Published in final edited form as:

Proc (IEEE Int Conf Healthc Inform). 2023 June ; 2023: 558–564. doi:10.1109/ichi57859.2023.00100.

Named Entity Recognition and Normalization for Alzheimer's Disease Eligibility Criteria

Zenan Sun[1], **Cui Tao**[1]

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas

Abstract

Alzheimer's Disease (AD) is a complex neurodegenerative disorder that affects millions of people worldwide. Finding effective treatments for this disease is crucial. Clinical trials play an essential role in developing and testing new treatments for AD. However, identifying eligible participants can be challenging, time-consuming, and costly. In recent years, the development of natural language processing (NLP) techniques, specifically named entity recognition (NER) and named entity normalization (NEN), have helped to automate the identification and extraction of relevant information from the eligibility criteria (EC) more efficiently, in order to facilitate semi-automatic patient recruitment and enable data FAIRness for clinical trial data. Nevertheless, most current biomedical NER models only provide annotations for a restricted set of entity types that may not be applicable to the clinical trial data. Additionally, accurately performing NEN on entities that are negated using a negative prefix currently lacks established techniques. In this paper, we introduce a pipeline designed for information extraction from AD clinical trial EC, which involves preprocessing of the EC data, clinical NER, and biomedical NEN to Unified Medical Language System (UMLS). Our NER model can identify named entities in seven pre-defined categories, while our NEN model employs a combination of exact match and partial match search strategies, as well as customized rules to accurately normalize entities with negative prefixes. To evaluate the performance of our pipeline, we measured the precision, recall, and F1 score for the NER component, and we manually reviewed the top five mapping results produced by the NEN component. Our evaluation of the pipeline's performance revealed that it can successfully normalize named entities in clinical trial ECs with optimal accuracies. The NER component achieved a overall F1 of 0.816, demonstrating its ability to accurately identify seven types of named entities in clinical text. The NEN component of the pipeline also demonstrated impressive performance, with customized rules and a combination of exact and partial match strategies leading to an accuracy of 0.940 for normalized entities.

Keywords

Alzheimer's Disease; clinical trial eligibility criteria; named entity recognition; named entity normalization

I. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects millions of people worldwide and is one of the leading causes of dementia in the elderly population

[1, 2]. Currently, there is no cure for AD, and treatments are primarily aimed at managing symptoms and slowing disease progression [3]. Clinical trials play a crucial role in developing and testing new treatments for AD and require rigorous screening of potential participants to ensure they meet the eligibility criteria (EC). EC is a vital aspect of clinical trials, consisting of a set of inclusion and exclusion criteria that outline the necessary characteristics a participant must have or lack to be eligible for enrollment in the study. These critera help to ensure that the study's results are reliable and applicable to the target patient population while also minimizing the potential risks to participants. ECs may include demographic information, medical history, clinical symptoms, laboratory results, and other relevant data that determine a participant's eligibility for a particular clinical trial. These EC are usually recorded in free-text format. Due to the absence of standardized formatting and clear definitions, EC is not easily interoperable or comprehensible. [4, 5]. In addition, manually reviewing EC by clinical research staff to identify eligible patients can be a costly and time-consuming process, often resulting in a high rate of screen failure and low recruitment rates [6, 7, 8].

In recent years, natural language processing (NLP) techniques, specifically named entity recognition (NER) and named entity normalization (NEN), have emerged as a solution to automatically identify and extract pertinent information from eligibility criteria [9]. This has facilitated researchers in processing EC more efficiently. NER is a fundamental task in NLP that involves identifying and classifying named entities, such as disease, lab test, and medication, from the text [10]. NER has numerous applications in various domains, including biomedical informatics, social media analysis, and information extraction [10, 11, 12]. Although NER has achieved significant success in extracting entities from biomedical literature, its application to clinical trial documents has received relatively little attention in previous studies [13]. Moreover, many existing NER tools only give annotations for a limited number of entity types [14]. This limitation can make it challenging to accurately identify all relevant information from clinical trial ECs. Another NLP task NEN in biomedical informatics is the process of mapping recognized named entities to a standardized terminology such as UMLS, MeSH, etc. [15]. In natural language, a named entity can have multiple names, which can result in ambiguity and difficulty in identification and extraction process [16, 17]. Conversely, a single name can refer to various named entities, adding to the complexity of NLP tasks [16]. The use of standardized terminologies is critical for data integration and interoperability. It helps to reduce ambiguity and promotes interoperability, making it easier to share and integrate data across different applications and domains [4, 5]. Although there have been advancements in NEN, appropriately normalizing entities that are negated using a negative prefix is still difficult. Additionally, there are times the NEN model matches to synonyms entities when an exact mapping is possible.

