CHIA: CHoosing Instances to Annotate for Machine Translation

Rajat Bhatnagar[*] **Ananya Ganesh**[*] **Katharina Kann**
{rajat.bhatnagar, ananya.ganesh, katharina.kann}@colorado.edu
University of Colorado Boulder

Abstract

Neural machine translation (MT) systems have been shown to perform poorly on low-resource language pairs, for which large-scale parallel data is unavailable. Making the data annotation process faster and cheaper is therefore important to ensure equitable access to MT systems. To make optimal use of a limited annotation budget, we present CHIA (*choosing instances to annotate*), a method for selecting instances to annotate for machine translation. Using an existing multi-way parallel dataset of high-resource languages, we first identify instances, based on model training dynamics, that are most informative for training MT models for high-resource languages. We find that there are cross-lingual commonalities in instances that are useful for MT model training, which we use to identify instances that will be useful to train models on a new target language. Evaluating on 20 languages from two corpora, we show that training on instances selected using our method provides an average performance improvement of 1.59 BLEU over training on randomly selected instances of the same size.

1 Introduction

Machine translation (MT) systems have been widely adopted into daily use and facilitate easy communication and access to information. Advances in neural MT have enabled systems to approach human performance (Hassan et al., 2018; Popel et al., 2020). However, such high-performing MT systems are only available for a small subset of the world's languages as they require large training corpora (Mueller et al., 2020; Koehn and Knowles, 2017). While unsupervised methods (Lample et al., 2018; Artetxe et al., 2018) that use limited or no parallel data are effective for many languages, they perform poorly on low-resource language pairs (Guzmán et al., 2019), and are outperformed by supervised methods (Kim et al., 2020).

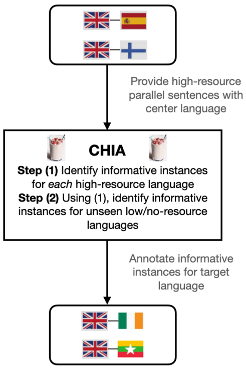
Provide high-resource parallel sentences with center language
CHIA
Step (1) Identify informative instances for *each* high-resource language
Step (2) Using (1), identify informative instances for unseen low/no-resource languages
Annotate informative instances for target language

Figure 1: Overview of our method, CHIA. Given data in high-resource languages such as English and Spanish, CHIA selects data to annotate for low-resource languages such as Irish or Burmese.

To ensure that advances in MT benefit all communities and users equally, we need efficient ways to collect parallel data. For high-resource languages with a large number of speakers, parallel data sources exist (Koehn, 2005; Agić and Vulić, 2019; McCarthy et al., 2020), and crowdsourcing has proved cheap and effective (Post et al., 2012). However, for low-resource languages, collecting sufficient parallel sentences is harder, as bilingual translators may be difficult to find or expensive. The amount of instances which can be manually translated for use during model development is thus limited either by time or monetary constraints.

Here, we explore how existing or easily obtainable parallel sentences in high-resource languages, e.g., English and Spanish, can help us construct high-quality datasets for a target language with low or no resources, under a limited annotation budget. We present CHIA (*choosing instances to annotate*), which requires a multi-way parallel dataset and automatically identifies those instances which will result in the strongest MT system if translated to construct a training set for a new language.

[*] Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7328 - 7344
December 7-11, 2022 ©2022 Association for Computational Linguistics

CHIA is based on cross-lingual information: First, we identify the most effective instances to train MT systems between the *center language* – the language from which we wish to translate into a low-resource language – and multiple high-resource languages. For this, we utilize MT model training dynamics to identify examples that help a model learn, as proposed by Swayamdipta et al. (2020). We extend their method, originally proposed for classification tasks, to sequence-to-sequence tasks. Second, we use the intersection between the sets of informative instances for different language pairs to determine which instances will be most beneficial for training an MT system for a new language pair, cf. Figure 1.

We perform experiments on two multi-way parallel datasets, the Europarl corpus (Koehn, 2005) and the JHU Bible corpus (McCarthy et al., 2020). Our language pairs consist of English – our center language – and 15 simulated low-resource languages as well as 5 truly low-resource languages. We show that, on average, training on examples selected by CHIA results in gains of 1.59 BLEU as compared to training on randomly selected instances. We further examine the characteristics of the selected training examples, and find that CHIA does not rely on simple properties such as sentence length or number of unique words.

CHIA is based on the two contributions we present in this paper: 1) a method to identify the most useful training instances for sequence-to-sequence tasks, and 2) an empirical demonstration that this method can be used to identify examples, which will be beneficial for a new low-resource language, based on a set of high-resource languages.

2 Method: CHIA

2.1 Background: Dataset Cartography

Individual instances in a training set have varying impact on a model's learning behavior (Lewis and Gale, 1994; Cohn et al., 1995). To identify the examples that contribute the most to the training process for classification tasks, Swayamdipta et al. (2020) propose to look at two metrics for each instance i, confidence c_i and variability v_i:

$$c_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^e}(y_i^*|x_i) \tag{1}$$

and

$$v_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta^e}(y_i^*|x_i) - c_i)^2}{E}}, \tag{2}$$

with E being the number of training epochs, θ^e being the model parameters at the end of epoch e, y_i^* being the true label in the training set, and x_i being the respective input. Thus, c_i corresponds to the average probability of the true label of an example over epochs, and v_i is the confidence's standard deviation.

Based on confidence and variability, Swayamdipta et al. (2020) partition the training set into three regions: instances with high confidence and low variability are designated as easy-to-learn, instances with high variability are designated as ambiguous, and instances with low confidence and low variability are designated as hard-to-learn. They show that instances in the ambiguous region contribute the most to the model's ability to generalize out of the training distribution, i.e., ambiguous instances are the most effective training examples.

2.2 Computing MT Training Dynamics

Our first contribution is a generalization of Swayamdipta et al. (2020)'s method for classification tasks to sequence-to-sequence tasks like machine translation. Importantly, we have more than one gold label per example: each ground-truth translation consists of a *sequence* of gold labels y_i^* of length T. We modify Equation 1 as follows to compute the confidence c_{S_i} for a gold sequence:

$$c_{S_i} = \frac{1}{ET}\sum_{e=1}^{E}\sum_{t=1}^{T} p_{\theta^e}(y_{it}^*|x_i) \tag{3}$$

We then compute the variance v_{S_i} as:

$$v_{S_i} = \sqrt{\frac{\sum_{e=1}^{E}(\frac{1}{T}\sum_{t=1}^{T} p_{\theta^e}(y_{it}^*|x_i) - c_{S_i})^2}{E}} \tag{4}$$

Based on v_S and c_S we group training instances into three sets: hard-to-learn examples $\mathcal{H}$, easy-to-learn $\mathcal{E}$, and ambiguous $\mathcal{A}$.

