An Empirical Study on Development Set Selection Strategy for Machine Translation Learning*

Cong Hui[12], **Hai Zhao**[12]†, **Yan Song**[3], **Bao-Liang Lu**[12]
[1]Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University, 800 Dong Chuan Rd., Shanghai 200240, China
[3]Department of Chinese, Translation and Linguistics, City University of Hong Kong
huicong@sjtu.edu.cn, {zhaohai,blu}@cs.sjtu.edu.cn

Abstract

This paper describes a statistical machine translation system for our participation for the WMT10 shared task. Based on MOSES, our system is capable of translating German, French and Spanish into English. Our main contribution in this work is about effective parameter tuning. We discover that there is a significant performance gap as different development sets are adopted. Finally, ten groups of development sets are used to optimize the model weights, and this does help us obtain a stable evaluation result.

1 Introduction

We present a machine translation system that represents our participation for the WMT10 shared task from Brain-like Computing and Machine Intelligence Lab of Shanghai Jiao Tong University (SJTU-BCMI Lab). The system is based on the state-of-the-art SMT toolkit MOSES (Koehn et al., 2007). We use it to translate German, French and Spanish into English. Though different development sets used for training parameter tuning will certainly lead to quite different performance, we empirically find that the more sets we combine together, the more stable the performance is, and a development set similar with test set will help the performance improvement.

2 System Description

The basic model of the our system is a log-linear model (Och and Ney, 2002). For given source language strings, the target language string t will be obtained by the following equation,

$$\hat{t}_1^I = \arg\max_{t_1^I}\{p_{\lambda_1^m}(t_1^I \mid s_1^J)\}$$
$$= \arg\max_{t_1^I}\{\frac{\exp[\sum_{m=1}^M \lambda_m h_m(t_1^I, s_1^J)]}{\sum_{\bar{t}_1^I}\exp[\sum_{m=1}^M \lambda_m h_m(\bar{t}_1^I, s_1^J)]}\},$$

where h_m is the m-th feature function and λ_m is the m-th model weight. There are four main parts of features in the model: translation model, language model, reordering model and word penalty. The whole model has been well implemented by the state-of-the-art statistical machine translation toolkit MOSES.

For each language that is required to translated into English, two sets of bilingual corpora are provided by the shared task organizer. The first set is the new release (version 5) of Europarl corpus which is the smaller. The second is a combination of other available data sets which is the larger. In detail, two corpora, *europarl-v5* and *news-commentary10* are for German, *europarl-v5* and *news-commentary10* plus *undoc* for French and Spanish, respectively. Details of training data are in Table 1. Only sentences with length 1 to 40 are acceptable for our task. We used the larger set for our primary submission.

We adopt word alignment toolkit GIZA++ (Och and Ney, 2003) to learn word-level alignment with its default setting and *grow-diag-final-and* parameters. Given a sentence pair and its corresponding word-level alignment, phrases will be extracted by using the approach in (Och and Ney, 2004). Phrase probability is estimated by its relative frequency in the training corpus. Lexical reordering is determined by using the default setting of MOSES with *msd-bidirectional* parameter.

For training the only language model (English), the data sets are extracted from monolingual parts of both *europarl-v5* and *news-commentary10*,

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 60773090 and Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the National High-Tech Research Program of China (Grant No.2008AA02Z315).

†corresponding author

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 67–71,
Uppsala, Sweden, 15-16 July 2010.

		sentences	words(s)	words(t)
de	small	1540549	35.76M	38.53M
	large	1640818	37.95M	40.64M
fr	small	1683156	44.02M	44.20M
	large	8997997	251.60M	228.50M
es	small	1650152	43.17M	41.25M
	large	7971200	236.24M	207.79M

Table 1: Bilingual training corpora from German(de), French(fr) and Spanish(es) to English.

which include 1968914 sentences and 47.48M words. And SRILM is adopted with *5-gram*, *interpolate* and *kndiscount* settings (Stolcke, 2002) .

The next step is to estimate feature weights by optimizing translation performance on a development set. We consider various combinations of 10 development sets with 18207 sentences to get a stable performance in our primary submission.

We use the default toolkits which are provided by WMT10 organizers for preprocessing (i.e., tokenize) and postprocessing (i.e., detokenize, recaser).

