Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering

Philipp Koehn
Computer Science Department
Johns Hopkins University
Baltimore, Maryland, United States
phi@jhu.edu

Huda Khayrallah
Computer Science Department
Johns Hopkins University
Baltimore, Maryland, United States
huda@jhu.edu

Kenneth Heafield
School of Informatics
University of Edinburgh
Edinburgh, Scotland, European Union
kheafiel@inf.ed.ac.uk

Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
03690 St. Vicent del Raspeig, Spain
mlf@dlsi.ua.es

Abstract

We posed the shared task of assigning sentence-level quality scores for a very noisy corpus of sentence pairs crawled from the web, with the goal of sub-selecting 1% and 10% of high-quality data to be used to train machine translation systems. Seventeen participants from companies, national research labs, and universities participated in this task.

1 Introduction

Training corpora for machine translation come in varying degrees of quality. On the one extreme end they are carefully professionally translated specifically for this purpose which may have done under the instruction to provide fairly literal translations and adherence to sentence-by-sentence correspondences. The other extreme are sentence pairs extracted with fully automatic processes from indiscriminate crawling of the World Wide Web.

The Shared Task on Parallel Corpus Filtering targets the second extreme, although the methods developed for this data condition should also carry over to less noisy parallel corpora. In setting this task, we were motivated by our ongoing efforts to create large publicly available parallel corpora from web sources and the recognition that noisy parallel data is especially a concern for neural machine translation (Khayrallah and Koehn, 2018).

This paper gives an overview of the task, presents its results and provides some analysis.

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] (Skadiņš et al., 2014).

We are currently engaged in a large-scale effort to crawl text from the web. This work has been funded by Google Faculty Awards and is also currently funded by the European Union via the Connecting Europe Facility.[2] In 2016, we organized a shared task on document alignment as part of this effort (Buck and Koehn, 2016).

Acquiring parallel corpora from the web typically goes through the stages of identifying web sites with parallel text, downloading the pages of the web site, aligning document pairs, and aligning sentence pairs. A final stage of the processing pipeline filters out bad sentence pairs. These exist either because the original web site did not have any actual parallel data (garbage in, garbage out), or due to failures of earlier processing steps.

In 2016, a shared task on sentence pair filtering[3] was organized, albeit in the context of cleaning translation memories which tend to be cleaner that 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

[1] http://opus.lingfil.uu.se/
[2] http://www.paracrawl.eu/
[3] NLP4TM 2016: Shared task http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 726–739

https://doi.org/10.18653/v1/W18-64080

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 demonstrates 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 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. 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.

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) address noisy *training* data and focus on types of noise occurring in web-crawled corpora. They carried out a study how noise that occurs in crawled parallel text impacts statistical and neural 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 (say, about cooking recipes) that are just not relevant for the targeted domain (say, software manuals). Our task is focused on data quality that is relevant for all domains.

3 Task

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

Specifically, we provided a very noisy 1 billion word (English token count) German–English corpus crawled from the web by the Paracrawl project. We asked participants to subselect sentence pairs that amount to (a) 10 million words, and (b) 100 million words, counted on the English side. The quality of the resulting subsets was determined by the quality of a statistical machine translation (Moses, phrase-based) and a neural machine translation system (Marian) trained on this data. The quality of the machine translation system was measured by BLEU score on the (a) official WMT 2018 news translation test set and (b) other undisclosed test sets.

Note that the task addressed the challenge of data quality and not domain-relatedness of the data for a particular use case. Hence, we discouraged participants from subsampling the corpus for relevance to the news domain. Thus, we place more emphasis on the undisclosed test sets, although we report both scores.

Participants in the shared task submitted a file with quality scores, one per line, corresponding to the sentence pairs. The scores do not have to be meaningful, except that higher scores indicate better quality. The scores were uploaded to a Google Drive folder which remains publicly accessible.[4]

Evaluation of the quality scores was done by subsampling 10 million and 100 million word corpora based on these scores, training statistical and neural machine translation systems with the subsampled corpora, and evaluation translation quality on blind test sets using the BLEU score.

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
- a Moses configuration file to train and test a statistical machine translation system
- Marian scripts to train and test a neural machine translation system

[4] https://drive.google.com/drive/folders/1zZNPlAThm-Rnvxsy8rXzChC49bc0_TGO

Type of Noise	Count
Okay	23%
Misaligned sentences	41%
Third language	3%
Both English	10%
Both German	10%
Untranslated sentences	4%
Short segments (≤2 tokens)	1%
Short segments (3–5 tokens)	5%
Non-linguistic characters	2%

Table 1: Noise in the raw Paracrawl corpus.

- the test set from the WMT 2016 Shared Task on Machine Translation of News as development set
- the test set from the WMT 2017 Shared Task on Machine Translation of News as development test set

The web site for the shared task[5] provided detailed instructions on how to use these tools to replicate the official testing environment.

4 Data

4.1 Training Data

The provided raw parallel corpus is the outcome of a processing pipeline that aimed at high recall at the cost of precision, so it is very noisy. It exhibits noise of all kinds (wrong language in source and target, sentence pairs that are not translations of each other, bad language, incomplete of bad translations, etc.).

A cursory inspection of the corpus is given in Table 1. According to analysis by Khayrallah and Koehn (2018), only about 23% of the data is *okay*, but even that fraction may be flawed in some way. Consider the following sentence pairs that we did count as *okay* even though they contain mostly untranslated names and numbers.

DE: Anonym 2 24.03.2010 um 20:55 314 Kommentare
EN: Anonymous 2 2010-03-24 at 20:55 314 Comments
DE: << erste < zurück Seite 3 mehr letzte >>
EN: << first < prev. page 3 next last >>

It is an open question if such data is also harmful, merely irrelevant, or maybe even beneficial.

[5] http://www.statmt.org/wmt18/parallel-corpus-filtering.html

The raw corpus consists of a billion words of English, paired with German on the sentence level. It was deduplicated from a subset of the raw Paracrawl Release 1.

