Nízko-Zdrojové Corpus filtrování pomocí vícejazyčné věty Embeddings

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

Abstraktní

V této práci popisujeme naše podání WMT19 s nízkým zdrojovým paralelním corpus filtrovánímsdíleného úkolu.Náš hlavní přístup je založen na nástrojové sadě LASER (Language-Agnostická věta Representations), která využívá architekturu enkodér-dekodér vyškolenou na paralelnímcorpusu k získání vícejazyčných reprezentací vět.Pak použijeme reprezentace-přímo k skóre a filtrování hlučné paralelní věty-bez dalšího tréninku bodování funkce.Náš přístup kontrastujeme s jinými slibnými metodami a ukazujeme, že LASER přináší silné výsledky.Nakonec vyrobímesoubor různých bodovacích metod a získáme další zisky.Naše podání dosáhlo nejlepšího celkového výkonu jak pro Nepálsko-anglické a Sinhala-anglické 1M úkoly o rozpětí 1,3 a 1,4 BLEU, v porovnání s druhým nejlepším systémem.Naše experimenty navíc ukazují, že tato technika je slibná pro scénáře s nízkými a dokonce bez zdrojů.

1 Úvod

Dostupnost vysoce kvalitních souběžných školicích dat je rozhodující pro získání dobrého výkonu překladu, protože systémy neuronového strojového překladu (NMT) jsou proti hlučným paralelním datům méně robustní než systémy statistického strojového překladu (Khayrallah a Koehn,2018).V poslední době existuje zvýšený zájem o filtrování hlučných paralelníchkorpusů (např. Paracrawl1)ke[zvýšení]množství údajů, které lze použít k výcviku překladatelských systémů (Koehn etal., 2018).

Zatímco nejmodernější metody, které používají modely NMT, se ukázaly jako účinné v těžbě Paralelní věty (Junczys— Dowmunt, 2018) pro jazyky s vysokými zdroji, jejich účinnost nebyla testována v nízkozdrojových jazycích.Důsledky nízké dostupnosti údajů o výcviku pro metody paralelního hodnocení nejsou dosud známy.

Pro úkol filtrování s nízkými zdroji (Koehnet al., 2019)máme k dispozici velmi hlučné 40,6 milionu slov (anglicky token count) Nepálsko—anglický corpus a 59,6 milionu slov Sinhala-anglický corpus vyšplhal z webu jako součást projektu Paracrawl.Výzva spočívá v poskytovánískóre pro každý pár vět v obou hlučné paralelní sady.Skóre budou použity na subsamplevěty páry, které činí 1 milionů a 5 milionů anglických slov.Kvalita výslednýchpodskupin je určena kvalitou statistického strojovéhopřekladu (Moses, na bázi frází (Koehn etal., 2007))aneurálního strojového překladového systému fairseq(Ott et al., 2019) vyškolených natěchto údajích.Kvalitasystému strojového překladu bude měřena skóre BLEU pomocí Sacrebleu(Post,2018) nasadě testů Wikipedie pro Sinhala-anglicky a nepálsko-angličtinu ze souboru flores (Guzman etal., 2019).

V našem podání k tomuto společnému úkolu používáme vícejazyčné vložky získané z LASERu,který[1] používá architekturu enkodér-dekodér kvýcviku vícejazyčného modelu reprezentace vět pomocí relativněmálého paralelního korpusu.Naše experimenty ukazují, že navrhovaný přístup překonává ostatní existující přístupy.Kromě toho využíváme soubor vícenásobnýchbodovacích funkcí pro další zvýšení filtračního výkonu.

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

Jednáníčtvrtékonference o strojovém překladu (WMT), svazek 3:Sdílené pracovní dokumenty (Den 2) str. 261—266Florence, Itálie, 1-2. srpna 2019.

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

2 Metodika

WMT 2018 sdílený úkol pro paralelní filtrování corpusu (Koehn et al., 2018)[3] představil několik-metod pro řešení vysokozdrojové německo-anglické datové podmínky.Zatímco mnohé z těchto metod se-podařilo odfiltrovat hlučné překlady,jen málokdo bylo vyzkoušeno za podmínek s nízkými zdrojovými zdroji.V této práci řešíme problém filtrování věty s nízkými zdroji pomocí reprezentací úrovně věty a porovnáváme je s dalšími populárními metodami používanými ve vysokozdrojových podmínkách.