3 Development Set Selection

3.1 Motivation

Given the previous feature functions, the model weights will be obtained by optimizing the following maximum mutual information criterion, which can be derived from the maximum entropy principle:

$$\hat{\lambda}_1^M = \arg\max_{\lambda_1^M}\{\sum_{i=1}^{S}\log p_{\lambda_1^M}(t_i \mid s_i)\}$$

As usual, minimum error rate training (MERT) is adopted for log-linear model parameter estimation (Och, 2003). There are many improvements on MERT in existing work (Bertoldi et al., 2009; Foster and Kuhn, 2009), but there is no demonstration that the weights with better performance on the development set would lead to a better result on the unseen test set. In our experiments, we found that different development sets will cause significant BLEU score differences, even as high as one percent. Thus the remained problem will be how to effectively choose the development set to obtain a better and more stable performance.

3.2 Experimental Settings

Our empirical study will be demonstrated through German to English translation on the smaller corpus. The development sets are all development sets and test sets from the previous WMT shared translation task as shown in Table 2, and labeled as dev-0 to dev-9. Meanwhile, we denote 10 batch sets from batch-0 to batch-9 where the batch-i set is the combination of dev- sets from dev-0 to dev-i. The test set is *newstest2009*, which includes 2525 sentences, 54K German words and 58K English words, and *news-test2008*, which includes 2051 sentences, 41K German words and 43K English words.

id	name	sent	w(de)	w(en)
dev-0	dev2006	2000	49K	53K
dev-1	devtest2006	2000	48K	52K
dev-2	nc-dev2007	1057	23K	23K
dev-3	nc-devtest2007	1064	24K	23K
dev-4	nc-test2007	2007	45K	44K
dev-5	nc-test2008	2028	45K	44K
dev-6	news-dev2009	2051	41K	43K
dev-7	test2006	2000	49K	54K
dev-8	test2007	2000	49K	54K
dev-9	test2008	2000	50K	54K

Table 2: Development data.

3.3 On the Scale of Development Set

Having 20 different development sets (10 dev- sets and batch- sets), 20 models are correspondingly trained.The decode results on the test set are summarized in Table 3 and Figure 1. The dotted lines are the performances of 10 different development sets on the two test sets, we will see that there is a huge gap between the highest and the lowest score, and there is not an obvious rule to follow. It will bring about unsatisfied results if a poor development set is chosen. The solid lines represents the performances of 10 incremental batch sets on the two test sets, the batch processing still gives a poor performance at the beginning, but the results become better and more stable when the development sets are continuously enlarged. This sort of results suggest that a combined development set may produce reliable results in the worst case. Our primary submission used the combined development set and the results as Table 4.

id	09-dev	09-batch	08-dev	08-batch
0	16.46	16.46	16.38	16.38
1	16.67	16.25	16.66	16.44
2	16.74	16.20	16.94	16.22
3	16.15	16.83	16.18	17.02
4	16.44	16.73	16.64	16.89
5	16.50	16.97	16.75	17.13
6	17.15	17.03	17.67	17.24
7	16.51	17.00	16.34	17.09
8	17.03	16.97	17.15	17.22
9	16.25	16.99	16.24	17.26

Table 3: BLEU scores on the two test sets(*newstest2009* & *news-test2008*), which use two data set sequences(dev- sequence & batch- sequence) to optimize model weights.

de-en	fr-en	es-en
18.90	24.30	26.40

Table 4: BLEU scores of our primary submission.

3.4 On BLEU Score Difference

To compare BLEU score differences between test set and development set, we consider two groups of BLEU score differences, For each development set, dev-i, the BLEU score difference will be computed between b_1 from which adopts itself as the development set and b_2 from which adopts test set as the development set. For the test set, the BLEU score difference will be computed between b'_1 from which adopts each development set, dev-i, as the development set and b'_2 from which adopts itself as the development set.

These two groups of results are illustrated in Figure 2 (the best score of the test set under self tuning, *newstest2009* is 17.91). The dotted lines have the inverse trend with the dotted in Figure 1(because the addition of these two values is constant), and the solid lines have the same trend with the dotted, which means that the good performance is mutual between test set and development sets: if tuning using A set could make a good result over B set, then vice versa.

