

Low-Resource Corpus Filtering mit mehrsprachigen Satzeinbettungen

Vishrav Chaudhary* Yuqing Tang* Francisco Guzman* Holger Schwenk* Philipp Koehn*

*Facebook KI, Johns Hopkins University
{vishrav,yuqtang,fguzman,schwenk}@fb.com phi@jhu.edu

Zusammenfassung

In diesem Beitrag beschreiben wir unsere Einreichung bei der WMT19 Low-Resource Parallel Corpus Filtering Shared Task. Unser Hauptansatz basiert auf dem LASER-Toolkit (Language-Agnostic sentence Representations), das eine auf einem parallelen Korpus trainierte Encoder-Decoder-Architektur verwendet, um mehrsprachige Satzdarstellungen zu erhalten. Anschließend nutzen wir die Darstellungen direkt, um die lauten parallelen Sätze zu punkten und zu filtern, ohne zusätzlich eine Scoring-Funktion zu trainieren. Wir kontrastieren unseren Ansatz mit anderen vielversprechenden Methoden und zeigen, dass LASER starke Ergebnisse erzielt. Schließlich produzieren wir ein Ensemble unterschiedlicher Scoring-Methoden und erhalten zusätzliche Gewinne. Unsere Einreichung erreichte die beste Gesamtleistung sowohl für die nepalesisch-englischen als auch für die singhalisch-englischen 1M-Aufgaben mit einer Marge von 1,3 bzw. 1,4 BLEU im Vergleich zu den zweitbesten Systemen. Darüber hinaus zeigen unsere Experimente, dass diese Technik für niedrige und sogar ressourcenlose Szenarien vielversprechend ist.

(Koehn et al., 2018).

Während sich die modernsten Methoden, die NMT-Modelle verwenden, im Bergbau als wirksam erwiesen haben

<http://www.paracrawl.eu/>

1 Einleitung

Die Verfügbarkeit hochwertiger paralleler Trainingsdaten ist entscheidend für eine gute Übersetzungsleistung, da neuronale maschinelle Übersetzungssysteme (NMT) gegenüber lauten parallelen Daten weniger robust sind als statistische maschinelle Übersetzungssysteme (Khayrallah und Koehn, 2018). In jüngster Zeit besteht ein erhöhtes Interesse an der Filterung von lärmenden Parallelkorpus (wie Paracrawl1) zur ^{Erhöhung} der Datenmenge, die zur Ausbildung von Übersetzungssystemen verwendet werden kann

Parallele Sätze (Junczys—Dowmunt, 2018) für hochkarätige Sprachen, ihre Wirksamkeit wurde nicht in niedrigkarätigen Sprachen getestet. Die Auswirkungen der geringen Verfügbarkeit von Trainingsdaten für parallele Abstützungsmethoden sind noch nicht bekannt.

Für die Aufgabe der ressourcenschonenden Filterung (Koehn et al., 2019) erhalten wir im Rahmen des Paracrawl-Projekts einen sehr lauten 40,6 Millionen Wort (englische Tokenzahl) nepalesisch-englischen Korpus und einen 59,6 Millionen Wortsinghala-englischen Korpus, der aus dem Netz gekrochen ist. Die Herausforderung besteht in der Bereitstellung von Partituren für jedes Satzpaar in beiden lauten parallelen Sätzen. Die Partituren werden zur Unterstichprobe von Satzpaaren verwendet, die sich auf 1 Millionen und 5 Millionen englische Wörter belaufen. Die Qualität der daraus resultierenden Teilmengen wird durch die Qualität einer statistischen maschinellen Übersetzung (Moses, Phrasenbasiert (Koehn et al., 2007)) und des neuronalen maschinellen Übersetzungssystems fairseq (Ott et al., 2019) bestimmt, das auf diesen Daten geschult wird. Die Qualität des maschinellen Übersetzungssystems wird anhand der BLEU-Scores mit Sacrebleu (Post, 2018) auf einem ausgehaltenen Testsatz von Wikipedia-Übersetzungen für singhalisch-englisch und nepalesisch-englisch aus dem flores-Datensatz gemessen (Guzman et al., 2019).