ModelLASER (Artetxe a Schwenk, 2018a) využívá mnohojazyčného vyjádření větyk určení podobnosti mezi zdrojem a cílovou větou.Poskytla nejmodernější výkon na těžebním úkolu Bucc corpus a byla účinná i při filtrování datWMT Paracrawl (Artetxe aSchwenk, 2018a) .Tyto úkoly se však týkaly pouze jazyků s vysokými zdroji, jmenovitě francouzštiny, němčiny, ruštiny a čínštiny.Naštěstí tato technika byla účinná i na nulovou střelu cross-lingual natural language inference v datovém souboru XNLI (Artetxe a Schwenk, 2018b), což ji činí slibným pro nízkoenergetický scénář zaměřený natento sdílený úkol.V této práci navrhujeme použít adaptaci LASER na podmínky s nízkými zdroji pro výpočet podobnosti skóre proodfiltrování hlučné věty.

Pro srovnání s LASER, také stanovíme počáteční-referenční hodnoty pomocí Bicleaner a Zipporah, dvě populární základní linie, které byly použity v projektu Paracrawl;A dvojí podmíněná crossentropie, která se ukázala jako nejmodernější pro vysokozdrojový filtrační úkol (Koehn etal.,2018).Zkoumáme výkonnost technik v podobných podmínkách předběžného zpracování,-pokud jde o filtrování jazykové identifikace a lexikální překrývání.Pozorujeme, že LASER skóre poskytují jasnou výhodu pro tento úkol.Nakonec provádíme sestavování výsledků vycházejících z-různých metod.Pozorujeme, že když jsou výsledky LASER zahrnuty do mixu, zvýšení výkonu je relativně menší.Ve zbytku této části diskutujeme nastavení pro každou z použitých metod.

[3]http://statmt.org/wmt18/paralelní-corpus-filtrování.html

2.1 Laserové vícejazyčné zastoupení

Základní myšlenkou je použít vzdálenosti mezi dvěma vícejazyčnými reprezentacemi jako pojem paralelismus mezi oběma vloženými větami (Schwenk, 2018).Abychom toho dosáhli, nejprve vycvičíme enkodér, který se naučí vytvářet vícejazyčné, pevné reprezentaci věty;a pak vypočítat vzdálenost mezi dvěma větami v učeném zabudování prostoru.Kromě toho používáme *maržové* kritérium, které používá přístup k nejbližším sousedům pro normalizaci skóre podobnosti vzhledem k tomu, že kosine podobnost není celosvětově konzistentní (Artetxea Schwenk, 2018a).

Enkodér Vícejazyčný enkodér se skládá z obousměrného LSTM a naše větase získává použitím max-poolingu nad jeho výstupem.V našem systému používáme jeden enkodér a dekodér, které jsou sdíleny všemi zúčastněnými jazyky.Za tímto účelem jsme vyškolili vícejazyčné vkládání vět jen na poskytnutých souběžných údajích (podrobnosti viz bod 3.2).

Marže Podle definice poměru *od*[2] (Artetxe aSchwenk, 2018a).Pomocí tohoto, podobnost skóre mezi dvěma větami (x, y) lze vypočítat jako

$$\frac{2k\cos(x,y)}{Sy'eNfc(x)\ ^{\cos}(^{x},^{y/}) + Sx'eNfc_{(}y)\cos\ (^{x'}, y^{)})}$$

Kde $NNK_{(}x)$ označuje k nejbližší sousedy x v jiném jazyce a obdobně pro NNK $(_{y})$.Všimněte si, že tento seznam nejbližších sousedů neobsahuje duplikáty, takže i v případě,že má uvedená věta více výskytů v corpusu, bude mít (nejvíce) jednu položku v seznamu.

