Data Selection for Discriminative Training in Statistical Machine Translation

Xingyi Song and **Lucia Specia**
Department of Computer Science
University of Sheffield
S1 4DP, UK
{xsong2,l.specia}@sheffield.ac.uk

Trevor Cohn
Computing and Information Systems
The University of Melbourne
VIC 3010, Australia
t.cohn@unimelb.edu.au

Abstract

The efficacy of discriminative training in Statistical Machine Translation is heavily dependent on the quality of the development corpus used, and on its similarity to the test set. This paper introduces a novel development corpus selection algorithm – the LA selection algorithm. It focuses on the selection of development corpora to achieve better translation quality on unseen test data and to make training more stable across different runs, particularly when hand-crafted development sets are not available, and for selection from noisy and potentially non-parallel, large scale web crawled data. LA does not require knowledge of the test set, nor the decoding of the candidate pool before the selection. In our experiments, development corpora selected by LA lead to improvements of over 2.5 BLEU points when compared to random development data selection from the same larger datasets.

1 Introduction

Discriminative training – also referred to as *tuning* – is an important step in log-linear model in Statistical Machine Translation (SMT) (Och and Ney, 2002). The efficacy of training is closely related to the quality of training samples in the development corpus, and to a certain extent, to the proximity between this corpus and the test set(s). Hui et al. (2010) in their experiments show that by using different development corpora to train the same SMT system, translation performance can vary up to 2.5 BLEU points (Papineni et al., 2002) with a standard phrase-based system (Koehn et al., 2007). How to build a 'suitable' development corpus is a important problem in SMT discriminative training.

A suitable development corpus should aid discriminative training achieve higher quality models, and thus yield better translations. Previous research on selecting training samples for the development corpus can be grouped into two categories: i) selecting samples based on the test set (transductive learning), or ii) selecting samples without knowing the test set (inductive learning). Research in the first category focuses on how to find similar samples to the ones the system will be tested on. Li et al. (2010), Lu et al. (2008), Zheng et al. (2010), and Tamchyna et al. (2012) measure similarity based on information retrieval methods, while Zhao et al. (2011) selects similar sentences based on edit distance. These similarity based approaches have been successfully applied to the local discriminative algorithm proposed in (Liu et al., 2012). The limitation of these approaches is that the test set needs to be known before model building, which is rarely true in practice.

Our research belongs to the second category. Previous work on development data selection for unknown test sets include Hui et al. (2010). They suggest that training samples with high oracle BLEU scores[1] will lead to better training quality. Cao and Khudanpur (2012) confirmed this and further showed that better training data will offer high variance in terms of BLEU scores and feature vector values between oracle and non-oracle hypotheses, since these are more easily separable by

[1] Oracle BLEU scores are those computed for the closest candidate translation to the reference in the n-best list of the development set.

the machine learning algorithms used for tuning. Both of the above studies achieved positive results, but these approaches require decoding the candidate development data to obtain BLEU scores and feature values, which may be difficult apply if the pool for data selection is extremely large.

Another potential way of improving training quality based on a development corpus is to increase the size of this corpus. However, high-quality sentence aligned parallel corpora are expensive to obtain. In contrast to data used for rule extraction in SMT, data used for SMT discriminative training is required to be of better quality for reliable training. Development data is therefore often created by professional translators. In addition, increasing the corpus size also increases the computational cost and the time required to train a model. Therefore, finding out how much data is enough to build a suitable development corpus is also an important question. Web crawled or crowdsourcing data are much cheaper than professionally translated data, and research towards exploiting such type of data (Zaidan and Callison-Burch, 2011; Uszkoreit et al., 2010; Smith et al., 2010; Resnik and Smith, 2003; Munteanu and Marcu, 2005) has already been successfully applied to machine translation, both in phrase extraction and discriminative training. However, they do not provide a direct comparison between their selected data and professionally built development corpora.

In order to address these problems, in this paper we introduce a novel development corpus selection algorithm, the **LA Selection** algorithm. It combines sentence length, bilingual alignment and other textual clues, as well as data diversity for sample sentence selection. It does not rely on knowledge of the test sets, nor on the decoding of the candidate sentences. Our results show that the proposed selection algorithm achieves improvements of over 2.5 BLEU points compared to random selection. We also present experiments with development corpora for various datasets to shed some light on aspects that might have an impact on translation quality, namely showing a substantial effect of the sentence length in the development corpus, and that with the right selection process large development corpora offer little benefits over smaller ones.

