Niskozasobowe filtrowanie Corpus za pomocą wielojęzycznych obudowań

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

Streszczenie

W niniejszym artykule opisujemy nasze zgłoszenie do równoległegozadania WMT19 filtrującego równoległe korpusy o niskim zasobach.Nasze główne podejście opiera się na zestawie narzędzi LASER (reprezentacje zdania językowo-agnostycznego), który wykorzystuje architekturę koder-dekoder wyszkoloną na-korpusie równoległym do uzyskania wielojęzycznych reprezentacji zdań.Następnie wykorzystujemy reprezentacjebezpośrednio do zdobywania punktów i filtrowania hałaśliwych równoległych zdańbez dodatkowego szkolenia funkcji punktowania.Kontrastujemy nasze podejście do innych obiecujących metod i pokazujemy, że LASER daje dobre wyniki.Wreszcie, produkujemyzespół różnych metod punktowania i uzyskujemy dodatkowe zyski.Nasza propozycja osiągnęła najlepszą ogólną wydajność zarówno dla nepalsko-angielsko-angielskich zadań 1M, jak i Sinhala-English o marży odpowiednio 1,3 i 1,4 BLEU, w porównaniu z drugimi najlepszymi systemami.Co więcej, nasze eksperymenty pokazują, że ta technika jest obiecująca dla scenariuszy niskich, a nawet bez zasobów.

1 Wprowadzenie

Dostępność wysokiej jakości równoległych danych treningowych ma kluczowe znaczenie dla uzyskania dobrej wydajności tłumaczenia, ponieważ systemy tłumaczenia maszynowego (NMT) są mniej wytrzymałe wobec głośnych danych równoległych niż statystyczne systemy tłumaczenia maszynowego (Khayrallah iKoehn, 2018).Ostatnio wzrasta zainteresowanie filtrowaniem głośnych równoległychciał (takich jak Paracrawl[1]) wcelu[zwiększenia]ilości danych, które mogą być wykorzystane do szkolenia systemów tłumaczeniowych (Koehn etal., 2018).

Podczas gdy najnowocześniejsze metody wykorzystujące modele NMT okazały się skuteczne w górnictwie

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

Obradyczwartej konferencji wsprawietłumaczenia maszynowego (WMT), tom 3:Wspólne dokumenty zadaniowe (Dzień 2) strony 261—
266Florence, Włochy, 1-2 sierpnia 2019.

Zdania równoległe (Junczys-Dowmunt, 2018) dla języków o wysokich zasobach, ich skuteczność nie została przetestowana w językach o niskich zasobach.Implikacje niskiej dostępności danych szkoleniowych dla metod równoległych nie są jeszcze znane.

Za zadanie niskozasobowego filtrowania (Koehn et al., 2019)otrzymujemy bardzo hałaśliwy 40,6 miliona słów (angielski token count) nepalski-angielski korpus i 59,6 miliona słów Sinhala-angielski korpus wyczołgał się z sieci w ramach projektu Paracrawl.Wyzwanie polega na zapewnieniupunktów dla każdej pary zdań w obu hałaśliwych zestawach równoległych.Wyniki zostaną wykorzystane do podpróbki parzdań, które wynoszą 1 milionów i 5 milionów angielskich słów.Jakość powstającychpodzbiorów zależy od jakości statystycznegotłumaczenia maszynowego (Moses, fraza-based (Koehn etal., 2007))ineural machine translation system fairseq(Ott et al.,2019) przeszkolonych natych danych.Jakośćsystemu tłumaczeń maszynowych będzie mierzona przez wynik BLEU za pomocą Sacrebleu(Post,2018) naprowadzonym przez Wikipedię zestawie tłumaczeń dla Sinhala-Angielski i Nepalski-Angielski z zestawu flores (Guzman etal., 2019).