2.3 Selecting Instances for New Languages

The above method for beneficial-instance detection requires that a model has already been trained on all available data, i.e., it chooses a subset of already existing data, and cannot be used to select *new*

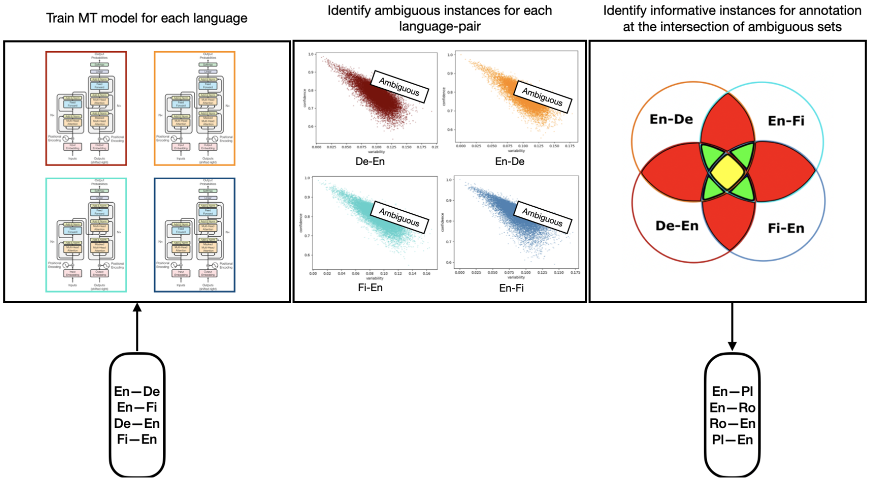
Train MT model for each language
Identify ambiguous instances for each language-pair
Identify informative instances for annotation at the intersection of ambiguous sets
Ambiguous
Ambiguous
Ambiguous
Ambiguous
De-En
En-De
Fi-En
En-Fi
En-De
En-Fi
De-En
Fi-En
En—De
En—Fi
De—En
Fi—En
En—Pl
En—Ro
Ro—En
Pl—En

Figure 2: CHIA steps: (1a) Train MT models for each high-resource language (Transformer model figure from Vaswani et al. (2017)). (1b) Using training dynamics, identify high variability or ambiguous instances for each language pair. For readability, we plot the probability here as confidence, whereas we use loss in our experiments. (2) Using all ambiguous sets, find instances at the intersection that are most informative. We first select instances in the yellow region, then the green, and finally the red.

instances for annotation. Thus, our second contribution is the second step of CHIA, which chooses instances to *annotate* (i.e., *translate*) for a new language. Importantly, CHIA does not rely on any existing parallel data between the two languages.

CHIA assumes that the following is given: (1) an n-way parallel corpus, (2) a center language L_c, which is one of the n languages in the corpus and from which we want to translate into a new language, and (3) separate models for translating between the center language and all $n-1$ other languages together with their training dynamics. The latter enables us to compute sets $\mathcal{A}_{L_sL_t}$, which contain the ambiguous instances for training an MT system between a source language L_s and a target language L_t. We assume that the center language is either the source or target language in all cases[1] .

Once we have the ambiguous sets corresponding to each language pair, we select instances that lie at the intersection of multiple ambiguous sets. Specifically, for each instance i in the n-way parallel dataset, we count the number of language pairs l_i for which $i \in \mathcal{A}_{L_sL_t}$, where $0 \leq l_i \leq n$. We rank each instance i by l_i, and select the top k instances, where k is the desired size of the dataset for a new target language. The selected instances in the source language can then be used for constructing a parallel dataset in the target language. In practice, a human translator would manually translate these examples, whereas in our experiments, we make use of existing gold translations in the target language.

3 Data and Languages

We use two corpora from different domains: the Europarl corpus (Koehn, 2005) and the JHU Bible corpus (McCarthy et al., 2020).

3.1 Europarl

The Europarl corpus covers parliamentary proceedings. It contains multi-way parallel sentences in 21 European languages, which we filter for those sentences that are parallel between all languages. Our final dataset contains 180,000 sentences per language. To explore the effectiveness of CHIA for different dataset sizes, we create subsets of our data with 20k, 40k, 80k, and 160k sentences.

Seen languages. We create a set of 10 seen language pairs, which we use to compute training dynamics and to identify ambiguous instances. Our seen language pairs consist of our center language English paired in two directions with Greek, German, Finnish, Spanish, and Slovak.

Unseen languages. We create a set of 30 unseen or evaluation language pairs, consisting of English, our center language, paired in two directions with Bulgarian, Czech, Danish, Estonian, French, Hungarian, Italian, Lithuanian, Latvian, Dutch, Polish, Portuguese, Romanian, Slovene, and Swedish.

[1] We set this restriction to limit the number of models we need to train, but an exploration of the effectiveness of other combinations could be interesting future work.

Validation and test sets. We randomly select 7.8k and 18.2k English sentences to be our validation and test set, respectively. We then choose parallel sentences from other languages corresponding to the above English sentences to keep the test and validation set the same for all languages.

3.2 JHU Bible Corpus

The Johns Hopkins University Bible corpus covers 1611 languages across 95 different language families, with dataset sizes for each language ranging from 8k sentences (verses, in the corpus) to 40k sentences. Similar to above, we create sets of seen and unseen languages to be used for identifying ambiguous instances and evaluation, respectively. We select languages that have at least 30k sentences, yielding a multi-way parallel dataset of 29k sentences.

Seen languages. We use 10 high-resource, European seen language pairs, consisting of English as our center language paired in two directions with Bulgarian, Italian, Finnish, German, and Greek.

Unseen languages. We evaluate on both European languages and low-resource languages from a typologically diverse set of families. The European languages contains 14 language pairs, consisting of English paired in two directions with Swedish, Portuguese, Lithuanian, Danish, Dutch, Czech, and French. The low-resource languages contain ten language pairs, consisting of English paired in two directions with Lashi (Sino-Tibetan), Tampulma (Niger-Congo), Cebuano (Austronesian), Yucatec Maya (Mayan), and Dyula (Niger-Congo).

Validation and test sets. We randomly select 1.8k sentences as the validation set and 7.4k sentences as the test set from the multi-way parallel corpus, similar to the Europarl setup.

