Rozhledy matematicko-fyzikální

Jindřich Libovický

Neuronové sítě a automatický překlad

Rozhledy matematicko-fyzikální, Vol. 94 (2019), No. 4, 30–40

Persistent URL: http://dml.cz/dmlcz/148014

Neuronové sítě a automatický překlad

Jindřich Libovický,
Ludwig-Maximillians-Universität München & MFF UK, Praha

Co je to strojový překlad

Pod strojovým překladem si většina lidí představí nejspíš Google Translate a většina lidí si také nejspíš na vlastní oči vyzkoušela, jak funguje. Ten, kdo překladač používá častěji, si mohl všimnout, že zhruba před třemi lety se kvalita překladu, kterou služba poskytuje, dramaticky zlepšila. Důvodem bylo, že se změnila technologie, na které překlad stojí: překlad založený na statistických metodách nahradily neuronové sítě. Hodně lidí také asi překvapí, že překladač od Googlu není jediný a už vůbec ne nejlepší na světě a že se v kvalitě strojového překladu pořádají každoroční soutěže.

Automatický překlad se v mnohém liší od jiných problémů, které informatika řeší. Pokud například chceme najít na mapě nejkratší cestu z bodu A do bodu B, existuje algoritmus, který spolehlivě tuto cestu najde. Může se stát, že mapa, na které hledáme, je příliš velká, nebo prostě jen chceme ušetřit výpočetní čas (a elektrický proud). V takové situaci můžeme použít nějaký jiný algoritmus, který ušetří výpočetní čas a neudělá chybu větší než třeba 10%. V případě strojového překladu, kde místo matematických struktur pracujeme s lidským jazykem, se něco takového dělá celkem těžko. Není tak úplně jasné ani to, co to překlad je, jak pořádně jeho kvalitu nějak exaktně měřit a co by třeba znamenalo, že nějaký překlad je o 10 % horší než jiný.

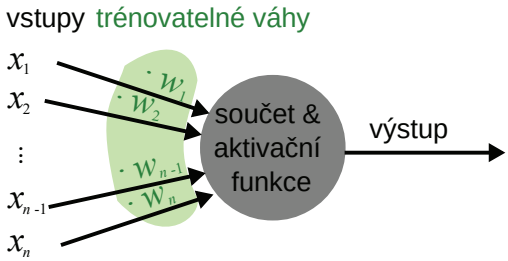
Nejspíš bychom se shodli na tom, že jedna věta je překladem do jiného jazyka, pokud tyto věty mají v těchto jazycích stejný význam. To ale vyvolává hned další otázku: co to ten význam, o kterém zde mluvíme, vlastně je. Jak se můžeme přesvědčit v lingvistice nebo filozofii jazyka, není to vůbec jednoduchá otázka. Odkazuje slovo „stůl“ k množině všech stolů světa v současnosti i v minulosti? Ke všem stolům, které by potenciálně mohly být vyrobeny? Nebo snad odkazuje k myšlence stolu? A je ta myšlenka u všech lidí stejná? Jaký význam má třeba slovo „vodník“ v jazyce Hindí, když v Indii o středoevropském vodníkovi jaktěživ neslyšeli? A mají vůbec slova sama nějaký význam nebo význam vzniká až v nějakém kontextu?

Ve strojovém překladu se naštěstí této složité filozofické otázce dokážeme docela dobře vyhnout. Lidé také zvládají překládat, aniž by měli dobře rozmyšleno, co je to význam. Když zvládneme dobře simulovat, *jak* překlad dělají lidé, nemusíme se věnovat otázce, *co* vlastně dělají. Tato myšlenka přímo navádí k tomu řešit strojový překlad pomocí metod strojového učení z vícejazyčných dat. Na Internetu je možné najít mnoho textů, které jsou ve více jazycích. Například oficiální dokumenty Evropské unie jsou výborným zdrojem takových dat. Pokud chceme dělat strojový překlad, stačí nám tato data posbírat (ne že by to byl sám o sobě jednoduchý úkol) a s jejich pomocí začít trénovat modely.