W złożeniu wniosku o to wspólne zadanie wykorzystujemy wielojęzyczne okładziny zdań uzyskane z LASER,które[1] wykorzystuje architekturę koder-dekoder doszkolenia wielojęzycznego modelu reprezentacji zdań przy użyciu stosunkowo-niewielkiego korpusu równoległego.Nasze eksperymenty pokazują, że proponowane podejście przewyższa inne istniejące podejścia.Ponadto wykorzystujemy zespół wielu funkcjipunktowania, aby jeszcze bardziej zwiększyć wydajność filtrowania.

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

2 Metodyka

Wspólne zadanieWMT 2018 dla równoległego filtrowania korpusu (Koehn et al., 2018)[3] wprowadziło kilkametod, aby sprostać wysokim zasobom niemiecko-angielskim warunkom danych.Choć wiele z tych metodudało się odfiltrować hałaśliwe tłumaczenia, niewiele z nich-zostało wypróbowanych w warunkach niskich zasobów.W tym artykule zajmujemy się problemem filtrowania zdania o niskich zasobach przy użyciu reprezentacji poziomu zdania i porównywania ich z innymi popularnymi metodami stosowanymi w warunkach o wysokich zasobach.

Model LASER (Artetxe i Schwenk, 2018a) wykorzystuje wielojęzyczne reprezentacje zdań,aby ocenić podobieństwo między źródłem i docelowym zdaniem.Zapewniał on najnowocześniejsze wyniki w zakresie zadania wydobywczego Bucc corpus, a także skutecznie filtrował daneWMT Paracrawl (Artetxe iSchwenk, 2018a) .Do zadań tych zaliczano jednak wyłącznie języki o wysokich zasobach, tj. francuski, niemiecki, rosyjski i chiński.Na szczęście technika ta okazała się również skuteczna przy zerowym, wielojęzycznym wniosku językowym w zbiorze danych XNLI (Artetxei Schwenk, 2018b), co czyni ją obiecującą dla scenariusza o niskim poziomie zasobów skupionychw tym wspólnym zadaniu.W tym dokumencie proponujemy zastosowanie dostosowania LASER do warunków o niskich zasobach, aby obliczyć wynik podobieństwa doodfiltrowania hałaśliwych zdań.
Dla porównania z LASER, ustalamy również początkowepoziomy odniesienia za pomocą Bicleaner i Zipporah, dwóch popularnych punktów odniesienia, które zostały wykorzystane w projekcie Paracrawl;Oraz podwójna krzyżowa warunkowa, która okazała się najnowocześniejsza w przypadku wysokozasobowego zadania filtrowania korpusu (Koehnet al.,2018).Badamy wydajność technik w podobnych warunkach wstępnego przetwarzania w-zakresie filtrowania identyfikacji językowej i nakładania się na siebie leksykalnych.Zauważamy, że wyniki LASER zapewniają wyraźną przewagę w tym zadaniu.Wreszcie, wykonujemy zestawianie partytur pochodzących zróżnych metod.Obserwujemy, że gdy wyniki LASER są zawarte w miksie, wzrost wydajności jest relatywnie niewielki.W dalszej części niniejszej sekcji omawiamy ustawienia dla każdej z zastosowanych metod.

[3] http://statmt.org/wmt18/równoległy-corpus-filtering.html

2.1 Wielojęzyczne reprezentacje laserowe

Podstawową ideą jest wykorzystanie odległości między dwoma wielojęzycznymi przedstawieniami jako pojęcia równoległości pomiędzy dwoma wbudowanymi zdaniami (Schwenk, 2018).Aby to zrobić, najpierw szkolimy koder, który uczy się produkować wielojęzyczne, stałe wyrażenie;a następnie obliczyć odległość między dwoma zdaniami w uczonej przestrzeni osadzania.Ponadto stosujemy kryterium *marginesu*, które wykorzystuje podejście k najbliższych sąsiadów do normalizacji wyników podobieństwa, biorąc pod uwagę, że podobieństwo cosinusowe nie jest globalnie spójne (Artetxe i Schwenk, 2018a).

Enkoder Wielojęzyczny enkoder składa się z dwukierunkowego LSTM, a nasze zdania osadzania-są uzyskiwane poprzez zastosowanie max-pooling nad jego wyjściem.Używamy jednego kodera i dekodera w naszym systemie, które są współdzielone przez wszystkie języki zaangażowane.W tym celu przeszkoliliśmy wielojęzyczne zdania wyłącznie na danych równoległych (szczegóły patrz punkt 3.2).

