International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2018.10143

Review Article

IJCRR
Section: Life Sciences
Sci. Journal Impact Factor: 5.385 (2017)
ICV: 71.54 (2015)

Natural Language Processing and Unsupervised Learning: It's Significance on Biomedical Literature

Kanika Gupta[1], **Ashok Kumar**[2]

[1]Research Scholar, Centre for Systems Biology and Bioinformatics, Punjab University, Chandigarh-160014, India; [2]Assistant Professor, Chairperson, Centre for Systems Biology and Bioinformatics, Punjab University, Chandigarh-160014., India

ABSTRACT

There is massive information hidden in the biomedical literature in the form of scientific publications, book chapters, conference reports, etc. This information is growing exponentially with the speed exceeding Moore's Law i.e. observations double in every two years. It is therefore not possible for researchers and practitioners to automatically extract and relate information from different written resources. Also the data present in the written recourses is unstructured i.e. free-text therefore it becomes very arduous and exorbitant to obtain annotated material for its literature. So in order to overcome these problems Natural Language Processing (NLP) and Unsupervised Learning approaches are used. Natural Language Processing approach is the part of text mining which is the discovery by computer of new, previously unknown information by automatically extracting and relating information from different written resources to reveal the otherwise 'hidden' meanings. The Unsupervised Learning approach is the part of machine learning where no annotated training is necessary and it is more about exploring the data to find insights. Both these approaches can be used to find knowledge from written textual data in the form different interactions like protein-protein, gene-gene, gene-protein, etc. These approaches could also be used to develop classifiers, databases, tools or softwares which in future would automatically extract the knowledgeable information from literature, answering questions arising in the biomedical research and would also help in the development of new hypothesis. So here we discuss 53 softwares, tools and databases developed using Natural Language Processing (NLP) and unsupervised learning approaches, which are involved in plain texts analyzing and processing, categorizes current work in biomedical information and entities extraction.

Key Words: Text Mining, Natural Language Processing (NLP), Unsupervised Learning and Biomedical Literature

INTRODUCTION

Textual data is considered as the building block upon which any research thrives. The extent of details and the rush of data providing information through the advancements in technologies and internet have increased tremendously. The exponential growth in research for biomedical sciences has led to an increase in its publications. The textual data in the published literature is unstructured or free-text. The unstructured data either does not have a pre-characterized information model or is not sorted out in a pre-characterized way. The information is commonly text-heavy, but may contain critical information in the form of dates, numbers, and facts like protein-protein interactions, gene-disease associations, etc. as well. As the data both communicated and hidden up in biomedical writings are developing exponentially and the composed content is unstructured information so it isn't workable for analysts and experts to naturally extricate and relate data from various compositions (1) (2). Therefore manual effort to transform unstructured text into structured is a laborious process and automatic techniques are the solution (3). There are various automatic techniques for solving the above mentioned issue viz. supervised and unsupervised machine learning, text mining, semantic analysis, artificial intelligence etc. In the current work we will discuss the importance of two automatic techniques i.e. unsupervised machine learning and natural language processing and the softwares, tools and databases developed using these techniques, so that these techniques could be implemented on any biomedical corpus. Natural language processing (NLP) is the ability of a computer program to comprehend human language as it is spoken. The progress in Natural Language

Corresponding Author:
Ashok Kumar, Chairperson, Centre for Systems Biology and Bioinformatics, Block-III, South Campus, Sector-25, Punjab University, Chandigarh-160014, India; Tel: +919216270057; E-mail: ashokbiotech@gmail.com & ashokkumar@pu.ac.in

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)
Received: 28.05.2018 **Revised:** 15.06.2018 **Accepted:** 07.07.2018

Processing (NLP) applications is provocative because computers commonly require humans to "speak" to them in a programming language that is accurate, explicit and highly organized, or through a limited number of clearly articulated voice commands. Most of the research being done on Natural Language Processing (NLP) rotates around search i.e. keyword search or searching relationship entities. This Natural Language Processing (NLP) method enables users to query data sets in the form of a question that they might pose to another person. The machine elucidates the critical components of the human language sentence, such as those that might correspond to specific features in a data set, and returns an answer. Natural Language Processing (NLP) can be utilized to interpret free text and make it analyzable (51). The unsupervised machine learning approaches are generally beneficial on the unstructured data i.e. the data where no labels are given to the learning algorithm, leaving it on its own to find structure in its input. This kind of learning can be a goal in itself by finding hidden patterns in data or a means towards an end i.e. feature learning used for the development of textual classifiers. The unsupervised learning problems can be further grouped into clustering which is problem is where you want to find the inherent groupings in the data (52). Many researchers have utilized these approaches for information extraction from biomedical literature, especially for discovery of protein–protein interactions, gene-protein interaction, gene-drug interaction, etc. In this paper we will discuss few softwares, databases or techniques which use Natural Language Processing (NLP) and unsupervised learning approaches for classification and entity recognition from biofilm literature.

