Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment

Philipp Koehn[*†], **Vishrav Chaudhary**[†], **Ahmed El-Kishky**[†]
Naman Goyal[†], **Peng-Jen Chen**[†], **Francisco Guzmán**[†]
phi@jhu.edu, vishrav@fb.com, ahelk@fb.com
naman@fb.com, pipibjc@fb.com, fguzman@fb.com
[*]Johns Hopkins University, Baltimore, Maryland, United States
[†]Facebook AI, Menlo Park, California, United States

Abstract

Following two preceding WMT Shared Tasks on Parallel Corpus Filtering (Koehn et al., 2018, 2019), we posed again the challenge of assigning sentence-level quality scores for very noisy corpora of sentence pairs crawled from the web, with the goal of sub-selecting the highest-quality data to be used to train machine translation systems. This year, the task tackled the low resource condition of Pashto–English and Khmer–English and also included the challenge of sentence alignment from document pairs. 10 participants from companies, national research labs, and universities participated in this task.

1 Introduction

The field of Machine Translation has experienced significant advances in recent years thanks to improvements in neural modeling (Bahdanau et al., 2015; Gehring et al., 2016; Vaswani et al., 2017), as well as the availability of large parallel corpora for training (Tiedemann, 2012; Smith et al., 2013; Bojar et al., 2017). Unfortunately, today's neural machine translation models, perform poorly on *low-resource* language pairs, for which clean, high-quality training data is lacking (Koehn and Knowles, 2017). Improving performance on low resource language pairs has high impact considering that these languages are spoken by a large fraction of the world population. This is a particular challenge for industrial machine translation systems that need to support hundreds of languages in order to provide adequate services to their multilingual user base.

While there have been advances in using monolingual corpora (Lample et al., 2018; Liu et al., 2020) and parallel corpora in multiple language pairs (Aharoni et al., 2019; Fan et al., 2020), the best training data for machine translation are still parallel corpora in the targeted language pair and domain.

Parallel corpora are typically gathered from any available source without much guarantees about quality. This is especially the case for parallel corpora that are extracted from the web without much control over which web sites are mined. Since noisy training data has been recognized as a challenge for neural machine translation training (Khayrallah and Koehn, 2018), an essential step in using such data is filtering or discounting noisy sentence pairs.

Recently, there is increased interest in the filtering of noisy parallel corpora to improve the data that can be used to train translation systems. The Shared Task on Parallel Corpus Filtering and Alignment at the Conference for Machine Translation (WMT 2020) was organized to promote research to make learning from noisy data more viable for low-resource languages. It is similar to the previous year's task but tackles different languages (Pashto and Khmer instead of Nepali and Sinhala) and also included the challenge to extract sentence pairs from document pairs.

The shared task is organized similarly to previous years (Koehn et al., 2018, 2019). We provide about 11.6 million word noisy parallel data for Pashto-English and 58.3 million word noisy parallel data for Khmer-English. We also provide small amounts of clean parallel data of varying quality and monolingual data from Wikipedia and CommonCrawl.

Participants developed methods to assign a quality score for each sentence pair. These scores are used to filter the web crawled corpora down to

Proceedings of the 5th Conference on Machine Translation (WMT), pages 726–742

a fixed size (5 million English words), train neural machine translation systems on these subsets, and measure their quality with the BLEU score on a test set of multi-domain Wikipedia content (Guzmán et al., 2019).

This paper gives an overview of the task, presents the results for the participating systems and provides analysis on additional subset sizes, the average sentence length of sub-selected data, and overlap between the submissions.

2 Related Work

Although the idea of crawling the web indiscriminately for parallel data goes back to the 20th century (Resnik, 1999), work in the academic community on extraction of parallel corpora from the web has so far mostly focused on large stashes of multilingual content in homogeneous form, such as the Canadian Hansards, Europarl (Koehn, 2005), the United Nations (Rafalovitch and Dale, 2009; Ziemski et al., 2015), or European Patents (Täger, 2011). A nice collection of the products of these efforts is the OPUS web site[1] (Tiedemann, 2012).

2.1 Parallel Corpus Acquisition

Noisy parallel documents and parallel sentences were sourced from the CCAligned[2] dataset (El-Kishky et al., 2020a), a massive collection of cross-lingual web documents covering over 8k language pairs aligned from 68 Common Crawl snapshots. Additional parallel data was sourced from the Paracrawl project – a large-scale effort to crawl text from the web[3] (Bañón et al., 2020).

Acquiring parallel corpora from the web (El-Kishky et al., 2020b) is an active area of research that typically involves identifying web sites with parallel text, downloading the documents from the web site, aligning document pairs (Buck and Koehn, 2016; Thompson and Koehn, 2020; El-Kishky and Guzmán, 2020), and aligning sentence pairs. A final stage of the processing pipeline filters out non-parallel sentence pairs. Such noise exists either because the original web site did not have any actual parallel data (garbage in, garbage out), only partially-parallel data, or due to failures of processing steps.

[1] http://opus.nlpl.eu
[2] http://statmt.org/cc-aligned
[3] http://www.paracrawl.eu/

2.2 Sentence Alignment

Sentence alignment has been a very active field of research since the early days of statistical machine translation. An influential early method is based on sentence length, measured in words (Gale and Church, 1993). Several researchers proposed including lexical information (Chen, 1993; Moore, 2002) with the emergence of tools that use provided bilingual dictionaries (Varga et al., 2005) or acquire them during in an unsupervised fashion (Braune and Fraser, 2010). Later work introduced scoring methods that use MT to get both documents into the same language (Sennrich and Volk, 2010) or use pruned phrase tables from a statistical MT system (Gomes and Pereira Lopes, 2016). Both methods anchor high-probability 1–1 alignments in the search space and then fill in and refine alignments. More recently, Thompson and Koehn (2019) introduced the use of sentence embeddings and a coarse-to-fine search method to the task (Vecalign).

2.3 Filtering Noisy Parallel Corpora

In 2016, a shared task on sentence pair filtering[4] was organized, albeit in the context of cleaning translation memories which tend to be cleaner than the data at the end of a pipeline that starts with web crawls.

There is a robust body of work on filtering out noise in parallel data. For example: Taghipour et al. (2011) use an outlier detection algorithm to filter a parallel corpus; Xu and Koehn (2017) generate synthetic noisy data (inadequate and non-fluent translations) and use this data to train a classifier to identify good sentence pairs from a noisy corpus; and Cui et al. (2013) use a graph-based random walk algorithm and extract phrase pair scores to weight the phrase translation probabilities to bias towards more trustworthy ones.

