A Survey on Biomedical Named Entity Recognition and Normalization

Ruoyao Ding* and Huaxing Chen

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangdong, P.R. China

Abstract

With a rapidly-growing amount of biomedical information available only in textual form, there is considerable interest in applying NLP techniques to extract such information from the biomedical literature. Much of the research has paid special attention to extracting information about biomedical named entities. In this paper, we conducted a survey on biomedical named entity recognition and normalization, focusing on gene mention recognition and normalization. We believe this can help researchers to find work of their interest and interpret their own research.

Keywords

Recognition; Normalization; Text Mining

Correspondence to:

Ruoyao Ding
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangdong, P.R. China
Email: ruoyaoding@163.com

EJBI 2019; 15(2):11-16
Received: April 29, 2019
Accepted: July 12, 2019
Published: July 19, 2019

1 Introduction

Biomedical researchers usually describe their experimental results in research publications. With the rapid growth of biomedical publications, the information of interest needs to be extracted automatically to avoid the time consuming and labor intensive process.

Named entity recognition and normalization are two common tasks in the biomedical text mining field. Together they provide a means to extract the unstructured information buried in the literature and put the extracted information to structured form. There already has been some work on the survey of biomedical named entity recognition and normalization. However, a more comprehensive and most updated version is still needed. In this paper, we conduct a survey to present current work on biomedical named entity recognition and normalization. Given the primary importance genes and their products play in biological and medical studies, this survey will focus on gene mention recognition and normalization. We hope this can help researchers in biomedical text mining field to find the information of their interest and interpret their own research.

2. Gene Mention Recognition

The task of gene mention (GM) recognition is to automatically recognize gene/protein names mentioned in text. This task has received wide attention, and has been used in several challenge evaluations such as BioCreative I [1] and BioCreative II [2]. Other annotated corpora have also been constructed for system development and evaluation purpose.

There are several challenges of the gene mention recognition task:

(1) **No. of genes:** The number of gene names is in the millions and new names are created continuously.

(2) **Name variations: Authors usually do not use proposed standardized gene names.**

(3) **Polysemy:** Gene names often also refer to other entities such as disease names.

2.1 Gene Mention Recognition systems

Approaches to gene mention recognition can be categorized into two major classes: rule-based approaches and machine learning-based approaches.

While rule-based gene mention recognition approaches do not require annotated data to train a system, they do require domain experts to be closely involved in developing the rules. The following three systems are examples of gene mention detectors that rely on manually developed rules.

Hanisch et al. [3] presented a dictionary matching based system that detects fly, mouse and yeast gene names from biomedical text. Fukuda et al. [4] proposed a method which incorporates two new concepts called c-term (a concept based on orthography) and f-term (a concept that is based on terms that correspond to types of biological entities) (details about those two terms will be introduced later in the Gene Normalization chapter). Narayanaswamy et al. [5] developed a system which extracts multiple types of named entities including gene names. Their system is based on a manually developed set of rules that rely upon some crucial lexical information, linguistic constraints of English, and contextual information and develop the notion of c-term and f-term in named entity recognition.

The machine learning-based gene mention recognition approaches require annotated data to train a system. Thus, domain expertise is now required in the development of the data annotation and less during the system training.

In the machine learning-based gene mention recognition approaches, the gene mention recognition task is often treated as a sequence labelling problem (label the tokens in the text using the tags). BIO (or IO) tags for the text are commonly used to represent the boundaries of gene mentions where B represents the beginning of the gene name in text, I is assigned to a token inside the gene mention and O is assigned to token that are outside the gene mentions.

Among the machine-learning based systems, Banner [6] is widely used for recognizing biomedical named entities including gene mentions. It is based on conditional random fields and applied orthographic, morphological and shallow syntax features. Liu et al. [7] trained a classification system using Conditional random field (CRF) [8] to classify each word in the literature to the BIO tags. They applied BioThesaurus [9], a comprehensive collection of gene names to entries in the UniProt Knowledgebase, for dictionary lookup and used the matching information as a feature. Huang et al. [10] considered the gene mention task as a classification problem and applied support vector machine (SVM) [11] to solve it. Chen et al. [12] proposed a gene mention recognition system for biomedical literature using a dictionary and Support Vector Machine. Zhou et al. [13] proposed an ensemble of classifiers for gene mention recognition. They combined three classifiers, one Support Vector Machine and two discriminative Hidden Markov Models using a simple majority voting strategy. Other machine learning based gene mention systems can be found in [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24].

