EVIDENCEMINER: Textual Evidence Discovery for Life Sciences

Xuan Wang[1], **Yingjun Guan**[2], **Weili Liu**[1], **Aabhas Chauhan**[1], **Enyi Jiang**[1], **Qi Li**[3], **David Liem**[4], **Dibakar Sigdel**[4], **J. Harry Caufield**[4], **Peipei Ping**[4], **Jiawei Han**[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]School of Information Sciences, University of Illinois at Urbana-Champaign
[3]Department of Computer Science, Iowa State University
[4]School of Medicine, University of California, Los Angeles
[1,2]{xwang174,yingjun2,weilil2,aabhasc2,enyij2,hanj}@illinois.edu, [3]qli@iastate.edu,
[4]{dliem,dsigdel,jcaufield,pping}@mednet.ucla.edu

Abstract

Traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements. Some of these statements may serve as textual evidence that is key to tasks such as hypothesis generation and new finding validation. We present EVIDENCEMINER, a web-based system that lets users query a natural language statement and automatically retrieves textual evidence from a background corpora for life sciences. EVIDENCEMINER is constructed in a completely automated way without any human effort for training data annotation. It is supported by novel data-driven methods for distantly supervised named entity recognition and open information extraction. The entities and patterns are pre-computed and indexed offline to support fast online evidence retrieval. The annotation results are also highlighted in the original document for better visualization. EVIDENCEMINER also includes analytic functionalities such as the most frequent entity and relation summarization. EVIDENCEMINER can help scientists uncover essential research issues, leading to more effective research and more in-depth quantitative analysis. The system of EVIDENCEMINER is available at https://evidenceminer.firebaseapp.com/[1].

1 Introduction

Search engines on scientific literature have been widely used by life scientists for discoveries based on prior knowledge. Each day, millions of users query PubMed[2] and PubMed Central[3] (PMC) for their information needs in biomedicine (Allot et al., 2019). However, traditional search engines for life sciences (e.g., PubMed) are designed for document retrieval and do not allow direct retrieval of specific statements (Lu, 2011; Ren et al., 2017; Shen et al., 2018). With the results from those search engines, scientists still need to read a large number of retrieved documents to find specific statements as textual evidence to validate the input query. This textual evidence is key to tasks such as developing new hypotheses, designing informative experiments, or comparing and validating new findings against previous knowledge.

While the last several years have witnessed substantial growth in interests and efforts in evidence mining (Lippi and Torroni, 2016; Wachsmuth et al., 2017; Stab et al., 2018; Chen et al., 2019; Majithia et al., 2019; Chernodub et al., 2019; Allot et al., 2019), little work has been done for evidence mining system development in the scientific literature. A significant difference between evidence in the scientific literature and evidence in other corpora (e.g., the online debate corpus) is that scientific evidence usually does not have a strong sentiment (i.e., positive, negative or neutral) in the opinion it holds. Most scientific evidence sentences are objective statements reflecting how strongly they support a query statement. Therefore, if scientists are interested in finding textual evidence for "*melanoma is treated with nivolumab*", they may expect a ranked list of statements with the top ones like "bicytopenia in primary lung melanoma treated with nivolumab" as the textual evidence that supports the input query.

This paper presents EVIDENCEMINER, a web-based system for textual evidence discovery for life sciences (Figure 1). Given a query as a natural language statement, EVIDENCEMINER automatically retrieves sentence-level textual evidence from a background corpora of biomedical literature. EVIDENCEMINER is constructed in a completely automated way without any human effort for training data annotation. It is supported by

[1]A brief demo of EVIDENCEMINER is available at https://youtu.be/iYuQ6gsr--I.
[2]https://www.ncbi.nlm.nih.gov/pubmed/
[3]https://www.ncbi.nlm.nih.gov/pmc/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 56–62
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

novel data-driven methods for distantly supervised named entity recognition and open information extraction. EVIDENCEMINER relies on external knowledge bases to provide distant supervision for named entity recognition (NER) (Shang et al., 2018b; Wang et al., 2018b, 2019). Based on the entity annotation results, it automatically extracts informative meta-patterns (textual patterns containing entity types, e.g., CHEMICAL inhibit DISEASE) from sentences in the background corpora. (Jiang et al., 2017; Wang et al., 2018a; Li et al., 2018a,b). Sentences with meta-patterns that better match the query statement is more likely to be textual evidence. The entities and patterns are precomputed and indexed offline to support fast online evidence retrieval. The annotation results are also highlighted in the original document for better visualization. EVIDENCEMINER also includes analytic functionalities such as the most frequent entity and relation summarization. The contributions and features of the EVIDENCEMINER system are summarized as follows.

