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

Antonio Jimeno Yepes, Ameer Albahem and Karin Verspoor
School of Computing and Information Systems, The University of Melbourne, Australia
School of Computing Technologies, RMIT University, Australia
{antonio.jimeno,ameer.albahem}@gmail.com, karin.verspoor@rmit.edu.au

Introduction

- Global monitoring of repositories of potentially harmful biological materials is an important component of ensuring the health and safety of our populations.
- We are building an information extraction system to identify information related to:
  - experimentation with potentially dangerous biological pathogens
  - detect facilities that may serve as repositories of harmful pathogens
- Differentiating mentions of actively studied organisms in the scientific literature from other, background or incidental mentions of organisms poses a deeper natural language processing challenge.
- We 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.

Datasets

We constructed a dataset based on information obtained from the Biological Material Information Program (BMIP) of the Defense Threat Reduction Agency (DTRA).

- **Pathogen entity list**: pathogens provided by BMIP, it contains taxonomic pathogens, mapped to the NCBI Taxonomy and prion proteins and toxins, which were assigned a custom id.
- **Gold standard dataset**: Manual annotations of relevant pathogens over PubMed citations.
  - *Relevance* is defined as evidence of an actively studied pathogen.
  - This set contains 87 PubMed citations 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, 32 unique pathogens, iacross remaining 52 citations.
  - Most pathogens belong to the *Influenza* virus family.

Methods

- **Pathogen identification**
  - We aim to find all pathogens mentioned in a PubMed citation.
  - We apply a dictionary method using a dictionary derived from the BMIP list of relevant pathogens, mapped to the NCBI Taxonomy.
  - Toxin mentions are identified using regular expressions. This has higher recall than dictionary matching.
  - Using the identification method, we detect 58 pathogens, of which 49 are focus entities (i.e. they match our manual annotation) and 9 mentions that we treat as background entities.
- **Pathogen characterisation**
  - Given the list of pathogens identified in a citation, the next step is to characterise which of these pathogens are focus entities, i.e. actively researched.
  - Association rule classification CAR M1 (Liu et al., 1998) was used to infer rules that predict that the pathogen is a focus entity using discourse labels.
- **Discourse segment labelling**
  - We use the scientific discourse tagger (Dasigi et al., 2017), 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.

Related Work

- **Identification of salient entities**: The study of discourse structure has been suggested in previous work on entity salience (Boguraev andKennedy, 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. This work does not address saliency of entity mentions.

Results

		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 1. 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.

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 2. 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.

Conclusions

- 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.
- We are developing a larger data set, which will support comprehensive exploration of more refined rules.