The remainder of this paper is structured as follows: We will describe our novel LA selection algorithm in Section 2. Experimental settings and results are presented in Sections 3 and 4, respectively, where we also discuss the training quality and scalability over different corpus size.

Algorithm 1 Development Data Selection

```
Require: Data Pool $D = (f^t, r^t, a^t)_{t=1}^T$, Number
    of words $N$, length limits $\lambda_{low}$ and $\lambda_{top}$
 1: Select = [], Cand = [], $L = 0$
 2: for $d_i = (f^i, r^i, a^i)$ in $D$ do
 3:   if $\lambda_{low} < \text{length}(f^i) < \lambda_{top}$ then
 4:     Calculate feature score
          $s^i = \text{score}(f^i, r^i, a^i)$
 5:     Add $(s^i, d^i)$ to Cand
 6:   end if
 7: end for
 8: Sort Cand by score from high to low
 9: while Selected length $L < N$ do
10:   for $d^i$ in Cand do
11:     if $\text{maxSim}(f^i, \text{Select}[f^j]_{j=J-200}^J) < 0.3$
        and $\text{sim}(f^i, r^i) < 0.6$ then
12:       Add $(f^i, r^i)$ to Select
13:       $L = L + \text{length}(f^i)$
14:     end if
15:   end for
16: end while
17: return $Selected$
```

2 Development Corpus Selection Algorithm

The proposed development corpus selection algorithm has two main steps: (i) selecting training sentence pairs by sentence **L**ength, and (ii) selecting training sentence pairs by **A**lignment and other textual clues. We call it **LA selection**. It also has an further step to reward diversity in the set of selected sentences in terms of the words they contain. The assumption of the LA algorithm is that a good training sample should have a "reasonable" length, be paired with a good quality translation, as mostly indicated by the word alignment clues between the candidate pair, and add to the existing set in terms of diversity.

LA selection is shown in Algorithm 1. Assume that we have T sentence pairs in our data set D. Each sentence pair d_i in D contains a foreign sentence f^i, a translation of the foreign sentence r^i and the word alignment between them a^i. We first filter out sentence pairs below the low length threshold λ_{low} and above the high length threshold λ_{top} (Line 3). Sentence length has a major im-

+/-	**Alignment Features**
+	Source/Target alignment ratio
-	Source/Target top three fertilities ratio
+	Source/Target largest contiguous span ratio
-	Source/Target largest discontiguous span
	Text only Features
+	Source and target length ratio
-	Target function word penalty

Table 1: Features used to score candidate sentence pairs.

pact on word alignment quality, which constitute the basis for the set of features we use in the next step. Shorter sentences tend to be easier to align than longer sentences and therefore our algorithm would naturally be biased to selecting shorter sentences. However, as we show later in our experiments, sentences that are either too short or too long often harm model accuracy. Therefore, is important to set both bottom and top limits on sentence length. Based on empirical results, we suggest set $\lambda_{low} = 10$ $\lambda_{top} = 50$, as we will further discuss in Section 4.1.

After filtering out sentences by the length thresholds, the next step is to extract the feature values for each remaining candidate sentence pair. The features used in this paper are listed in Table 1. The first column of the Table is an indicator of the sign of the feature value, where a negative sign indicates that the feature will return a negative value, and positive sign indicates that the feature will return a positive value. The actual features, which we describe below, are given in the second column. These include word alignment features, which are computed based on GIZA++ alignments for the candidate development set, and simpler textual features. The alignment features used here are mostly adapted from (Munteanu and Marcu, 2005).

The **alignment ratio** is the ratio between the number of aligned words and length of the sentence in words:

$$\text{Alignment Ratio} = \frac{\text{No. Aligned Words}}{\text{Sentence Length}}$$

A low alignment ratio means that the data is most likely non-parallel, or else a highly non-literal translation. Either way, these are likely to prove detrimental.

Word fertility is the number of foreign words aligned to each target word. The **word fertility ratio** is the ratio between word fertility and sentence length. We use the top three largest fertility ratio as three features:

$$\text{Fertility Ratio} = -\frac{\text{Word fertility}}{\text{Sentence Length}}$$

This feature can detect garbage collection, where the aligner uses a rare word to erroneously account for many difficult words in the parallel sentence.

