Distinguishing between focus and background entities in biomedical corpora using discourse structure and transformers

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

[1]School of Computing Technologies, RMIT University
[2]School of Computing and Information systems, The University of Melbourne
Melbourne, VIC, Australia
[1]{antonio.jose.jimeno.yepes,karin.verspoor}@rmit.edu.au

Abstract

Scientific documents typically contain numerous entity mentions, while only a subset are directly relevant to the key contributions of the paper. Distinguishing these focus entities from background ones effectively could improve the recovery of relevant documents and the extraction of information from documents. To study the identification of focus entities, we developed two large datasets of disease-causing biological pathogens using MEDLINE, the largest collection of biomedical citations, and PubMed Central, a collection of full text articles. The focus entities were identified using human-curated indexing on these collections. Experiments with machine learning methods to identify focus entities show that transformer methods achieve high precision and recall and that document discourse information is relevant. The work lays the foundation for more targeted retrieval/summarisation of entity-relevant documents.

1 Introduction

Scientific documents typically discuss one or more topics linked to key entities of interest. However, entities may also be mentioned incidentally to support argumentation, in discussing related work, or be used in comparison with focus entities of direct interest. Distinguishing between these focus and background entities might improve the selection of information most relevant to a user.

The automatic identification of entities in text is typically achieved using named entity recognition or entity linking methods based on dictionary, rule-based and/or machine learning methods, and aims to identify *all* mentions of entities of the target type(s). However, not all entities correctly identified in a text may be entities relevant for further processing or important to the main conclusions of a document. For example, it has been suggested that only ~10% of chemical mentions play a major role within a chemical patent (Akhondi et al., 2019). Strategies for identifying entities that are in focus in a document enable honing in on critical document information, and can support filtering out entities that are ancillary to the main objectives of the work, e.g. for literature-based discovery applications (Henry and McInnes, 2017).

In this work, we introduce two large datasets annotated with focus and background entities that support experimentation with methods for distinguishing these two types of entity mentions[1]. We evaluated several machine learning algorithms on these dataset, setting baseline results for future work to be done on this task, and laying the foundation for more nuanced treatment of document entities in document retrieval or in summarisation.

2 Related work

Entity salience, relevant to identifying focus entities, has been discussed in previous work. Use 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. One hypothesis is that the focus and background entities are distributed in specific argumentative sections of a document (Ruch et al., 2007; Jimeno Yepes et al., 2021).

The identification of focus entities has multiple relevant applications. In information retrieval (IR), the objective is to recover documents that are relevant to the user information needs, which is challenging for long documents (Webber et al., 2012) as a larger number of entities are being mentioned. In information extraction (IE), we find the task of named entity recognition (NER), in which the objective is to identify entities of interest, from people and locations to proteins and genes, depending on the domain. In NER, all entities of a certain type

[1]https://zenodo.org/record/5866759

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 35 - 40

are identified, even the ones that are not the main focus (Dunietz and Gillick, 2014).

Our study relates specifically to identification of biological pathogen entities in scientific literature. Pathogen NER has been studied in the Bacteria Biotope shared task (Bossy et al., 2019). The GeoBoost tool (Tahsin et al., 2018) addresses the identification of entities from the gene database GenBank (Benson et al., 2012) and largely includes information about viruses and bacteria.

3 Datasets

Development of large corpora is costly since human annotation is slow and expensive. There are biomedical datasets that have been manually annotated and could be considered as proxy for manual annotation. For this work, we have developed two large corpora automatically using existing resources from the National Center for Biotechnology Information (NCBI) at the NLM. The corpora are targeted to microbial pathogens, some of the most relevant entities for infectious diseases (Balloux and van Dorp, 2017), such as COVID-19.