Brief Description of Techniques

This section presents a brief discussion on the Natural Language Processing (NLP) and unsupervised techniques and its general method for linguistics analysis to find different interactions (4).

Natural Language Processing (NLP) methods. Knowledgeable discovery from unstructured text utilizes computational linguistics and philosophy, like syntactic parsing or semantic parsing to analyze sentence structures. Methods of this category define grammars to describe sentence structures and utilize parsers to extract syntactic information and internal dependencies within individual sentences. Approaches in this category can be applied to different knowledge domains after being carefully tuned to the specific problems. But, there is still no guarantee that the performance in the field of biomedicine can achieve comparable performance after tuning. Until recently, methods based on computational linguistics still could not generate satisfactory results (5) (6).

Unsupervised Machine learning. Machine learning broaches to the potentiality of a machine to grasp from experience to extract knowledge from data corpora. As opposed to the aforementioned technique which needs laborious effort to define a set of rules or grammars, machine learning techniques are able to extract protein–protein interaction patterns without human intervention. Statistical approaches are based on word occurrences in a large text corpus. Significant features or patterns are detected and used to classify the abstracts or sentences containing protein–protein interactions, gene-protein interaction and characterize the corresponding relations among genes or proteins. They also define a set of rules for possible textual relationships, called patterns, which encode similar structures in expressing relationships. When combined with statistical methods, scoring schemes depending on the occurrences of patterns to describe the confidence of the relationship are normally used. Similar to computational linguistics methods, rule-based approaches can make use of syntactic information to achieve better performance, although it can also work without prior parsing and tagging of the text (7) (8).

The **Figure 1** shows the general outlook of information extraction system from any Biomedical Literature. In this the data is collected from various sources like published articles, scientific journals, books and technical reports, etc and the collected data is in unstructured format. Then using automatic techniques like text mining, text units i.e. words, sentences, paragraphs containing relevant information are generated which needs to be analyzed to get knowledgeable data. Then these text units are further processed and analyzed using unsupervised learning and natural language processing which are used for text classification or clustering on certain textual features and entity recognition like gene-protein interaction, protein-protein interaction, gene-disease interaction, gene-drug interaction, etc. This gathered information can be used for the development of databases, classifiers, softwares, tools or pipelines for future use.

DISCUSSION

The softwares, tools, databases and pipelines which are involved in information extraction in the form of relationship entities in biofilm literature using Natural Language Processing and Unsupervised Learning approaches are shown in **Table 1**.

The above mentioned softwares, tools, databases and pipelines can be used for information extraction by initially identifying an item or concept in textual resource and then detecting links between the concepts obtained from the text. By linking the concepts together additional context is given to the concepts, which results in valuable knowledge that can be used for downstream analysis like genome and gene expression annotation, drug-target discovery, drug repositioning, protein-protein interactions, construction of ontologies etc (9). These techniques also help researchers in formulating hypothesis of their future studies as they could find new concepts while analyzing text.

CONCLUSION

The importance of natural language processing and unsupervised learning depends on the fact that it not only extract information hidden in the biomedical textual data but could also be used for the development of new servers, softwares, databases, etc. These approaches could be used on any biomedical literature. If the above mentioned tools meet all the challenges like specific ontologies describing single disease at various levels, individual pathways and genes for particular diseases, appropriate gene-disease interactions, quality of tool to distinguish between false negative results, etc. being faced in analysis of textual data they will continue to be an indispensable asset for researchers in the biomedical domain (9).