Our definition of **contiguous span** differs from that in (Munteanu and Marcu, 2005): we define it as a substring in which all words have an alignment to words in the other language. A **discontiguous span** is defined as a substring in which all words have no alignment to any word in the other language. The **contiguous span ratio**, CSR, is the length of the largest contiguous span over the length of the sentence:

$$CSR = \frac{LC}{\text{Sentence Length}}$$

The **discontiguous span ratio**, $DCSR$, is the length of the largest discontiguous span over the length of the sentence:

$$DCSR = -\frac{LDC}{\text{Sentence Length}}$$

where LC is the length of the contiguous span and LDC is the length of the discontiguous span.

In addition to the word alignment features, we use **source and target length ratio**, LR, to measure how close the source and target sentences in the pair are in terms of length:

$$LR = \begin{cases} \frac{TL}{SL} & \text{if } SL > TL \\ \frac{SL}{TL} & \text{if } TL > SL \end{cases}$$

where TL is target sentence length and SL is source sentence length.

Finally, the **target function words penalty**, FP, penalises sentences with a large proportion of function words or punctuation:

$$FP = -\exp\left(-\frac{n_{\text{func}}}{TL}\right)$$

where n_{func} is number of function words and punctuation symbols, and TL is the target sentence length. We only consider a target language penalty, but a source language penalty could also be used.

Once we obtained these feature values for all candidate sentence pairs, we apply two approaches to calculate an overall score for the candidate. The first is a heuristic approach, which simply sums over the scores of all features for each sentence (with some features negated as shown in Table 1). The second approach uses machine learning to combine these features, similar to what was done in (Munteanu and Marcu, 2005) to distinguish between parallel and non-parallel sentences. Here a binary SVM classifier is trained to predict samples that are more similar to professionally created sentences. The labelling of the data was therefore done by contrasting professionally created translations against badly aligned translations from web crawled data. The heuristic approach achieved better performance than the machine learning approach, as we will discuss in Section 4.2.

Lines 8 through 16 in Algorithm 1 describe the sentence pair selection procedure based on this overall feature score. The candidate sentence pair and its features are stored in the Cand list, and sorted from high to low according to their overall feature scores. The algorithm takes candidate sentence pairs from the Cand list until the number of words in the selected training corpus Select reaches the limit N. If the candidate sentence pair passes the condition in Line 11, the sentence pair is added to the selected corpus Select.

Line 11 has two purposes: first, it aims at increasing the diversity of the selected training corpus. Based on our experiments, candidate sentence pairs with similar feature scores (and thus similar rankings) may be very similar sentences, with most of their words being identical. We therefore only select a sentence pair whose source sentence has less than 0.3 BLEU similarity as compared to the source sentences in last 200 selected sentence pairs.[2] The second purpose is to filter out sentence pairs that are not translated, i.e., sentence pairs with same words in the source and target sides. Untranslated or partially untranslated sentence pairs are common in web crawled data. We therefore filter out the sentence pairs whose source and target have a BLEU similarity score of over 0.6.

3 Experimental Settings

SMT system: We build standard phrase-based SMT systems for each corpus using Moses with its 14 default features. The word alignment and language models were learned using GIZA++ and IRSTLM with Moses default settings. A trigram language model was trained on English side of the parallel data. For discriminative training we use the popular MERT (Och, 2003) algorithm.

Two language pairs are used in the experiments, French to English and Chinese to English, with the following corpora:

French-English Corpora: To build a French to English system we used the Common Crawl corpus (Smith et al., 2013). We filtered out sentence with length over 80 words and split the corpus into training (Common Crawl training) and tuning (Common Crawl tuning). The **training** subset was used for phrase table, language model and reordering table training. It contains $3,158,523$ sentence pairs (over 161M words) and average source sentence length of 27 words. The **tuning** subset is used as "Noisy Data Pool" to test our LA selection algorithm. It contains $31,929$ sentence pairs (over 1.6M words), and average source sentence length of 27 words. We compare the performance of our selected corpora against a concatenation of four professionally created development corpora (Professional Data Pool) for the news test sets distributed as part of the WMT evaluation (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010): 'newssyscomb2009', 'news-test2008', 'newstest2009' and 'newstest2010'. Altogether, they contain $7,518$ sentence pairs (over 392K words) with average source sentence length of 27 words. As **test data**, we take the WMT13 (average source sentence length = 24 words) and WMT14 (average source sentence length = 27 words) news test sets.