2.2 Gene Mention Recognition Corpora

High quality gene mention corpora are important for the development of any type of gene mention recognition system. Even for the rule-based system, more accurate rules can be made by analyzing the instances in the corpora.

The GENIA corpus [25] is a collection of 2000 abstracts extracted from Medline database. Multiple biomedical named entities, including gene names, are annotated. It is focused on a subset of human hematology. The PennBioIE corpus [26] consists of 1414 Medline abstracts on cancer. 24 types of biomedical named entities, including gene names, are annotated. The BioCreative 1 GM corpus [1] contains 15,000 sentences from Medline abstracts. Genes and related entities mentions are annotated. The BioCreative 2 GM corpus [2] contains 20,000 sentences from Medline abstracts (15,000 of which were used previously in BioCreative 1).

3. Gene Normalization

The task of gene normalization (GN) is to automatically link a gene mention to a database entry for the gene (product). Other than the challenges stated in the gene mention recognition task, the challenges for the gene normalization task also include:

(1) identifying the species for the gene mentions since most gene (product) knowledge bases contain species-specific entries, and

(2) disambiguation since multiple gene entries may share the same short name (symbol).

3.1 Gene Normalization Systems

The following were the top performing systems in the BioCreative I [27] and BioCreative II [28] Challenge GN Tasks. ProMiner [29] is a dictionary-based GN system which is characterized by the inclusion of different biomedical dictionaries and manual clean-up of a dictionary. BioTagger [7] tackles the GN problem with the steps: (1) dictionary lookup to obtain a list of mapping pairs of gene mention and database identifier, (2) machine learning that considers features such as the gene mention recognition, name ambiguity, and token shape information, and (3) a similarity based method to associate Entrez gene records with phrases detected by the gene mention tagger. GNAT [30] is a GN system encompassing four steps: named entity recognition for genes and species, validation of gene mentions, correlating gene mentions with species, and finally gene mention disambiguation. GeNo [31] tackles the GN problem by employing a carefully crafted suite of symbolic and statistical methods.

In BioCreative III [32], the GN task was further extended to cover genes of all relevant species in the literature corpora. Among the systems, Bhattacharya et al. [33] tried to associate a species name with a gene name by considering their proximity to the gene mention. Dai et al. [34] employed a multistage GN procedure and selected dictionary entries from only the top 22 most common species in NCBI (from 7283 species) to speed up the GN process. A document-level gene normalization system, called GeneTUKit [35], employed features from the local context as well as the global context of the whole full-text article. GenNorm [36] follows three steps: gene name recognition, species assignment, and species-specific gene normalization, and uses SR4GN [37] for assigning species to gene mentions. GenNorm has been widely used in text mining systems that require GN, such as in PubTator [38] and in an event extraction pipeline [39]. GNormPlus [40], as an updated version of GenNorm, refined the gene mention process by training the mention recognizer on a new corpus with gene, gene family and protein domain annotations. It also integrates several advanced text mining techniques, including SimConcept for resolving composite gene names.

3.2 Gene Normalization Corpora

High quality gene normalization corpora are important for the development of any type of gene normalization system.

The BioCreative I gene normalization corpus [27] and the BioCreative II gene normalization corpus [28] focused on the GN task for yeast, fly, and mouse genes and human genes respectively. Both of these corpora annotate gene mentions found in abstracts. In contrast, the BioCreative III gene normalization corpus [33] annotates full length articles and is not limited to specific species.

The BioCreative I gene normalization corpus consists 15,000 abstracts for training, 468 abstracts for developing, and 750 abstracts for testing. All these abstracts are annotated in abstract level, not mention level (only a list of database identifiers is given for each abstract, without any location information). No corresponding gene name in the abstracts for the database identifier is provided in this corpus. The BioCreative II gene normalization corpus consists 281 abstracts for training, and 262 abstracts for testing. All these abstracts are also annotated in abstract level, but the corresponding gene names in the abstracts are given for each database identifier. The BioCreative III gene normalization corpus consists 32 fully annotated articles and 500 partially annotated articles for training. For testing, it provides 50 articles as gold standard and 507 articles as silver standard.

