Abstract

This dissertation investigates methods for transforming unstructured biomedical literature into structured, computationally actionable knowledge that can support the construction of Petri net-based models. The work is motivated by the rapidly growing volume of biomedical publications and the increasing difficulty of integrating fragmented scientific findings into coherent representations of complex biological systems. Modern systems biology requires models that capture interactions among multiple molecular, cellular, and physiological processes, yet constructing such models manually remains highly time-consuming and dependent on extensive expert knowledge. Petri nets provide an attractive modeling formalism in this context because they combine an intuitive graphical representation of biological processes with rigorous mathematical semantics and can be applied even when detailed quantitative parameters are unavailable.

The central objective of this work was therefore to develop computational methods that reduce the effort required to extract biologically meaningful entities and relations from scientific literature while preserving the precision, interpretability, and traceability required for scientific modeling. Rather than aiming for complete automation, the research adopts a semi-automated, human-in-the-loop approach in which artificial intelligence supports domain experts by identifying and organizing candidate knowledge that can subsequently be validated and incorporated into formal models.

The methodological contributions cover several stages of biomedical information extraction. First, the research investigates Named Entity Recognition, including improvements to an existing model and the development of a new model family for multi-type, span-aware entity extraction. These studies demonstrate that biomedical entity recognition can be made computationally efficient and scalable while retaining high predictive performance. At the same time, the results show that accurate identification of entities alone is insufficient for reconstructing coherent biological processes, highlighting the need for subsequent stages of information extraction and interpretation.

The work then addresses Entity Linking through the development of a new model and a modular framework combining contextual semantic representations, candidate retrieval, disambiguation, and reranking. While these approaches achieve competitive linking performance with relatively modest training resources, their application to biological modeling reveals an important conceptual limitation: many biologically relevant entities are inherently context-dependent and cannot always be mapped to fixed database identifiers without losing information relevant to the modeled process. Consequently, strict normalization may in some cases reduce, rather than improve, the structural correctness of computational models.

A further contribution concerns the transition from traditional multi-stage information extraction pipelines based on Named Entity Recognition, Entity Linking, and Relation Extraction toward unified, relation-centered extraction enabled by Large Language Models. Reasoning-capable models make it possible to identify entities and relations jointly, thereby improving coherence and reducing error propagation between separate pipeline components. These findings motivated the development of a platform supporting the semi-automatic extraction of biologically meaningful interactions and their incorporation into Petri net-based models. Evaluation showed that a substantial portion of relevant interactions can be recovered directly from biomedical text, while also revealing limitations associated with implicit knowledge, information contained in figures, document segmentation, and long-range dependencies.

To improve the reliability of the extraction process, the dissertation further introduces an evidence-grounded annotation approach that preserves document context and improves traceability. This methodology is extended by incorporating verification across multiple language models.

Together, these developments shift the extraction process toward a precision-oriented regime in which fewer relations are produced, but individual candidates are better supported by textual evidence and are therefore more suitable for expert validation and scientific modeling.

The research was developed and evaluated in the biomedical domain of high-density lipoprotein metabolism and function, including processes related to lipid transport, inflammation, and cardiovascular biology. A Petri net model describing HDL-related processes affected by local inflammation and oxidative stress was also constructed.

Overall, the dissertation demonstrates both the potential and the current limitations of artificial intelligence for knowledge extraction and biological model reconstruction. Its main conclusion is that fully automated reconstruction of complete biological pathways from scientific literature alone is neither presently feasible nor necessarily desirable. A more effective strategy is to use artificial intelligence to generate reliable, evidence-grounded candidate relations that accelerate model development while preserving expert control and scientific rigor. Future work should extend this approach by incorporating additional information sources and improving the representation of implicit knowledge and long-range contextual dependencies.

Streszczenie

Niniejsza rozprawa doktorska dotyczy metod przekształcania nieustrukturyzowanej literatury biomedycznej w ustrukturyzowaną wiedzę, wspierającą budowę modeli opartych na sieciach Petriego. Motywacją do podjęcia tych badań jest szybko rosnąca liczba publikacji biomedycznych oraz wynikające z tego coraz większe trudności w integrowaniu rozproszonych wyników badań w spójne reprezentacje złożonych systemów biologicznych. Współczesna biologia systemowa wymaga modeli zdolnych do odwzorowania interakcji pomiędzy wieloma procesami molekularnymi, komórkowymi i fizjologicznymi, jednak ich ręczne konstruowanie pozostaje procesem bardzo czasochłonnym i wymagającym rozległej wiedzy eksperckiej. Sieci Petriego stanowią w tym kontekście atrakcyjny formalizm modelowania, ponieważ łączą intuicyjną reprezentację graficzną procesów biologicznych z rygorystyczną semantyką matematyczną, a jednocześnie mogą być stosowane również wtedy, gdy szczegółowe parametry ilościowe nie są dostępne.

