Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy

Learning Machine Translation from In-domain and Out-of-domain Data

Marco Turchi
European Commission JRC,
IPSC - GlobeSec
Via Fermi 2749,
21020 Ispra (VA), Italy
marco.turchi@jrc.ec.europa.eu

Cyril Goutte
Interactive Language Tech.,
National Research Council Canada,
283 Boulevard Alexandre-Taché,
Gatineau QC J8X3X7, Canada
Cyril.Goutte@nrc.ca

Nello Cristianini
Intelligent Systems Lab.,
University of Bristol,
MVB, Woodland Rd,
BS8-1 UB, Bristol, UK
Nello@support-vector.net

Abstract

The performance of Phrase-Based Statistical Machine Translation (PBSMT) systems mostly depends on training data. Many papers have investigated how to create new resources in order to increase the size of the training corpus in an attempt to improve PBSMT performance. In this work, we analyse and characterize the way in which the in-domain and out-of-domain performance of PBSMT is impacted when the amount of training data increases. Two different PBSMT systems, Moses and Portage, two of the largest parallel corpora, Giga (French-English) and UN (Chinese-English) datasets and several in- and out-of-domain test sets were used to build high quality learning curves showing consistent logarithmic growth in performance. These results are stable across language pairs, PBSMT systems and domains. We also analyse the respective impact of additional training data for estimating the language and translation models. Our proposed model approximates learning curves very well and indicates the translation model contributes about 30% more to the performance gain than the language model.

1 Introduction

With the growing availability of bilingual parallel corpora, the past two decades saw the development and widespread adoption of *statistical* machine translation (SMT) models. Given a source ("foreign") language sentence $\mathbf{f}$ and a target ("english") language translation $\mathbf{e}$, the relationship between $\mathbf{e}$ and $\mathbf{f}$ is modelled using a statistical or probabilistic model which is estimated from a large amount of textual data, comprising bilingual and monolingual corpora. The most popular class of SMT systems is Phrase-Based SMT (PBSMT, (Koehn et al., 2003)).

In this paper, we are concerned with analyzing and characterizing the way in which the performance of PBSMT models evolves with increasing amounts of training data. In the SMT community, it is a common belief that learning curves follow logarithmic laws. However, there are few large-scale systematic analyses of the growth rate of the PBSMT performance. Early work (Al-Onaizan et al., 1999) used a relatively small training set and perplexity as evaluation metric. (Koehn et al., 2003) and (Suresh, 2010) show that BLEU score has a log-linear dependency with training corpus size, but this is limited to 350k training sentence pairs. Learning curves were also presented in order to motivate the use of active learning for MT (Bloodgood and Callison-Burch, 2010; Haffari et al., 2011). They attempt to address the challenge of "diminishing returns" in learning MT, although this is again done with small training corpora (<90k sentence pairs), and, on a log-scale, performance seems again to increase linearly. (Brants et al., 2007) produced a large-scale study, but focused on the language model training only, with billions of (monolingual) tokens.

The first complete and systematic analysis of PBSMT learning curves was obtained by (Turchi et al., 2008) using the Spanish-English Europarl, and recently extended to larger training data and more systems by (Turchi et al., 2011). In their work, accurate learning curves obtained over a large range of data sizes confirm that performance grows linearly in the log domain.

The reason why relatively few systematic studies have been reported may be that producing accurate learning curves up to large data sizes with state-of-the-art systems requires the use of high performance computing in a carefully set up environment. This may seem dispensable when typical SMT research is usually focused on maximizing the performance that can be extracted from a given data set, rather than analysing how this performance evolves. However, we believe that the analysis and quantification of the way machine translation systems learn from data are important steps to identify critical situations which affect the overall translation performance. We also wish to characterize PBSMT performance up to data sizes more typical of current large-scale bilingual corpora.

In the following we pursue three purposes:

1. We confirm, in a systematic way, previous findings that PBSMT performance gains constant improvements for each doubling of the data. This holds across systems, language pairs and over a large range of data sizes.

2. We show that, somewhat surprisingly, this extends to out-of-domain data, although the growth is weaker in that case.

3. We analyse and quantify the relative importance of training data in language and translation model training, and show that the latter contributes about 30% more to the gains in performance.