ACKNOWLEDGEMENT

The authors acknowledge the support provided by the online servers, tools and softwares for the compilation of this review. Authors also acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors / editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

Source of Funding: None

Conflict of Interest: None

REFERENCES

1. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews Genetics. 2006;7(2):119-29.
2. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends in biotechnology. 2006;24(12):571-9.
3. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, et al. Literature-curated protein interaction datasets. Nature methods. 2009;6(1):39-46.
4. Zhou D, He Y. Extracting interactions between proteins from the literature. Journal of biomedical informatics. 2008;41(2):393-407.
5. Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug discovery today. 2005;10(6):439-45.
6. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Briefings in bioinformatics. 2005;6(1):57-71.
7. Rinaldi F, Schneider G, Kaljurand K, Hess M, Andronis C, Konstandi O, et al. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artificial intelligence in medicine. 2007;39(2):127-36.
8. Huang M, Zhu X, Hao Y, Payan DG, Qu K, Li M. Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics. 2004;20(18):3604-12.
9. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. Methods. 2015;74:97-106.
10. Baral C, Gonzalez G, Gitter A, Teegarden C, Zeigler A, Joshi-Tope G. CBioC: beyond a prototype for collaborative annotation of molecular interactions from the literature. Computational systems bioinformatics Computational Systems Bioinformatics Conference. 2007;6:381-4.
11. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics. 2004;5:147.
12. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic acids research. 2005;33(Web Server issue):W783-6.
13. Hoffmann R, Valencia A. A gene network for navigating the literature. Nature genetics. 2004;36(7):664.
14. Fontelo P, Liu F, Ackerman M. askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC medical informatics and decision making. 2005;5:5.
15. Lewis J, Ossowski S, Hicks J, Errami M, Garner HR. Text similarity: an alternative way to search MEDLINE. Bioinformatics. 2006;22(18):2298-304.
16. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA. MedlineRanker: flexible ranking of biomedical literature. Nucleic acids research. 2009;37(Web Server issue):W141-6.
17. States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD. MiSearch adaptive pubMed search tool. Bioinformatics. 2009;25(7):974-6.
18. Huang KC, Chiang IJ, Xiao F, Liao CC, Liu CC, Wong JM. PICO element detection in medical text without metadata: are first sentences enough? Journal of biomedical informatics. 2013;46(5):940-6.
19. Hokamp K, Wolfe KH. PubCrawler: keeping up comfortably with PubMed and GenBank. Nucleic acids research. 2004;32(Web Server issue):W16-9.
20. Plikus MV, Zhang Z, Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC bioinformatics. 2006;7:424.
21. Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, et al. PubMatrix: a tool for multiplex literature mining. BMC bioinformatics. 2003;4:61.
22. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome biology. 2005;6(9):R80.
23. Brancotte B, Biton A, Bernard-Pierrot I, Radvanyi F, Reyal F, Cohen-Boulakia S. Gene List significance at-a-glance with GeneValorization. Bioinformatics. 2011;27(8):1187-9.
24. De S, Zhang Y, Garner JR, Wang SA, Becker KG. Disease and phenotype gene set analysis of disease-based gene expression in mouse and human. Physiological genomics. 2010;42A(2):162-7.
25. Li C, Jimeno-Yepes A, Arregui M, Kirsch H, Rebholz-Schuhmann D. PCorral--interactive mining of protein interactions from MEDLINE. Database : the journal of biological databases and curation. 2013;2013:bat030.
26. Glynn RW, Kerin MJ, Sweeney KJ. Authorship trends in the surgical literature. The British journal of surgery. 2010;97(8):1304-8.
27. Xuan W, Dai M, Mirel B, Wilson J, Athey B, Watson SJ, et al. An active visual search interface for Medline. Computational systems bioinformatics Computational Systems Bioinformatics Conference. 2007;6:359-69.