Neuronové sítě pro strojový překlad

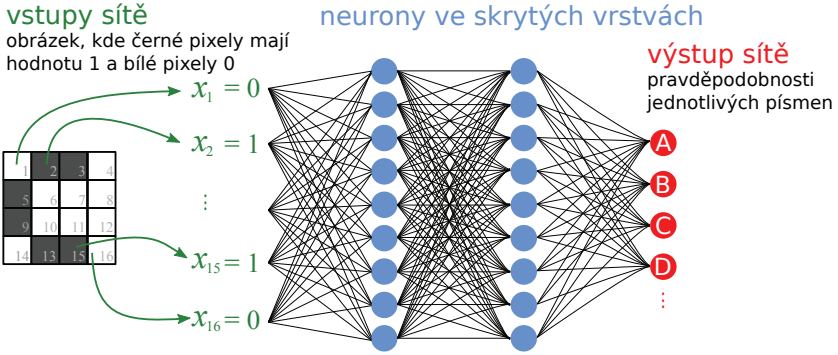
Začneme tím, jak vypadá to, čemu se v informatice říká neuronová síť. Umělé neurony, ze kterých se neuronová síť skládá, jakkoli by to jejich název mohl naznačovat, nemají dnes prakticky nic společného s nervovými buňkami živočichů. Neuronové sítě vznikly v padesátých letech 20. století skutečně jako jednoduchý výpočetní model biologického neuronu. Od té doby se ale výrazně změnila naše představa, jak biologické neurony fungují. Vývoj umělých neuronových sítí ale nešel cestou napodobování toho, co se děje v tělech živočichů, ale soustředil se na zlepšení úspěšnosti neuronových sítí ve strojovém učení. Mediálně vděčná srovnání počtu neuronů v mozku různých savců a umělých neuronových sítí používaných ve strojovém učení je potřeba brát s velkou rezervou.

Umělý neuron má mnoho vstupů, kde přijímá reálná čísla. Pro každý vstup má jednu trénovatelnou váhu, reálné číslo, kterým vstup vynásobí a všechny vstupy sečte. Když tento součet přesáhne určitou hodnotu, vydá na svůj výstup nějakou hodnotu. Právě váhy se v průběhu učení mění tak, aby neuron dával požadovaný výstup.



Obr. 1: Perceptron

Abychom vytvořili modely, které se mohou naučit složitější vztahy mezi vstupy a výstupy, spojujeme neurony do vrstev. Vrstvy je možné skládat dále za sebe. Takto nějak například může vypadat neuronová síť pro klasifikaci znaků při rozpoznávání textu.



Obr. 2: Vícevrstvá neuronová síť pro rozpoznávání znaků

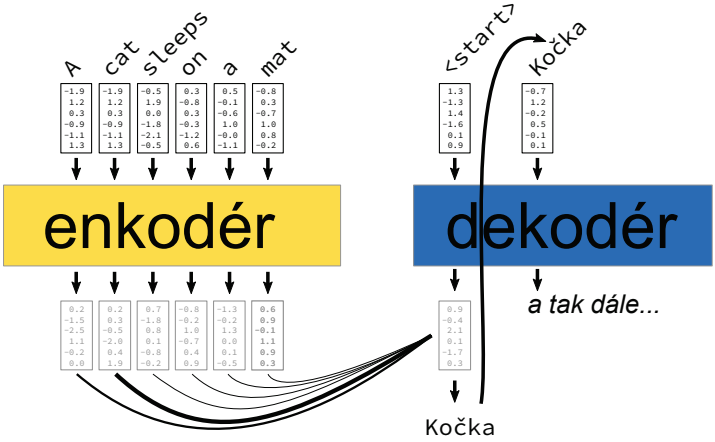
Na vstupu jsou hodnoty jednotlivých pixelů obrázku (u černobílého obrázku 0 a 1). Pro každý možný znak je na konci sítě neuron, jehož výstupem je pravděpodobnost, že na vstupu sítě byl ten který znak.

V tomto jednoduchém případě byl na vstupu obrázek, který měl vždy stejnou velikost (nebo je možné ho příslušně zvětšit nebo zmenšit). V případě strojového překladu je situace složitější, protože po neuronové síti chceme, aby byla schopná zpracovat vstupy různé délky a produkovat výstupy různé délky. Navíc délka vstupu a výstupu spolu sice souvisí, ale ne tak, že by bylo možné pro ně napsat nějaký jednoduchý vzoreček.