In this paper, we present the development of a pipeline that combines the strengths of prevailing clinical NLP tools and deep learning technology to effectively extract information from Alzheimer's Disease clinical trial EC. The pipeline consists of three major components: (1) pre-process eligibility criteria data, (2) perform clinical NER, and (3) conduct NEN and link to the UMLS Concept Unique Identifier (CUI). We adopted CLAMP [18], an NLP tool designed specifically for clinical text processing, for data preprocessing and NER modeling. In particular, our NER model can detect seven types of named entities (Problem, Medication, Lab, Rating Criteria, Social Determinant Health, Procedure, and Fertilize.) We constructed our NEN component using SapBERT [19], a cutting-edge deep learning model that normalizes biomedical entities to the UMLS. By integrating deep learning strategies and rules into our NEN model, we were able to achieve optimal accuracies in normalizing entities with negative prefixes and minimizing instances where the model maps to synonym entities. By assessing the precision, recall, and F1 score for NER and manually reviewing the top 5 mapping results for NEN, we demonstrated our proficiency in identifying various entity types in the AD clinical trial EC and precisely mapping them to the UMLS standard terminologies. Although the current study focused on AD, our pipeline is not limited to AD clinical trial ECs and can be applied to other clinical trial ECs or medical texts in general. This flexibility makes our pipeline valuable tool for clinical research, providing a more efficient and accurate approach to information extraction and annotation.

The paper structure will be presented below: in Section II, some existing related research will be briefly introduced; in Section III, we demonstrate the entire architecture of such information extraction pipeline and evaluation metrics; in Section IV, the detailed results and discussion will be reported; in Section V, an overview of all sections in this paper, as well as limitation and potential future works, will be described.

II. Related Works

A. Clinical Named Entity Recognition

NER has been applied extensively to biomedical data to extract important information from electronic health records (EHRs), clinical notes, and other types of medical text [ref Biomedical NER application]. In recent years, there have been numerous studies that have investigated various NER methods and techniques for extracting clinical entities [10]. For example, Chen et al. developed a multi-level rule-based natural language processing system to address the challenge of identifying suitable patient cohorts based on the ECs for clinical trials [20]. It is a rule-based model and achieved an overall micro-F1 score of 0.90 on the 2018 n2c2-1 challenge dataset. [21] proposed a machine learning approach for NER in clinical texts, which combines segment representation models using SVM and conditional random forest algorithms. It reported an F1-score of 0.77 on three entity types (Problem, Treatment, Test) from the i2b2 dataset. While these NER models have exhibited commendable performance, the quantity of captured entity types is insufficient to extract information from clinical trial EC.

Apart from standalone NER models, several NLP tools have been used for clinical NER. One of the the most widely used tools is cTAKES (clinical Text Analysis and Knowledge Extraction System), which aims to perform various NLP tasks such as sentence boundary detection, tokenization, part-of-speech tagging, shallow parsing, NER, and NEN [22]. Its NER model has been reported to achieve an F1 score of 0.715 on 160 clinical notes from the Mayo Clinic EMR. CLAMP (Clinical Language Annotation, Modeling, and Processing) is an NLP tool designed specially for clinical text processing [18]. CLAMP is capable of performing multiple tasks similar to cTAKES, while also provides a user-friendly graphical interface that enables non-technical users to create customized NLP pipelines for a range of applications. The pre-trained NER component (CRF based model) of CLAMP can accurately identify Problem, Treatment, and Test entities, achieving impressive F1 scores of 0.94, 0.91, and 0.95 on the i2b2 dataset, MTSamples corpus, and UTNotes, respectively.

B. Biomedical Named Entity Normalization

NEN is an important task in NLP that involves mapping identified named entities to standard terminology or ontology concepts. In the past, early NEN methods relied on manually crafted rules and dictionaries and were typically limited to a single entity type [23]. For instance, GNormPlus was specifically designed for gene normalization [24], SR4GN for species normalization [25], and tmVar for mutation normalization [26, 27]. These rule-based techniques are widely used in practical applications due to their flexibility and straightforward interpretation [28]. However, such methods have limitations in addressing intricate morphological variations of biomedical entity references and can be challenging to adapt to new data [14]. As a result, machine learning and deep learning techniques have gained prominence in recent research and outperformed rule-based methods in terms of performance in the biomedical domain.