28. Fleuren WW, Verhoeven S, Frijters R, Heupers B, Polman J, van Schaik R, et al. CoPub update: CoPub 5.0 a text mining system to answer biological questions. Nucleic acids research. 2011;39(Web Server issue):W450-4.
29. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011;27(13):i111-9.
30. Raja K, Subramani S, Natarajan J. PPInterFinder--a mining tool for extracting causal relations on human proteins from literature. Database : the journal of biological databases and curation. 2013;2013:bas052.
31. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. Bioinformatics. 2006;22(19):2444-5.
32. Soldatos TG, O'Donoghue SI, Satagopam VP, Jensen LJ, Brown NP, Barbosa-Silva A, et al. Martini: using literature keywords to compare gene sets. Nucleic acids research. 2010;38(1):26-38.
33. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research. 2013;41(Database issue):D808-15.
34. Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic acids research. 2014;42(Web Server issue):W137-46.
35. Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015;74:83-9.
36. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS. NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. BMC bioinformatics. 2016;17:32.
37. Tao C, Song D, Sharma D, Chute CG. Semantator: semantic annotator for converting biomedical text to linked data. Journal of biomedical informatics. 2013;46(5):882-93.
38. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2009;37(Web Server issue):W170-3.
39. Stokes TH, Wang MD. SimplevisGrid: grid services for visualization of diverse biomedical knowledge and molecular systems data. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2009;2009:4178-81.
40. Gupta S, Ross KE, Tudor CO, Wu CH, Schmidt CJ, Vijay-Shanker K. miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases. Journal of biomedical semantics. 2016;7(1):9.
41. Lee K, Lee S, Park S, Kim S, Kim S, Choi K, et al. BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. Database : the journal of biological databases and curation. 2016;2016.
42. Blank CE, Cui H, Moore LR, Walls RL. MicrO: an ontology of phenotypic and metabolic characters, assays, and culture media found in prokaryotic taxonomic descriptions. Journal of biomedical semantics. 2016;7:18.
43. Mahmood AS, Wu TJ, Mazumder R, Vijay-Shanker K. DiMeX: A Text Mining System for Mutation-Disease Association Extraction. PloS one. 2016;11(4):e0152725.
44. Wei CH, Leaman R, Lu Z. SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine. ACM-BCB : the ACM Conference on Bioinformatics, Computational Biology and Biomedicine ACM Conference on Bioinformatics, Computational Biology and Biomedicine. 2014;2014:138-46.
45. Finch DK, McCart JA, Luther SL. TagLine: Information Extraction for Semi-Structured Text in Medical Progress Notes. AMIA Annual Symposium proceedings AMIA Symposium. 2014;2014:534-43.
46. Sharma VK, Kumar N, Prakash T, Taylor TD. MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets. Nucleic acids research. 2010;38(Database issue):D468-72.
47. Kuo CJ, Ling MH, Lin KT, Hsu CN. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. BMC bioinformatics. 2009;10 Suppl 15:S7.
48. Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R. PepBank--a database of peptides based on sequence text mining and public peptide data sources. BMC bioinformatics. 2007;8:280.
49. Kim S, Shin SY, Lee IH, Kim SJ, Sriram R, Zhang BT. PIE: an online prediction system for protein-protein interactions from text. Nucleic acids research. 2008;36(Web Server issue):W411-5.
50. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G. Inter-species normalization of gene mentions with GNAT. Bioinformatics. 2008;24(16):i126-32.
51. TechTarget (2017). Available at: http://searchbusinessanalytics.techtarget.com/definition/ natural-language-processing-NLP [Accessed 08 Feburary 2018].
52. Machine Learning Mastery (2016). Available at : https://machinelearningmastery.com/ supervised-and-unsupervised-machine-learning-algorithms/ [Accessed 07 Feburary 2018]

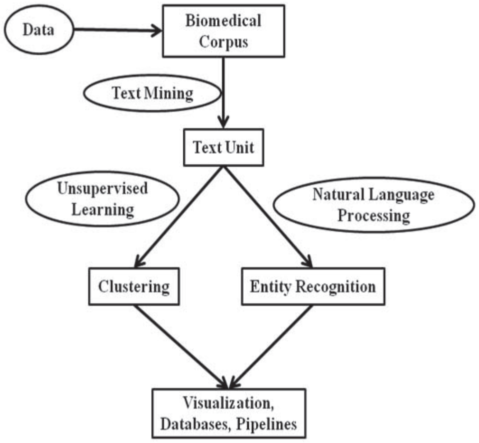

Figure 1: A general architecture of an information extraction system from Biomedical Literature which could be further used for the development of databases or pipelines.

Table 1: Overview of Natural Language Processing and Unsupervised Learning applications for the biomedical (e.g. biofilm) domain. The table lists the tools/software/database/pipeline name, its brief description and URL. The approach used for the development is given in the bracket where NLP stands for Natural Language Processing and UL stands for Unsupervised Learning.

