Nízko-zdrojové filtrovanie korpusu pomocou viacjazyčných vložiek

Vishrav Chaudhary* Yuqing Tang* Francisco Guzman* Holger Schwenk* Philipp Koehn“
*Facebook AI „Johns Hopkins University
{vishrav,yuqtang, fguzman,schwenk}@ fb.com phi@jhu.edu

Abstraktné

V tomto dokumente popisujeme naše podanie dozdieľanej úlohy filtrovania nízkozdrojového korpusu WMT19.Náš hlavný prístup je založený na súbore nástrojov LASER (Jazykovo-agnostické zastúpenia vety), ktorý používa encoder-dekoderovú architektúru vyškolenú na paralelnomkorpuse na získanie viacjazyčných vetových reprezentácií.Potom používame reprezentáciepriamo na skóre a filtrovať hlučné paralelné vetybez ďalšieho tréningu funkcie bodovania.Porovnávame náš prístup k iným sľubným metódam a dokazujeme, že LASER prináša silné výsledky.Nakoniecvytvoríme súbor rôznych skórovacích metód a získame ďalšie zisky.V porovnaní s druhými najlepšími systémami dosiahla naša ponuka najlepší celkový výkon tak pre nepálsko-anglické úlohy 1M, ako aj pre úlohy Sinhala-English 1M, resp. 1,4 BLEU.Navyše, naše experimenty ukazujú, že táto technika je sľubná pre scenáre s nízkymi a dokonca žiadnymi zdrojmi.

1 Úvod

Dostupnosť vysokokvalitných paralelných školiacich údajov je rozhodujúca pre dosiahnutie dobrého výkonu prekladu, keďže systémy NMT sú menej odolné voči hlučným paralelným údajom ako systémy štatistického strojového prekladu (Khayrallah aKoehn, 2018).Nedávno sa zvýšil záujem o filtrovanie hlučných paralelnýchkorpusov (ako je Paracrawl1)ozvýšenie množstva dát, ktoré možno použiť na výcvik prekladateľských systémov (Koehn akol., 2018).

Zatiaľ čo najmodernejšie metódy, ktoré používajú NMT modely, sa ukázali ako účinné pri ťažbe Paralelné vety (Junczys-Dowmunt, 2018) pre jazyky s vysokým zdrojom, ich účinnosť nebola testovaná v jazykoch s nízkym zdrojom.Dôsledky nízkej dostupnosti údajov o odbornej príprave pre metódy paralelného označovania zatiaľ nie sú známe.

Za úlohu nízkozdrojového filtrovania (Koehnet al., 2019)sme dostali veľmi hlučné 40,6 milióna slov (počet anglických tokenov) Nepali- anglický korpus a 59,6 milióna slov Sinhala- anglický korpus vyliezol z webu ako súčasť projektu Paracrawl.Problém spočíva v poskytovaníbodov pre každý pár viet v oboch hlučných paralelných súboroch.Skóre sa použije na podvzorky párovviet, ktoré predstavujú 1 miliónov a 5 miliónov anglických slov.Kvalita výslednýchpodmnožín je určená kvalitou štatistického strojovéhoprekladu (Mozes, frázy (Koehn a kol., 2007))aneurálneho strojového prekladu Fairseq(Ott a kol., 2019) vyškoleným natieto údaje.Kvalitasystému strojového prekladu bude meraná podľa BLEU skóre pomocou Sacrebleu(Post,2018) navybranej testovacej sade Wikipédie preklady pre Sinhala-anglický a nepali-anglický z flores dataset (Guzman a kol., 2019).

V našom podaní pre túto zdieľanú úlohu používame viacjazyčné vkladanie vety získané od LASER,ktorý[1] používa encoder-dekodér architektúru natrénovanie viacjazyčného zastúpenia vety s použitím relatívnemalého paralelného korpusu.Naše experimenty ukazujú, že navrhovaný prístup prekonáva iné existujúce prístupy.Okrem toho využívame súbor viacerýchbodovacích funkcií na ďalšie zvýšenie filtračného výkonu.