With the development of deep learning techniques, pretrained models—specifically BioBERT [29], PubMedBERT [30], Med-BERT [31], BlueBERT [32], ClinicalBERT [33]—have shown remarkable success in biomedical entity representation which then can be effectively applied in various biomedical tasks. BioSyn is one of the state-of-the-art SOTA methods for biomedical NEN that can represent biomedical entities using Synonym Marginalization [34]. It combines distributional semantics and probabilistic modeling to create a representation for each entity that considers all of its synonyms. Based on BioSyn, SapBERT was designed to enhance the representation of biomedical entities [19]. SapBERT employed a new metric learning framework, so-called Self-Alignment, and pre-trained on the UMLS to learn biomedical entities from a cluster of synonyms with the same concept. While NEN models have shown good performance, they currently lack the ability to accurately normalize biomedical entities with a negative prefix. Furthermore, pretraining these models on synonyms may result a synonym-normalized entity instead of a precisely matched normalized entity, leading to potential errors in downstream applications.

C. Existing NLP Tools for Information Extraction in EC

EliIE is an open-source information extraction system that is particularly designed for clinical trial EC [35]. It comprises of 4 sequential components: NER, negation detection, relation extraction, and NEN that links entities to OMOP standard terminology. The NER model of EliIE employs a conditional random field (cRF) algorithm to identify four different entity types (Condition, Observation, Procedure, Drug) and their associated attributes (Qualifier, Temporal, and Measurement). It has been reported to achieve an overall F1 score of 0.786 on a randomly selected sample of 230 AD trials from ClinicalTrials.gov. Criteria2Query is a tool that transforms clinical trial EC into a structured and computable format [36]. It performs various tasks such as NER, negation detection, relation extraction, logic detection, and NEN to achieve this. The NER model of Criteria2Query is developed using the conditional random fields method in CoreNLP2. It identifies five entity types (Condition, Drug, Measurement, Procedure, and Observation) and two attribute types

(Value, Temporal) with an F1 score of 0.804. Additionally, Criteria2Query utilizes Usagi to map identified entities to OMOP standard concepts with an accuracy of 0.447. Tseo et al. introduced another information extraction approach for clinical trial EC [37]. The authors utilized an attention-based conditional random field method to identify ten entity classes, including Treatment, Chronic disease, Cancer, Gender, Pregnancy, Allergy, Contraception consent, Language literacy, Technology access, and Ethnicity. They also employed word2vec and DBSCAN embedding clustering methods to connect these entities to the MesH knowledge source. The reported F1 for NER and accuracy for NEN were 0.802 and 0.485, respectively.

Although these NLP tools are capable of recognizing various entity types in EC and have demonstrated good performance, they rely on fixed algorithms for NEN tasks which may restrict our ability to incorporate our customized rules and handle diverse situations, as highlighted in section II.B.

To address the concerns mentioned above, we have developed a pipeline for extracting information from Alzheimer's Disease clinical trial eligibility criteria. To begin with, we leveraged CLAMP for data pre-processing and to train a NER model that identifies seven entity types. We selected CLAMP due to its exceptional performance and user-friendly interface. To further enhance the pipeline's NEN performance, we incorporated SapBERT as the base model and implemented our own rules for normalizing named entities. This allowed us to overcome situations where an exact match was present but failed to be identified, and improve the mapping accuracy on entities with negative prefixes. Our pipeline demonstrated optimal performance and was able to extract more valuable information from EC.

III. Methodology

A. Dataset

We retrieved 1508 AD clinical trials from ClinicalTrials.gov and extracted the eligibility criteria text from each trial file for annotation. A total of 300 randomly selected clinical trials were manually annotated, and seven entity types were identified, including Problem, Medication, Lab, Rating Criteria, Social Determinant Health, Procedure, and Fertilize. TABLE I shows the descriptive analysis of our annotated EC dataset.

B. Named Entity Recognition Component

We developed a customized machine learning model for NER of clinical text data using the CLAMP toolkit. The toolkit was applied for data preprocessing and preparation for NER training, which involved tasks such as removing punctuation, detecting biomedical abbreviations, and converting all text to lowercase. We utilized the Conditional Random Field (CRF) algorithm to implement the NER model. During CRF model training, features such as word shape, part-of-speech tags, and distinguishing sentence patterns were employed. The NER model was trained through 5-fold cross-validation on 250 annotated clinical trial files and assessed on the remaining 50 annotated files as a test set to evaluate its performance. overall, the NER component enable us to accurately identify and extract named entities from clinical text.