To se řeší tak, že model rozdělíme na dvě části: enkodér a dekodér. Enkodér zpracuje vstupní větu do číselné reprezentace. Dekodér potom na základě této reprezentace generuje jedno slovo za druhým, dokud nevygeneruje speciální symbol pro konce věty (viz obr. 3).

Model pracuje s omezeným slovníkem. Pro každé slovo, se kterým model dovede pracovat, má vstupní vektor. Už jsme zmínili, že tím, že použijeme neuronovou síť, se tak trochu vzdáváme možnosti vědět přesně, co model dělá. Jednou z mála věcí, kterou můžeme s jistotou tvrdit, je, že do těchto vektorů si ukládá informaci o významu slova, přesněji o tom, jak se může používat ve vztahu k jiným slovům. Slova s podobným významem se podobně používají a tak většinou mají podobné vektory. Podobné reprezentace mají také slova, která patří do stejných mluvnic-

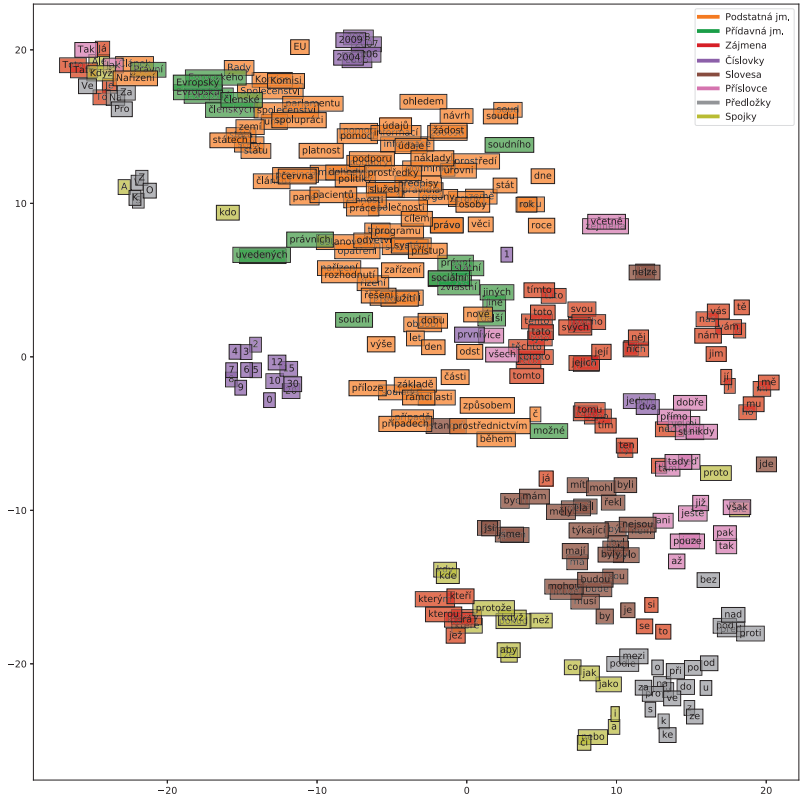
kých kategorií jako třeba pád nebo číslo. Pozoruhodnou vlastností je především to, že model nikdo neučí tyto kategorie rozlišovat. Neuronová síť se je naučí jaksi mimochodem, jako prostředek k tomu dělat dobrý překlad.



Obr. 3: Architektura enkodér–dekodér

Na obr. 4 vidíme dvourozměrnou projekci reprezentací 300 běžných českých slov, která jsou obarvená podle slovních druhů. Vidíme, že ne vždy model kategorizuje slova přesně tak, jak jsme se učili na základní škole. Například v levém horním rohu vidíme slova, která bývají často na začátku vět a mají tak pro překladač speciální význam. Hned vedle nich jsou podstatná a přídavná jména, která nějak souvisí s Evropskou unií. Tato zvláštnost je dána tím, že úřední dokumenty Evropské unie jsou důležitým zdrojem trénovacích dat pro strojový překlad. Jednou barvou vidíme odlišené číslovky, které se přirozeně rozdělily do dvou skupin: obecné číslovky (v dolní části obr. 4) a číslovky, které označují roky (v horní části obr. 4). Za povšimnutí stojí také to, že tvary vztažných zájmen „který“ a „jenž“ model přiřadil k neohebným spojkám, protože ve větě plní podobnou funkci.