Sousedství Dodatečně jsme prozkoumali dva způsoby vzorkování k nejbližších sousedů.Nejprve *globální* metoda, ve které jsme použili okolí,které se skládalo z hlučných dat spolu s čistými daty.Druhá *lokální*metoda, ve které jsme získali pouze hlučná data za použití hlučné čtvrti, nebo čisté údaje pomocí čistého sousedí. [3]

2.2 Ostatní metody simulace

Zipporah (Xu a Koehn, 2017;Khayrallah et al., 2018), který se často používá jako výchozí srovnání, používá jazykový model a slovní překladskóre, s váhami optimalizovány pro oddělení čistých a syntetických zvukových dat.V naší sestavě jsme trénovali Zipporah modely pro oba jazykové páry Sinhala-angličtina a nepálská angličtina.Použili jsme open source release6nástroje Zipporah bez úprav.Všechny složky modelu Zipporah (pravděpodobné překladové slovníkya jazykové

[2]Zkoumali jsme absolutní, *vzdálenost* a *poměr*marže kritéria, ale druhý pracoval nejlépe

[3] Tato poslední část byla provedena pouze pro školení souboru

modely) byly vyškoleny na poskytnutých čistých údajích (s výjimkou slovníků).Jazykové modely byly vyškoleny pomocí KenLM (Heafield et al., 2013) na základě čistých paralelních dat.Nepoužíváme poskytnuté monolingvální údaje podle výchozího nastavení.Protrénink váhy jsme použili sadu `vývojových` dat z květin.

Bicleaner (Sanchez-Cartagena et al., 2018) používá lexikální překlady a skóre jazykových modelů a několik mělkých rysů, jako jsou:příslušná délka, odpovídající čísla a interpunkce.Stejně jako u Zip-porah, jsme použili open source Bicleaner7toolkit[unmodified] out-of-the-box.K výcviku tohoto modelu byly použity pouze poskytnuté čisté paralelní údaje.Bicleaner používá komponentu založené na pravidlech k identifikaci hlučnějších příkladů v paralelních údajích a trénuje klasifikátora, aby se naučil, jak je oddělit od ostatních údajů o výcviku.Použití jazykových modelů je volitelné.Používali jsme pouze modely bez komponenty pro bodování jazykových modelů.8

Duální podmíněná Cross-Entropie Jedna z Nejlepšími metodami na tomto úkolu bylo dvojí podmíněné filtrování křížové entropie (Junczys—Dowmunt, 2018), které používá kombinaci dopředu a zpětných modelů pro výpočet křížové podobnosti skóre.V našich experimentech, pro každou jazykovou dvojici, jsme použili poskytnuté čisté školicí data k výcviku neurální stroje translační modely v obou směrech překladu:zdroj k cíli a cíl ke zdroji.Vzhledem k takovému překladumodelu M, jsme síla-dekódovat věty páry (x, y) zhlučné paralelní corpus a získat kříž-entropie skóre

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

[6]https://github.com/hainan-xv/zipporah

[7]https://github.com/bitextor/bicleaner

[8]Zjistilijsme, že včetně LM jako funkce vyústilo v téměř všechny větové dvojice, které obdržely skóre 0.

Předem a zpětně křížová entropie skóre, $H_F(y|x)$ a Hb ($_x$ | y), se pak průměrují sdodatečným trestem na velkém rozdílu mezi oběma skóre | Hf (y |$_x$) - Hb (xM.

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

Přední a zadní modely jsou pětivrstvé enkodéry/dekodérové transformátory vyškolené pomocí fairseq s parametry identickými jako u základního modelu `flores` [45].Modelybyly vyškoleny na čistých paralelních datech pro 100 epochy.Pro nepálsko-anglický úkol jsme také prozkoumali použití hindsko-anglických dat bez zásadních rozdílů ve výsledcích.Použili jsmesadu `flores` development k výběru modelu, který maximalizuje skóre BLEU.

2.3 Soubor

Abychom využili silné a slabé stránky různých bodovacích systémů, zkoumali jsme využití binárního klasifikátoru k vytvoření souboru.I když je triviální získat pozitivní (např. čisté údaje o tréninku), těžební negativy mohou být skličující úkol.Proto používáme kladně neoznačené (PU) učení (Mordeleta Vert, 2014), které nám umožňuje získat-klasifikátory, aniž bychom museli kurátorovat soubor explicitních pozitivních a negativů.V tomto nastavení pocházejí naše pozitivní štítky z čistých paralelních dat, zatímco neoznačená data pocházejí z hlučné sady.