Most of this work was done in the context of statistical machine translation, but more recent work targets neural models. Carpuat et al. (2017) focus on identifying semantic differences in translation pairs using cross-lingual textual entailment and additional length-based features, and demonstrate that removing such sentences improves neural machine translation performance.

As Rarrick et al. (2011) point out, one type of noise in parallel corpora extracted from the web

[4] NLP4TM 2016: Shared task http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/

are translations that have been created by machine translation. Venugopal et al. (2011) propose a method to watermark the output of machine translation systems to aid this distinction, with a negligible loss of quality. Antonova and Misyurev (2011) report that rule-based machine translation output can be detected due to certain word choices, and statistical machine translation output can be detected due to lack of reordering. It is notable that none of the participants in our shared task have tried to detect machine translation.

There is a rich literature on data selection which aims at sub-sampling parallel data relevant for a task-specific machine translation system (Axelrod et al., 2011). Van der Wees et al. (2017) find that the existing data selection methods developed for statistical machine translation are less effective for neural machine translation. This is different from our goals of handling noise since those methods tend to discard perfectly fine sentence pairs that are just not relevant for the targeted domain. Our task is focused on data quality that is relevant for all domains.

2.4 Impact of Noise on Neural Machine Translation

Belinkov and Bisk (2017) investigate the impact of noise on neural machine translation. They focus on creating systems that can *translate* the kinds of orthographic errors (typos, misspellings, etc.) that humans can comprehend. In contrast, Khayrallah and Koehn (2018) examine noisy *training* data and focus on types of noise occurring in web-crawled corpora. They carried out a study about how noise that occurs in crawled parallel text impacts statistical and neural machine translation.

Neural machine translation model training may combine data selection and model training, taking advantage of the increasing quality of the model to better detect noisy data or to increasingly focus on cleaner parts of the data (Wang et al., 2018; Kumar et al., 2019).

2.5 Findings of Previous Shared Tasks

We organized versions of this shared task in the previous two years. In 2018, we started with a high-resource language pair (German–English) and a very large web-crawled parallel corpus, a subset of the Paracrawl corpus consisting of 1 billion English words (Koehn et al., 2018). The best-performing submission (Junczys-Dowmunt, 2018) used neural machine translation systems in both translation directions to score sentence pairs with dual cross-entropy.

Last year, we moved the focus to low resource languages (Koehn et al., 2019) with smaller noisy parallel corpora, comprising 50-60 million words for Nepali–English and Sinhala–English. For these languages much less clean parallel data was available and hence many of the methods developed for high-resource languages are less reliable. The best-performing submission that year (Chaudhary et al., 2019) also considered dual cross-entropy but found that matching multilingual sentence embeddings (Schwenk, 2018) gave better results.

2.6 Monolingual Pre-Training

By now, neural machine translation systems are rarely trained only on the parallel corpus of the desired language pair. Common foundations are pre-trained models trained on multiple language pairs which share the source or target language (Aharoni et al., 2019; Fan et al., 2020) or monolingual pre-training methods (Liu et al., 2020). Often, the models are also improved by a second stage of training that uses back-translated synthetic parallel data that was generated from first stage model — a process that may be iterated (Hoang et al., 2018).

To reflect such a more realistic training setup, we provided pre-trained models that were trained on monolingual data using a denoising auto-encoder method called mBART (Liu et al., 2020). Here, monolingual data is converted into input and output pairs by (a) masking out words in the input, forcing the model to learn the correct word or word sequence from the context, and (b) shuffling the order of a few concatenated sentence pairs.

3 Shared Task Definition

The shared task tackled the problem of filtering parallel corpora. Given a noisy parallel corpus (crawled from the web), participants developed methods to align sentences in document pairs and to filter it to a smaller size of high quality sentence pairs.

3.1 Filtering

For the filtering-only task, we provided a very noisy 58.3 million word corpus for Khmer–English (English token count) and a 11.6 million word corpus for Pashto–English, crawled from the

web (see Section 4.3 for details). We asked participants to generate sentence-level quality scores that allow selecting subsets of sentence pairs that amount to 5 million words, counted on the English side. This amount was chosen based on preliminary experiments (we report below on additional subset sizes).

Participants in the shared task submitted a file with quality scores, one score per line, corresponding to the sentence pairs. Scores are only required to have the property that higher scores indicate better quality. The scores were uploaded to a Google Drive folder which remains publicly accessible.[5]

3.2 Alignment

We also released the document pairs from which we extracted the sentence pairs. For Khmer–English, we released 391,250 document pairs, for Pashto-English 45,312 document pairs.

Participants were encouraged to develop novel methods for sentence alignment. The resulting sentence pairs also had to be annotated with quality scores, as in the filtering-only tasks, and uploaded with quality scores to the same Google Drive folder.

3.3 Evaluation

The submissions were scored by building a neural machine translation system (Ott et al., 2019) trained on this data, and then measuring their BLEU score on the `flores` Wikipedia test sets (Guzmán et al., 2019). The neural machine translation model was either randomly initialized or initialized by monolingual pre-training (mBART).

For development purposes, we released configuration files and scripts that mirror the official testing procedure with a development test set. The development pack consists of:

- A script to subsample corpora based on quality scores.
- `fairseq` scripts to train and test a neural machine translation system.
- A pre-trained mBART model for continued training.
- The `flores-dev` set of Wikipedia translations as development set.
- The `flores-devtest` set of Wikipedia translations as development test set.

[5] https://bit.ly/2IoOXOr

Corpus	Sentence Pairs	English Words
Pashto-English GNOME	95,312	277,188
KDE4	3,377	8,881
Tatoeba	31	239
Ubuntu	9,645	26,626
Bible	13,432	298,522
TED Talks	664	11,157
Wikimedia	737	37,566

Table 1: Provided clean parallel data for Pashto–English.

The web site for the shared task[6] provided detailed instructions on how to use these tools to replicate the official testing environment.

4 Data

We provided three types of data for this shared task: (1) clean parallel and monolingual data, including related language data in Hindi, to train models that aid with the filtering task, (2) the noisy parallel data crawled from the web which participants have to score for filtering, and (3) development and test sets that are used to evaluate translation systems trained on filtered data.

4.1 Clean Parallel Data

For Pashto (see Table 1 for detailed statistics), the largest data sets are the Bible (prepared for us by Arya McCarthy and David Yarowsky), various data sets from OPUS[7] (GNOME, KDE4, and Unbuntu software localization; Tatoeba volunteer translations; and Wikimedia), and a TED Talks corpus created for this task, crawled from TED web site, and sentence-aligned with Vecalign (Thompson and Koehn, 2019).