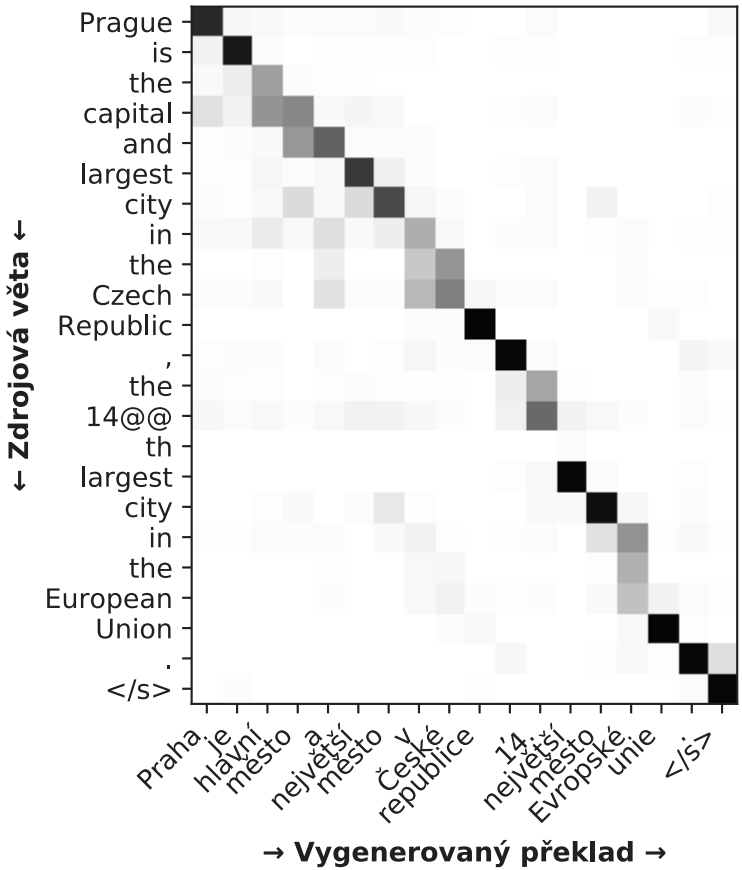
Tyto slovní reprezentace jsou vstupem do enkodéru. Pro ten se používají různé architektury. V současnosti nejlepších výsledků dosahuje architektura, kterou Google pojmenoval atraktivním jménem Transformer, která v každé vrstvě zohledňuje všechny možné vztahy mezi slovy vstupní věty.



Obr. 4: Dvourozměrná projekce reprezentací 300 českých slov z enkodéru natrénovaného překladače

Nejsložitější částí modelu je dekodér, který generuje větu v cílovém jazyce. Dekodér v každém kroku posbírá informaci ze vstupní věty a z textu, který už vygeneroval. Podle toho, co už je na výstupu, přiřadí dekodér váhu stavům enkodéru, tedy vlastně vybere nějaká vstupní slova a podle toho vygeneruje slovo na výstup.

Obr. 5 ukazuje, na jaká vstupní slova se dekodér zaměřoval při generování jednotlivých výstupních slov. Ukazuje se, že dekodér se učí přibližně párovat slova, která jsou navzájem svými překlady. V průběhu trénování nemá neuronová síť žádnou informaci o tom, jaká slova by si mohla ve vstupní a výstupní větě odpovídat, toto slovní párování je jenom jakýsi vedlejší produkt učení sítě.



Obr. 5: Párování slov z překladu a zdrojové věty, které dělá dekodér v neuronové síti

Celá tato celkem složitá neuronová síť, kterou jsme právě popsali, se musí naučit, co má dělat pouze pomocí trénovacích dat, která tvoří páry vět, které jsou navzájem svými překlady. Síť se skládá z desetitisíců neuronů, které mají stovky milionů vah – slovní vektory a váhy, kterými jednotlivé neurony váží svoje vstupy.

