Polska Akademia Nauk
Instytut Badań Systemowych

mgr inż. Sławomir Dadas

Iteracyjne metody wykrywania hierarchicznych struktur jednostek nazewniczych

Rozprawa doktorska

Promotor:
prof. dr hab. inż. Witold Pedrycz

Promotor pomocniczy:
dr inż. Jarosław Protasiewicz

Warszawa, 2021

Spis treści

Przedmowa

Praktyczne zastosowania uczenia maszynowego często wiążą się z przetwarzaniem danych sekwencyjnych. W problemach sekwencyjnych mamy do czynienia z listą obserwacji o ustalonej kolejności, na podstawie której model ma za zadanie dokonać predykcji. Powszechnie spotykanymi zadaniami z zakresu przetwarzania sekwencji są między innymi przewidywanie kolejnego elementu na podstawie elementów go poprzedzających, wykrywanie charakterystycznych wzorców w danych (np. w zadaniach związanych z detekcją anomalii), klasyfikacja całej sekwencji lub jej części, czy generowanie zupełnie nowej sekwencji z oryginalnych danych. Przykładami danych, które można przedstawić w postaci uporządkowanego ciągu obserwacji, są szeregi czasowe jako sekwencje danych numerycznych, filmy jako sekwencje obrazów czy dokumenty tekstowe jako sekwencje pojedynczych znaków, grup znaków lub słów.

Niniejsza praca skupia się na wykrywaniu jednostek nazewniczych (ang. *named entity recognition, NER*), czyli zadaniu należącym do dziedziny przetwarzania języka naturalnego, którego głównym założeniem jest identyfikacja i klasyfikacja fraz w tekście odwołujących się różnych kategorii obiektów lub pojęć. Definicja jednostki nazewniczej, jak również lista dopuszczalnych kategorii, jest uzależniona od danych, którymi dysponujemy oraz celu, których chcemy osiągnąć. Przykładowo, dla artykułów prasowych typowymi klasami mogą być imiona i nazwiska osób, nazwy organizacji, nazwy określające obiekty geograficzne i administracyjne, wyrażenia związane z datą i czasem. Natomiast w przypadku artykułów biomedycznych lista klas może uwzględniać na przykład nazwy substancji i związków chemicznych, nazwy wirusów, typy komórek i tkanek, symbole i nazwy genów. Na podstawie powyższych przykładów można zauważyć, że klasy jednostek nazewniczych często związane są z nazwami własnymi, bowiem podobnie jak one służą identyfikacji konkretnych obiektów. Nie jest to jednak reguła - w niektórych instancjach problemu wykrywania jednostek nazewniczych mogą występować klasy uwzględniające również nazwy pospolite, odnoszące się do szerszego zakresu obiektów.

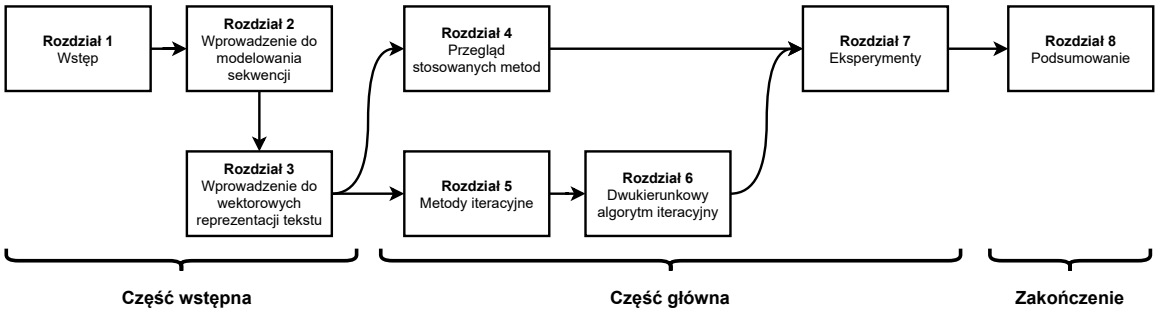
Najczęściej zadanie wykrywania jednostek nazewniczych jest sprowadzane do postaci problemu tagowania sekwencji (ang. *sequence tagging*). Jest to rodzaj uczenia pod nadzorem, w którym model przypisuje klasę każdemu elementowi sekwencji niezależnie - w przeciwieństwie do standardowej klasyfikacji, w której klasyfikowany jest tekst jako całość. Tak zdefiniowany problem ma również zastosowanie przy identyfikacji fraz. W uproszczeniu, kilka następujących po sobie elementów przynależących do tej samej klasy może zostać wyodrębnionych jako samodzielna jednostka nazewnicza. Przyjęcie powyższych założeń pozwala na budowanie modeli, które w sposób sekwencyjny identyfikują frazy w tekście. Do niedawna zainteresowanie naukowców

