JU-USAAR: A Domain Adaptive MT System

Koushik Pahari[1], Alapan Kuila[2], Santanu Pal[3], Sudip Kumar Naskar[2], Sivaji Bandyopadhyay[2], Josef van Genabith[3,4]
[1] Indian Institute of Engineering Science and Technology, Shibpur, India
[2]Jadavpur University, Kolkata, India
[3]Universität des Saarlandes, Saarbrücken, Germany
[4]German Research Center for Artificial Intelligence (DFKI), Germany
{pahari.koushik,alapan.cse}@gmail.com,
sudip.naskar@jdvu.ac.in, sivaji_cse_ju@yahoo.com
{santanu.pal, josef.vangenabith}@uni-saarland.de

Abstract

This paper presents the JU-USAAR English–German domain adaptive machine translation (MT) system submitted to the IT domain translation task organized in WMT-2016 . Our system brings improvements over the in-domain baseline system by incorporating out-domain knowledge. We applied two methodologies to accelerate the performance of our in-domain MT system: (i) additional training material extraction from out-domain data using data selection method, and (ii) language model and translation model adaptation through interpolation. Our primary submission obtained a BLEU score of 34.5 (14.5 absolute and 72.5% relative improvements over baseline) and a TER score of 54.0 (14.7 absolute and 21.4% relative improvements over baseline).

1 Introduction

Statistical Machine Translation (SMT) is the currently dominant MT technology. The underlying statistical models in SMT always tend to closely approximate the empirical distributions of the bilingual training data and monolingual target-language text. However, the performance of SMT systems quickly degrades when testing conditions deviate from training conditions. In order to achieve optimal performance, an SMT system should be trained on data from the same domain. Now-a-days domain adaptation has gained interest in SMT to cope with this performance drop. The basic aim of domain adaptation is to maintain the identity of the in-domain data while using the best of the out-domain data. However, large amount of additional out-domain data may bias the resultant distribution towards the out-domain. In practice, it is often difficult to obtain sufficient amount of in-domain parallel data to train a system which can provide good performance in a specific domain. The performance of an in-domain model can be improved by selecting a subset from the out-domain data which is very similar to the in-domain data (Matsoukas et al., 2009; Moore and Lewis, 2010), or by re-weighting the probability distributions (Foster et al., 2006; Sennrich et al., 2013) in favor of the in-domain data.

In this task, the information technology (IT) domain English–German parallel corpus released in the WMT-2016 IT-domain shared task serves as the in-domain data and the Europarl, News and Common Crawl English–German parallel corpus released in the Translation Task are treated as out-domain data.

In this paper we describe the joint submission of Jadavpur University (JU) and Saarland University (USAAR) English–German machine translation (MT) system (JU-USAAR) to the shared task on IT domain translation organized in WMT-2016. In our approach we initially applied data selection method where we directly measured cross entropy for the source side of the text; successively we applied Moore and Lewis (2010) method of data selection and ranked the out-domain bilingual parallel data according to cross entropy difference. Finally, we built domain specific language models on both in-domain and selected out-domain target language monolingual corpus, linearly interpolate them choosing weights that minimize perplexity on a held out in-domain development set. In addition, we also interpolated the translation models trained on the in-domain and selected out-domain parallel corpora. However, instead of using bilingual cross-entropy difference, we applied bilingual cross-perplexity difference to model our data selection process.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 442–448,
Berlin, Germany, August 11-12, 2016.

2 Related Work

Koehn (2004; Koehn (2005) first proposed domain adaptation in SMT by integrating terminological lexicons in the translation model, as a result of which there was a significant reduction in word error rate (WER). Over the last decade, many researchers (Foster and Kuhn, 2007; Duh et al., 2010; Banerjee et al., 2011; Bisazza and Federico, 2012; Sennrich, 2012; Sennrich et al., 2013; Haddow and Koehn, 2012) investigated the problem of combining multi-domain datasets.

To construct a good domain-specific language model, sentences which are similar to the target domain should be included (Sethy et al., 2006) in the monolingual target language corupus on which the language model is trained. Lü et al. (2007) identified those sentences using the tf/idf method and they increased the count of such sentences.

Domain adaptation in MT have been explored in many different directions, ranging from adapating language models and translation models to alignment adaptation approach to improve domain-specific word alignment.

Koehn et al. (2007) used multiple decoding paths for combining multiple domain-specific translation tables in the state-of-the-art PB-SMT decoder MOSES. Banerjee et al. (2013) combined an in-domain model (translation and reordering model) with an out-of-domain model into MOSES and they derived log-linear features to distinguish between phrases of multiple domains by applying the data-source indicator features and showed modest improvement in translation quality.