Na začátku učení jsou váhy náhodné. V průběhu učení se síti postupně předkládají trénovací věty po malých skupinách a váhy sítě vždy upraví tak, aby se trochu zvýšila pravděpodobnost, kterou síť přiřadí překladům v trénovacích datech. Tímto pozvolným učením se docílí toho, že síť

funguje v průměru dobře na všech datech a ne jen že si pouze zapamatuje svoje trénovací data. Pravděpodobnost trénovacích vět je jediný signál, který síť v průběhu trénování dostává. To jí stačí k tomu, aby se naučila ukládat do slovních reprezentací gramatické vlastnosti slov a navíc se naučila párovat slova mezi jazyky.

Jak vyhrát soutěž v překladu

Od roku 2006 se ve strojovém překladu pořádá každoroční soutěž, které se účastní univerzitní týmy z celého světa a v posledních letech také týmy velkých firem. Soutěží se v poměrně velkém počtu disciplín. Kromě překladu mezi různými jazykovými páry se soutěží také v překladu jazyků, ke kterým je k dispozici pouze omezené množství dat, nebo v překladu vět, které obsahují gramatické chyby a překlepy. Potom, co soutěžní týmy odešlou překlady vět, které dostaly od organizátorů, probíhá lidské hodnocení.

Hodnocení od lidí se pak také používá v další zajímavé soutěžní disciplíně, a tou je automatický odhad, jak lidé budou kvalitu překladu hodnotit. Jak už jsme se zmínili v úvodu, poznat, že dvě věty mají stejný význam, není vůbec jednoduché. I přesto, že současné metriky dosahují poměrně vysoké shody s lidským hodnocením, otázka, jak měřit kvalitu překladu tak, jak o ní přemýšlí lidé, zůstává stále otevřená. Částečně také proto, že nejen při překladu jako takovém, ale v případě automatického hodnocení jeho kvality, se snažíme vyhnout otázce, co je to kvalita překladu, a snažíme se pouze o dobrou simulaci toho, jak to dělají lidé.

Za hlavní soutěžní disciplínu se většinou považuje překlad mezi angličtinou a němčinou. Věnuje se mu nejvíce týmů a tento jazykový pár je navíc zajímavý i pro soukromé firmy. V roce 2019 vyvinula nejlepší systém společnost Amazon a umístila se těsně před Microsoftem, který zvítězil rok předtím. V několika předchozích letech pak soutěži dominovala univerzita ze skotského Edinburghu, která vítězila téměř ve všech disciplínách.

Především díky aktivitě Matematicko-fyzikální fakulty v oblasti strojového překladu od začátku nechyběl v soutěži překlad mezi češtinou a angličtinou. Čeština je pro výzkum strojového překladu zajímavá především svým poměrně složitým tvaroslovím. Systém, který překládá do češtiny, nemusí jenom správně vybrat slova ve správném pořadí, ale také musí zařídit, aby slova byla ve správném tvaru. Právě v překladu z angličtiny do češtiny se daří dlouhodobě vítězit týmu Matematicko-fyzikální fakulty.

Klíčovou metodou, která pomáhá vylepšovat kvalitu automatického překladu, je generování umělých trénovacích dat takzvaným zpětným překladem. Pokud například chceme natrénovat překladač z jazyka A do jazyka B, můžeme si pomoci už hotovým automatickým překladačem z jazyka B do jazyka A. K autentickým větám z jazyka B si totiž můžeme vygenerovat automatické překlady do jazyka A a rozšířit si tak trénovací data. Tyto překlady samozřejmě nejsou tak kvalitní, jako by byl lidský překlad, důležité ale je, že na výstupu jsou autentické věty z jazyka B, takže se překladač učí generovat vždy pouze plynulé autentické věty. Překladač z A do B, který jsme vylepšili pomocí syntetických dat, můžeme dále použít k vygenerování umělých dat pro trénování překladu na druhou stranu z B do A. Takhle je možné jazyky střídat a překlad postupně zlepšovat.

V umělé inteligenci se jedná o poměrně častý princip. V principu podobně se učil systém AlphaGo, který v roce 2016 porazil jednoho z nejlepších světových hráčů v deskové hře Go. AlphaGo je založeno na neuronové síti, která se učí na základě zpětné vazby z opakovaného hraní sama se sebou. V případě zpětného překladu se jedná o podobnou situaci, kde se dvě neuronové sítě navzájem vylepšují tím, že jedna druhé připravuje stále kvalitnější trénovací data.