[1]http://www.paracrawl.eu/

Rokovaniaštvrtejkonferencie o strojovom preklade (WMT), zväzok 3:Dokumenty o zdieľanej úlohe (Deň 2) s. 261 – 266Florence , Taliansko, 1. – 2. augusta 2019.

[1]https://github.com/facebookresearch/LASER

2 Metodika

Zdieľaná úlohaWMT 2018 pre paralelné filtrovanie korpusu (Koehnakol., 2018)[3] zaviedla niekoľko-metód na riešenie vysokozdrojového nemeckého a anglického dátového stavu.Zatiaľ čo mnohé z týchto metód boliúspešné pri filtrovaní hlučných prekladov, málo znich bolo vyskúšaných za podmienok s nízkym zdrojom.V tomto dokumente riešime problém nízkozdrojového filtrovania viet pomocou reprezentácií na úrovni viet a porovnávame ich s inými populárnymi metódami používanými vo vysoko-zdrojových podmienkach.

Model LASER (Artetxe a Schwenk, 2018a) využíva viacjazyčné predstavy vetyna meranie podobnosti medzi zdrojom a cieľovou vetou.Poskytuje najmodernejšie výkony na baníckej úlohe Bucc corpus a tiež účinne filtruje dátaWMT Paracrawl (Artetxe aSchwenk, 2018a) .Tieto úlohy sa však týkali len jazykov s vysokým zdrojom, konkrétne francúzštiny, nemčiny, ruštiny a čínštiny.Našťastie, táto technika bola účinná aj v prípade nula-shot cross-lingual natural language indukcie v súbore dát XNLI (Artetxe a Schwenk, 2018b), čo sľubuje, že scenár s nízkymi zdrojmi bude zameranýna túto spoločnú úlohu.V tomto dokumente navrhujeme použiť adaptáciu LASER na nízke zdrojové podmienky na výpočet skóre podobnosti na filtrovaniehlučných viet.
Pre porovnanie s LASER, sme tiež stanoviť počiatočnéreferenčné hodnoty pomocou Bicleaner a Zipporah, dve populárne základné línie, ktoré boli použité v projekte Paracrawl;A dvojitá podmienená crossentropia, ktorá sa ukázala ako najmodernejšia pre úlohu filtrovania vysokozdrojového korpusu (Koehn a kol.,2018).Skúmame výkon techník za podobných predbežných podmienoktýkajúcich sa filtrovania jazykovej identifikácie a lexikálneho prekrývania.Pozorujeme, že skóre LASER poskytuje jasnú výhodu pre túto úlohu.Nakoniec vykonávame zhromaždenie bodov pochádzajúcich zrôznych metód.Pozorujeme, že keď sú v zmesi zahrnuté skóre LASER, zvýšenie výkonu je relatívne malé.Vo zvyšku tejto časti diskutujeme o nastaveniach pre každú z použitej metódy.

[3]http://statmt.org/wmt18/paralelný-korpus-filtering.html

2.1 Laserové viacjazyčné zastúpenie

Základnou myšlienkou je použiť vzdialenosti medzi dvoma viacjazyčnými reprezentáciami ako pojem paralelizmu medzi dvoma vloženými vetami (Schwenk, 2018).Aby sme to dosiahli, najprv vycvičíme kódovač, ktorý sa naučí vytvárať viacjazyčné, pevné zastúpenie vety;a potom vypočítajte vzdialenosť medzi dvoma vetami v učenom priestore.Okrem toho používame *maržové* kritérium, ktoré používa prístup k najbližším susedom normalizovať skóre podobnosti vzhľadom na to, že cosine podobnosť nie je globálne konzistentná (Artetxea Schwenk, 2018a).