4.2 Provided Meta Information

The provided corpus file contains three items per line, separated by a TAB character:

- English sentence
- German sentence
- Hunalign score

The Hunalign scores were obtained from the sentence aligner (Varga et al., 2005). They may be a useful feature for sentence filtering, but they do not by themselves correlate strongly with sentence pair quality. None of the participants generally used this score.

Participant's systems may take the source of the data into account, e.g., by discounting sentence pairs that come from a web domain with generally low quality scores. To this end, we released the URL sources for each sentence pair as additional data set. Note that due to de-duplication a single sentence pair may have several URL pairs associated it, since it may appear on multiple web pages.

Participants were also allowed to use existing tools and external training data to build their filtering methods. Specifically, they were permitted to use the WMT 2018 news translation task data for German-English (without the Paracrawl parallel corpus) to train components of their method.

4.3 Test Sets

The goal of the task is to filter down to high-quality sentence pairs, but not to sentence pairs that are most fitting to a specific domain. During the submission period of the task, we only announced that we will use the official new translation test set from the WMT 2018 Shared Task of Machine Translation of News,[6] which was not released at that time yet.

In total, we used six test sets. For statistics see Table 2. Two of them were taken from existing evaluation campaigns, four were created for this shared task.

NEWSTEST2018 The test set from the WMT 2018 Shared Task of Machine Translation of

[6] http://www.statmt.org/wmt18/translation-task.html

News. It contains news stories that were either translated from German to English or from English to German.

IWSLT2017 The test set from the IWSLT 2017 evaluation campaign. It consists of transcripts of talks given at the TED conference. They cover generally accessible topics in the area of technology, entertainment, and design.

ACQUIS This test set was extracted from the Acquis Communtaire corpus, which is available on OPUS[7] (Tiedemann, 2012) (which was the source to create the subsequent 3 test sets). The test set consists of laws of the European Union that have to be incorporated into the national laws of the EU member countries. We only used sentences with 15 to 80 words, and removed any duplicate sentence pairs.

EMEA This test set was extracted from documents European Medicines Agency, which consist of public health announcements and descriptions of medications. We only used sentences with 20 to 80 words, and removed any duplicate sentence pairs.

GLOBALVOICES This test set was extracted from news stories posted and translated on Global Voices, an international and multilingual community of bloggers, journalists, translators, academics, and human rights activists. We selected several complete stories from this corpus.

KDE This test set was extracted from KDE4 localization files, which is open source software for Linux. We only used sentences with 15 to 80 words, and removed any duplicate sentence pairs.

For all the test sets, we checked for overlap with the training data, to prevent the possibility of having the test set being contained in the released noisy parallel data. We originally considered a test set based on the PHP documentation but removed it because that was contained in Paracrawl.

The official scoring of machine translation systems generated from the subsampled data sources is the average of the individual BLEU scores for each test set.

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

Test set	Sentences	English Words
NEWSTEST2018	2998	58,628
IWSLT2017	1138	18,162
ACQUIS	2862	98,624
EMEA	3000	93,071
GLOBALVOICES	3000	54,930
KDE	3000	109,716

Table 2: Statistics for the test sets used to evaluate the machine translation systems trained on the subsampled data sets. Word counts are obtained with `wc` on untokenized text.

5 Evaluation Protocol

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

5.1 Particpants

We received submissions from 17 different organizations. See Table 3 for the complete list of participants. The participant's organizations are quite diverse, with 3 participants from Spain, 3 participants from the United States, 2 participants from Germany, 1 participant each from Canada, Greece, China, Japan, France, Latvia, Estonia, United Kingdom, and Brazil. 9 of the participants are companies, 3 are national research organizations, and 5 were universities.

Each participant submitted up to 5 different sets of scores, resulting in a total of 44 different submissions that we scored.

5.2 Subset Selection

We provided to the participants a file containing one sentence pair per line. 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.

Using the score file, we selected subsets of a pre-defined size, defined by the number of English words. We chose the number of English words instead of German words, since the latter would allow selection of sentence pairs with very few German words and many English words which are beneficial for language model training but do not count much towards the German word total.

Subselecting 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 num-

Acronym	Participant and System Description Citation
AFRL	Air Force Research Lab, USA (Erdmann and Gwinnup, 2018)
Alibaba	Machine Intelligence Technology Lab, Alibaba Group, China (Lu et al., 2018)
ARC	Inst. for Language and Speech Proc./Athena RC, Greece (Papavassiliou et al., 2018)
U Tartu	University of Tartu, Estonia (Barbu and Barbu Mititelu, 2018)
JHU	Johns Hopkins University, USA (Khayrallah et al., 2018)
LMU	Ludwig Maximilian University of Munich, Germany (Hangya and Fraser, 2018)
MAJE	WebInterpret, Spain (Fomicheva and González-Rubio, 2018)
Microsoft	Microsoft Corp., USA (Junczys-Dowmunt, 2018)
NICT	National Inst. of Information and Communications Tech., Japan (Wang et al., 2018)
NRC	National Research Council, Canada (Littell et al., 2018; Lo et al., 2018)
Prompsit	Prompsit, Spain (Sánchez-Cartagena et al., 2018)
RWTH	Rheinland-Westphälische Technical University, Germany (Rossenbach et al., 2018)
Speechmatics	Speechmatics, United Kingdom (Ash et al., 2018)
Systran	Systran, France (Pham et al., 2018)
Tilde	Tilde, Latvia (Pinnis, 2018)
UTFPR	Federal University of Technology, Paranà, Brazil (Paetzold, 2018)
Vicomtech	Vicomtech, Spain (Azpeitia et al., 2018)

Table 3: Participants in the shared task.

ber of sentence pairs. Including all of them would yield a too large subset, excluding them yields a too small subset. Hence, we randomly included some of the sentence pairs to get the desired size in this case.

5.3 System Training

Given a selected subset of given size for a system submission, we built statistical (SMT) and neural machine translation (NMT) systems to evaluate the quality of the selected sentence pairs.