zajmujących się przetwarzaniem języka naturalnego skupiało się na prostych przykładach wykrywania jednostek nazewniczych. W szczególności przyjmowano założenie, że jednostki nie mogą się na siebie nakładać. Tymczasem niektóre praktyczne zastosowania wymagają wykrywania również jednostek zagnieżdżonych, będących częścią jednostek nadrzędnych. W takim przypadku efektem działania modelu powinna być wielowarstwowa, hierarchiczna struktura jednostek nazewniczych. Liczba warstw w strukturze jest uzależniona od liczby klas i szczegółowości postawionego problemu. W niektórych zastosowaniach praktycznych wymagana jest identyfikacja wielu kategorii jednostek o drobnej granularności, co w znaczący sposób zwiększa prawdopodobieństwo nakładania się jednostek.

Temat hierarchicznej identyfikacji jednostek nazewniczych zyskał na popularności w ostatnich dziesięciu latach. Zaproponowane w literaturze ujęcia początkowo opierały się na wykorzystaniu modeli statystycznych takich jak conditional random fields (CRF). Próbowano też sprowadzać ten problem do zadania generowania hipergrafu reprezentującego zagnieżdżoną strukturę encji. Wraz z rozwojem metod uczenia głębokiego (ang. *deep learning*) zaczęły pojawiać się dedykowane architektury neuronowe, które obecnie stały się dominującą grupą metod stosowanych dla tego typu problemów. Pomimo dynamicznego rozwoju, wiele z zaproponowanych rozwiązań charakteryzuje się wysoką złożonością obliczeniową, uzależnioną od takich parametrów zbioru danych jak liczba zdefiniowanych klas, głębokość zagnieżdżenia czy maksymalna długość jednostki nazewniczej. Ponadto niektóre z modeli narzucają sztywne ograniczenia na powyższe parametry i wymagają ich ustalenia przed rozpoczęciem procesu uczenia modelu. Współczesne rozwiązania bazujące na sieciach neuronowych często są złożonymi systemami, wymagającymi wytrenowania kilku wyspecjalizowanych modeli.

Przedmiotem niniejszej pracy jest zaproponowanie nowego iteracyjnego ujęcia problemu wykrywania hierarchicznych struktur jednostek nazewniczych. W ujęciu tym zakładamy, że problem predykcji struktury hierarchicznej można rozwiązać w sposób iteracyjny, zaczynając od pustego zbioru predykcji i rozszerzając go z każdą kolejną iteracją do momentu, gdy nie zostaną wykryte żadne nowe jednostki nazewnicze. Zaletą zaproponowanej metody jest możliwość łatwego dostosowania istniejących modeli stosowanych w problemach niezagnieżdżonych do działania w sposób iteracyjny, a tym samym rozwiązywania za ich pomocą bardziej złożonych problemów hierarchicznej predykcji. Ponadto przedstawiona metoda nie narzuca żadnych ograniczeń na głębokość zagnieżdżeń czy maksymalną długość jednostki nazewniczej - model iteracyjny automatycznie dostosowuje się do charakterystyki zbioru treningowego w procesie uczenia.