3.5 On the Similarity between Development Set and Test Set

This experiment is motivated by (Utiyama et al., 2009), where they used BLEU score to measure the similarity of a sentences pair and then extracted sentences similar with those in test set to

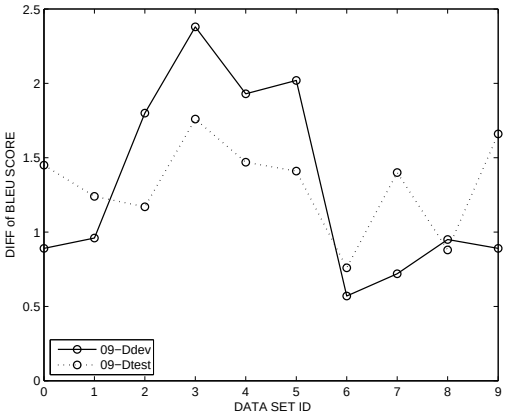

Figure 2: The trend of BLEU score differences

construct a specific tuning set. In our experiment, we will try to measure data set similarity instead. Given two sets of sentences, one is called as candidate(cnd) set and the other reference(ref) set. For any cnd sentence, we let the whole ref set to be its reference and then multi-references BLEU score is computed for cnd set. There comes a problem that the sentence penalty will be constant for any cnd sentence, we turn to calculate the average length of whose sentences which have common n-gram with the given cnd sentence.

Now we may define three measures. The measure which uses dev- and batch- sets as cnd sets and *news-test2009* set as ref set is defined as precision-BLEU , and the measure which uses the above sets on the contrary way is defined as recall-BLEU. Then F1-BLEU is defined as the harmonic mean of precision-BLEU and recall-BLEU. These results are illustrated in Figure 3. From the figure, we find that F1-BLEU plays an important role to predict the goodness of a development set, F1-BLEU scores of batch- sets have an ascending curve and batch data set sequence will cause a stable good test performance, the point on dev- sets which has high F1-BLEU(eg, dev-0,4,5) would also has a good test performance.

3.6 Related Work

The special challenge of the WMT shared task is domain adaptation, which is a hot topic in recent years and more relative to our experiments. Many existing works are about this topic (Koehn and Schroeder, 2007; Nakov, 2008; Nakov and Ng, 2009; Paul et al., 2009; Haque et al., 2009). However, most of previous works focus on language

model, translation phrase table, lexicons model and factored translation model, few of them pay attention to the domain adaptation on the development set. For future work we consider to use some machine learning approaches to select sentences in development sets more relevant with the test set in order to further improve translation performance.

4 Conclusion

In this paper, we present our machine translation system for the WMT10 shared task and perform an empirical study on the development set selection. According to our experimental results, Choosing different development sets would play an important role for translation performance. We find that a development set with higher F1-BLEU yields better and more stable results.

References

Nicola Bertoldi, Barry Haddow, and Jean Baptiste Fouet. 2009. Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91:7–16.

George Foster and Roland Kuhn. 2009. Stabilizing minimum error rate training. In *Proceedings of the 4th Workshop on Statistical Machine Translation(WMT)*, Boulder, Colorado, USA.

Rejwanul Haque, Sudip Kumar Naskar, Josef Van Genabith, and Andy Way. 2009. Experiments on Domain Adaptation for EnglishнHindi SMT. In *7th International Conference on Natural Language Processing(ICNLP)*, Hyderabad, India.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation(WMT)*, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, 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 the 45th Annual Meeting of the Association for Computational Linguistics(ACL)*, Prague, Czech Republic.

Preslav Nakov and Hwee Tou Ng. 2009. NUS at WMT09: domain adaptation experiments for English-Spanish machine translation of news commentary text. In *Proceedings of the 4th Workshop on Statistical Machine Translation(WMT)*, Singapore.

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the 3rd Workshop on Statistical Machine Translation(WMT)*, Columbus, Ohio, USA.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL)*, Philadelphia, Pennsylvanian, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics(ACL)*, Sapporo, Japan.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2009. NICT@ WMT09: model adaptation and transliteration for Spanish-English SMT. In *Proceedings of the 4th Workshop on Statistical Machine Translation(WMT)*, Singapore.

Andreas Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing(ICSLP)*, Denver, Colorado, USA.

Masao Utiyama, Hirofumi Yamamoto, and Eiichiro Sumita. 2009. Two methods for stabilizing MERT: NICT at IWSLT 2009. In *Proceedings of International Workshop on Spoken Language Translation(IWSLT)*, Tokyo, Japan.

