Pre-Training to Identify Immunization-Related Entities from Systematic Reviews

Bahar İlgen
Center for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Nordufer 20, Berlin, 13353, Germany
ilgenb@rki.de

Antonia Pilic
Immunization Unit, Robert Koch Institute, Seestrasse 10, Berlin, 13353, Germany
pilica@rki.de

Thomas Harder
Immunization Unit, Robert Koch Institute, Seestrasse 10, Berlin, 13353, Germany
hardert@rki.de

Georges Hattab
Center for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Nordufer 20, Berlin, 13353, Germany; Department of Mathematics and Computer science, Freie Universität, Arnimallee 14, Berlin, 14195, Germany
hattabg@rki.de

ABSTRACT

Entity recognition from semi or unstructured systematic reviews is one of the most essential processes for evidence-based decision-making systems. The task involves collecting information from diverse studies concerning PICO (Population, Intervention, Comparison, and Outcomes) elements with additional domain-related information using named entity recognition (NER) as it is the fundamental task for extracting the structured data. In this study, we create an adapted immunization-related dataset and evaluate its performance in the extraction of relevant entities from systematic reviews. We conducted experiments to investigate several models for entity recognition performance using language models pre-trained in the biomedical domain. Our results suggest that PubMedBERT and BertNER results are superior to the other models, and the immunization-related entities can be successfully recognized with a 76% F1 score and 92% accuracy.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial Intelligence**; • **Natural language processing**; • **Information extraction**;

KEYWORDS

Named entity recognition, Systematic reviews, Immunization, Vaccination, BERT, PICO, Evidence based medicine

ACM Reference Format:
Bahar İlgen, Antonia Pilic, Thomas Harder, and Georges Hattab. 2023. Pre-Training to Identify Immunization-Related Entities from Systematic Reviews. In *2023 7th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2023), December 15–17, 2023, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3639233.3639355

NLPIR 2023, December 15–17, 2023, Seoul, Republic of Korea

ACM ISBN 979-8-4007-0922-7/23/12
https://doi.org/10.1145/3639233.3639355

1 INTRODUCTION

Evidence-based medicine (EBM) aims to improve medical outcomes based on the highest quality evidence available to support healthcare decision systems. It combines the best available evidence, clinical expertise, and patient values to enhance patient outcomes, minimize the risk of harm, and optimize resource utilization by focusing on effective interventions [1-2]. Employing EBM in public health management plays a key role in making the most effective decisions in a reasonable time, especially on subjects such as immunization. National Immunization Technical Advisory Groups (NITAGs) play a pivotal role in informing immunization policy and program decisions by delivering evidence-based recommendations on immunization-related issues to policymakers and program managers in their countries [3]. These recommendations are developed through a rigorous process drawing upon empirical evidence, with systematic reviews serving as valuable sources of information [4]. Health-related information systems regulate the activities with extraction, assessment, monitoring, and interpretation of beneficial information from unstructured ad hoc systems regarding health events or risks that may represent a potential and acute risk to human health [5]. The main activities of these systems consist of employing tasks such as text classification, geographical and temporal analysis, temporal information extraction, and text summarization. Appropriate organization of articles is critical to prioritize the existing and more urgent threat types. Geographical analysis of health-related events helps determine the threatened and in-risk populations. Temporal information extraction and reasoning are crucial tasks to distinguish between cases in the temporal domain. These systems regularly extract and report new cases to monitor the progress of a particular disease. Summarization systems provide easy access to essential information by extracting it from a large volume of text. This task successfully reduces the time required to determine the threat type [6]. One of the leading issues for evidence-based surveillance systems is the exponential growth

of health-related texts. From this perspective, the development of efficient *Information Retrieval* (IE) and *Natural Language Processing* (NLP) methods gained a significant role in the process. In the scope of this study, we focus on extracting immunization-related information through systematic reviews. Systematic reviews, comprehensively summarize the evidence in a given area, use systematic and transparent methods to identify all studies that are potentially relevant to a research question, select studies for inclusion, appraise the quality of included studies, and synthesize study results [7]. The PICO (Population, Intervention, Comparison, Outcome) framework is widely accepted to formalize essential information through the IE process of systematic reviews. Formulating the relevant questions is an initial step in the IE domain [8]. The population refers to the targeted population for the vaccination as characterized by age groups, sex, ethnicity, etc. Intervention is the intervention being considered for a particular case (e.g., a dose or type of a vaccine). Comparison (or absence of exposure) represents the alternative intervention or "action" to be compared (e.g., placebo, no vaccination, other prevention options). The outcome addresses the positive or negative (in terms of harm or adverse events) endpoints.