C. Named Entity Normalization Component

Our NEN model was built using SapBERT as the foundation. SapBERT was pre-trained on the biomedical knowledge graph of UMLS [38], which provides an extensive repository of biomedical synonyms in different formats. We first employed SapBERT to generate 768-sized vectors for all entity names from the UMLs dictionary, and all labeled queries. The objective was to identify the entity names that were the closest neighbors to the query in the embedding space, in other words, to find the most similar embeddings, as the normalization results. To compute similarity scores, we utilized Facebook AI similarity search (Faiss), which is optimized for rapid similarity search and clustering of high-dimensional vectors, to compute similarity scores [39]. The vectors are compared based on Euclidean distance (l_2 distance) to determine their similarity, where smaller distances indicate greater similarity.

Due to the existence of multiple synonyms in the UMLS, it is plausible that SapBERT may not return an exact matching entity as the best result. To address this issue, we devised a solution that involves conducting an exact match search before executing the Faiss similarity test. If an exact match is not found, we proceed with Faiss as a partial match search strategy. In this case, either an exact match result or the five most similar outcomes will be retrieved.

Negative prefixes such as “in-”, “un-”, “non-”, “de-”, “dis-”, “a-”, “anti-”, “im-”, “il-”, and “ir-” are commonly used with named entities. In our study, when a query begins with any of these prefixes, we create a negated query by removing the prefix. We then compare the best match results of both the original query and the negated query. If the negated query has a higher similarity score (lower l_2 distance) than the original query, we add a negation before the normalized result of the negated query to obtain the final outcome. Otherwise, we just remove this negated query. The architecture of the entire pipeline can be found in Fig. 1.

D. Evaluation Metrics

We utilized precision, recall, and F1 score to evaluate NER model performance, and manually reviewed the NEN model accuracy. The precision score is a key metric in machine learning and statistics used to assess the quality of a model’s positive predictions. It is defined as

$$Precision = \frac{TP}{TP + FP},$$

where TP is true positive, FP is false positive. A high precision score indicates that the model is making fewer false positive predictions and has a high level of accuracy in detecting positive cases. Recall score is a commonly used metric to assess the proportion of true positive predictions concerning all actual positive cases:

$$Recall = \frac{TP}{TP + FN},$$

where FN is false negative. Similarly, a high recall score implies that the model can accurately identify a significant portion of the positive cases. F1 score is a metric that helps to balance both precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}.$$

IV. Result and Discussion

The performance of the NER model is presented in TABLE II. The results indicate that the category Fertilize has the highest precision of 0.908, recall of 0.857, and F1 of 0.882 among all other entity types, demonstrating the NER model’s excellent performance in identifying positive cases of Fertilize. Except for Fertility, Medication has the second highest precision of 0.872, whereas Problem has the second highest recall of 0.823 and F1 of 0.832. Overall, the NER model achieved a micro average F1 of 0.816, indicates the proficiency in recognizing various entity types present in the AD clinical trial EC.

Algorithm 1

Named Entity Normalization Algorithm

```
Require:
    Query set: $Q = \{q_i|_{i=1}^{N}\}$, N is the total number of queries;
    Negative prefix set: $S = \{s_j|_{j=1}^{M}\}$, $M$ is the total number of negative prefixes;
    Remove negative prefix: $R$;
    UMLS dictionary: $D$;
    $l_2$ loss: $\mathscr{L}$;
    Exact match search strategy: ExactS;
    Partial similarity search strategy: PartS;
Output:
    Best normalized results: $Y = \{y_i|_{i=1}^{N}\}$
    Initialisation :
for $i = 1$ to $N$ do
  if ExactS($q_i$) $< >$ $Null$ then
    $P_i = \text{ExactS}(q_i)$
  else
    $P_i = \text{PartS}(q_i)$ // get partial search result
    // check if query starts with a negative prefix
    if $q_i$ starts with $s_j$, $\forall_j \in 1, \ldots M$ then
      if ExactS($R(q_i)$) $< >$ $Null$ then
        $P_i = \text{ExactS}(R(q_i))$
      else
        $T_i = \text{PartS}(R(q_i))$ // partial search for negated query
        if $\mathscr{L}(T_{i[1]}) < \mathscr{L}(P_{i[1]})$ then
          $P_i = T_i$
        end if
```

```
            end if
        end if
    end if
    $y_i = P_i$
end for
return $Y$
```