Abychom toho dosáhli, použijeme balení 100 slabých, zkreslených klasifikátorů (tj. s 2:1 předpojatostí pro neoznačená data vs. pozitivní údaje).Používáme podpůrné vektorové stroje (SVM) s radiálním jádrem základů, a náhodně pod-vzorek sadu funkcí pro výcvikkaždého základního klasifikátoru, pomáhá udržet je různorodé a nízké kapacity.

Projeli jsme dvě iterace výcviku tohoto souboru.V první iteraci jsme použili původní pozitivnía neoznačená data popsaná výše.Pro druhou iteraci jsme použili naučený klasifikátor k přeznačení výcvikových dat.Prozkoumali jsme několik přeznačovacíchpřístupů (např. nastavení prahové hodnoty, která maximalizuje *skóre F1*).Nicméně jsme zjistili, že nastavení hranice třídy zachovat původní pozitivní-k-neoznačený poměr pracoval nejlépe.Také jsme zaznamenali, že výkon se po dvou iteracích zhoršil.

3 Experimentální nastavení

Experimentovalijsme s různými metodami pomocí nastavení, které úzce odráží oficiální bodování sdíleného úkolu.Všechny metody jsou vyškoleny na poskytnutých čistých souběžných údajích (viz tabulka 1).Nepoužili jsme uvedené monolingvální údaje.Provývoj jsme použili zajišťovanou sadu flores dev.Pro vyhodnocení jsme vyškolili systém strojového překladu na vybraných podskupinách (1M, 5M) hlučných paralelních školicích dat pomocí fairseq s výchozí konfigurací parametrů školení flores.Hlásíme Sacrebleu skóre na Flores DevTest set.Vybrali jsme náš hlavní systém na základě nejlepších výsledků na DevTest sadě pro podmínku 1M.

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

	si-en	ne-en	Hi-en
Věty	646k	573k	1,5 M
Anglická slova	3,7 M	3,7 M	20,7 M

Tabulka 1:Dostupné bitexty pro výcvik filtračních přístupů.

3.1 Předzpracování

Použilijsme sadu filtračních technik podobnou těm, které se používají v LASER (Artetxea Schwenk, 2018a) a přiřadili jsme skóre -1 na hlučné věty založené na nesprávném jazyce buď na zdroji nebo na cílové straně nebo s překrytím nejméně 60 % mezi zdrojem a cílovými žetony.Profiltrování[jazyka]jsme použili fastText10.-Vzhledem k tomu, že LASER vypočítává podobnost bodů pro větový pár pomocí těchto filtračních technik, experimentovali jsme s jejich přidáním k ostatním modelům, které jsme použili pro tento společný úkol.

3.2 Školení laserového enkodéru

Pro naše experimenty a oficiální podání jsme vycvičili vícejazyčný enkodér vět s použitím-povolených zdrojů v tabulce 1. Trénovali jsme jeden enkodér s využitím všech paralelních dat pro Sinhala-English, nepálsko-anglický a hindsky-anglicky.Vzhledem k tomu, že Hindi a Nepálí sdílejí stejný scénář, spojili jsme jejich sbor do jediného paralelního korpusu.Abychom zohlednili rozdíl ve velikosti souběžných školicích dat, převlekli jsme Sinhala-English a Nepálí/Hindi-anglické bitextyv poměru 5:3.To mělo za následek zhruba 3,2M výcvikové věty pro každý jazykový směr, tj. Sinhala a kombinované Nepálsko-Hindi.

[109]https://fasttext.cc/docs/en/identifikace jazyka.html

Modely byly vyškoleny stejným nastavením jako veřejný kódátor LASER, který zahrnuje normalizaci-textů a tokenizace pomocí nástrojů Mojžíše (spadá zpět do anglického módu).Nejprve jsme se naučili společný 50k BPE slovní zásobu o koncatenated training data pomocí fastBPE.[6]Enkodér vidí na vstupu věty Sinhala, Nepálí, hindština a angličtina, aniž by měl žádné informace o aktuálním jazyce.Tento vstup je vždy přeloženo angličtiny.[7]Experimentovalijsme s různými-technikami pro přidání hluku do anglických vstupních vět, podobné tomu, co se používá v nekontrolovaném neuronovém překladu, např. (Artetxeet al., 2018; Lample et al., 2018), to však-výsledky nezlepšilo.