SMT For statistical machine translation, we used Moses (Koehn et al., 2007) with fairly basic settings, such as Good-Turing smoothing of phrase table probabilities, maximum phrase length of 5, maximum sentence length of 80, lexicalized reordering (*hier-mslr-bidirectional-fe*), fast-align for word alignment with *grow-diag-final-and* symmetrization, tuning with batch-MIRA, no operation sequence model, 5-gram language model trained on the English side of the subset with no additional data, and decoder beam size of 5,000.

NMT For neural machine translation, we used Marian (Junczys-Dowmunt et al., 2018). It uses the default settings of version 1.5, with 50,000 BPE operations, maximum sentence length of 100, layer normalization, dropout of 0.2 for RNN states, 0.1 for source embeddings and 0.1 for target embeddings, exponential smoothing, and decoding with beam size 12 and length normalization (1). Training a system for the 10 million word subset was limited to 20 epochs and took about 10 hours. Training a system for the 100 million word subset was limited to 10 epochs and took about 2 days.

Scores on the test sets were computed with `multi-bleu-detok.perl` included in Moses. We report case-insensitive scores.

6 Results

6.1 Core Results

The official results are reported in Table 4. The table contains the average BLEU score over all the 6 test sets for the 4 different setups

- statistical machine translation for 10 million word corpus
- statistical machine translation for 100 million word corpus
- neural machine translation for 10 million word corpus
- neural machine translation for 100 million word corpus

In the table, we highlight cells for the best scores for each of these settings, as well as scores that are close to it.

One striking observation is that the scores differ much more for the 10 million word subset than for

Participant	System	SMT 10M	SMT 100M	NMT 10M	NMT 100M
AFRL	afrl-cvg-large	21.9	25.2	13.8	30.2
AFRL	afrl-cvg-mix-meteor	23.4	25.3	27.1	30.3
AFRL	afrl-cvg-mix	22.5	25.2	19.8	30.1
AFRL	afrl-cvg-small	21.9	22.9	13.5	21.1
AFRL	afrl-cyn-mix	22.4	25.0	25.1	29.6
Alibaba	alibaba-div	24.1	26.4	27.6	31.9
Alibaba	alibaba	24.1	26.4	27.6	31.9
ARC	arc-11	22.7	26.1	19.8	31.3
ARC	arc-13	22.4	26.1	25.8	31.3
ARC	arc-9	21.9	26.0	24.0	31.3
U Tartu	tartu-hybrid-pipeline	22.3	25.7	25.2	30.6
JHU	zipporah-10000	22.6	25.8	25.3	30.2
JHU	zipporah	22.6	25.8	25.4	29.8
LMU	lmu-ds-lm-si	23.1	25.4	22.1	29.0
LMU	lmu-ds-lm	23.3	25.6	23.6	29.5
LMU	lmu-ds	23.3	25.5	23.6	29.5
LMU	lmu	21.5	25.6	23.0	30.5
MAJE	webinterpet	22.5	26.1	24.8	31.2
Microsoft	microsoft	24.4	26.5	28.6	32.1
NICT	nict	23.5	26.0	25.9	30.0
NRC	nrc-mono-bicov	21.0	26.2	23.1	31.6
NRC	nrc-mono	19.8	26.0	20.7	31.2
NRC	nrc-seve-bicov	22.1	26.2	25.3	31.7
NRC	nrc-yisi-bicov	23.9	26.4	27.4	31.9
NRC	nrc-yisi	23.5	26.4	26.5	31.8
Prompsit	prompsit-al	22.8	26.4	25.6	31.7
Prompsit	prompsit-lm	21.3	26.3	19.4	31.8
Prompsit	prompsit-lm-nota	20.1	26.2	19.3	31.7
Prompsit	prompsit-sat	22.9	26.3	26.1	31.7
RWTH	rwth-count	23.9	25.9	26.6	31.1
RWTH	rwth-nn	24.5	26.2	28.0	31.2
RWTH	rwth-nn-redundant	24.6	26.2	28.0	31.3
Speechmatics	balanced-scoring	23.8	25.8	27.9	31.0
Speechmatics	prime-neural	23.9	25.9	28.0	30.8
Speechmatics	purely-neural	18.1	25.8	18.0	30.0
Systran	systran	21.8	25.4	24.3	29.9
Tilde	tilde-max-rescored	23.0	26.0	26.6	31.2
Tilde	tilde-max	21.4	26.2	23.6	31.2
Tilde	tilde-isolated	21.0	25.9	22.6	30.8
UTFPR	utfpr-tree	17.6	20.7	11.4	11.9
UTFPR	uftpr-regression	20.8	22.4	21.8	22.2
UTFPR	utfpr-forest	13.2	17.0	6.6	6.2
Vicomtech	vicomtech	23.2	25.9	26.4	30.4
Vicomtech	vicomtech-ngsat	23.3	25.8	25.6	24.9

Table 4: Main results. BLEU scores (case-insensitive) are reported on the average of 6 test sets. Best performance on a test set is reported in bright green, scores within 0.5 BLEU points off the best in light green, and scores within 1 BLEU point off the best in light yellow.

the 100 million word subset. Scores also differ more for neural machine translation systems than for statistical machine translation systems.

For the 10 million word subset, there are only 2 submissions within 0.5 BLEU of the best system for statistical machine translation, and 0 for neural machine translation. For the 100 million word subset, there are 15 submissions within 0.5 BLEU of the best system for statistical machine translation, and 9 submissions within 0.5 for neural machine translation. Note that many of these submissions come from the same participants.

For both data sets, scores for neural machine translation are significantly higher. For the 10 million word subsets, the best NMT score is 28.6, while the best SMT score is 24.6. For the 100 million word subsets, the best NMT score is 32.1, while the best SMT score is 26.5. To be fair, statistical machine translation is typically trained with large monolingual corpora for language modelling that are essential for good performance.

6.2 Results by Test Set

Table 5 and 6 break out the results by each of the test sets, for statistical machine translation and neural machine translation, respectively.

The use of multiple test sets was motivated by the objective to discourage participants to filter sentence pairs for a specific domain, instead of filtering for general quality. Some participants used domain-specific data for training some elements of their filtering systems, such as monolingual news data sets to train language models but argued that these are broad domains that do not lead to domain over-fitting.