Table III presents the accuracy of correctly normalized entities in the top 5 candidate entities. The highest accuracy of 0.890 at the 1st mapping was achieved for the Problem entity type, while the Social Determinant Health entity had the lowest performance with an accuracy of 0.580. In terms of total accuracy among all top 5 candidates, Lab and Procedure entities performed the best, while Social Determinant Health still had the lowest accuracy. It should be noted that some correct mapping may appear in multiple candidate position due to the ranking by different synonyms. In such cases, we have selected the frontmost candidate as the most suitable mapping result. In addition, for entity types as Problem, Medication, Rating Criteria, Lab, Procedure, and Fertilize, none of the 4th or 5th candidate mappings were correct (denoted as " - " in Table III). This suggests that either the correct answer was found within the top 3 candidates, or it was not identified at all.

After reviewing the entities and their mapping candidates, we have identified several instances where mapping was unsuccessful:

- **The correct mapping is among the top 5 candidates but not ranked as the 1st.**

  While some of the 1st mapping candidates may share similar spellings with the queried entities, they differ in either meaning or semantic type. For example: the 1st candidate of "allergy" under the Problem type is denoted by the CUI code C0002111 and is described as "Allergy Specialty". Although there are similarities between these two entities, it is important to note that "Allergy Specialty" belongs to the Biomedical Occupation or Discipline semantic type and denotes a specific subject or field of study. Therefore, after considering the remaining candidates, we selected "AllergyAllergy: hypersensitivity" with the CUI code C0020517 as the accurate mapping. We believe that taking semantic type into account during the mapping process can reduce such cases.

- **The correct mapping does not appear within the top 5 candidates.**

  1. Complex phrases or expressions may not always be defined in the UMLS, leading to cases where a search does not yield a matching term. For instance, terms like "age is between 60-80", "BACE inhibitors", "ninds-airen" are not present in the current version of UMLS. This is particularly common in the domain of Social Determinants of Health, which involves ambiguous definitions and differnt levels of granularities, making it difficult to define terms or concepts accurately.

2. The NEN model's capacity to identify synonyms and recognize abbreviations is limited. For example, "tacs" is short for "Transcranial Alternating Current Stimulation", and "ct" may refer to "computed tomography" or "cognitive therapy", etc. The NEN model was unable to identify all the abbreviations used in the clinical trial EC. One way to mitigate this issue is by incorporating additional data into the model training process and enhancing the abbreviation detection dictionary.

V. Conclusion

In conclusion, we have presented a pipeline that integrates data pre-processing, clinical NER, and a NEN for linking named entity to the UMLS, which enable us to extract information from AD clinical trial EC more efficiently. The NER component of our pipeline can identify seven types of named entities, while the NEN component uses a combination of exact match and partial match search strategies, along with customized rules to accurately normalize entities with negative prefixes. We evaluated the performance of the NER component using using precision, recall, and F1 score metrics, and mannually reviewed the top five mapping results from the NEN component. our evaluation of the pipeline's performance revealed that it can succesfully extract and normalize named entities in clinical trial eligibility criteria with optimal accuracies.

Although our pipeline has several benefits, there is still room for improvement, particularly in integrating relation extraction. By adding relation extraction to the pipeline, we can not only improve the NEN model's performance but also gain more comprehensive information. For example, if an entity such as "Age is greater than 35" cannot be accurately mapped, we may break it down into three parts: "Age", "is greater than", and "35". We then easily locate the correct mapping for "Age", while adding a relation "is greater than" and a value "45" to the mapped entity. This type of work will also assist in constructing Ontology or Knowledge Graph, further improving the overall process.

Acknowledgment

This study was supported by the National Institutes of Health (R56AG074604, RF1AG072799). We thank Ms. Yi Zheng and Dr. Manqi Li for helping annotating the data sets used in this project.

References

[1]. Weller Jason and Budson Andrew. "Current understanding of Alzheimer's disease diagnosis and treatment". In: F1000Research 7 (July 2018), p. 1161. DOI: 10.12688/f1000research.14506.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6073093/.

[2]. Khan Sahil, Barve Kalyani H, and Kumar Maushmi S. "Recent advancements in pathogenesis, diagnostics and treatment of alzheimer's disease". In: Current Neuropharmacology 18 (May 2020). DOI: 10.2174/1570159x18666200528142429.

[3]. Yiannopoulou Konstantina G and Papageorgiou Sokratis G. "Current and future treatments in alzheimer disease: An update". In: Journal of Central Nervous System Disease 12 (Jan. 2020), p. 117957352090739. DOI: 10.1177/1179573520907397. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050025/.