4. Other Biomedical Named Entity Recognition and Normalization

There has been considerable interest in the detection and normalization of other types of biomedical entities such as diseases, chemical compounds and drugs.

4.1 Other Biomedical Named Entity Recognition

tmChem [41] is a chemical named entity recognition system created by combining two Conditional random field (CRF) models in an ensemble. The two models in the system used different tokenization methods, feature sets, CRF implementations, CRF parameters. Lu et al. [42] developed a chemical named entity recognition system based on mixed CRFs with word clustering. Lowe et al. [43] proposed a system for chemical entity recognition based on grammar and dictionary. Their system uses a mixture of expertly curated grammars and dictionaries, as well as dictionaries automatically derived from public resources.

Chowdhury et al. [44] presented a CRF based approach for disease mention recognition. The features they used include disease specific contextual features, orthographic features, general linguistic features, syntactic dependency features and dictionary lookup features. Kaewphan et al. [45] developed a system for disease mention recognition. Their system was based on an existing named entity system, NERsuite, supplemented with UMLS dictionary features.

Other biomedical named entity recognition systems can be found in [46, 47, 48, 49, 50].

4.2 Other Biomedical Named Entity Normalization

Leaman et al. [41] paired their chemical named entity recognition system with a dictionary approach for normalization. They used a dictionary of chemical entities and their names that was collected from MeSH and ChEBI. DNorm [51] is a disease normalization system, which uses a linear model to score the similarity between mentions and concept names. DNorm has an interesting approach of learning term variation directly from training data. Kaewphan et al. [45] developed a disease normalization system, which was based on their disease mention system. They combined compositional word vector representations with CRF to map the recognized mentions to the UMLS concepts. Other biomedical named entity normalization works can be found in [52, 53, 54, 55].

5. Discussion

1.1 Biomedical Named Entity Recognition based on Deep Learning

In recent years, deep learning has drawn much attention in biomedical named entity recognition. Hence, we will next describe some novel work of biomedical named entity recognition in the last three years based on deep learning. Hemati et al. [56] combined Long Short Term Memory (LSTM) nerual networks and CRF to detect drug named entity, and achieved state-of-the-art performace. Korvigo et al. [57] first used Convolutional Neural Network (CNN) to encode the text, then applied Recurrent Neural Networks (RNN) to recognize drug named entity. Xu et al. [58] constructed a LSTM+CRF network to tackle the task of disease named entity recognition. Zhao et al. [59] developed a nerural network based on CNN to recognize disease mention. Zhang et al. [60] constructed a network using LSTM+CRF structure to recognize the named entities in electronic health records. Zhu et al. [61] used n-gram and context as input of CNN to detect named entities in biomedical text. Luo et al. [62] developed an attention based bidirectional-LSTM+CRF model to recognize drug named entity. Sinilarly, Habibi et al. [63] compared LSTM+CRF with pure CRF model, and shown that deep learnin model outperformed traditional machine learning model in the tasks of recognizing gene mention, chemical mention, species mention, and disease mention. Lyv et al. [64] constructed three models based on RNN, RNN+CRF, and BiLSTM+RNN. The three models were compared in the task of gene mention recognition. Experimental results shown BiLSTM+RNN model outperformed the other two.

5. Conclusion

Named entity recognition and normalization are tasks to recognize entities mentioned in natural language text and link them to database IDs. We have conducted a survey of works related to biomedical named entity recognition and normalization, focusing on gene mention recognition and normalization. We believe this work will assist researchers to find the information of their interest and interpret their own research. In the further, we plan to conduct another study on biomedical named entity relation extraction.

References

1. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics. 2005; 6 Suppl 1:S2.

2. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008; 9 Suppl 2: S2.

3. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. ProMiner: Organism-specific protein name detection using approximate string matching. Proceedings of the BioCreative: Critical Assessment for Information Extraction in Biology, Granada, Spain. 2004.

