Strojový překlad přirozeného jazyka

Martin Wörgötter

příspěvek pro konferenci Technického muzea v Brně „Od strojového kódu k programování neuronových sítí“ v září 2018 a pro web prog-story.technicalmuseum.cz

V tomto článku se budeme zabývat počítačovými programy, které dostávají na vstup text v přirozeném jazyce jako je čeština, angličtina, čínština, apod. a na výstup dávají text v určeném cílovém jazyce. Protože k jednomu zdrojovému vstupu přirozeného jazyka existuje mnohdy více možných překladů, nelze ani výsledek strojového překladu hodnotit binárně jako správný či špatný, ale je vhodné hodnotit jej pomocí metriky, která srovnává výstup programu s jedním či více referenčními překlady. Cílem je maximalizovat kvalitu, aby výstup byl v ideálním případě nerozeznatelný od překladu profesionálním překladatelem. Strojový překlad je obtížnou úlohou, což vysvětluje i kombinatorická exploze, ke které by vedlo prohledávání celého stavového prostoru. Uveďme pro příklad velikosti slovníků některých jazyků. Indoevropské jazyky jako čeština, angličtina, francouzština mají celkový počet slovních variant v řádu několika desítek milionů. S odstupem více má např. němčina s přibližně 90 miliony, což je způsobeno téměř libovolným skládáním substantiv do složenin.[1] Z tohoto pohledu jednodušší pro překlad se zdají být jazyky využívající znakové písmo. Japonština nebo čínština mají několik desítek tisíc znaků. Znaky se ovšem skládají do slov jejichž počet lze odhadnout na několik jednotek milionů. Navíc každý znak má zásadní vliv na význam celé výpovědi, což činí špatný překlad nesrozumitelným. V tomto srovnání jednodušší pro překlad je francouzština s angličtinou. Oba jazyky mají okolo 500 000 základních slov spisovného jazyka a variabilita slovních tvarů je nízká. Neexistence formální gramatiky přirozeného jazyka vede k tomu, že žádnou permutaci a výběr slov nemůžeme předem vyloučit.

Samozřejmě historicky první snahy k realizaci strojového překladu byly založené na pravidlech a formálních gramatikách. Víme, že často lze složité jevy v přírodě popsat pomocí matematických rovnic a modelů. Proto bychom chtěli podobně nahlížet na přirozený jazyk a popsat jej několika pravidly, která budou zohledňovat známé gramatické jevy, abychom dostali sice idealizovaný, ale použitelný model. Toho se bohužel nepodařilo dosáhnout. Nabízí se ale možnost jazyk modelovat na úrovni nižší, tedy z pohledu vzorů, které vytvářejí jeho stavební jednotky (slova, ale i části slov až po samotné znaky). Budeme dále nahlížet na text jako na posloupnost těchto jednotek.

Úspěšnou iniciativou se stal statistický přístup k modelování přirozeného jazyka. Pro predikci v posloupnosti následujícího slova se počítá pravděpodobnost každého

[1] Údaje vychází z webového korpusu.

Zdroj:
4. Within the overall amount, an indicative amount of 3 % will be used to cover the human and material resources required for effective administration and su- pervision of the assistance.
Referenční překlad:
4. Z celkové částky se použije orientační částka 3 % na zajištění lidských a materiálních zdrojů potřebných pro účinnou správu pomoci a dohled nad ní.
Strojový překlad:
4. v rámci celkové částky se orientační částka ve výši 3% budou použity na pokrytí lidské a materiální zdroje nezbytné pro účinnou správu pomoci a dohled nad ní.