3.1 MEDLINE citation dataset

MEDLINE[2] is the largest biomedical citation database with over 30 million citations from more than 5,000 journals. MEDLINE is indexed semi-manually (Mork et al., 2013) with the MeSH (Medical Subject Headings) controlled vocabulary[3], providing a resource to identify focus entities in biomedical articles. To identify the pathogens in MEDLINE, we created a dictionary of pathogens and collected MEDLINE citations that indexed these pathogens. MeSH contains 360 of the 2.8k pathogens of interest in our work, which constitutes our focus entities. We applied a dictionary-based approach using ConceptMapper (Tanenblatt et al., 2010; Funk et al., 2014) with evaluation available from Jimeno Yepes and Verspoor (2022).

With the list of PubMed identifiers (PMIDs) obtained using MeSH indexing, we recovered their citations from MEDLINE and annotated the text with the dictionary. Overlapping mentions of the same entity were removed and removed pathogen mentions that could not be identified in MeSH. From the set of selected pathogens identified in the citations, the ones that appeared in the MeSH indexing of the citation were considered focus entities, while the pathogens not mentioned in the indexing were consider background entities. We considered both major and minor MeSH headings. For each pathogen identified in a citation, all of its mentions in text were changed to the string @PATHOGEN$. Table 1 presents the corpus statistics, divided into $2/3$ for training $1/3$ for testing.

3.2 PubMed Central full text dataset

In addition to MEDLINE citations, we also consider full text articles from PubMed Central (Roberts, 2001), a collection of full text articles made available from the NLM. To collect the full text articles from PubMed Central, we used the PMIDs obtained using MeSH indexing and mapped these identifiers to PubMed Central identifiers (PMCIDs). We applied the same methodology to highlight the mentions of a specific pathogen as with the MEDLINE citations. Statistics of the full text collection are available in table 1.

MEDLINE dataset	Training	Testing
Unique citations	622,447	320,318
With more than one pathogen	136,546	70,670
Focus entities	661,470	340,991
Background entities	160,540	82,470
Document avg entities	1.3206	1.3220
Document avg focus entities	1.1250	1.1268
Full text dataset	Training	Testing
Unique articles	79,352	39,677
With more than one pathogen	53,003	26,551
Focus entities	82,922	41,602
Background entities	157,072	78,148
Document avg entities	3.0244	3.0181
Document avg focus entities	1.0450	1.0485

Table 1: Frequency of example documents and statistics on focus and background pathogen entities in MEDLINE and full text datasets.

Full text articles are already divided into discourse sections. We process these sections in two ways, first by concatenating the text in the article following the order in the PMC XML file, in which each section is prefixed by the name of the section starting with the character "@ "and ending with ":", e.g. "@title:". Second, we keep each information in a separate section, which allows only considering text in a specific section and can be used with learning algorithms that leverage this organization. Table 2 shows entities distribution in full text.

[2] https://www.nlm.nih.gov/medline/medline_overview.html

[3] https://www.ncbi.nlm.nih.gov/mesh

Background	Count	Focus	Count
introduction	53,574	abstract	68,098
discussion	53,486	title	46,971
results	37,768	introduction	44,466
abstract	19,860	results	27,177
methods	18,674	discussion	21,313
background	11,483	methods	11,813
title	5,789	background	11,637
conclusions	3,526	conclusions	6,155
the study	969	the study	785
case layout	745	abbreviations	705
all	157,072	all	82,922

Table 2: Frequency of background and focus entities in training full text sections

4 Methods

4.1 Baseline methods

We consider two baselines. The first baseline selects a single focus entity per document on the basis of frequency. We utilised the inverted document frequency of entity mentions to evaluate if frequent entities in the collection should be discounted. The second baseline annotates all entities mentioned as focus entities.

4.2 Bag-of-words entity categorization

In our work, focus entities are identified at the document level. In a sense, we would be categorising the mentions of the entity within a citation as focus or background. In our datasets, the entity of interest has been renamed to @PATHOGEN$.