1. We build EVIDENCEMINER, a web-based system for textual evidence discovery for life sciences. EVIDENCEMINER is supported by novel methods for distantly supervised named entity recognition and pattern-based open information extraction.

2. The retrieved evidence sentences can be easily located in the original text. The entity and relation annotation results are also highlighted in the original document for better visualization.

3. Analytic functionalities are included such as finding the most frequent entities/relations for given entity/relation types and finding the most frequent entities given a relation type with another entity.

2 Related Work

Search engines performing sentence-level retrieval have been developed in the biomedical domain. For example, Textpresso (Müller et al., 2004) highlights the query-related sentences in the retrieved documents. However, the sentence highlighting is only based on query word matching, which does not necessarily find sentences semantically related to the input query. Another example is LitSense (Allot et al., 2019), which retrieves semantically similar sentences in biomedical literature given

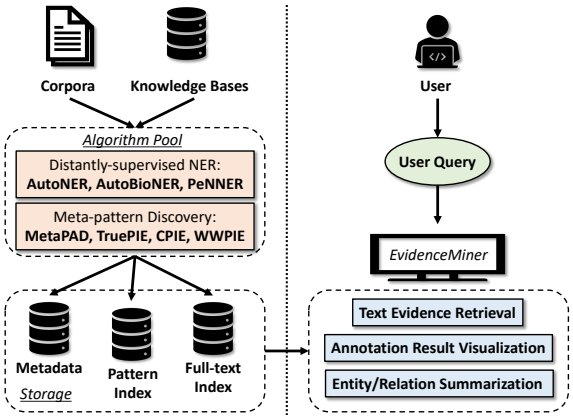


Figure 1: System architecture of EVIDENCEMINER.

a query sentence. It returns best-matching sentences using a combined approach of traditional word matching and neural embedding. However, their neural embeddings are noisy and thus negatively impact the effectiveness in retrieving query-specific evidence sentences. EVIDENCEMINER is more effective compared with LitSense for textual evidence retrieval in biomedical literature.

Similar tools are also developed for other domains, such as claim mining and argument mining tools on Twitter or news articles. PerspectroScope (Chen et al., 2019) allows users to query a natural language claim and extract textual evidence in support or against the claim. ClaimPortal (Majithia et al., 2019) is an integrated infrastructure for searching and checking factual claims on Twitter. TARGER (Chernodub et al., 2019) is an argument mining framework for tagging arguments in the free input text and keyword-based retrieval of arguments from the argument-tagged corpus. Most of these tools rely on fully supervised methods that require human-annotated training data. It is difficult to directly apply these systems to other domains, such as life sciences since it is non-trivial to retrieve the set of human-annotated articles and the annotations are prone to errors (Levy et al., 2017).

3 System Description

EVIDENCEMINER consists of two major components: an open information extraction pipeline and a textual evidence retrieval and analysis pipeline. The open information extraction pipeline includes two functional modules: (1) distantly supervised NER, and (2) meta-pattern-based open information extraction; whereas the textual evidence retrieval and analysis pipeline includes three functional modules: (1) textual evidence search, (2) annotation

Background corpora	Cancers	Heart Diseases
#PubMed abstracts	48,201	11,766
#PMC full-text papers	7,130	1,151
#Sentences in total	1,466,091	246,106
#Entity instances	3,315,092	400,327
#Relation instances	29,160	9,576

Table 1: Basic statistics of background corpora. It includes PubMed abstracts and PMC full-text papers related to cancers and heart diseases published in 2019.

result visualization in the original document, and (3) the most frequent entity and relation summarization. Figure 1 shows the system architecture of EVIDENCEMINER. The functional modules are introduced in the following sections.