Soutěžní systémy se ale liší od toho, co vidíme, když si pustíme Google Translate nebo Bing Translator od Microsoftu, kde výsledky překladu vidíme okamžitě. Soutěžní systémy jsou většinou kombinací více systémů, místo jedné věty generují více kandidátů, které se následně hodnotí pomocí jiných modelů. Překlad tak rozhodně nemůžeme vidět plynule vznikat v reálném čase, jak to vidíme v překladači od Googlu. Od pomyslného stisknutí enteru do okamžiku, než se vypíše překlad, tak mohou uplynout i více než dvě sekundy.

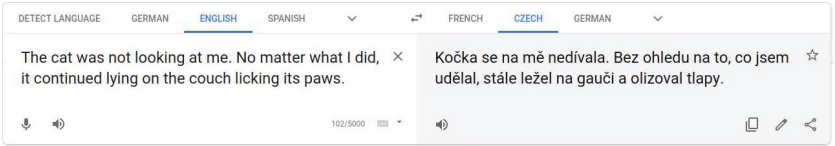
Kdy překlad nefunguje dobře

Strojový překlad samozřejmě není dokonalý a pravděpodobně také nikdy dokonalý nebude. Cílem strojového překladu je především ušetřit práci překladatelům s běžnými texty a pomoci lidem orientovat se v textech v cizím jazyce.

V současnosti největším problémem strojového překladu je, že se provádí na úrovni jednotlivých vět. Věty se ale vždy vyskytují v nějakém kontextu, který však překladač nemá k dispozici.

Jak vidíme na příkladu z Google Translate (obr. 6), systém nemá jak zjistit, že podmět „it“ ve druhé větě je ve skutečnosti kočka, která

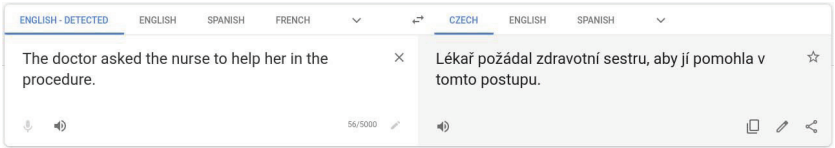
je v češtině ženského rodu. Jednoduché řešení, jakým by se mohlo zdát trénovat síť, která překládá celé dokumenty a ne jednotlivé věty, selhává, protože slova, která závisí na relativně vzdálených slovech v minulosti, se v textu vyskytují poměrně řídce a síť je při učení ignoruje. Jak řešit zapojení kontextu tak zůstává otevřenou otázkou.



Obr. 6: Překlad přes hranice vět (screenshot z Google Translate)

Jak už jsme několikrát zmínili, strojový překlad se řeší strojovým učením. Z toho plynou jeho silné i slabé stránky. V průběhu učení se překladač snaží najít statisticky nejjednodušší vysvětlení dat, na kterých se učí. Kromě toho, že to vede k tomu, že se síť omylem naučí rozlišovat slovní druhy, vede to také k tomu, že z dat vyvodí stereotypy, které mohou vést k chybným překladům. Příkladem může být přílišná asociace pojmů s genderem.

Když dáme do Google překladače větu: „The doctor asked the nurse to help her in the procedure," správný překlad by měl být něco jako: „Doktorka požádala sestru, aby jí pomohla se zákrokem." Že se jedná o lékařku a ne lékaře, je možné poznat podle zájmena „her" ve vedlejší větě. To je zjevně pro překladač příliš daleko a „doctor", který o něco svého asistenta či asistentku žádá, je v trénovacích datech výrazně častěji muž než žena. Výstup vidíme na obr. 7.

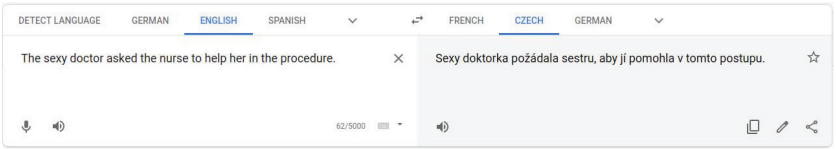


Obr. 7: Genderové stereotypy v překladu (screenshot z Google Translate)

Teď se podívejme, jak se změní výstup, když řekneme, že je lékařka „sexy" (obr. 8). Tento dodatečný přívlastek (a stereotyp, který se k němu váže) modelu pomůže pochopit, že se jedná o lékařku a ne lékaře.