In unserer Einreichung für diese gemeinsame Aufgabe verwenden wir mehrsprachige Satzeinbettungen von LASER, die mittels¹ einer Encoder-Decoder-Architektur ein mehrsprachiges Satzdarstellungsmodell mit einem relativ kleinen parallelen Korpus trainieren. Unsere Experimente zeigen, dass der vorgeschlagene Ansatz andere bestehende Ansätze übertrifft. Darüber hinaus nutzen wir ein Ensemble von mehreren Scoring-Funktionen, um die Filterleistung weiter zu steigern.

¹<https://github.com/facebookresearch/LASER>

2 Methodik

Die gemeinsame Aufgabe der WMT 2018 zur parallelen Korpusfilterung (Koehn et al., 2018)³ führte mehrere Methoden ein, um eine hochressourcenreiche deutsch-englische Datenbedingung zu bekämpfen. Während viele dieser Methoden es geschafft haben, laute Übersetzungen herauszufiltern, wurden nur wenige unter ressourcenschwachen Bedingungen versucht. In diesem Beitrag befassen wir uns mit dem Problem der ressourcenschonenden Satzfilterung mittels Satzrepräsentationen und vergleichen sie mit anderen gängigen Methoden, die unter ressourcenschonenden Bedingungen eingesetzt werden.

Das LASER-Modell (Artetxe und Schwenk, 2018a) nutzt mehrsprachige Satzdarstellungen, um die Ähnlichkeit zwischen Quelle und Zielsatz abzuschätzen. Sie lieferte hochmoderne Performance bei der Bucc corpus-Mining-Aufgabe und war auch effektiv bei der Filterung von WMT Paracrawl-Daten (Artetxe und Schwenk, 2018a). Bei diesen Aufgaben handelte es sich jedoch nur um hochkarätige Sprachen, nämlich Französisch, Deutsch, Russisch und Chinesisch. Glücklicherweise hat diese Technik auch eine Wirkung auf null-shot Cross-lingual Natural Language Inference im XNLI-Datensatz (Artetxe und Schwenk, 2018b), was es vielversprechend macht, dass das Low-Ressource-Szenario in dieser gemeinsamen Aufgabe fokussiert wird. In diesem Papier schlagen wir vor, eine Anpassung von LASER an ressourcenschwache Bedingungen zu verwenden, um die Ähnlichkeitspunkte zu berechnen, um laute Sätze herauszufiltern.

Für den Vergleich mit LASER legen wir auch erste Benchmarks mit Bicleaner und Zipporah fest, zwei populäre Basislinien, die im Paracrawl-Projekt verwendet wurden; Und dual bedingte Kreuzentropie, die sich für die hochressourcenreiche Korpusfilterung als State-of-the-Art erwiesen hat (Koehn et al., 2018). Wir untersuchen die Leistungsfähigkeit der Techniken unter ähnlichen Vorverarbeitungsbedingungen in Bezug auf Sprachfilterung und lexikalische Überlappung. Wir beobachten, dass LASER-Scores einen klaren Vorteil für diese Aufgabe bieten. Schließlich führen wir die Einbettung der Partituren aus verschiedenen Methoden durch. Wir beobachten, dass, wenn LASER-Scores im Mix enthalten sind, der Leistungsschub relativ gering ist. Im Rest dieses Abschnitts besprechen wir die Einstellungen für jede der angewandten Methoden.

³<http://statmt.org/wmt18/parallel-corpus-filtering.html>

2.1 Laser mehrsprachige Darstellungen

Die zugrunde liegende Idee ist es, die Distanzen zwischen zwei mehrsprachigen Darstellungen als Begriff der Parallelität zwischen den beiden eingebetteten Sätzen zu nutzen (Schwenk, 2018). Um dies zu tun, bilden wir zunächst einen Encoder aus, der lernt, eine mehrsprachige, feste Satzdarstellung zu erstellen; und dann einen Abstand zwischen zwei Sätzen im gelernten Einbettraum berechnen. Darüber hinaus verwenden wir ein *Margenkriterium*, das einen nächsten Nachbarnansatz nutzt, um die Ähnlichkeitspunkte zu normalisieren, da die Kosinähnlichkeit nicht global konsistent ist (Artetxe und Schwenk, 2018a).

