A Survey of Information Extraction in Biology

Haniya Ali
University of California, Berkeley

Abstract

Information Extraction (IE) is a natural language processing (NLP) concept which refers to the task of retrieving information regarding a topic from a given text. IE turns unstructured information in texts into structured data. In the past couple of decades, we have observed an exponential increase in the quantity of scientific literature in the biological domain. Although these texts contain useful insights that can help further biomedical research, the process of individually reading and extracting relevant information from each document is an incredibly laborious task. To solve this issue, many efforts have been taken to build IE methods specifically for biological texts. As such, this paper will present an overview of some of the major challenges faced and tools built to use the process of IE in the biological domain.

1 Introduction

The turn of the 21st century marked the completion of the Human Genome Project (HGP) which allowed scientists to sequence and read all the genes in the human body. The impact of the findings of the HPG were tremendous as it led to the emergence of new research areas which in turn resulted in an exponential increase in the quantity of scientific literature. To manage and process these, various efforts have been taken to employ the NLP technique of IE to read and extract information from biological text. IE is the process wherein a database is populated from unstructured or loosely structured text (McCallum, 2005). While this concept has been used in various domains, its application in biology remains a continuous work in progress due to the complex nature of biological documents. By highlighting some of the challenges and discussing their potential solutions, this paper will present a survey of how the process of IE has been used in the biological domain.

2 Initial Endeavors

Since the HGP was completed in 2003, many research efforts prior to this were focused on curating databases on enzymes and bacteria. Many of these models, however, were not generalizable as they were either based on dictionaries or arbitrary handwritten rules. The former is typically the “first step in extracting information from biomedical documents because it can provide ID information on recognized terms” (Tsuruoka and Tsujii, 2004). But since dictionaries were susceptible to spelling issues, there was growing trend to use a corpus-based, machine-learning approach. To achieve this an ontology of the biological domain was created in which substances were classified according to their chemical characteristics (Tateisi et al., 2000). Since ontologies represent the relationship between exactly two concepts or entities, this classification allowed the representation of a substance beyond its biological role. This made the annotation task more complicated as the introduction of this type of nested structure which gave importance to a substance’s chemical and biological function allowed for a more robust classification mechanism. Although this was a creative decision, it fell victim to the most significant challenge posed by biological corpora - the naming conventions. Unlike other written documents, biological abstracts contain various non-proper names that may begin with capital letters such as CD4 or RelB, chemical and numeric expressions that may include various types of punctuations, for example, beta-(1,3)-glucan, etc. (Tateisi and Tsujii, 2004). Additionally, there are various terms that require domain specific knowledge that can make

2166 words

annotation a challenging task involving a high degree of error. These issues led to efforts to create resources that consolidated biologically relevant terms and definitions.

2.1 Consolidating Biological Terms

Some of the most initial efforts to build an annotated corpora for biology included the use of a part-of-speech (POS) tagged corpus. While POS describes how words are used in a sentence, a POS tagged corpus would contain a list of all terms and how they may occur in a sentence (D'Souza, 2018). These efforts used the GENIA corpus which is a collection of biomedical literature that contains several annotations of biological terms. By assigning POS to each word, the researchers "made use of the existing term annotation of the GENIA corpus to annotate the POS to the constituents of technical terms" (Tateisi and Tsujii, 2004). Although this model did not require much domain specific knowledge, to handle abbreviations, non-proper nouns, etc it highlighted that a "more intelligent preprocessor using exhaustive dictionary might be necessary" (Tateisi and Tsujii, 2004).

Since manually creating dictionaries had already proven to be unsuccessful, more creative efforts had to be taken to achieve this task. As such a subfield of named entity recognition called biological entity recognition (BER) gained popularity in which concepts of interest in biological texts are extracted by mapping relevant words to a set of predefined categories (Hem, 2008). Liu (2005) demonstrated that biological entity tagging can create a system to automatically generate a protein entity dictionary. To do so, online resources such as publicly available protein databases were used to find repetitions of terms and definitions. These findings were then computationally curated in a raw dictionary called the BioThesauras. While the effort to generate such a dictionary automatically was pivotal, these online databases often lacked annotations with regards to various protein characteristics such as subcellular localization or function (Fyshe and Szafron, 2006). So, the information in the BioThesauras only mapped protein names to their entries. Proteins generally have a language of their own (Devi et al., 2017) and so identifying the complex relationships between proteins with the amino acids that make up those proteins further complicated the task of unifying the various scientific literature.