For Khmer (see Table 2 for detailed statistics), the largest data sets are the alignment of 2 English with 4 Khmer Bibles, various data sets from OPUS (GNOME, KDE4, and Ubuntu software localization; GlobalVoices citizen journalism articles; Tatoeba volunteer translations; and Wikimedia). We also re-aligned the Jehova's Witness corpus (JW300), a collection of religious texts, with Vecalign.

[6] http://www.statmt.org/wmt20/parallel-corpus-filtering.html

[7] http://opus.nlpl.eu/

Corpus	Sentence Pairs	English Words
GNOME	56	233
GlobalVoices	793	14,294
KDE4	120,087	767,919
Tatoeba	748	3,491
Ubuntu	6,987	27,413
Bible	54,222	1,176,418
JW300	107,156	1,827,348

Table 2: Provided clean parallel data for Khmer–English.

	Wikipedia	CommonCrawl
Pashto	76,557	6,558,180
Khmer	132,666	13,832,947
English	67,796,935	1,806,450,728

Table 3: Provided clean monolingual data (number of sentences).

For both language pairs, the available clean parallel data is rather small and mostly out-of-domain. It is not sufficient to build reasonable machine translation systems. In fact, even the provided raw unfiltered noisy parallel data gives better results when used directly for training.

4.2 Clean Monolingual Data

Monolingual data is always available in much larger quantities, and we provided data from two sources: Wikipedia and CommonCrawl. Both contain language that is similar to what is expected in the noisy web data to be filtered.

We filtered the data to eliminate overlap with the development and test sets. See Table 3 for detailed statistics.

4.3 Noisy Parallel Data

Noisy parallel data sourced from CCAligned and Paracrawl follow different philosophies. While CCAligned mines bitexts from a high-precision set of aligned web-documents yielding cleaner parallel bitexts, the noisy parallel corpora from Paracrawl are the outcome of a processing pipeline aimed at high recall at the cost of precision, yielding noisy bitexts. They exhibit noise of all kinds: wrong language in source and target, sentence pairs that are not translations of each other, bad language (incoherent mix of words and non-words), incomplete or bad translations, etc.

To ensure that CCAligned yields additional noisy pairs, we don't perform any filtering after mining bitexts from the CCAligned corpus.

We used the processing pipeline of the Paracrawl project to create the data, using the clean parallel data to train underlying models such as the dictionary used by Hunalign (Varga et al., 2007) and a statistical translation model used by the document aligner. The provided parallel corpus is the raw output of the crawling pipeline, with sentence pairs de-duplicated but otherwise no further filtering performed. See Table 4 for statistics of the corpus and Tables 5 and 6 for some example sentences.

4.4 Development and Test Sets

For test and development purposes, we use the `flores` Wikipedia data sets (Guzmán et al., 2019). These sets are multi-domain, that is they were sampled from Wikipedia documents with a diverse set of topics. In Table 7 we present the statistics of these sets. The official scoring of machine translation systems generated from the sub-sampled data sources is done on the *test* set.

5 Evaluation Protocol

The testing setup mirrors the development environment that we provided to the participants.

5.1 Participants

We received submissions from 10 different organizations, and an additional baseline LASER submission that was posted on the website. See Table 8 for the complete list of participants. The participant's organizations are quite diverse, with 3 participants from the United States, 2 participants from China, and 1 participant each from Canada, Egypt, Turkey/China, Scotland, and Spain. 3 of the participants are universities, 4 are companies, 1 is a joint company/university participant, and 2 are national research organizations. There was little participant overlap between this year's shared task and last year's shared task. Only AFRL and NRC participated also last year.

Each participant submitted up to 3 different sets of scores, not all participants addressed both languages, resulting in a total of 16 different submissions for Pashto and 11 different submissions for Khmer, including a baseline submission of using

	Sentence Pairs	English Words	Document Pairs
Pashto–English	1,022,883	11,551,009	45,312
Khmer–English	4,169,574	58,347,212	391,250

Table 4: Noisy parallel data to be filtered (de-duplicated raw output). Data is made available as aligned sentence pairs (see table for number of English words) and as document pairs for which sentence alignment has to be performed.

Pashto	English
د Mikoyan-Gurevich MiG-29 (ناټو کوډ مګ‌ځنۍ‌ای) د شوروي اتحاد د هوايي غلبې لپاره هېله‌ل‌ومړی‌و 1970s چې او د چا ل‌ومړنی الوتنه ترسره اکتوبر 6، 1977 یې چنګ‌یالی الوتکې. دا په 1983 کې د شوروي پوځ د خدمتونو ته ننوتل او اوس هم د روسیې د هوايي ځواک او نورو له خوا نړ کارول. 1,100 زیات نسخې تر اوسه جوړ شوې دي.	The Mikoyan-Gurevich MiG-29 (NATO code Fulcrum) is a fighter aircraft of the Soviet air supremacy developed in the early 1970s and whose first flight took place October 6, 1977. It entered service in the Soviet army in 1983 and is still used today by the Russian Air Force and many others. More than 1,100 copies have so far been built.
تشنابشاورونه جوړونکي لپاره د یونان عرضه کوونکی (1)	Bathroom faucets manufacturer supplier wholesale for Yerevan (1)
تیلیفون: 4663278-0543	Telefon: 0543-4663278
بې‌اکتنه شوې: تیموتی شیم	Reviewed by: Timothy Shim
ستاسو IP پته 54.227.76.35 ده. My-ip-is.com د لیدلو لپاره کار کیدی شي IP پتهد موندلو لپاره جیو تایمز د آی IP پته، د پراکسي کشف، د بریښنالیک بریښنالیک او د تور لیست چکونه. نوی: زمونږ سره د انټرنیټ چټک وګورئ چټک ټسټ. غواړئ خپل پی پی رومن شمیري په اړه پوه شئ؟ خپل چک وګورئ د رومن شمیره IP.	Your IP address is 54.227.76.35. My-ip-is.com can be handy for looking up IP addresses, to find out the GeoLocation of a IP address, proxy detection, email tracing and blacklist checks. New: Check your Internet Speed with our Speed Test. Want to know your IP in Roman Numerals? Check Your Roman Numerals IP.

Table 5: Examples of relatively good sentence pairs from the noisy corpus for Pashto–English. Note that unreliable sentence splitting for Pashto led to merging of sentence pairs.