Chinese-English Corpora: To build the Chinese to English translation system we use the non-UN and non-HK Hansards portions of the FBIS (LDC2003E14) training corpus ($1,624,512$ sentence pairs, over 83M words, average source sentence = 24) and **tuning** ($33,154$ sentence pairs, over 1.7M words, average sentence length = 24). The professionally created development corpus in this case is the NIST MT06 test set[3] ($1,664$ sentence pairs, 86K words, average sentence length = 23 words). As **test data**, we use the NIST

[2] The 200 sentence pair limit is used to reduce the runtime on large datasets.

[3] It contains 4 references, but we only apply the first reference to make it comparable to our selection algorithm.

MT08 test set (average source sentence length = 24 words).

Note that for both language pairs, the test sets and professionally created development corpora belong to the same domain: news, for both French-English and Chinese-English. In addition, the test and development corpora for each language pair have been created in the same fashion, following the same guidelines. Our pool of noisy data, however, includes not only a multitude of domains different from news, but also translations created in various ways and noisy data.

4 Results

Our experiments are split in three parts: Section 4.1 examines how sentence length in development corpora affects the training quality. Section 4.2 compares our LA selection algorithm against randomly selected corpora and against professionally created corpora. Section 4.3 discusses the effect of development corpus size by testing translation performance with corpora of different sizes.

4.1 Selection by Sentence Length

In order to test how sentence length affects the quality of discriminative training, we split the tuning corpus into six parts according to source sentence length ranges (in words): [1-10], [10-20], [20-30], [30-40], [40-50] and [50-60]. For each range, we randomly select sentences to total $30,000$ words as a small training set, train a discriminative model based on the small training set, and test the translation performance on WMT13 and NIST MT08 test sets. We repeat the random selection and training procedure five times and report average BLEU scores in Table 2.

The top half of Table 2 shows the results for French-English translation. From this Table, we can see that corpora with sentence lengths of [30-40] and [30-50] lead to better translation quality than random selection, with a maximum average BLEU score of 25.62 for sentence length [30-40], outperforming random length selection by 1.26 BLEU points. Corpora with sentences in [10-20] and [20-30] perform slightly worse than random selection. The worst performance is obtained for corpora with very short or very long sentences.

The lower half of Table 2 shows the results for Chinese-English translation. Lengths [10-20], [20-30], [30-40] and [40-50] lead to better translation performance than random selection. As for French-English translation, the worst performance is obtained for corpora with very short or very long sentences, with a lower BLEU score than random selection.

According to above results, the best sentence length for discriminative training is not fixed, as it may depend on language pairs and corpus type. However, sentences below 10 words or above 50 words lead to poor results for both language pairs. We conduct another experiment selecting development corpora excluding sentences with length below 10 or above 50. Results are shown in column [10-50] of both Tables. Compared to random selection, [10-50] improved BLEU scores by 1.18 for French-English, and by 0.54 for Chinese-English. Note that our systems were developed on corpora with average sentence length of around 25 words, which is typical in most freely available training corpora,[4] the thresholds may differ for corpora with very different sentence lengths.

4.2 Selection by LA Algorithm

In what follows we compare the performance of our LA selection algorithm against randomly selected and professionally created corpora. We set $\lambda_{low} = 10$ and $\lambda_{top} = 50$ and select a development corpus with no more than $30,000$ words. Results are reported in Table 3, again with averages over five runs.

Considering first the results for the French-English WMT13 test set, the LA selection improves BLEU by 1.36 points compared with random selection, and also improves over sentence length-based selection (10-50). The performance of the LA selected corpus is only slightly lower (0.1 BLEU) than that of the professionally created corpus (Prof.), but the system is much more robust with much lower standard deviation (std). This is a surprising outcome as the professionally created development sets are drawn from the same domain as the test sets (news), and were created using the same translation guidelines as the test set, and therefore better results were expected for these corpora. We have similar findings for the French-English WMT14 and Chinese-English MT08 test sets. Systems trained on corpora selected by LA increase 1.21 and 2.53 BLEU points over random selection, respectively. For the WMT14 test set, the corpus selected by LA show slight im-

[4] For example, both Europarl and News-Commentary WMT corpora have an average of 25 words on their English side.

		Rand.	1-10	10-20	20-30	30-40	40-50	50-60	10-50
WMT13	**avg.**	24.36	22.85	23.61	24.43	25.62	24.62	22.94	25.54
	std.	0.84	0.65	0.80	0.51	0.40	1.06	0.99	0.84
MT08	**avg.**	18.79	18.11	20.00	19.63	18.85	19.29	18.53	19.33
	std.	0.83	0.29	1.45	1.00	0.85	1.38	0.81	1.16

Table 2: Average BLEU scores and standard deviation on French to English (WMT13) and Chinese to English (MT08) test sets for different ranges of sentence length. The leftmost **Rand.** column has no length restrictions.