The results do not evoke the impression that some systems are doing better on some domains than others, at least not more than random variance would lead to expect. The closest test sets to the development sets are NEWSTEST2018, GLOBALVOICES, and maybe IWSLT2018. Only the 10 million word submissions *rwth-nn* and *rwth-nn-redundant* seem to do much better on these sets than others, relative to other submissions.

6.3 Additional Subset Sizes

Since we were interested in the shape of the curve of how different corpus sizes impact machine translation performance, we subselected additional subset size. Specifically, in addition to the 10 and 100 million word corpora, we also subselected 20, 30, 50, 80, 150, and 200 million words.

See Figure 1 for results for neural machine translation systems (also broken down by each individual test set) and Figure 2 for statistical machine translation systems. We only computed results for six systems due to the computational cost involved.

The scoring on additional subset sizes was not announced before the submission deadline for the shared task, so none of the participants optimized for these. In fact, some participants assigned the same low value for almost all sentence pairs that would be ignored when subselecting the 100 million word corpus. So, when subsampling larger corpora (150 and 200 million words, as we have done), the resulting system scores collapse.

The curves for neural machine translation system scores peak almost always at 100 million words, although also occasionally at 80 or 150 million words. Since we did not plot these curves when setting up the shared task, we cannot say if 100 million words is just a optimal value for this corpus or if participants overfitted their system to this value, although we would guess the first.

The performance between the submissions are quite similar on the different test sets. None of the submissions we show in the figures has overly optimized on the news test set.

7 Methods used by Participants

Not surprising due to the large number of submissions, many different approaches were explored for this task. However, most participants used a system using three components: (1) pre-filtering rules, (2) scoring functions for sentence pairs, and (3) a classifier that learned weights for feature functions.

Pre-filtering rules. Some of the training data can be discarded based on simple deterministic filtering rules. These may include rules to remove

- too short or too long sentences
- sentences that have too few words (tokens with letters instead of just special characters), either absolute or relative to the total number of tokens
- sentences whose average token length is too short or too long
- sentence pairs with mismatched lengths in terms of number of tokens