Name	Description	URL
CBioC (NLP)	Collaborative Bio Curation uses automatic text extraction as a starting point to initialize the interaction database. After that, researchers in biomedical domain contribute to the curation process by subsequent edits (10).	cbioc.eas.asu.edu
Chilibot (NLP)	This software for MEDLINE literature database to rapidly identify relationships between genes, proteins, or any keywords that the user might be interested (11).	www.chilibot.net
GoPubMed (NLP)	This is a search engineer that allows users to explore PubMed search results with the Gene Ontology (GO), a hierarchically structured vocabulary for molecular biology (12).	www.gopubmed.org
iHOP (NLP)	Information Hyperlinked over Proteins constructs a gene network by converting the information in MEDLINE into one navigable resource using genes and proteins as hyperlinks between sentences and abstracts (13).	www.ihop-net.org/UniPub/iHOP
iProLINK (NLP and UL)	This is a resource to facilitate text mining in the area of literature-based database curation, named entity recognition, and protein ontology development. It can be utilized by computational and biomedical researchers to explore the literature information on proteins and their features or properties (4).	pir.georgetown.edu/iprolink
PreBIND (NLP and UL)	This tool helps researchers locate bio-molecular interaction information in the scientific literature. It identifies papers describing interactions using a support vector machine (4).	prebind.bind.ca
PubGene (NLP)	This is constructed to identify the relationships between genes and proteins, diseases, cell processes, and so on based on their co-occurrences in the abstracts of scientific papers, their sequence homology, and statistical probability of their co-occurrences (4).	www.pubgene.org
Whatizit (NLP)	It is a text processing tool that can identify molecular biology terms and linking them to publicly available databases. Identified terms are enveloped with XML tags that carry additional data, such as the primary keys to the databases where all the applicable information is kept. It is also a MEDLINE abstracts search engine (4).	www.ebi.ac.uk/webservices/whatizit/info.jsf
askMEDLINE (NLP)	Free-text, natural language (English only) query for MEDLINE/PubMed (14).	askmedline.nlm.nih.gov/ask/ask.php
XplorMed (NLP)	The system provides the main associations between the words in groups of abstracts (9).	xplormed.ogic.ca/
eTBLAST (NLP)	A text-similarity based search engine, using all words in a paragraph to match similar documents (15).	etest.vbi.vt.edu/etblast3/
Medline Ranker (NLP)	Ranks MEDLINE abstracts based on user defined queries (16).	cbdm.mdc-berlin.de/~medlineranker/cms/medline-ranker
MiSearch (NLP)	Ranks retrieved articles from PubMed based on a customized personal profile (17).	portal.ncibi.org/gateway/misearch.html
PICO (NLP and UL)	Search engine for MEDLINE/PubMed with an integrated spelling checker (18).	pubmedhh.nlm.nih.gov/nlmd/pico/piconew.php
PubCrawler (NLP)	Free alerting service that scans daily updates of the NCBI MEDLINE/PubMed and GenBank databases (19).	pubcrawler.gen.tcd.ie/