We trained a linear Support Vector Machine (SVM) (Vapnik, 2013) with modified Huber loss (Zhang, 2004) suited for imbalanced data and AdaBoostM1 (Freund and Schapire, 1997), (both from the MTIMLExtension package[4] optimised for large datasets and using uni-grams and bi-grams) and FastText (Joulin et al., 2017)[5], using default parameters as well for classification.

4.3 Transformer based methods

Focus entities might appear in specific contexts in comparison to background entities. Bag-of-words methods have a limited coverage of the context in which these entities might appear. Recent advances in deep learning have delivered self-attention methods that have led to the Transformer architecture (Vaswani et al., 2017).

BERT (Devlin et al., 2019) is a transformer based method that encodes the input tokens into contextualised embeddings trained on large corpora. Classification is achieved using the output from BERT, pooled on the [CLS] character, and a fully connected layer to predict if an entity is a focus or background one.

BERT supports a maximum size of 512 tokens, while other methods developed using the BERT architecture, such as the Longformer (Beltagy et al., 2020), allow for longer documents. Longformer achieves this by using a sliding window instead of attending to all tokens and by using a global attention mask which we set to the [CLS] token used in text categorisation settings.

Our MEDLINE corpus has an average of 308 tokens per document, with just a 6% of the citations with length above 512 tokens. We have used the SciBERT (Beltagy et al., 2019) pre-trained model[6], truncating documents at 512 tokens. When using Longformer, we considered a maximum document length of 1,250 tokens due to memory limitations. Transformer methods were trained using 80% of the training set for training purposes and 20% as validation set. We used Adam (Kingma and Ba, 2015) with an initial learning rate of 2e-5 for 30 epochs. The model with best performance on the validation set after each epoch was selected.

4.4 Scientific discourse focus entity selection

Scientific articles follow a discourse structure, with information organised into different rhetorical sections. The mention of an entity in a certain section can indicate the relevance of that entity in the document. Only a small number of MEDLINE citations have an explicit discourse structure (Ripple et al., 2011). Hence, we apply a discourse tagger (Li et al., 2021) to annotate sentences of a citation relevant to a discourse section, except to the *title* which is explicitly marked in the metadata. Table 3 shows the frequency of each of the categories.

5 Results

Table 4 shows the results of using the different methods. We observe that the baseline based on classifying all entities identified by our dictionary method as focus entities has maximum recall and already has a high precision. The most-frequent mentioned entity baseline has better precision, with de-

[4] https://github.com/READ-BioMed/MTIMLExtension

[5] https://fasttext.cc

[6] We have used Huggingface's (Wolf et al., 2020) implementations of tranformer methods.

Category	Background	Focus
fact	33,044	139,290
goal	9,295	56,100
hypothesis	7,544	21,433
implication	14,077	42,828
method	44,132	203,225
none	1,026	4,452
problem	2,858	11,429
result	61,317	181,691
title	44,884	435,100
all	160,540	661,470

Table 3: Frequency of each discourse category in the training MEDLINE dataset

Average	Prec.	Recall	F1
All-focus entities	0.3474	**1.0000**	0.5157
tf baseline	0.7475	0.7078	0.7271
tf-idf baseline	0.7587	0.7184	0.7380
SVM-tag	0.8110	0.6440	0.7179
SVM-all	0.6525	0.7761	0.7090
AdaBoostM1-tag	0.8447	0.8824	0.8631
AdaBoostM1-all	0.7845	0.7580	0.7710
fastText	0.8557	0.7374	0.7922
SciBERT	0.9115	**0.9314**	0.9213
Longformer	**0.9410**	0.9269	**0.9339**

Table 5: Focus entity prediction in PubMed Central. The *All-focus* baseline trivially has perfect Recall.

creased recall. Considering the learning algorithms, SciBERT and Longformer perform better than the bag-of-words algorithms, which is expected since these algorithms do not consider the context of the pathogen mention, even with bigrams. The two deep learning algorithms have similar performance.