Simultaneously, there work was also being done with genetic information where the occurrences of genes were identified and then normalized by text matching with dictionaries (Fang et al., 2006). Normalization in IE refers to putting information in a standard format such that it can be reliably compared (McCallum, 2005). These efforts to use genetic and proteomic data to generate dictionaries, however, could not accurately capture the diversity of the various biological terms and definitions. Since normalization is an important subtask of IE, the application of this step on biological text would often result in the exclusion of important information. Likewise, while successful efforts were made to extract information about the bacterial genome and create automated dictionaries (Deléger et al., 2016), the issue in normalizing proteomic, genetic, and bacterial data consistently produced insufficient inter-annotator agreement values (Hahn et al., 2008).

This changed in 2008, however, when the BioLexicon was implemented. Being one the most significant breakthroughs in consolidating the various biological terms, the BioLexicon was "a large-scale lexical-terminological resource encoding different information types in one single integrated resource" (Quochi et al., 2008). Since previous efforts focused on creating isolated dictionaries and ontologies, the introduction of the BioLexicon consolidated different terms and their variants of form and of meaning by automatically extracting information from literature. This availability of vast amounts of text on proteins and the genes that code them triggered newer areas of research that were previously never explored. For example, comprehensive dictionaries for viral species could be generated as information on proteins that compose them was more readily available (Cook et al., 2017).

3 State of the Art

After the task of creating a resource for biological terms had been mainly resolved, much of the recent efforts in using the process of IE in biology focus on extracting more complicated relationships from text such as events. A biological event refers to a biochemical process, e.g., a protein-protein interaction or a chemical protein interaction, within a signaling pathway or a metabolic pathway (Li et al., 2015). While previous endeavors successfully identified either binary relations or simple occurrences, these findings did not present a complete picture of the complex interactions between biological components (Bui and Sloot, 2011).

Some of the initial tools developed for event extraction can be categorized into two major categories, a rule-based system, and a machine learning (ML) approach. In the latter implementation, a system "performs logical inference over the semantic structures by using handcrafted inference rules and extracts target information from the results of the inference" (Hahn et al., 2009). These rules are created by utilizing domain specific information. The ML approaches require more syntactic processing steps such as POS-tagging, chunking, etc. Experiments from these two different types of experiments revealed that the F-scores for rule-based and ML-based system were 34 % and 19% respectively (Hahn et al., 2009). Despite these figures, however, an ML-based system is more generalizable. As such it is important to discuss some crucial ML-based models that have been created in recent years.

Reranking

Some more complex ML tools to extract events from biological text include the use of a reranking architecture to "incorporate truly global features to the model of named entity tagging" (Yoshida and Tsujii, 2007). In NLP, reranking refers to the generation of N-best candidates which are ranked by using local and global features (Shen and Joshi, 2005). Since biological text typically contains very long names, reranking successfully allows for a more careful extraction of these terms by creating stricter boundaries.

Concept Recognizer

Other tools in the field of event extraction approach the problem as one of concept recognition and analysis. Cohen (2009) explains, "concept recognition can be equivalent to the named entity recognition task when it is limited to locating mentions of particular semantic types in text". By employing readily available ontologies in the biomedical domain, the use of concept recognition successfully captured both the mentions of the events as well as their triggers.

Knowledge Driven Tree

Despite the creativity of these tools, however, they fail to capture the more complicated instances of events in biological text. As such, Li (2019) proposed a tree structure based long short-term memory (TreeLSTM) network. The tree structure creates more complicated relationships between each trigger and its corresponding event.

Conditional Random Fields (CRFs)

There have also been efforts to create more robust and scalable tools for event extraction such at the Conditional Random Fields (CRFs) proposed by Rao (2017). Compared to the other models that have been introduced for event extraction, this model is more robust in that it scans text for trigger words and only then identifies the proceeding event.

3.1 Other Forms of Extraction

As biological events do not appear in isolation within an organism, various efforts are underway to extract biological processes which are a series of events from a given text. These endeavors are necessary as they can help answer non-factual questions. For example, while event extraction is limited to answering where an event occurs in a text, process extraction can answer how a complicated event takes place by looking at all the various upstream and downstream steps. The models that answer this do not consider single words or sentences but rather look at associations between multiple sentences (Scaria et al., 2013). An example of such model is the clustering-based inference model which utilizes different linking techniques to enable joint entity linking predictions (Angell et al., 2021); this model does not require any domain specific knowledge.