Khmer	English
21:13 និងព្រះយេស៊ូវចូលមកជិត, ហើយគាត់បានយកនំប៉័ង, ហើយគាត់បានផ្តល់ឱ្យវាដើម្បីឱ្យពួកគេ, ដូចគ្នានេះដែរជាមួយនឹងត្រី.	21:13 And Jesus approached, and he took bread, and he gave it to them, and similarly with the fish.
ឌានយ៉ូអាន់ • ខែមករា 8, 2015 នៅ 9:54 ល្ងាច • ឆ្លើយតប	JoAnna • January 8, 2015 at 9:54 pm • Reply
មនុស្សដែល, ដោយអាយុឬពិការភាព, មិនអាចទៅយកសំបុត្ររបស់ពួកគេនៅក្នុងប្រអប់សំបុត្រជាការlantbrevbärarservice ទៀងទាត់ផងដែរអាចអនុវត្តសម្រាប់សេវាកម្មខាងក្រោម:	people who, because of age or disability, unable to retrieve their mail in the mailbox as regular lantbrevbärarservice can also apply for the following services:
/កំណត់ហេតុផ្សេងៗ/IDFP/IDFP (615250-02-7): ប៉ែបផែន, កិតើ, ពិនិត្យ	/Blog/IDFP/IDFP (615250-02-7): Effects, Dosage, Reviews
A Quiet Place (ភាសាអង់គ្លេស, ភាសាខ្មែរ)	A Quiet Place (English, Khmer)
KMSPICO ប្រព័ន្ធប្រតិបត្តិការ Windows 10 - ទាញយកសកម្មគាំទ្រដោយកតគិតថ្លៃ - Softkelo - រកឃើញកម្មវិធីដែលគ្មានដែនកំណត់, ការបង្ក្រាប & ការ Hack	KMSPICO Windows 10 - Free Download Pro Activator - Softkelo - Find Unlimited Softwares, Cracks & Hacks0

Table 6: Examples of relatively good sentence pairs from the noisy corpus for Khmer–English. Note the lack of word segmentation in Khmer leads to very long tokens.

just the LASER scores that was provided to participants at the outset.

5.2 Methods used by Participants

This year, participants in general used a broader range of features and more sophisticated classifier approaches than previously. We first provide an overview of methods and then give a short summary of each submission.

5.2.1 Methods

Pre-filtering Almost all participants employ pre-filtering rules, based on the length of sentences in terms of tokens or characters, ratio of the lengths, ratio of alpha-numerical tokens, overlap between the English and the foreign sentence (to

	Pashto		Khmer	
	Sentence Pairs	**English Words**	**Sentence Pairs**	**English Words**
dev	3,162	55,439	2,378	40,436
dev test	2,698	46,175	2,309	44,471
test	2,719	47,695	2,320	40,341

Table 7: Statistics for the `flores` test sets used to evaluate the machine translation systems trained on the subsampled data sets. Word counts are obtained with `wc` on tokenized text.

Short Name	**Participant and System Description Citation**
AFRL	Air Force Research Lab, USA
Alibaba	Alibaba, China (Lu et al., 2020)
Bytedance	Bytedance, China (Xu et al., 2020)
Edinburgh	University of Edinburgh, Scotland
Huawei	Huawei, Turkey/China (Açarçiçek et al., 2020)
JHU-Kejriwal	Ankur Kejriwal, Johns Hopkins University, USA (Kejriwal and Koehn, 2020)
JHU-Koerner	Felicia Koerner, Johns Hopkins University, USA (Koerner and Koehn, 2020)
Microsoft	Microsoft, Egypt Development Center, Egypt (Nokrashy et al., 2020)
NRC	National Research Council, Canada (Lo and Joanis, 2020)
UA-Prompsit	University of Alicante and Prompsit, Spain (Esplà-Gomis et al., 2020)
LASER	Officially provided baseline

Table 8: Participants in the shared task.

avoid copy noise), or mismatched email addresses, URLs or numbers.

A common pre-filtering method is also language ID. However mixed results were reported and some participants decided to not use it for Pashto (Açarçiçek et al., 2020).

Some participants worked on morphological segmentation of Khmer but this did not lead to any improvements (Esplà-Gomis et al., 2020; Koerner and Koehn, 2020).

LASER We provided LASER scores that performed well in previous year's filtering task. LASER sentence embeddings are trained as a bottleneck feature for a neural machine translation model and trained on a large collection of parallel corpora in 93 languages[8] which include Khmer but not Pashto. A similarity score for a sentence pair is computed as the cosine distance between the English sentence embedding and the foreign sentence embedding (Nokrashy et al., 2020; Kejriwal and Koehn, 2020; Koerner and Koehn, 2020).

[8] https://github.com/facebookresearch/LASER#supported-languages

Dual cross entropy Neural machine translation systems trained on the provided clean parallel data can be used by feeding in the English sentence and computing the probability of the foreign sentence according to the model, and vice versa. Junczys-Dowmunt (2018) proposed a metric that uses not only the individual computed cross entropy scores but also the difference between them (Lu et al., 2020; Koerner and Koehn, 2020).

Language models To assess the quality of sentences by themselves, i.e., preferring sentences that are fluent in the language, statistical or neural language models are trained, typically using provided Wikipedia and CommonCrawl corpora (Lu et al., 2020; Esplà-Gomis et al., 2020; Kejriwal and Koehn, 2020; Koerner and Koehn, 2020; Lo and Joanis, 2020).

Statistical word translation scores Words in the two sentences should be translation of each other. To what degree this is the case can be assessed with classic word translation models which are learned with the EM algorithm over the clean parallel data (Lu et al., 2020; Lo and Joanis, 2020; Esplà-Gomis et al., 2020).

Classifier An increasing number of participants framed the quality estimation problem as a classification task. This requires positive examples drawn from the provided clean parallel text and negative examples created by corrupting these examples. Typically this involves mismatched sentences, truncated sentences, sentences with swapped word order (Esplà-Gomis et al., 2020; Açarçiçek et al., 2020; Nokrashy et al., 2020; Xu et al., 2020). To create harder negative examples for the classifier, a sentence is paired not with a random sentence from the foreign corpus but with a neighboring sentence of the correctly paired sentence and sentences that have 60% similarity (measured by fuzzy match score) to the correct translation (Açarçiçek et al., 2020).

5.2.2 Individual Submissions

AFRL use their corpus-building method (Erdmann and Gwinnup, 2019) but with a bidirectional quality metric that nearly eliminates pre-filtering (used only for the limit on training line length). The coverage metric encourages the addition of a sentence that improves corpus-level bilingual vocabulary frequencies. The new quality metric is the average of sentence-level NMT scores ("log-likelihoods") in both directions.

Alibaba (Lu et al., 2020) use a number of features that are combined linearly: a bilingual GPT-2 model trained on source-target language pairs as well as monolingual GPT-2 model each of the languages, dual cross entropy from neural machine translation models trained in both directions and statistical word translation model scores. They report that they experimented with classifiers to weight features but found this to be not beneficial.