Bach et al. (2008) suggested that sentences may be weighted by how much it matches with the target domain. A comparison among different domain adaptation methods for different subject matters in patent translation was carried out by (Ceauşfu et al., 2011) which led to a small gain over the baseline.

In order to select supplementary out-of- domain data relevant to the target domain, a variety of criteria have been explored ranging from information retrieval techniques to perplexity on in-domain datasets. Banerjee et al. (2011) proposed a prediction based data selection technique using an incremental translation model merging approach.

3 System Description

3.1 Data selection Approach

Among the different approaches proposed for data selection, the two most popular and successful methodologies are based on monolingual cross-entropy difference (Moore and Lewis, 2010) and bilingual cross-entropy difference (Axelrod et al., 2011). The data selection approach taken in the present work is also motivated by the bilingual cross-entropy difference (Axelrod et al., 2011) based data selection. However, instead of using bilingual cross-entropy difference, we applied bilingual cross-perplexity difference to model our data selection process. The difference in cross-entropy is computed on two language models (LM); the domain-specific LM is estimated from the entire in-domain corpus (lm_{in}) and the second LM (lm_o) is estimated from the out-domain corpus. Mathematically, the cross-entropy $H(P_{lm})$ of language model probability P_{lm} is defined as in Equation 1 considering a k-gram language model.

$$H(P_{lm}) = -\frac{1}{N}\sum_{i=1}^{N} \log P_{lm}(w_i|w_{i-k+1}...w_{i-1}) \quad (1)$$

We calculated perplexity ($PP = 2^H$) of individual sentences of out-domain with respect to in-domain LM and out-domain LM for both source (sl) and target (tl) language.

The score, i.e., sum of the two cross-perplexity differences, for the j^{th} sentence pair $[s_j - t_j]$ is calculated based on Equation 2.

$$score = |PP_{in_{sl}}(s_j) - PP_{o_{sl}}(s_j)| + |PP_{in_{tl}}(t_j) - PP_{o_{tl}}(t_j)| \quad (2)$$

Subsequently, sentence pairs $[s - t]$ from the out-domain corpus (o) are ranked based on this score.

3.2 Interpolation Approach

To combine multiple translation and language models, a common approach is to linearly interpolate them. The language model interpolation weights are automatically learnt by minimizing the perplexity on the development set. For interpolating the translation models, we use moses training pipeline which selects the interpolation weights that optimizes performance on the development set. These weights are subsequently used

to combine the individual feature values for every phrase pair from two different phrase-tables (i.e., in-domain phrase table $p_{in}(e|f)$ and out-domain phrase table $p_o(e|f)$) using the formula in Equation 3 where f and e are source and target phrases respectively and the value of λ ranges between 0 and 1.

$$p(f|e) = \lambda \times p_{in}(f|e) + (1 - \lambda) \times p_o(f|e) \quad (3)$$

4 Experiments and Results

We first accumulate all the domain specific corpus and clean them. We also use out of domain data to accelerate the performance of the in-domain MT system. The following subsections describe the datasets used for the experiments, detailed experimental settings and systematic evaluation on both the development set and test set.

4.1 Datasets

In-domain Data: The detailed statistics of in-domain data is reported in Table 1. We considered all the data provided by the WMT-2016 organizers for the IT translation task. We combined all data and performed cleaning in two steps: (i) Cleaning-1: following the cleaning process described in (Pal et al., 2015), and (ii) Cleaning 2: using the Moses (Koehn et al., 2007) corpus cleaning scripts with minimum and maximum number of tokens set to 1 and 80 respectively. Additionally, 1000 sentences are used for development set ('Batch 1' in Table 3) and anther 1000 sentences are used for development test set ('Batch2' in Table 3).

Out-domain Data: We utilized all the parallel training data provided by the WMT-2016 shared task organizers for the English-German translation task. The out of domain training data includes Europarl, News Commentary and Common Crawl. this corpus is noisy and contains some non-German, as well as, non-English words and sentences. Therefore, we applied a language identifier (Shuyo, 2010) on both bilingual English–German parallel data and monolingual German corpora. We discarded those parallel sentences from the bilingual training data which were detected as belonging to some different languages by the language identifier. The same method was also applied to the monolingual data. Successively, the corpus cleaning process was carried out first by calculating the global mean ratio of the number of characters in a source sentence to that in the corresponding target sentence and then filtering out sentence pairs that exceed or fall below 20% of the global ratio (Tan and Pal, 2014). Tokenization and punctuation normalization were performed using Moses scripts. In the final step of cleaning, we filtered the parallel training data on maximum allowable sentence length of 80 and sentence length ratio of 1:2 (either direction). Approximately 36% sentences were removed from the total training data during the cleaning process. Table 2 shows the out-domain data statistics after filtering.