In contrast with previous work, we build our learning curves using two of the largest available parallel training sets: the French-English Giga corpus and the Chinese-English UN corpus. In addition to being large corpora, these also cover two very distinct language pairs. We also use two PBSMT systems: Moses (Koehn et al., 2007) and Portage (Ueffing et al., 2007). Finally, we analyze in- and out-of-domain learning curves in order to better understand and investigate the growth rate.

The following section gives a quick overview of the models and systems we used in our experiments. We then briefly describe the experimental settings and data we used. Section 4 shows and analyzes the learning curves we obtained on French-English and on Chinese-English, and section 5 presents our results on the relative importance of LM and TM in the performance increase.

2 Translation Models and Systems

The standard phrase-based machine translation systems which we analyse here rely on a log-linear model and a set of baseline features functions. Translations of a source sentence $\mathbf{f}$ is obtained by:

$$\widehat{\mathrm{e}}(\mathbf{f}) = \operatorname*{argmax}_{\mathbf{e}} \sum_{i} \lambda_i h_i(\mathbf{e}, \mathbf{f}).$$

where the $h_i(\mathbf{e}, \mathbf{a}, \mathbf{f})$ are *feature functions* involving both the source and target sentences, and the λ_i are the weights of those feature functions. Typical examples of feature functions that compose a basic phrase-based MT system are:

- phrase translation feature, e.g.: $h_T(\mathbf{e}, \mathbf{f})$ = $\sum_k \log p(f_k|e_k)$;

- language model feature, e.g.: $h_L(\mathbf{e}, \mathbf{f})$ = $\sum_j \log p(w_j|w_{j-1}, \ldots w_1)$

- distortion feature, e.g.: $h_D(\mathbf{e}, \mathbf{f})$ = $\sum_k \|\text{start}(f_k)$ - $\text{end}(f_{k-1}) - 1\|$

- Word penalty and/or phrase penalty features.

where e_k and f_k are contiguous subsequences of words in the source and target sentences and w_j are target words.

Parameter estimation is crucial for both the translation and language model features. Conditional probabilities are estimated from a large training corpus using empirical counts and various smoothing strategies. In addition, the weights λ_i are also estimated from a (usually disjoint) corpus of source and target sentence pairs. The size and composition of the training data will therefore have an influence on the quality of the predictions $\widehat{\mathrm{e}}$ through the estimation of both the log-linear parameters and the feature functions.

Note that alternate models such as hierarchical (Chiang, 2007) or syntax based (Zollman and Venugopal, 2006) have been developed and could also be studied. However their use on the large scale necessary for creating accurate learning curves would require solving a number of practical issues and we focus instead on the straight PBSMT approach, which has been shown in recent MT evaluations (Callison-Burch et al., 2009; Callison-Burch et al., 2011) to offer competitive performance.

2.1 PBSMT Software

Several software packages are available for training PBSMT systems. In this work, we use Moses (Koehn et al., 2007) and Portage (Ueffing et al., 2007), two state-of-the-art systems capable of learning translation tables, language models and decoding parameters from one or several parallel corpora. Moses is a complete open-source phrase-based translation toolkit available for academic purposes, while Portage is a similar package, available to partners of the National Research Council Canada.

Given a parallel training corpus, both perform basic preprocessing (tokenization, lowercasing, etc.) if necessary, and build the various components of the model. Both use standard external tools for training the language model, such as SRILM (Stolcke, 2002). Moses uses GIZA++ (Och and Ney, 2003) for word alignments, while Portage uses an in-house IBM model and HMM implementation. The parameters of the log-linear models are tuned using minimum error rate training (MERT, (Och, 2003)).

Earlier experiments performed on the Europarl corpus with both systems showed (Turchi et al., 2011) that despite small differences in observed performance, both systems produce very similar learning curves.

3 Experimental Setting

3.1 Corpora

We experiment with large corpora in two language pairs: French-English and Chinese-English.

For French-English, we use the Giga corpus (Callison-Burch et al., 2009) to provide the training, development and one in-domain test set. As out-of-domain test set, we use two different samples from the EMEA corpus (Tiedemann, 2009), which contains parallel documents from the European Medicines Agency, and two News test sets from the 2009 (Callison-Burch et al., 2009) and 2011 (Callison-Burch et al., 2011) editions of the Workshop on Statistical Machine Translation, containing news articles drawn from a variety of sources and languages in different periods and translated by human translators.