3.1 Open Information Extraction

The open information extraction pipeline extracts entities with distant supervision from knowledge bases and relations with automatic meta-pattern discovery methods. In particular, to extract high-quality entities and relations, we design noise-robust neural models for distantly supervised named entity recognition (Shang et al., 2018b; Wang et al., 2019) and wide-window meta-pattern discovery methods to deal with the long and complex sentences in biomedical literature (Wang et al., 2018a; Li et al., 2018b).

Data Collection. To obtain the background corpora for EVIDENCEMINER, we collect the titles and abstracts of 26M papers from the entire PubMed[4] dump, and the full-text contents of 2.2M papers from PubMed Central[5] (PMC). For the demonstration purpose, we select a subset of documents published in 2019 that are specifically related to two important diseases (cancers and heart diseases) to form the background corpora. The subset of documents are selected by concept matching on MeSH[6], a biomedical concept ontology with the concepts related to cancers (Neoplasms) and heart diseases (Cardiovascular Diseases). Table 1 summarizes the statistics of the background corpora.

Distantly Supervised Named Entity Recognition. EVIDENCEMINER relies on UMLS[7], a comprehensive biomedical knowledge base to provide distant supervision for named entity recognition. We select 5 major biomedical entity types (Organism, Fully Formed Anatomical Structure, Chemical, Physiologic Function, and Pathologic Function) including 17 fine-grained entity types (Archaeon, Bacterium, Eukaryote, Virus, Body Part/Organ/Organ Component, Tissue, Cell, Cell Component, Gene or Genome, Chemical, Organism Function, Organ or Tissue Function, Cell Function, Molecular Function, Disease or Syndrome, Cell or Molecular Dysfunction, Experimental Model of Disease, and Pathological Function) from UMLS as the entity types to be annotated. To tackle the problem of limited coverage of the input dictionary, we first apply a data-driven phrase mining algorithm, AutoPhrase (Shang et al., 2018a), to extract high-quality phrases as additional entity candidates. Then we automatically expand the dictionary with a novel dictionary expansion method (Wang et al., 2019). The expanded dictionary is used to label the input corpora with the 17 fine-grained entity types to train a neural model. We apply AutoNER (Shang et al., 2018b), a state-of-the-art distantly supervised NER method that effectively deals with noisy distant supervision. Comparing with PubTator (Wei et al., 2013), a state-of-the-art BioNER system trained with extensive human annotation on 5 biomedical entity types, EVIDENCEMINER can automatically annotate 17 fine-grained entity types with high quality without any human effort for training data annotation.

Meta-pattern Discovery. Based on the entity annotation results above, meta-patterns can be automatically discovered from the corpora to support textual evidence discovery. Meta-patterns are defined as sub-sequences in an entity-type-replaced corpus with at least one entity type token in it. For example, "PPAR gamma agonist" and "caspase 1 agonist" are two word-sequences in the raw corpus. If we replace all the entities (i.e., "PPAR gamma" and "caspase 1") with their corresponding entity types (i.e., $GENE) in the raw corpus, "PPAR gamma agonist" and "caspase 1 agonist" are represented as one meta-pattern "$GENE agonist" in the entity-type-replaced corpus. Meta-patterns containing at least two entity types (e.g., "$CHEMICAL induce $DISEASE") are relational meta-patterns. Quality relational meta-patterns can serve as informative textual patterns that guide textual evidence discovery. We apply two state-of-the-art meta-pattern discovery methods, CPIE (Wang et al., 2018a) and WW-PIE (Li et al., 2018b), to extract high-quality meta-patterns from the NER-tagged corpora. Both methods are specifically de-

[4] https://pubmed.gov/pubmed
[5] https://pubmed.gov/pmc
[6] https://www.nlm.nih.gov/mesh/
[7] https://www.nlm.nih.gov/research/umls/index.html

signed to better deal with the long and complex sentence structures in the biomedical literature. In EVIDENCEMINER, we combine the meta-pattern extraction results from CPIE and WW-PIE as our informative meta-patterns to guide textual evidence retrieval. We use Elasticsearch[8] to create the index for each sentence for fast online retrieval. In addition to indexing the keywords, we index each sentence with the meta-patterns it matches and the corresponding entities extracted by the meta-patterns in the sentence.

