Published online: 2020-08-21

A Year of Papers Using Biomedical Texts:
Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook

Cyril Grouin[1], Natalia Grabar[1, 2], Section Editors for the IMIA Yearbook Section on Natural Language Processing
[1] Université Paris-Saclay, CNRS, LIMSI, Orsay, France
[2] STL, CNRS, Université de Lille, Villeneuve-d'Ascq, France

Summary

Objectives: Analyze papers published in 2019 within the medical natural language processing (NLP) domain in order to select the best works of the field.

Methods: We performed an automatic and manual pre-selection of papers to be reviewed and finally selected the best NLP papers of the year. We also propose an analysis of the content of NLP publications in 2019.

Results: Three best papers have been selected this year including the generation of synthetic record texts in Chinese, a method to identify contradictions in the literature, and the BioBERT word representation.

Conclusions: The year 2019 was very rich and various NLP issues and topics were addressed by research teams. This shows the will and capacity of researchers to move towards robust and reproducible results. Researchers also prove to be creative in addressing original issues with relevant approaches.

Keywords

Natural Language Processing; social media; state-of-the-art review; Artificial Intelligence

Yearb Med Inform 2020:221-5
http://dx.doi.org/10.1055/s-0040-1701997

1 Introduction

Natural Language Processing (NLP) aims at providing methods, tools and resources designed to mine textual and narrative documents, and to make it possible to access the information they convey [1]. While human languages are complex (as an example, learning a human language requires many years in order to be fluent), the importance of using NLP approaches to mine medical and health documents produced by humans has been pointed out since a long time [2]. In this synopsis, we present the selection process applied in 2020 to select the best NLP papers published in 2019, and we provide an analysis of the content of relevant publications. More particularly, we focus on several important issues such as robustness of methods, reproducibility of results, as well as the trends and originality of the research questions addressed in 2019.

2 Selection Process

In order to identify all papers published during the year 2019 in the field of Natural Language Processing, we queried two databases: Medline[1], specifically dedicated to the biomedical domain, and the Association for Computational Linguistics (ACL) anthology[2], a database that brings together the major NLP conferences (ACL, Coling, Empirical Methods in Natural Language Processing (EMNLP), Language Resources and Evaluation Conference (LREC), North American Chapter of the Association for Computational Linguistics (NAACL), etc.) and journals, since some NLP studies concerning the biomedical domain are published in conferences and journals which are not indexed by PubMed. We applied the basic query we defined last year for MEDLINE (Figure 1) to retrieve journal papers published in English in 2019, having abstract, and composed of the sequences "clinical language processing" or "medical language processing" or "natural language processing".

As of 2020, January 9th, we collected 767 entries. We applied a similar query on the ACL anthology database and collected 10 additional entries. In order to process those 777 papers, we automatically scored the papers: indeed, all candidate papers are not specifically related to the NLP domain despite the use of one of the three sequences from the query. For instance, they may be related to other sections of the IMIA Yearbook of Medical Informatics (*e.g.*, Public Health and Epidemiology Informatics, Bioinformatics and Translational Informatics, Knowledge Representation and Management) and not address the major issues of the NLP section.

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://www.aclweb.org/anthology/

Hence, we applied three sets of rules that we previously defined to identify best papers published in 2018, in order to compute global scores for each publication.

The first set of rules is based on the name of the journal (both full name and concepts found in the name of the journal):

- a positive score is assigned to the main journals in which biomedical NLP research is usually published by the NLP community (Biomedical Informatics insights, International Journal of Medical Informatics, Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, BMJ Bioinformatics);
- a negative score is assigned to journals not specifically related to NLP, but to other domains such as Cognitive studies and Communication disorders (*e.g.*, Neuroscience, Human brain mapping, Operative neurosurgery, Speech therapy). We manually defined this set of journals in order to rule out those false positives.

The second set of rules relies on the concepts found in both the title and the abstract of papers:

- a positive score is assigned to concepts typically involved in papers related to NLP. Those concepts may be related to objectives, resources, and tools (such as *natural language processing (NLP), named entity recognition (NER), part of speech (POS), tagged words, semantic, syntax, biomedical entity, meanings, electronic health record (EHR), reports, notes, clinical text, text corpus, free text, unstructured text, tweets, PubMed, Social Media, MedDRA, UMLS, annotated data, Metamap*);
- a negative score is assigned to concepts that are usually involved in studies related to disorders involving anatomical parts or language abilities (such as *word processing, language production, language comprehension, voice quality, posterior superior temporal gyrus (pSTG), posterior superior temporal sulcus (pSTS), inferior fronto-occipital fasciculus (IFOF), dorsolateral prefrontal cortex, cortex, language lateralization, chemical fragment, fragment chemistry, brain structures, verbal intelligence, cerebral, positive mismatch responses (pMMRs), prelingual, postlingual, cochlear, aphasia, SAPS, cortical, language function, infants*).