Bytedance (Xu et al., 2020) tackle only the combined alignment/filtering task. The sentence alignment methods draws on statistical lexical translation scores, as used in YiSi-2. They iteratively improve the lexical model by adding high-quality mined sentence pair to its training data. Their filtering method is a classifier based on monolingual language models and a cross-lingual language model (XLM), followed by an added convolutional layer. They also use language ID and n-gram coverage during a re-ranking stage and ensemble model variations (different architectures, hyper parameters).

Huawei (Açarçiçek et al., 2020) focus on an end-to-end classifier approach that learns to distinguish clean parallel data from misaligned sentence pairs. The model first uses a Transformer model to obtain sentence representations, followed either by a classifier (Siamese network) or additional layers that are fine-tuned. They report better performance with a RoBERTa-style Transformer setup over a BERT-style Transformer. A relatively small training corpus is used (2,000 or 10,000 sentence pairs) with 10x over-sampled negatives.

JHU-Kejriwal (Kejriwal and Koehn, 2020) use LASER scores with some novel transformation of score ranges, language ID confidence scores, monolingual language models trained on words and characters, and length-based filters.

JHU-Koerner (Koerner and Koehn, 2020) employ a linear combination of LASER scores, monolingual language model scores, dual cross entropy, and use a sentence duplication penalty.

Microsoft (Nokrashy et al., 2020) focus on the LASER scores, using both the provided LASER scores, custom LASER scores using a model trained on the provided clean parallel data (which are better for Pashto but worse for Khmer), and a classifier built on a pair of LASER sentence embeddings trained to distinguish between clean sentence pairs and artificially bad sentence pairs. While these three scores fare differently for the two languages pairs, a combination of them performs best.

NRC (Lo and Joanis, 2020) tackle both filtering and alignment. Their filtering score is mainly based on Yisi-2 (Lo, 2019), a language model trained on the target side, and representations obtained with XLM-RoBERTa (Conneau et al., 2020) pre-trained for Pashto, Khmer, and English. Sentence alignment is based on the approach by Moore (2002), first applied to align paragraphs and then sentences.

UA-Prompsit (Esplà-Gomis et al., 2020) use an extended version of the established Bicleaner tool which is a classifier that uses several features ranging from coarse (e.g., statistical word translation models scores) to shallow (e.g., average token length, length ratio, punctuation count). The classifier uses the extremely randomized tree algorithm. They also use a 7-gram character language model as refinement.

LASER scores were provided to participants, with filtering for language ID and maximum 60% overlap between source and target sentence.

Edinburgh did not submit a system description paper.

5.3 Subset Selection

We provided to the participants a file containing one sentence pair per line (see Section 4.3) each for the two languages. A submission to the shared task consists of a file with the same number of lines, with one score per line corresponding to the quality of the corresponding sentence pair.

To evaluate a submitted score file, we selected subsets of a predefined size, defined by the number of English words (5 million). We chose the number of English words instead of Pashto or Khmer words, since the latter would allow selection of sentence pairs with very few non-English words and many English words which are beneficial for decoder training but do not count much towards the non-English word total.

Selecting a subset of sentence pairs is done by finding a threshold score, so that the sentence pairs that will be included in the subset have a quality score at and above this threshold. In some cases, a submission assigned this threshold score to a large number of sentence pairs. Including all of them would yield too large a subset, excluding them yields too small a subset. Hence, we randomly included some of the sentence pairs with the exact threshold score to get the desired size in this case.

5.4 Evaluation System Training

Given a selected subset of a given size for a system submission, we built neural machine translation systems from scratch (SCRATCH) and by continued training on a pre-trained model (MBART) to evaluate the quality of the selected sentence pairs.

SCRATCH For from-scratch training, we used the `fairseq` (Ott et al., 2019) transformer model with the parameter settings shown in Figure 1. Preprocessing was done with sentence piece for a 5000 subword vocabulary on tokenized text using the Moses tokenizer (but no truecasing was used). Decoding was done with beam size 5 and length normalization 1.2. Training a system for the 5 million subsets took about 13 hours, on a single GTX 1080ti GPU. Scores on the test sets were computed with Sacrebleu (Post, 2018). We report case-insensitive scores.

```
--arch transformer
--share-all-embeddings
--encoder-layers 5
--decoder-layers 5
--encoder-embed-dim 512
--decoder-embed-dim 512
--encoder-ffn-embed-dim 2048
--decoder-ffn-embed-dim 2048
--encoder-attention-heads 2
--decoder-attention-heads 2
--encoder-normalize-before
--decoder-normalize-before
--dropout 0.4
--attention-dropout 0.2
--relu-dropout 0.2
--weight-decay 0.0001
--label-smoothing 0.2
--criterion label_smoothed_cross_entropy
--optimizer adam
--adam-betas '(0.9, 0.98)'
--clip-norm 0
--lr-scheduler inverse_sqrt
--warmup-update 4000
--warmup-init-lr 1e-7
--lr 1e-3 --min-lr 1e-9
--max-tokens 4000
--update-freq 4
--max-epoch 100
--save-interval 10
```

Figure 1: The baseline `flores` model settings[9] for the NMT training from scratch with `fairseq`

MBART For mBART evaluation, we initialize the weights of transformer with the mBART bilingual pre-training. We used monolingual text from CommonCrawl with denoising objective to pre-train the transformer. We trained 2 bilingual mBART models, one with English and Pashto text and another with English and Khmer text. Both these models were pre-trained with batch size of 256 for 500,000 updates, which took about 57 hours on 16 V100 GPUs.

Continued training on the filtered subsets uses some different parameter settings, as listed in Figure 2. This continued training is faster; it takes about half as much time.

6 Results

In this section we present the results of the shared task evaluation. We added additional unofficial condition at 2, 3, and 7 million English words, to better observe tendencies.

6.1 Core Results

The results are reported in Table 9 (Pashto) and Table 10 (Khmer). The tables contains the BLEU

[9] https://github.com/facebookresearch/flores#train-a-baseline-transformer-model

```
--dropout 0.1
--attention-dropout 0.1
--relu-dropout 0.0
--weight-decay 0.0
--label-smoothing 0.1
--adam-eps 1e-06
--lr 0.0001
--max-update 100000
--patience 10
```

Figure 2: Different model settings for continued training of the provided mBART model. The other settings are the same.

scores for

- development test set and final test set
- neural machine translation from scratch and mBART pre-training
- 2, 3, 5 and 7 million word subsets.

The official scoring is for the 5 million word data settings on the final test set. In the table, we highlight cells for the best scores for each of these settings, as well as scores that are close to it. Results for the unofficial 2, 3 and 7 million word baseline are shown without highlighting.