Tento možná úsměvný příklad ukazuje velké nebezpečí, které s sebou nese používání neuronových sítí, které se učí z velkých dat, obzvlášť

za situace, kdy přesně nevíme, jakým způsobem se model dopracuje ke svému výstupu. Neuronová síť se snaží jenom napodobovat trénovací data. Přestože se často mluví o takzvaném hlubokém učení, často to dělá poměrně povrchním způsobem. Nemá a ani nemůže mít pochopení společenských okolností, za kterých data vznikají. Může se tak snadno stát, že z dat odvodí vzorce, které považujeme za stereotypní nebo urážlivé.



Obr. 8: Genderové stereotypy v překladu (screenshot z Google Translate)

U strojového překladu to nemusí tolik vadit. Ale pokud bychom neuronovou síť chtěli použít například k odhadování toho, jestli někdo bude schopen platit bankovní půjčku, je dost pravděpodobné, že se budeme na základě těchto stereotypů skrytých v neuronové síti dopouštět vážné diskriminace.

Vyzkoušejte si sami

Jak neuronový strojový překlad funguje nebo nefunguje si ostatně můžete vyzkoušet sami na webu. Kromě všeobecně známého překladače od Googlu se dá na webu najít i několik dalších, které překládají kvalitněji než Google. Důvodem není to, že by Google lépe překládat nedovedl, ale protože svůj překladač nabízí zdarma – výrazně větší modely by potřebovaly výrazně víc výpočetní síly, což by znamenalo výrazně větší náklady za něco, za co se neplatí. Jak už jsme zmínili výše, všechny online ukázky jsou také méně kvalitní než překlady, se kterými se soutěží.

ÚFAL MFF UK

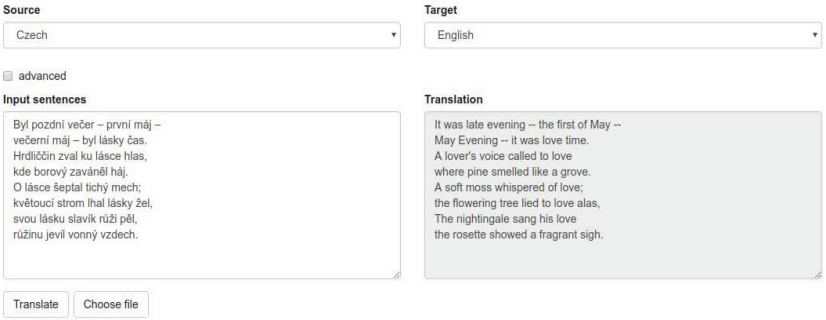
(`https://lindat.mff.cuni.cz/services/translation`)
Demonstraci toho, jak funguje strojový překlad s využitím modelu Transformer, můžete najít na webu Ústavu formální a aplikované lingvistiky MFF UK. Model, který si můžete vyzkoušet, je v současnosti nejkvalitnějším strojovým překladem mezi češtinou a angličtinou.

University of Edinburgh
(`http://demo.statmt.org`)
Na univerzitě ve skotském Edinburghu působí skupina, která se strojovému překladu věnuje dlouhodobě, a často jsou to edinburghské systémy, které vítězí v pravidelných soutěžích. Jejich soutěžní systémy si také můžete vyzkoušet na webu.

Systran
(`https://translate.systran.net/translationTools/text`)
Systran je společnost, která se komerčně věnuje strojovému překladu už od konce šedesátých let dvacátého století. Kromě obecných modelů připravuje modely, které se specializují na překlady finančních, právních nebo technických textů. Za kvalitní překlad si ale uživatelé musí připlatit a zdarma je možné vyzkoušet si pouze desítky vět.

Bing Translator
(`https://www.bing.com/translator`)
Velmi kvalitní automatický překlad nabízí poslední dobou také Bing Translator od společnosti Microsoft. Důvodem je, že se Microsoft nedávno ujal vývoje open-sourcového nástroje Marian, který se do té doby vyvíjel především na univerzitě v Edinburghu.

Posuďte sami:



Obr. 9: Máchův Máj v překladači z MFF UK (screenshot z LINDAT Translation)

Není to úplně špatné, jen se to moc nerýmuje.