Překladač z Matfyzu dohání v kvalitě běžné překladatele

Univerzita Karlova
Ovocný trh 5, Praha 1, 116 36
www.cuni.cz

Praha 8. září 2020 - Prestižní vědecký časopis Nature Communications publikoval studii realizovanou na Matematicko-fyzikální fakultě Univerzity Karlovy, která představila anglicko-český překladač CUBBITT založený na neuronových sítích, jenž při překladu novinových zpráv dosahuje kvality srovnatelné s výstupem profesionálních překladatelů. V zaslepeném testu byly automatické překlady hodnotiteli označeny jako v průměru o něco méně plynulé, ale obsahově mírně přesnější než překlady lidské.

Jednou z nejpopulárnějších aplikací metod umělé inteligence (AI) v oblasti zpracování přirozeného jazyka je automatický překlad z jednoho jazyka do jiného. Donedávna se předpokládalo, že pro kvalitní překlad je nutné velmi důkladné porozumění obsahu překládaného textu a že automatizovaný překlad kvalitou srovnatelný s výsledkem lidského překladatele tedy ještě dlouho nebude na dohled. V automatizovaném překladu, stejně jako v jiných oblastech AI, nastala ale díky pokroku v tzv. hlubokém učení v posledních letech doslova změna paradigmatu, která tuto dosavadní představu mění.
Jako významný úspěch se jeví výsledek experimentu, který byl pro překladový směr angličtina-čeština realizovaný na Matematicko-fyzikální fakultě Univerzity Karlovy ve spolupráci s vědci z Univerzity v Oxfordu (oba též absolventi MFF UK) a z týmu Google Brain. Autoři natrénovali neuronovou síť na česko-anglickém paralelním korpusu, což je kolekce autentických anglických textů a jejich protějšků přeložených do češtiny o celkové velikosti 58 milionů párů vět.

Výsledný překladač nazvaný CUBBITT autoři použili k přeložení vzorku anglických novinových textů. Tentýž vzorek byl nezávisle přeložen profesionálními překladateli z překladové agentury. Kvalita výsledných automatických i ručních překladů byla následně hodnocena 15 rodilými mluvčími češtiny, kteří měli posoudit přesnost a plynulost překladu. Hodnocení bylo slepé, tj. hodnotitelé neměli informaci o tom, kdo věty překládal.
„Výsledek srovnání můžeme považovat za průlomový. Automatický překladač sice nepatrně pokulhával za lidskými překladateli v hodnocení plynulosti, byl ale v průměru o něco přesnější, pokud jde o obsahovou správnost překladu. Naměřený výsledek byl statisticky signifikantní", uvedl hlavní autor studie **Mgr. Martin Popel, Ph.D.** z MFF UK. Podobné pozorování autoři učinili již v roce 2018, ovšem tehdy byly hodnoceny jen izolované věty (bez kontextu celého článku).

Jedna z nových myšlenek, díky které překladač dosáhl výrazného zlepšení oproti předchozím verzím, spočívala ve způsobu, jakým byla překladači při trénování střídavě předkládána autentická a syntetická paralelní data (páry českých vět a jejich automatických překladů do angličtiny). Velikost existujících autentických dat, tj. lidmi vytvořených anglicko-českých překladů, je z principu omezená a roste relativně pomalu. Proto se k nim přimíchávají ještě syntetická paralelní data, kde pro existující autentické texty v češtině byly jejich anglické protějšky vygenerovány automatickým překladem v opačném směru (tzv. backtranslation; nižší kvalita na straně vstupního jazyka, zde angličtiny, totiž při trénování překladače vadí méně). Velmi překvapivé experimentální pozorování spočívalo v tom, že je výhodnější neuronové síti překládat autentická a syntetická data nikoli rovnoměrně promísená, ale ve specificky vyváženém rytmu střídajících se autentických a syntetických bloků. Prvotní impuls pro zkoumání tohoto směru vznikl vlastně náhodou, když mísení zůstalo omylem vypnuté a tato „chyba"? způsobila okamžitý růst úspěšnosti překladače.

Autoři studie upozorňují, že i přes představený pokrok se situace zatím výrazně liší od jiných oblastí, kde se AI v posledních letech úspěšně utkává s člověkem. Zatímco například v šachu dnes AI poráží víceméně rutinně i nejlepší hráče světa, zde šlo o „soutěž" s běžnými (byť profesionálními) překladateli, kteří v danou chvíli ani nevěděli, že „soutěží". Měření navíc proběhlo pouze na specifickém žánru novinových textů a výsledky rozhodně nelze zobecňovat na překladatelskou práci jako celek.

Pro více informací o studii kontaktujte:

Mgr. Martin Popel, Ph.D.
Ústav formální a aplikované lingvistiky MFF UK
tel.: 951 554 278
e-mail: popel@ufal.mff.cuni.cz

ZA SPRÁVNOST:
Mgr. Václav Hájek

Tiskový mluvčí UK
Odbor vnějších vztahů
Univerzita Karlova
tel: +420 224 491 248
mob: 721 285 565
e-mail: pr@cuni.cz

POZNÁMKY:
Článek v Nature Communications je k dispozici zde: https://www.nature.com/articles/s41467-020-18073-9

Experiment proběhl v Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK. Bylo jej možné realizovat díky špičkovému technickému zázemí pracoviště, které disponuje výpočetním clusterem (přes 100 GPU s velkou pamětí a 2000 CPU) a mimo jiné provozuje uzel jazykové výzkumné infrastruktury LINDAT/CLARIAH-CZ, ve které je také experimentální verze překladače k dispozici veřejnosti na adrese https://lindat.cz/services/translation . Pracoviště nabízí studijní programy zaměřené na zpracování přirozeného jazyka a vede své studenty, aby byli zvyklí od začátku srovnávat své síly se zahraniční konkurencí. Hlavní autor studie Mgr. Martin Popel, Ph.D., implementoval představený překladač ještě jako doktorand a úspěšně se zúčastnil několika ročníků soutěž v automatickém překladu WMT Shared Task. Díky zmíněnému vybavení pracoviště například mohl v krátké době na jeden přípravný experiment spotřebovat 4 roky strojového času.

Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy byl založen v roce 1990 jako pokračování výzkumné a pedagogické činnosti bývalé Laboratoře algebraické lingvistiky, existující od počátku 60. let na Filozofické fakultě a později na Matematicko-fyzikální fakultě Univerzity Karlovy. Ústav je především výzkumnou institucí, která se zabývá mnoha tématy v oblasti počítačové lingvistiky a zpracování přirozeného jazyka a která se účastní mnoha výzkumných projektů na národní i mezinárodní úrovni. Je také koordinačním pracovištěm velké výzkumné infrastruktury LINDAT/CLARIAH-CZ, která podporuje výzkum v České republice i ve světě poskytováním jazykových zdrojů, nástrojů a služeb v oblasti jazykových technologií a digitálních humanitních věd. Ústav formální a aplikované lingvistiky nabízí komplexní výukový program jak pro bakalářský a magisterský stupeň (Bc., Mgr.), tak pro doktorské studium (Ph.D.) v oboru počítačové lingvistiky. Všechny programy se vyučují v češtině a angličtině. Ústav je také členem konsorcia evropských univerzit, které poskytují magisterský "double degree" program LCT (https://lct-master.org/).