Margines Stosujemy definicję*stosunku z* [2] (Artetxe iSchwenk, 2018a).Używając tego, wynik podobieństwa pomiędzy dwoma zdaniami (x, y) można obliczyć jako

$$\frac{2k \cos(x,y)}{Sy'eNNfc(x)\ ^{\cos}(^{x,\,y/}) + Sx_{'eNNfc(}y)\ ^{\cos(}x^{'},\ y^{)}}$$

Gdzie $NNK_{(}$ x) oznacza k najbliższych sąsiadów x w innym języku i analogicznie dla NNK ($_{y}$).Zauważ, że ta lista najbliższych sąsiadów nie zawiera duplikatów, więc nawet jeśli dane zdaniema wiele wystąpień w korpusie, to miałoby (co najwyżej) jedną pozycję na liście.

Sąsiedztwo Dodatkowo zbadaliśmy dwa sposoby pobierania próbek k najbliższych sąsiadów.Najpierw *globalna* metoda, w której wykorzystaliśmy okolicę-składającą się z głośnych danych wraz z czystymi danymi.Po drugie, metoda *lokalna*, w której zdobyliśmy tylko hałaśliwe dane używając hałaśliwej okolicy, lub czyste dane używając czystego sąsiada- borhood. [3]

2.2 Inne metody podobieństwa

[2] Zbadaliśmy kryteria marginesu *bezwzględnego, odległości* i *stosunku*, ale te ostatnie sprawdziły się najlepiej.

[3] Ta ostatnia część została zrobiona tylko do szkolenia zespołu

Zipporah (Xu i Koehn, 2017;Khayrallah et al., 2018), który jest często używany jako porównanie bazowe, wykorzystuje model językowy iwyniki tłumaczeń słownych, z wagami zoptymalizowanymi w celu oddzielenia czystych i syntetycznych danych dotyczących hałasu.W naszej konfiguracji szkoliliśmy modele Zipporah dla obu par językowych Sinhala-English i Nepali-Angielski.Użyliśmy open source release6 narzędziaZipporah bez modyfikacji.Wszystkie komponenty modelu Zipporah (prawdopodobne słowniki tłumaczeniowei modele językowe) zostały przeszkolone na dostarczonych czystych danych (z wyłączeniem słowników).Modele językowe zostały przeszkolone za pomocą KenLM (Heafield et al., 2013) na czystych równoległych danych.Nie używamy dostarczonych danych jednojęzycznych, zgodnie z ustawieniami domyślnymi.Dotreningu siłowego wykorzystaliśmy zestaw rozwojowy z zestawu danych Flores.

Bicleaner (Sanchez-Cartagena et al., 2018) używa tłumaczeń leksykalnych i wyników modelu językowego oraz kilku płytkich cech, takich jak:odpowiednia długość, pasujące numery i interpunkcja.Tak jak w przypadku Zip-porah, użyliśmy otwartegozestawunarzędzi Bicleaner7[niezmodyfikowanego] poza skrzynką.Do szkolenia tego modelu wykorzystano jedynie czyste dane równoległe.Bicleaner wykorzystuje element oparty na zasadach do identyfikacji bardziej hałaśliwych przykładów w danych równoległych i szkoli klasyfikującego, aby dowiedzieć się, jak oddzielić je od reszty danych szkoleniowych.Korzystanie z funkcji modelu językowego jest opcjonalne.Używaliśmy tylko modeli bez komponentu oceniającego model językowy.8

Podwójna warunkowa krzyżowa Entropia Jedna z Najlepszymi metodami w tym zadaniu było podwójne warunkowe filtrowanie krzyżowo-entropijne (Junczys—Dowmunt, 2018), które wykorzystuje kombinację modeli do przodu i do tyłu, aby obliczyć wynik podobieństwa między językami.W naszych eksperymentach, dla każdej pary językowej, użyliśmy dostarczonych czystych danych treningowych do szkolenia neuronowych maszyn tłumaczeń modeli w obu kierunkach tłumaczenia:źródło-do-cel i cel-do-źródła.Biorąc pod uwagę takimodel tłumaczenia M, zmuszamy do dekodowania par zdań (x, y) zgłośnego równoległego korpusu i otrzymujemy wynik krzyżowy