Obrázek 1: Ukázka statistického strojového překladu z angličtiny

možného výstupního slova podmíněná kontextem, který je z výpočetních důvodů omezený na n předchozích slov. Přitom se maximalizuje pravděpodobnost celé posloupnosti. Elementární pravděpodobnosti se vypočítají z četností souvýskytu slov. Tyto četnosti se spočítají na *korpusu*, který je ideálně textovým záznamem celého jazyka[2]. Jedná se o výpočetně velmi náročný proces a velikost kontextu, tzv. *n-gramu* je často menší než 5, což je zcela nedostačující pro zachycení gramatických vazeb uvnitř věty. Tímto způsobem modelované věty mohou znít lokálně přirozeně, ale celkově nemusí dávat smysl viz příklad 1.

Právě řešení problému podmíněné predikce výstupu celým kontextem přinesly neuronové sítě. Budeme se zde zabývat pouze *state-of-the-art* řešením a vynecháme různé jiné architektury neuronových sítí, podobně jako hybridní statistické systémy.

Zřejmě nejrozšířenější pro problém strojového překladu je architektura zvaná *kodér-dekodér*. Jedná se v zásadě o dvě rekurentní neuronové sítě, z nichž první – *kodér* zpracuje po částech vstup a vytvoří jeho interní reprezentaci, tzv. *kontextový vektor*. Tato reprezentace je pak čtena *dekodérem*, který generuje výstup. Generování nového výstupního slova je podmíněno *kontextovým vektorem* a předchozími vygenerovanými slovy.

Nyní popíšeme podrobněji, jak rekurentní neuronová síť funguje. Předpokládáme, že všechna vstupní slova w náleží do slovníku $w \in D$. Slova kódujeme jako binární vektory způsobem *1 z N*. Zavedeme projekční matici C a vynásobením $Cx^\intercal$ dostaneme *embedding* slova v prostoru $\mathbb{R}^m$, kde m je velikost *embeddingu*. Nyní v každém kroku výpočtu čte kodér postupně sekvenci $x = (x_1, \ldots, x_n)$ a aktualizuje *skrytý stav* sítě podle následujícího vzorce 1.

$$h_t = f(UCx_t^\intercal, Wh_{t-1}), \tag{1}$$

kde f je diferencovatelná, nelineární funkce. U, W jsou matice vah. Definujeme h_0 jako stav reprezentující začátek sekvence. Po přečtení x_n je *skrytý stav* sítě reprezentací celého vstupu.

Dekodér je koncepčně podobný *kodéru*. Navíc ale obsahuje výstupní vrstvu, která počítá pravděpodobnosti slov nad slovníkem cílového jazyka a generuje slovo s nej-

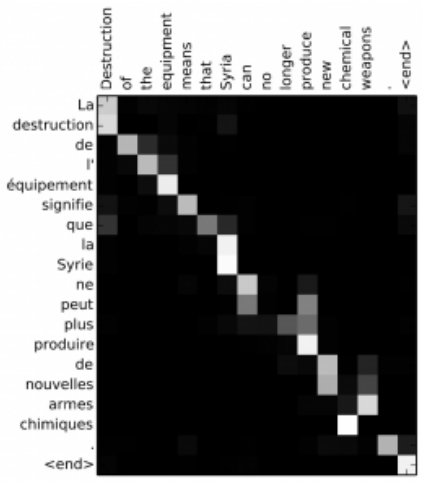
[2]Využívají se webové korpusy, které mají velikost v počtu slov řádově 10^9 a předpokládáme o nich, že se v nich všechny jazykové jevy vyskytují.

Zdroj:
the funds allocated and disbursed by the Union for each project of common interest.
Referenční překlad:
finanční prostředky vyčleněné a uhrazené Unií na každý projekt společného zájmu.
Strojový překlad:
finanční prostředky přidělené a vyplacené unií pro každý projekt společného zájmu .

Obrázek 3: Ukázka překladu pomocí neuronové sítě

vyšší pravděpodobností. Protože tento přístup často nevede ke globálnímu optimu – maximalizaci pravděpodobnosti celé výstupní sekvence, tak se využívá algoritmus *paprskového prohledávání*, jak znázorňuje schéma 4 na překladu věty 3.