[4]. Wu Danny TY et al. “Assessing the readability of ClinicalTrials.gov”. In: Journal of the American Medical Informatics Association 23 (Aug. 2015), pp. 269–275. DOI: 10.1093/jamia/ocv062. [PubMed: 26269536]

[5]. Si Yuqi and Weng Chunhua. “An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria”. In: MEDINFO 2017: Precision Healthcare through Informatics 245 (2017), pp. 950–954. DOI: 10.3233/978-1-61499-830-3-950. URL: https://ebooks.iospress.nl/publication/48293 (visited on 03/15/2023).

[6]. Penberthy Lynne T et al. “Effort required in eligibility screening for clinical trials”. In: Journal of Oncology Practice 8 (Nov. 2012), pp. 365–370. DOI: 10.1200/jop.2012.000646. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3500483/ (visited on 08/2019). [PubMed: 23598846]

[7]. Thadani SR et al. “Electronic screening improves efficiency in clinical trial recruitment”. In: Journal of the American Medical Informatics Association 16 (Aug. 2009), pp. 869–873. DOI: 10.1197/jamia.m3119. (Visited on 01/2022). [PubMed: 19717797]

[8]. Dobbins Nicholas J et al. “The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria”. In: Scientific Data 9 (Aug. 2022). DOI: 10.1038/s41597-022-01521-0. (Visited on 03/2023).

[9]. Idnay Betina et al. “A systematic review on natural language processing systems for eligibility prescreening in clinical research”. In: Journal of the American Medical Informatics Association 29 (Nov. 2021), pp. 197–206. DOI: 10.1093/jamia/ocab228. (Visited on 04/2022). [PubMed: 34725689]

[10]. Song Bosheng et al. “Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison”. In: Briefings in Bioinformatics 22 (July 2021). DOI: 10.1093/bib/bbab282. (Visited on 03/2023).

[11]. Nie Yuyang et al. “Named entity recognition for social media texts with semantic augmentation”. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020). DOI: 10.18653/v1/2020.emnlp-main.107. (Visited on 03/2023).

[12]. Guo Jiafeng et al. “Named entity recognition in query”. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’09 (2009). DOI: 10.1145/1571941.1571989. (Visited on 12/2022).

[13]. Li Jianfu et al. “A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora”. In: BMC Medical Informatics and Decision Making 22 (Sept. 2022). DOI: 10.1186/s12911-022-01967-7.

[14]. Sung Mujeen et al. “BERN2: an advanced neural biomedical named entity recognition and normalization tool”. In: Bioinformatics 38 (Sept. 2022). Ed. by Karsten Borgwardt, pp. 4837–4839. DOI: 10.1093/bioinformatics/btac598. (Visited on 03/2023). [PubMed: 36053172]

[15]. Peng Hao et al. “Biomedical named entity normalization via interaction-based synonym marginalization”. In: Journal of Biomedical Informatics 136 (Dec. 2022), p. 104238. DOI: 10.1016/j.jbi.2022.104238. (Visited on 03/16/2023). [PubMed: 36400329]

[16]. Shen Wei, Wang Jianyong, and Han Jiawei. “Entity linking with a knowledge base: Issues, techniques, and solutions”. In: IEEE Transactions on Knowledge and Data Engineering 27 (Feb. 2015), pp. 443–460. DOI: 10.1109/tkde.2014.2327028. URL: http://dbgroup.cs.tsinghua.edu.cn/wangjy/papers/TKDE14-entitylinking.pdf.

[17]. Leaman Robert, Khare Ritu, and Lu Zhiyong. “Challenges in clinical natural language processing for automated disorder normalization”. In: Journal of Biomedical Informatics 57 (Oct. 2015), pp. 28–37. DOI: 10.1016/j.jbi.2015.07.010. (Visited on 11/14/2020). [PubMed: 26187250]

[18]. Soysal Ergin et al. “CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines”. In: Journal of the American Medical Informatics Association 25 (Nov. 2017), pp. 331–336. DOI: 10.1093/jamia/ocx132.

[19]. Liu Fangyu et al. “Self-Alignment Pretraining for Biomedical Entity Representations”. In: arXiv:2010.11784 [cs] (Apr. 2021). URL: https://arxiv.org/abs/2010.11784.