Encoder Viacjazyčný kódovačsa skladá z obojsmerného LSTM, a naša veta embeddingssú získané použitím max-pooling cez jeho výstup.Používame jediný kódovač a dekodér v našom systéme, ktoré zdieľajú všetky zúčastnené jazyky.Na tento účel sme trénovali viacjazyčné vety len na poskytnutých paralelných údajoch (podrobnosti pozri v časti 3.2).

Marža Sledujeme definíciu*pomeru z* [2] (Artetxe aSchwenk, 2018a).Použitím tohto, môže byť skóre podobnosti medzi dvoma vetami (x, y) vypočítané ako

$$\frac{2\,ks\,(x, y)}{Sy'eNNfc(x)\ ^{cos}(^{x,\ y/}) + Sx'_{eNNfc(}y)\ cos\ ^{(}x', y^{)}}$$

Kde $NNK_{(}$ x) označuje k najbližších susedov x v inom jazyku, a analogicky pre NNK ($_{y}$).Všimnite si, že tento zoznam najbližších susedov neobsahuje duplikáty, takže aj v prípade, že daná vetamá viac výskytov v korpuse, to by malo (najviac) jeden záznam v zozname.

Okolie Okrem toho sme preskúmali dva spôsoby odberu vzoriek k najbližším susedom.Najprv *globálna* metóda, v ktorej sme využívali okolie, pozostávali z hlučných dát spolu s čistými dátami.Po druhé, *miestna*metóda, v ktorej sme ohodnotili len hlučné dáta pomocou hlučnej štvrte, alebo čisté dáta pomocou čistej štvrte. [3]

2.2 Iné metódy podobnosti

Zipporah (Xu a Koehn, 2017;Khayrallah et al., 2018), ktorý sa často používa ako východiskové porovnanie, používa jazykový model askóre prekladu slov, s váhami optimalizovanými pre oddelenie čistých a syntetických údajov o hluku.V našom nastavení sme trénovali Zipporah modely pre oba jazykové páry Sinhala-anglický a nepálsky-anglický.Použili sme open source verziu6nástroja Zipporah bez úprav.Všetky komponenty modelu

[2]Preskúmali sme absolútne, *vzdialenosť* a *pomer*maržové kritériá, ale tie fungovali najlepšie

[3] Táto posledná časť bola vykonaná len pre tréning súboru

Zippora (probabilistické prekladové slovníkya jazykové modely) boli školené na poskytnutých čistých údajoch (s výnimkou slovníkov).Jazykové modely boli vycvičené pomocou KenLM (Heafield a kol., 2013) cez čisté paralelné dáta.Nepoužijeme poskytnuté jednojazyčné dáta ako predvolené nastavenie.Na cvičenie sme použili vývojovú sadu z floresových dát.

Bicleaner (Sanchez-Cartagena et al., 2018) používa lexikálny preklad a skóre jazykových modelov a niekoľko plytkých funkcií, ako sú:príslušná dĺžka, zodpovedajúce čísla a interpunkcia.Rovnako ako u Zip-porah, sme použili open source Bicleaner7toolkit[nemodifikovaný] out-of-the-box.Na výcvik tohto modelu sa použili iba poskytnuté paralelné údaje.Bicleaner používa komponent založený na pravidlách na identifikáciu hlučnejších príkladov v paralelných dátach a školí klasifikátora, aby sa naučil, ako ich oddeliť od ostatných údajov o výcviku.Použitie funkcií jazykového modelu je voliteľné.Používali sme len modely bez skórovacieho komponentu jazykového modelu.8

Dual Conditional Cross-Entropy Jeden z Najlepšie výkonné metódy na túto úlohu boli duálne podmienené cross-entropie filtrovanie(Junczys—Dowmunt, 2018), ktoré používa kombináciu dopredu a dozadu modely na výpočet krížovej lingválnej podobnosti skóre.V našich experimentoch sme pre každú jazykovú dvojicu použili poskytnuté čisté školiace údaje na výcvik modelov prekladu neurálneho stroja v oboch smeroch prekladu:zdroj k cieľu a cieľ k zdroju.Vzhľadom na takýto preklad-modelu M, sme silový dekódovať vety páry (x, y) zhlučného paralelného korpusu a získať skóre cross-entropy

$$HM(y|x) = \frac{1}{|y|}\ {}^{\log pm}(\text{YTLY}[i,t\text{-}i],x) \quad (1)$$

[6]https://github.com/hainan-xv/zipporah
[7]https://github.com/bitextor/bicleaner
[8]Zistili sme, že zahrnutie LM ako extra hry viedlo k tomu, že takmer všetky páry viet dostali skóre 0.