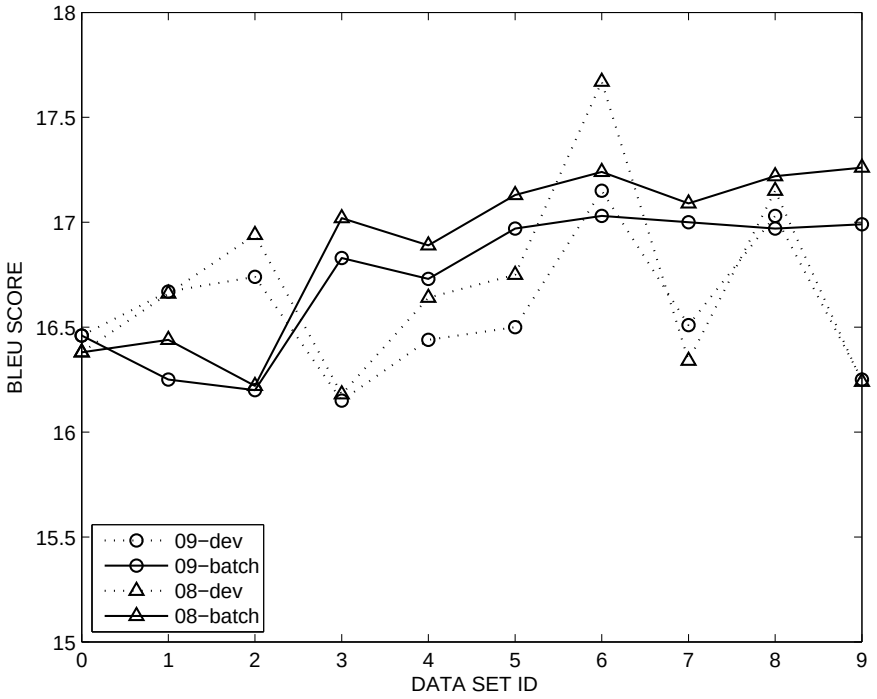

Figure 1: The BLEU score trend in Tabel 3, we will see that the batch lines output a stable and good performance.

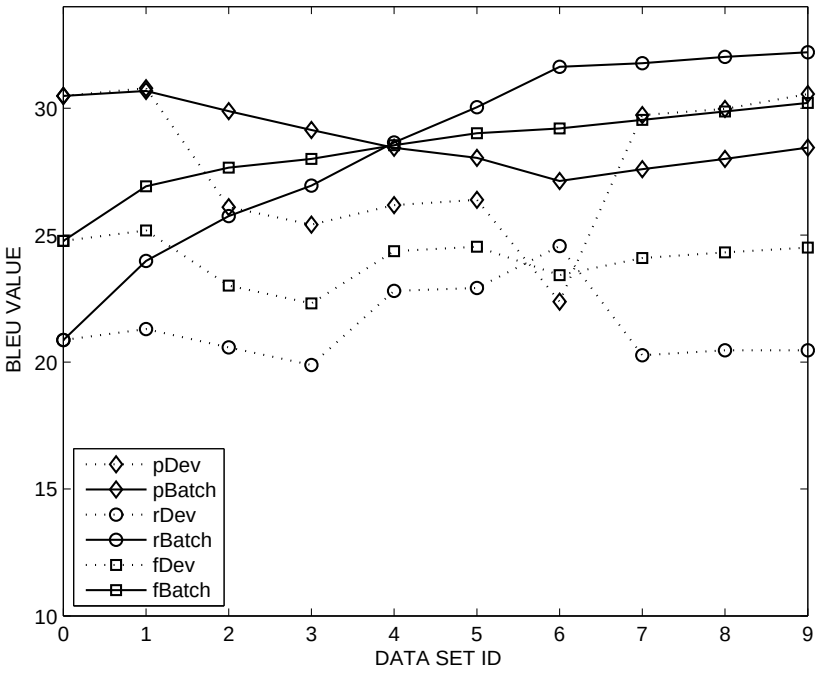

Figure 3: The precision(p), recall(r) and F1(f) BLEU score on the dev(Dev) and batch(Batch) sets based on the comparison with *news-test2009* set.