Participant	System	10M							100M						
		AVERAGE	NEWSTEST2018	IWSLT2017	ACQUIS	EMEA	GLOBALVOICES	KDE	**AVERAGE**	NEWSTEST2018	IWSLT2017	ACQUIS	EMEA	GLOBALVOICES	KDE
AFRL	afrl-cvg-large	21.9	26.1	18.9	18.3	26.0	20.0	22.1	25.2	29.9	22.2	21.5	29.7	22.6	25.5
AFRL	afrl-cvg-mix-meteor	23.4	27.7	20.0	20.6	26.8	21.1	24.0	25.3	29.9	22.3	21.5	29.9	22.7	25.6
AFRL	afrl-cvg-mix	22.5	26.5	19.4	20.2	25.5	20.4	22.8	25.2	29.8	22.3	21.5	29.7	22.6	25.4
AFRL	afrl-cvg-small	21.9	26.2	18.9	18.3	26.0	20.1	22.1	22.9	27.1	20.0	20.9	25.8	21.2	22.4
AFRL	afrl-cyn-mix	22.4	26.6	19.5	19.7	25.7	20.3	22.8	25.0	29.4	22.2	21.3	29.5	22.4	25.3
Alibaba	alibaba-div	24.1	29.1	22.2	20.6	26.7	22.0	24.2	26.4	31.2	22.9	22.4	31.2	24.0	26.8
Alibaba	alibaba	24.1	28.9	22.1	20.5	26.8	22.0	24.2	26.4	31.1	23.0	22.5	31.2	24.0	26.8
ARC	arc-11	22.7	26.9	18.9	19.3	27.2	20.4	23.3	26.1	30.8	22.7	22.4	30.9	23.5	26.6
ARC	arc-13	22.4	26.3	18.8	18.7	26.5	20.2	23.8	26.1	30.6	22.8	22.3	30.9	23.4	26.7
ARC	arc-9	21.9	26.0	18.1	18.5	25.8	20.0	23.2	26.0	30.7	22.7	22.1	30.9	23.4	26.3
U Tartu	tartu-hybrid-pipeline	22.3	26.8	19.5	18.7	24.8	20.6	23.5	25.7	30.4	22.3	21.9	30.5	23.1	26.1
JHU	zipporah-10000	22.6	26.3	20.2	19.9	24.7	20.3	24.3	25.8	30.2	22.6	22.1	29.9	23.4	26.4
JHU	zipporah	22.6	26.3	20.4	19.3	24.8	20.4	24.3	25.8	30.4	22.6	22.1	30.1	23.3	26.5
LMU	lmu-ds-lm-si	23.1	27.6	20.8	17.7	26.6	21.5	24.4	25.4	30.0	22.3	21.4	29.9	23.1	26.0
LMU	lmu-ds-lm	23.3	28.0	20.6	18.0	26.9	21.4	24.7	25.6	30.1	22.4	21.5	30.1	23.1	26.2
LMU	lmu-ds	23.3	28.0	20.6	18.0	27.0	21.5	24.6	25.5	30.0	22.3	21.2	30.2	23.2	26.1
LMU	lmu	21.5	25.4	19.7	15.3	25.3	20.0	23.1	25.6	30.3	22.4	21.0	30.4	23.3	26.2
MAJE	webinterpet	22.5	27.2	21.3	19.1	24.5	21.2	22.0	26.1	30.7	22.9	22.4	30.6	23.7	26.2
Microsoft	microsoft	24.4	29.5	21.6	19.7	28.7	22.5	24.7	26.5	31.4	23.2	22.3	31.4	23.9	26.9
NICT	nict	23.5	27.8	20.9	19.3	25.9	21.4	25.5	26.0	30.8	22.8	22.0	30.4	23.4	26.6
NRC	nrc-mono-bicov	21.0	25.1	17.9	16.6	24.2	20.0	22.1	26.2	31.1	22.8	22.4	31.1	23.8	26.2
NRC	nrc-mono	19.8	23.5	16.6	15.5	23.1	18.6	21.4	26.0	30.6	22.7	22.1	30.7	23.7	26.2
NRC	nrc-seve-bicov	22.1	26.0	18.6	18.8	27.9	20.1	21.4	26.2	31.1	22.8	22.2	31.2	23.7	26.5
NRC	nrc-yisi-bicov	23.9	28.7	21.3	19.7	26.4	22.1	25.2	26.4	31.4	22.8	22.4	31.1	23.8	26.9
NRC	nrc-yisi	23.5	28.0	21.1	19.3	26.0	21.8	25.0	26.4	31.0	23.2	22.5	30.8	23.9	26.8
Prompsit	prompsit-al	22.8	26.0	19.9	19.1	27.0	20.1	24.3	26.4	31.2	22.8	22.5	31.3	23.8	26.9
Prompsit	prompsit-lm	21.3	25.4	19.5	16.9	23.2	19.3	23.3	26.3	31.1	22.8	22.5	31.0	23.6	26.6
Prompsit	prompsit-lm-nota	20.1	24.9	19.4	15.9	19.7	18.6	21.9	26.2	31.0	22.9	22.2	30.9	23.5	26.5
Prompsit	prompsit-sat	22.9	27.0	19.0	19.0	27.4	20.6	24.6	26.3	31.0	22.8	22.5	31.1	23.6	26.9
RWTH	rwth-count	23.9	28.6	21.8	21.0	26.8	22.0	22.8	25.9	30.7	22.9	22.0	30.2	23.5	26.3
RWTH	rwth-nn	24.5	29.6	21.8	21.4	28.0	22.7	23.8	26.2	30.8	23.2	22.2	30.9	23.4	26.6
RWTH	rwth-nn-redundant	24.6	29.6	21.8	21.4	28.1	22.6	23.9	26.2	30.8	23.1	22.1	30.9	23.6	26.8
Speechmatics	balanced-scoring	23.8	28.2	21.0	19.7	27.6	21.5	24.7	25.8	30.3	22.6	22.0	30.5	23.3	26.3
Speechmatics	prime-neural	23.9	28.2	20.5	19.6	28.3	21.4	25.3	25.9	30.4	22.5	21.9	30.7	23.3	26.4
Speechmatics	purely-neural	18.1	20.4	15.1	13.6	22.2	16.3	21.0	25.8	30.3	22.5	21.9	30.6	23.2	26.2
Systran	systran	21.8	25.4	19.4	16.7	25.7	19.9	23.9	25.4	30.0	22.3	21.5	30.1	22.7	26.1
Tilde	tilde-max-rescored	23.0	27.3	19.8	18.3	27.7	21.0	24.1	26.0	30.6	22.8	21.9	30.9	23.4	26.2
Tilde	tilde-max	21.4	25.0	18.2	16.6	25.6	19.7	23.6	26.2	30.8	22.8	22.1	31.1	23.6	26.6
Tilde	tilde-isolated	21.0	24.3	17.4	16.2	25.1	19.4	23.5	25.9	30.6	22.5	22.0	30.8	23.2	26.5
UTFPR	utfpr-tree	17.6	20.5	14.7	14.0	21.0	16.1	19.0	20.7	23.7	18.2	17.1	23.9	18.9	22.3
UTFPR	uftpr-regression	20.8	25.1	18.6	16.2	23.7	19.1	22.2	22.4	26.5	20.2	17.4	26.0	20.5	23.5
UTFPR	utfpr-forest	13.2	14.9	9.9	10.6	16.8	12.1	15.0	17.0	18.7	14.4	13.9	20.4	15.2	19.2
Vicomtech	vicomtech	23.2	27.5	20.4	19.3	26.5	21.2	24.6	25.9	30.5	22.5	22.2	30.3	23.4	26.6
Vicomtech	vicomtech-ngsat	23.3	27.5	19.8	19.3	26.8	21.1	25.1	25.8	30.2	22.4	22.1	30.0	23.4	26.7

Table 5: Detailed results for SMT performance. BLEU scores (case-insensitive) are reported on all the 6 test sets. The best performance on a test set is reported in bright green, scores within 0.5 BLEU points off the best in light green, and scores within 1 BLEU point off the best in light yellow.