Encoder Der mehrsprachige Encoder besteht aus einem bidirektionalen LSTM, und unsere Satzeinbettungen werden durch die Anwendung von max-pooling über seine Ausgabe erhalten. Wir verwenden einen einzigen Encoder und Decoder in unserem System, die von allen beteiligten Sprachen geteilt werden. Zu diesem Zweck haben wir nur mehrsprachige Satzeinbettungen auf die zur Verfügung gestellten parallelen Daten geschult (Details siehe Abschnitt 3.2).

Marge Wir folgen der Definition des *Verhältnisses* von² (Artetxe und Schwenk, 2018a). Damit kann der Ähnlichkeitspunkt zwischen zwei Sätzen (x, y) berechnet werden als

$$\frac{2k \cos(x, y)}{S_y' eNNfc(x) \cos(x, y) + S_x' eNNfc(y) \cos(x, y)}$$

NNK_(x) bezeichnet die nächsten Nachbarn von x in der anderen Sprache, und analog für NNK_(y). Beachten Sie, dass diese Liste der nächsten Nachbarn keine Duplikate enthält, also selbst wenn ein gegebener Satz mehrere Vorkommen im Corpus hat, hätte er (höchstens) einen Eintrag in der Liste.

Nachbarschaft Zusätzlich erforschten wir zwei Möglichkeiten der Probenahme k nächsten Nachbarn. Zunächst eine *globale* Methode, in der wir die Nachbarschaft benutzten, bestand aus den lauten Daten zusammen mit den sauberen Daten. Zweitens eine *lokale* Methode, bei der wir nur die lauten Daten mit der lauten Nachbarschaft, oder die sauberen Daten mit der sauberen Nachbarschaft

²Wir untersuchten die *absoluten*, *Abstand* und *Verhältnis* Marge Kriterien, aber letztere funktionierte am besten

erzielten.³

2.2 Andere Methoden der Ähnlichkeit

Zippora (Xu und Koehn, 2017; Khayrallah et al., 2018), der oft als Basisvergleich verwendet wird, verwendet Sprachmodell- und Wortübersetzungs-Scores, wobei Gewichte optimiert werden, um saubere und synthetische Rauschdaten zu trennen. In unserem Setup haben wir Zipporah-Modelle für beide Sprachpaare Sinhala-Englisch und Nepali-Englisch trainiert. Wir haben die Open-Source-Release6des ^{Zipporah}Tools ohne Modifikationen verwendet. Alle Komponenten des Zipporah-Modells (probabilistische Übersetzungswörterbücher und Sprachmodelle) wurden auf den zur Verfügung gestellten sauberen Daten (ohne die Wörterbücher) geschult. Sprachmodelle wurden mittels KenLM (Heafield et al., 2013) über die sauberen Paralleldaten trainiert. Wir verwenden die zur Verfügung gestellten monolingualen Daten nicht nach der Standardeinstellung. Für das Gewichtstraining haben wir das Entwicklungsset aus dem Flores — Datensatz verwendet.

Bicleaner (Sanchez-Cartagena et al., 2018) verwendet lexische Übersetzung und Sprachmodell Partituren, und mehrere flache Funktionen wie: entsprechende Länge, passende Zahlen und Interpunktion. Wie bei Zip- porah haben wir das Open Source Bicleaner7Toolkit unmodified^{out} —of-the-box verwendet. Nur die zur Verfügung gestellten sauberen parallelen Daten wurden für die Ausbildung dieses Modells verwendet. Bicleaner verwendet eine regelbasierte Komponente, um lautere Beispiele in den parallelen Daten zu identifizieren und trainiert einen Klassifikator, um zu lernen, wie man sie von den restlichen Trainingsdaten trennt. Die Verwendung von Sprachmodellfunktionen ist optional. Wir haben nur Modelle ohne Sprachmodell-Scoring-Komponente verwendet.⁸

Dual Conditional Cross-Entropy Einer der

Die besten Methoden bei dieser Aufgabe waren die dual bedingte Cross-Entropie-Filterung (Junczys — Dowmunt, 2018), die eine Kombination von Vor- und Rückwärtsmodellen zur Berechnung eines lingualen Vergleichspunktes verwendet. In unseren Experimenten nutzten wir für jedes Sprachpaar die zur Verfügung gestellten sauberen Trainingsdaten, um neuronale maschinelle Übersetzungsmodelle in beide Übersetzungsrichtungen zu trainieren: Quelle-zu-Ziel und Ziel-zu-Quelle. Bei einem solchen Übersetzungsmodell M erzwingen wir Satzpaare $(x,$