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

[12To]znamená, že musíme trénovat anglický autoenkodér.-Nezdálo se, že by to bolelo, protože stejný enkodér také ovládá tři další jazyky.

Enkodér je pětivrstvý BLSTM s 512 dimenzionálními vrstvami.Dekodér LSTM má jednu skrytou vrstvu velikosti 2048, vyškolenou s Adamem optimalizátorem.Pro vývoj počítámechybu podobnosti na koncatenaci sad flores dev pro Sinhala-English a nepálsko-angličtinu.Naše modely byly vyškoleny na sedm epochy po dobu cca 2,5 hodiny na 8 Nvidia GPU.

4 Výsledky

Z výsledků v tabulce 2 sledujeme několik trendů:(i) skóre pro podmínku 5M jsou obecněnižší než u podmínky 1 M.Tato podmínkase zdá být zhoršena použitím jazykové id a překrývání filtrace.II) LASER vykazuje trvale dobrý výkon.*Místní* čtvrť funguje lépe než ta *globální*.V tomto nastavení je LASER v průměru 0,71 BLEU nad nejlepším systémem, který není LASER.Tyto mezery jsou vyšší pro podmínku 1M (0.94 BLEU).(*iii*) Nejlepší konfigurace souboru poskytuje malá vylepšení nad nejlepší konfigurace LASER.Pro Sinhala-English nejlepší konfiguracezahrnuje všechny ostatní metody bodování (ALL).Pro nepálsko-anglickou nejlepší konfiguraci je soubor LASER skóre.(iv) Dual cross entropieukazuje smíšené výsledky.Pro Sinhala-Angličtinu funguje pouze po zapnutém filtrování jazyka, což odpovídá předchozím pozorováním-(Junczys —Dowmunt, 2018).Pro nepálskou angličtinu poskytuje skóre hluboko pod ostatními bodovacími metodami.Všimněte si, že jsme neprovedli průzkum architektury.

Metoda	ne, ne.— Ne,		si-en	
	1 M	5 M	1 M	5 M
Zipporah				
základna	5.03	2.09	4.86	4.53
+ VÍKO	5.30	1.53	5.53	3.16
+ Překrývání	5.35	1.34	5.18	3.14
Duální X-Ent.				
základna	2.83	1.88	0.33	4.63^{+}
+ VÍKO	2.19	0.82	6.42	3.68
+ Překrývání	2.23	0.91	6.65	4.31
Bicleaner				
základna	5.91	2.54^{+}	6.20	4.25
+ VÍKO	5.88	2.09	6.36	3.95
+ Překrývání	6.12 +	2.14	6.66^{+}	3.26
LASER				
místní	*7.37* *	**3.15**	*7.49* *	5.01
globální	6.98	*2.98* *	7.27	4.76
Soubor				
VŠECHNY	6.17	2.53	**7.64**	**5.12**
— *Laserový glob.*— +	**7.49**	2.76	7.27	*5.08* *

Tabulka 2:Sacrebleu skóruje na Flores DevTest.Tučnězvýrazníme nejlepší skóre pro každou podmínku.*Kurzívou* zvýrazníme* běžce.Také signalizujeme nejlepší non-LASER metoda s +.

4.1 Diskuse

Přirozenou otázkou je, jak by prospěla metoda LASER, kdyby měla přístup k dodatečným-

údajům.Abychom to prozkoumali, použili jsme open-source toolkit LASER, který poskytuje vyškolený enkodérpokrývající 93 jazyků, ale nezahrnuje nepálštinu.V tabulce 4 poznamenáváme, že předškolený modelLASER překonává místní *model LASERo* 0.4 BLEU.Pro nepálsko-anglickou situaci se situace obrátí:Laser *lokální* poskytuje mnohem lepší výsledky.Výsledky předškoleného LASERujsou však jen mírně horší než výsledky Bicleaner (6.12), což je nejlepší metoda, která není LASER.To naznačuje, že LASER může dobře fungovat v nulových scénářích (tj. nepálsko-anglicky), ale funguje ještě lépe, když má další dohlednad jazyky, na kterých se testuje.

