Název práce: Interpretace a přizpůsobování jazykových jevů ve vícejazyčných modelech

Autor: Tomasz Limisiewicz

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí práce: David Mareček,
Ústav formální a aplikované lingvistiky

Abstrakt:

Jazykové modely založené na neuronových sítích se staly základem pro řešení nejrůznějších úloh, jejich vnitřní fungování však zůstává nejasné. Tato disertační práce zkoumá, které komponenty jazykových modelů jsou klíčové pro reprezentaci a zpracování textových informací. Zaměřujeme se především na vícejazyčné modely, které mohou využívat reprezentaci pro zpracování úloh napříč různými jazyky. Předpokládáme, že pochopení toho, jak modely reprezentují jazykové jevy, je nezbytné pro jejich následné přizpůsobování a zmírňování problémů, kterými tyto modely trpí. Za tímto účelem navrhujeme nové techniky pro interpretaci jednotlivých komponent jazykových modelů. Naše metody umožňují lokalizovat reprezentaci různých typů signálů v těchto modelech. Tato lokalizace nám umožňuje omezit naši analýzu na konkrétní komponenty a aplikovat cílené opravy modelů ke zmírnění genderové zaujatosti nebo zlepšení mezijazykového přenosu.

Klíčová slova: zpracování přirozeného jazyka, jazykové modely, vícejazyčnost, interpretovatelné strojové učení