$y)$ aus dem lauten Parallelkorpus und erhalten die Cross-Entropie-Score

$$HM(y|x) = \frac{1}{|y|} \log_{pm}(YTLy[i,t-i],x) \quad (1)$$

⁶<https://github.com/hainan-xv/zipporah>

⁷<https://github.com/bitextor/bicleaner>

⁸Wir fanden heraus, dass ein LM als Funktion dazu führte, dass fast alle Satzpaare eine Punktzahl von 0 erhielten.

Vorwärts und rückwärts werden die Punkte $H_f(y|x)$ und $H_b(x|y)$ mit einer zusätzlichen Strafe auf einen großen Unterschied zwischen den beiden Punkten gemittelt $|H_f(y|x) - H_b(x|y)|$.

$$\text{Score}(x, y) = \frac{H_f(y|x) + H_b(x|y)}{2} - |H_f(y|x) - H_b(x|y)| \quad (2)$$

Bei den Vorwärts- und Rückwärtsmodellen handelt es sich um Fünfschicht-Encoder/Decoder-Transformatoren, die mit Fairseq trainiert werden und mit den Parametern identisch sind, die den im Basismodell verwendeten Parametern entsprechen⁴⁵. Die Modelle wurden auf den sauberen Paralleldaten für 100 Epochen geschult. Für die nepalesisch-englische Aufgabe haben wir auch Hindi-englische Daten ohne große Unterschiede in den Ergebnissen untersucht. Wir nutzten die Flores— Entwicklung, um das Modell auszuwählen, das BLEU-Scores maximiert.

2.3 Ensemble

Um die Stärken und Schwächen verschiedener Scoring-Systeme zu nutzen, untersuchten wir die Verwendung eines binären Klassifikators zum Aufbau eines Ensembles. Während es trivial ist, Positive (z. B. die sauberen Trainingsdaten) zu erhalten, können Mining Negative eine erschreckende Aufgabe sein. Daher verwenden wir Positiv-unlabeled (PU) Learning (Mordet und Vert, 2014), was uns erlaubt, Klassifikatoren zu erhalten, ohne einen Datensatz von expliziten Positiv- und Negativdaten zu kuratieren. In dieser Einstellung kommen unsere positiven Etiketten von den sauberen Paralleldaten, während die nicht markierten Daten vom lauten Set stammen.

Um dies zu erreichen, verwenden wir Beutel von 100 schwachen, voreingenommenen Klassifikatoren (d. h. mit einer 2: 1-Bias für nicht markierte Daten vs. positive Etikettendaten). Wir verwenden Support-Vektor-Maschinen (SVM) mit einem radialen Basiskernel, und wir untersuchen zufällig den Satz von Funktionen für das Training jeder Basisklassifikator, was dazu beiträgt, sie vielfältig

³ Dieser letzte Teil wurde nur für die Ausbildung eines Ensembles getan

⁴<https://github.com/facebookresearch/Flores#train-a-baseline-transformer-model>

und gering zu halten.

Wir führten zwei Iterationen der Ausbildung dieses Ensembles durch. In der ersten Iteration haben wir die oben beschriebenen ursprünglichen positiven und nicht markierten Daten verwendet. Für die zweite Iteration nutzten wir den gelernten Klassifikator, um die Trainingsdaten neu zu markieren. Wir untersuchten mehrere Relabeling—Ansätze (z. B. die Festlegung eines Schwellenwerts, der den F1—Score maximiert). Wir fanden jedoch heraus, dass die Festlegung einer Klassengrenze, um das ursprüngliche Positiv-zu-unmarkierte Verhältnis zu erhalten, am besten funktionierte. Wir beobachteten auch, dass sich die Leistung nach zwei Iterationen verschlechterte.