The third set of rules is also applied on the titles and the abstracts, and targets the concepts describing the methodology used in papers:

- a positive score is assigned to papers using classical NLP methods or evaluation metrics (such as *annotation tool, text-mining, rule-based, regular expression, lexicon, conditional random fields (CRFs), recall, precision, F1-score, F-measure, accuracy, inter-annotator agreement, Kappa, classify/classifier, detect, extract, extraction, predict, predicting, text simplification, lexical simplification*);
- a negative score is assigned to papers claiming to use the NLP methods, such as pointed out by sequences like *using natural language processing, using NLP,* or *perform a Natural Language Processing analysis*. Such papers are downgraded because NLP claims are usually limited to the use of existing and ready-to-use NLP tools, while the main contribution of papers is related to the analysis of tool results rather than to the improvements made to NLP methods and issues. Notice that such papers are usually related to other areas from biomedical informatics: researchers take advantage of existing tools.

For each of the 777 candidate papers, the final score ranked from 0.25 to 0.9 (cf. Figure 2). On this Figure, the violet bars indicate the total number of papers for each computed grade, while the pink bars indicate the papers we kept in the top 15 best papers list. This score has been used as a meta-element during the manual selection of the top 15 papers. Indeed, section editors did not fully rely on the scores but only used them as additional information. Hence, for each of the 777 papers, both section editors independently browsed the abstracts, keywords, and automatic scores, and then assigned the *Yes / Maybe / No* score of inclusion into the IMIA Yearbook as candidate best papers. All papers having at least one *Yes* or *Maybe* score have been kept for the next step of the selection. At this stage, 48 candidate papers remained (*i.e.*, a subset of 6.3% of the whole dataset). We then performed an adjudication process in order to choose the final 15 candidate best papers to be proofread by external reviewers. We paid attention to the topics addressed by researchers and to their geographic origin so as to provide enough diversity. As a result, out of the 15 papers, seven come from the USA, three from China, and one from each of the following countries: Belgium, France, Italy, Spain, and South Korea. In the next section, we present the main issues and approaches addressed in the 15 preselected publications.

```
(English[LA] AND journal article[PT] AND 2019[dp] AND ((medical OR
clinical OR natural) AND "language processing"))
```

Fig. 1 Query used for collecting candidate papers for review

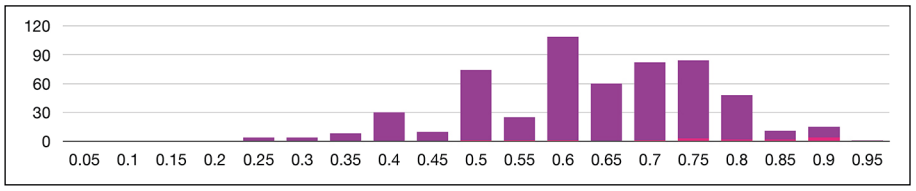

Fig. 2 Distribution of papers according to the filter scores (violet bars concern the total number of papers and pink bars concern papers kept in the top-15 best papers).

3 Principal Findings

3.1 The Languages Addressed

We identified 78 papers in which the processed language was clearly indicated in the abstract. Among the languages found, we observe the following distribution:

- English was still the first language considered in studies with 26 explicit mentions. Yet, we can consider that the papers that did not explicitly indicate the language should also be dedicated to the processing of data in English ;
- Chinese became the second language processed in medical NLP papers with 17 mentions. Among the papers published in 2019, we can mention Guan *et al.* [3] working on the generation of synthetic medical record texts, Chen *et al.* [4] aiming at identifying named entities, and Zheng *et al.* [5] interested by the detection of medical text similarity ;
- French was the third language (seven mentions) as in the work by Lerner *et al.* [6], followed by three other European languages with less than five mentions: German, Italian [7], and Spanish [8] ;
- Other languages identified in the abstracts accounted for one or two papers and included both languages spoken by millions of people (Arabic, Portuguese, Russian) and languages spoken by small communities (Basque, Danish, Japanese, Korean, Lithuanian, Persian, Romanian, Turkish, and Urdu).

