidence of research on those pathogens, thereby supporting gathering of information from public resources for biosecurity purposes (Jarrad et al., 2015).

A key requirement for automated characterisation of research on pathogens using text-based information sources, including the scientific literature, is to identify pathogens that are *actively studied*. An actively studied pathogen is defined as an organism that is subjected to direct physical experimentation in the research.

Recognition of potentially relevant entities is relatively advanced using biomedical named entity recognition tools that detect biological nomenclature such as the names of biological organisms (e.g. as studied in the context of BioCreative (Smith et al., 2008)). However, differentiating mentions of actively studied organisms from other, background or incidental mentions of organisms poses a deeper natural language processing challenge. In the context of chemical patents, it has been suggested that only ~10% of chemical mentions play a major role within the patent (Akhondi et al., 2019). It is insufficient to simply detect a mention of a potentially relevant pathogen name; it must also be decided whether that pathogen is a focus of the experiments. The main goal of our *pathogen characterisation* task is therefore to enable filtering out pathogens that are mentioned in articles but not considered to be actively studied in the described research.

Publications may refer to pathogens in various ways. In addition to mentions in the context of direct experimentation, pathogens may be mentioned as part of background knowledge or in the context of discussion or comparison. We propose that a key element of identifying actively studied pathogens is understanding where in a publication a pathogen is described (e.g. in a Methods segment vs. in the Background segment of the paper), and how the pathogen is relevant to the research (e.g. mentions of the pathogen being subjected to specific tests or examinations that reveal experimentation).

In this paper, we therefore explore the hypothesis that the context in a scientific paper where a potentially relevant entity is mentioned can provide clues about whether that entity is a *focus* (foregrounded)

entity, or an entity in the *background*; our notion of an actively studied entity assumes that it is a focus of the research described in a paper.

We investigate this hypothesis by comparing the distribution of focus and background entities across discourse segments, and apply association rule mining to identify combinations of segments that are relevant to identify focus entities. We present a small case study illustrating the proposed methodology, providing preliminary evidence of the value of discourse structure – consideration of *where* entities are mentioned – for identifying focus entities.

2 Related work

Identifying salient entities is a relevant component of information retrieval and text summarisation. The study of discourse structure has been suggested in previous work on entity salience (Boguraev and Kennedy, 1999; Walker and Walker, 1998). The work of (Dunietz and Gillick, 2014) evaluates a comprehensive set of features, showing that the discourse structure and centrality may support predicting entity salience. Our task differs in that we adopt a narrower focus specifically on identification of actively studied pathogens in scientific research papers.

Pathogen characterisation has been studied in recent shared tasks, such as the Bacteria Biotope task (Bossy et al., 2019). The tool GeoBoost (Tahsin et al., 2018) also addresses the identification of entities from GenBank, which includes largely information about viruses and bacteria. The main role of GeoBoost is to identify the location of these biological entities, which requires performing natural language processing tasks in addition to combining information from NCBI resources. This work does not address saliency of entity mentions.

In our work, we evaluate discourse features for direct identification of actively researched pathogens, covering a broad set of pathogen types.

3 Datasets

In our experiments, we constructed a dataset based on information obtained from the Biological Material Information Program (BMIP)[1] of the Defense Threat Reduction Agency (DTRA)[2].

[1]BMIP media article:
https://globalbiodefense.com/2017/05/08/bmip-pathogen-repositories-worldwide

[2]https://www.dtra.mil/

3.1 Pathogen entity list

We were provided with a list of all pathogens tracked in the BMIP database, which we refer to as the BMIP list. To align these pathogens to publicly available resources, and normalise their representation, we mapped each pathogen in the list to the NCBI Taxonomy (Federhen, 2012) via direct lookup. These pathogens include viruses, bacteria, viroids, fungi and protozoa. In addition, there are mentions of toxins and PrPSc prions that were assigned a custom identifier.

3.2 Gold standard dataset

We have a small initial gold standard dataset that we use for our investigation. It consists of manual annotations of relevant pathogens over PubMed citations. *Relevance* is defined here as evidence of an actively studied pathogen, or focus entity.

This gold standard contains 87 PubMed citations (publication metadata) including titles and abstracts, each with an associated list of relevant pathogens. Out of these 87 citations, 35 have no actively studied pathogen, so we consider 52 citations in this study. There are a total of 69 relevant pathogen mentions, corresponding to 32 unique pathogens (individual NCBI Taxonomy IDs), identified across the remaining 52 articles. The maximum number of relevant pathogens annotated for a document is 5. Nineteen (19) pathogens are annotated only once; the pathogen with the largest number of annotations is *H1N1*, with total frequency of 11 (i.e. 11 citations are annotated with this pathogen). Most pathogens in the gold standard belong to the Influenza virus family.

4 Methods

We approach identifying focus entities of scientific articles as a two-stage process: pathogen identification and pathogen characterisation. Here, we describe our approach to each stage.