3.3 Experimental Setup

Machine translation model. Our MT model is a standard transformer (Vaswani et al., 2017) implemented in PyTorch (Paszke et al., 2019)[2] . Our hyperparameters are as follows: 6 layers, 4 attention heads, an embedding size of 512, and a hidden dimension of 1024. During training, we use dropout (Srivastava et al., 2014) with a probability of 0.3 on embedding layer and dropout with a probability of 0.2 on the attention layers. We train the models for a maximum of 100 epochs using early stopping with a patience of 15. We employ an Adam optimizer (**?**) with beta values of 0.9, 0.98 and a learning rate of 0.0005. Each model was trained for a maximum of six hours on a single nVIDIA V100 GPU.

CHIA hyperparameters. We select 33% of the instances with the highest variability as our ambiguous sets, following Swayamdipta et al. (2020). For the Europarl corpus, this results in new training sets of size 6.6k, 13.2k, 26.4k and 52.8k. For the Bible, this gives a new training set of size 6.6k.

Metric. We evaluate our MT models with BLEU (Papineni et al., 2002), using the SacreBLEU tool (Post, 2018).

Random baseline. We compare CHIA to random sampling of instances, or *Random*. In order to account for variation caused by randomness, for each language pair and data size, we create *three* independent random sets. Results we report for the random baseline are average performances of the models trained on the random sets.

4 Results

4.1 Europarl: MT Training Dynamics

Figure 3 shows performance of models trained on the seen languages. For each data subset, we report the difference in BLEU between a model trained on instances selected using CHIA and *Random*. The exact scores for both methods, as well as model performance on the entire dataset can be found in the appendix.

In 39 out of the 40 models we investigate, models trained on instances chosen by CHIA outperform models trained on randomly selected instances. Further, we observe that the improvement in performance is greater when the size of the dataset is small. On dataset sizes of 6.6k, 13.2k, 26.4k, and 52.8k, the average BLEU score improvements are 4.30, 4.41, 2.48, and 1.29 respectively.

Looking at individual subsets, we observe that ambiguous instances that are most effective change with dataset size. For subsets of size 6.6k, with CHIA, we see the largest improvements of 10.76 BLEU on the Spanish–English model. In contrast, with the Finnish–English dataset, there is a drop of 1.09 BLEU when instances are selected using CHIA. For subsets of size 13.2k, the English–German model shows the largest improvement with

[2] Code and models are at https://nala-cub.github.io/resources/

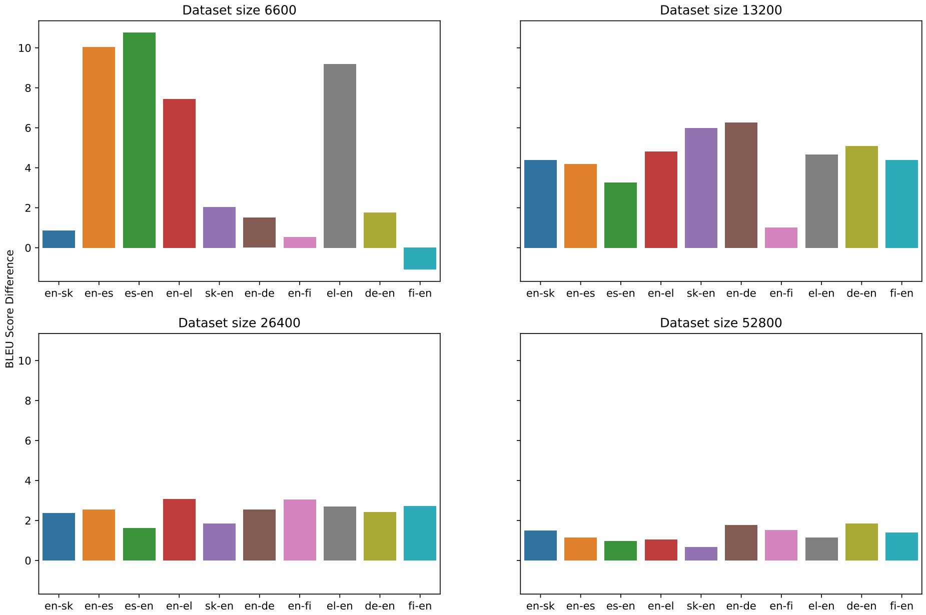

Figure 3: Difference in BLEU score between models trained on sentences chosen using CHIA and randomly selected sentences. The languages reported here are from our seen set of Europarl languages.

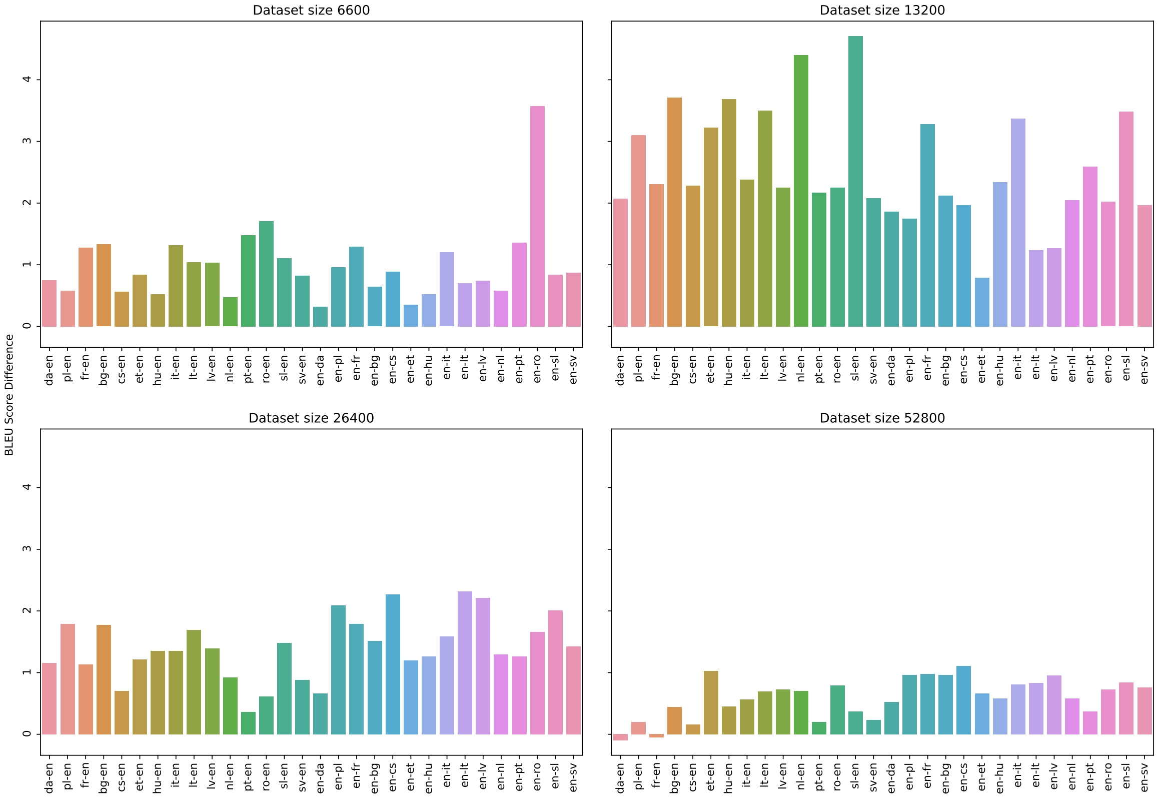

Figure 4: Difference in BLEU score between models trained on sentences chosen using CHIA and randomly selected sentences. The languages reported here are on our unseen set of Europarl languages.