Average	Prec.	Recall	F1
All-focus entities	0.8052	**1.0000**	0.8921
tf baseline	0.9047	0.8508	0.8770
tf-idf baseline	0.8838	0.8311	0.8566
SVM	0.8975	0.9450	0.9206
AdaBoostM1	0.8654	**0.9682**	0.9139
fastText	0.8608	0.9572	0.9064
SciBERT	**0.9359**	0.9631	**0.9493**
Longformer	0.9285	0.9679	0.9478

Table 4: Focus entity prediction results on MEDLINE. The *All-focus* baseline trivially has perfect Recall.

Table 5 shows the result of the learning algorithms on the full text dataset. Compared to the MEDLINE corpus, we identify that the baseline methods suffer a substantial drop in performance. This is expected since there are more background entities in the full texts, and the most frequent entity is not always in focus. Bag-of-words methods have a lower performance as well, AdaBoostM1 with tag related words outperforms the other methods, indicating the effectiveness of linking words to article sections. In this set, documents are longer and longformer improves over the SciBERT model, which has a limit of 512 tokens.

6 Discussion

The datasets we have constructed for the identification of focus entities are large, supporting evaluation of a variety of methods and comparison of performance in both short and large documents.

Full text is more challenging compared to citations, consistent with findings on other tasks (Cohen et al., 2010), and mostly due to the higher proportion of focus entities in citations. Machine learning approaches based on bags-of-words tend to improve over simple baseline methods but underperform transformer methods.

The distribution of entities in article sections (table 2) and prediction results in full text (table 5) show that the discourse sections in which entities appear are relevant for the identification of focus entities in scientific articles.

7 Conclusions and future work

We have developed two large datasets of scientific documents for the study of the identification of focus entities. We find that short documents, represented by MEDLINE citations, are easier to process than longer (full-text) documents. Transformer methods showed higher performance.

Future work will address using the proposed methods in scenarios in which focus entities become relevant, and comparing our approach with other existing methods (Lu and Choi, 2021; Dunietz and Gillick, 2014).

8 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. Experiments were done using the LIEF HPC-GPGPU Facility, supported by LIEF Grant LE170100200, at the University of Melbourne.

References

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.

Francois Balloux and Lucy van Dorp. 2017. Q&a: What are pathogens, and what have they done to and for us? *BMC Biology*, 15(1):1–6.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2012. Genbank. *Nucleic acids research*, 41(D1):D36–D42.

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.

K Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

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):1–29.

Sam Henry and Bridget T. McInnes. 2017. Literature based discovery: Models, methods, and trends. *Journal of Biomedical Informatics*, 74:20–32.

Antonio Jimeno Yepes, Ameer Albahem, and Karin Verspoor. 2021. Using discourse structure of scientific literature to differentiate focusfrom background entities in pathogen characterisation. In *Australasian Language Technology Association*.

Antonio Jimeno Yepes and Karin Verspoor. 2022. Classifying literature mentions of biological pathogens as experimentally studied using natural language processing. *Journal of Biomedical Semantics*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (Poster volume)*.

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.

Jiaying Lu and Jinho D Choi. 2021. Evaluation of unsupervised entity and event salience estimation. In *The International FLAIRS Conference Proceedings*, volume 34.

James G Mork, Antonio Jimeno-Yepes, Alan R Aronson, et al. 2013. The nlm medical text indexer system for indexing biomedical literature. *BioASQ@ CLEF*, 1.

Anna M Ripple, James G Mork, Lou S Knecht, and Betsy L Humphreys. 2011. A retrospective cohort study of structured abstracts in medline, 1992–2006. *Journal of the Medical Library Association: JMLA*, 99(2):160.

Richard J Roberts. 2001. Pubmed central: The genbank of the published literature. *Proceedings of the National Academy of Sciences*, 98(2):381–382.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2-3):195–200.

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 LREC'10*.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer Science & Business Media.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 5998–6008.

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

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 116.