For Chinese-English, we use various parallel corpora obtained from the Linguistic Data Consortium for the NIST evaluations. The training, development and in-domain test sets are sampled from the United Nations corpus (UN, LDC2004E12). As out-of-domain test sets, we used a sample from the Hong-Kong Hansard (HKH, LDC2000T50), a corpus of Chinese News translations (LDC2005T06) and the NIST 2008 Chinese evaluation set (LDC2009E09). Basic statistics are given in Table 1.

src	**Training Set**	**Sentences**	**Words**
fr	Giga	18.276 M	482,744k
ch	UN	4.968 M	163,960k
	Dev. Set		
fr	Giga	1,000	62k
ch	UN	2,000	32k
	Test Set		
fr	Giga	3,000	109k
fr	Emea	3,051	45.4k
fr	Emea2	3,051	46.7k
fr	News 2009	2,489	70.7k
fr	News 2011	3,030	85.1k
ch	UN	10,000	332k
ch	HKH	5,000	153k
ch	NIST	1,357	42k
ch	News	10,317	320k

Table 1: Number of sentences and words (source side) for the training, dev and various test sets.

In order to analyse the way MT performance evolves with increasing data, we subsample (without replacement) the training sets at various sizes, averaging performance (estimated by BLEU, cf. section 3.3) over several samples. Learning curves are then obtained by plotting the average BLEU score, with error bars, as training data sizes increases. The relatively large amount of sentences in most test sets will allow us to reduce the uncertainty on the estimated test error, therefore producing smaller error bars.

For the French-English data, we followed the methodology proposed in (Turchi et al., 2008) and sampled 20 different sizes representing 5%, 10%, etc. of the original training corpus. Due to the large size of the corpus, only three random subsets are sampled at each size. For the Chinese-English dataset, we sampled at sizes corresponding to one half, one quarter, etc. down to $1/512^{th}$ ($\sim 0.2\%$) of the full size. At each size we produced 10 random samples. Each random subsample produces a model (cf. below) which is used to translate the various test sets. The learning curves will therefore cover the range from around 900 thousand to 18.3 million sentences for French-English, and from around 10 thousand to 5 million sentences for Chinese-English.

Note that the corpora, in addition to differing in language pair, also differ in domain and homogeneity. The UN data contains only material from the United Nations, covering a wide range of themes, but fairly homogeneous in terms of style and genre. The Giga corpus, on the other hand, was obtained through a targeted web crawl of bilingual web sites from the Canadian government, the European Union, the United Nations, and other international organizations. In addition to covering a wide range of themes, they also contain documents with different styles and genres. Moreover, we estimated in an independent study that the rate of misaligned sentence pairs in the Giga corpus is as high as 13%.

The choice of source languages is driven by the desire to analyze two very different languages and by the scarcity of large publicly available bilingual corpora, especially outside European languages. UN data is also available in Russian or Arabic, but by definition would be the same domain and homogeneity as the Chinese-English corpus.

3.2 PBSMT System Training

For both systems, Portage and Moses, we used the basic configuration and features: phrase extraction is done by aligning the corpus at the word level (IBM models 1, 2, 3 and 4 for Moses, HMM and IBM2 models for Portage), the parameters of the log-linear model are set using an implementation of Och's MERT algorithm (Och, 2003), n-gram language modelling uses Kneser-Ney smoothing (3-gram using SRILM for Moses and 4-gram for Portage) and the maximum phrase length is 7 tokens. In Portage, phrase pairs were filtered so that the top 30 translations for each source phrase were retained. In both systems, the MERT algorithm was independently run on each sampled training set for each experiment.

Note that we expect that there will be differences in the quality of the translation depending on the source language. However, we are not so much interested in the actual translation performance as in the way this performance evolves with increasing data under various conditions.

3.3 Evaluation metrics

We report performance in terms of BLEU score (Papineni et al., 2001), the well accepted and widespread automatic MT metric. We are well aware that maximizing BLEU may neither be necessary for, nor guarantee good translation performance, and that automatic MT metrics may not tell the whole story as far as translation quality is concerned. However, our systematic study aims at characterizing the behaviour of PBSMT systems that are built by maximizing such metrics, and this maximization is part of the learning system we analyze. Deriving learning curves for human evaluations of translation quality would be interesting, but is clearly impractical at his point.

4 Learning Curve Analysis

We now present the results obtained under the general framework outlined above.