3 Experimentelle Einrichtung

Wir experimentierten mit verschiedenen Methoden mit einem Setup, das die offizielle Wertung der gemeinsamen Aufgabe genau widerspiegelt. Alle Methoden werden auf den zur Verfügung gestellten sauberen Paralleldaten geschult (siehe Tabelle 1). Wir haben die gegebenen monolingualen Daten nicht verwendet. Für Entwicklungszwecke haben wir das zur Verfügung gestellte flores dev set verwendet. Zur Auswertung haben wir maschinelle Übersetzungssysteme auf den ausgewählten Teilmengen (1M, 5M) der lauten parallelen Trainingsdaten mittels fairseq mit der Default-Flores-Trainingsparameterkonfiguration geschult. Wir berichten über Sacrebleu-Scores auf dem Volores DevTest Set. Wir haben unser Hauptsystem auf der Grundlage der besten Punkte auf dem DevTest Set für den 1M Zustand ausgewählt.

	SI-en	ne-en	Hi-en
Sätze	646k	573k	1.5M
Englische Wörter	3.7M	3.7M	20.7M

Tabelle 1: Verfügbare bitexte, um die Filteransätze zu trainieren.

3.1 Vorverarbeitung

Wir haben eine Reihe von Filtertechniken angewendet, die denen von LASER ähneln (Artetxe und Schwenk, 2018a) und den lauten Sätzen, die auf falscher Sprache auf der Quelle oder auf der Zielseite basieren oder eine Überschneidung von mindestens 60 % zwischen der Quelle und den Zielmarken aufweisen, eine Punktzahl von -1 zugewiesen. Wir haben fastText10 für die Sprach-ID-Filterung verwendet. Da LASER Ähnlichkeitspunkte für ein Satzpaar mit diesen Filtertechniken berechnet, experimentierten wir, indem wir diese zu den anderen Modellen hinzufügten, die wir für diese gemeinsame Aufgabe benutzten.

3.2 Laser Encoder Training

Für unsere Experimente und die offizielle Einreichung haben wir einen mehrsprachigen Satz-Encoder mit den zulässigen Ressourcen in Tabelle 1 trainiert. Wir trainierten einen einzelnen Encoder mit allen parallelen Daten für Sinhala-Englisch, Nepali-Englisch und Hindi-Englisch. Da Hindi und Nepali das gleiche Drehbuch teilen, verketteten wir ihre Korpus zu einem einzigen parallelen Korpus. Um die Größenunterschiede der parallelen Trainingsdaten zu berücksichtigen, haben wir die singhalisch-englischen und nepalesisch/hindi-englischen Bitexte im Verhältnis 5: 3 überprobiert. Dies führte zu ca. 3,2M Trainingssätzen für jede Sprachrichtung, d. h. Sinhala und kombiniert Nepali-Hindi.

¹⁰⁹<https://fasttext.cc/docs/en/Sprachidentifikation.html>

Die Modelle wurden mit der gleichen Einstellung wie der öffentliche LASER-Encoder trainiert, der die Normalisierung von Texten und Tokenisierung mit Moses-Tools beinhaltet (zurück in den englischen Modus). Mit fastBPE lernen wir zunächst ein gemeinsames 50k BPE Vokabular auf den verketteten Trainingsdaten.⁶ Der Encoder sieht Sinhala, Nepali, Hindi und Englisch Sätze an der Eingabe, ohne irgendwelche Informationen über die aktuelle Sprache zu haben. Diese Eingabe wird immer in Englisch übersetzt.⁷ Wir haben mit verschiedenen Techniken experimentiert, um den englischen Eingabesätzen Rauschen hinzuzufügen, ähnlich dem, was bei unbeaufsichtigten neuronalen maschinellen Übersetzungen verwendet wird, z. B. (Artetxe et al., 2018); Lample et al., 2018), aber dies hat die Ergebnisse nicht verbessert.

Der Encoder ist ein fünfzügiger BLSTM mit 512-dimensionalen Schichten. Der LSTM Decoder hat eine versteckte Schicht der Größe 2048, trainiert mit dem Adam Optimizer. Für die Entwicklung berechnen wir Ähnlichkeitsfehler auf der Verkettung der Flores dev-Sets für Sinhala-Englisch und Nepali-Englisch. Unsere Modelle wurden für sieben Epochen für etwa 2,5 Stunden auf 8 Nvidia GPUs trainiert.