4.1 Pathogen identification

In the pathogen identification stage, the objective is to find all pathogens mentioned in a citation, irrespective of whether they are focus or background entities. Despite some pathogen mentions are available in author keywords and MeSH indexing, this information is sparse within the citations in MEDLINE or not mentioned at all. Both dictionary lookup and machine learning models learned from annotated data are possible for this step. Lacking

annotated data specifically for the BMIP pathogen list, we utilise the dictionary-based ConceptMapper tool (Tanenblatt et al., 2010), found by Funk et al. (2014) to outperform other methods. We leverage the NCBI Taxonomy ConceptMapper annotation pipelines for the CRAFT corpus[3]. We construct the dictionary based on the BMIP list of relevant pathogens, mapped to the NCBI Taxonomy using the database downloaded from the OBO Foundry[4].

The BMIP list of pathogens also includes mentions of pathogens that are either toxins generated by pathogens or PrPSc prions, which are proteins with a pathological folding. Toxin mentions are identified using regular expressions that have higher recall than just using a dictionary matching while obtaining the same level of precision.

Using these strategies for identifying pathogens, we detect 49 mentions annotated as focus entities (out of the 69 from the 52 citations) and 9 mentions that we treat as background entities.

4.2 Pathogen characterisation

Given the list of pathogens in an abstract, the next step is to characterise which of these pathogens are focus entities, i.e. actively researched.

As described before, we hypothesise that focus pathogens are more likely to appear in some segments than others (e.g. in Methods segments vs. Fact segments), and that therefore the mention patterns of actively studied (focus) pathogens across segments are different from the mention patterns of not-actively studied (background) pathogens.

To model mention patterns, we adopt the method of association rule classification Liu et al. (1998)[5] to infer rules based on which discourse segments a pathogen is mentioned in that predict that the pathogen is a focus entity.

We treat the event of mentioning a pathogen once or more in a scientific article as a transaction event. Each transaction consists of items corresponding to the discourse structure labels of the different mentions of the pathogen. For instance, if the pathogen *Bacillus anthracis* is mentioned once in the title and once in the methods segment of a citation, we add a transaction to our dataset with the itemset (TITLE, METHOD). Given this transaction dataset, we employ association rule mining to mine top association rules for focus entities.

The class association rules (CAR) are obtained using a two-part algorithm. First, rules are generated using the APRIORI algorithm (Agrawal et al., 1994). The algorithm generates association rules that have enough support and confidence. The rules are generated without any target classification task under consideration, i.e. mention patterns for both focus and background entities are considered.

In the second part, the generated rules are used to build a classifier using the CAR M1 algorithm (Agrawal et al., 1994). The rules are sorted by confidence and then by support. Following this order, if a rule correctly classifies examples in the instance set, the rule is selected and those examples are removed. The total number of errors is recorded for the rule as the error of the rule on the instance set and the error of the default class (selected using the majority class of the remaining examples). Additional rules are selected using the remaining examples and this process continues until there are no more rules or examples. From the set of selected rules, the one with the lowest total numbers of errors is identified and the rules after that one are discarded, which reduces the error of the set. A rule is added at the end that returns the default class, which is the most frequent class not covered by the selected rules.

4.3 Discourse segment labeling

Ideally, we would hope to access articles with explicit discourse structure such as the introduction, methods, and results headings. However, such labelling is available for less than one quarter of PubMed abstracts (Jimeno Yepes et al., 2013). We therefore use automated discourse structure tagging to label segments in each abstract.

We build on existing work in scientific discourse tagging (Dasigi et al., 2017), which utilises a deep learning sequence-labeling model that identifies structure within experiment narratives in the scientific literature. A seven-label taxonomy is adopted from de Waard and pan der Maat (2012), containing GOAL, FACT, RESULT, HYPOTHESIS, METHOD, PROBLEM, and IMPLICATION. Li et al. (2019, 2021) extends the previous work, training on their SciDT dataset that contains 634 paragraphs and 6124 clauses. Their method combines a SciBERT (Beltagy et al., 2019) feature generator with a recurrent neural network to predict the scientific discourse labels.

[3] https://github.com/UCDenver-ccp/ccp-nlp-pipelines

[4] OBO NCBI taxonomy: http://www.obofoundry.org/ontology/ncbitaxon.html

[5] Using: https://pypi.org/project/pyarc/

Rule	Sup	Conf
method=1,title=1	0.21	1.00
title=1,result=1,goal=0	0.21	1.00
implication=1	0.17	1.00
title=1,result=1,fact=0	0.14	1.00
method=1,fact=1	0.10	1.00
title=1,fact=0,goal=1	0.10	1.00
title=1,fact=0	0.34	0.95
fact=1,goal=0	0.26	0.94
method=0,result=1,goal=0	0.24	0.93

Table 1: CAR M1 rules predicting that the pathogen is a focus entity. A value of 1 indicates that the pathogen appears in the corresponding discourse segment, while 0 indicates that the pathogen is absent from that type of segment. Rules have been selected and sorted based on the confidence (Conf) and support (Sup) values.

