Název: Strojový překlad na základě jednojazyčných textů

Autor: Ivana Kvapilíková

Ústav: Ústav aplikované a formální lingvistiky

Vedoucí: doc. RNDr. Ondřej Bojar, Ph.D., Ústav aplikované a formální lingvistiky

Abstrakt: Současné systémy strojového překladu (SP) jsou závislé na existenci paralelních dat, tedy textů, které byly dříve přeloženy lidmi. Tento typ dat je drahý a je dostupný pouze pro několik jazykových párů v omezených doménách. Vznikl tedy nový výzkumný směr zaměřený na navrhování modelů schopných naučit se překládat z jednojazyčných textů, které jsou výrazně dostupnější než texty paralelní, např. z internetu. I když je působivé, že takové modely překládat skutečně dokáží, kvalita jimi vyprodukovaných výstupů je pro praktické aplikace stále nedostatečná. Tato disertační práce se snaží vylepšit jejich výkonnost. Zkoumáme stávající přístupy používání jednojazyčných zdrojů k trénování překladových modelů a navrhujeme novou techniku generování pseudo-paralelních trénovacích dat uměle, bez drahého lidského vstupu. Automaticky hledáme podobné věty v jednojazyčný korpusech v různých jazycích a ukazujeme, že jejich použití v počátečních fázích trénování SP vede k výraznému zvýšení kvality překladu. Poukazujeme také na omezení stávajících modelů SP založených na jednojazyčných textech, které si často nedokáží poradit s překladem pojmenovaných entit a obecně produkují nekvalitní překlady, zejména v podmínkách s opravdu omezenými zdroji, kde je k dispozici pouze malé množství jednojazyčných textů, které navíc patří do odlišných domén.

Klíčová slova: strojový překlad, neřízené učení, hluboké neuronové sítě, nízkozdrojové jazyky, zpracování přirozeného jazyka