		Rand.	10-50	LA_{10-50}	Prof.
WMT13	**avg.**	24.36	25.54	25.72	25.82
	std.	0.84	0.84	0.01	0.23
WMT14	**avg.**	25.19	25.31	26.40	26.31
	std.	0.30	0.14	0.04	0.16
MT08	**avg.**	18.79	19.33	21.32	23.49
	std.	0.83	1.16	0.83	0.31

Table 3: Average BLEU scores and standard deviation for French-English (WMT13, WMT14) news test sets and Chinese-English (MT08) test set with development corpora selected by length (10-50), LA algorithm (LA_{10-50}), randomly (Rand.), or created by professionals (Prof.).

	WMT13	WMT14
avg.	25.42	26.08
std.	0.08	0.08

Table 4: Average BLEU scores and standard deviation for SVM-based LA selection on French-English WMT13 and WMT14 test sets.

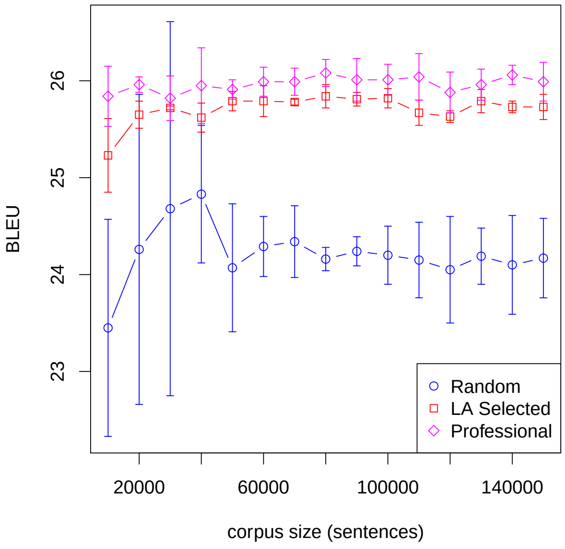

Figure 1: BLEU score changes for development corpora of different sizes with the French-English WMT13 corpus. The horizontal axis shows corpus size, and the vertical axis, BLEU scores. Points show the mean results and whiskers denote ± one standard deviation.

provements over the professionally created corpus (26.40 vs. 26.31) with a lower variance.

We also experiment with using the SVM classifier to combine features in the LA selection algorithm, as previously discussed. The classifier was trained using the SVMlight[5] toolkit with RBF kernel with its default parameter settings. We selected $30,000$ words from the professionally created WMT development corpus as positive training samples, and used as negative examples $30,000$ words from our corpus with the lowest LA selection score. Different from the LA selection method, here sentence length is not limited to 10-50, but rather the sentence length is provided as a feature to the classifier. The motivation was to test the ability of the algorithm in learning a suitable sentence length for tuning. Nevertheless, on average sentences have similar lengths: 16 for the corpus selected with the SVM classifier against 18 for the corpus selected with the heuristic method. Results for sentence selection using the highest classification scores are shown in Table 4.

LA selection with the SVM classifier outperforms random selection, but does worse than our heuristic approach (compare to LA_{10-50} in Table 3). The reason may be the quality of the training data: both our positive and negative training examples will contain considerable noise. The WMT professionally created corpora includes some odd translations, so the alignment features will be less reliable. Also, we stress that this is a harder problem than the one introduced in (Munteanu and Marcu, 2005), since their pool of candidate samples contained either parallel or non-parallel sentences, which are easier to label and to distinguish based on word alignment features. Our pool of candidate samples is assumed to be parallel, with our selection procedure aiming at selecting from this the highest quality translations.

[5] `http://svmlight.joachims.org/`

4.3 Effect of Training Corpus Size

Next, we consider the question of how much development data is needed to train a phrase-based SMT system. To test this we experiment with corpora ranging in size from $10,000$ words to $150,000$ words, with an incremental step of $10,000$ words. At each step we run MERT training five times and report the average BLEU scores. The test set is the WMT13.