Skóre dopredu a dozadu krížovej entropie, $H_F(y|x)$ a Hb ($_x$ | y), sú potom v priemere sdodatočným trestom na veľký rozdiel medzi dvoma bodmi | Hf (y |$_x$) – Hb (xM.

$$\text{Score}(x, y) = \frac{\sout{H_F(y|x)} + \sout{H_B(x|y)}}{-|H_F(y|x) - H_B(x|y)|} \quad (2)$$

Vpredu a dozadu sú päťvrstvové enkodérové/dekodérové transformátory trénované pomocou fairseq sparametrami, ktoré sú identické s parametrami používanými vmodeli základných flores [45].Modelyboli vycvičené na čisté paralelné dáta pre 100 epoch.Pre nepálsko-anglickú úlohu sme tiež skúmali pomocou hindsko-anglických dát bez väčších rozdielov vo výsledkoch.Použili sme vývoj flores set na výber modelu, ktorý maximalizuje skóre BLEU.

2.3 Súbor

Aby sme využili silné a slabé stránky rôznych bodovacích systémov, preskúmali sme použitie binárneho klasifikátora na zostavenie súboru.Hoci je banálne získať klady (napr. čisté školiace údaje), ťažba negatívov môže byť skľučujúcou úlohou.Preto používame pozitívne neoznačené (PU) učenie (Mordeleta Vert, 2014), ktoré nám umožňuje získať triedičov bez toho, aby sme museli spracovať súbor údajov explicitne pozitívnych a záporných.V tomto nastavení pochádzajú naše pozitívne označenia z čistých paralelných údajov, zatiaľ čo neoznačené údaje pochádzajú z hlučného súboru.

Na dosiahnutie tohto cieľa používame balenie 100 slabých, zaujatých klasifikátorov (t. j. s 2:1 pre neoznačené údaje oproti pozitívnym údajom značenia).Používame podporné vektorové stroje (SVM) s radiálnym jadrom a náhodne podvzorujeme sadu funkcií pre tréningkaždého základného klasifikátora, čím pomáhame udržať ich rôznorodú a nízku kapacitu.

Absolvovali sme dve iterácie tréningu tohto súboru.V prvej iterácii sme použili pôvodné pozitívnea neoznačené údaje popísané vyššie.Na druhú iteráciu sme použili poučeného klasifikátora na preznačenie údajov o výcviku.Preskúmali sme niekoľkoprístupov k opätovnému označeniu (napr. stanovenie prahovej hodnoty, ktorá maximalizuje *skóre F1*).Zistili sme však, že stanovenie triednej hranice na zachovanie pôvodného pomeru pozitív k neoznačeniu najlepšie fungovalo.Pozorovali sme tiež, že výkon sa po dvoch iteráciách zhoršil.

3 Nastavenie experimentu

Experimentovali sme s rôznymi metódami pomocou nastavenia, ktoré presne odráža oficiálne skóre zdieľanej úlohy.Všetky metódy sú školené na poskytnutých čistých paralelných údajoch (pozri tabuľku 1).Nepoužili sme dané jednojazyčné údaje.Na vývojovéúčely sme použili dodávané flores dev set.Na vyhodnotenie sme vyškolili systémy strojového prekladu na vybraných podmnožinách (1M, 5M) hlučných paralelných tréningových dát pomocou fairseq s predvolenou konfiguráciou flores tréningového parametra.Nahlasujeme Sacrebleu skóre na flores DevTest set.Vybrali sme si náš

[4]https://github.com/facebookresearch/Flores# Vlak-a-základný-transformer-model

hlavný systém na základe najlepších skóre na DevTest sade pre 1M stav.

	si-sk	ne-sk	hi-en
Rozsudky	646k	573k	1.5
Anglické slová	3.7	3.7 MIL.	20.7

Tabuľka 1:Dostupné bitexty na trénovanie filtračných priblížení.

