Vytvorenie a natrénovanie strojového prekladu do slovenčiny pomocou neurónových sietí

Autor: Mgr. Matúš Kleštinec
Školiteľka: prof. RNDr. Daša Munková, PhD.
Konzultant: Mgr. František Forgáč

Strojový preklad

Strojový preklad je automatické konvertovanie textu z jedného prirodzeného jazyka do druhého prirodzeného jazyka pomocou modelu, ktorý je nutné natrénovať. Tento proces sa označuje aj ako Neural Machine Translation (NMT).

V rámci práce sa vykonali nasledovné kroky:

1. Získanie dvojjazyčného (angličtina-slovenčina) korpusu,
2. Predspracovanie korpusu- čistenie dát, úprava a tokenizácia,
3. Rozdelenie korpusu na trénovaciu, testovaciu a validačnú množinu,
4. Trénovanie a testovanie NMT modelu pomocou frameworku OpenNMT-py,
5. Evalvovanie NMT modelu pomocou metrík BLEU, METEOR a COMET,
6. Porovnanie modelov (pomenované V+číslo) s odlišnými parametrami trénovania a na základe porovnania vyhodnotiť vplyv sledovaných parametrov.

Predspracovanie korpusu Europarl

Popis kroku	Zostatok riadkov	Odstránené alebo upravené riadky
Vloženie viet do datasetu	640 715	-
Odstránenie NaN hodnôt	639 954	761
Normalizácia na základe unicode znakov	639 954	ANG(17) / SK(13)
ANG text = SK text	636 704	3250
Odstránenie duplicitných riadkov	620 889	15815
Odstránenie príliš dlhých viet	620 873	16
Dodatočné odstránenie NaN hodnôt	620 869	4
Kontrola jazyka pomocou Langdetect	611 479	9390
Zmena poradia riadkov	611 479	-

Vplyv veľkosti testovacej a validačnej množiny

Na základe pokusov s veľkosťou rozdelenia testovacej a validačnej množiny sme dosiahli výsledok, že najvhodnejšie rozdelenie sa pohybuje v rozmedzí ≈ 0,66% do ≈ 4,91%. Hodnoty sú orientačné, ale predstavujú približné rozdelenie, ktoré nám prinieslo najlepšie výsledky. Je nutné podotknúť, že v prípade väčšieho modelu by toto rozmedzie mohlo byť odlišné. Pracovali sme s relatívne malým korpusom, čiže nebolo možné experimentovať s väčším množstvom viet rovnakej kvality.

	BLEU	METEOR	COMET	Test/valid	Test/valid v %
V10	0,39361	0,6744	0,8978	2000	≈ 0,33
V5	0,39908	0,7038	0,8993	4000	≈ 0,66
V2	0,39889	0,6765	0,9008	6000	≈ 0,98
V3	0,39959	0,6890	0,9009	30 000	≈ 4,91
V4	0,40082	0,4942	0,8992	60 000	≈ 9,81

Vplyv veľkosti slovnej zásoby na kvalitu modelu

Pokusy s veľkosťou slovnej zásoby ukázali, že veľkosť slovnej zásoby môže ovplyvniť model. Príliš malá slovná zásoby môže negatívne ovplyvniť výsledný model, ale to platí aj pre príliš veľkú slovnú zásobu. Na základe našich poznatkov sme usúdili, že je vždy nutné zvážiť ideálnu veľkosť slovnej zásoby, keďže pre každé riešenie môže byť hodnota odlišná. Experimentovali sme len v jazykoch angličtina a slovenčina a s jednou veľkosťou korpusu, čiže naše hodnoty platia v prípade len tejto dvojice jazykov a pre približnú veľkosť korpusu. Jazyky majú odlišne bohatú slovnú zásobu, čo by tiež mohlo ovplyvniť vhodnú veľkosť slovnej zásoby modelu strojového prekladu. Väčší korpus by mohol obsahovať bohatšiu slovnú zásobu, čo by tiež mohlo ovplyvniť vhodnú veľkosť slovnej zásoby.

	BLEU	METEOR	COMET	Slovná zásoba	Kroky trénovania
V12	0,39867	0,7843	0,9024	2000	100 000
V9	0,42028	0,6905	0,9058	4000	100 000
V11	0,45228	0,6905	0,9131	8000	100 000
V6	0,46227	0,6932	0,9147	16 000	80 000
V7	0,47777	0,7937	0,9147	32 000	55 000
V8	0,44686	0,7351	0,9194	50 000	35 000

Vplyv algoritmu tokenizácie na kvalitu modelu

Na základe porovnania modelov s odlišnými algoritmami tokenizácie sme sa dopracovali k výsledku, že byte pair encoding (skr. BPE) vykazuje lepšie výsledky ako unigram. Pokus s algoritmami tokenizácie bol vykonaný na malej vzorke, respektíve sme porovnávali len dva modely medzi sebou. Vzorka je malá, ale všetky metriky, hlavne METEOR poukazujú na prepad kvality prekladu v prípade použitia algoritmu unigram.

	BLEU	METEOR	COMET	Algorit. token.
V5	0,39908	0,7038	0,8993	BPE
V15	0,38804	0,6220	0,8819	unigram

Vplyv veľkosti korpusu na model

Pre získanie predstavy ako veľmi ovplyvní model veľkosť korpusu sme Europarl rozdelili na približne polovicu. Mohli sme vidieť prepad u všetkých troch metrík, najvýraznejšie u metriky METEOR. Model mal problém preložiť aj vetu z testovacej množiny a tým zmenil úplne celý kontext vety. Týmto pokusom sme chceli poukázať na dôležitosť počtu viet, na ktorých trénujeme. Je nutné podotknúť, že u väčšieho korpusu by sme tak výrazný prepad kvality nemuseli vidieť, ale to by si vyžadovalo väčší korpus na ktorom by sme to mohli otestovať.

	BLEU	METEOR	COMET	Kroky trénovania	Počet viet
V5	0,39908	0,7038	0,8993	100 000	611 479
V17	0,33371	0,1195	0,8819	85 000	305 456

Záver

Tieto výsledky môžu poskytnúť hlbšie pochopenie fungovania strojového prekladu a ukázať, aké dôležité je správne nastavenie jednotlivých parametrov trénovania a prekladu. Je náročné určiť, ktoré parametre sú potrebné a aké sú ideálne ich nastavenia, aby sme dosiahli najlepší model. Výsledný model môže byť ovplyvnený veľkým množstvom faktorov ako kvalita korpusu, veľkosť korpusu, spôsob predspracovania korpus, tokenizovanie korpus, rozdelenie viet na množiny a mnoho iných parametrov, ktoré môžu ovplyvniť samotné trénovanie. Možností na experimentovanie je toľko, že by bolo časovo a výpočtovo náročné všetko vyskúšať, čo môžeme považovať zároveň za najväčšiu limitáciu tejto práce. Ďalšou limitáciou je relatívne malé množstvo kvalitných dvojjazyčných textov v jazykoch angličtina a slovenčina. S väčším objemom viet, by sme dokázali lepšie porovnávať preklady bez rizika, že náš preklad je ovplyvnený malým množstvom viet, na ktorých by sa mohol model strojového prekladu učiť.