We stress that in these experiments, we focus on the growth rate of the learning curves. In particular we are interested in 1) confirming that learning curves have logarithmic growth, and 2) possible differences between domains, languages and systems. A common, but poorly supported belief in PBSMT is that each doubling of the data yields a more or less constant increase in performance. In order to analyze and support this belief, we show all learning curves on a log scale, where we can check if the curve has a linear behaviour.

Note that sampling without replacement results in an increasing overlap between samples as their sizes grow. The size of the error bars therefore decreases as the training set size grows, because the training sets, and therefore the resulting models, are not independent. This must be kept in mind, although we still believe that the presence of error bars helps to better understand the stability of the MT system's performance.

The resulting learning curves are shown in Figures 1 and 2 for the French-English and Chinese-English data, respectively. The plots show the performance, averaged over samples (marks, connected with dotted lines), the error bars (vertical lines) indicating the natural variance in the performance, and a least-squares linear fit of these points (dashed or solid line). It is very clear that the learning curves are almost exactly linear on the log scale in most cases (Chinese-English and most French-English curves). The EMEA 2 and News 2009 curves display a worse fit, but the empirical results are within error bars of the linear fit, showing that the deviation from linearity is not statistically significant. The instability in these last two curves

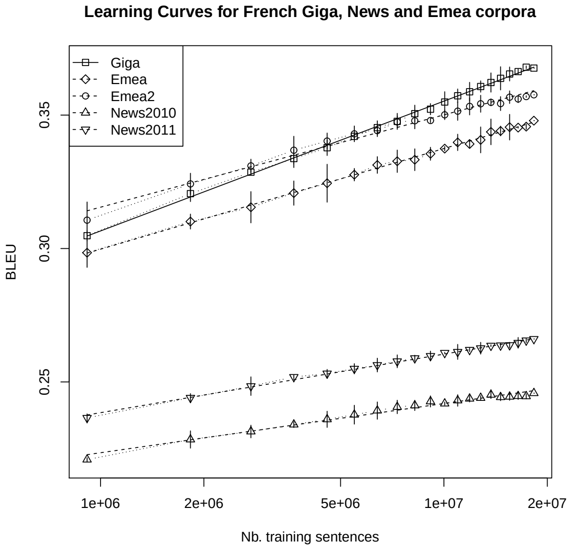

Figure 1: French-English learning curves obtained using the Giga corpus for training Moses on five test sets: one in-domain and four out-of-domain.

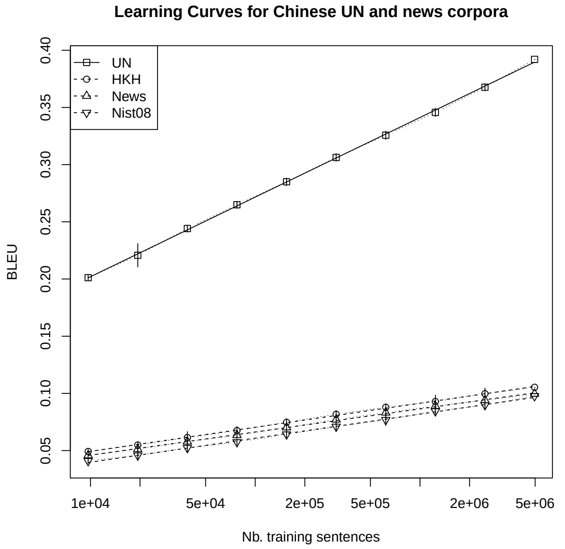

Figure 2: Chinese-English learning curves obtained using the UN corpus for training Portage on four test sets: one in-domain and three out-of-domain.

may actually be due to the fact mentioned earlier that the dependency between the performance estimates increases for large training sizes, which may lead to an increasing bias in the average.

These results confirm the findings of (Turchi et al., 2008) and extend them to more language pairs and much larger data sizes. These experiments supports the following claims:

- The increase in performance for PBSMT systems is essentially constant for each doubling of the data, over a wide range of training data sizes. Note that the growth does not seem to slow down as we near 20M training sentence pairs.
- A corollary of that first claim is that minor, even statistically significant increases in performance due to model "tweaking" are likely to be dwarfed by moderate increases in data sizes. For our Chinese system, for example, a 10% increase in data produces a 0.43 BLEU gain.
- On a linear scale, however, the addition of massive amounts of data from the same domain will result in diminishing improvements ("diminishing returns") in the performance after an initial fast growth (Turchi et al., 2008; Bloodgood and Callison-Burch, 2010).
- Interestingly, the general shape of the learning curves is essentially the same across different language pairs, different PBSMT systems, and also over different sources of test data (in-domain or out-of-domain).
- In particular, although the *performance* on out-of-domain data may greatly suffer (cf. Figure 2), the rate of increase is still linear in the log domain, up to large data sizes.