3.2 Textual Evidence Retrieval and Analysis

The textual evidence retrieval and analysis pipeline retrieves textual evidence given a user-input query statement and the indexed corpora. The retrieved evidence sentence can be easily located in the original text. The entity and relation annotation results are also highlighted in the text for better visualization. EVIDENCEMINER also includes analytic functionalities such as finding the most frequent entities and relations as summarization.

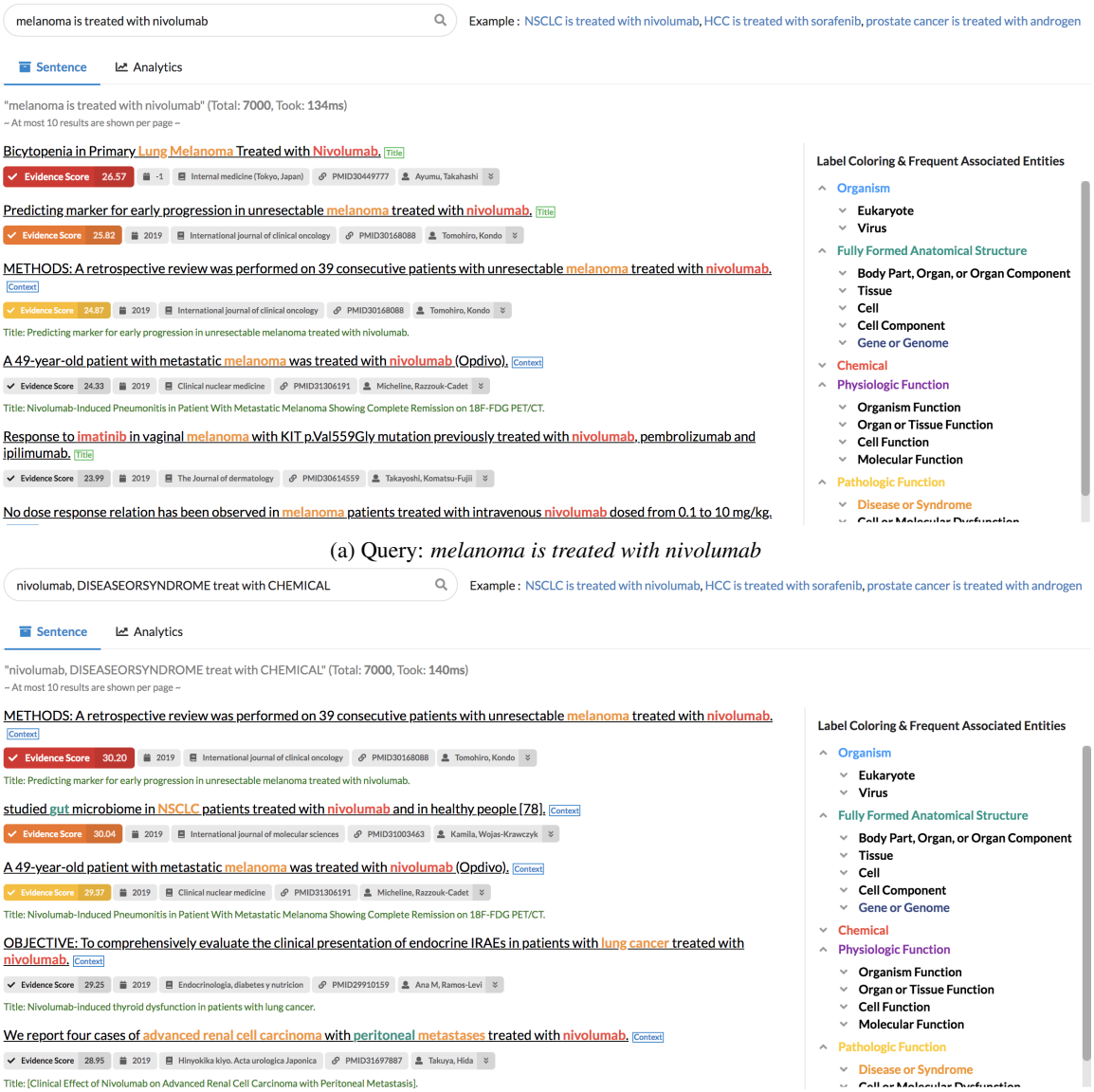
Textual Evidence Search. Given a user-input query statement and the indexed corpora, EVIDENCEMINER retrieves and ranks the candidate sentences with a combined approach of keyword weighting and meta-pattern weighting. Sentences with meta-patterns that better match the query statement are ranked higher as textual evidence. This ranking mechanism is more effective compared with existing methods (e.g., LitSense) for textual evidence retrieval in biomedical literature (see Section 4). We use Elasticsearch to support keyword and meta-pattern search over the indexed corpora.