		Background		Focus	
Label	S.	Freq	%	Freq	%
METHOD	73	3	33.33	17	34.69
RESULT	186	4	44.44	29	59.18
FACT	51	2	22.22	21	42.86
IMPLICATION	44	0	0.00	10	20.41
GOAL	25	3	33.33	15	30.61
PROBLEM	8	0	0.00	3	6.12
HYPOTHESIS	15	0	0.00	1	2.04
TITLE	52	3	33.33	39	79.59
NONE	3	0	0.00	1	0.00
Pathogens	-	9	100.00	49	100.0

Table 2: Frequency (Freq) of the mentions of background and focus entities in various discourse segments of PubMed citations. The percentages indicate the proportion of pathogen mentions of each type occurring in each scientific discourse segment. "S." stands for the overall number of sentences per type in the 52 citations.

The scientific discourse tagger obtained an F1 of 0.841 on the SciDT dataset. They also added NONE label to allow for *none of the above*. We apply the scientific discourse tagger to assign one of the eight discourse labels to each sentence in an abstract. The TITLE label was assigned using the available citation metadata.

5 Results

We ran the CAR M1 algorithm on our data set annotated with pathogen mentions and present the inferred rules in Table 1. There are 9 rules that predict focus entities. The first rule means that if the pathogen is mentioned in the METHOD and TITLE segments of the citation, then it is a focus pathogen. The second rule means that if the pathogen is mentioned in the TITLE and RESULT segment but not in the GOAL segment, then it is a focus pathogen.

Doing an analysis of the rules, we find that most of the rules indicate that a pathogen being mentioned in the title is a sign that it is a focus pathogen, which is expected since the title denotes the most important concepts of the article. The rules also indicate that a mention of a pathogen in the results segment is relevant to the classification of the entity as a focus pathogen. Consideration of combinations of segments is more effective to identify focus entities than occurrence in any individual segment, apart from IMPLICATION.

Table 2 shows the frequency of pathogen mentions in the various discourse segments. We find that the focus pathogens are significantly more prevalent in the TITLE, RESULT and FACT segments, which correlates with the predicates of the inferred rules. Background pathogens seem to be equally prevalent in both the METHOD and GOAL segments when compared to the focus pathogens. Some of the labels, such as HYPOTHESIS, PROBLEM and NONE, have low frequency in our data set and did not participate in any of the generated rules.

6 Conclusion

We have proposed an approach to the problem of detecting focus versus ground entities using class association rules over entity mentions in discourse segments, specifically examining its use for pathogen characterisation. Focus pathogens tend to appear in the title and results segments of abstracts, where the key findings of research are highlighted. Our case study suggests that discourse information provides valuable cues to identify focus pathogens.

Given the small-scale data we have available, this work is only indicative of the promise of the approach. We are developing a larger data set, which will support comprehensive exploration of more refined rules. This data set would also support the exploration of additional existing methods, such as centrality and transformer based methods.

Acknowledgments

We acknowledge the funding support of the US Army International Pacific Centre, and the support of the US Defence Threat Reduction Agency Biological Materials Information Project team. We also thank Dr. Leyla Roohi for her work on an early version of this paper.

References

Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer.

Saber A Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, Marius Doornenbal, Michelle Gregory, and Jan A Kors. 2019. Automatic identification of relevant chemical compounds from patents. *Database*, 2019. Baz001.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Branimir Boguraev and Christopher Kennedy. 1999. Salience-based content characterisation of text documents. *Advances in automatic text summarization*, pages 99–110.

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. 2019. Bacteria biotope at BioNLP open shared tasks 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 121–131, Hong Kong, China. Association for Computational Linguistics.

Pradeep Dasigi, Gully APC Burns, Eduard Hovy, and Anita de Waard. 2017. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *arXiv preprint arXiv:1702.05398*.

Jesse Dunietz and Daniel Gillick. 2014. A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209, Gothenburg, Sweden. Association for Computational Linguistics.

Scott Federhen. 2012. The NCBI taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):59.

Frith Jarrad, Samantha Low-Choy, and Kerrie Mengersen. 2015. *Biosecurity surveillance: quantitative approaches*. Cabi.

Antonio Jimeno Yepes, James Mork, and Alan Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 102–110, Sofia, Bulgaria. Association for Computational Linguistics.

Xiangci Li, Gully Burns, and Nanyun Peng. 2019. Discourse tagging for scientific evidence extraction. *arXiv preprint arXiv:1909.04758*.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.

Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 80–86. AAAI Press.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(2):1–19.

Tasnia Tahsin, Davy Weissenbacher, Karen O'Connor, Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. GeoBoost: Accelerating research involving the geospatial metadata of virus GenBank records. *Bioinformatics*, 34(9):1606–1608.

Michael Tanenblatt, Anni Coden, and Igor Sominsky. 2010. The ConceptMapper approach to named entity recognition. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Anita de Waard and Henk pan der Maat. 2012. Verb form indicates discourse segment type in biological research papers: Experimental evidence. *Journal of English for academic purposes*, 11(4):357–366.

Joshi Prince Walker and Marilyn I Walker. 1998. *Centering theory in discourse*. Oxford University Press.