Named entity recognition (NER) is one of the fundamental tasks in NLP to automate the recognition and extraction of a set of predefined information known as named entities from semi or unstructured large volumes of text. These entities refer to the fundamental information found in predefined categories such as *Person* (PER), *Organization* (ORG), *Location* (LOC), *Geographical* (Geo), *Time* (Tim), *Event* (Eve), *Monetary* (Mon) values as well as other domain-specific entities. Apart from being a standalone application, the NER task improves the accuracy of several other NLP tasks functioning as automatic summarization, machine translation, natural language understanding, sentiment analysis, chatbots, text simplification systems, search engines, and knowledge base construction. A NER task is approached using different methodologies involving (1) Rule-based methods rely on a set of hand-crafted rules with no need for annotated data. (2) Feature engineering and supervised learning algorithms, (3) unsupervised algorithms employed using unannotated data, (4) deep learning algorithms to automatically discover representations required for the classification and detection from raw input in an end-to-end manner. Recently, deep learning approaches attracted significant interest among domain researchers and achieved state-of-the-art performance on entity recognition tasks [9-10]. However, these methods are usually built upon large, high-quality labeled data which is labor intensive as it requires expertise in the area. Transfer learning and pre-training paradigms in the context of large language models have widely been accepted to overcome the lack of training data and help boost the performance of many downstream tasks such as named entity recognition, sentiment analysis, question answering, and several NLP tasks.

In this study, we perform named entity recognition on systematic reviews to extract relevant information on the intersection of public health and the immunization domain. We use a PICO dataset compiled from PubMed articles. Since not all the abstracts are in the immunization domain, we also prepare and annotate a subset of the dataset to introduce our immunization-related entities. We fine-tune the transformer models [11] for the NER task and present the performance on the dataset. Since most of the previous studies on NER and PICO extraction focused on sentence-level annotation, a newly adapted PICO dataset using word-level annotations will be useful for future work, especially for analyzing more complex inner word relations. The remaining sections are organized as follows: Section 2 reviews the related work on information extraction and event-based decision-making systems. Section 3 provides the methodology and the model we follow in the study. Section 4 is the experimental setup, and Section 5 concludes the study.

2 RELATED WORK

Systematic reviews aim to reach and identify as many relevant studies and documents as possible to answer a proposed research question. NER is one of the fundamental methods in text processing for the biomedical domain to recognize essential information, drug names, protein and gene names, and many others. Supervised learning methods approach the NER problem as a multi-class classification and sequence labeling task. They benefit from feature engineering where features are abstracted over texts and represented using Boolean, numeric, nominal values, or word-level features such as part-of-speech tags [12]. Hidden Markov Models (HMMs) [13], Conditional Random Fields (CRFs) [14], Decision Trees [15], and Support Vector Machines [16] are some of the ML algorithms experimented in the scope of feature-based approaches. Previous studies model the task of PICO element identification as a sequence labeling task. In [8], bidirectional long short-term memory (LSTM) [17] is adopted as the base model and decoded with a linear chain CRF in the output layer. Bidirectional long short-term memory (Bi-LSTM) [17-18] is used to learn vector representations and as an input to conditional random fields (CRF) [14]. Deep neural network approaches have gained attraction and success as they can learn parameters in an end-to-end fashion without the need for feature engineering. In this regard, they are preferred over the traditional ML models in recent studies. The fully connected self-attention architecture (i.e., transformers) [11] is widely accepted since they have the ability of parallelism to model the long-range context. The transformer model solely relies on the use of a self-attention mechanism where the representation of sequence (or sentence) is computed by relating different words in the same sequence. In the scope of more recent studies, language models such as ELMo (Embeddings from language models) [19], and bidirectional encoder representations from transformers (BERT) [20] achieved state-of-the-art performance. BERT and ELMo learn contextual embeddings, and ELMo uses Bi-LSTM as an encoder. We conduct experiments using language models and BioBERT [21] which is a BERT model pre-trained on biomedical corpora. These models are trained on large-scale corpora such as Wikipedia and BookCorpus [22] and achieve promising results on tasks including natural language understanding and generation [20].