4. Fukuda KI, Tsunoda T, Tamura A, Takagi T. Toward information extraction: identifying protein names from biological papers. Pacific Symposium on Biocomputing. Proceedings of the Pacific Symposium on Biocomputing; 1998. p. 707-718.

5. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. Pac Symp Biocomput. 2003: 427-438.

6. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008: 652-663.

7. Liu H, Torii M, Hu ZZ, Wu C. Gene mention and gene normalization based on machine learning and online resources. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 135-140.

8. Lafferty J, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conference on Machine Learning; 2001 Jun 28 - Jul 01; San Francisco, CA, USA. Morgan Kaufmann Publishers Inc; 2001. p. 282-289.

9. Liu H, Hu ZZ, Zhang J, Wu C. BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics. 2006; 22(1): 103-105.

10. Huang HS, Lin YS, Lin KT, Kuo CJ, Chang YM, Yang BH, et al. High-recall gene mention recognition by unification of multiple backward parsing models. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 109-111.

11. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory; 1992 Jul 27-29; Pittsburgh, Pennsylvania, USA. ACM; 1992. p. 144-152.

12. Chen Y, Liu F, Manderick B. Improving the performance of gene mention recognition system using reformed lexicon-based support vector machine. Margin. 2007; 500: 2.

13. Zhou G, Shen D, Zhang J, Su J, Tan S. Recognition of protein/gene names from text using an ensemble of classifiers. BMC bioinformatics. 2005; 6(1): S7.

14. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 21(14): 3191-3192.

15. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C. Exploring the boundaries: gene and protein identification in biomedical text. BMC Bioinformatics. 2005; 6 Suppl 1: S5.

16. McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics. 2005; 6 Suppl 1: S6.

17. Kinoshita S, Ogren P, Cohen KB, Hunter L. Entity identification in the molecular biology domain with a stochastic POS tagger: the BioCreative task. In: Proceedings of the BioCreAtIvE Workshop; 2004 Mar 28-31; Granada, Spain.

18. Ando RK. BioCreative II genes mention tagging system at IBM Watson. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 101-103.

19. Kuo CJ, Chang YM, Huang HS, Lin KT, Yang BH, Lin YS, et al. Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 105-107.

20. Klinger R, Friedrich CM, Fluck J, Hofmann-Apitius M. Named entity recognition with combinations of conditional random fields. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 89-92.

21. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc. 2009; 16(2): 247-255.

22. Struble CA, Povinelli RJ, Johnson MT, Berchanskiy D, Tao J, Trawicki M. Combined conditional random fields and n-gram language models for gene mention recognition. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 81-83.

23. Baumgartner Jr WA, Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A, et al. An integrated approach to concept recognition in biomedical text. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 257-271.

24. Tsai RT, Sung CL, Dai HJ, Hung HC, Sung TY, Hsu WL. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical

named entity recognition. BMC Bioinformatics. 2006; 7 Suppl 5: S11.

25. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 2003; 19 Suppl 1: i180-i182.

26. Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, et al. Integrated annotation for biomedical information extraction. In: Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting; 2004 May 02-07; Boston, MA, USA. HLT/NAACL; 2004. p. 61-68.

27. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics. 2005; 6 Suppl 1: S11.

28. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome Biol. 2008; 9 Suppl 2: S3.

29. Fluck J, Mevissen HT, Dach H, Oster M, Hofmann-Apitius M. ProMiner: recognition of human gene and protein names using regularly updated dictionaries. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop; 2007 Apr 23-25; Madrid, Spain. CNIO; 2007. p. 149-151.

30. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G. Inter-species normalization of gene mentions with GNAT. Bioinformatics. 2008; 24(16): i126-132.

31. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. Bioinformatics. 2009; 25(6): 815-821.

32. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, et al. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011; 12 Suppl 8: S2.

33. Bhattacharya S, Sehgal AK, Srinivasan P. Cross-species gene normalization at the University of Iowa. In: Proceedings of the BioCreative III workshop; 2010 Sep 13 -15; Bethesda, MD USA. University of Delaware; 2010. p. 55-59.

34. Dai HJ, Lai PT, Tsai RT. Multistage gene normalization and SVM-based ranking for protein interactor extraction in full-text articles. IEEE/ACM Trans Comput Biol Bioinform. 2010; 7(3): 412-420.

35. Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. Bioinformatics. 2011; 27(7): 1032-1033.

36. Wei CH, Kao HY. Cross-species gene normalization by species inference. BMC Bioinformatics. 2011; 12 Suppl 8: S5.

37. Wei CH, Kao HY, Lu Z. SR4GN: a species recognition software tool for gene normalization. PLoS One. 2012; 7(6): e38460.

38. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013; 41(Web Server issue): W518-522.

39. Van Landeghem S, Björne J, Wei CH, Hakala K, Pyysalo S, Ananiadou S, et al. Large-scale event extraction from literature with multi-level gene normalization. PLoS One. 2013; 8(4): e55814.

40. Wei CH, Kao HY, Lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. Biomed Res Int. 2015; 2015: 918710.

41. Leaman R, Wei CH, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S3.

42. Lu Y, Ji D, Yao X, Wei X, Liang X. CHEMDNER system with mixed conditional random fields and multi-scale word clustering. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S4.

43. Lowe DM, Sayle RA. LeadMine: a grammar and dictionary driven approach to entity recognition. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S5.

44. Chowdhury M, Faisal M. (2010). Disease Mention Recognition with Specific Features. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing; 2010 Jul 15; Uppsala, Sweden. Assoc Comput Linguist; 2010. p. 83-90.

45. Kaewphan S, Hakala K, Ginter F. UTU: disease mention recognition and normalization with CRFs and vector space representations. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014); 2014 Aug 23-24; Dublin, Ireland. QCRI; 2014. p. 807-811.

46. Batista-Navarro R, Rak R, Ananiadou S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S6.

47. Munkhdalai T, Li M, Batsuren K, Park HA, Choi NH, Ryu KH. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S9.

48. Akhondi SA, Hettne KM, van der Horst E, van Mulligen EM, Kors JA. Recognition of chemical entities: combining dictionary-based and grammar-based approaches. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S10.

49. Khabsa M, Giles CL. Chemical entity extraction using CRF and an ensemble of extractors. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S12.

50. Xu S, An X, Zhu L, Zhang Y, Zhang H. A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. J Cheminform. 2015; 7(Suppl 1 Text mining for chemistry and the CHEMDNER track): S11.

51. Leaman R, Lu Z. Automated disease normalization with low rank approximations. In: Proceedings of BioNLP; 2014 Jun 27-28; Baltimore, Maryland, USA. ACL; 2014. p. 24-28.

52. Dogan RI, LuZ. An Inference Method for Disease Name Normalization. In: Proceedings of AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text; 2012 Nov 2-4; Arlington, VA, USA. AAAI PRESS; 2012.

53. Kang N1, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 2013; 20(5): 876-881.

54. Lee HC, Hsu YY, Kao HY. An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; 2017 Apr 26-27; Barcelona, Spain. CNIO; 2017. p. 226-233.

55. Deleger L, Grouin C, Bossy R. Hybrid approaches for the DNER task at BioCreative V: the INRA/LIMSI system. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; 2017 Apr 26-27; Barcelona, Spain. CNIO; 2017. p. 154-166.

56. Hemati W, Mehler A. LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools. J Cheminform. 2019; 11(1): 3.

57. Korvigo I, Holmatov M, Zaikovskii A, Skoblov M. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. J Cheminform. 2018; 10(1): 28.

58. Xu K, Zhou Z, Gong T, Hao T, Liu W. SBLC: a hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields. BMC Med Inform Decis Mak. 2018; 18: 114.

59. Zhao Z, Yang Z, Luo L, Wang L, Zhang Y, Lin H, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Med Genomics. 2017; 10: 73.

60. Zhang Y, Wang X, Hou Z, Li J. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR Med Inform. 2018; 6(4): e50.

61. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics. 2017; 34(9): 1547-1554.

62. Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics. 2018; 34(8):1381-1388.

63. Habibi M, Weber L, Neves M, Wiegandt D, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017; 33(14): i37-i48.

64. Lyu C, Chen B, Ren Y, Ji D. Long short-term memory RNN for biomedical named entity recognition. BMC Bioinform. 2017; 18(1): 462.