On a similar note, the identification of biological pathways is also very crucial in biomedical research. So, the research in event and process extraction has resulted in the creation of software tools like CellDesigner that extract NLP event representations and convert them to standard pathway representations (Spranger et al., 2015).

3.2 Datasets and Databases

While several tools have been implemented for extracting various types of information from biological texts, there is a lack of high-quality benchmark datasets which can be used for the robust comparative evaluation of existing approaches. Khachatrian (2019) proposed the BioRelEx which is a new dataset that contains fully annotated sentences from biomedical literature. Several baselines have been evaluated using this dataset to create more accurate methods for event extraction.

Apart from these datasets, multi-modal protein-protein interaction databases have also been curated by manually annotating biomedical archives (Dutta and Saha, 2020). This is a more robust collection of protein entities and interactions compared to the BioLexicon resource.

4 Future areas of research

As there are several models that are being developed independently to achieve various biological tasks, efforts are underway to build tools that can either merge these models or find similarities between them. One such approach is the Mixture-of-Partitions models which can take as input very large knowledge graphs and infuse their knowledge into various BERT models (Meng et al., 2021). But since BERT is a bidirectional encoder that conditions on both left and right context (analyticsvidhya, 2020), this model often fails to work on biological text.

There is also an active area of research that seeks to capture more domain-specific semantics that are more generalizable for various types of biological corpora (Fivez et al., 2021). These attempts focus on using neural architectures such as a Deep Averaging Network which is a simple deep neural network that works very effectively on datasets with high syntactic variance, a common challenge faced in the biological domain (Iyyer et al., 2015).

5 Conclusion

The extraction of relevant information from biological documents is crucial for breakthroughs in biomedical research. As such, this paper focused on the major challenges and milestones in applying IE in the biological domain.

References

Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany Harding, and Christopher D. Manning. 2013. Learning Biological Processes with Global Constraints. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1710–1720, Seattle, Washington, USA. Association for Computational Linguistics.

Alona Fyshe and Duane Szafron. 2006. Term Generalization and Synonym Resolution for Biological Abstracts: Using the Gene Ontology for Subcellular Localization Prediction. In Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, pages 17–24, New York, New York. Association for Computational Linguistics.

Andrew McCallum. ACM Queue, volume 3, Number 9, November 2005.

Anon. 2020. What is Bert: Bert for text classification. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Anon. 2020. What is Bert: Bert for text classification. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/

Chen Li, Runqing Song, Maria Liakata, Andreas Vlachos, Stephanie Seneff, and Xiangrong Zhang. 2015. Using word embedding for bio-event extraction. In Proceedings of BioNLP 15, pages 121–126, Beijing, China. Association for Computational Linguistics.

Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. 2019. Biomedical Event Extraction based on Knowledge-driven Tree-LSTM. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

2166 words

Technologies, Volume 1 (Long and Short Papers), pages 1421–1430, Minneapolis, Minnesota. Association for Computational Linguistics.

G.Devi, Ashish V. Tendulkar, Sutanu Chakraborti. 2017. Protein Word Detection using Text Segmentation Techniques. In Proceedings of the BioNLP 2017 workshop, pages 238–246, Vancouver, Canada. Association for Computational Linguistics.

Gang Li, Cathy Wu, and K. Vijay-Shanker. 2017. Noise Reduction Methods for Distantly Supervised Biomedical Relation Extraction. In BioNLP 2017, pages 184–193, Vancouver, Canada,. Association for Computational Linguistics.

Haw-ren Fang, Kevin Murphy, Yang Jin, Jessica Kim, and Peter White. 2006. Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries. In Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, pages 41–48, New York, New York. Association for Computational Linguistics.

Helen Cook, Rūdolfs Bērziņš, Cristina Leal Rodrıguez, Juan Miguel Cejuela, and Lars Juhl Jensen. 2017. Creation and evaluation of a dictionary-based tagger for virus species and proteins. In BioNLP 2017, pages 91–98, Vancouver, Canada,. Association for Computational Linguistics.