In order to quantify these findings, we estimate the gain per each doubling of the training set size by fitting a simple linear model on the learning curves in the log domain. For the Chinese-English data, each doubling of the data yields a gain of around 2.1 BLEU points on the in-domain data, and only 0.6 on the out-of-domain test sets. For the French-English data, the BLEU gain per training data doubling is around 1.5 points for the in-domain data, 1.1 for the EMEA test sets and 0.6 for the News test sets.

One may wonder why the out-of-domain EMEA test sets yield such high learning curves. Although the EMEA data comes from a European agency, we have verified that the sentences it contains are not contained in the Giga corpus. However, it turns out that the EMEA data is actually fairly easy to translate. The language is relatively constrained and repetitive, sentences are much shorter (on average ∼15 words against more than 28 for the other corpora), and the number of out-of-vocabulary words much lower than in the other test sets.

By contrast, all out-of-domain learning curves on Chinese-English are much lower than the in-domain curves (we have corroborated this with a dozen different test sets taken from various sources available for NIST evaluations, but omitted here for clarity). We believe this reflects differences between the sources of our training data. The UN corpus covers a number of topics but is very homogeneous and rather limited in genre. By contrast, the Giga corpus contains a wide range of documents covering many themes and genres. As a consequence, any test set that does not come from the UN data is distinctively different and "far" out-of-domain. On the other hand, it is not inconceivable that even for French text that does not come from the same sources, the larger and more diverse Giga corpus provides some measure of overlap in topics and genre.

5 Relative Importance of TM and LM

In the previous Section, experiments have been run using the same training set size for language and translation models. However, there is a large difference in the cost of training data for language and translation models. The former can be trained using monolingual data only while the latter requires bilingual texts. In recent years, several parallel corpora have been produced, e.g. Europarl (Koehn, 2005), JRC Acquis (Steinberger et al., 2006), and others, but they are not comparable to the amount of freely available monolingual data.

(Brants et al., 2007) have shown that performance improves linearly with the log of the number of tokens in the language model training set when this quantity is huge (from billions to trillions of tokens). In this section, we are interested in understanding the trade-off between the training data size used to build language and translation models, as well as in how performance is affected by that difference. We propose a mathematical model to estimate the variation in BLEU score according to the size of the training data used by the language model vs. that use by the translation model. The previous section shows that the overall performance of a PBSMT system grows in the logarithm of the training data size. We therefore modelled this relation in the following way:

$$\begin{aligned} BLEU(d_{LM}, d_{TM}) \ &= \\ \alpha_{LM} * log_2(d_{LM}) \ &+ \ \alpha_{TM} * log_2(d_{TM}) + \epsilon \end{aligned}$$

where d_{LM} is the amount of training data used to build the language model, d_{TM} is the amount of training data used to build the Translation Model. α_{LM} and α_{TM} are weighting factors that identify the contribution of language and translation training data to the BLEU score, and ϵ is the residual. Note that when $d_{LM} = d_{TM}$, we recover a simple logarithmic relationship between performance and data size, as illustrated in the previous section.

In order to evaluate the relation between the amount of training data used to build language and translation models we estimate α_{LM} and α_{TM} from data. We focus on the French-English data, and use the training data subsets at every 10% of the full data size (10%, 20%, etc.), using the same development and test sets as before. One instance of a PBSMT model is learned for each combination of language and translation training data sizes, and we compute the resulting BLEU on the test sets. We estimate the parameters α_{LM} and α_{TM} using multivariate linear regression based on least squares (Draper and Smith, 1981), with the BLEU scores as response variables and the log values of the LM and TM training sizes as explanatory variables. This is done for three French-English test sets: the in-domain Giga, Emea and News 2009. The Emea2 and News 2011 test sets were qualitatively very similar.