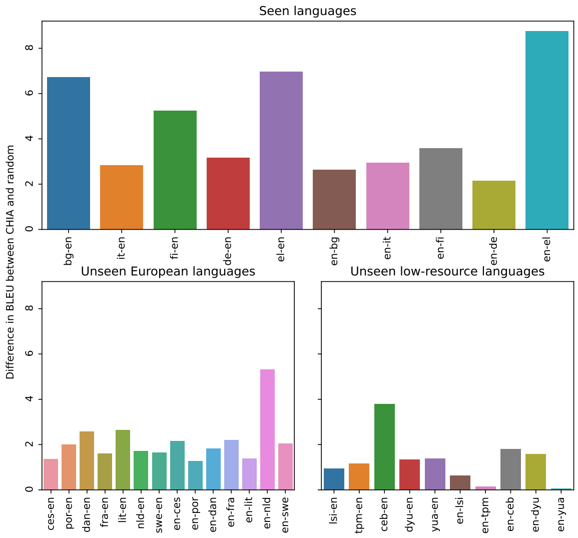

Figure 5: Difference in BLEU score between models trained on sentences chosen using CHIA and randomly selected sentences. All results are reported on the Bible corpus.

6.26 BLEU, whereas the models trained on English–Finish have the smallest improvement with 1.02 BLEU. For subsets of size 26.4k, the English–Greek model shows the largest improvement with 3.06 BLEU, whereas the Spanish–English model has the smallest improvement of 1.61 BLEU. For subsets of size 52.8k, the German–English shows the largest improvement with 1.85 BLEU, whereas the Slovak–English model has the smallest improvement with 0.67 BLEU.

4.2 Europarl: Selecting Effective Instances for New Languages

Figure 4 shows the performance of models on the unseen languages when trained on instances selected using CHIA, in comparison to randomly selected instances. The exact scores for both methods can be found in the Appendix.

We see that out of the 120 models we investigate, 118 of the models outperform *Random* when trained on instances selected using CHIA. The correlation between BLEU score improvements and dataset sizes is less pronounced in this case than with the seen languages – on average, the BLEU score improvements obtained by selecting instances through CHIA on dataset sizes of 6.6k, 13.2k, 26.4k, and 52.8k are 0.98, 2.54, 1.41, and 0.60, respectively.

As with the performance on the seen languages, we observe changes with dataset size. For subsets of size 6.6k, the English–Romanian model shows the largest improvement with 3.57 BLEU, whereas the models trained on English–Finnish have the smallest improvement with 0.32 BLEU. For subsets of size 13.2k, the Slovene–English model shows the largest improvement with 4.71 BLEU, whereas the models trained on English–Estonian have the smallest improvement with 0.79 BLEU. For subsets of size 26.4k, the English–Lithuanian model shows the largest improvement with 2.32 BLEU, whereas the models trained on Portuguese–English have the smallest improvement with 0.32 BLEU. For subsets of size 52.8k, the English–Croatian model shows the largest improvement with 1.11 BLEU, whereas models trained with CHIA on Danish–English and French–English underperform randomly chosen instances by 0.1 and 0.05 points respectively.

Overall, our results show the benefit of using CHIA to select sentences that should be translated for an unseen target language. Depending on the desired size of the dataset, selecting sentences with CHIA can result in MT model performance improvements of up to 4.71 points over randomly selecting sentences.

4.3 Bible

Figure 5 shows the performance of CHIA on the Bible; exact scores can be found in the appendix.

From the first subplot, we see that for all seen languages, *all* models trained on instances chosen by CHIA outperform those trained on randomly selected instances. The maximum improvement (8.76 BLEU) is seen for English–Greek, and the average improvement is 4.50 BLEU.

The second and third subplots show the performance on unseen languages. First, for European languages, we see a maximum improvement of 5.31 BLEU for English–Dutch and an average improvement of 2.12 BLEU. However, more importantly, we also obtain consistent improvements for languages from low-resource language families: the largest improvement (3.8 BLEU) is for Cebuano–English, while the smallest improvement (0.14 BLEU) is for English–Yucatan Maya. On average, we observe an improvement of 1.28 BLEU when instances are chosen through CHIA.

5 Analysis

5.1 Number of Overlapping Ambiguous Sets

As described in Section 3.3, we determine our final informative instances using ambiguous sets from all 10 seen language pairs. We perform a case study on the Europarl corpus to investigate the benefit of using CHIA when fewer seen languages are available for identifying ambiguous instances. Specifically, for each language pair in our unseen set, we train MT models on the ambiguous sets from each of the 10 language pairs in our seen set.

The results are presented in Table 1. The last row indicates the average performance across all unseen languages when each seen language is used as the informative set. The overall average performance in this setting, computed as the average of the last row, is -0.60, indicating that models trained on instances selected by CHIA underperform randomly selected instances. In contrast, for the same dataset size of 6.6k, when all 10 seen languages are used, the average BLEU score *increases* by 0.99 for models trained on instances selected using CHIA vs *Random* (as described in Section 4.2).

Looking at individual unseen language pairs, we see that for 22 out of 30 language pairs, using the intersection of all 10 seen languages (*All*) gives an improvement over using any single seen language pair. Out of the eight exceptions, for five of them, using ambiguous sentences from the English–Finish dataset gives the largest improvement: these are Danish–English, Czech–English, Lithuanian–English, Swedish–English and English–Danish. Additionally, for Estonian–English and English–Slovene, using the Finnish–English dataset gives the largest improvement. Finally, for English–Swedish, using the German–English dataset gives the largest improvement.