In Figure 2, we show an example of our search interface. For example, if scientists are interested in finding the textual evidence for "*melanoma is treated with nivolumab*", they can search it in EVIDENCEMINER and see the top results such as "bicytopenia in primary lung melanoma treated with nivolumab" (Figure 2a). If they click one of the top results, the retrieved sentence is highlighted in the original article (Figure 3) on the annotation interface. Moreover, EVIDENCEMINER allows more flexible queries, such as a mixture of keywords and relational patterns. For example, if scientists are interested in finding the diseases that can be treated with the chemical "nivolumab", but are not sure which disease to search, they may input a query like "*nivolumab, DISEASEORSYNDROME treat with CHEMICAL*". EVIDENCEMINER automatically finds all the textual evidence indicating a "treatment" relationship with the chemical "nivolumab" (Figure 2b).

Annotation Result Visualization. The annotation interface shows all the annotated entities and relations for better visualization. For example, in Figure 3, we color all the annotated entities with different colors for different types. We use five different colors for the five major biomedical entity types and two additional colors for two specific fine-grained types, "Gene or Genome" and "Disease or Syndrome", since those two are the most frequent biomedical entity types. In Figure 3, we see that the "melanoma" is colored as a "Disease or Syndrome" and "nivolumab" is colored as a "Chemical". We also list all the meta-pattern instances and meta-patterns that match the sentences in the article. If the user clicks the meta-pattern instances, the corresponding sentences are also highlighted in the article. In Figure 3, a meta-pattern "DISEASEORSYNDROME patient treat with CHEMICAL" captures the entity pair "melanoma" and "nivolumab" in the article.

Entity and Relation Summarization. To make our system more user-friendly and interesting, we add analytic functionalities for the most frequent entity and relation summarization. For example, in Figure 4, if scientists are interested in finding the most frequent diseases, they can search "entity_type = DISEASEORSYNDROME" in our analytic interface and see the top entities such as *tumor* and *breast cancer*. Similarly, if scientists are interested in finding the most frequent chemical-disease pairs with a treatment relation, they can search "pattern = DISEASEORSYNDROME treat with CHEMICAL" in our analytic interface and see the top entity pairs such as *HCC&sorafenib*. More interestingly, if researchers are interested in finding the most frequent diseases that can be treated by a specific chemical (e.g., nivolumab), they can search "entity = nivolumab & pattern = DISEASEORSYNDROME treat with CHEMICAL" in our analytic interface and see the most frequent diseases, such as *melanoma* and *NSCLC*, that can be treated with nivolumab. With these analytic functionalities, EVIDENCEMINER can help scientists uncover important research issues, leading to more effective research and more in-depth quantitative analysis.

[8] https://www.elastic.co/



(a) Query: *melanoma is treated with nivolumab*

(b) Query: *(nivolumab, DISEASEORSYNDROME treat with CHEMICAL)*

Figure 2: The search interface with the textual evidence retrieved. The evidence score indicates the confidence of each retrieved sentence being a supporting evidence of the input query.