3.1 Predbežné spracovanie

Použili sme sadu filtračných techník podobných tým, ktoré sa používajú v LASER (Artetxea Schwenk, 2018a) a priradili sme skóre —1 na hlučné vety založené na nesprávnom jazyku buď na zdroji alebo na cieľovej strane, alebo s prekrývaním aspoň 60 % medzi zdrojom a cieľovými tokenmi.Na filtrovaniejazykových id sme použili fastText10.-Vzhľadom k tomu, LASER počíta skóre podobnosti pre pár viet pomocou týchto filtračných techník, experimentovali sme pridaním týchto k ostatným modelom, ktoré sme použili pre túto zdieľanú úlohu.

3.2 Školenie laserového Encoderu

Pre naše experimenty a oficiálne podanie sme trénovali viacjazyčný kódovač viet s použitím-povolených zdrojov v tabuľke 1. Vyškolili sme jediný encoder využívajúci všetky paralelné dáta pre Sinhala-anglický, nepálsky-anglický a hindský jazyk.Keďže Hindi a Nepali zdieľajú rovnaké písmo, zhromaždili sme ich korpusy do jedného paralelného korpusu.Aby sme zohľadnili rozdiel vo veľkosti paralelných tréningových údajov, preverili sme bitexty Sinhala-anglického a Nepálu/Hindi-angličtinyhov pomere 5:3.To malo za následok zhruba 3,2 M výcvikové vety pre každý smer jazyka, tj. Sinhala a kombinovať Nepali-Hindi.

[109]https://fasttext.cc/docs/en/jazyková identifikácia. html

Modely boli vycvičené s použitím rovnakého nastavenia ako verejný kód LASER, ktorý zahŕňa normalizáciutextov a tokenizáciu s Mojžišovými nástrojmi (späť do anglického režimu).Najprv sa naučíme spoločnú 50k BPE slovnú zásobu na skrátené školiace dáta pomocou fastBPE.[6]Encoder vidí Sinhala, Nepali, hindčina a anglické vety na vstupe, bez akýchkoľvek informácií o aktuálnom jazyku.Tento vstup je vždy preloženýdo angličtiny.[7]Experimentovalisme s rôznymi-technikami na pridanie šumu do anglických vstupných viet podobných tomu, čo sa používa v neurálnom strojovom preklade bez dozoru, napr. (Artetxe et al., 2018; Lample a kol., 2018), ale to-nezlepšilo výsledky.

Enkodér je päťvrstvový BLSTM s 512 rozmerovými vrstvami.Dekodér LSTM má jednu skrytú vrstvu veľkosti 2048, trénovaný s optimalizátorom Adam.Pre vývoj vypočítamechybu podobnosti na zoskupenie flores dev sady pre Sinhala-anglický a nepálsky-anglický.Naše modely boli trénované pre sedem epoch asi 2,5 hodiny na 8 GPU Nvidia.