4 Ergebnisse

Aus den Ergebnissen in Tabelle 2 lassen sich mehrere Trends ablesen: (i) Die Werte für den 5M-Zustand sind im Allgemeinen niedriger als für den 1M-Zustand. Diese Bedingung scheint durch die Anwendung von Sprach-ID und Überlappungs-Filterung verschärft zu werden. (ii) LASER zeigt eine

⁶ <https://github.com/glample/fastBPE>

¹² Das bedeutet, dass wir einen englischen Autoencoder trainieren müssen. Dies schien nicht zu schaden, da der gleiche Encoder auch die drei anderen Sprachen behandelt

durchweg gute Leistung. Die *lokale* Nachbarschaft funktioniert besser als die *globale*. In dieser Einstellung liegt LASER im Durchschnitt 0,71 BLEU über dem besten Nicht-LASER-System. Diese Lücken sind für den 1M-Zustand höher (0,94 BLEU). (iii) Die beste Ensemble-Konfiguration bietet kleine Verbesserungen gegenüber der besten LASER-Konfiguration. Für Sinhala-Englisch umfasst die beste Konfiguration jede andere Scoring-Methode (ALL). Für nepalesisch-englisch ist die beste Konfiguration ein Ensemble von LASER Partituren. (iv) Duale Kreuzentropie zeigt gemischte Ergebnisse. Für Sinhala-Englisch funktioniert es erst, wenn die Sprach-ID-Filterung aktiviert ist, die mit früheren Beobachtungen übereinstimmt (Junczys—Dowmunt, 2018). Für nepalesisch-englisch, bietet es Punkte weit unter dem Rest der Scoring-Methoden. Beachten Sie, dass wir keine Architektorexploration durchgeführt haben.

Methode	ne EN		SI-en	
	1M	5M	1M	5M
Zipporah				
Basis	5.03	2.09	4.86	4.53
+ DECKEL	5.30	1.53	5.53	3.16
+ Überlappung	5.35	1.34	5.18	3.14
Dual X-Ent.				
Basis	2.83	1.88	0.33	4.63 ⁺
+ DECKEL	2.19	0.82	6.42	3.68
+ Überlappung	2.23	0.91	6.65	4.31
Bicleaner				
Basis	5.91	2.54 ⁺	6.20	4.25
+ DECKEL	5.88	2.09	6.36	3.95
+ Überlappung	6.12 ⁺	2.14	6.66 ⁺	3.26
LASER				
<i>lokale</i>	7.37*	3.15	7.49*	5.01
<i>Global</i>	6.98	2.98*	7.27	4.76
Ensemble				
ALLES	6.17	2.53	7.64	5.12
Laserglob.Laserglob.	7.49	2.76	7.27	5.08*

Tabelle 2: Sacrebleu punktet auf dem Volores DevTest Set. In fett, wir markieren die besten Punkte für jede Bedingung. In kursiv*unterstreichen wir den Zweitplatzierten. Wir signalisieren auch die beste Nicht-LASER-Methode mit +.

Einreichung Für die offizielle Einreichung nutzten wir das *ALLE* Ensemble für die singhalisch-englische Aufgabe und das LASER *global* + *local* ensemble für die nepalisch-englische Aufgabe. Wir haben auch den LASER *local* als kontrastierendes System eingereicht. Wie wir in Tabelle 3 sehen können, liegen die Ergebnisse der wichtigsten und-

4.1 Diskussion

Eine natürliche Frage ist, wie die LASER-Methode profitieren würde, wenn sie Zugriff auf zusätzliche-Daten hätte. Um dies zu erforschen, haben wir das LASER Open-Source-Toolkit verwendet, das einen geschulten Encoder für 93 Sprachen bereitstellt, aber Nepali nicht einschließt. In Tabelle 4 stellen wir fest, dass das vortrainierte LASER-Modell das LASER-Lokalmodell um 0,4 BLEU übertrifft. Für nepalesisch-englisch kehrt sich die Situation um: Laser *local* liefert viel bessere Ergebnisse. Die Ergebnisse des vortrainierten LASERs sind jedoch nur geringfügig schlechter als die von Bicleaner (6.12), das die beste Nicht-LASER-Methode ist. Dies deutet darauf hin, dass LASER in Zero-Shot-Szenarien (d. h. Nepalisch-Englisch) gut funktionieren kann, aber es funktioniert noch besser, wenn es zusätzliche Überwachung für die Sprachen hat, an denen es getestet wird.