Metoda	**ne, ne.**	**— Ne, ne**	**— Ne, ne ne**	**— Ne, ne ne**
	1 M	**5 M**	**1 M**	**5 M**
Předškolený LASER	6.06	1.49	**7.82**	**5.56**
*Místní*laser	**7.37**	**3.15**	7.49	5.01

Tabulka 4:Porovnání výsledků nasadě `flores` DevTest s využitím omezujících a předškolených vesion LASER.

Podání Pro oficiální podání jsme použili soubor ALL pro úkol Sinhala-English a LASER Global + *lokální* soubor pro nepálsko-anglický úkol.Také jsme předložili LASER *lokální* jako kontrastní systém.Jak je vidět v tabulce 3, výsledky hlavních a kontrastníchpodání jsou velmi blízké.V jednom případě, kontrastní řešení (jednorázový LASER) model přináší lepší výsledky než soubor.Tyto výsledky umístily naše 1M příspěvky 1.3 a 1.4 BLEU bodů nad běžci pro Nepálsko-anglicky úkoly, resp. Sinhala-English.Jak již bylo uvedeno, naše systémy jsou horší ve stavu 5M.Zaznamenali jsme také, že čísla v tabulce 2 se mírně liší od čísel uvedených v (Koehn et al., 2019).Tento rozdíl připisujeme efektu tréninku v 4 (naše) GPU vs. 1 (jejich).

Metoda	**ne, ne.**	**— Ne, ne**	**— Ne, ne ne**	**— Ne, ne ne**
	1 M	**5 M**	**1 M**	**5 M**
Hlavní - Soubor	6.8	2.8	**6.4**	4.0
— To je Constr.—	**6.9**	2.5	6.2	3.8
Nejlepší (jiné)	5.5	**3.4**	5.0	**4.4**

Tabulka 3:Oficiální výsledky hlavních a sekundárních podání na sadě `floresových` testů vyhodnocených s konfigurací NMT.Pro srovnání, zahrneme nejlepší skóre podle jiného systému.

5 Závěry a budoucí práce

V této práci popisujeme naše podání WMT s nízkými zdroji paralelního filtrování corpusu.Používáme vícejazyčné vložky z LASERu k filtrování hlučné věty.Pozorujeme, že LASER může získat lepší výsledky než základní linies širokým rozpětím.Začlenění skóre z jiných technik a vytvoření souboru poskytujedalší zisky.Naše hlavní odevzdání sdílenému úkolu je založeno na nejlepší-konfiguraci souboru a naše kontrastní podání je založeno na nejlepší konfiguraci LASER.Naše systémyfungují nejlépe na 1M stavu pro nepálsko-anglické a Sinhala-anglické úkoly.Analyzujeme výkon předškolené verze LASER a pozorujeme, že filtračníúkol dokáže dobře plnit i při nulových zdrojových scénářích, což je velmi slibné.

V budoucnu chceme tuto techniku vyhodnotit pro scénáře s vysokými zdroji a sledovat, zda stejné výsledky přenášejí do tohoto stavu.Kromě toho-plánujeme zjistit, jak velikost školicích dat (paralelní, monolingvální) dopad na filtrování vět s nízkými zdroji.

Odkazy

Mikel Artetxe, Gorka Labaka, Eneko Agirre a Kyunghyun Cho.2018 LET.Bez dozorupřeklad neurálního stroje.Na *mezinárodní konferenci o reprezentacích vzdělávání (ICLR).*

Mikel Artetxe a Holger Schwenk.— 2018A. - Dobře.Margin— based Parallel Corpus Mining with MultilingualSentence Embeddings. arXivpreprint arXiv:1811.01136.

Mikel Artetxe a Holger Schwenk.— 2018b. b.Masivně vícejazyčné věty 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 LET.Dva nové sady vyhodnocovacích dat pro strojový překlad s nízkými zdroji:Nepálsko-anglický a sinhala-english.arXiv preprint arXiv:1902.01382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark a Philipp Koehn.2013 LET.Škálovatelný odhad-modelového modelu Kneser-Ney.V *řízení na 51. výročním zasedání Asociace provýpočetní lingvistiku, s. 690*—696, Sofia, Bulharsko.