Further, we observe that out of all individual seen language pairs, using the English–Finnish dataset achieves the largest improvement over using randomly selected sentences in 19 out of 30 unseen language pairs. Next, for 5 out of 30 unseen language pairs, using the Finnish–English dataset achieves the largest improvement. This is notable since from looking at Figure 3, we see that models trained on ambiguous sets from the Finnish–English and English–Finnish dataset achieve the lowest performance improvement on their respective validation sets.

5.2 Analysis of Ambiguous Instances

To investigate the characteristics of sentences selected by CHIA, we analyze if ambiguous sentences are similar to randomly sampled sentences in length and number of types. We examine sentences in the training set of the Europarl corpus, and the results are presented in Table 2.

We find that sentences chosen by CHIA are the same length as sentences chosen randomly – in the source set, sentences chosen by CHIA are 1.07% longer on average, and in the target set, sentences chosen by CHIA are 0.09% shorter on average. We also find that the number of distinct word types are comparable between both – in the source set, sentences chosen by CHIA have 1.58% fewer types than sentences chosen randomly, whereas in the target set, sentences chosen by CHIA have 3.84% fewer types than sentences chosen randomly.

This indicates that CHIA does not just rely on simple characteristics of the training sentences, and instead identifies sentences that are truly beneficial to a model's learning ability. We leave it to future work to design more complex probes that can characterize the nature of the sentences selected by CHIA.

6 Related Work

Active learning. Active learning (Cohn et al., 1995; Settles, 2009; Ren et al., 2021) provides a way to identify the most useful instances that help a model learn, thereby optimizing limited labeled training data or annotation budgets. The most relevant active learning strategy to ours is uncertainty sampling (Lewis and Gale, 1994; Lewis and Catlett, 1994), where a learner trained on seed data is used to choose examples whose labels it is least certain about, and passing those examples to an oracle for annotation, typically a human annotator. This strategy has been utilized in NLP for tasks including text classification (Lewis and Gale, 1994; Zhu et al., 2008), named entity recognition (Shen et al., 2018), dependency parsing (Li et al., 2016), *inter alia*.

Active learning has also been used for identifying informative instances for MT (Haffari et al., 2009; Haffari and Sarkar, 2009; Bloodgood and Callison-Burch, 2010; Zeng et al., 2019). These methods require seed labeled or unlabeled data in the target language, from which informative instances are identified by an MT model, translated by a human expert, and added back to the seed data to improve the model iteratively. In contrast, CHIA identifies informative instances for a target language based only on a multi-way parallel corpus containing the source language. Moreover, our motivation is to present a method to curate parallel data for low-resource or no resource languages without an *existing* model in the target language

Language pair	All	en-es	en-fi	en-de	en-sk	en-el	de-en	el-en	es-en	fi-en	sk-en
da-en	0.75	-0.95	**2.33**	0.27	-0.87	-0.55	0.26	-0.61	-1.40	0.50	-0.60
pl-en	0.58	-0.49	**0.34**	-0.42	-0.88	-1.14	0.17	-1.56	-1.61	-0.21	-1.21
fr-en	1.28	-1.88	**-0.04**	-0.09	-1.07	-1.51	-0.31	-1.45	-1.71	-0.28	-0.99
bg-en	1.33	-0.84	**0.86**	0.11	-1.57	-1.99	-0.52	-2.01	-2.28	-0.01	-1.10
cs-en	0.56	0.23	**0.83**	0.08	-0.90	0.71	-0.22	-1.12	-1.25	0.29	-0.87
et-en	0.84	-1.06	-0.08	-0.30	-0.79	-0.33	-0.01	-1.26	-1.20	**1.09**	-0.87
hu-en	0.52	-0.71	-0.01	**0.31**	-0.66	-0.66	-0.40	-0.81	-1.13	-0.28	-0.72
it-en	1.32	-1.60	**0.74**	0.50	-1.54	-1.37	-0.06	-1.80	-1.44	0.36	-0.66
lt-en	1.04	-0.83	**1.59**	-0.02	-0.90	-0.70	0.28	-0.77	-0.42	-0.22	-0.91
lv-en	1.03	-0.51	**1.00**	-0.46	-1.28	-1.23	-0.54	-1.20	-1.34	-0.08	-1.33
nl-en	0.47	-1.03	0.22	-0.02	-1.02	-0.58	0.13	-0.97	-1.26	**0.23**	-0.47
pt-en	1.48	-1.55	**0.85**	0.15	-2.03	-1.89	-0.05	-1.71	-2.36	0.37	-0.19
ro-en	1.71	-1.73	-2.09	-0.40	-2.31	-1.66	-0.63	-1.64	-2.41	**0.22**	-1.27
sl-en	1.11	-1.33	0.08	0.28	-1.15	-1.10	0.36	-1.41	-1.89	**0.50**	-0.87
sv-en	0.82	-0.69	**1.47**	0.08	-1.10	-0.73	-0.22	-1.15	-1.33	0.69	-0.49
en-da	0.32	-0.72	**1.10**	0.11	-1.00	-0.88	-0.35	-1.21	-0.89	0.40	0.75
en-pl	0.96	-0.53	**0.31**	-0.10	-0.39	-0.29	-0.43	-0.70	-0.99	-0.08	-0.39
en-fr	1.29	-1.75	**-0.37**	-0.51	-2.26	-2.04	-1.17	-1.72	-2.00	-0.63	-0.60
en-bg	0.64	-1.56	**0.45**	-0.04	-1.54	-2.00	-0.54	-1.81	-1.55	0.09	-0.65
en-cs	0.89	-0.82	**0.02**	-0.41	-0.32	-0.35	-0.26	-1.07	-1.05	0.00	-0.08
en-et	0.35	-0.43	-0.11	-0.24	-0.26	-0.26	-0.38	-0.38	-0.67	-0.29	**0.15**
en-hu	0.52	-0.26	-0.14	**-0.05**	-0.15	-0.18	-0.31	-0.51	-0.37	-0.21	-0.33
en-it	1.20	-1.10	**0.09**	-0.75	-0.87	-1.29	-0.21	-1.89	-1.63	-0.51	-0.99
en-lt	0.70	-0.40	-0.22	-0.62	-0.46	**-0.11**	-0.23	-0.60	-0.82	-0.20	-0.56
en-lv	0.74	-0.29	**-0.01**	-0.29	-0.41	-0.62	-0.38	-0.89	-0.80	-0.21	-0.01
en-nl	0.58	-0.58	**0.18**	0.10	-0.53	-0.74	-0.60	-0.29	-0.12	-0.01	-0.93
en-pt	1.36	-1.56	**0.90**	0.19	-1.02	-1.05	0.17	-1.14	-1.96	0.62	-1.09
en-ro	3.57	-3.87	0.73	**1.34**	-3.69	-1.25	-1.86	-1.49	-6.44	-1.04	-0.73
en-sl	0.84	-0.32	0.57	0.38	-0.75	-0.79	-0.45	-0.94	-0.80	**1.68**	-0.44
en-sv	0.87	-0.87	0.07	0.35	-1.32	-0.50	**1.77**	-0.84	-1.40	1.03	-0.54
Average	0.99	-1.00	0.39	-0.02	-1.10	-0.90	-0.23	-1.17	-1.48	0.13	-0.63

Table 1: Analysis of the effect of using a single seen language-pair (columns) for identifying instances in unseen languages (rows). Values are difference in BLEU score between models trained on sentences selected by CHIA and randomly. The *All* column indicates results when all 10 seen language-pairs are used. Bold numbers are the highest scores among the 10 seen language-pairs, underlined numbers are the highest overall (including *All*).

