

Abstrakt:

Systémy pro generování textu z dat by měly generovat texty odpovídající co nej-
přesněji vstupním datům. Pravidlové systémy tento aspekt zaručují, ale zaostávají
v plynulosti výstupů a možnostech přizpůsobení pro nové domény. Naopak neu-
ronové jazykové modely zvládají snadno generovat plynulé texty a přizpůsobovat
se novým doménám, ale jsou notoricky náchylné k produkci nepřesných výstupů.
V této práci zkoumáme, jak efektivně zakomponovat do systémů pro generování
textu z dat neuronové modely tak, abychom propojili výhody obou typů systémů.
Naše přístupy zakládáme na předtrénovaných jazykových modelech architektury
transformer. Tyto modely primárně používáme jako stavební bloky, díky kterým
mohou být systémy pro generování textu robustní a efektivně se učit z trénovacích
dat. Spolu s tím představujeme automatické evaluační metriky pro odhalování chyb ve
výstupech a sadu nástrojů pro předzpracování a vizualizaci datasetů pro generování
textu z dat. Analyzujeme také chování předtrénovaných a velkých jazykových modelů
ve specifických případech jako je popis jednotlivých relací ve znalostních grafech a
generování textů ze standardních datových formátů. Z našich experimentů vyplývá,
že ačkoli k použití neuronových jazykových modelů při generování textu z dat je
potřeba přistupovat s rozmyslem, neuronové komponenty mohou zlepšit plynulost
výstupních textů a přizpůsobitelnost systémů novým doménám. Přesnost výstupů
přitom může zůstat vysoká, pokud jsou modely používány pro konkrétní dílčí úkoly
pro zlepšení kvality textu. Cílem budoucího výzkumu by mělo být vyhodnocování
systémů pomocí vhodných evaluačních metrik na reálných problémech.