3 METHODOLOGY

3.1 Task Definition and Model Details

A word or a word phrase is defined as a named entity (NE) if it identifies an item from a group of other items sharing similar attributes [23]. Given a sequence of tokens $s = \{w_1, w_2, \ldots, w_n\}$, where w_i is the i-th word/token and N represents the length of the sentence, the aim of the NER task is to categorize each word/token in S. It assigns it to an appropriate label $t \in T$, where T is a predefined list of all

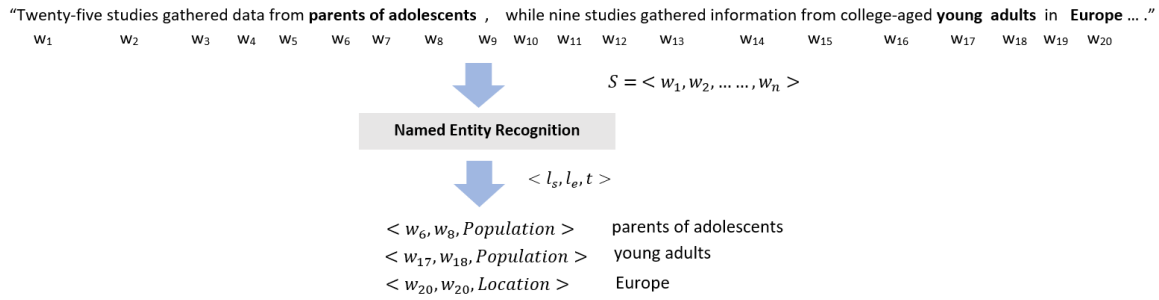

Figure 1: NER task definition.

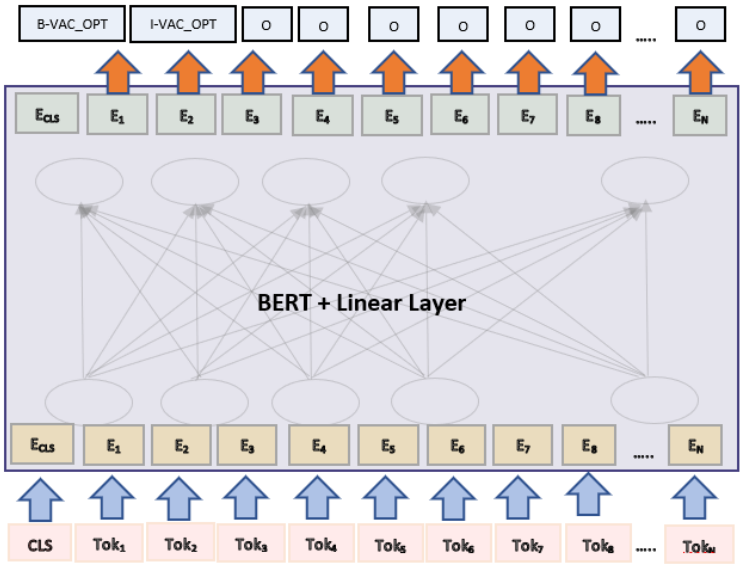

Figure 2: Using BERT to perform NER task on Biomedical domain.

available label types (*e.g., disease, location, organization, vaccine type, and more*) and outputs a list of tuples {I_s, I_e, t} where each list represents a named entity mentioned in S. Figure 1 illustrates the task of NER where $I_s \in [1, N]$ and $I_e \in [1, N]$ are words with given indices and t is the entity type from a predefined category set.

We exploited the BERT model [20] as our model backbone and used a pre-trained model from HuggingFace [24]. BERT is a multi-layer transformer encoder-decoder model implemented using a self-attention mechanism [11] and pre-trained by combining masked language modeling and next-sentence prediction tasks. BERT has led to successful results in many NLP tasks regarding text classification tasks. The original BERT pre-trained on 800M words of BookCorpus [22] and English Wikipedia with 2500M words. On the other hand, BioBERT is initialized with weights from the original BERT and then pre-trained on PubMed Abstracts and PubMed Central Full-text articles. In this research, we opted to recognize entities in the biomedical domain, so we employed BioBERT [21] and variants. BERT can be fine-tuned on tasks that use the whole sentence to make decisions, such as sequence classification, token classification, or question answering. The PICO classification model is usually employed as a sentence classification task in previous studies [25]. In this study, we approach the PICO classification at the word level and use BERT for token classification. Figure 2 shows the BioNER task implemented using BERT. To use BERT for a token classification task such as NER, we adjust the embedding vector to get output from all the tokens. In this study, we used pre-trained models including BioBERT, BERTNer, PubMedBERT, ClinicalBERT, and BlueBERT.