Marcin Junczys-Dowmunt.2018 LET.Dvojí podmíněná filtrace zkříženého entropie hlučné paralelní korpusy.V *řízení o třetí konferenci o strojovém překladu, Svazek 2:Sdílené pracovní dokumenty*, s. 901-908, Belgie, Brusel.Asociace pro výpočetnílingvistiku.

Huda Khayrallah a Philipp Koehn.2018 LET.Na dopad různých typů hluku na překlad neurálního stroje.V *Proceedings of the 2. Workshop on Neural Machine Translation and Generation*, pages 74-83, Melbourne, Australia.Asociace pro výpočetnílingvistiku.

Huda Khayrallah, Hainan Xu a Philipp Koehn.2018 LET.Paralelní corpus filtrační systémyJHU pro WMT

2018.V *řízení o třetí konferenci o strojovém překladu, Svazek 2:Sdílené pracovní dokumenty*, s. 909-912, Belgie, Brusel.Asociacepro výpočetní lingvistiku.

Philipp Koehn, Francisco Guzman, Vishrav Chaud- Hary a Juan M. Pino.2019 LET.Výsledky wmt 2019 sdílený úkol na paralelní corpus filtrování pro nízkozdrojové podmínky.V *řízení ze čtvrté konference o strojovém překladu, Svazek 2:Sdílené pracovní dokumenty*, Florencie, Itálie.Asociace pro výpočetní lingvistiku.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondřej Bojar Chris Dyer, Alexandra Constantin a Evan Herbst.2007 LET.— Ne,ne, ne.Open source toolkit pro statistické strojové překlady.— Ne, ne. *Výroční zasedání Asociace pro výpočetnílingvistiku (ACL), demo session.*

Philipp Koehn, Huda Khayrallah, Kenneth Heafield a Mikel L Forcada.2018 LET.Výsledky sdíleného úkolu WMT 2018 o paralelním filtrování corpusu.V *řízenío třetí konferenci o strojovém překladu,Svazek 2:Sdílené pracovní dokumenty*, strany 726739, Belgie, Brusel.Asociace pro výpočetnílingvistiku.

Guillaume Lample, Myle Ott, Alexis Conneau, Lu... Dovic Denoyer a Marc'Aurelio Ranzato.2018 LET.Fráze-založené & neurální bez dozoru strojový překlad.V *Empirical Methods in Natural Language Processing (EMNLP)*, str. 5039-5049, Belgie, Brusel.Asociace pro výpočetní lingvistiku.

Fantine Mordeletová a J-P Vert.2014 LET.Balení svm učit se z pozitivních a neoznačených příkladů.*Vzor rozpoznávání dopisů,* 37:201-209.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier a Michael Auli.2019. fairseq:Rychlý, rozšiřitelný sad nástrojů pro sekvenční modelování.V *řízení NAACL-HLT 2019:— To jsou ukázky.*

— Matt Post. - Matt Post.2018 LET.Výzva k jasnosti ve vykazování skóre bleu.V *řízení třetí konference o strojovém překladu (WMT), svazek 1:Výzkumné dokumenty*, svazek 1804.08771, s. 186-191, Belgie, Brusel.Asociace pro výpočetní lingvistiku.

Victor M Sanchez-Cartagena, Marta Banon, Sergio Ortiz-Rojas a Gema Ramirez.2018 LET.Soubor Prompsit na WMT 2018 paralelní corpus filtrování sdílený úkol.V *řízení o třetí konferenci o strojovém překladu:Sdílené pracovní dokumenty*, s. 955-962, Belgie, Brusel.Asociace pro výpočetnílingvistiku.

— Holger Schwenk.2018 LET.Filtrování a získávání paralelních dat ve společném vícejazyčném prostoru.V *řízení na 56. výročním zasedání Asociace pro výpočetnílingvistiku (Short Papers)*, s.228234,- Austrálie, Melbourne.Asociace pro výpočetní-lingvistiku.

Hainan Xu a Philipp Koehn.2017 LET.— Ne,ne, ne, ne.Rychlý a škálovatelný systém čištění dat pro hlučné pavučiny.V *řízení z roku 2017 konference o Empirical Methods in Natural Language Processing*, pages 2945-2950, Dánsko, Cophenhagen.Asociace pro výpočetní lingvistiku.