For almost all submission the highest BLEU scores is reached with subsets of 5 million words. There is also fairly high consistency between relative performance under training from scratch and mBART training. The best showings are by Alibaba and Huawei, followed by NRC and UA-Prompsit, with Microsoft still competitive. Other submissions score at least 1 BLEU points behind these.

Participants that also worked on sentence alignment of the provided document pairs were able to outperform the provided sentence pairs. The peak for these submissions shifts in most cases to the 7 million word subset. So, they were able to extract more useful sentence pairs. The best submissions for this setup comes from Bytedance. They outperform the provided sentence pairs and LASER scores by +3.8 BLEU (from 7.7 to 11.5) for Pashto from-scratch, +2.6 BLEU (from 10.3 to 12.9) for Pashto mBART, +4.3 BLEU (from 8.4 to 12.7) for Khmer from-scratch, +2.6 BLEU (from 12.9 to 15.5) for Khmer mBART.

6.2 Variance in the Evaluation

During the exploration of the evaluation protocol, we had some concerns about the stability of the BLEU scores obtained from training runs on a data set. This concern was reinforced by feedback from participants who did not match the baseline scores that we reported on the shared task web page.

To assess this, we executed three training runs for each subset of 5 million words selected from participant submissions. The resulting scores vary at most by 0.3 BLEU points for an identical training corpus, and differ most frequently just 0.1 BLEU point difference or are identical across all runs. The official reported results in Tables 9 and 10 are the average score across these three runs.

There may be higher differences for training on different hardware. We used a single NVidia GeForce GTX 1080ti GPU.

6.3 Average Sentence Length

Given the quality scores, subsets are selected by including the highest ranked sentence pairs until the total number of English words in these sentences reaches the specified size. So, if a quality scores prefers shorter sentences, more sentences are selected. It is not clear in general, all things being otherwise equal, if shorter or longer sentences are better for training machine translation systems.

What choices did the participants make in their quality scores? Table 11 and Table 12 show the number of sentences and the corresponding average number of words per sentence for the official subsets for all submissions. The average sentence length differs quite significantly, ranging from 12.3 to 29.0 words per sentence for Pashto, and 17.0 to 27.3 words per sentence for Khmer. Cross-referencing this against the effectiveness of the scores, methods that selected shorter sentences on average performed better.

In contrast to this, the average sentence length of submissions that also tackled sentence alignment is longer when compared to each participant's filtering-only submission.

6.4 Diversity of Submissions

The different submissions subselect different sentences, but how different are they?

Tables 13 and 14 give detailed statistics about how many sentence pairs the subsets of any two submissions for the two languages and two data conditions have in common.

The tables show for the 5 million word subset selected for each submission how many sentence pairs it contains (e.g., AFRL: 172,145), how many

Pashto	2 million				3 million				5 million				7 million			
	SCRATCH		MBART		SCRATCH		MBART		SCRATCH		MBART		SCRATCH		MBART	
	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST
AFRL	6.2	4.8	9.6	8.7	7.4	5.9	10.7	9.8	9.4	8.2	11.2	10.1	9.3	7.4	11.0	9.1
Alibaba	9.9	8.4	12.0	11.0	10.3	9.4	12.6	11.6	10.8	9.5	13.1	12.2	10.0	8.8	12.8	11.6
Edinburgh	9.6	8.5	11.4	10.8	10.3	8.5	11.6	10.5	10.0	8.3	11.3	10.5	9.6	7.7	11.6	9.7
Huawei	9.7	8.6	11.5	10.6	10.7	9.3	12.3	11.7	10.9	9.7	13.3	12.2	-	-	12.6	10.1
JHU-Kejriwal 0.8-5	8.0	6.7	10.5	9.9	9.1	7.7	10.8	10.1	9.7	7.8	11.3	10.2	9.4	7.2	11.5	10.0
JHU-Kejriwal 0.9-0	7.9	6.7	10.4	9.9	9.1	7.3	11.0	10.5	9.6	8.0	11.7	10.2	9.4	7.9	11.6	10.3
JHU-Kejriwal 0.9-5	8.2	6.9	10.2	9.8	9.1	7.6	11.0	10.2	9.6	7.7	11.6	10.4	9.4	7.5	11.4	9.9
JHU-Koerner dual-xent	7.7	6.0	9.4	9.4	8.9	7.6	11.3	10.7	9.8	8.0	10.9	10.3	9.5	7.4	11.3	9.5
JHU-Koerner laser-lm	9.1	7.7	11.0	10.4	9.7	8.4	11.4	10.3	9.9	8.3	11.1	10.0	9.5	7.8	11.0	9.6
LASER	9.1	7.6	10.9	10.2	9.4	7.8	11.0	10.3	9.7	7.7	11.4	10.3	9.7	8.2	11.1	9.8
Microsoft	9.4	8.5	11.2	10.6	10.5	9.2	11.7	11.1	10.1	8.5	12.8	11.6	9.9	8.5	11.7	10.3
NRC	8.7	7.5	9.3	8.8	10.2	8.6	11.4	10.7	10.5	8.9	12.9	12.0	9.8	8.5	12.5	11.5
UA-Prompsit	9.9	9.2	11.2	10.8	10.3	9.5	11.8	11.1	10.8	9.2	12.6	11.7	10.2	8.4	11.7	10.3
Alibaba alignment	9.1	8.8	11.7	10.9	10.8	10.0	12.2	11.8	11.7	10.4	13.2	12.4	11.2	9.8	12.8	11.8
Bytedance alignment	11.2	9.9	12.1	11.4	11.7	10.7	12.8	12.3	12.2	11.4	13.4	12.8	12.9	11.5	13.6	12.9
NRC alignment	11.4	10.1	12.2	11.1	12.0	10.5	12.7	11.7	11.8	10.5	13.4	12.4	11.1	10.0	13.1	11.9

Table 9: Results for Pashto: BLEU scores are reported for systems trained on 2, 3, 5 and 7 million word subsets of the data, subsampled based on the quality scores provided by the participants.