Participant	System	10M							100M						
		AVERAGE	NEWSTEST2018	IWSLT2017	ACQUIS	EMEA	GLOBALVOICES	KDE	**AVERAGE**	NEWSTEST2018	IWSLT2017	ACQUIS	EMEA	GLOBALVOICES	KDE
AFRL	afrl-cvg-large	13.8	11.2	6.1	15.5	23.8	8.9	17.4	30.2	37.0	26.3	26.5	35.1	28.0	28.2
AFRL	afrl-cvg-mix-meteor	27.1	33.4	23.3	25.6	29.9	25.4	25.0	30.3	37.4	26.0	26.6	35.2	28.1	28.4
AFRL	afrl-cvg-mix	19.8	19.7	10.9	23.9	26.9	14.8	22.7	30.1	37.4	26.1	26.4	34.8	28.1	28.1
AFRL	afrl-cvg-small	13.5	10.9	5.6	15.3	23.7	8.5	16.9	21.1	23.3	16.8	22.9	26.2	19.0	18.1
AFRL	afrl-cyn-mix	25.1	29.2	21.4	24.2	29.0	22.7	24.0	29.6	36.2	25.1	26.2	35.0	27.4	27.7
Alibaba	alibaba-div	27.6	35.0	25.2	24.1	29.8	25.8	25.7	31.9	39.5	27.1	28.4	36.7	29.1	30.7
Alibaba	alibaba	27.6	35.2	25.6	24.2	29.4	25.6	25.5	31.9	39.7	27.3	28.4	36.4	29.1	30.6
ARC	arc-11	19.8	20.3	11.4	21.1	27.4	14.7	23.7	31.3	39.0	26.6	27.8	35.9	28.3	30.4
ARC	arc-13	25.8	31.3	21.2	22.9	30.2	23.4	25.7	31.3	39.0	26.6	27.6	36.0	28.2	30.6
ARC	arc-9	24.0	30.4	20.2	21.5	28.8	22.9	20.0	31.3	39.0	26.5	27.6	35.8	28.3	30.7
U Tartu	tartu-hybrid-pipeline	25.2	31.6	21.8	21.8	28.1	24.0	23.6	30.6	38.2	26.2	27.5	35.8	28.1	27.8
JHU	zipporah-10000	25.3	31.4	23.1	22.8	26.3	24.0	24.3	30.2	36.8	24.2	27.6	35.4	27.7	29.3
JHU	zipporah	25.4	31.3	23.1	22.5	26.6	24.4	24.5	29.8	36.4	23.2	27.3	35.1	27.3	29.2
LMU	lmu-ds-lm-si	22.1	31.2	22.0	16.8	24.0	23.8	14.7	29.0	36.2	25.7	24.4	33.2	27.5	27.1
LMU	lmu-ds-lm	23.6	31.9	22.4	18.5	27.0	24.6	17.5	29.5	37.0	25.5	25.2	33.5	27.5	28.2
LMU	lmu-ds	23.6	31.8	22.1	18.4	27.1	24.5	17.9	29.5	36.7	25.5	25.2	34.1	27.7	27.9
LMU	lmu	23.0	28.8	21.1	16.0	27.0	23.3	21.6	30.5	37.8	25.9	25.8	35.6	28.5	29.6
MAJE	webinterpet	24.8	32.4	24.8	22.6	24.6	24.3	20.2	31.2	38.7	26.9	27.9	35.6	28.9	29.2
Microsoft	microsoft	28.6	35.7	25.1	23.7	32.7	26.7	27.8	32.1	39.9	27.4	28.3	36.7	29.3	30.8
NICT	nict	25.9	32.9	23.7	21.7	27.6	25.1	24.6	30.0	37.3	25.8	26.1	34.1	27.6	29.2
NRC	nrc-mono-bicov	23.1	27.9	19.3	19.0	26.4	22.0	23.7	31.6	38.9	27.1	28.1	36.0	28.9	30.4
NRC	nrc-mono	20.7	25.0	17.2	16.6	23.8	19.8	21.9	31.2	38.4	26.8	27.9	35.7	28.0	30.3
NRC	nrc-seve-bicov	25.3	30.3	21.5	22.6	31.7	23.1	22.9	31.7	39.4	27.1	28.3	36.3	28.9	30.1
NRC	nrc-yisi-bicov	27.4	33.9	24.4	23.2	29.8	25.4	27.8	31.9	39.6	26.9	28.4	36.6	29.1	30.7
NRC	nrc-yisi	26.5	32.7	23.9	22.2	28.6	24.8	26.8	31.8	39.3	27.1	27.9	36.3	29.0	30.9
Prompsit	prompsit-al	25.6	31.1	22.4	21.8	30.0	23.2	24.9	31.7	39.4	27.0	28.1	36.6	28.6	30.6
Prompsit	prompsit-lm	19.4	26.5	20.2	18.9	17.4	19.5	14.2	31.8	39.5	27.3	28.4	36.6	28.9	30.4
Prompsit	prompsit-lm-nota	19.3	26.1	20.0	18.8	17.3	19.8	14.0	31.7	39.8	26.7	28.3	36.4	29.1	30.0
Prompsit	prompsit-sat	26.1	31.6	20.8	22.1	31.2	23.7	26.8	31.7	39.2	26.7	28.2	36.4	28.7	30.8
RWTH	rwth-count	26.6	34.8	25.0	24.4	27.7	25.9	22.1	31.1	38.6	26.9	27.5	35.4	29.0	28.9
RWTH	rwth-nn	28.0	36.0	25.2	25.2	31.1	26.7	23.7	31.2	38.8	26.7	27.7	36.1	28.7	29.3
RWTH	rwth-nn-redundant	28.0	36.0	25.2	25.3	31.1	26.6	23.9	31.3	39.2	26.5	27.4	36.3	28.7	29.6
Speechmatics	balanced-scoring	27.9	34.0	24.6	24.7	30.9	25.0	28.0	31.0	37.8	26.5	27.9	35.4	28.2	30.1
Speechmatics	prime-neural	28.0	34.7	24.1	24.4	31.4	24.9	28.2	30.8	37.4	26.5	27.8	35.1	28.2	30.1
Speechmatics	purely-neural	18.0	21.8	15.6	13.1	21.3	17.6	18.4	30.0	35.2	25.8	26.9	35.1	27.4	29.8
Systran	systran	24.3	29.6	21.3	19.1	28.3	23.0	24.6	29.9	36.3	25.1	26.2	35.1	26.9	29.8
Tilde	tilde-max-rescored	26.6	32.4	22.1	22.1	31.3	24.4	27.1	31.2	38.6	26.8	27.5	36.6	28.2	29.6
Tilde	tilde-max	23.6	28.0	19.5	17.9	28.5	22.3	25.1	31.2	38.6	26.4	27.3	36.3	28.6	30.3
Tilde	tilde-isolated	22.6	26.6	18.9	16.8	27.4	21.8	24.2	30.8	38.0	25.8	26.7	35.7	27.9	30.4
UTFPR	utfpr-tree	11.4	13.2	7.8	10.4	17.5	9.9	9.8	11.9	10.5	6.8	11.7	18.2	10.1	13.9
UTFPR	uftpr-regression	21.8	27.2	18.5	18.6	24.9	19.2	22.1	22.2	25.0	16.7	19.1	28.8	19.7	24.1
UTFPR	utfpr-forest	6.6	6.5	2.9	4.2	11.5	5.9	8.3	6.2	4.7	2.1	3.5	12.3	5.0	9.3
Vicomtech	vicomtech	26.4	32.3	22.6	22.6	29.0	24.3	27.4	30.4	37.1	26.4	26.8	34.5	27.7	29.9
Vicomtech	vicomtech-ngsat	25.6	31.2	21.8	20.7	29.1	23.5	27.6	24.9	27.2	22.4	23.1	26.9	22.9	26.8

Table 6: Detailed results for NMT performance. BLEU scores (case-insensitive) are reported on all the 6 test sets. The best performance on a test set is reported in bright green, scores within 0.5 BLEU points off the best in light green, and scores within 1 BLEU point off the best in light yellow.

