

Neural language models for clinical trial eligibility criteria

Klaudia Kantor

Under the supervision of Professor Mikolaj Morzy

DOCTORAL THESIS
Poznan University of Technology

August 2024

Abstract

Clinical trials are crucial for advancing medical knowledge and introducing new therapies to the market. However, the recruitment of patients is a significant challenge, often leading to delays and trial terminations. A major contributor to this issue is the manual process of patient screening, which is highly inefficient due to the stringent and complex nature of eligibility criteria. These criteria are written in free text, making them difficult to process with existing algorithms. Currently, there is no tool capable of automating this process effectively. This thesis explores the potential of neural language models, particularly recent advances in natural language processing (NLP), to transform clinical trial eligibility criteria into a machine-readable format, and consequently automate and optimize the recruitment process.

The study begins by investigating the impact of trial design on trial execution performance, followed by a comprehensive exploration of available biomedical NLP resources. A critical scoping review reveals that the majority of approaches to parsing eligibility criteria rely on outdated and non-generalizable methods such as regular expressions, rule-based systems, and early word embedding techniques like word2vec and FastText. Surprisingly, only a small portion of recent studies have employed pretrained transformer models, with a few leveraging generative AI, despite its rapid advancements.

The experimental phase of this research involves evaluating text embeddings in the biomedical domain for semantic textual similarity (STS) tasks and comparing the performance of BERT models with GPT-4 in named entity recognition (NER) from eligibility criteria. This work is conducted within the context of an industrial PhD program, where scientific findings are applied in practice. The implementation phase includes organizing a hackathon at Roche to assess the potential of large language models (LLMs) in parsing eligibility criteria and developing a Proof of Concept (PoC) tool using GPT-4o, along with advanced prompt engineering. The tool was developed in collaboration with domain experts and evaluated in a carefully designed experiment involving their participation. The results emphasize its practical utility in the pharmaceutical industry.

The findings indicate that the problem of criteria processing is complex and extends beyond basic STS or NER tasks. Effective parsing requires the extraction of more nuanced information, including relationships between entities and advanced reasoning—a task where generative LLMs demonstrate significant potential, particularly in the context of limited annotated data. Additionally, the support of domain experts is crucial for the successful implementation of such tools. The research also identifies a gap in benchmark datasets for evaluating models that process eligibility criteria and a lack of reliable methods for evaluating LLMs in information retrieval tasks, highlighting an area for future exploration.

The contributions of this thesis include the most updated and extensive scoping review in the field, practical requirements for implementing NLP tools in the pharmaceutical industry, a comparison of BERT models and LLMs in biomedical applications, and the design and development of a PoC tool guided by domain expert input and advanced prompt engineering techniques. The conclusions underscore the need for further refinement of the PoC for broader applicability across different therapeutic areas and criteria types, more rigorous evaluation under diverse conditions, and enhanced collaboration with domain experts. Despite these challenges, the research demonstrates that integrating NLP with domain expertise can significantly improve patient recruitment processes in clinical trials.

Keywords: generative AI, large language models, transformers, named entity recognition, information retrieval, eligibility criteria, clinical trials

Streszczenie

Badania kliniczne są kluczowe dla rozwoju medycyny i wprowadzania nowych metod leczenia na rynek. Rekrutacja pacjentów do badań klinicznych stanowi istotne wyzwanie, które często powoduje opóźnienia lub nawet przerwanie badań. Jedną z głównych przyczyn tego problemu jest ręczny proces selekcji pacjentów, który jest wysoce nieefektywny ze względu na bardzo rygorystyczne i złożone kryteria kwalifikacji. Kryteria te są pisane hermetycznym językiem medycznym, co utrudnia ich bezpośrednie przetwarzanie przez istniejące algorytmy. Obecnie nie istnieje żadne narzędzie, które skutecznie zautomatyzowałoby proces porównywania danych pacjentów z tymi kryteriami. Niniejsza praca doktorska bada potencjał modeli językowych opartych na sieciach neuronowych w przekształcaniu kryteriów kwalifikacji do badań klinicznych w ustrukturyzowaną formę czytelną dla algorytmów, w konsekwencji badając, czy te modele mogą pomóc w automatyzacji i optymalizacji procesu rekrutacji.

Rozprawa rozpoczyna się od oceny wpływu elementów projektu badania klinicznego na jego przebieg, skupiając się na istotności cech modelu uczenia maszynowego przewidującego metryki efektywności operacyjnej. Wyniki wskazują, że cechy projektowe mogą stanowić wiarygodne predyktory efektywności operacyjnej badań. Jednocześnie badanie identyfikuje brak zaawansowanych metod analizy treści kryteriów kwalifikacyjnych, co jest kluczową motywacją do dalszego rozwijania narzędzi automatycznego przetwarzania i analizy tych kryteriów w ramach kolejnych etapów pracy doktorskiej.

Kolejny rozdział zawiera opis dostępnych zasobów do przetwarzania tekstu w dziedzinie biomedycyny, takich jak bazy danych, ontologie oraz biomedyczne modele językowe. Omówiony jest również dynamiczny rozwój generatywnej sztucznej inteligencji, z podkreśleniem jej potencjału w przetwarzaniu tekstu biomedycznego.