[20]. Chen Long et al. “Clinical trial cohort selection based on multi-level rule-based natural language processing system”. In: Journal of the American Medical Informatics Association : JAMIA 26 (July 2019), pp. 1218–1226. DOI: 10.1093/jamia/ocz109. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647235/ (visited on 12/2021). [PubMed: 31300825]

[21]. Improving NER for clinical texts by ensemble approach using segment representations. NLP Association of India, Dec. 2017, pp. 197–204. URL: https://aclanthology.org/W17-7525.

[22]. Savova Guergana K et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: Journal of the American Medical Informatics Association 17 (Sept. 2010), pp. 507–513. DOI: 10.1136/jamia.2009.001560. URL: https://academic.oup.com/jamia/article/17/5/507/830823?login=true. [PubMed: 20819853]

[23]. Cuffy Clint et al. "Exploring Representations for Singular and Multi-Concept Relations for Biomedical Named Entity Normalization". In: Companion Proceedings of the Web Conference 2022 (Apr. 2022). DOI: 10.1145/3487553.3524701. (Visited on 03/08/2023).

[24]. Wei Chih-Hsuan, Kao Hung-Yu, and Lu Zhiyong. "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains". In: BioMed Research International 2015 (2015), pp. 1–7. DOI: 10.1155/2015/918710. (Visited on 03/08/2023).

[25]. Wei Chih-Hsuan, Kao Hung-Yu, and Lu Zhiyong. "SR4GN: A Species Recognition Software Tool for Gene Normalization". In: PLoS ONE 7 (June 2012). Ed. by Jan Aerts, e38460. DOI: 10.1371/journal.pone.0038460. (Visited on 11/16/2021). [PubMed: 22679507]

[26]. Wei C-H et al. "tmVar: a text mining approach for extracting sequence variants in biomedical literature". In: Bioinformatics 29 (Apr. 2013), pp. 1433–1439. DOI: 10.1093/bioinformatics/btt156. (Visited on 03/08/2023). [PubMed: 23564842]

[27]. Wei Chih-Hsuan et al. "tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine". In: Bioinformatics 34 (Sept. 2017). Ed. by Jonathan Wren, pp. 80–87. DOI: 10.1093/bioinformatics/btx541. (Visited on 08/26/2020).

[28]. Wen Andrew et al. "Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation". In: npj Digital Medicine 2 (Dec. 2019). DOI: 10.1038/s41746-019-0208-8. (Visited on 02/29/2020).

[29]. Lee Jinhyuk et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: Bioinformatics 36 (Sept. 2019). Ed. by Jonathan Wren. DOI: 10.1093/bioinformatics/btz682. URL: 10.1093/bioinformatics/btz682/5566506.

[30]. Gu Yu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: ACM Transactions on Computing for Healthcare 3 (Jan. 2022), pp. 1–23. DOI: 10.1145/3458754.

[31]. Rasmy Laila et al. "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: npj Digital Medicine 4 (May 2021). DOI: 10.1038/s41746-021-00455-y.

[32]. Peng Yifan, Yan Shankai, and Lu Zhiyong. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. ACLWeb, Aug. 2019. DOI: 10.18653/v1/W19-5006. URL: https://www.aclweb.org/anthology/W19-5006/ (visited on 10/22/2020).

[33]. Huang Kexin, Altosaar Jaan, and Ranganath Rajesh. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: arXiv:1904.05342 [cs] (Apr. 2019). URL: https://arxiv.org/abs/1904.05342.

[34]. Sung Mujeen et al. "Biomedical Entity Representations with Synonym Marginalization". In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). DOI: 10.18653/v1/2020.acl-main.335. (Visited on 03/08/2023).

[35]. Kang Tian et al. "EliIE: An open-source information extraction system for clinical trial eligibility criteria". In: Journal of the American Medical Informatics Association 24 (Apr. 2017), pp. 1062–1071. DOI: 10.1093/jamia/ocx019. URL: https://academic.oup.com/jamia/article/24/6/1062/3098256?login=true. [PubMed: 28379377]

[36]. Yuan Chi et al. "Criteria2Query: a natural language interface to clinical databases for cohort definition". In: Journal of the American Medical Informatics Association 26 (Feb. 2019), pp. 294–305. DOI: 10.1093/jamia/ocy178. (Visited on 03/15/2023). [PubMed: 30753493]

[37]. Tseo Yitong et al. "Information Extraction of Clinical Trial Eligibility Criteria". In: arXiv:2006.07296 [cs] (July 2020). URL: https://arxiv.org/abs/2006.07296 (visited on 03/15/2023).