Methode	ne EN		SI—EN	
	1M	5M	1M	5M
Vortrainierte LASER	6.06	1.49	7.82	5.56
Laser <i>lokal</i>	7.37	3.15	7.49	5.01

Tabelle 4: Vergleich der Ergebnisse auf den Fores DevTest Set unter Verwendung der eingeschränkten und vortrainierten Versionen von LASER.

kontrastreichen Einreichungen sehr nahe. In einem Fall liefert das kontrastreiche Lösungsmodell (ein einziges LASER) bessere Ergebnisse als das Ensemble. Mit diesen Ergebnissen lagen unsere 1M-Einreichungen 1,3 bzw. 1,4 BLEU-Punkte über den Vize-Ups der nepalesisch-englischen bzw. singhalasisch-englischen Aufgaben. Wie bereits

erwähnt, schneiden unsere Systeme im 5M Zustand schlechter ab. Wir stellten auch fest, dass sich die Zahlen in Tabelle 2 leicht von den Zahlen unterscheiden (Koehn et al., 2019). Wir schreiben diesen Unterschied dem Effekt der Ausbildung in 4 (unsere) GPUs vs. 1 (ihre) zu.

Methode	ne EN		SI—EN	
	1M	5M	1M	5M
Hauptdarsteller –	6.8	2.8	6.4	4.0
Constr.—LASER vor	6.9	2.5	6.2	3.8
Am besten (andere)	5.5	3.4	5.0	4.4

Tabelle 3: Offizielle Ergebnisse der wichtigsten und sekundären Einreichungen des mit der NMT-Konfiguration bewerteten Testsatzes. Zum Vergleich bringen wir die besten Punkte eines anderen Systems mit ein.

5 Schlussfolgerungen und künftige Arbeiten

In diesem Beitrag beschreiben wir unsere Unterwerfung bei der WMT Low-Resource Parallel Corpus Filtering Task. Wir verwenden mehrsprachige Satzeinbettungen von LASER, um laute Sätze zu filtern. Wir beobachten, dass LASER mit großem Abstand bessere Ergebnisse als die Basislinien erzielen kann. Die Einbeziehung von Partituren aus anderen Techniken und die Schaffung eines Ensembles bietet zusätzliche Gewinne. Unsere Hauptunterwerfung zur gemeinsamen Aufgabe basiert auf dem Besten der Ensemble-Konfiguration und unsere kontrastive Submission basiert auf der Besten LASER-Konfiguration. Unsere Systemerfüllen das Beste auf der 1M-Kondition für die nepalesisch-englischen und singhalasisch-englischen Aufgaben. Wir analysieren die Leistung einer vortrainierten Version von LASER und beobachten, dass sie die Filteraufgabe auch in Null-Ressource-Szenarien gut ausführen kann, was sehr vielversprechend ist.

In Zukunft wollen wir diese Technik für hochkarätige Szenarien evaluieren und beobachten, ob die gleichen Ergebnisse in diese Bedingung übergehen. Darüber hinaus wollen wir untersuchen, wie sich die Größe der Trainingsdaten (parallel, monolingual) auf die ressourcenschonende Satzfilterung auswirkt.

Referenzen

- Mikel Artetxe, Gorka Labaka, Eneko Agirre und Kyunghyun Cho. 2018. [Unbeaufsichtigte neurale maschinelle Übersetzung](#). Auf der Internationalen Konferenz über Lernrepräsentationen (ICLR).
- Mikel Artetxe und Holger Schwenk. 2018a. [Margin—basiertes Parallel Corpus Mining mit mehrsprachigen](#)

[Sentence Embeddings](#). arXiv preprint arXiv: 1811.01136.

- Mikel Artetxe und Holger Schwenk. 2018b. [Massive mehrsprachige Satzeinbettungen für Zero-Shot Cross-Lingual Transfer und darüber hinaus](#). arXiv preprint arXiv: 1812.10464.

Francisco Guzman, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary und Marc’Aurelio Ranzato. 2019. [Zwei neue Auswertungsdatensätze für die ressourcenschonende maschinelle Übersetzung: Nepalesisch-english und sinhala-english](#). arXiv preprint arXiv: 1902.01382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark und Philipp Koehn. 2013. [Skalierbare modifizierte-Kneser-Ney-Sprachmodellschätzung](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Seiten 690—696, Sofia, Bulgarien.