4 Výsledky

Z výsledkov v tabuľke 2 pozorujeme niekoľko trendov:I) skóre pre 5M stav je vo všeobecnosti-nižšie ako pri 1 M stave.Zdá sa, že táto podmienka sa zhoršuje použitím jazykového id a prekrývania filtrovania.(ii) LASER vykazuje trvale dobrý výkon.*Miestna* štvrť funguje lepšie ako tá *globálna*.V tomto nastavení je LASER v priemere 0,71 BLEU nad najlepším systémom non-LASER.Tieto medzery sú vyššie pri 1M stave (0,94 BLEU).(*iii*) Najlepšia konfigurácia súboru poskytuje malé vylepšenia v porovnaní s najlepšou konfiguráciou LASER.Pre Sinhalu-anglickú najlepšiu konfiguráciuzahŕňa všetky ostatné bodovacie metódy (ALL).Pre nepali-anglický je najlepšou konfiguráciou súbor LASER skóre.(iv) Dvojitá krížová entropiavykazuje zmiešané výsledky.V prípade Sinhaly-anglického jazyka funguje iba vtedy, keď je povolené filtrovanie id jazyka, čo je v súlade s predchádzajúcimi pozorovaniami (Junczys-Dowmunt, 2018).Pre Nepali – angličtinu, poskytuje skóre hlboko pod ostatnými bodovacími metódami.Všimnite si, že sme nevykonali prieskum architektúry.

Metóda	česk		si-sk	
	1M	5M	1M	5M
Zipporah				
základňa	5.03	2.09	4.86	4.53
+ VEKO	5.30	1.53	5.53	3.16
+ Prekrytie	5.35	1.34	5.18	3.14
Duálny X-Ent.				
základňa	2.83	1.88	0.33	4.63+
+ VEKO	2.19	0.82	6.42	3.68
+ Prekrytie	2.23	0.91	6.65	4.31
Bicleaner				
základňa	5.91	2.54+	6.20	4.25
+ VEKO	5.88	2.09	6.36	3.95
+ Prekrytie	6.12+	2.14	6.66+	3.26
LASER				
miestne	*7.37*	**3.15**	*7.49*	5.01
globálne	6.98	*2.98*	7.27	4.76
Súbor				
VŠETKY	6.17	2.53	**7.64**	**5.12**
Laserový *glob.*+	**7.49**	2.76	7.27	*5.08*

Tabuľka 2:Sacrebleu skóre na `flores` DevTest set.Tučným `písmom`zvýrazňujeme najlepšie skóre pre každú podmienku.*Kurzívou* zvýrazníme* bežca hore.Tiež signalizujeme najlepšiu non-LASER metódu s +.

[6] Https://github.com/glample/fastBPE

[12To]znamená, že musíme trénovať anglický autoenkodér.-Nezdalo sa, že by to bolelo, pretože ten istý encoder ovláda aj ďalšie tri jazyky.

4.1 Diskusia

Jedna prirodzená otázka, ktorú treba preskúmať, je, ako by metóda LASER mala prospech, ak by mala prístup k dodatočnýmúdajom.Aby sme to preskúmali, použili sme súbor LASER open-source toolkit, ktorý poskytuje vyškolený encoder-pokrývajúci 93 jazykov, ale nezahŕňa nepali.V tabuľke 4 pozorujeme, že vopred vyškolený model-LASER prekonáva miestny *model LASERo* 0.4 BLEU.Pre nepálsko-anglický stav je situácia opačná:Laser *lokálne* poskytuje oveľa lepšie výsledky.Avšak výsledky predtrénovaného LASERu sú leno niečo horšie ako výsledky Bicleaner (6.12), ktorý je najlepšou metódou, ktorá nepatrí do systémuLASER.To naznačuje, že LASER môže dobre fungovať v scenároch s nulovým záberom (t.j. nepálsky-anglický), ale funguje ešte lepšie, keď má dodatočný dohľadnad jazykmi, na ktorých sa testuje.

Metóda	česk		sisk	
	1M	5M	1M	5M
Vopred vyškolený	6.06	1.49	**7.82**	**5.56**
Laserová *lokalita*	**7.37**	**3.15**	7.49	5.01

Tabuľka 4:Porovnanie výsledkov na `flores` DevTest set s použitím obmedzených a vopred vyškolených veziónov LASER.