[38]. Bodenreider O. “The Unified Medical Language System (UMLS): integrating biomedical terminology”. In: Nucleic Acids Research 32 (Jan. 2004), pp. 267D–270. DOI: 10.1093/nar/gkh061.

[39]. Johnson Jeff, Douze Matthijs, and Jegou Herve. “Billion-scale similarity search with GPUs”. In: IEEE Transactions on Big Data 7 (Mar. 2021), pp. 1–1. DOI: 10.1109/tbdata.2019.2921572. (Visited on 08/21/2019).

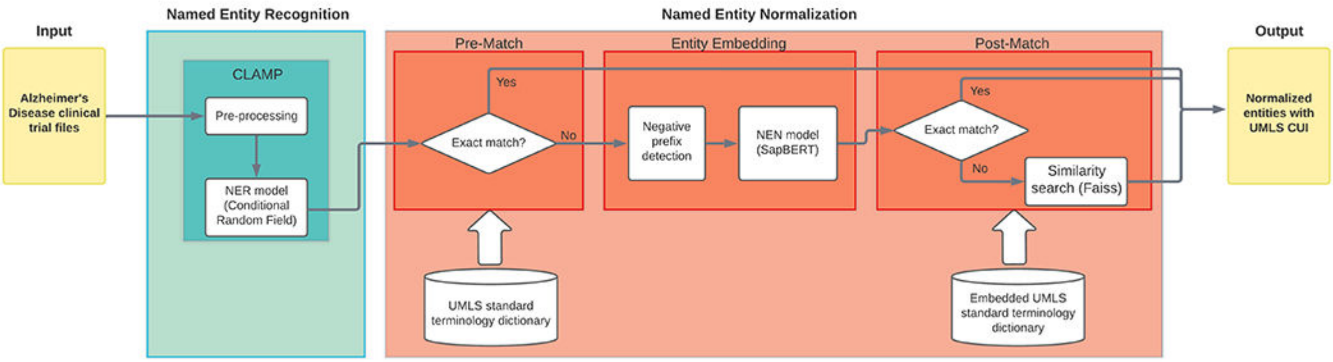
Input
Alzheimer's Disease clinical trial files
Named Entity Recognition
CLAMP
Pre-processing
NER model (Conditional Random Field)
Named Entity Normalization
Pre-Match
Yes
Exact match?
No
UMLS standard terminology dictionary
Entity Embedding
Negative prefix detection
NEN model (SapBERT)
Post-Match
Yes
Exact match?
No
Similarity search (Faiss)
Embedded UMLS standard terminology dictionary
Output
Normalized entities with UMLS CUI

Fig. 1.

The basic architecture of our information extraction pipeline for Alzheimer's Disease clinical trial eligibility criteria. We utlized the CLAMP toolkit for data pre-processing and named entity recognition. In addition, we integrated SapBERT as the base model and implemented a combination of exact match and partial match search techniques, along with negative-prefix detection, to effectively normalize entities.

TABLE I

DESCRIPTIVE STATISTICS OF ANNOTATED ELIGIBILITY CRITERIA

Main entity type	**unique count**
Problem	1,178
Medication	658
Rating Criteria	317
Lab	282
Social Determinant Health	264
Procedure	109
Fertilize	31
Total	**2,839**

TABLE II

NER MODEL PERFORMANCE

Main entity type	**Precision**	**Recall**	**F1**
Problem	0.840	0.823	0.832
Medication	0.872	0.794	0.831
Rating Criteria	0.798	0.736	0.766
Lab	0.808	0.720	0.762
Social Determinant Health	0.808	0.768	0.787
Procedure	0.868	0.740	0.799
Fertilize	0.908	0.857	0.882
Micro Average			**0.816**

TABLE III

NEN Top 5 mapping accuracy

Main entity type	**1st**	**2nd**	**3rd**	**4th**	**5th**	**Total**
Problem	0.890	0.051	0.012	-	-	**0.953**
Medication	0.864	0.045	0.015	-	-	**0.924**
Rating Criteria	0.780	0.060	0.060	-	-	**0.900**
Lab	0.760	0.140	0.080	-	-	**0.980**
Social Determinant Health	0.580	0.160	0.080	0.020	0.020	**0.860**
Procedure	0.800	0.140	0.040	-	-	**0.980**
Fertilize	0.806	0.161	-	-	-	**0.967**
Overall Average	**0.802**	**0.094**	**0.039**	**0.002**	**0.002**	**0.940**