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

[6]https://github.com/hainan-xv/zipporah
[7]https://github.com/bitextor/bicleaner
[8]Stwierdziliśmy, że włączenie LM jako funkcji spowodowało, że prawie wszystkie pary zdań otrzymały wynik 0.

Do przodu i do tyłu wyniki krzyżowej entropii, odpowiednioHF(y|x) i HB(x | y) są uśredniane z-dodatkową karą za dużą różnicę między dwoma punktami | Hf (y |x) - Hb (xM.

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

Modele do przodu i do tyłu są pięciowarstwowe transformatory enkoder/dekoder przeszkoloneprzy użyciu fairseqo parametrach identycznych z tymi stosowanymi w podstawowym modeluflores [45].Modelezostały przeszkolone na czystych danych równoległych dla 100 epok.W przypadku nepalsko-angielskiego zadania zbadaliśmy również wykorzystanie danych hindi-angielskich bez znaczących różnic w wynikach.Użyliśmyzestawu do tworzenia flores, aby wybrać model, który maksymalizuje wyniki BLEU.

2.3 Zespół

Aby wykorzystać mocne i słabe strony różnych systemów punktowania, zbadaliśmy użycie klasyfikatora binarnego do zbudowania zespołu.Chociaż uzyskanie pozytywnych wyników jest trywialne (np. czystych danych treningowych), negatywy górnicze mogą być zniechęcającym zadaniem.Stąd korzystamy z nieoznakowanego pozytywnego (PU) uczenia (Mordeleti Vert, 2014), co pozwala nam na uzyskanieklasyfikatorów bez konieczności kuratorowania zbioru danych wyraźnie pozytywnych i negatywnych.W tym ustawieniu nasze pozytywne etykiety pochodzą z czystych równoległych danych, podczas gdy nieoznakowane dane pochodzą z zestawu hałaśliwych.

Aby to osiągnąć, stosujemy pakowanie 100 słabych, stronniczych klasyfikatorów (tj. 2:1 stronniczość dla nieoznakowanych danych vs. pozytywnych danych na etykietach).Używamy wektorów pomocniczych (SVM) z jądrem radialnym i losowo podpróbujemy zestaw funkcji do szkolenia-każdego klasyfikatora bazowego, pomagając utrzymać je zróżnicowane i niskiej pojemności.

Przeprowadziliśmy dwie iteracje szkolenia tego zespołu.W pierwszej iteracji wykorzystaliśmy oryginalne pozytywnei nieoznakowane dane opisane powyżej.W drugiej iteracji użyliśmy uczonego

[4]https://github.com/facebookresearch/Flores#pociąg-a-podstawowy-transformator-model

klasyfikatora do ponownego oznaczania danych treningowych.Zbadaliśmy kilkametod ponownego znakowania (np. ustalenie progu maksymalizującego *wynik F1*).Okazało się jednak, że ustalenie granicy klasy w celu zachowania początkowego stosunku dodatniego do nieoznakowanego zadziałało najlepiej.Zaobserwowaliśmy również, że wydajność pogorszyła się po dwóch iteracjach.

3 Konfiguracja eksperymentalna

Eksperymentowaliśmy z różnymi metodami, używając konfiguracji, która dokładnie odzwierciedla oficjalną punktację wspólnego zadania.Wszystkie metody są przeszkolone w oparciu o dostarczone czyste równoległe dane (patrz tabela 1).Nie korzystaliśmy z danych jednojęzycznych.Docelów rozwojowych używaliśmy zestawu flores dev.Do oceny przeszkoliliśmy systemy tłumaczeń maszynowych na wybranych podzespołach (1M, 5M) hałaśliwych danych treningowych z wykorzystaniem fairseq z domyślną konfiguracją parametrów treningowych flores.Zgłaszamy wyniki Sacrebleu na zestawie flores DevTest.Wybraliśmy nasz główny system na podstawie najlepszych wyników na zestawie DevTest dla stanu 1M.