Feature	Percentage difference
Source length	1.07
Target length	-0.09
Source number of types	-1.58
Target number of types	-3.84

Table 2: Analysis of the characteristics of chosen sentences. Percentage difference is between sentences selected by CHIA and sentences selected randomly.

Low-resource machine translation. While our work provides a way to obtain high-quality *parallel* data for low-resource machine translation, existing research has investigated how monolingual data can be used to develop MT systems (Pytlik and Yarowsky, 2006; Klementiev et al., 2012; Gülçehre et al., 2015; Sennrich et al., 2016; Zhang and Zong, 2016; Domhan and Hieber, 2017; Gibadullin et al., 2019). Monolingual corpora can be used to generate pseudo-parallel data through back-translation (Sennrich et al., 2016; Hoang et al., 2018), round-trip training (Cheng et al., 2016), or copying target language sentences to the source (Currey et al., 2017). Monolingual corpora have also been exploited by unsupervised methods (Lample et al., 2018; Artetxe et al., 2018; Liu et al., 2020) that need limited or no parallel data. However, on truly low-resource languages, existing unsupervised methods have been found to perform poorly (Guzmán et al., 2019; Marchisio et al., 2020).

Collecting parallel data in an efficient and cost-effective manner is thus important for building and evaluating MT systems for low-resource languages. For medium and high-resource languages, parallel data can be collected through scraping the web (Bañón et al., 2020; Ramesh et al., 2021), through religious corpora (Resnik et al., 1999; Agić and Vulić, 2019; McCarthy et al., 2020), or parliament proceedings (Koehn, 2005). Since such resources are not available for low-resources languages, other techniques such as crowdsourcing have been used, using bilingual speakers (Post et al., 2012), as well as monolingual speakers with images or GIFs as a pivot (Madaan et al., 2020; Bhatnagar et al., 2021).

7 Conclusion

We propose CHIA, an algorithm for choosing informative instances to annotate when creating MT datasets, thereby maximizing a limited annotation budget. CHIA is based on our two contributions: 1) By extending prior work we propose a method to identify beneficial training instances for sequence-to-sequence tasks. 2) Using this method, we show that we can leverage existing parallel data in high-resource languages to identify informative instances for new languages. We find that in comparison to randomly selected data, MT models trained on data selected using CHIA achieve average improvements in BLEU score of 1.59 points. Notably, CHIA is effective even when evaluating on low-resource languages, providing an efficient data annotation strategy.

Limitations and Future Work In our experiments we use English as a center language. In future work, we will investigate alternate center languages, as well as the effectiveness of our method without a center language. Additionally, a limitation of CHIA is that a multi-way parallel dataset containing the center language is required, which might be difficult to find for specific domains. Our case study further indicates that selecting data using a single high-resource language alone may not be adequate to achieve performance improvements. In future work, we will investigate the effect of different numbers and combinations of seen languages on our method.

Furthermore, while we apply CHIA to select useful data for machine translation, our method can potentially be applied for collecting data in a low-resource language for any NLP task, provided a multi-way parallel dataset, which may be investigated in future work.

Acknowledgments

We would like to thank the reviewers and the members of the NALA research group for their feedback and suggestions. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

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.

Rajat Bhatnagar, Ananya Ganesh, and Katharina Kann. 2021. Don't rule out monolingual speakers: A method for crowdsourcing machine translation data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1099–1106, Online. Association for Computational Linguistics.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864, Uppsala, Sweden. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

David Cohn, Zoubin Ghahramani, and Michael Jordan. 1995. Active learning with statistical models. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Ilshat Gibadullin, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. A survey of methods to leverage monolingual data in low-resource neural machine translation. *ArXiv*, abs/1910.00373.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 181–189.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

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, Melbourne, Australia. Association for Computational Linguistics.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140, Avignon, France. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 3–12, Berlin, Heidelberg. Springer-Verlag.

Zhenghua Li, Min Zhang, Yue Zhang, Zhanyi Liu, Wenliang Chen, Hua Wu, and Haifeng Wang. 2016. Active learning for dependency parsing with partial annotation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 344–354, Berlin, Germany. 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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Aman Madaan, Shruti Rijhwani, Antonios Anastasopoulos, Yiming Yang, and Graham Neubig. 2020. Practical comparable data collection for low-resource languages via images.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

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, Brussels, Belgium. Association for Computational Linguistics.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

Brock Pytlik and David Yarowsky. 2006. Machine translation for languages lacking bitext via multilingual gloss transduction. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 156–165, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Comput. Surv.*, 54(9).

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the "book of 2000 tongues". *Computers and the Humanities*, 33(1/2):129–153.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I.*

A Appendix

A.1 BLEU Scores for Europarl

The exact values for each model, trained on ambiguous subsets chosen by CHIA, three randomly selected subsets, as well as the entire available dataset, can be found below in Tables 3, 4, 5, and 6. The first ten rows are from our seen languages, and the next 30 are for our unseen languages.

A.2 BLEU scores for Bible

The exact BLEU scores for CHIA and random are reported below in Table 7. The first ten rows correspond to the seen languages, and the rest to the low-resource languages.