4 EXPERIMENTAL SETUP

4.1 Dataset

PICO elements are widely used in the biomedical domain to recognize entities that carry the essential information among semi-structured or unstructured texts. Although it is a generally accepted and convenient framework, related studies have an extensive scope

Table 1: An example IOB-Tagging in the systematic review dataset.

Tokens	IOB-Tag
Adherence	O
to	O
lifestyle	O
recommendations	O
by	O
patients	B-PARTICIPANT
with	I-PARTICIPANT
depression	I-PARTICIPANT
.	O

covering subject areas from names in clinical notes to protein and/or gene names. In this regard, we prepared our immunization-focused dataset to merge it with a general PICO dataset of retrieved PubMed abstracts. Hence, in addition to the existing entities, we prepared the additional entities involving population, outcome, disease, options for vaccination, and immunization topics. These additional entities are determined based on the manual process of extraction and annotation process from systematic reviews by the project team. We included texts from the dataset [25] that is basically extracted from PubMed abstracts and obtained a merged dataset of 10,000 samples with the newly created portion of 3,000 entity mentions. The annotation of entity mentions of the immunization dataset was made by annotators who are domain experts and computer scientists.

We used the inside-outside-beginning (IOB) tagging format for the tokens to represent the IOB information for the entities. The B-prefix before a tag refers to the beginning of a chunk while I-prefix indicates that the tag is inside a chunk. "O" represents that the token does not belong to a chunk. The data frame we used consists of texts and labels where the label information corresponds to the predefined categories of entities for each word in the text. The following sentence given in Table 1 is an example of the tagging scheme for the PARTICIPANT entity that refers to specific participant groups. The overall dataset consists of entity types of POPULATION, LOCATION, DISEASE, OUTCOME, INTERVENTION, VACCINE_OPTION, and PARTICIPANT. Table 2 shows the main categories of our adapted dataset. In Table 3, the statistics of the dataset are given.

4.2 Investigation of Different Models on the Performance

Pre-training large neural language models, such as BERT, has led to impressive gains on many NLP tasks. However, most of the efforts focus on the general scope of corpora, such as Wikipedia and the Web. A pre-trained BERT model can be applied as the token classification task for NER by reinitializing the output layer and fine-tuning the model on NER data. Hence, we conducted experiments to investigate several models for the performance of BERT-NER and compare the models in the biomedical domain, i.e., BioBERT [21], PubMedBERT, ClinicalBERT, and BlueBERT within the Hugging Face Transformers framework [24]. BioBERT is a biomedical language representation model designed for biomedical text mining tasks such as named entity recognition, relation extraction, question answering, and so on. ClinicalBERT is built

Table 2: Immunization related entities

Entity Type	Definition	Sample
POPULATION	**Target population.** All age groups, newborn, children, adolescents, adults, elderly, pregnant women, healthcare workers, parents/caregivers, travelers.	"Twenty-five studies gathered data from **parents of adolescents**, while nine studies gathered information from college-aged **young adults** and. . .."
TOPIC	**Immunization topics.** Efficacy/effectiveness, immunogenicity, safety, acceptance, coverage, administration, economic aspects, ethical issues, logistics, modelling	"Barriers to Human Papillomavirus Vaccine uptake among racial/ethnic minority groups. . ."
DISEASE	**Disease, pathogen.** HPV etc.	"Moreover, little is known about barriers to **HPV vaccination** in racial/ethnic minority groups."
LOCATION	**Geographical region.** WHO region, Country.	"Cost-effectiveness of human papillomavirus vaccine in **China** were included for . . .."
OUTCOME	**Outcomes.** Infection, hospitalization, death, other	"We planned to investigate the effect of an HPV catch-up vaccination **on overall and cancer mortality**, and on cervical cancer incidence."
VACCINE_OPTION	**Type of vaccine.** Live, non-live, adjuvants, non-adjuvanted, etc	"Impact and effectiveness of the **Quadrivalent** Human Papillomavirus vaccine. . .."