	SI-en	ne-en	Hi-en
Zdania	646k	573k	1.5M
Angielskie	3.7M	3.7M	20.7M

Tabela 1:Dostępne bitexty do szkolenia podejścia filtrowania.

3.1 Przetwarzanie wstępne

Zastosowaliśmy zestaw technik filtrowania podobnych do stosowanych w LASER (Artetxei Schwenk, 2018a) i przypisaliśmy wynik —1 do hałaśliwych zdań opartych na błędnym języku albo po stronie źródła, albo po stronie docelowej, albo o nakładaniu się co najmniej 60 % pomiędzy źródłem a tokenami docelowymi.Używaliśmy fastText10dofiltrowania id[języka].Ponieważ LASER oblicza punkty podobieństwa dla pary zdań przy użyciu tych technik filtrowania, eksperymentowaliśmy dodając je do innych modeli, których użyliśmy do tego wspólnego zadania.

3.2 Szkolenie laserowe Encoder

Do naszych eksperymentów i oficjalnego zgłoszenia przeszkoliliśmy wielojęzyczny koder zdania używającdozwolonych zasobów w Tabeli 1. Przeszkoliliśmy jeden koder wykorzystując wszystkie równoległe dane dla Sinhala-English, Nepali-Angielski i Hindi-Angielski.Ponieważ hindi i Nepalczycy mają ten sam scenariusz, połączyliśmy ich korpus w jeden korpus równoległy.Aby uwzględnić różnice w wielkości równoległych danych treningowych, przepróbowaliśmy bitexty Sinhala-Angielski i nepalski/hindi-angielskiw stosunku 5:3.Zaowocowało to w przybliżeniu 3,2 mln zdań treningowych dla każdego kierunku językowego, tj. Sinhala i połączyło nepalsko-hindi.

[109]https://fasttext.cc/docs/en/language-identification.html

Modele zostały przeszkolone przy użyciu tego samego ustawienia co publiczny koder LASER, który polega na normalizacjitekstów i tokenizacji za pomocą narzędzi Mojżesza (cofając się do trybu angielskiego).Najpierw uczymy się wspólnego słownictwa 50k BPE na sprzężonych danych treningowych za pomocą fastBPE.[6]Koder widzi zdania syngalejskie, nepalskie, hindi i angielskie przy wejściu, nie mając żadnych informacji na temat aktualnego języka.To wejście jest zawsze tłumaczonena język angielski.[7]Eksperymentowaliśmy z różnymi-technikami dodawania szumu do angielskich zdań wejściowych, podobnych do tego, co jest używane w nienadzorowanych maszynowych tłumaczeniach neuronowych, np. (Artetxe et al., 2018; Lample et al., 2018), ale nie poprawiło towyników.

Koder to pięciowarstwowy BLSTM z 512 wymiarowymi warstwami.Dekoder LSTM posiada jedną ukrytą warstwę rozmiaru 2048, przeszkoloną z optymalizatorem Adama.Dla rozwoju, obliczamy-błąd podobieństwa przy konkatacji zestawów flores dev dla Sinhala-Angielski i nepalsko-angielski.Nasze modele zostały przeszkolone przez siedem epok na około 2,5 godziny na 8 procesorach Nvidia.

4 Wyniki

Na podstawie wyników w Tabeli 2 obserwujemy kilka trendów:I) wyniki dla stanu 5M są generalnie-niższe niż dla warunku 1 M.Warunek tenwydaje się być zaostrzony przez zastosowanie identyfikatora języka i nakładanie się filtrowania.(ii) LASER wykazuje konsekwentnie dobre wyniki.*Lokalna* okolica działa lepiej niż globalna .W tym ustawieniu, LASER jest średnio 0,71 BLEU ponad najlepszym systemem nielaser.Luki te są wyższe dla stanu 1M (0,94 BLEU).(*iii*) Najlepsza konfiguracja zespołu zapewnia niewielkie ulepszenia w stosunku do najlepszej konfiguracji LASER.Dla Sinhala-English najlepszą konfiguracjąjest każda inna metoda punktowania (ALL).Dla nepalsko-angielskiego najlepszą konfiguracją jest zespół partytur