Przedstawione w pracy ujęcie iteracyjne stanowi oryginalny wkład autora w problem wykrywania zagnieżdżonych struktur jednostek nazewniczych. W ramach badań opisanych w tej pracy opracowane zostały również nowe metody wykrywania jednostek będące praktyczną implementacją tego ujęcia. Podstawowym z zaproponowanych rozwiązań jest neuronowy model iteracyjny, którego główna idea polega na przekazywaniu zakodowanego wektora stanu pomiędzy poszczególnymi iteracjami, co pozwala na wykorzystanie informacji o relacjach hierarchicznych pomiędzy jednostkami przy dokonywaniu predykcji. W dalszej części pracy przedstawiono bardziej zaawansowane metody iteracyjne oparte na wykorzystaniu dwóch modeli trenowanych w przeciwnych kierunkach. W ramach predykcji dwukierunkowej rozpatrywane są różne sposoby łączenia wyników modeli (tzw. funkcje selekcji) oraz propagacji wektorów stanu pomiędzy



Rysunek 1: Struktura pracy.

modelami.

Niniejsza praca podzielona została na osiem rozdziałów. Struktura pracy została przedstawiona w postaci schematu graficznego na Rysunku 1.

Pierwszy rozdział stanowi wprowadzenie do zagadnienia wykrywania jednostek nazewniczych. Pokazane zostaną przykładowe praktyczne zastosowania modeli identyfikacji jednostek. Przedstawiona zostanie również formalna definicja problemu wykrywania jednostek nazewniczych. Zdefiniowane zostaną cele oraz hipotezy badawcze będące przedmiotem niniejszej rozprawy.

Kolejne dwa rozdziały zawierają niezbędne podstawy teoretyczne wprowadzające czytelnika w problem wykrywania jednostek nazewniczych. Część wprowadzająca została podzielona na dwie grupy zagadnień. W rozdziale drugim opisane zostaną metody wykorzystywane przy modelowaniu danych sekwencyjnych, w tym modele graficzne (ang. *graphical models*) oraz architektury sieci neuronowych dedykowanych do przetwarzania sekwencji. Rozdział trzeci natomiast dotyczy tematyki związanej z reprezentowaniem tekstu w modelach uczenia maszynowego. Omówione zostaną między innymi wektorowe reprezentacje tekstu, autoregresyjne i maskowane modele języka oraz metody podziału tekstu na sekwencje.

Rozdział czwarty przestawia aktualny stan wiedzy w kontekście problemu hierarchicznego wykrywania jednostek nazewniczych. Uwzględnia przegląd i krytyczną analizę stosowanych do tej pory metod. Wprowadzona zostanie taksonomia ujęć, wyróżniająca trzy główne grupy: metody statystyczne, metody oparte na hipergrafach oraz metody oparte na sieciach neuronowych.

W rozdziale piątym przedstawiono nowe iteracyjne ujęcie problemu hierarchicznego wykrywania jednostek nazewniczych będące zasadniczym przedmiotem tej rozprawy. Zaproponowana zostanie konkretna neuronowa architektura iteracyjna jako praktyczna implementacja tego ujęcia.

W rozdziale szóstym wprowadzono dwukierunkowy algorytm iteracyjny i omówiono poszczególne jego elementy, w tym funkcje selekcji oraz sposób integracji modeli iteracyjnych w ramach algorytmu.

Rozdział siódmy zawiera omówienie i krytyczną analizę wyników doświadczeń przeprowadzonych przy użyciu zaproponowanej metody. Wyniki zostaną porównane z wynikami innych metod omówionych w rozdziale czwartym. W rozdziale zidentyfikowane zostaną również kluczowe cechy algorytmu wpływające na jakość predykcji (*ablation study*). Porównane zostaną warianty zaproponowanego rozwiązania różniące się funk-

cją selekcji oraz wybranymi hiperparametrami.

Rozdział ósmy stanowi podsumowanie zasadniczej części rozprawy. Zawiera dyskusję na temat osiągniętych rezultatów, wkładu pracy w rozwój dziedziny oraz praktycznych możliwości zastosowania przedstawionych w pracy rozwiązań.