Table 3: Dataset Statistics

	Train	Validation	Test	Total
TEXTS #	8,000	1,000	1,000	10,000
%	80.0	10.0	10.0	100
SENTENCES #	76,273	10,037	9,944	96,254
%	79.25	10.42	10.33	100
TOKENS #	1,755.891	233,572	230,277	2,219.740
%	79.11	10.52	10.37	100
NEs #	427,564	53,638	51,740	532,942
%	80.22	10.06	9.72	100

Table 4: Hyper-parameters used in the experiments.

Hyper-parameter	Value
Optimizer	Stochastic gradient descent (SGD)
Learning rate	5e-3
Batch size	8-16

Table 5: Experimental results (%) of five models.

Models	F1	Acc
Bert-base-NER	0.72	0.91
BioBERT	0.70	0.90
BlueBERT	0.72	0.90
PubMedBERT	0.73	0.92
ClinicalBERT	0.69	0.88

and proposed to demonstrate the impact of using clinical-specific contextual embeddings [26]. BlueBERT [27] was pre-trained and fine-tuned for five NLP tasks (i.e., relation extraction, sentence similarity, NER, and document classification and similarity) and ten datasets.

4.3 Experimental Results

We evaluate the performance of NER using exact matching criteria (i.e., correct identification type for each token) for entity types in terms of macro F1 and accuracy scores. The BERT model was fined-tuned for 5 epochs using the initial learning rate of 5.10-3 and set to batch size between 8 to 16. Table 4 summarizes the parameters we used in the experiments. Table 5 presents the performance of NER models on the test set. We also fine-tuned the models using immunization-related subsets only and evaluated them on a small set of test instances shown in Fig 3. In addition to the precision, recall, and F1 scores shown in Figure 3, Table 6 displays the accuracy results of the experiment for all matching entities. Table 7 displays entity extraction samples predicted by different models used in the experiments. Our experiments show that the disease, population, and location entities are better recognized than the Vaccine Option and Topic entities. On the other hand, the size of the test set that we used in this additional experiment was relatively limited and the performance of these categories performs well on larger sets. Additionally, Bert-NER and PubMedBERT scores outperform the other model results in general.

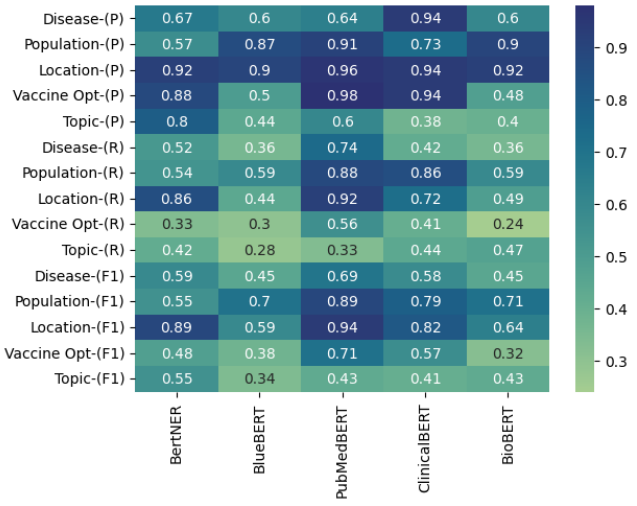

Figure 3: Entity based scores of models on test set.

5 CONCLUSION

This study presents an adapted dataset built for the PICO extraction task on immunization and the public health domain with experimental evaluation results for the NER task. The PICO framework is widely used in the health and biomedical domain. However, there are not enough resources as publicly available annotated datasets, particularly datasets focusing on immunization topics. The entities in our dataset include the information for population, immunization topics, and location. Although most of the available datasets utilize sentence-level annotations for the PICO retrieval task, we opted to annotate PICO elements at the word and phrase level. To automate the extraction of immunization-related entities, we used state-of-the-art learning models, with the best-performing model yielding an exact match of 92% accuracy and 76% F1 score. Our results for the involved language models have shown that the entities can be successfully recognized for a token classification task. Both sets of experiments yielded consistent results in terms of model performance and entity type score. In our future work, we aim to focus on developing fine-grained NER models besides extracting

Table 6: Experimental results of five models for all matching entities on the test set.