We estimated the weighting factors using all the data. The results in Table 2 empirically confirm the common belief that adding data to the translation model is more important than to the language model ($\alpha_{TM} > \alpha_{LM}$). The values of α_{LM} and α_{TM} vary across the test sets, and correspond to an increase of 1 to 1.3 BLEU point per doubling of the training data for the LM and 1.2 to 1.8 BLEU point per doubling for the TM. However, the ratio is rather stable, indicating that the relative importance of the TM w.r.t. the LM is stable across domains. Not surprisingly, the more similar the test set is to the training data, the larger is the BLEU point growth. Our results are qualitatively compatible with the observations reported in a tutorial by (Och, 2005), although the increments in BLEU with each doubling of the training data size are reported 0.5 and 2.5 points for the language and translation models, respectively, in the context of Arabic-English translation. The ratio we observed in our experiments is lot more favourable to the language model.

In order to validate this finding, we performed

Test Set	α_{LM}	α_{TM}	α_{TM}/α_{LM}
Giga	0.0133	0.0182	1.368
Emea	0.0134	0.0168	1.2563
News 2009	0.0097	0.0122	1.2532

Table 2: Empirical estimation of the contributions α_{LM} and α_{TM} of the LM and TM, respectively, (ϵ is smaller than 1×10^{-4}), in BLEU per $\log_2$ in size. Experiments have been performed independently on the three test sets.

two simple experiments where we added a fairly large, 10 million sentence corpus of monolingual data (not included in the Giga corpus) to our LM training data, starting with around 5 million sentence of bilingual data from the Giga corpus. This produced a 1.79 BLEU increase in performance on News 2009 and 1.38 BLEU increase on News 2011, which is roughly consistent with a tripling in LM training data size according to the rate estimated in Table 2 ($0.97 \times \log_2 3 \approx 1.54$).

6 Discussion

Although limited to two language pairs, our results investigate the behaviour of PBSMT as a learning system over a range of different conditions: very different language pairs, in-domain and out-of-domain data, differing level of corpus homogeneity. etc. We emphasize that obtaining systematic and accurate learning curves requires a significant effort, even with an high performance computing architecture (Figure 2 requires translating more than 3 million test sentences with 91 models).

The learning curves obtained here suggest that, on an absolute (linear) scale, performance gains per fixed amount of additional data decrease. The diminishing improvements in performance after an early fast growth was also reported by (Uszkoreit et al., 2010) who mined the Web to extract very large sets of parallel documents. Starting with two corpora (French/Spanish to English) similar in dimensions to the Giga training set and using the News 2009 test sets, they report that adding more than 4,800 M words from a different domain resulted in relative small performance gains ($<$ 2 BLEU points).

On a log-scale, on the other hand, there is no sign that performance gains decrease as we keep doubling the training corpus size, at least up to 20M sentence pairs. Note that although usual MT metrics have natural bounds (0 for error-based metrics such as TER, 1 for BLEU), this has little practical relevance to the results presented here. Indeed, assuming we could extrapolate the very stable growth rates observed here, taking the performance of the out-of-domain HKH test set to where the in-domain UN data starts (for 10k sentence pairs only) would require close to 180 *billion* sentence pairs. For all practical purpose, we would run out of data long before we reached even half of the theoretical maximum BLEU score.

Finally, the analysis of the relative importance of TM and LM estimation shows that the translation model contributes about 30% more to the increase in performance than the language model. Considering the crucial role of the phrase table in the translation process, this contribution is maybe less than one would expected. This means that the massive addition of training data to the language model has a substantial impact in terms of performance, as shown by (Brants et al., 2007). It is interesting that the ratio of α_{TM} and α_{LM} seems stable across different domains. The relation between the translation and language model contribution to the final BLEU score does not change whether we translate in- or out-of-domain data.

7 Conclusion

Using state-of-the-art Phrase-Based Statistical Machine Translation packages and large parallel corpora, we derived very accurate learning curves for a number of language pairs and domains. Our results suggest that performance, as measured by BLEU, increases by a constant factor for each doubling of the data. Although that factor varies depending on corpus and language pair, this result seems consistent over all experimental conditions we tried. Our findings confirm the results reported for example by (Brants et al., 2007) and (Och, 2005), and extend and complete the findings of (Turchi et al., 2008).

We propose a study of how performance is influenced by difference sizes of data used for training the language and translation models. Our model gives more importance to the translation model than the language model every doubling of training data, but we are lot more favourable to the language model compared to other reported results in the literature.

Even if we do not currently provide any result that is immediately actionable to improve current

PBSMT performance, we believe it is important to analyse and quantify the way Machine Translation systems learn. In addition, the markedly different rates of performance increase for in-domain and out-of-domain data may provide a clue to better characterise the suitability of a MT model to translate a given test set. Investigating features that help us differentiate out-of-domain from in-domain data may prove very useful to improve practical performance of PBSMT systems.