Podanie Pre oficiálne podanie sme použili súbor ALL pre úlohu Sinhala-anglický a LASER *globálny* + *lokálny* súbor pre nepálsko-anglickú úlohu.Tiež sme predložili LASER *lokálny* ako kontrastný systém.Ako môžeme vidieť v tabuľke 3, výsledky hlavných a kontrastnýchpodaní sú veľmi blízke.V jednom prípade, kontrastný roztok (jedno LASER) model prináša lepšie výsledky ako súbor.Tieto výsledky umiestnili naše 1M príspevky 1.3 a 1,4 bodov BLEU nad bežcov pre úlohy nepálčiny – angličtiny a Sinhaly-anglického.Ako už bolo spomenuté, naše systémy dosahujú horšie výsledky pri 5M stave.Takisto sme poznamenali, že čísla v tabuľke 2 sa mierne líšia od čísel vykázaných v (Koehn a kol., 2019).Tento rozdiel pripisujeme efektu tréningu v 4 (náš) GPU vs. 1 (ich).

Metóda	česk		sisk	
	1M	5M	1M	5M
Main – Ensemble	6.8	2.8	**6.4**	4.0
To je Constr.— LASER	**6.9**	2.5	6.2	3.8
Najlepší (iný)	5.5	**3.4**	5.0	**4.4**

Tabuľka 3:Oficiálne výsledky hlavných a sekundárnych podaní na sadetestov `flores` vyhodnotených s konfiguráciou NMT.Pre porovnanie, zahrňujeme najlepšie skóre iným systémom.

5 Závery a budúca práca

V tomto dokumente popisujeme naše podanie do úlohy filtrovania nízkozdrojového paralelného korpusu WMT.Používame viacjazyčné vety z LASER na filtrovanie hlučných viet.Pozorujeme, že LASER môže dosiahnuť lepšie výsledky ako základné línieso širokým rozpätím.Zahrnutie skóre z iných techník a vytvorenie súboru prinášaďalšie zisky.Náš hlavný príspevok k zdieľanej úlohe je založený na tom najlepšom zokonfigurácie súboru a naše kontrastné podanie je založené na najlepšej konfigurácii LASER.Naše systémyplnia najlepšie 1M podmienky pre nepálsko-anglické a sinhalsko-anglické úlohy.Analyzujeme výkonnosť vopred vycvičenej verzie LASER a pozorujeme, že dokáže vykonávať filtračnúúlohu dobre aj v scenároch s nulovými zdrojmi, čo je veľmi sľubné.

V budúcnosti chceme túto techniku vyhodnotiť pre scenáre s vysokým zdrojom a sledovať, či sa rovnaké výsledky prenesú do tohto stavu.Okrem tohoplánujeme preskúmať, ako veľkosť školiacich dát (paralelné, jednojazyčné) ovplyvňuje úlohu nízkozdrojového filtrovania viet.

Referencie

Mikel Artetxe, Gorka Labaka, Eneko Agirre a Kyunghyun Cho.2018.Nekontrolovaný neurálny strojový-preklad.Na *medzinárodnej konferencii o zastúpeniach vzdelávania (ICLR).*

Mikel Artetxe a Holger Schwenk.2018a.Maržová paralelná baníctvo korpusu s viacjazyčným vložením-trestu. arXivpreprint arXiv:1811.01136.

Mikel Artetxe a Holger Schwenk.2018b.Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond.arXiv preprint arXiv:1812.10464.

Francisco Guzman, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary a Marc'Aurelio Ranzato.2019.Dva nové hodnotiace súbory údajov pre nízkozdrojový strojový preklad:Nepali-angličtina a sinhala-english.arXiv preprint arXiv:1902.01382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark a Philipp Koehn.2013.Scalable modifikovanýmodel jazyka Kneser-Ney.V *pokračovaní51. výročného zasadnutia Asociácie pre výpočtovú lingvistiku, strany 690*— 696, Sofia, Bulharsko.

Marcin Junczys-Dowmunt.2018.Dvojitá podmienená cross-entropia filtrácia hlučných paralelných korpusov.In *Proceedings of the Third Conference on Machine Translation, zväzok 2:Dokumenty o zdieľanej*úlohe, strany 901 – 908, Belgicko, Brusel.Asociácia pre počítačovúlingvistiku.