Głównym celem pracy było opracowanie metod obliczeniowych ograniczających nakład pracy konieczny do wydobywania z literatury naukowej biologicznie istotnych obiektów i relacji, przy jednoczesnym zachowaniu precyzji, interpretowalności i możliwości śledzenia pochodzenia informacji, wymaganych w modelowaniu naukowym. Zamiast dążyć do pełnej automatyzacji, przyjęto podejście półautomatyczne, oparte na współpracy człowieka z systemem, w którym sztuczna inteligencja wspiera ekspertów dziedzinowych poprzez identyfikowanie i porządkowanie potencjalnie istotnej wiedzy, która następnie może zostać zweryfikowana i włączona do modeli formalnych.

Wkład metodyczny obejmuje kilka etapów ekstrakcji informacji biomedycznej. W pierwszej kolejności analizowano rozpoznawanie nazwanych obiektów, w tym ulepszenia istniejącego modelu oraz opracowanie nowej rodziny modeli umożliwiających wielotypową ekstrakcję obiektów z uwzględnieniem ich zakresów tekstowych. Badania te pokazują, że rozpoznawanie obiektów biomedycznych może być realizowane w sposób wydajny obliczeniowo i skalowalny przy zachowaniu wysokiej jakości predykcji. Jednocześnie wyniki wskazują, że sama poprawna identyfikacja obiektów nie jest wystarczająca do odtwarzania spójnych procesów biologicznych, co podkreśla potrzebę stosowania kolejnych etapów ekstrakcji i interpretacji informacji.

W dalszej części pracy podjęto zagadnienie normalizacji obiektów poprzez opracowanie nowego modelu oraz modularnego środowiska łączącego kontekstowe reprezentacje semantyczne, wyszukiwanie kandydatów, ujednoznacznianie oraz ranking. Chociaż zaproponowane rozwiązania osiągają konkurencyjną skuteczność przy relatywnie niewielkich zasobach treningowych, ich zastosowanie do modelowania biologicznego ujawnia istotne ograniczenie koncepcyjne: wiele biologicznie istotnych obiektów ma charakter silnie zależny od kontekstu i nie zawsze może być odwzorowane na stałe identyfikatory baz danych bez utraty informacji ważnych z punktu widzenia modelowanego procesu. W konsekwencji ścisła normalizacja może w niektórych przypadkach pogarszać, zamiast poprawiać, poprawność strukturalną modeli obliczeniowych.

Kolejny istotny element pracy dotyczy przejścia od tradycyjnych, wieloetapowych potoków ekstrakcji informacji, opartych na rozpoznawaniu obiektów, normalizacji obiektów i ekstrakcji relacji, do zintegrowanej ekstrakcji skoncentrowanej na relacjach, umożliwionej przez duże modele językowe. Modele zdolne do rozumowania umożliwiają jednoczesną identyfikację obiektów i relacji, co sprzyja zachowaniu większej spójności oraz ogranicza propagację błędów pomiędzy kolejnymi komponentami potoku. Wyniki te stały się podstawą do opracowania platformy wspierającej półautomatyczną ekstrakcję biologicznie istotnych interakcji oraz ich włączanie do modeli opartych na sieciach Petriego. Ewaluacja wykazała, że znaczną część istotnych interakcji można wydobyć bezpośrednio z tekstów biomedycznych, a jednocześnie ujawniła ograniczenia związane z wiedzą niejawną, informacjami zawartymi na rycinach, segmentacją dokumentów oraz zależności-

ami nielokalnymi.

W celu zwiększenia wiarygodności procesu ekstrakcji w rozprawie zaproponowano również podejście do adnotacji oparte na dowodach, które zachowuje kontekst dokumentu i poprawia możliwość śledzenia źródeł informacji. Metodologia ta została następnie rozszerzona o mechanizmy weryfikacji wykorzystujące wiele modeli językowych. Łącznie rozwiązania te przesuwają proces ekstrakcji w kierunku podejścia zorientowanego na precyzję, w którym generowanych jest mniej relacji, lecz poszczególni kandydaci są lepiej uzasadnieni materiałem źródłowym, a tym samym bardziej odpowiedni do walidacji i wykorzystania w modelowaniu.

Badania opracowano i oceniono w domenie biomedycznej związanej z metabolizmem i funkcją lipoprotein o wysokiej gęstości, obejmującej między innymi procesy transportu lipidów, zapalenia oraz biologii układu sercowo-naczyniowego. Opracowano również model sieci Petriego opisujący procesy związane z HDL pod wpływem lokalnego zapalenia oraz stresu oksydacyjnego.

Podsumowując, rozprawa pokazuje zarówno potencjał, jak i aktualne ograniczenia sztucznej inteligencji w zakresie ekstrakcji wiedzy i rekonstrukcji modeli biologicznych. Głównym wnioskiem jest stwierdzenie, że w pełni automatyczna rekonstrukcja kompletnych szlaków biologicznych wyłącznie na podstawie literatury naukowej nie jest obecnie ani w pełni wykonalna, ani koniecznie pożądana. Bardziej efektywną strategią jest wykorzystanie sztucznej inteligencji do generowania wiarygodnych, opartych na dowodach kandydackich relacji, które przyspieszają rozwój modeli, jednocześnie zachowując kontrolę eksperta i rygor naukowy. Dalsze prace powinny koncentrować się na włączaniu dodatkowych źródeł informacji oraz doskonaleniu sposobów reprezentowania wiedzy niejawnej i nielokalnych zależności kontekstowych.