Language pair	Sub count	All 20000	Random 1	Random 2	Random 3	CHIA
en-es	6600	37.51	20.06	20.11	20.33	**30.26**
en-fi	6600	13.6	7.64	7.67	7.54	**8.13**
en-de	6600	22.24	10.8	10.59	10.69	**12.27**
en-sk	6600	23.84	9.53	9.6	9.53	**10.34**
en-el	6600	31.89	14.72	18.59	14.6	**23.52**
de-en	6600	28.28	15.96	15.54	15.46	**17.36**
el-en	6600	35.39	19.36	19.26	19.63	**28.81**
es-en	6600	37.77	20.52	20.82	20.29	**31.38**
fi-en	6600	21.08	11.32	11.06	**11.38**	10.29
sk-en	6600	28.41	14.94	14.58	14.56	**16.95**
da-en	6600	-	16.95	17.27	16.81	**17.76**
pl-en	6600	-	13.29	13.42	13.42	**13.96**
fr-en	6600	-	16.47	17.11	17.4	**18.27**
bg-en	6600	-	18.06	17.47	17.17	**18.9**
cs-en	6600	-	14.61	15.22	14.94	**15.48**
et-en	6600	-	11.81	11.94	11.48	**12.58**
hu-en	6600	-	11.63	11.52	11.73	**12.15**
it-en	6600	-	14.6	14.58	14.41	**15.85**
lt-en	6600	-	12.36	12.57	12.73	**13.59**
lv-en	6600	-	13.42	13.26	13.38	**14.38**
nl-en	6600	-	13.43	13.45	13.87	**14.05**
pt-en	6600	-	18.14	17.99	17.99	**19.52**
ro-en	6600	-	18.59	18.53	18.48	**20.24**
sl-en	6600	-	14.64	14.47	14.78	**15.74**
sv-en	6600	-	16.56	16.35	16.55	**17.31**
en-da	6600	-	15.17	14.91	15.4	**15.48**
en-pl	6600	-	7.31	7.58	7.58	**8.45**
en-fr	6600	-	15.85	16.29	15.08	**17.03**
en-bg	6600	-	14.84	14.31	14.38	**15.15**
en-cs	6600	-	8.26	8.75	8.39	**9.36**
en-et	6600	-	6.01	6.17	5.89	**6.37**
en-hu	6600	-	8.44	8.37	7.99	**8.79**
en-it	6600	-	11.22	10.88	10.83	**12.18**
en-lt	6600	-	8.32	8.28	8.05	**8.92**
en-lv	6600	-	9.07	8.79	9.28	**9.79**
en-nl	6600	-	12.27	12.02	12.07	**12.7**
en-pt	6600	-	15.92	15.76	15.77	**17.18**
en-ro	6600	-	13.65	13.39	12.98	**16.91**
en-sl	6600	-	10.19	10.46	10.15	**11.11**
en-sv	6600	-	12.87	13.09	12.63	**13.73**

Table 3: BLEU scores for language pairs for each setting where the base dataset count is 20000

Language pair	**Sub count**	**All 40000**	**Random 1**	**Random 2**	**Random 3**	**CHIA**
en-es	13200	44.05	33.19	33.93	33.58	**37.53**
en-fi	13200	20.83	9.36	**11.99**	9.7	11.36
en-de	13200	29.47	16.89	17.06	17.06	**23.28**
en-sk	13200	30.82	19.23	16.14	19.28	**22.67**
en-el	13200	38.27	26.11	27.67	27.85	**31.87**
de-en	13200	35.86	23.51	23.95	23.6	**28.71**
el-en	13200	42.44	30.21	33.19	33.21	**36.79**
es-en	13200	44.34	34.84	33.78	34.87	**37.82**
fi-en	13200	28.41	14.02	14.95	15.14	**19.11**
sk-en	13200	35.48	22.79	23.19	23.15	**29.05**
da-en	13200	-	28.68	29.33	28.94	**31.05**
pl-en	13200	-	19.7	19.85	17.79	**22.21**
fr-en	13200	-	26.96	26.48	26.91	**29.09**
bg-en	13200	-	29.58	29.44	29.24	**33.13**
cs-en	13200	-	22.54	22.82	22.25	**24.82**
et-en	13200	-	15.51	14.81	15.66	**18.55**
hu-en	13200	-	14.25	14.87	14.14	**18.11**
it-en	13200	-	22.32	22.09	22.41	**24.65**
lt-en	13200	-	15.94	15.57	16.27	**19.43**
lv-en	13200	-	19.78	20.18	20.23	**22.31**
nl-en	13200	-	17.43	18.04	20.35	**23.01**
pt-en	13200	-	29.44	32.29	31.82	**33.35**
ro-en	13200	-	33.2	31.98	33.6	**35.18**
sl-en	13200	-	23.78	23.06	23.1	**28.02**
sv-en	13200	-	26.81	26.45	27.71	**29.07**
en-da	13200	-	28.76	28.95	28.91	**30.73**
en-pl	13200	-	12.18	12.29	12.2	**13.97**
en-fr	13200	-	25.73	25.85	26.04	**29.15**
en-bg	13200	-	25.49	26.02	27.44	**28.44**
en-cs	13200	-	13.99	14.13	13.93	**15.99**
en-et	13200	-	8.28	7.92	8.03	**8.87**
en-hu	13200	-	10.52	10.57	9.86	**12.66**
en-it	13200	-	17.96	17.99	17.94	**21.33**
en-lt	13200	-	10.44	10.63	10.5	**11.76**
en-lv	13200	-	14.51	14.01	13.95	**15.43**
en-nl	13200	-	18.04	17.89	18.13	**20.07**
en-pt	13200	-	26.74	28.25	27.64	**30.13**
en-ro	13200	-	25.61	26.03	25.67	**27.79**
en-sl	13200	-	17.01	19.84	20.08	**22.46**
en-sv	13200	-	25.42	23.25	25.21	**26.6**

Table 4: BLEU scores for language pairs for each setting where the base dataset count is 40000