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

[12]Oznacza to, że musimy trenować angielski autoencoder.To nie bolało, ponieważ ten sam koder obsługuje również trzy inne języki

LASER.IV) Podwójna entropia krzyżowawykazuje mieszane wyniki.W przypadku języka Sinhala-angielskiego, działa on tylko po włączeniu id języka, co jest zgodne z wcześniejszymi obserwacjami(Junczys—Dowmunt, 2018).Dla nepalsko-angielskiego, zapewnia wyniki znacznie poniżej pozostałych metod punktowania.Zauważ, że nie przeprowadziliśmy eksploracji architektury.

Metoda	ne po		SI-en	
	1M	5M	1M	5M
Zipporah				
baza	5.03	2.09	4.86	4.53
+ POKRYWA	5.30	1.53	5.53	3.16
+ nakładanie się na	5.35	1.34	5.18	3.14
Podwójny X-Ent.				
baza	2.83	1.88	0.33	4.63^{+}
+ POKRYWA	2.19	0.82	6.42	3.68
+ nakładanie się na	2.23	0.91	6.65	4.31
Bicleaner				
baza	5.91	2.54^{+}	6.20	4.25
+ POKRYWA	5.88	2.09	6.36	3.95
+ nakładanie się na	6.12+	2.14	6.66^{+}	3.26
LASER				
lokalne	*7.37**	**3.15**	*7.49**	5.01
globalny	6.98	*2.98**	7.27	4.76
Zespół				
WSZYSTKIE	6.17	2.53	**7.64**	**5.12**
Glob *laserowy.*+ *loc.*	**7.49**	2.76	7.27	*5.08**

Tabela 2:Sacrebleu zdobywa punkty na zestawie `flores` DevTest.`Pogrubione`, podkreślamy najlepsze wyniki dla każdego warunku.*Kursywą** *podświetlamy* zwycięzcę.Sygnalizujemy również najlepszą metodę non-LASER za pomocą +.

Zgłoszenie Do oficjalnego zgłoszenia wykorzystaliśmy zespół*ALL*do zadania Sinhala-angielskiego oraz zespół LASER *globalny* + *lokalny* do zadania nepalsko-angielskiego.Złożyliśmy również system LASER *local* jako system kontrastowy.Jak widać w Tabeli 3, wyniki z głównych i kontrastowychwypowiedzi są bardzo bliskie.W jednym przypadku rozwiązanie kontrastowe (pojedynczy model LASER) daje lepsze rezultaty niż zespół.Wyniki te umieściły nasze zgłoszenia 1M 1.3 i 1.4 punktów BLEU powyżej up runnerów do zadań nepalsko-angielsko-angielskich, odpowiednio.Jak już wspomniano, nasze systemy działają gorzej w warunkach 5M.Zauważyliśmy również, że liczby w Tabeli 2 różnią się nieznacznie od liczb podanych w (Koehn et al., 2019).Tę różnicę przypisujemy efektowi treningu w 4 (naszych) GPU vs. 1 (ich).

4.1 Dyskusja

Jednym z naturalnych pytań do zbadania jest to, w jaki sposób metoda LASER byłaby korzystna, gdyby miała dostęp do dodatkowychdanych.Aby to zbadać, użyliśmy otwartego zestawu narzędzi LASER, który zapewnia wyszkolony koderobejmujący 93 języków, ale nie obejmuje nepalskiego.W Tabeli 4 zauważamy, że wstępnie przeszkolony model-LASER przewyższa *model lokalny LASERo* 0.4 BLEU.W przypadku nepalsko-angielskiego sytuacja się odwróciła:Laser *lokalny* zapewnia znacznie lepsze wyniki.Jednak wyniki wstępnie przeszkolonegoLASER są tylko nieznacznie gorsze od wyników stosowania produktu Bicleaner (6.12), który jest najlepszą metodą non-LASER.Sugeruje to, że LASER może dobrze funkcjonować w scenariuszach zero-shot (tj. nepalsko-angielskich), ale działa jeszcze lepiej, gdy ma dodatkowy nadzór-nad językami, na których jest testowany.