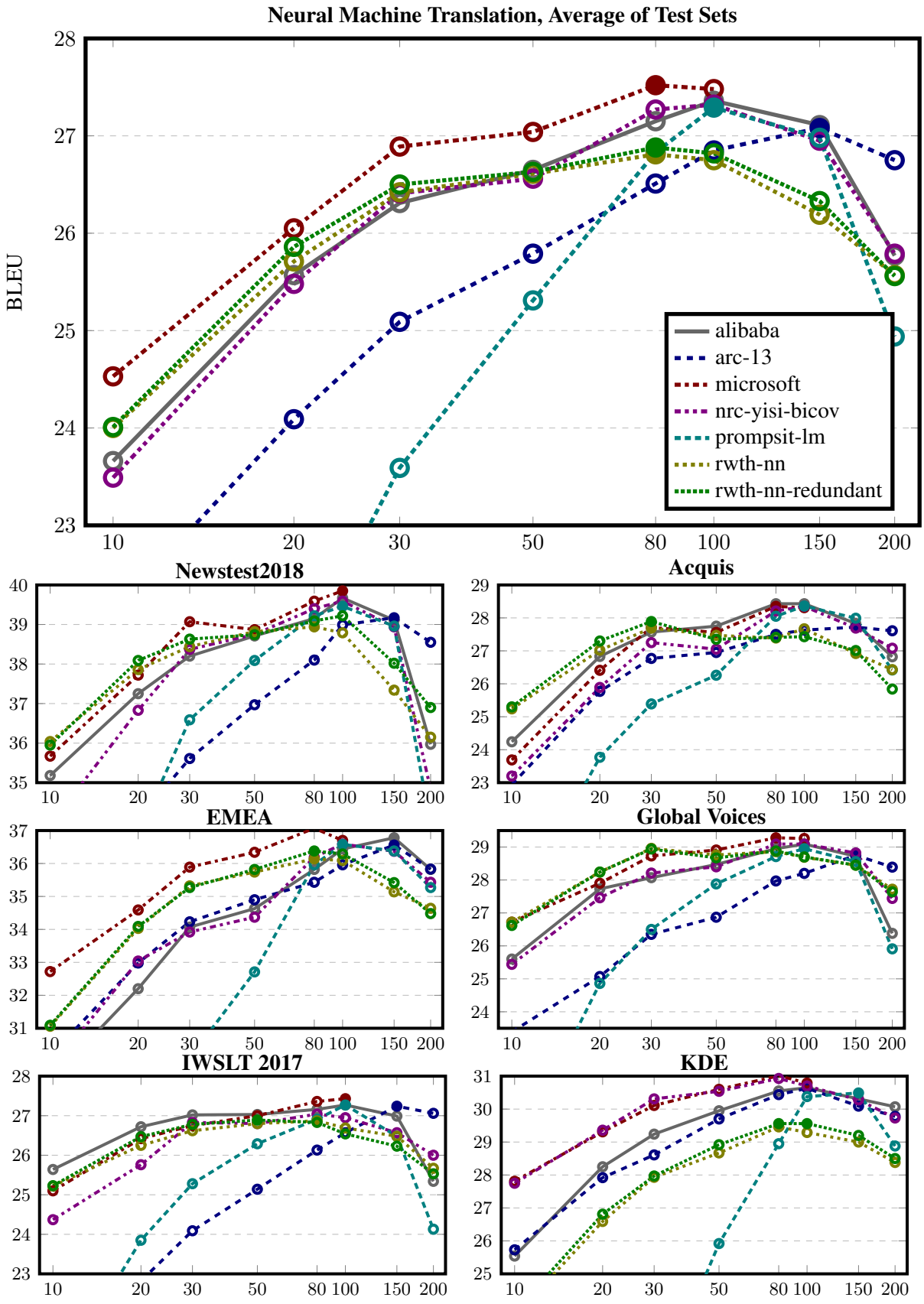

Figure 1: Additional corpus sizes, with breakdown by individual test set for some high-performing submissions. The charts plot BLEU scores against the size of the subselected corpus (in millions of words). The curves peak around 100 million words.