Khmer	2 million				3 million				5 million				7 million			
	SCRATCH		MBART		SCRATCH		MBART		SCRATCH		MBART		SCRATCH		MBART	
	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST	DEVT	TEST
Alibaba	8.2	9.3	10.3	12.5	8.7	10.3	10.9	12.9	8.9	11.0	11.5	14.0	7.8	10.1	10.6	13.2
Huawei	8.5	9.8	10.2	13.0	8.8	10.5	11.1	13.8	8.8	10.8	11.4	14.0	8.2	10.5	11.1	14.0
JHU Kejriwal 0.8-6	6.6	7.9	9.4	11.3	6.9	8.3	9.7	12.0	7.1	8.3	9.8	12.5	6.7	7.8	10.1	12.1
JHU-Kejriwal 0.8-5-filt	6.4	7.6	9.2	11.4	6.6	7.9	10.1	12.2	7.1	8.4	9.9	12.7	6.3	7.6	10.1	12.2
JHU-Kejriwal 0.8-5	5.5	6.1	6.0	8.1	5.9	6.8	6.8	7.9	6.5	7.4	10.0	12.1	6.5	7.8	9.8	12.2
LASER	6.4	7.7	9.2	10.9	7.0	8.0	9.7	12.0	7.1	8.4	10.5	12.9	6.7	8.6	10.5	12.6
Microsoft	7.2	8.7	9.7	11.9	8.0	9.3	10.3	12.5	7.8	9.3	11.2	13.3	7.8	9.7	11.1	13.7
NRC	7.7	9.5	10.4	12.6	8.5	10.6	10.5	13.4	8.7	10.8	11.2	13.7	8.4	10.3	11.2	13.8
UA-Prompsit	7.9	9.1	10.0	12.2	8.4	9.7	10.7	13.0	8.4	10.0	10.8	13.8	7.6	9.4	10.9	13.2
Bytedance alignment	9.3	11.2	11.2	14.0	9.8	11.8	11.7	14.6	10.5	12.7	12.3	14.9	10.3	12.5	12.7	15.5
NRC alignment	8.3	9.9	10.3	12.6	8.5	10.8	11.0	13.2	9.1	11.3	11.5	14.2	9.4	11.9	11.7	14.5

Table 10: Results for Khmer: BLEU scores are reported for systems trained on 2, 3, 5 and 7 million word subsets of the data, subsampled based on the quality scores provided by the participants.

Pashto	Sentences	Words/S
AFRL	172,145	29.0
Alibaba	375,507	13.3
Edinburgh	274,021	18.2
Huawei	383,554	13.0
JHU Kejriwal 0.8-5	208,922	23.9
JHU Kejriwal 0.9-0	257,060	19.5
JHU Kejriwal 0.9-5	209,059	23.9
JHU-Koerner laser-lm	225,750	22.1
JHU-Koerner dual-xent	205,346	24.3
LASER	225,725	22.2
Microsoft	238,612	21.0
NRC	405,330	12.3
UA-Prompsit	315,133	15.9
Alibaba alignment	222,539	22.5
Bytedance alignment	219,887	22.7
NRC alignment	244,622	20.4

Table 11: Number of sentences and the corresponding average sentence length (counting English words) for Pashto.

Khmer	Sentences	Words/S
Alibaba	258,044	19.4
Huawei	278,534	18.0
JHU Kejriwal 0.8-5	218,851	22.7
JHU Kejriwal 0.8-5-filt	191,864	26.0
JHU Kejriwal 0.8-6	182,126	27.3
LASER	240,978	20.7
Microsoft	256,762	19.4
NRC	293,414	17.0
UA-Prompsit	206,018	24.3
Bytedance alignment	169,492	29.5
NRC alignment	264,796	18.8

Table 12: Number of sentences and the corresponding average sentence length (counting English words) for Khmer.

sentence pairs are unique to this submission's subset (e.g., AFRL: 7.6% of the 172,145 sentence pairs) and how many are in common with other submission (e.g,, 59.2% of AFRL's subset are also in Alibaba's subset).

The leading submissions show mostly about 60% overlap, although there are also more similar submissons (Alibaba's and Huawai's share around 80% of sentence pairs). The alignment submissions tend to be quite different, not surprisingly.

7 Conclusion

We report on the findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. Ten participants used a variety of methods that gave quite different results, as measured by translation quality, optimal subset sizes, sentence length, etc. We hope that this task provides a benchmark for future research and improvements on this task.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandra Antonova and Alexey Misyurev. 2011. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Haluk Açarçiçek, Talha Çolakoğlu, pınar ece aktan hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Submission	Total	Unique	AFRL	Alibaba	Alibaba alignment	Bytedance alignment	Edinburgh	Huawei	JHU-Kejriwal 0.8-5	JHU-Kejriwal 0.9-0	JHU-Kejriwal 0.9-5	JHU-Koerner dual-xent	JHU-Koerner laser-lm	LASER	Microsoft	NRC	NRC alignment	UA-Prompsit
AFRL	172145	7.6%	-	59.2%	6.0%	19.2%	61.3%	57.2%	42.6%	43.3%	42.7%	53.8%	47.5%	47.5%	51.6%	60.5%	30.9%	60.2%
Alibaba	375507	4.6%	27.1%	-	3.9%	12.5%	45.9%	82.2%	36.6%	44.6%	36.8%	41.6%	36.8%	36.8%	42.8%	66.4%	31.1%	59.2%
Alibaba alignment	222539	62.8%	4.6%	6.6%	-	32.4%	5.9%	6.6%	3.5%	3.4%	3.5%	5.2%	3.9%	3.9%	5.1%	6.5%	7.9%	6.0%
Bytedance alignment	219887	45.2%	15.1%	21.3%	32.8%	-	19.1%	22.0%	14.0%	13.6%	14.1%	18.1%	15.0%	15.0%	18.1%	21.1%	19.4%	19.2%
Edinburgh	274021	5.6%	38.5%	62.9%	4.8%	15.4%	-	60.3%	50.4%	51.9%	50.6%	46.3%	58.7%	58.7%	53.4%	56.3%	32.4%	59.9%
Huawei	383554	3.0%	25.7%	80.5%	3.9%	12.6%	43.1%	-	36.4%	44.0%	36.6%	42.3%	34.1%	34.1%	42.1%	72.3%	31.2%	59.2%
JHU-Kejriwal 0.8-5	208922	0.2%	35.1%	65.8%	3.7%	14.7%	66.1%	66.8%	-	95.7%	98.5%	63.8%	74.8%	74.8%	64.8%	57.6%	31.1%	58.6%
JHU-Kejriwal 0.9-0	257060	2.0%	29.0%	65.2%	3.0%	11.7%	55.4%	65.6%	77.7%	-	78.0%	56.5%	68.3%	68.3%	59.0%	55.2%	26.4%	54.0%
JHU-Kejriwal 0.9-5	209059	0.0%	35.2%	66.2%	3.7%	14.8%	66.3%	67.2%	98.5%	95.9%	-	63.9%	75.1%	75.1%	65.1%	58.0%	31.4%	58.9%
JHU-Koerner dual-xent	205346	1.6%	45.1%	76.0%	5.6%	19.4%	61.8%	79.1%	64.9%	70.7%	65.1%	-	56.6%	56.6%	62.3%	68.5%	39.0%	68.1%
JHU-Koerner laser-lm	225750	0.0%	36.3%	61.2%	3.9%	14.6%	71.3%	58.0%	69.2%	77.8%	69.5%	51.5%	-	100.0%	77.3%	50.8%	28.5%	54.8%
LASER	225725	0.0%	36.3%	61.2%	3.9%	14.6%	71.3%	58.0%	69.2%	77.8%	69.5%	51.5%	100.0%	-	77.3%	50.8%	28.5%	54.8%
Microsoft	238612	0.9%	37.2%	67.4%	4.7%	16.7%	61.3%	67.7%	56.7%	63.5%	57.0%	53.6%	73.1%	73.1%	-	66.6%	32.5%	58.5%
NRC	405330	12.7%	25.7%	61.5%	3.6%	11.4%	38.1%	68.4%	29.7%	35.0%	29.9%	34.7%	28.3%	28.3%	39.2%	-	29.7%	50.1%
NRC alignment	244622	42.2%	21.7%	47.8%	7.2%	17.5%	36.3%	48.9%	26.6%	27.7%	26.8%	32.8%	26.3%	26.3%	31.7%	49.1%	-	41.7%
UA-Prompsit	315133	7.4%	32.9%	70.6%	4.3%	13.4%	52.1%	72.0%	38.8%	44.1%	39.1%	44.4%	39.2%	39.2%	44.3%	64.5%	32.4%	-