Models	BioBERT	Bert-base-NER	PubMedBERT	ClinicalBERT	BlueBert
Accuracy (%)	0.70	0.74	0.76	0.75	0.70

Table 7: Entity extraction by different models

Model	**Text**
Bert-NER	`The Cost Effectiveness of Human` **Papillomavirus Vaccines** `.` `['O', 'O', 'O', 'O', 'O', 'I-DISEASE', 'I-DISEASE', 'O']`
BlueBert	`The Cost Effectiveness of` **Human Papillomavirus** `Vaccines .` `['O', 'O', 'O', 'O', 'I-DISEASE', 'I-DISEASE', 'O', 'O']`
PubMedBERT	`The Cost Effectiveness of` **Human Papillomavirus Vaccines** `.` `['O', 'O', 'O', 'O', 'B-DISEASE', 'I-DISEASE', 'I-DISEASE', 'O']`
ClinicalBERT	`The Cost Effectiveness of Human` **Papillomavirus** `Vaccines .` `['O', 'O', 'O', 'O', 'O', 'I-DISEASE', 'O', 'O']`
BioBERT	`The Cost Effectiveness of Human Papillomavirus Vaccines .` `['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']`

the relations between immunization-related entities. We hope that the results of this study will be useful for our future experiments to construct more complex inner entity relations. The dataset will be publicly available with appropriate licensing.

ACKNOWLEDGMENTS

The authors thank the financial support granted by the Germany Federal Ministry of Health (BMG) under grant No.: ZMI5-2523GHP027 (project "Strengthening National Immunization Technical Advisory Groups and their Evidence-based Decision-making in the WHO European Region and globally", SENSE) part of the Global Health Protection Programme, GHPP.

REFERENCES

[1] Sackett, D. L. (1997, February). Evidence-based medicine. In Seminars in perinatology (Vol. 21, No. 1, pp. 3-5). WB Saunders.
[2] Masic, I., Miokovic, M., & Muhamedagic, B. (2008). Evidence based medicine–new approaches and challenges. Acta Informatica Medica, 16(4), 219.
[3] https://www.who.int/europe/groups/national-immunization-technical-advisory-groups-(nitags)
[4] Bero, L. A., & Jadad, A. R. (1997). How consumers and policymakers can use systematic reviews for decision making. Annals of internal medicine, 127(1), 37-42.
[5] World Health Organization. (2014). Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version (No. WHO/HSE/GCR/LYO/2014.4). World Health Organization.
[6] Ng, V., Rees, E. E., Niu, J., Zaghool, A., Ghiasbeglou, H., & Verster, A. (2020). Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. Canada Communicable Disease Report= Releve des Maladies Transmissibles au Canada, 46(6), 186-191.
[7] Pilic, A., Reda, S., Jo, C. L., Burchett, H., Bastías, M., Campbell, P., ... & Harder, T. (2023). Use of existing systematic reviews for the development of evidence-based vaccination recommendations: Guidance from the SYSVAC expert panel. Vaccine, 41(12), 1968-1978.
[8] Kang, T., Zou, S., & Weng, C. (2019). Pretraining to recognize PICO elements from randomized controlled trial literature. Studies in health technology and informatics, 264, 188.
[9] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016, June). Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 260-270).
[10] Jin, D., & Szolovits, P. (2018, July). Pico element detection in medical text via long short-term memory neural networks. In Proceedings of the BioNLP 2018 workshop (pp. 67-75).
[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[12] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.
[13] Eddy, S. R. (1996). Hidden markov models. Current opinion in structural biology, 6(3), 361-365.
[14] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
[15] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.
[16] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.
[17] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[18] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
[19] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations." arXiv, v2, March 22. Accessed 2019-10-14.
[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.
[21] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
[22] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19-27).
[23] Sharnagat, R. (2014). Named entity recognition: A literature survey. Center for Indian Language Technology, 1-27.
[24] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
[25] Jin, D., & Szolovits, P. (2018, July). Pico element detection in medical text via long short-term memory neural networks. In Proceedings of the BioNLP 2018 workshop (pp. 67-75).
[26] Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72-78).
[27] Peng, Y., Yan, S., & Lu, Z. (2019, August). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 58-65).