Comparing to the previous year, trends were modified in 2019: if English was still the first language mentioned in papers (18 mentions), German was the second language (nine mentions), and Chinese, French, Italian, and Japanese were following with three mentions each. Thus, we can observe a noticeable increase of papers dealing with data in Chinese and the emergence of works dedicated to other languages. We expect that these trends will be developed in the future.

3.2 NLP and Application Contexts

The main issue when performing NLP research in the biomedical domain is the access to data. As a consequence, social media, such as Reddit and Twitter, are still widely used because of their easy access by researchers. Besides, specific resources have been developed from and for social media, due to the specificity of this kind of data. Let's mention the work by Lavertu and Altman [9] who designed the Redmed system in order to produce a specific drug lexicon to be used in social media applications. When clinical data are available and accessible to researchers, NLP technics may be applied on all types of textual data: electronic health records for the identification of adverse drug events [8], discharge summaries [10], and more rarely triage notes [11] for performing named entity recognition.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 in the section 'Natural Language Processing'. The articles are listed in alphabetical order of the first author's surname.

Section Natural Language Processing
▪ Guan J, Li R, Yu S, Zhang X. A Method for Generating Synthetic Electronic Medical Record Text. IEEE/ACM Trans Comput Biol Bioinform 2019.
▪ Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234-40.
▪ Rosemblat G, Fiszman M, Shin D, Kılıçoğlu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. J Biomed Inform 2019;98:103275.

3.3 Original Issues

Among the preselected papers, we can draw two main final objectives for which NLP methods are used.

The first objective, which is present in a large amount of published papers, addresses the improvement of the medical care process. Hence, papers published in 2019 focus on the identification of patients with obesity and several comorbidities from clinical texts [10], the prediction of emergency department patient disposition from triage notes [11], the identification of drug discontinuation events from EHR [12], and the help for monitoring patients in intensive care unit (ICU) [13].

The second objective can be characterized as "the research for the research". Hence, NLP researchers are using NLP methods to improve access to the knowledge contained in scientific papers. In this perspective, Rosemblat *et al.* [14] designed a methodology to identify apparent contradictions in the literature. They applied their method automatically on a sub-set of scientific papers (related to around 20 common diseases and pathologies, signs or symptoms) and identified five types of contradictions among which 58 were real ones.

3.4 Original Methods and Approaches

Neural networks and word embeddings are now widely used to process data from the biomedical domain, and this year the survey paper of the IMIA Yearbook NLP section also addresses this issue [15]. For instance, among the 2019 papers, Lee *et al.* [16] trained a Bidirectional Encoder Representations from Transformers (BERT) model on biomedical data in order to produce the BioBERT resource, which is a word representation specifically tuned to process biomedical data now widely used within the Medical Informatics area. Chen *et al.* [4] performed a named entity recognition using several models trained through BiLSTM, while Si et al.[17] produced contextual embeddings to improve their concept extraction method. A similar idea to make NLP methods more robust consists in using semantic composition to extract concepts from clinical texts [18]. Overall, the use of BioBERT word embeddings and of neural network methods allows to improve results on several tasks dedicated to named entity recognition, relation extraction, and question-answering, such as experienced by Lee *et al.* [16].

4 Conclusion

We identified 777 papers published in 2019 that involved the use or the application of NLP methods and tools in the biomedical domain. After a first rapid manual reviewing process, we obtained a short list of 49 candidates which undergone a human consensus in order to identify the 15 best candidate papers. Those papers have been peer-reviewed by a set of external reviewers. Based on the evaluation of these reviewers, we selected the three best papers of the NLP section. Out of the main findings from papers published in 2019, we observed an important increase of papers dealing with data in Chinese. As for the methodological issues, word embeddings tailored for the biomedical domain (BioBERT) and neural networks will certainly result in an increasing number of publications in the years to come. Due to the exceptional sanitary situation in 2020, which witnessed the emergence and expansion of the Covid-19 pandemics through the planet, we also expect that a huge number of publications in 2020 will specifically focus on pandemics, viruses, and the Covid-19 more particularly. Through different initiatives and needs that emerged from the clinical, research, and industrial areas, we expect that these publications will deal with various related research questions, and mainly with (1) the identification of key findings in all types of data for improving the research for a vaccine development and use, and (2) the early detection and prevention of pandemics.