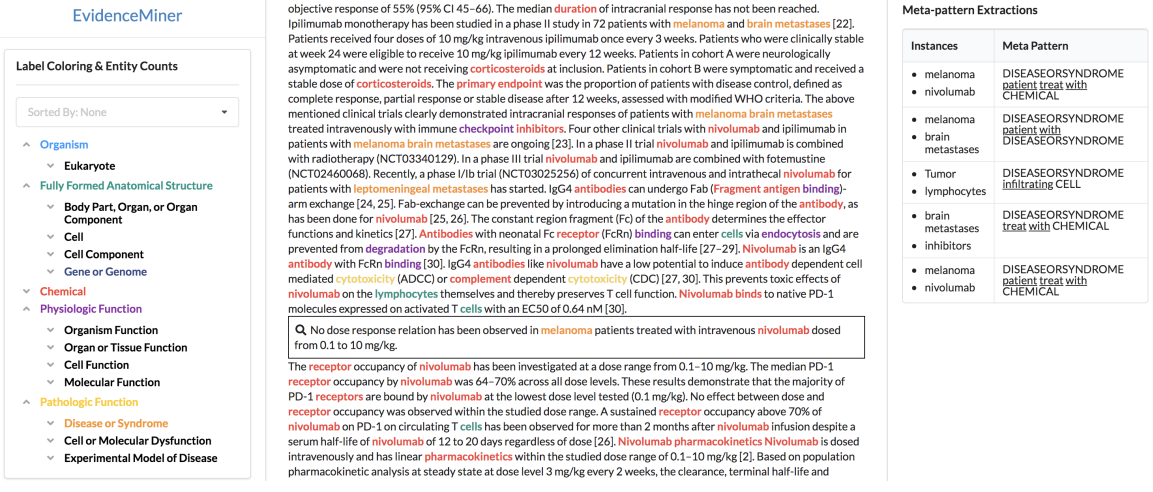
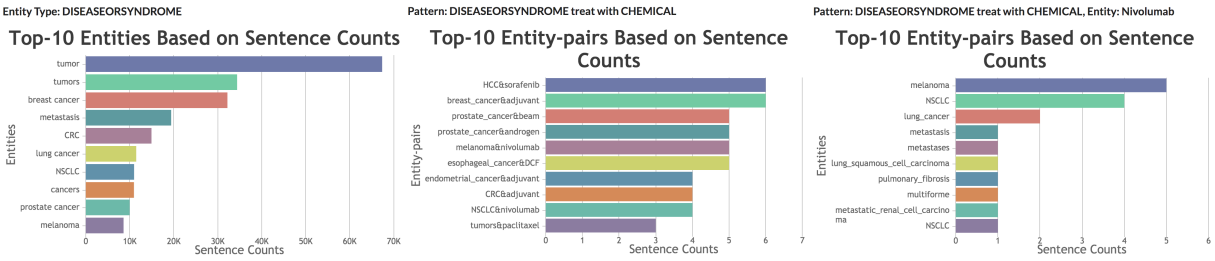


Figure 3: The annotation interface with all the entity and relation annotation results.



(a) (b) (c)

Figure 4: The analytic interface with the entity and relation summarization results. The queries used are (a) entity_type=DISEASEORSYNDROME, (b) pattern=DISEASEORSYNDROME treat with CHEMICAL, and (c) entity=nivolumab&pattern=DISEASEORSYNDROME treat with CHEMICAL.

Method / nDCG	@1	@5	@10
BM25	0.714	0.720	0.746
LitSense	0.599	0.624	0.658
EVIDENCEMINER	**0.855**	**0.861**	**0.889**

Table 2: Performance comparison of the textual evidence retrieval systems with nDCG@1,5,10.

4 Evaluation

To demonstrate the effectiveness of EVIDENCEMINER in textual evidence retrieval, we compare its performance with the traditional BM25 (Robertson et al., 2009) and a recent sentence-level search engine, LitSense (Allot et al., 2019). The background corpus is the same PubMed subset for all the compared methods. We first ask domain experts to generate 50 query statements based on the relationships between three biomedical entity types (gene, chemical, and disease) in the Comparative Toxicogenomics Database[9]. Then we ask domain experts to manually label the top-10 retrieved evidence sentences by each method with three grades indicating the confidence of the evidence. We use the average normalized Discounted Cumulative Gain (nDCG) score to evaluate the textual evidence retrieval performance. In Table 2, we observe that EVIDENCEMINER always achieves the best performance compared with other methods. It demonstrates the effectiveness of using meta-patterns to guide textual evidence discovery in biomedical literature.