Następnie rozprawa przedstawia szczegółowy przegląd literatury dotyczącej przetwarzania tekstu kryteriów kwalifikacji do badań klinicznych, ze szczególnym uwzględnieniem stosowanych metod, osiągniętych wyników i wyciągniętych wniosków. Analiza ujawnia, że większość badań wokół przetwarzania kryteriów kwalifikacji używa przestarzałych metod, takich jak wyrażenia regularne, algorytmy oparte na regułach oraz stare techniki wektoryzacji tekstów, takie jak word2vec i FastText. Co ciekawe, tylko niewielka część publikacji wykorzystuje modele oparte o architekturę Transformer, a zaledwie kilka bazuje na generatywnych modelach językowych, które bardzo szybko rozwijają się w ostatnich latach.

Faza eksperymentalna rozprawy obejmuje ocenę metod wektoryzacji tekstów biomedycznych w zadaniu podobieństwa semantycznego (STS) oraz porównanie modeli opartych o BERT i modelu GPT-4 w zadaniu rozpoznawania nazwanych

jednostek (ang. *NER, named entity recognition*) w tekstach kryteriów kwalifikacji. W eksperymencie STS testowane są różne reprezentacje tekstowe, takie jak statyczne wektory GloVe, pretrenowane biomedyczne modele BERT i transformery zdaniowe (ang. *Sentence Transformers*) w celu zidentyfikowania najbardziej efektywnej metody wektoryzacji zdań dla analizy tekstów biomedycznych. Wyniki wskazują, że transformery zdaniowe są najbardziej uniwersalne i efektywne w zadaniach związanych z semantycznym podobieństwem tekstów biomedycznych.

Eksperyment NER ma na celu ocenę skuteczności inżynierii zachęt (ang. *prompt engineering*) w ekstrakcji kluczowych informacji z kryteriów kwalifikacyjnych do badań klinicznych. Badanie przeprowadzone jest w warunkach bardzo ograniczonej liczby danych uczących i minimalnego wsparcia ekspertów domenowych. W przypadku modelu GPT zastosowana jest prosta zachęta bez dodatkowych wskazówek dla modelu GPT i podejście few-shot, a modele oparte o BERT są dotrenowane na niewielkim zbiorze danych. Wyniki pokazują, że podejście few-shot może być skuteczne w zadaniach NER, szczególnie gdy brak jest wystarczających danych uczących. Niemniej jednak, modele BERT, takie jak CODER, osiągają wyraźnie lepsze rezultaty w tym zadaniu, zwłaszcza gdy możliwe jest ich dostrojenie.

Ze względu na to, że praca ta jest doktoratem wdrożeniowym, jej wyniki naukowe są wdrażane w przemyśle. Faza implementacyjna obejmuje organizację hackathonu w firmie Roche w celu oceny potencjału dużych modeli językowych (LLM) w przetwarzaniu kryteriów kwalifikacji oraz opracowanie prototypu narzędzia (PoC) parsującego te kryteria. Narzędzie to wykorzystuje model GPT-4o oraz zaawansowane techniki inżynierii zachęt, uwzględniające wyspecjalizowaną wiedzę domenową. Narzędzie zostało zbudowane we współpracy z ekspertami dziedzinowymi, a jego efektywność oceniono w starannie zaprojektowanym eksperymencie ewaluacyjnym z udziałem tychże ekspertów. Wyniki ewaluacji podkreślają jego praktyczną użyteczność w branży farmaceutycznej.

Rozprawa dowiodła, że problem przetwarzania kryteriów kwalifikacji jest bardzo złożony i wykracza poza podstawowe zadania NLP, takie jak STS, czy NER. Przetworzenie kryteriów do formatu użytecznego w rekrutacji pacjentów wymaga wydobywania bardziej złożonych informacji, na przykład relacji między kryteriami, i zaawansowanego rozumowania. Jest to zadanie, w którym generatywne duże modele językowe osiągają wysokie wyniki, szczególnie w przypadku braku dużych zbiorów uczących. Ponadto wsparcie ekspertów domenowych jest kluczowe dla skutecznego rozwoju i wdrożenia takich narzędzi. W rozprawie rozpoznano deficyt dostępnych referencyjnych zbiorów danych, które mogłyby być użyte do oceny modeli przetwarzających kryteria kwalifikacji, a także brak rzetelnych metod mierzenia dokładności modeli generatywnych w zadaniach ekstrakcji informacji z tekstu. Zidentyfikowano tutaj potencjalny obszar dalszych badań naukowych.

Kontrybucja tej pracy obejmuje najbardziej aktualny i rozległy przegląd literaturowy w tej dziedzinie, zebranie praktycznych wymagań dotyczących wdrożenia narzędzia NLP w przemyśle farmaceutycznym, porównanie modeli BERT i LLM w dziedzinie biomedycznej oraz zaprojektowanie i zbudowanie narzędzia PoC, które integruje wiedzę domenową z najnowszą technologią. Dalsze badania w tym ob-

szarze powinny skupić się na poszerzeniu zakresu zastosowania opracowanego narzędzia o nowe obszary terapeutyczne i typy kryteriów, opracowaniu bardziej rygorystycznego sposobu ewaluacji tego narzędzia oraz włączenie większej liczby ekspertów domenowych. Pomimo ograniczeń, wyniki rozprawy pokazują, że integracja NLP z wiedzą ekspercką może znacząco poprawić efektywność rekrutacji pacjentów do badań klinicznych.

Słowa kluczowe: generatywna sztuczna inteligencja, duże modele językowe, transformer, BERT, GPT, NER, ekstrakcja informacji, kryteria kwalifikacji, badania kliniczne