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural Language Processing: an introduction. J Am Med Inform Assoc 2011;18:544–51.
2. Friedman C, Hripcsak G. Natural Language Processing and its future in medicine. Acad Med 1999;74(8):890–5.
3. Guan J, Li R, Yu S, Zhang X. A method for generating synthetic electronic medical record text. IEEE/ACM Trans Comput Biol Bioinform 2019.
4. Chen Y, Zhou C, Li T, Wu H, Zhao X, Ye K, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. J Biomed Inform 2019 Aug;96:103252.
5. Zheng T, Gao Y, Wang F, Fan C, Fu X, Li M, et al. Detection of medical text semantic similarity based on convolutional neural network. BMC Med Inform Decis Mak 2019;19(1):156.
6. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. J Biomed Inform 2020 Feb;102:103356.
7. Viani N, Miller TA, Napolitano C, Priori SG, Savova GK, Bellazzi R, et al. Supervised methods to extract clinical events from cardiology reports in italian. J Biomed Inform 2019;95:103219.
8. Santiso S, Pérez A, Casillas A. Smoothing dense spaces for improved relation extraction between drugs and adverse reactions. Int J Med Inform 2019;128:39–45.
9. Lavertu A, Altman RB. Redmed: Extending drug lexicons for social media applications. J Biomed Inform 2019;99:103307.
10. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmusen LV, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 2019;99:103310.
11. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. Int J Med Inform 2019;129:184–8.
12. Liu F, Pradhan R, Druhl E, Freund E, Liu W, Sauer BC, et al. Learning to detect and understand drug discontinuation events from clinical narratives. J Am Med Inform Assoc 2019;26(10):943–51.
13. Khadanga S, Aggrawa K, Joty S, Srivastava J. Using clinical notes with time series data for ICU management. In ACL, editor. Proceedings of EMNLP 2019. Hongkong, China; 2019. p. 6431–6.
14. Rosemblat G, Fiszman M, Shin D, Kilicoglu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. J Biomed Inform 2019;98:103275.
15. Hahn U, Oleynik M. Medical information extraction in the age of deep learning: Methodological foundations and neural network applications. 2020 Yearb Med Inform (in press).
16. Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234-40.
17. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019;26(11):1297–304.
18. Tulkens S, Šuster S, Daelemans W. Unsupervised concept extraction from clinical text through semantic composition. J Biomed Inform 2019;91:103120.

Correspondence to:
Cyril Grouin
Université Paris-Saclay
CNRS, LIMSI
Campus universitaire
91405 Orsay
France
E-mail: cyril.grouin@limsi.fr

Content Summaries of Best Papers for the Natural Language Processing Section of the 2020 IMIA Yearbook

Guan J, Li R, Yu S, Zhang X

A Method for Generating Synthetic Electronic Medical Record Text

IEEE/ACM Transact on Comput Biology and Inform 2019

The main problem to perform Natural Language Processing in the biomedical domain is the access to clinical texts for non-medical staff, and more accurately for languages other than English. This paper presents a method based on neural networks (GAN + reinforce algorithm) to produce clinical documents in Chinese, for a given disease (either pneumonia or lung cancer). The authors used a corpus of 2,216 clinical notes written in Chinese, using the 'History of Present Illness' section as input and the 'Admission Diagnosis' section as tags. The authors report an accuracy of 0.7635 for generated data. They also defined three types of errors in their generated content: repetitions, inconsistent content ("temperature of 39.5°C; no fever"), and improper word matching ("body temperature paroxysmal cough").

Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, Kang J

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Bioinformatics 2019;36(4):1234-40

Current NLP methods rely on word representations to improve results, among which BERT is the most commonly used resource. Nevertheless, while general resources exist, a domain-specific language needs specific resources. This paper introduces BioBERT, a BERT model tuned for the biomedical domain. In order to produce this model, the authors used several corpora in English (Wikipedia, BooksCorpus, PubMed abstracts, and PMC full texts). They compared results achieved by the BioBERT model with the BERT general model on three tasks (named entity recognition, relation extraction, and question-answering). For each task, better results were achieved when using the BioBERT model.

Rosemblat G, Fiszman M, Shin D, Kılıçoğlu H

Towards a characterization of apparent contradictions in the biomedical literature using context analysis

J Biomed Inform 2019;98:103275

This paper aims at identifying contradictions in scientific papers. The authors defined five categories of contradictions: (a) internal to patient, such as comorbidities, (b) external to patient, such as dosage, (c) endogenous and exogenous, (d) known controversy, and (e) contradictions in literature. They used the SemRep tool to identify relationships between 20 common diseases and pathologies, or sign or symptoms. Then, they assessed the level of certainty based on the SemMedDB repository (from PubMed) which contains subject-relation-object predications. On 117,000 instances (from 62,000 abstracts), they identified 2,236 apparent contradictions, among which 58 contradictions were real ones.