Language pair	**Sub count**	**All 80000**	**Random 1**	**Random 2**	**Random 3**	**CHIA**
en-es	26400	47.74	40.86	40.41	40.48	**43.25**
en-fi	26400	25.96	15.67	16.37	16.48	**19.34**
en-de	26400	34.31	25.07	25.75	25.89	**28.12**
en-sk	26400	36.08	26.98	27.03	26.83	**29.04**
en-el	26400	42.83	34.45	34.68	34.28	**37.46**
de-en	26400	40.81	31.8	32.12	31.95	**34.29**
el-en	26400	46.74	38.41	39.28	39.32	**41.6**
es-en	26400	47.97	41.44	40.65	40.82	**42.67**
fi-en	26400	33.68	23.92	23.69	24.14	**26.51**
sk-en	26400	40.35	32.08	31.38	31.54	**33.67**
da-en	26400	-	35.75	35.83	36.24	**37.1**
pl-en	26400	-	27.86	27.51	27.91	**29.55**
fr-en	26400	-	34.92	34.82	35.14	**36.09**
bg-en	26400	-	37.1	36.43	36.17	**38.34**
cs-en	26400	-	30.84	30.67	31.02	**31.54**
et-en	26400	-	24.76	24.99	24.37	**25.92**
hu-en	26400	-	24.48	24.38	24.67	**25.86**
it-en	26400	-	30.56	30.47	30.62	**31.9**
lt-en	26400	-	24.29	24.5	24.76	**26.21**
lv-en	26400	-	28.05	28.02	27.89	**29.38**
nl-en	26400	-	28.31	28.27	28.54	**29.29**
pt-en	26400	-	38.34	38.22	38.21	**38.62**
ro-en	26400	-	39.93	39.4	40.04	**40.4**
sl-en	26400	-	32.32	31.96	31.64	**33.45**
sv-en	26400	-	34.33	34.33	34.33	**35.21**
en-da	26400	-	35.42	35.3	35.44	**36.05**
en-pl	26400	-	18.41	18.94	18.94	**20.85**
en-fr	26400	-	35.34	34.84	34.83	**36.79**
en-bg	26400	-	34.48	34.43	34.76	**36.07**
en-cs	26400	-	22.31	21.72	21.37	**24.07**
en-et	26400	-	14.54	14.93	14.27	**15.78**
en-hu	26400	-	17.37	17.16	17.59	**18.63**
en-it	26400	-	26.63	26.95	26.08	**28.14**
en-lt	26400	-	18.03	18.23	18.23	**20.48**
en-lv	26400	-	22.25	20.97	21.57	**23.81**
en-nl	26400	-	24.55	24.89	24.89	**26.07**
en-pt	26400	-	36.12	35.58	35.67	**37.05**
en-ro	26400	-	33.07	32.44	32.59	**34.36**
en-sl	26400	-	27.99	27.18	27.38	**29.53**
en-sv	26400	-	30.87	31.04	31.25	**32.47**

Table 5: BLEU scores for language pairs for each setting where the base dataset count is 80000

Language pair	**Sub count**	**All 160000**	**Random 1**	**Random 2**	**Random 3**	**CHIA**
en-es	52800	50.18	45.72	45.8	45.59	**46.87**
en-fi	52800	28.97	23.12	23.09	23	**24.57**
en-de	52800	36.51	31.8	31.64	31.51	**33.26**
en-sk	52800	38.82	32.89	32.35	33.1	**34.09**
en-el	52800	45.16	40.49	40.25	40.53	**41.49**
de-en	52800	43.91	38.33	38.2	38	**39.86**
el-en	52800	49.47	44.36	44.54	44.13	**45.51**
es-en	52800	50.59	45.85	46.14	45.53	**46.86**
fi-en	52800	37.19	30.88	30.64	30.73	**31.94**
sk-en	52800	43.48	37.91	37.86	38.25	**38.5**
da-en	52800	-	41.2	41.32	**41.5**	41.24
pl-en	52800	-	33.67	33.86	33.75	**33.96**
fr-en	52800	-	39.95	39.92	**40.11**	39.94
bg-en	52800	-	41.9	42.29	41.87	**42.46**
cs-en	52800	-	36.67	36.82	**37.12**	37.03
et-en	52800	-	31.25	31.69	31.54	**32.52**
hu-en	52800	-	31.74	31.83	32.04	**32.32**
it-en	52800	-	35.46	35.31	35.82	**36.09**
lt-en	52800	-	30.58	30.69	30.54	**31.29**
lv-en	52800	-	33.93	33.63	34.09	**34.61**
nl-en	52800	-	32.67	33	33	**33.59**
pt-en	52800	-	43.39	43.31	43.36	**43.55**
ro-en	52800	-	45.36	45.2	45.18	**46.04**
sl-en	52800	-	37.96	38.1	38.1	**38.42**
sv-en	52800	-	39.64	39.58	39.77	**39.89**
en-da	52800	-	40.34	40.52	40.88	**41.1**
en-pl	52800	-	25.06	24.99	24.81	**25.91**
en-fr	52800	-	40.7	40.92	40.61	**41.72**
en-bg	52800	-	40.57	40.57	40.6	**41.54**
en-cs	52800	-	28.5	28.41	28.56	**29.6**
en-et	52800	-	20.49	20.77	20.7	**21.31**
en-hu	52800	-	23.68	23.61	23.5	**24.18**
en-it	52800	-	32.05	32.29	32.19	**32.99**
en-lt	52800	-	24.98	24.62	24.91	**25.67**
en-lv	52800	-	28.34	28.53	28.53	**29.42**
en-nl	52800	-	29.66	29.5	29.6	**30.17**
en-pt	52800	-	41.18	40.96	41.3	**41.52**
en-ro	52800	-	38.64	38.48	38.01	**39.11**
en-sl	52800	-	34.64	34.36	34.62	**35.38**
en-sv	52800	-	36.3	36.53	36.5	**37.2**

Table 6: BLEU scores for language pairs for each setting where the base dataset count is 160000

Language pair	**Subcount**	**All 20000**	**Random**	**CHIA**
bg-en	6600	39.86	22.17	**28.89**
it-en	6600	28.78	16.01	**18.86**
fi-en	6600	30.88	17.29	**22.53**
de-en	6600	24.71	15.85	**19.01**
el-en	6600	45.4	28.7	**35.67**
en-bg	6600	34.07	15.29	**17.94**
en-it	6600	18.12	9.79	**12.74**
en-fi	6600	20.05	10.17	**13.77**
en-de	6600	17.55	8.41	**10.57**
en-el	6600	40.04	21.82	**30.58**
ces-en	6600	-	18.26	**19.61**
lsi-en	6600	-	7.1	**8.05**
por-en	6600	-	9.98	**11.98**
tpm-en	6600	-	7.21	**8.37**
ceb-en	6600	-	26.07	**29.87**
dan-en	6600	-	17.84	**20.42**
dyu-en	6600	-	10.96	**12.3**
fra-en	6600	-	16.39	**18.0**
lit-en	6600	-	12.88	**15.51**
nld-en	6600	-	28.75	**30.47**
swe-en	6600	-	18.75	**20.4**
yua-en	6600	-	7.81	**9.18**
en-ces	6600	-	8.52	**10.67**
en-lsi	6600	-	12.22	**12.85**
en-por	6600	-	6.62	**7.89**
en-tpm	6600	-	5.7	**5.84**
en-ceb	6600	-	29.25	**31.06**
en-dan	6600	-	12.75	**14.58**
en-dyu	6600	-	11.26	**12.83**
en-fra	6600	-	10.21	**12.41**
en-lit	6600	-	5.86	**7.23**
en-nld	6600	-	20.8	**26.11**
en-swe	6600	-	11.48	**13.53**
en-yua	6600	-	5.2	**5.26**

Table 7: BLEU scores for language pairs on the Bible dataset, base size is 20000