4.2 Experimental Settings

We used the standard log-linear PB-SMT model for our experiments. All the experiments were carried out using a maximum phrase length of 7 for the translation model and 5-gram language models. The other experimental settings involved word alignment model between EN–DE trained with Berkeley Aligner (Liang et al., 2006). The phrase-extraction heuristics of (Koehn et al., 2003) were used to build the phrase-based SMT systems. The reordering model was trained with the hierarchical, monotone, swap, left to right bidirectional (hier-mslr-bidirectional) (Galley and Manning, 2008) method and conditioned on both the source and target languages. The 5-gram language models were built using KenLM (Heafield, 2011). Phrase pairs that occur only once in the training data are assigned an unduly high probability mass (i.e., 1). To alleviate this shortcoming, we performed smoothing of the phrase table using the Good-Turing smoothing technique (Foster et al., 2006). System tuning was carried out using Minimum Error Rate Training (MERT) (Och, 2003) on a held out development set (Batch1 in Table 3) of size 1,000 sentences provided by the WMT-2016 task organizers. After the parameters were tuned, decoding was carried out on the held out development test set (Batch2 in Table 3) as well as test set released by the shared task organizers. We evaluated the systems using three well known automatic MT evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and TER (Snover et al., 2006). The evaluation results of our baseline systems trained on in-domain and out-domain data are reported in Table 3.

Data Source		Sentences	Tokens	
			EN	DE
Localization	-	157,414	860,169	814,863
IT Term	-	23,136	52,201	45,773
Technical documentation	Liboffice	95,997	794,498	760,444
	Drupal	4,682	41,081	41,081
	Ubuntu	6,320	120,274	113,792
	Chromium	6,306	38,278	37,631
	Undoc	167,627	5,105,968	4,949,335
Total	-	461,479	7,012,469	6,762,919
Cleaning-1	-	456,042	9,105,378	8,958,348
Cleaning-2	-	440,780	7,553,659	7,426,095

Table 1: In-domain data statistics, Cleaning-1: tokenization and cleaning (Pal et al., 2015) and Cleaning-2 is MOSES cleaner with minimum token is set to 1 and maximum 80

Data	Sentences	Tokens	
		EN	DE
Europarl and news	1,623,546	36,050,888	34,564,547
Common crawl	1,811,826	37,456,978	35,172,840
Total	3,435,372	73,507,866	69,737,387

Table 2: Out-domain cleaned data statistics

5 Result and Analysis

We have taken various attempts to enhance the quality of translation for the English–German IT domain translation task.

Figure 1 shows how data selection method helps to enhance the in-domain baseline system by incrementally adding a subset of data from the out-domain corpus as additional training material.

We applied bilingual cross-perplexity difference based method (cf. Section 3.1) to rank the out-domain sentences according to their proximity to the in-domain data from which we incrementally select top ranking sentence pairs and add them as additional training material to our in-domain training set. We trained the incremental in-domain PB-SMT models in an iterative manner for each incremental batch size of 100K top ranked additional parallel data from the remaining 'ranked' out-domain data. The iterative process is stopped when the learning curve falls down in two successive iterations. BLEU is considered as the objective function for the learning curve experiment. Finally, we selected 400K sentence pairs as additional training material from the entire out-domain data as it provided the optimum result in BLEU on the development test set. The rest of our experiments are carried out with this 400K additional training data. Therefore, our submitted JU-USAAR system is built on 440,780 in-domain training data, as well as 400K additional training data selected from the out-domain parallel corpus.

We made use of the out-domain data selected by the data selection method (Moore and Lewis, 2010; Axelrod et al., 2011) using simple merging as well as interpolation technique (Sennrich, 2012).

Linear interpolation with instant weighting (Sennrich, 2012) was used for interpolating the translation and language models.

Our baseline system was trained on the in-domain English–German parallel corpus containing 440,780 sentence pairs. As reported in Table 4, the baseline system obtained a BLEU score of 20 and TER of 68.7 on the test set. We developed two different systems.

System1: System1 is trained on 440,780 in-domain training data combined with additional 400K parallel sentences selected from the out-domain dataset. This system produced a BLEU score of 31.9 and a TER of 66.6 on the test set which are far better than the baseline scores.