Huda Khayrallah a Philipp Koehn.2018.O vplyve rôznych

typov hluku na neurálny strojový preklad.In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, strany 74 – 83, Melbourne, Austrália.Asociácia pre počítačovú-lingvistiku.

Huda Khayrallah, Hainan Xu a Philipp Koehn.2018.Paralelné korpusové systémy JHU pre WMT 2018.In *Proceedings of the Third Conference on Machine Translation, zväzok 2:Dokumenty o zdieľanejúlohe, strany* 909 – 912, Belgicko, Brusel.Asociáciapre počítačovú lingvistiku.

Philipp Koehn, Francisco Guzman, Vishrav Chaud- Hary a Juan M. Pino.2019.Zistenia zdieľanej úlohy WMT 2019 o paralelnom korpusovom filtrovaní pre nízke zdroje.In *Proceedings of the Fourth Conference on Machine Translation, zväzok 2:Common Task Papers*, Florencia, Taliansko.Asociácia pre počítačovú lingvistiku.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar Chris Dyer, Alexandra Constantin, a Evan Herbst.2007.Mojžiš:Súbor nástrojov s otvoreným zdrojovým kódom na preklad štatistických strojov.Če
Výročné zasadnutie Asociácie pre počítačovú lingvistiku(ACL), demo zasadnutie.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield a Mikel L Forcada.2018.Výsledky zdieľanej úlohy WMT 2018 o paralelnom korpusovom filtrovaní.In *Proceedingsof the Third Conference on Machine Translation,zväzok 2:Common Task Papers*, s. 726739, Belgicko, Brusel.Asociácia pre počítačovú-lingvistiku.

Guillaume Lample, Myle Ott, Alexis Conneau, Lu- dovic Denoyer, a Marc'Aurelio Ranzato.2018.Frázy založené na & neurálny bez dozoru strojový preklad-.*Empirical Methods in Natural Language Processing (EMNLP)*, strany 5039 – 5049, Belgicko, Brusel.Asociácia pre počítačovú lingvistiku.

Fantine Mordelet a J-P Vert.2014.Bagging svm učiť sa z pozitívnych a neoznačených príkladov.*Listy o rozpoznávaní*vzorov, 37:201 – 209.

Myle Ott, Sergej Edunov, Alexej Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier a Michael Auli.2019. fairseq:Rýchly, rozšíriteľný súbor nástrojov pre sekvenčné modelovanie.V *konaní o NAACL-HLT 2019:Ukážky.*

Matt Post.2018.Výzva na zrozumiteľnosť pri oznamovaní výsledkov bleu.In *Proceedings of the Third Conference on Machine Translation (WMT), zväzok 1:Research Papers*, zväzok 1804.08771, strany 186 – 191, Belgicko, Brusel.Asociácia pre počítačovú lingvistiku.

Victor M. Sanchez-Cartagena, Marta Banon, Sergio Ortiz-Rojas a Gema Ramirez.2018.Prompsitove podanie do WMT 2018 paralelného korpusu zdieľanej úlohy.V *konaní o tretej konferencii o strojovom preklade:Dokumenty o zdieľanej*úlohe, strany 955 – 962, Belgicko, Brusel.Asociácia pre počítačovú-lingvistiku.

Holger Schwenk.2018.Filtrovanie a ťažba paralelných údajov v spoločnom viacjazyčnom priestore.V *Proceedings of the 56th Annual Meeting of the Association forComputational Linguistics (Krátke dokumenty)*,strany 228234,Austrália, Melbourne.Asociácia pre počítačovúlingvistiku.

Hainan Xu a Philipp Koehn.2017.Zipporah:Rýchly a škálovateľný systém čistenia dát pre hlučnú web—plazil sa paralelné korpusy.In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945 – 2950, Dánsko, Cophenhagen.Asociácia pre počítačovú lingvistiku.