Figure 1 shows how BLEU changes as we increase the training corpus size. The three lines represent the BLEU scores of three systems: Random selection from the French-English tuning dataset (blue line), LA selection from the same pool (red line), and WMT professionally created development corpus (green line). According to this Figure, performance increases as corpora sizes increase, for all techniques, but only up to $70,000$ words, after which performance is stable. The professionally created corpus achieves the best performance for any corpus size. Note however that the LA selection technique is only slightly worse, with less than 0.1 BLEU difference, for corpora sizes $\geq 30,000$ words. Random selection clearly performs poorly compared to both.

Also shown in Figure 1 are the standard deviation from five runs of the experiment. Random selection presents the largest standard deviation (greater than 0.6 BLEU) for training corpora of sizes below $50,000$ words. The maximum standard deviation is 1.93 at $30,000$ words. With larger training corpus sizes, the standard deviation of random selection is still higher than that of LA selected and professional data. LA selection has a much lower average standard deviation, even lower than the professionally created data. This is important for real application settings, where repeated runs are not practical and robust performance from a single run is imperative.

These results confirm some findings of previous research (Hui et al., 2010), namely that enlarging the tuning corpus leads to more accurate models. However we find that increasing the amount of data is not the best solution when creating a development corpus: much greater improvements are possible by instead focusing on selecting better quality data. Using data selection reduces the need for large development sets, in fact as few as 70k words is sufficient for robust tuning.

5 Conclusions

In this paper we have shown how the choice of the development corpus is critical for machine translation systems' performance. The standard practice of resorting to expensive human translations is not practical for many SMT application scenarios, and consequently making better use of existing parallel resources is paramount. Length is the most important single criterion for selecting effective sentences for discriminative training: overly short and overly long training sentences often harm training performance. Using large development sets brings only small improvements in accuracy, and a modest development set of 30k-70k words is sufficient for good performance. The key innovation in this paper was the LA sentence selection algorithm, which selects high quality and diverse sentence pair for translation. We have shown large improvements over random selection, of up to 2.53 BLEU points (Chinese-English). The approach is competitive with using manually translated development sets, despite having no knowledge of the test set, test set domain, nor using expensive expert translators. In future work, we plan to improve the classification technique for automatically predicting training quality through alternative methods for extracting training examples and additional features to distinguish between good and bad translations.

6 Acknowledgement

Dr. Specia has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347 (QT-LaunchPad). Dr. Cohn is the recipient of an Australian Research Council Future Fellowship (project number FT130101105).

References

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Cao, Yuan and Sanjeev Khudanpur. 2012. Sample selection for large-scale mt discriminative training. In *AMTA*.

Hui, Cong, Hai Zhao, Yan Song, and Bao-Liang Lu. 2010. An empirical study on development set selection strategy for machine translation learning. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 67–71, Uppsala, Sweden. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Bertoldi Nicola Federico, Marcello, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007, Demonstration Session*, Prague, Czech Republic.

Li, Mu, Yinggong Zhao, Dongdong Zhang, and Ming Zhou. 2010. Adaptive development data selection for log-linear model in statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 662–670, Beijing, China. Association for Computational Linguistics.

Liu, Lemao, Hailong Cao, Taro Watanabe, Tiejun Zhao, Mo Yu, and CongHui Zhu. 2012. Locally training the log-linear model for smt. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 402–411, Jeju Island, Korea.

Lu, Yajuan, Jin Huang, and Qun Liu. 2008. Improving statistical machine translation performance by training data selection and optimization.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504, December.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 295–302, Philadelphia, Pennsylvania.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Sapporo, Japan.

Papineni, Kishore, 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, pages 311–318, Philadelphia, Pennsylvania.

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, September.

Smith, Jason R., Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Los Angeles, California.

Smith, Jason R., Philipp Koehn, Herve Saint-Amand, Chris Callison-Burch, Magdalena Plamada, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013)*.

Tamchyna, Aleš, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting data for english-to-czech machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 374–381, Montreal, Canada.

Uszkoreit, Jakob, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109, Beijing, China.

Zaidan, Omar F. and Chris Callison-Burch. 2011. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Portland, Oregon.

Zhao, Yinggong, Yangsheng Ji, Ning Xi, Shujian Huang, and Jiajun Chen. 2011. Language model weight adaptation based on cross-entropy for statistical machine translation. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 20–30, Singapore, December. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Zheng, Zhongguang, Zhongjun He, Yao Meng, and Hao Yu. 2010. Domain adaptation for statistical machine translation in development corpus selection. In *Universal Communication Symposium (IUCS), 2010 4th International*.