Metoda	ne po		— Nie, po	
	1M	5M	1M	5M
Wstępnie przeszkolony	6.06	1.49	**7.82**	**5.56**
Laser *lokalny*	**7.37**	**3.15**	7.49	5.01

Tabela 4:Porównanie wynikówzestawu `flores` DevTest z wykorzystaniem ograniczonych i przeszkolonych zaworów LASER.

Metoda	ne po		— po	
	1M	5M	1M	5M
Główny - zespół	6.8	2.8	**6.4**	4.0
Constr.— LASER	**6.9**	2.5	6.2	3.8
Najlepszy (inny)	5.5	**3.4**	5.0	**4.4**

Tabela 3:Oficjalne wyniki wniosków głównych i wtórnych dotyczących zestawu `testów` floresa ocenione w konfiguracji NMT.Dla porównania, zawieramy najlepsze wyniki innego systemu.

5 Wnioski i przyszłe prace

W tym artykule opisujemy nasze poddanie się niskozasobowemu zadaniu filtrowania korpusu równoległego WMT.Używamy wielojęzycznych okładzin zdań z LASER do filtrowania hałaśliwych zdań.Obserwujemy, że LASER może uzyskać lepsze wyniki niż wartości bazoweo szerokim marginesie.Wprowadzenie partytur z innych technik i stworzenie zespołu zapewniadodatkowe

korzyści.Nasze główne poddanie się wspólnemu zadaniu opiera się na najlepszejkonfiguracji zespołu, a nasze kontrastowe poddanie opiera się na najlepszej konfiguracji LASER.Nasze systemy najlepiejsprawdzają się w warunkach 1M dla zadań nepalsko-angielsko-angielskich i syngaleczno-angielskich.Analizujemy wydajność przeszkolonej wersji LASER i obserwujemy, że potrafi ona dobrze wykonywaćzadanie filtrowania nawet w scenariuszach zerowych, co jest bardzo obiecujące.

W przyszłości chcemy ocenić tę technikę dla scenariuszy o wysokich zasobach i zaobserwować, czy te same wyniki przenoszą się do tego warunku.Ponadtoplanujemy zbadać, w jaki sposób wielkość danych szkoleniowych (równoległe, jednojęzyczne) wpływa na zadanie filtrowania zdania o niskich zasobach.

Odniesienia

Mikel Artetxe, Gorka Labaka, Eneko Agirre i Kyunghyun Cho.2018.Bez nadzoru nad maszynowym-tłumaczeniem neuronowym.Na *Międzynarodowej Konferencji Reprezentacji Nauki (ICLR).*

Mikel Artetxe i Holger Schwenk.2018a.Margin— oparte równoległe Corpus Mining with MultilingualSentence Embeddings. arXivpreprint arXiv:1811.01136.

Mikel Artetxe i 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 i Marc'Aurelio Ranzato.2019.Dwa nowe zestawy danych ewaluacyjnych dla niskozasobowego tłumaczenia maszynowego:Nepalsko-angielskie i sinhala-english.arXiv preprint arXiv:1902.01382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark i Philipp Koehn.2013.Skalowalna zmodyfikowana-estymacja modelu języka Kneser-Ney.W *Proceedings-of the 51st Annual Meeting of the Association for-Computational Linguistics, strony 690*—696, Sofia, Bułgaria.

Marcin Junczys-Dowmunt.2018.Podwójne warunkowe filtrowanie krzyżowo-entropijne głośnego równoległego korpusu.W *wyniku trzeciej konferencji na temat tłumaczenia maszynowego, tom 2:Wspólne dokumenty zadaniowe*, strony 901-908, Belgia, Bruksela.Stowarzyszenielingwistyki obliczeniowej.