References

Y. Al-Onaizan and J. Curin and M. Jahr and K. Knight et al. 1999. *Statistical Machine Translation: Final Report*. JHU 1999 Summer Workshop on Language Engineering, CSLP.

M. Bloodgood and C. Callison-Burch. 2010. *Bucking the trend: large-scale cost-focused active learning for statistical machine translation*. 48th Meeting of the ACL, pp. 854–864.

T. Brants and A. C. Popat and P. Xu and F. J. Och and J. Dean. 2007. *Large Language Models in Machine Translation*. Proc. EMNLP-CoNLL 2007, pp. 858–867.

C. Callison-Burch, and P. Koehn and C. Monz and J. Schroeder. 2009. *Findings of the 2009 Workshop on Statistical Machine Translation*. Fourth Workshop on Statistical Machine Translation, pp. 1–28.

C. Callison-Burch, and P. Koehn and C. Monz and O. Zaidan. 2011. *Findings of the 2011 Workshop on Statistical Machine Translation*. Sixth Workshop on Statistical Machine Translation, pp. 22–64.

D. Chiang. 2007. *Hierarchical Phrase-Based Translation*. Computational Linguistics, 33(2):201–228.

N.R. Draper and H. Smith. 1981. *Applied regression analysis*. Wiley, New York, USA.

G. Haffari and M. Roy and A. Sarkar. 2009. *Active Learning for Statistical Phrase-based Machine Translation*. Proc. HLT-NAACL, pp. 415–423.

P. Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. Proc. MT-Summit X, pp. 79-86.

P. Koehn and H. Hoang and A. Birch and C. Callison-Burch et al. 2007. *Moses: Open source toolkit for statistical machine translation*. 45th Meeting of the ACL demo, pp. 177–180.

P. Koehn and F. J. Och and D. Marcu. 2003. *Statistical phrase-based translation*. Proc. NAACL-HLT, pp. 48–54. Edmonton, Canada.

F. J. Och 2005. *Statistical machine translation: Foundations and recent advances*. Proc. MT-Summit X tutorial.

F. J. Och 2003. *Minimum error rate training in statistical machine translation*. 41st Meeting of the ACL, pp. 160–167.

F. J. Och and H. Ney 2003. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, 29(1): pages 19–51. Sapporo, Japan.

F. J. Och and H. Ney 2002. *Discriminative training and maximum entropy models for statistical machine translation*. 40th Meeting of the ACL, pp. 295–302.

K. Papineni and S. Roukos and T. Ward and W. J. Zhu 2002. *BLEU: a method for automatic evaluation of machine translation*. 40th Meeting of the ACL, pp. 311–318.

A. Stolcke. 2002. *SRILM – An extensible language modeling toolkit*. Intl. Conf. Spoken Language Processing.

R. Steinberger and B. Pouliquen and A. Widiger and C. Ignat and T. Erjavec and D. Tufiş and D. Varga. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. 5th LREC, pp. 2142–2147.

B. Suresh. 2010 *Inclusion of large input corpora in Statistical Machine Translation*. Technical report, Stanford University.

J. Tiedemann. 2009. *News from OPUS—A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. RANLP (vol V), pp. 237-248.

M. Turchi and T. DeBie and N. Cristianini. 2008. *Learning Performance of a Machine Translation System: a Statistical and Computational Analysis.* Third Workshop on Statistical Machine Translation, pp. 35–43.

M. Turchi and T. DeBie and C. Goutte and N. Cristianini. 2012. *Learning to Translate: a statistical and computational analysis.* Advances in Artificial Intelligence, in press.

N. Ueffing and M. Simard and S. Larkin and J. Howard Johnson. 2007. *NRC's PORTAGE system for WMT*. Second Workshop on Statistical Machine Translation, pp. 185–188.

J. Uszkoreit and J.M. Ponte and A.C. Popat and M. Dubiner. 2010. *Large scale parallel document mining for machine translation*. 23rd COLING, pp. 1101–1109.

A. Zollman and A. Venugopal. 2006. *Syntax augmented machine translation via chart parsing.* Proc. NAACL Workshop on Machine Translation.

R. Zens and F. J.Och and H. Ney. 2002. *Phrase-Based Statistical Machine Translation*. Proc. KI '02: Advances in Artificial Intelligence, pp. 18–32.