Marcin Junczys-Dowmunt. 2018. [Dual Conditional Cross-Entropy-Filterung von lauten Parallelkorpus](#). In *den Proceedings of the Third Conference on Machine Translation, Band 2: Gemeinsame Task Papers*, Seiten 901-908, Belgien, Brüssel. Vereinigung für Computational Linguistics.

Huda Khayrallah und Philipp Koehn. 2018. [Über die Auswirkungen von verschiedenen Arten von Lärm auf neuronale maschinelle Übersetzung](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Seiten 74-83, Melbourne, Australien. Vereinigung für Computational Linguistics.

Huda Khayrallah, Hainan Xu und Philipp Koehn. 2018. [Die JHU-Parallelkorpus-Filterssysteme für WMT 2018](#). In *den Proceedings of the Third Conference on Machine Translation, Band 2: Gemeinsame Task Papers*, Seiten 909-912, Belgien, Brüssel. Vereinigung für Computational Linguistics.

Philipp Koehn, Francisco Guzman, Vishrav Chaudhary und Juan M. Pino. 2019. Ergebnisse der wmt 2019 gemeinsame Aufgabe zur parallelen Korpusfilterung für ressourcenschonende Bedingungen. In *den Proceedings of the Vierte Konferenz über maschinelle Übersetzung, Band 2: Gemeinsame Task Papers*, Florenz, Italien. Vereinigung für Computational Linguistics.

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 und Evan Herbst. 2007. [Moses: Open Source Toolkit für statistische maschinelle Übersetzung](#). In *Jahrestagung der Association for Computational Linguistics (ACL), Demositzung*.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield und Mikel L Forcada. 2018. [Die Ergebnisse der WMT 2018 teilen die Aufgabe zur parallelen Korpusfilterung](#). In *den Proceedings of the Third Conference on Machine Translation, Band 2: Gemeinsame Task Papers*, Seite

726739, Belgien, Brüssel. Vereinigung für ComputationalLinguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer und Marc'Aurelio Ranzato.2018.[Phrase-basierte & neurale unbeaufsichtigte maschinelle Übersetzung](#).In *Empirical Methods in Natural Language Processing (EMNLP)*, Seiten 5039-5049, Belgien, Brüssel. Vereinigung für Computational Linguistics.

Fantine Mordelet und J-P Vert.2014. Ein Absacken svm von positiven und nicht gekennzeichneten Beispielen zu lernen.*Mustererkennungsbriefe*, 37: 201-209.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier und Michael Auli.2019. Fairseq: Ein schnelles, erweiterbares Toolkit für das Sequenzmodellieren. In *Verfahren von NAACL-HLT 2019: — Vorführungen*.

Matt Post.2018.[Ein Aufruf zur Klarheit in der Berichterstattung bleu punktet](#).In *Proceedings of the Third Conference on Machine Translation (WMT), Band 1: Research Papers*, Band 1804.08771, Seiten 186-191, Belgien, Brüssel. Vereinigung für Computational Linguistics.

Victor M Sanchez-Cartagena, Marta Banon, Sergio Ortiz-Rojas und Gema Ramirez.2018.[Prompsit's Einreichung an WMT 2018 parallele Korpus Filterung gemeinsame Aufgabe](#).In den *Proceedings der dritten Konferenz über maschinelle Übersetzung: Gemeinsame Task Papers*, Seiten 955-962, Belgien, Brüssel. Vereinigung für ComputationalLinguistics.

Holger Schwenk.2018.[Parallele Daten in einem gemeinsamen mehrsprachigen Raumfiltern und abbauen](#).In *Proceedings of the 56th Annual Meeting of the Association for ComputationalLinguistics (Short Papers)*,Seite 228234,Australien, Melbourne. Vereinigung für ComputationalLinguistics.

Hainan Xu und Philipp Koehn.2017.[Zipporah: Ein schnelles und skalierbares Datenreinigungssystem für laute Web— Crawled Parallel Corpora](#).In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, S. 2945-2950, Dänemark, Copenhagen. Vereinigung für Computational Linguistics.