Huda Khayrallah i Philipp Koehn.2018.Na wpływ różnego rodzaju szumów na przekład maszyny neuronowej.W *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, strony 74-83, Melbourne, Australia.Stowarzyszenie-lingwistyki obliczeniowej.

Huda Khayrallah, Hainan Xu i Philipp Koehn.2018.Systemy filtrowania równoległego korpusu JHU dla WMT 2018.W *wyniku trzeciej konferencji na temat tłumaczenia maszynowego, tom 2:Wspólne dokumenty zadaniowe*, strony 909-912, Belgia, Bruksela.Stowarzyszenielingwistyki obliczeniowej.

Philipp Koehn, Francisco Guzman, Vishrav Chaud- hary i Juan M. Pino.2019.Ustalenia dotyczące wspólnego zadania wmt 2019 w zakresie równoległego filtrowania korpusu w warunkach niskich zasobów.W *postępowaniu z czwartą konferencją na temat tłumaczenia maszynowego, tom 2:Wspólne Prace Zadania*, Florencja, Włochy.Stowarzyszenie lingwistyki obliczeniowej.

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 i Evan Herbst.2007.Mojżesz:Otwarty zestaw narzędzi do statystycznego tłumaczenia maszynowego.W *Coroczne spotkanie StowarzyszeniaJęzykoznawstwa Obliczeniowego (ACL), sesja demonstracyjna.*

Philipp Koehn, Huda Khayrallah, Kenneth Heafield i Mikel L Forcada.2018.Wyniki badań WMT 2018 podzieliły się wspólnym zadaniem w zakresie filtrowania równoległego korpusu.W *wyniku trzeciej konferencji na temat tłumaczenia maszynowego,tom 2:Wspólne dokumenty zadaniowe*, strony 726739, Belgia, Bruksela.Stowarzyszenielingwistyki obliczeniowej.

Guillaume Lample, Myle Ott, Alexis Conneau, Lu-dovic Denoyer i Marc'Aurelio Ranzato.2018.Tłumaczenie maszynowe oparte na frazach i neuronach.In *Empirical Methods in Natural Language Processing (EMNLP)*, strony 5039-5049, Belgia, Bruksela.Stowarzyszenie lingwistyki obliczeniowej.

Fantine Mordelet i J-P Vert.2014.Pakowanie svm do uczenia się z pozytywnych i nieoznaczonych przykładów.List do*rozpoznawania*wzoru, 37:201-209.

Myle Ott, Siergiej Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier i Michael Auli.2019. Fairseq:Szybki, rozszerzalny zestaw narzędzi do modelowania sekwencji.W *postępowaniu z NAACL-HLT 2019:Demonstracje.*

Matt Post.2018.Wezwanie do jasności w raportowaniu wyników bleu.W *wyniku trzeciej konferencji na temat tłumaczenia maszynowego (WMT), tom 1:Dokumenty badawcze*, tom 1804.08771, strony 186-191, Belgia, Bruksela.Stowarzyszenie lingwistyki obliczeniowej.

Victor M Sanchez-Cartagena, Marta Banon, Sergio Ortiz-Rojas i Gema Ramirez.2018. Zgłoszenie Prompsit do równoległego korpusu WMT 2018 filtrującego wspólne zadanie.W *wyniku trzeciej konferencji na temat tłumaczenia maszynowego:Wspólne dokumenty zadaniowe*, strony 955-962, Belgia, Bruksela.Stowarzyszenielingwistyki obliczeniowej.

Holger Schwenk.2018.Filtrowanie i wydobywanie

równoległych danych we wspólnej przestrzeni wielojęzycznej.W *Proceedings of the 56th Annual Meeting of the Association for Computational-Linguistics (Short Papers)*,strony 228234,Australia, Melbourne.Stowarzyszenielingwistyki obliczeniowej.

Hainan Xu i Philipp Koehn.2017.Zipporah:Szybki i skalowalny system czyszczenia danych dla hałaśliwej sieci...W *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2017*, strony 2945-2950, Dania, Cophenhagen.Stowarzyszenie lingwistyki obliczeniowej.