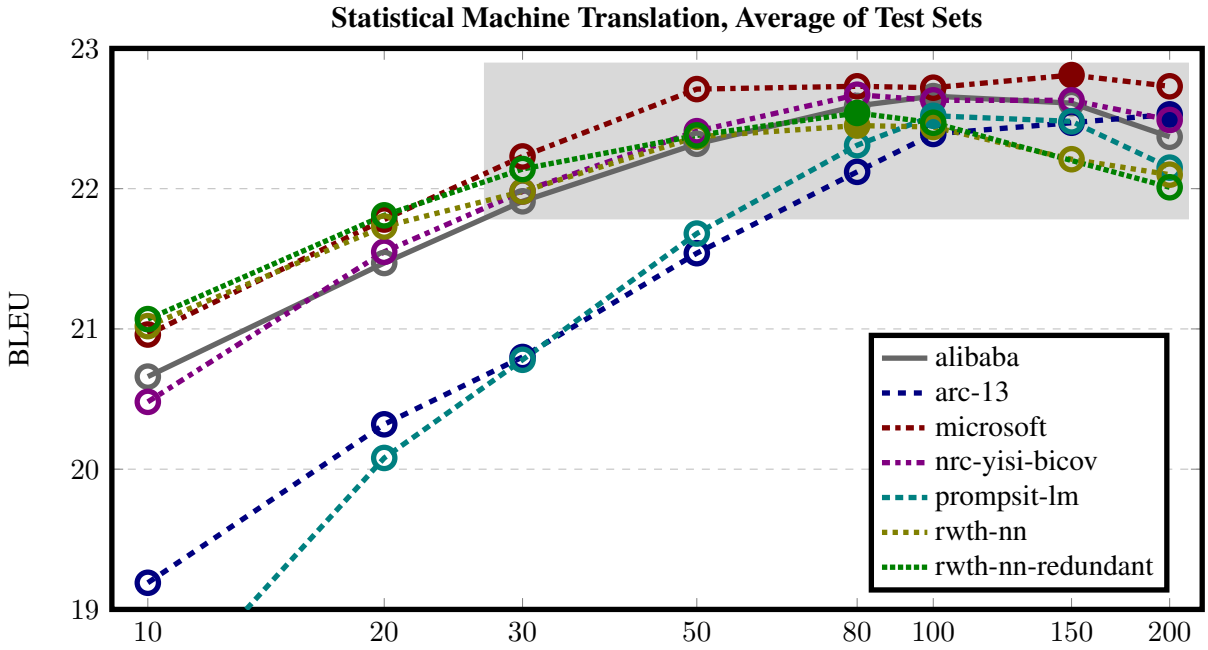

Figure 2: Version of Figure 1 for statistical machine translation systems built from the subselected data. Note that the curves are flatter, and the several systems score in a narrow band of 1 BLEU point across a wide range of corpus sizes (30-200 million words), indicated in grey.

- sentence pairs where names, numbers, email addresses, URLs do not match between both sides
- sentence pairs that are too similar, indicating simple copying instead of translating
- sentences where language identifier do not detect the required language

Scoring functions. Sentence pairs that pass the pre-filtering stage are assessed with scoring functions which provide scores that hopefully correlate with quality of sentence pairs. Participants used a variety of such scoring functions, including

- n-gram or neural language models on clean data
- language models trained on the provided raw data as contrast
- neural translation models
- bag-of-words lexical translation probabilities

Note that the raw scores provided by these models may be also refined in several ways. For instance, we may desire that the language model perplexities of a German sentence and its paired English sentence are similar. Or, we may contrast the translation model score for a sentence and its given paired sentence with the translation model score for the sentence and its best translation according to the model.

Learning weights for scoring functions. Given a large number of scoring functions, simply averaging their resulting scores may be inadequate. Learning weights to optimize machine translation system quality is computationally intractable due to the high cost of training these systems to evaluate different weight settings. A few participants used instead a classifier that learns how to distinguish between good and bad sentence pairs. Good sentence pairs are selected from existing high-quality parallel corpora, while bad sentence pairs are either synthesized by scrambling good sentence pairs or by using the raw crawled data.

Some participants made a distinction between unsupervised methods that did not use existing parallel corpora to train parts of the system, and supervise methods that did. Unsupervised methods have the advantage that they can be readily deployed for language pairs for which no seed parallel corpora exist.

Acknowledgements

The shared task was supported by a Google Faculty Research Award to Johns Hopkins University and by the European Union through the Connected Europe Facility project *Provision of Web-*

Scale Parallel Corpora for Official European Languages (Paracrawl).

References

Antonova, A. and Misyurev, A. (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.

Ash, T., Francis, R., and Williams, W. (2018). The speechmatics parallel corpus filtering system for wmt18. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Axelrod, A., He, X., and Gao, J. (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.

Azpeitia, A., Etchegoyhen, T., and Martínez garcia, E. (2018). Stacc, oov density and n-gram saturation: Vicomtech's participation in the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Barbu, E. and Barbu Mititelu, V. (2018). A hybrid pipeline of rules and machine learning to filter web-crawled parallel corpora. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

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

Buck, C. and Koehn, P. (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.

Carpuat, M., Vyas, Y., and Niu, X. (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.

Cui, L., Zhang, D., Liu, S., Li, M., and Zhou, M. (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.

Erdmann, G. and Gwinnup, J. (2018). Coverage and cynicism: The afrl submission to the wmt 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Fomicheva, M. and González-Rubio, J. (2018). Maje submission to the wmt2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Hangya, V. and Fraser, A. (2018). An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Junczys-Dowmunt, M. (2018). Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Khayrallah, H. and Koehn, P. (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.

Khayrallah, H., Xu, H., and Koehn, P. (2018). The jhu parallel corpus filtering systems for wmt 2018. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Littell, P., Larkin, S., Stewart, D., Simard, M., Goutte, C., and Lo, C.-k. (2018). Measuring sentence parallelism using mahalanobis distances: The nrc unsupervised submissions to the wmt18 parallel corpus

filtering shared task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Lo, C.-k., Simard, M., Stewart, D., Larkin, S., Goutte, C., and Littell, P. (2018). Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The nrc supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Lu, J., Lv, X., Shi, Y., and Chen, B. (2018). Alibaba submission to the wmt18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Paetzold, G. (2018). Utfpr at wmt 2018: Minimalistic supervised corpora filtering for machine translation. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Papavassiliou, V., Sofianopoulos, S., Prokopidis, P., and Piperidis, S. (2018). The ilsp/arc submission to the wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Pham, M. Q., Crego, J., and Senellart, J. (2018). Systran participation to the wmt2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Pinnis, M. (2018). Tilde's parallel corpus filtering methods for wmt 2018. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Rafalovitch, A. and Dale, R. (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.

Rarrick, S., Quirk, C., and Lewis, W. (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.

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

Rossenbach, N., Rosendahl, J., Kim, Y., Graça, M., Gokrani, A., and Ney, H. (2018). The rwth aachen university filtering system for the wmt 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sánchez-Cartagena, V. M., Bañón, M., Ortiz Rojas, S., and Ramírez, G. (2018). Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Täger, W. (2011). The sentence-aligned european patent corpus. In Forcada, M. L., Depraetere, H., and Vandeghinste, V., editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 177–184.

Taghipour, K., Khadivi, S., and Xu, J. (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.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

van der Wees, M., Bisazza, A., and Monz, C. (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.

Varga, D., Halaácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. (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.

Wang, R., Marie, B., Utiyama, M., and Sumita, E. (2018). Nict's corpus filtering systems for the wmt18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels. Association for Computational Linguistics.

Xu, H. and Koehn, P. (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.

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