Table 1: *(Continued)*

Name	Description	URL
PubFocus (NLP)	Statistical analysis of the MEDLINE/PubMed search queries enriched with additional information from journal rank database and forward referencing database (20).	www.pubfocus.com/
PubGet (NLP)	Retrieves PDFs directly based on a user defined query in PubMed (9).	pubget.com/
PubMatrix (NLP and UL)	Allows simple text based mining of the NCBI literature search service PubMed using any two lists of keywords terms, resulting in a frequency matrix of term co-occurrence (21).	pubmatrix.grc.nia.nih.gov/secure-bin/index.pl
PubNet (NLP)	A flexible system for visualizing literature-derived networks (22).	pubnet.gersteinlab.org/
GeneValorization (NLP)	GeneValorization gives a very clear and handful overview of the bibliography corresponding to user uploaded gene lists (23).	bioguide-project.net/gv/start_geneval.php
DPWP (NLP)	Disease/phenotype PAGE is a disease focused gene set analysis web tool to analyze microarray gene expression data with predefined groups of disease related genes (24).	dpwebpage.nia.nih.gov/
Anne O'Tate (NLP and UL)	An overview of the set of articles retrieved by a PubMed query is generated.	arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi
ProteinCorral (NLP and UL)	Combines information retrieval and extraction from MEDLINE (25).	www.ebi.ac.uk/Rebholz-srv/pcorral/
MEDIE (NLP and UL)	Retrieve biomedical correlations from MEDLINE, based on indexing by natural language processing and text mining techniques (9).	www.nactem.ac.uk/medie/
PubReMiner (NLP and UL)	Breaks down a results of a Pubmed query into Categories (26).	hgserver2.amc.nl/cgi-bin/miner/miner2.cgi
PubViz (NLP)	An interactive MEDLINE search engine utilizing external knowledge (27).	brainarray.mbni.med.umich.edu/
Quertle (NLP)	Using an amalgamation of linguistic systems, Quertle finds facts defined within documents, creating its own database of about 300 million relationships, and is able to report the ones that are relevant to your query.	www.quertle.info/
CoPub (NLP)	Web application with gene focussed retrieval of co-occurring keywords from MEDLINE (28).	www.copub.org
COREMINE Medical (NLP)	Presents results about health, medicine and biology in a dashboard format comprised of panels containing various categories of information ranging from introductory sources to the latest scientific articles (9).	www.coremine.com/
FACTA+ (NLP and UL)	The associated concepts with text analysis based on a user queries are found (29).	www.nactem.ac.uk/facta/
PPInterFinder (NLP)	PPInterFinder uses relation keyword co-occurrences with protein names to extract information on protein–protein Interactions from MEDLINE abstracts (30).	biomining-bu.in/ppinterfinder/html/action.pl
Reflect (NLP)	Reflect highlights protein and small molecule names, such as IL-5 and rapamycin in text.	reflect.embl.de
AliBaba (NLP)	The PubMed abstracts parses for biological objects and their relations (31).	alibaba.informatik.hu-berlin.de/
Martini (NLP)	Martini uses literature keywords to compare gene sets (32).	martini.embl.de
STRING (NLP)	STRING is a database of known and predicted protein interactions (33).	string-db.org/

Table 1: *(Continued)*

Name	Description	URL
DiseaseConnect (NLP)	A comprehensive web server for mechanism-based disease–disease connections (34).	http://disease-connect.org
DISEASES (NLP)	Text mining and data integration of disease–gene associations (35).	http://diseases.jensenlab.org/
NOBLE (NLP)	Flexible concept recognition for large-scale biomedical natural language Processing (36).	https://omictools.com/noble-coder-tool
Semantator (NLP)	Semantic annotator for converting biomedical text to linked data (37).	https://sbmi.uth.edu/ontology/project/se-mantator.htm
BioPortal (NLP)	An open repository of biomedical ontologies that provides access via Web services and Web browsers to ontologies developed in OWL, RDF, OBO format and Protégé frames (38).	http://bioportal.bioontology.org
SimplevisGrid (NLP)	Grid services for visualization of diverse biomedical knowledge and molecular systems data (39).	-
miRiaD (NLP)	A Text Mining Tool for Detecting Associations of microRNAs with Diseases (40).	http://biotm.cis.udel.edu/miRiaD
BRONCO (NLP)	Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations (41).	http://infos.korea.ac.kr/bronco
MicrO (NLP)	An ontology of phenotypic and metabolic characters, assays, and culture media found in prokaryotic taxonomic descriptions (42).	http://purl.obolibrary.org/obo/MicrO.owl https://github.com/carrineblank/MicrO http://www.obofoundry.org/ontology/micro.html
DiMeX (NLP)	A Text Mining System for Mutation-Disease Association Extraction (43).	http://biotm.cis.udel.edu/dimex/
SimConcept (NLP and UL)	A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine (44).	https://www.ncbi.nlm.nih.gov/research/bi-onlp/Tools/simconcept/
TagLine (NLP and UL)	Information Extraction for Semi-Structured Text in Medical Progress Notes (45).	-
MetaBioME (NLP)	A database to explore commercially useful enzymes in metagenomic datasets (46).	http://metasystems.riken.jp/metabiome/
BIOADI (NLP and UL)	A machine learning approach to identifying abbreviations and definitions in biological literature (47).	http://bioagent.iis.sinica.edu.tw/BIOADI/
PepBank (NLP)	A database of peptides based on sequence text mining and public peptide data sources (48).	http://pepbank.mgh.harvard.edu/
PIE (NLP and UL)	An online prediction system for protein-protein interactions from text (49).	http://bi.snu.ac.kr/pie/
GNAT (NLP)	Inter-species normalization of gene mentions (50).	http://cbioc.eas.asu.edu/gnat/ http://bcms.bioinfo.cnio.es