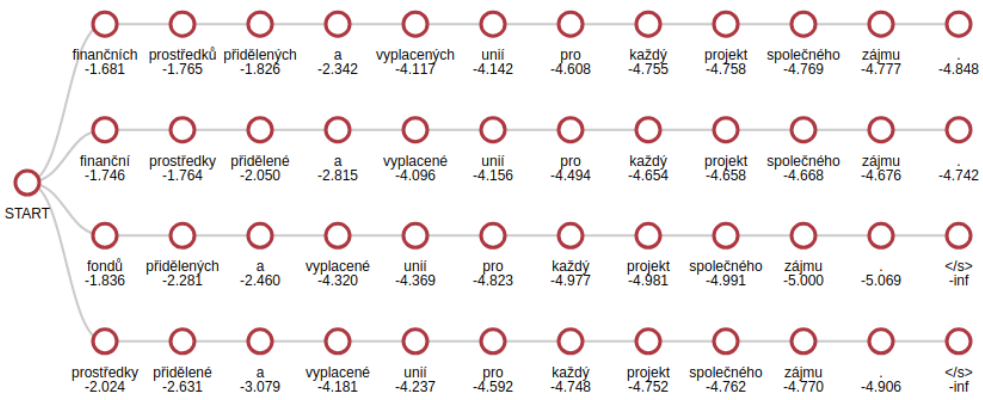
Kontextový vektor má v procesu překladu zásadní roli, protože kóduje veškerou informaci obsaženou ve vstupu. Protože *kontextový vektor* má fixní délku, může kódovat pouze omezené množství informace. Tento problém vyvstává zejména při překladu delších vět a způsobuje, že výstup je zkrácený a vynechává informace obsažené ve zdroji. Navíc informace ze začátku sekvence se při aktualizaci *skrytého stavu* postupně *ztrácejí*. Jako řešení problému byla navržena poměrně komplexní komponenta zvaná *zarovnávací model*, která může být zapojena přímo do neuronové sítě. Namísto zakódování vstupu do jednoho fixního vektoru je vstup reprezentován proměnným počtem vektorů, kde počet odpovídá délce vstupní sekvence, ze kterých následně zarovnávací model vybírá jejich podmnožinu, a to adaptivně při generování každého výstupního slova. Nejlépe ilustrujeme fungování zarovnávacího modelu na vizualizaci 2.



Obrázek 2: Matice zarovnání pro překlad z francouzštiny do angličtiny. Světlejší pole značí vyšší míru zarovnání mezi danými slovy. (Zdroj: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/)

Celá neuronová síť je trénována na *paralelním korpusu*, kde ke každé zdrojové větě existuje její překlad v cílovém jazyce. Trénování probíhá standardně metodou *gradientního sestupu*. Velký význam má přitom výpočetní síla hardwaru, kterou poskytují moderní grafické karty. Musíme zmínit, že trénování rekurentních neuronových sítí není přímo paralelizovatelné.

Na závěr se zamyslíme nad budoucností strojového překladu. Podle optimistických odhadů je strojový překlad téměř vyřešeným problémem. Stačí mít k dispozici dostatečné množství dat a výpočetní kapacity. Mezi aktuální překážky na cestě ke kvalitnímu překladu jsou omezená velikost slovníku a délka vět, obojí je spojeno s



Obrázek 4: Paprskové prohledávání s šířkou paprsku 4. (Vlastní ilustrace autora)

výpočetními nároky na hardware. Fundamentálnější problém do budoucna představuje práce s víceznačností v textu. V takových případech, kdy je věta víceznačná, jsou nutné *znalosti o světě* pro správnou interpretaci a až poté lze větu překládat. V tomto ohledu očekáváme, že systémy budou schopné zpracovávat větší množství vstupu a budou schopny disambiguovat jednotlivé věty na základě širšího kontextu.