5 Further Development

In some cases, a strict query matching may not find sufficiently high-quality answers due to the stringent search requirements or limited available entities that match the search queries. In this case, a smart query processor should automatically kick-in to do an approximate match, such as a graph-based approximate match or an embedding-based semantic match. In other cases, a user may query a set of entities (e.g., genes or diseases) or a timeline. We need to conduct a summary of the major differences among the set of entities or over time by analyzing large text.

[9] http://ctdbase.org

6 Conclusion

We build EVIDENCEMINER, a web-based system for textual evidence discovery for life sciences. The retrieved evidence sentences can be easily located in the background corpora for better visualization. EVIDENCEMINER also includes analytic functionalities such as the most frequent entity and relation summarization. We incorporated another corpus on COVID-19 in EVIDENCEMINER to help boost the scientific discoveries (Wang et al., 2020b,a). We are further developing EVIDENCEMINER to be a more intelligent system that can assist in more efficient and in-depth scientific discoveries.

Acknowledgment

Research was sponsored in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and SocialSim Program No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, and DTRA HDTRA11810026. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation hereon.

References

Alexis Allot, Qingyu Chen, Sun Kim, Roberto Vera Alvarez, Donald C Comeau, W John Wilbur, and Zhiyong Lu. 2019. Litsense: making sense of biomedical literature at sentence level. *Nucleic acids research*.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019. Perspectroscope: A window to the world of diverse perspectives. page 129134.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Targer: Neural argument mining at your fingertips. In *ACL: System Demonstrations*, pages 195–200.

Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *KDD*, pages 877–886. ACM.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proc. Work. Arg. Min.*, pages 79–84.

Qi Li, Meng Jiang, Xikun Zhang, Meng Qu, Timothy P Hanratty, Jing Gao, and Jiawei Han. 2018a. Truepie: Discovering reliable patterns in pattern-based information extraction. In *KDD*, pages 1675–1684. ACM.

Qi Li, Xuan Wang, Yu Zhang, Qi Li, Fei Ling, Cathy Wu H, and Jiawei Han. 2018b. Pattern discovery for wide-window open information extraction in biomedical literature. In *BIBM*. IEEE.

Marco Lippi and Paolo Torroni. 2016. Margot: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011.

Sarthak Majithia, Fatma Arslan, Sumeet Lubal, Damian Jimenez, Priyank Arora, Josue Caraballo, and Chengkai Li. 2019. Claimportal: Integrated monitoring, searching, checking, and analytics of factual claims on twitter. In *ACL: System Demonstrations*, pages 153–158.

Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, 2(11):e309.

Xiang Ren, Jiaming Shen, Meng Qu, Xuan Wang, Zeqiu Wu, Qi Zhu, Meng Jiang, Fangbo Tao, Saurabh Sinha, David Liem, et al. 2017. Life-inet: A structured network-based knowledge exploration and analytics system for life sciences. In *ACL: System Demonstrations*, pages 55–60.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *FnT Inf. Ret.*, 3(4):333–389.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018a. Automated phrase mining from massive text corpora. *TKDE*.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018b. Learning named entity tagger using domain-specific dictionary. In *EMNLP*. ACL.

Jiaming Shen, Jinfeng Xiao, Xinwei He, Jingbo Shang, Saurabh Sinha, and Jiawei Han. 2018. Entity set search of scientific literature: An unsupervised ranking approach. In *SIGIR*, pages 565–574. ACM.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 21–25.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan, and Jiawei Han. 2020a. Automatic textual evidence mining in covid-19 literature.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020b. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*.

Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018a. Open information extraction with meta-pattern discovery in biomedical literature. In *BCB*, pages 291–300. ACM.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503. IEEE.

Xuan Wang, Yu Zhang, Qi Li, Cathy H. Wu, and Jiawei Han. 2018b. PENNER: pattern-enhanced nested named entity recognition in biomedical literature. In *BIBM*, pages 540–547.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, 41(W1):W518–W522.