Table 13: **Overlap for Pashto.** For each submission, a row in the table lists the total number of sentence pairs, the ratio of unique sentence pairs that are in included in no other submission, and the ratio of sentence pairs shared with each of the other submissions.

Submission	Total	Unique	Alibaba	Bytedance	Huawei	JHU Kejriwal 0.8-6	JHU Kejriwal 0.8-5	JHU Kejriwal 0.8-5-filt	LASER	Microsoft	NRC	NRC alignment	UA-Prompsit
Alibaba	258044	13.7%	-	19.1%	68.7%	35.7%	35.3%	37.1%	41.9%	54.8%	59.6%	32.7%	41.1%
Bytedance alignment	169492	62.6%	29.0%	-	29.6%	18.7%	15.9%	18.8%	19.4%	24.9%	27.1%	24.0%	21.0%
Huawei	278534	11.4%	63.7%	18.0%	-	33.7%	38.1%	35.2%	41.8%	53.5%	58.1%	30.7%	42.4%
JHU Kejriwal 0.8-6	182126	0.2%	50.6%	17.4%	51.5%	-	78.0%	99.0%	91.1%	82.6%	47.2%	26.0%	31.3%
JHU-Kejriwal 0.8-5	218851	11.9%	41.7%	12.3%	48.5%	64.9%	-	68.8%	66.3%	60.8%	43.9%	20.3%	24.4%
JHU-Kejriwal 0.8-5-filt	191864	0.1%	49.9%	16.6%	51.1%	94.0%	78.5%	-	91.6%	82.8%	46.4%	25.3%	30.7%
LASER	240978	6.2%	44.8%	13.6%	48.3%	68.8%	60.2%	72.9%	-	82.1%	42.3%	22.4%	28.7%
Microsoft	256762	4.4%	55.1%	16.4%	58.0%	58.6%	51.9%	61.8%	77.0%	-	48.7%	26.9%	33.8%
NRC	293414	26.1%	52.4%	15.7%	55.1%	29.3%	32.8%	30.3%	34.8%	42.6%	-	32.1%	34.5%
NRC alignment	264796	58.9%	31.9%	15.3%	32.3%	17.9%	16.8%	18.4%	20.4%	26.1%	35.6%	-	21.6%
UA-Prompsit	206018	28.1%	51.5%	17.3%	57.4%	27.7%	25.9%	28.6%	33.6%	42.1%	49.2%	27.8%	-

Table 14: **Overlap for Khmer.** For each submission, a row in the table lists the total number of sentence pairs, the ratio of unique sentence pairs that are in included in no other submission, and the ratio of sentence pairs shared with each of the other submissions.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *CoRR*, abs/1711.02173.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Christian Buck and Philipp Koehn. 2016. Findings of the wmt 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (WMT)*.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Lei Cui, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Bilingual data cleaning for SMT using graph-based random walk. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–345, Sofia, Bulgaria. Association for Computational Linguistics.

Ahmed El-Kishky, Ahmed Chaudhary, Francisco Guzman, and Philipp Koehn. 2020a. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. *arXiv preprint arXiv:2002.00761*.

Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk. 2020b. Searching the web for cross-lingual parallel data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2417–2420.

Grant Erdmann and Jeremy Gwinnup. 2019. Quality and coverage: The afrl submission to the wmt19 parallel corpus filtering for low-resource conditions task. In *Proceedings of the Fourth Conference on Machine Translation (WMT)*.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant–prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

William A. Gale and Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1).

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine

translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Ankur Kejriwal and Philipp Koehn. 2020. An exploratory approach to the parallel corpus filtering shared task WMT20. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Felicia Koerner and Philipp Koehn. 2020. Dual conditional cross entropy scores and laser similarity scores for the WMT20 parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2054–2061, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo and Eric Joanis. 2020. Iteratively refined statistical sentence alignment and improved bilingual mappings of pretrained multilingual language model for identifying bette rparallel MT training data. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings*, volume 2499 of *Lecture Notes in Computer Science*. Springer.

Muhammad El Nokrashy, Amr Hendy, Mohamed Abdelghaffar, Mohamed Afify, Ahmed Tawfik, and Hany Hassan Awadalla. 2020. Score combination for improved parallel corpus filtering for low resource conditions. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alexandre Rafalovitch and Robert Dale. 2009. United Nations General Assembly resolutions: A six-language parallel corpus. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Spencer Rarrick, Chris Quirk, and Will Lewis. 2011. MT detection in web-scraped parallel corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 422–430. International Association for Machine Translation.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1374–1383.

Wolfgang Täger. 2011. The sentence-aligned european patent corpus. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 177–184.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 414–421. International Association for Machine Translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. *arXiv preprint arXiv:2004.14523*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1246.

Dániel Varga, Péter Halaácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. 2011. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Belgium, Brussels. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1411–1421. Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2940. Association for Computational Linguistics.

Runxin Xu, Zhuo Zhi, Jun Cao, Mingxuan Wang, and Lei Li. 2020. Volctrans parallel corpus filtering system for WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2015. The united nations parallel corpus v1.0. In *International Conference on Language Resources and Evaluation (LREC)*.