Hrant Khachatrian, Lilit Nersisyan, Karen Hambardzumyan, Tigran Galstyan, Anna Hakobyan, Arsen Arakelyan, Andrey Rzhetsky, and Aram Galstyan. 2019. BioRelEx 1.0: Biological Relation Extraction Benchmark. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 176–190, Florence, Italy. Association for Computational Linguistics.

Hongfang Liu, Zhangzhi Hu, and Cathy Wu. 2005. Dynamically Generating a Protein Entity Dictionary Using Online Resources. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 17–20, Ann Arbor, Michigan. Association for Computational Linguistics.

K. Bretonnel Cohen, Karin Verspoor, Helen Johnson, Chris Roeder, Philip Ogren, William Baumgartner, Elizabeth White, and Lawrence Hunter. 2009. High-precision biological event extraction with a concept recognizer. In Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, pages 50–58, Boulder, Colorado. Association for Computational Linguistics.

Jocelyn D'Souza. 2018. Learning POS Tagging & Chunking in NLP. Medium.

Kazuhiro Yoshida and Jun'ichi Tsujii. 2007. Reranking for Biomedical Named-Entity Recognition. In Biological, translational, and clinical language processing, pages 209–216, Prague, Czech Republic. Association for Computational Linguistics.

Libin Shen and Aravind K. Joshi. 2005. Ranking and Reranking with Perceptron. Department of Computer and Information Science, University of Pennsylvania.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In Proceedings of the 4th BioNLP Shared Task Workshop, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Michael Spranger, Sucheendra K. Palaniappan, Samik Ghosh. 2015. Extracting Biological Pathway Models From NLP Event Representations. In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 42–51, Beijing, China. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Pattabhi RK Rao, Sindhuja Gopalan, and Sobha Lalitha Devi. 2017. Scalable Bio-

2166 words

Molecular Event Extraction System towards Knowledge Acquisition. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 383–391, Kolkata, India. NLP Association of India.

Pratik Dutta and Sriparna Saha. 2020. Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6396–6407, Online. Association for Computational Linguistics.

Pieter Fivez, Simon Suster, and Walter Daelemans. 2021. Conceptual Grounding Constraints for Truly Robust Biomedical Name Representations. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2440–2450, Online. Association for Computational Linguistics.

Quoc-Chinh Bui, Peter M.A. Sloot. 2011. Extracting Biological Events from Text Using Simple Syntactic Patterns. In Proceedings of BioNLP Shared Task 2011 Workshop, pages 143–146, Portland, Oregon. Association for Computational Linguistics

Rico Angell , Nicholas Monath , Sunil Mohan , Nishant Yadav , and Andrew McCallum. 2021. Clustering-based Inference for Biomedical Entity Linking. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2598–2608. Association for Computational Linguistics.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Michael Poprat, Katrin Tomanek, and Joachim Wermter. 2008. Semantic Annotations for Biology: a Corpus Development Initiative at the Jena University Language & Information Engineering (JULIE) Lab. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Udo Hahn, Katrin Tomanek, Ekaterina Buyko, Jung-jae Kim, and Dietrich Rebholz-Schuhmann. 2009. How Feasible and Robust is the Automatic Extraction of Gene Regulation Events? A Cross-Method Evaluation under Lab and Real-Life Conditions. In Proceedings of the BioNLP 2009 Workshop, pages 37–45, Boulder, Colorado. Association for Computational Linguistics.

Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. 2008. A lexicon for biology and bioinformatics: the BOOTStrep experience. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Ying Hem, Mehmet Kayaalp. 2008. Biological Entity Recognition with Conditional Random Fields. National Library of Medicine.

Yoshimasa Tsuruoka, Jun'ichi Tsujii. 2004. Improving the performance of dictionary-based approaches in protein name recognition. In Journal of Biomedical Informatics, pages 461-470. ScienceDirect

Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. Building an Annotated Corpus in the Molecular-Biology Domain. In Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, pages 28–34, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.

Yuka Tateisi and Jun-ichi Tsujii. 2004. Part-of-Speech Annotation of Biology Research Abstracts. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Zaiqiao Meng, Fangyu Liu, Thomas Hikaru Clark, Ehsan Shareghi, Nigel Collier. 2021. Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. In Proceedings of the 2021

2166 words

Conference on Empirical Methods in Natural Language Processing, pages 4672–4681. Association for Computational Linguistics.