System2: System2 uses exactly the same amount of training data as System1, however, in this case instead of simply merging the two datasets (440,780 in-domain and 400K selected out-domain sentence pairs) separate translation

Data		BLEU	METEOR	TER
Out-domain	Batch1	18.47	24.03	63.18
	Batch2	16.54	24.04	60.33
In-domain	Batch1	26.12	28.48	59.18
	Batch2	30.76	32.67	48.66

Table 3: Experiment result of Baseline system trained on in-domain and out-domain data respectively

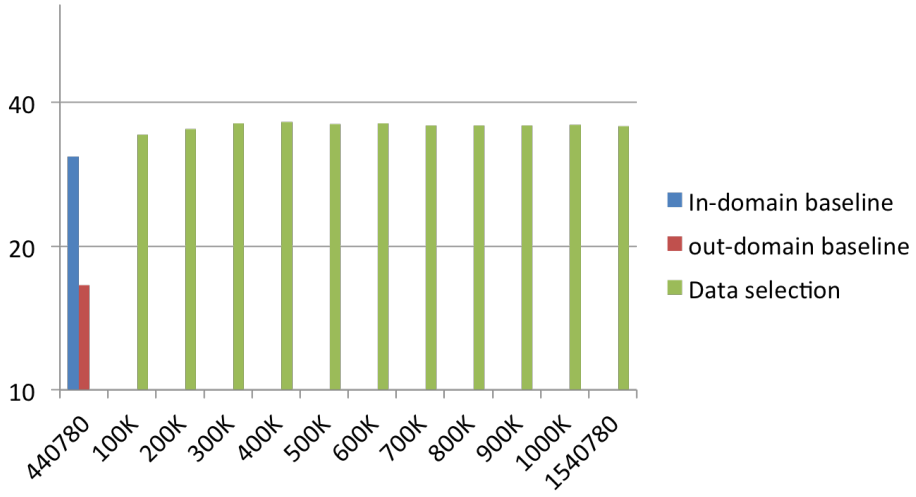

Figure 1: Learning curve experiments on BLEU by incremental data selection of 100K batch size from out-domain data

Systems	BLEU	BLEU (cased)	TER
Baseline	20.0	18.7	68.7
System1	31.9	29.4	66.6
System2	**34.5**	**33.7**	**54.0**

Table 4: Systematic evaluation on test set

models and language models are built on each dataset and they are interpolated based on instant weighting. Before decoding we forced the decoder to avoid translation of URLs. System2 resulted in 34.5 BLEU (14.5 absolute and 72.5% relative improvements over baseline) and 54.0 TER (14.7 absolute and 21.4% relative improvements over baseline) scores. System2 represents our primary submission.

6 Conclusions and Future Work

The JU-USAAR system employs two techniques for improving the performance of MT in the English–German translation task for the IT domain. We used bilingual cross-perplexity difference based data selection method and carried out learning curve experiments to identify additional "in-domain like" training material from the out-domain dataset. We made use of the selected additional training data using both simple merging and interpolation. Simple merging yielded in significant improvements over the baseline while linear interpolation of the translation and language models with instant weighting produced further improvements. Our primary submission (data selection and interpolation based model combination) resulted in 14.5 absolute and 72.5% relative improvements in BLEU and 14.7 absolute and 21.4% relative improvements in TER over the baseline system trained on just the in-domain dataset.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362.

Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Improving word alignment with language model based confidence scores. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 151–154, Columbus, Ohio, June. Association for Computational Linguistics.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2011. Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 285–292. International Association for Machine Translation.

Pratyush Banerjee, Raphael Rubino, Johann Roturier, and Josef van Genabith. 2013. Quality estimation-guided data selection for domain adaptation of SMT. In *Machine Translation Summit XIV*, pages 101–108.

Arianna Bisazza and Marcello Federico. 2012. Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 439–448, Avignon, France, April. Association for Computational Linguistics.

Alexandru Ceauşfu, John Tinsley, Jian Zhang, and Andy Way. 2011. Experiments on domain adaptation for patent machine translation in the PLuTO project. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 21–28.

Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Analysis of Translation Model Adaptation in Statistical Machine Translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 243–250.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135. Association for Computational Linguistics.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 53–61.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856.

Barry Haddow and Philipp Koehn. 2012. Analysing the Effect of Out-of-Domain Data on SMT Systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 175–185, Montreal, Canada, June. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 708–717. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167.

Santanu Pal, Sudip Naskar, and Josef van Genabith. 2015. UdS-Sant: English–German Hybrid Machine Translation System. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 152–157, Lisbon, Portugal, September. Association for Computational Linguistics.

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, pages 311–318.

Rico Sennrich, Holger Schwenk, and Walid Aransa. 2013. A Multi-Domain Translation Model Framework for Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 832–840.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France, April. Association for Computational Linguistics.

Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan. 2006. Selecting relevant text subsets from web-data for building topic specific language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 145–148, New York City, USA, June. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language Detection Library for Java.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Liling Tan and Santanu Pal. 2014. Manawi: Using Multi-word Expressions and Named Entities to Improve Machine Translation. In *Proceedings of Ninth Workshop on Statistical Machine Translation*.